STAT 3601

Data Analysis

Fall, 2005

Lecture Notes

Contents

1 Preliminaries

Complimentary reading from Rao: Chapters 1-4 and 7.

1.1 Introduction

We first provide some terminology that we will use throughout the course.

TERMINOLOGY: An **experiment** is a planned study where individuals are subjected to **treatments**.

TERMINOLOGY: An **experimental unit** is the unit of material to which a treatment is applied. An experimental unit may be an individual or a collection of individuals.

TERMINOLOGY: **Statistics** is the development and application of theory and methods to the collection (design), analysis, and interpretation of observed information from planned (or unplanned) experiments.

TERMINOLOGY: **Biometry** is the development and application of statistical methods for biological experiments (which are often planned).

TERMINOLOGY: In a statistical problem, the **population** is the entire group of individuals that we want to make some statement about. A **sample** is a part of the population that we actually observe.

TERMINOLOGY: A **variable** is a characteristic (e.g., temperature, age, CD4 count, growth, etc.) that we would like to measure on individuals.

TERMINOLOGY: Measurements of a variable observed from individuals in the sample are called **data**.

TERMINOLOGY: The process of generalising the results in our sample to that of the entire population is known as **statistical inference**.

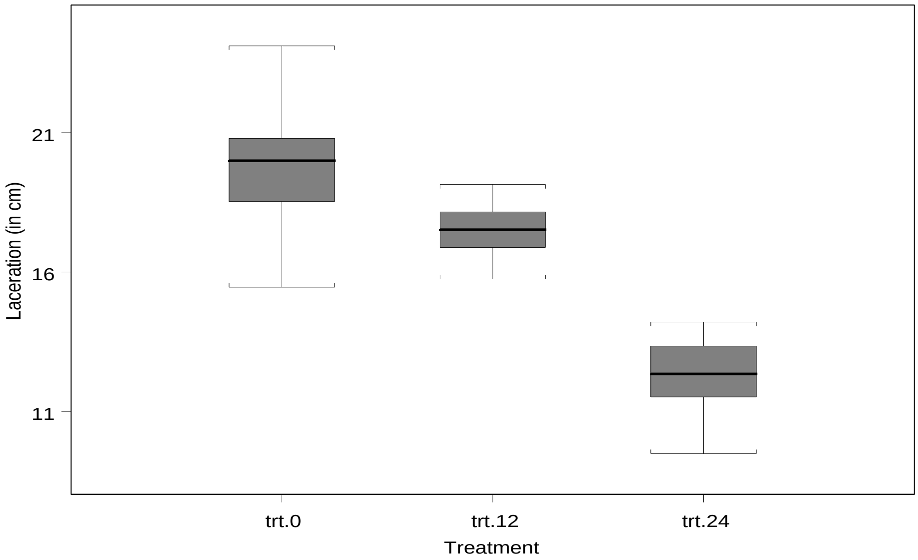

Figure 1.1: *Salmonella experiment: Laceration length for the three treatments.*

Example 1.1. Salmonella bacteria are widespread in human and animal populations; in particular, some serotypes can cause disease in swine. A food scientist wants to see how withholding feed from pigs prior to slaughter can reduce the number and size of gastrointestinal tract lacerations during the actual slaughtering process. This is an important issue since pigs infected with salmonellosis may contaminate the food supply.

- **Experimental units** = pigs
- **Population** = all market-bound pigs, say
- **Sample** = 45 pigs from 3 farms (15 per farm) assigned to three treatments:
  - Treatment 1: no food withheld prior to transport,
  - Treatment 2: food withheld 12 hours prior to transport, and
  - Treatment 3: food withheld 24 hours prior to transport.
- Data were measured on many variables, including body temperature prior to slaughter, weight prior to slaughter, treatment assignment, the farm from which each pig

originated, number of lacerations recorded, and size of laceration (cm). Boxplots of the lacerations lengths (by treatment) are in Figure 1.1.

QUESTIONS OF INTEREST:

- How should we assign pigs to one of the three treatments?
- What are the **sources of variation**? That is, what systematic components might affect laceration size or number of lacerations?
- Why would one want to use animals from three farms? Why might body temperature or prior weight be of interest?

SOME GENERAL COMMENTS:

- In agricultural, medical, and other biological applications, the most common objective is to compare two or more treatments. In light of this, we will often talk about statistical inference in the context of comparing treatments in an experimental setting. For example, in the salmonella experiment, one goal is to compare the three withholding times (0 hours, 12 hours, and 24 hours).
- Since populations are usually large, the sample we observe is just one of many possible samples that are possible to observe. That is, samples may be similar, but they are by no means identical. Because of this, *there will always be a degree of uncertainty about the decisions that we make concerning the population of interest.*
- A main objective of this course is to learn how to design controlled experiments and how to analyse data from these experiments. We would like to make conclusions based on the data we observe, and, of course, we would like our conclusions to apply for the entire population of interest.

A ONE-WAY MODEL: Let Y_{ij} denote the laceration length for the jth pig on the ith withholding time. We may consider modelling the lengths as

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij},$$

for $i = 1, 2, 3$, $j = 1, 2, ..., 15$, where μ denotes the overall mean length (ignoring withholding times), and τ_i denotes the effect associated with the ith withholding time. All of μ and the τ_i's are assumed to be fixed **parameters**. This model is an example of a **one-way linear model**. The model consists of two parts. The first part is the **deterministic** part $Y_{ij} = \mu + \tau_i$. The second part, ϵ_{ij}, is the **random** error associated with the jth pig on the ith withholding time. This error could arise from measurement inconsistencies, inherent biological differences in pigs, sampling error, etc.

AN ANCOVA MODEL: Alternatively, one may consider the model

$$Y_{ij} = \mu + \tau_i + \gamma_i x_{ij} + \epsilon_{ij},$$

for $i = 1, 2, 3$, $j = 1, 2, ..., 15$, where γ_i are fixed parameters and x_{ij} denotes the weight of the jth pig on the ith withholding time prior to slaughter. This is an example of an **analysis of covariance** (ANCOVA) **linear model**. Here, the **covariate** is x, the weight prior to slaughter. When would this model be preferred over the one-way model?

A TWO-WAY MODEL: Suppose that for each of the 15 pigs on withholding time i, 5 pigs were taken from each of 3 farms. In this case, we might consider modelling the laceration lengths as

$$Y_{ijk} = \mu + \tau_i + \beta_j + \epsilon_{ijk},$$

for $i = 1, 2, 3$, $j = 1, 2, ..., 3$, $k = 1, 2, ..., 5$. Here, Y_{ijk} denotes the laceration length for the kth pig at the jth farm on the ith withholding time. This is an example of **two-way linear model**. In this example, we might treat the farms as **blocks**. In experimental design, two variables (here, treatment assignment and farm) are said to be **confounded** when their effects cannot be distinguished from each other. One way to eliminate possible confounding effects is with the use of **blocking**; it allows the experimenter to make treatment comparisons under more homogeneous conditions.

A MULTIPLE REGRESSION MODEL: Let x_{i1} and x_{i2} denote the weight and body temperature of pig i, respectively, prior to slaughter. Also, we define the two indicator variables x_{i3} and x_{i4} as follows:

$$x_{i3} = \begin{cases} 1, & \text{pig } i \text{ is assigned to treatment 0 hrs.} \\ 0, & \text{otherwise} \end{cases}$$

$$x_{i4} = \begin{cases} 1, & \text{pig } i \text{ is assigned to treatment 12 hrs.} \\ 0, & \text{otherwise.} \end{cases}$$

A **multiple linear regression model** takes the form

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i,$$

for $i = 1, 2, ..., 45$. The β_j terms; $j = 0, 1, ..., 4$, are called **regression coefficients** and ϵ_i denotes the random error associated with pig i. Note that there are two continuous predictor variables in this model (x_1 and x_2). Also, note that if $\beta_2 = 0$, this model is simply a **reparameterisation** of the ANCOVA model when $\gamma_1 = \gamma_2 = \gamma_3$ (equal slopes).

QUESTION: For each model, which hypotheses (in terms of the model parameters) might the researcher be interested in testing?

1.2 Describing populations using distributions

TERMINOLOGY: A **random variable**, say, Y, is a variable whose value is determined by chance. A random sample is usually denoted by $Y_1, Y_2, ..., Y_n$; i.e., n measurements of the variable Y. In statistics, we usually think of data as random variables or realisations of random variables.

TERMINOLOGY: A random variable Y is called **discrete** if there are only a finite (really, a countable) number of possible values that Y can assume. A random variable Y is called **continuous** if it can assume any number in a certain interval of numbers.

TERMINOLOGY: For any random variable Y, discrete or continuous, there are two functions that describe probabilities involving Y: a **probability density function** (pdf) and a **cumulative distribution function** (cdf). The pdf and cdf are usually denoted by $f_Y(y)$ and $F_Y(y)$, respectively. The cdf (regardless of whether or not Y is discrete or continuous) gives the probability

$$F_Y(y) = P(Y \leq y).$$

1.2.1 Discrete distributions

THE DISCRETE CASE: The pdf gives probabilities of the form $f_Y(y) = P(Y = y)$. It must be that $f_Y(y) \geq 0$, for all y, and $\sum_A f_Y(y) = 1$, where A denotes the set of all possible values of Y.

NOTATION: We will (with rare exception) denote a random variable Y with a capital letter; we denote an observed (or realised) value of Y as y, a lowercase letter. This is standard notation.

REMARK: The function $f_Y(y)$ is sometimes called a **probability model**, because it serves as a mathematical description of a real life phenomenon. There are many different probability models! Examples of **discrete** probability models include the binomial, Poisson, negative binomial, hypergeometric, etc. Examples of **continuous** probability models include the normal, gamma, Weibull, t, χ^2, and F. In addition to the normal distribution, we'll pay particular attention to the t, χ^2, and F distributions since they are pervasive in applied statistics.

Example 1.2. The **Poisson** distribution is often used to model count data. Mathematics can show that the pdf of a Poisson distribution, with parameter λ, is given by

$$f_Y(y) = \begin{cases} \frac{\lambda^y e^{-\lambda}}{y!}, & y = 0, 1, 2, ... \\ 0, & \text{otherwise.} \end{cases}$$

The parameter λ determines the location and variability of the distribution of Y. Figure 1.2 displays the pdf of Y when $\lambda = 5$. The height of each bar equals $f_Y(y) = P(Y = y)$. For example, $f_Y(2) = P(Y = 2) \approx 0.084$. What would a graph of the cumulative distribution function (cdf) of Y look like?

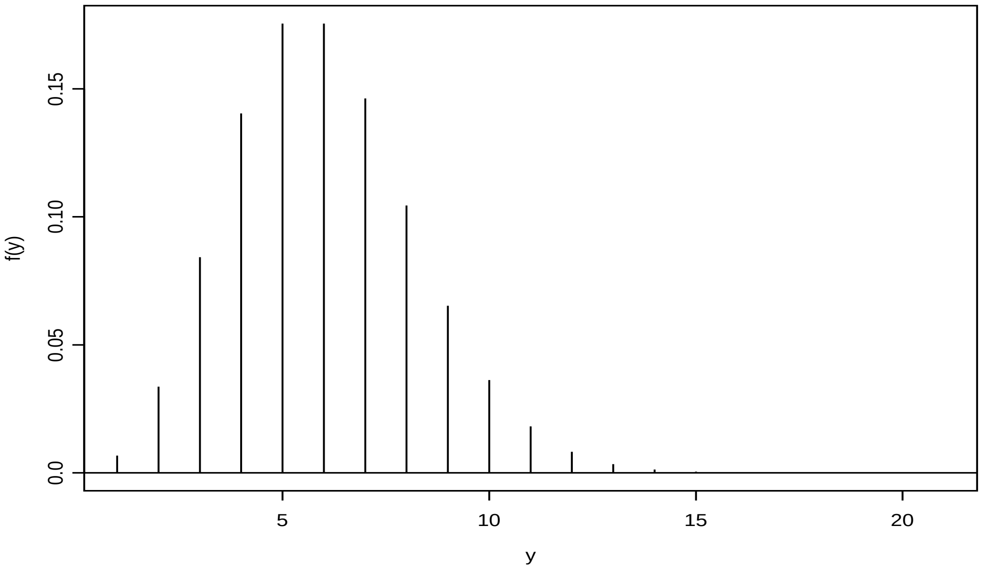

Figure 1.2: *The Poisson probability density function when* $\lambda = 5$.

TERMINOLOGY: Let Y be a discrete random variable with pdf $f_Y(y)$. The **expected value** or **population mean** of Y is given by

$$\mu \equiv E(Y) = \sum_{\text{all } y} y f_Y(y).$$

EXPECTATIONS OF FUNCTIONS OF Y. Let g be a real-valued function and let Y be a discrete random variable. Then, $g(Y)$ is a random variable and

$$E[g(Y)] = \sum_{\text{all } y} g(y) f_Y(y).$$

TERMINOLOGY: Let Y be a discrete random variable with pdf $f_Y(y)$. The **population variance** of Y is given by

$$\sigma^2 \equiv V(Y) = E[(Y - \mu_Y)^2] = \sum_{\text{all } y} (y - \mu_Y)^2 f_Y(y),$$

where $\mu_Y = E(Y)$.

1.2.2 Continuous distributions

CONTINUOUS CASE: Let Y be a **continuous** random variable with cdf $F_Y(y)$. The **probability density function** (**pdf**) for Y, denoted $f_Y(y)$, is given by

$$f_Y(y) = \frac{d}{dy}F_Y(y),$$

provided that $\frac{d}{dy}F_Y(y) \equiv F'_Y(y)$ exists. Furthermore,

$$F_Y(y) = \int_{-\infty}^{y} f_Y(t)dt.$$

The function $f_Y(y)$ is said to be **valid** if $f_Y(y) \geq 0$ for all y and $\int_A f_Y(y) = 1$, where A denotes the set of all possible values of Y.

RESULT: If Y is a **continuous** random variable with pdf $f_Y(y)$ and cdf $F_Y(y)$, then

$$P(a \leq Y \leq b) = \int_a^b f_Y(y)dy = F_Y(b) - F_Y(a).$$

Furthermore, $P(Y = a) = 0$, for any real constant a, since

$$P(Y = a) = P(a \leq Y \leq a) = \int_a^a f_Y(y)dy = 0.$$

Thus, for continuous random variables, probabilities are assigned to intervals with non-negative probability, and specific points with zero probability. *This is the key difference between discrete and continuous random variables.*

TERMINOLOGY: Let Y be a **continuous** random variable with pdf $f_Y(y)$. The **expected value** or **population mean** of Y is given by

$$\mu \equiv E(Y) = \int_{-\infty}^{\infty} y f_Y(y)dy.$$

EXPECTATIONS OF FUNCTIONS OF Y. Let g be a real-valued function and let Y be a **continuous** random variable. Then, $g(Y)$ is a random variable and

$$E[g(Y)] = \int_{-\infty}^{\infty} g(y)f_Y(y)dy.$$

TERMINOLOGY: Let Y be a **continuous** random variable with pdf $f_Y(y)$. The **population variance** of Y is given by

$$\sigma^2 \equiv V(Y) = E[(Y-\mu_Y)^2] = \int_{-\infty}^{\infty} (y-\mu_Y)^2 f_Y(y)dy,$$

where $\mu_Y = E(Y)$.

PROPERTIES OF EXPECTATIONS: Let Y be a random variable with pdf $f_Y(y)$, let $g, g_1, g_2, ..., g_k$ denote real functions, and let $c \in \mathcal{R}$ be any constant. Then,

(a) $E(c) = c$

(b) $E[cg(Y)] = cE[g(Y)]$

(c) $E[\sum_{j=1}^{k} g_j(Y)] = \sum_{j=1}^{k} E[g_j(Y)]$.

PROPERTIES OF VARIANCES: Let Y be a random variable with pdf $f_Y(y)$, and suppose a and b are real constants. Then,

(a) $V(a) = 0$

(b) $V(a+bY) = b^2V(Y)$.

1.3 The normal distribution

TERMINOLOGY: A random variable Y is said to have a **normal distribution** if its pdf is given by

$$f_Y(y) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}, & -\infty < y < \infty \\ 0, & \text{otherwise.} \end{cases}$$

NOTATION: Shorthand notation is $Y \sim \mathcal{N}(\mu, \sigma^2)$. It is not difficult to show that $E(Y) = \mu$ and $V(Y) = \sigma^2$.

FACTS ABOUT ANY NORMAL DISTRIBUTION:

(a) The pdf is **symmetric** about μ; that is, for any real constant a, $f_Y(\mu - a) = f_Y(\mu + a)$. The points of inflection are located at $y = \mu \pm \sigma$.

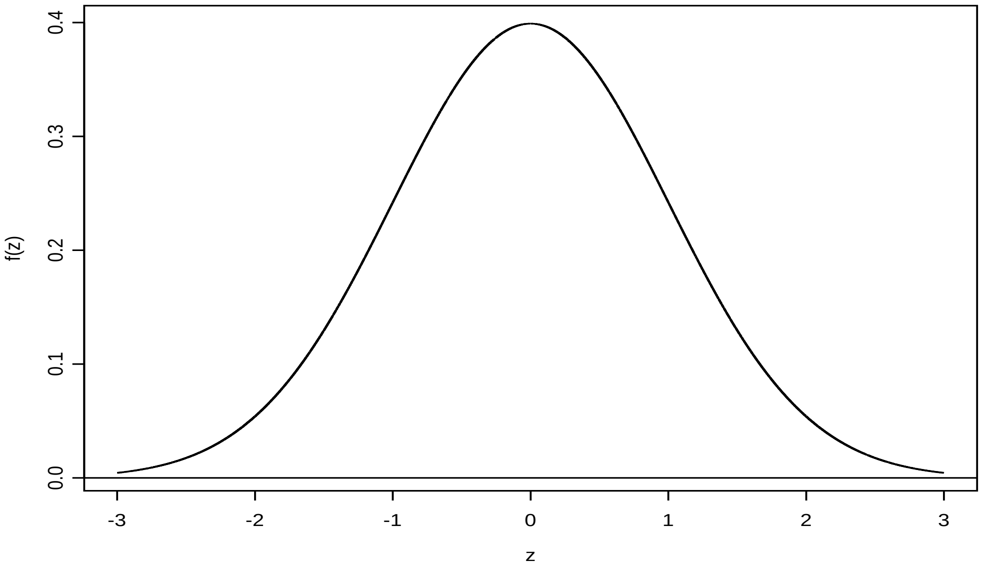

Figure 1.3: *The standard normal distribution.*

(b) A normal distribution can be "transformed" to a **standard normal distribution**.

(c) $\lim_{y\to\pm\infty} f_Y(y) = 0$.

TERMINOLOGY: A $\mathcal{N}(0,1)$ distribution is called the **standard normal distribution**. It is conventional notation to let Z denote the standard normal random variable; we often write $Z \sim \mathcal{N}(0,1)$. The pdf of Z is given in Figure 1.3.

IMPORTANT: Tabled values of the cdf of Z are given in Appendix C (Table C.1) of Rao. This table gives values of

$$1 - F_Z(z) = P(Z \geq z) = \int_z^\infty \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz,$$

for values of $z \geq 0$.

STANDARDISATION: Suppose that $Y \sim \mathcal{N}(\mu, \sigma^2)$. Then, the random variable

$$Z = \frac{Y - \mu}{\sigma} \sim \mathcal{N}(0,1).$$

FINDING NORMAL PROBABILITIES: Standarising each component of an event of interest involving a normal random variable Y, we see that

$$P(y_1 < Y < y_2) = P\left(\frac{y_1-\mu}{\sigma} < \frac{Y-\mu}{\sigma} < \frac{y_2-\mu}{\sigma}\right) = P\left(\frac{y_1-\mu}{\sigma} < Z < \frac{y_2-\mu}{\sigma}\right).$$

Similarly, we have

$$P(Y < y) = P\left(\frac{Y-\mu}{\sigma} < \frac{y-\mu}{\sigma}\right) P\left(Z < \frac{y-\mu}{\sigma}\right),$$

and

$$P(Y > y) = P\left(\frac{Y-\mu}{\sigma} > \frac{y-\mu}{\sigma}\right) = P\left(Z > \frac{y-\mu}{\sigma}\right).$$

1.4 Independence, covariance and correlation

TERMINOLOGY: Two random variables Y_1 and Y_2 are said to be **independent** if any event involving only Y_1 is independent of any event involving only Y_2.

REMARK: Independence is an important concept in statistics. This means that the random aspect of one observation contains no information about the random aspect of any other observation. There are mathematical ways to formalise this; however, for most purposes in applied statistics, this intuitive understanding of independence is sufficient.

TERMINOLOGY: Suppose that Y_1 and Y_2 are random variables with means μ_1 and μ_2, respectively. The **covariance** between Y_1 and Y_2 is given by

$$\text{Cov}(Y_1, Y_2) = E[(Y_1-\mu_1)(Y_2-\mu_2)] = E(Y_1Y_2) - \mu_1\mu_2$$

The covariance gives us information about how Y_1 and Y_2 are **linearly** related.

NOTES ON THE COVARIANCE:

- If $\text{Cov}(Y_1, Y_2) > 0$, then Y_1 and Y_2 are **positively** linearly related.
- If $\text{Cov}(Y_1, Y_2) < 0$, then Y_1 and Y_2 are **negatively** linearly related.
- If $\text{Cov}(Y_1, Y_2) = 0$, then Y_1 and Y_2 are **not** linearly related.

FACT: If two random variables Y_1 and Y_2 are independent, then $\text{Cov}(Y_1, Y_2) = 0$.

TERMINOLOGY: Suppose that Y_1 and Y_2 are random variables with variances σ_1^2 and σ_2^2, respectively. The quantity

$$\rho_{Y_1,Y_2} = \frac{\text{Cov}(Y_1, Y_2)}{\sigma_1\sigma_2}.$$

is called the **correlation** between Y_1 and Y_2. The correlation is often preferred to the covariance since ρ_{Y_1,Y_2} is always between -1 and 1.

1.5 Means, variances, and covariances of linear combinations

TERMINOLOGY: Suppose that $Y_1, Y_2, ..., Y_n$ are random variables and that $a_1, a_2, ..., a_n$ are constants. Then,

$$U_1 = \sum_{i=1}^{n} a_i Y_i = a_1Y_1 + a_2Y_2 + \cdots + a_nY_n$$

is called a **linear combination** of the random variables $Y_1, Y_2, ..., Y_n$.

MEAN OF A LINEAR COMBINATION:

$$E(U_1) = E\left(\sum_{i=1}^{n} a_iY_i\right) = \sum_{i=1}^{n} a_iE(Y_i)$$

VARIANCE OF A LINEAR COMBINATION:

$$\begin{aligned} V(U_1) = V\left(\sum_{i=1}^{n} a_iY_i\right) &= \sum_{i=1}^{n} a_i^2V(Y_i) + 2\sum_{i<j} a_ia_j\text{Cov}(Y_i, Y_j) \\ &= \sum_{i=1}^{n} a_i^2V(Y_i) + \sum_{i\neq j} a_ia_j\text{Cov}(Y_i, Y_j) \end{aligned}$$

If $Y_1, Y_2, ..., Y_n$ are **independent** random variables, then

$$V(U_1) = V\left(\sum_{i=1}^{n} a_iY_i\right) = \sum_{i=1}^{n} a_i^2V(Y_i),$$

since all the covariance terms are zero.

COROLLARY: Suppose that Y_1 and Y_2 are random variables. Then

$$V(Y_1 + Y_2) = V(Y_1) + V(Y_2) + 2\text{Cov}(Y_1, Y_2)$$

$$V(Y_1 - Y_2) = V(Y_1) + V(Y_2) - 2\text{Cov}(Y_1, Y_2).$$

Note that when Y_1 and Y_2 are independent, $V(Y_1 + Y_2) = V(Y_1 - Y_2) = V(Y_1) + V(Y_2)$.

COVARIANCE BETWEEN TWO LINEAR COMBINATIONS: Suppose that

$$U_2 = \sum_{j=1}^{m} b_j X_j = b_1 X_1 + b_2 X_2 + \cdots + b_m X_m,$$

where $X_1, X_2, ..., X_m$ are random variables $b_1, b_2, ..., b_m$ are constants. Then, it follows that

$$\text{Cov}(U_1, U_2) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j \text{Cov}(Y_i, X_j).$$

MISCELLANEOUS FACTS: Suppose that Y, Y_1, and Y_2 are random variables. Then

(a) $\text{Cov}(a_1 + b_1 Y_1, a_2 + b_2 Y_2) = b_1 b_2 \text{Cov}(Y_1, Y_2)$.

(b) $\text{Cov}(Y_1, Y_2) = \text{Cov}(Y_2, Y_1)$

(c) $\text{Cov}(Y, Y) = V(Y)$.

IMPORTANT FACT: Suppose that $Y_1, Y_2, ..., Y_n$ are independent $\mathcal{N}(\mu_i, \sigma_i^2)$ random variables for $i = 1, 2, ..., n$, and let $a_1, a_2, ..., a_n$ be non-random real constants. Then,

$$U_1 = a_1 Y_1 + a_2 Y_2 + \cdots + a_n Y_n \sim \mathcal{N}\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right).$$

IMPORTANT PUNCHLINE: The distribution of a linear combination of independent normal random variables is itself normally distributed! In fact, even if the random variables are **not** independent, the distribution of a linear combination of normal random variables is still normally distributed. In this case, what are the mean and variance for the linear combination?

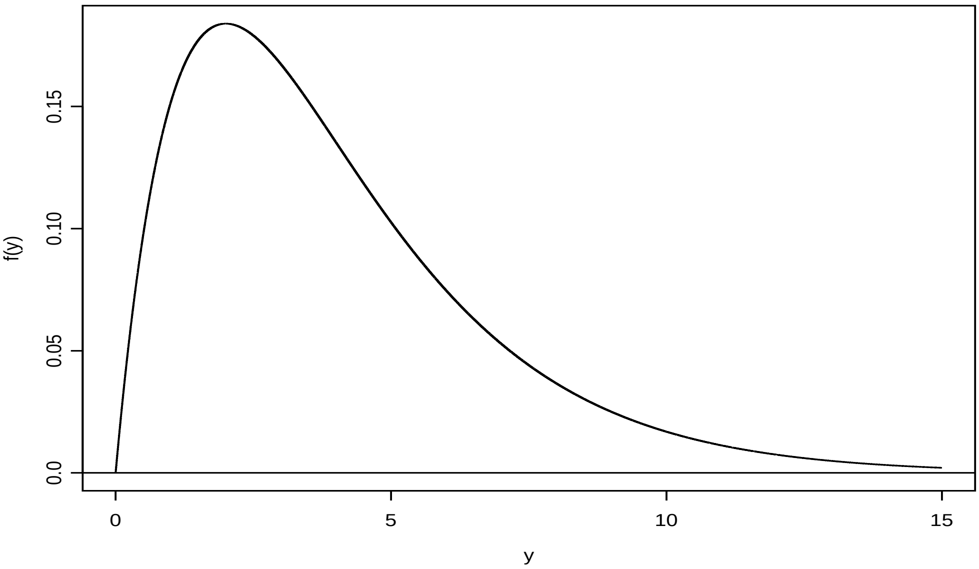

Figure 1.4: *The χ^2 distribution with 4 degrees of freedom.*

1.6 The χ^2, t, and F distributions

THE χ^2 DISTRIBUTION: The family of χ^2 distributions possesses the following properties. We denote a χ^2 distribution with ν degrees of freedom by χ^2_ν.

- The family is indexed by a degree of freedom parameter ν (which often depends on the sample size).
- Each χ^2 distribution is continuous and, in general, skewed right.
- $E(\chi^2_\nu) = \nu$ and $V(\chi^2_\nu) = 2\nu$.

IMPORTANT: Tabled values of the **quantiles** for the χ^2 distributions are given in Appendix C (Table C.3) of Rao. The table gives values of $\chi^2_{\nu,\alpha}$ which satisfy

$$P(\chi^2_\nu \geq \chi^2_{\nu,\alpha}) = \alpha$$

for different values of α.

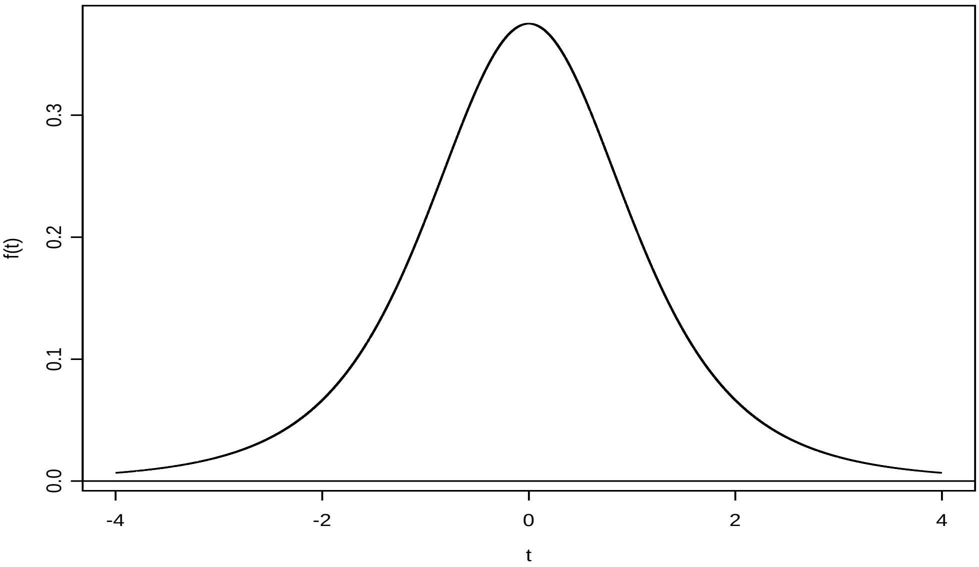

Figure 1.5: *The t distribution with 4 degrees of freedom.*

THEORETICAL FACTS:

- If $Y \sim \mathcal{N}(0,1)$, then $Y^2 \sim \chi^2_1$.
- If $Y_1, Y_2, ..., Y_n$ are independent $\mathcal{N}(\mu_i, \sigma_i^2)$ random variables, then

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \left(\frac{Y_i - \mu_i}{\sigma_i}\right)^2 \sim \chi^2_n.$$

THE t DISTRIBUTION: Suppose that $Z \sim \mathcal{N}(0,1)$ and that $W \sim \chi^2_\nu$. If Z and W are **independent**, then the quantity

$$T = \frac{Z}{\sqrt{W/\nu}}$$

has a t **distribution** with ν degrees of freedom, hereafter denoted as t_ν.

FACTS ABOUT THE t FAMILY OF DISTRIBUTIONS:

- The family is indexed by a degree of freedom parameter ν (which often depends on the sample size).
- Each distribution is continuous and **symmetric** about 0.

- As $\nu \to \infty$, $t_\nu \to \mathcal{N}(0,1)$; thus, when ν becomes larger, the t_ν and the $\mathcal{N}(0,1)$ distributions look more alike.
- $E(t_\nu) = 0$ and $V(t_\nu) = \frac{\nu}{\nu-2}$ for $\nu > 2$.
- When compared to the standard normal distribution, the t distribution, in general, is less peaked, and has more mass in the tails. Note that $V(t_\nu) > 1$.

IMPORTANT: Tabled values of the quantiles for the t distributions are given in Appendix C (Table C.2) of Rao. The table gives values of $t_{\nu,\alpha}$ which satisfy

$$P(t_\nu \geq t_{\nu,\alpha}) = \alpha$$

for different values of α.

THE F DISTRIBUTION: Suppose that $W_1 \sim \chi^2_{\nu_1}$ and that $W_2 \sim \chi^2_{\nu_2}$. If W_1 and W_2 are **independent**, then the quantity

$$F = \frac{W_1/\nu_1}{W_2/\nu_2}$$

has an F **distribution** with ν_1 (numerator) and ν_2 (denominator) degrees of freedom, hereafter denoted by F_{ν_1,ν_2}.

FACTS ABOUT THE F FAMILY OF DISTRIBUTIONS:

- The family is indexed by two degree of freedom parameters ν_1 and ν_2.
- Each distribution is continuous and, in general, skewed right.
- $t^2_\nu \sim F_{1,\nu}$.

IMPORTANT: Tabled values of the quantiles for the F distributions are given in Appendix C (Table C.4) of Rao. The table gives values of $F_{\nu_1,\nu_2,\alpha}$ which satisfy

$$P(F_{\nu_1,\nu_2} \geq F_{\nu_1,\nu_2,\alpha}) = \alpha$$

for different values of α. It is worth noting that

$$F_{\nu_1,\nu_2,\alpha} = \frac{1}{F_{\nu_2,\nu_1,1-\alpha}}.$$

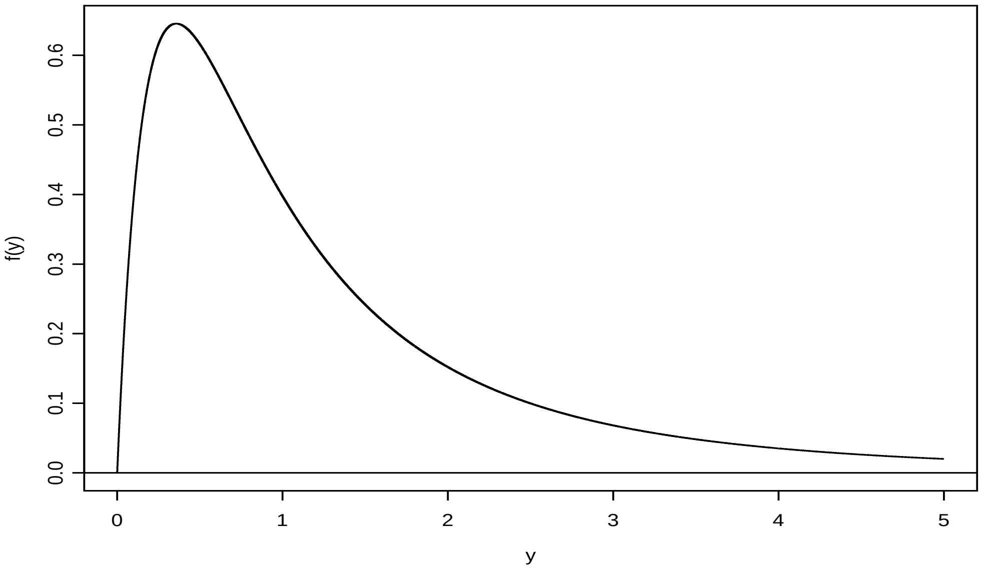

Figure 1.6: *The $F_{4,5}$ distribution.*

1.7 Sampling distributions

TERMINOLOGY: Suppose that $Y_1, Y_2, ..., Y_n$ is a **random sample** from some population modelled by $f_Y(y)$. This means precisely that

(a) the random variables $Y_1, Y_2, ..., Y_n$ are **independent**, and

(b) the random variables $Y_1, Y_2, ..., Y_n$ all follow the same probability model $f_Y(y)$; that is, each Y_i has the **identical** distribution.

The expression "$Y_1, Y_2, ..., Y_n$ is an **iid sample** from $f_Y(y)$," means that $Y_1, Y_2, ..., Y_n$ is a random sample from a population where $f_Y(y)$ is used to model Y.

NOTE: Throughout this course, we will assume that the population is infinite in size.

REMARK: For the remainder of the course, unless otherwise stated, we will assume that $Y_1, Y_2, ..., Y_n$ may be viewed as an iid sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution.

DEFINITION: A **statistic** is a function of the random variables $Y_1, Y_2, ..., Y_n$. It could possibly depend on non-random constants, but it can not depend on unknown parameters.

THREE IMPORTANT STATISTICS:

- the sample mean

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

- the sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

- the sample standard deviation

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$$

AN IMPORTANT IDEA: Since $Y_1, Y_2, ..., Y_n$ are random variables, any statistic is also a random variable! Thus, every statistic has, among other characteristics, a mean, a variance, and its **own probability distribution**!

DEFINITION: The **sampling distribution** of a statistic is simply the probability distribution of it. It characterises how the statistic varies in repeated sampling. *One of the major goals in statistics is to construct and use sampling distributions!*

ONE-SAMPLE RESULTS: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample of $\mathcal{N}(\mu, \sigma^2)$ observations. Then

$$\overline{Y} \sim \mathcal{N}(\mu, \sigma^2/n) \quad \text{and} \quad Z = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

Furthermore,

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} \quad \text{and} \quad t = \frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

TWO-SAMPLE RESULTS: Suppose that we have two independent samples:

$$Y_{11}, Y_{12}, ..., Y_{1n_1} \sim \text{iid } \mathcal{N}(\mu_1, \sigma_1^2)$$
$$Y_{21}, Y_{22}, ..., Y_{2n_2} \sim \text{iid } \mathcal{N}(\mu_2, \sigma_2^2).$$

Define

$$
\begin{aligned}
\overline{Y}_{1+} = \frac{1}{n_1}\sum_{j=1}^{n_1} Y_{1j} &= \text{sample mean for sample 1}\\
\overline{Y}_{2+} = \frac{1}{n_2}\sum_{j=1}^{n_2} Y_{2j} &= \text{sample mean for sample 2}\\
S_1^2 = \frac{1}{n_1-1}\sum_{j=1}^{n_1}(Y_{1j}-\overline{Y}_{1+})^2 &= \text{sample variance for sample 1}\\
S_2^2 = \frac{1}{n_2-1}\sum_{j=1}^{n_2}(Y_{2j}-\overline{Y}_{2+})^2 &= \text{sample variance for sample 2.}
\end{aligned}
$$

Mathematics can show that

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1};$$

furthermore, if the two population variances σ_1^2 and σ_2^2 are equal; i.e., $\sigma_1^2=\sigma_2^2=\sigma^2$, then

$$\frac{S_1^2}{S_2^2} \sim F_{n_1-1,n_2-1}.$$

QUESTION: Each of these results assumes that we are sampling from normal distributions. What happens when $Y_1, Y_2, ..., Y_n$ do not follow a normal distribution? In this case, it turns out that $\overline{Y}$ is still **approximately** normal when n is large.

THE CENTRAL LIMIT THEOREM: Suppose that $Y_1, Y_2, ..., Y_n$ are iid random variables with $E(Y_i)=\mu$ and $V(Y_i)=\sigma^2<\infty$ (note that we are not specifying a normal population). Denote the sample mean by $\overline{Y}=\frac{1}{n}\sum_{i=1}^n Y_i$ and define

$$Z_n = \frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}.$$

Then, as $n\to\infty$, the cumulative distribution function of Z_n converges to the $\mathcal{N}(0,1)$ distribution function.

INTERPRETATION: The CLT implies that even if $Y_1, Y_2, ..., Y_n$ are non-normal, the quantity Z_n will still have an **approximate** $\mathcal{N}(0,1)$ distribution, when n is large, and, thus, the sample mean $\overline{Y}$ will still have an **approximate** $\mathcal{N}(\mu,\sigma^2/n)$ sampling distribution. In this case, is common to write $\overline{Y}\sim\mathcal{AN}(\mu,\sigma^2/n)$.

1.8 Confidence intervals

TERMINOLOGY: A $100(1-\alpha)$ percent **confidence interval** for a parameter θ is a pair of statistics $\widehat{\theta}_L$ and $\widehat{\theta}_U$ such that

$$P(\widehat{\theta}_L \leq \theta \leq \widehat{\theta}_U) = 1-\alpha,$$

where $0<\alpha<1$. We call $1-\alpha$ the **confidence coefficient**.

FREQUENTIST INTERPRETATION: Before we see the data $Y_1, Y_2, ..., Y_n$, the interval $(\widehat{\theta}_L, \widehat{\theta}_U)$ is **random**. This follows since $\widehat{\theta}_L$ and $\widehat{\theta}_U$ are random quantities. On the other hand, θ is treated as a fixed **parameter**; it does not change. Thus, after we observe the data $y_1, y_2, ..., y_n$, the interval $(\widehat{\theta}_L, \widehat{\theta}_U)$ is no longer random. Thus, we think about confidence intervals in a repeated sampling context; namely, *in repeated sampling, approximately* $100(1-\alpha)$ *percent of the confidence intervals will contain the true parameter.*

TERMINOLOGY: We call Q a **pivot** if its sampling distribution does not depend on any unknown parameters.

CONFIDENCE INTERVAL FOR A NORMAL MEAN, σ^2 KNOWN: Suppose that $Y_1, Y_2, ..., Y_n$ iid $\mathcal{N}(\mu, \sigma_0^2)$, where μ is unknown and σ_0^2 is known. The quantity

$$Q = \frac{\overline{Y}-\mu}{\sigma_0/\sqrt{n}} \sim \mathcal{N}(0,1),$$

and, hence, is a pivot. Since $Q \sim \mathcal{N}(0,1)$, there exists a value $z_{\alpha/2}$ such that

$$\begin{aligned} 1-\alpha &= P(-z_{\alpha/2} < Q < z_{\alpha/2}) \\ &= P\left(-z_{\alpha/2} < \frac{\overline{Y}-\mu}{\sigma_0/\sqrt{n}} < z_{\alpha/2}\right) \\ &= P\Bigg(\underbrace{\overline{Y} - z_{\alpha/2}\times\sigma_0/\sqrt{n}}_{\widehat{\mu}_L} < \mu < \underbrace{\overline{Y} + z_{\alpha/2}\times\sigma_0/\sqrt{n}}_{\widehat{\mu}_U}\Bigg) \end{aligned}$$

Thus, $\overline{Y} \pm z_{\alpha/2}\times\sigma_0/\sqrt{n}$ is a $100(1-\alpha)$ percent confidence interval for μ.

CONFIDENCE INTERVAL FOR A NORMAL MEAN, σ^2 UNKNOWN: Suppose that $Y_1, Y_2, ..., Y_n$ are iid $\mathcal{N}(\mu, \sigma^2)$ observations, where both parameters are unknown. Then, $\overline{Y} \pm t_{n-1,\alpha/2}\times S/\sqrt{n}$ is a $100(1-\alpha)$ percent confidence interval for μ.

CONFIDENCE INTERVAL FOR A NORMAL VARIANCE: Suppose that $Y_1, Y_2, ..., Y_n$ are iid $\mathcal{N}(\mu, \sigma^2)$ observations. The quantity

$$Q = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1},$$

and, hence, is a pivot. Thus, there exists values $\chi^2_{n-1,1-\alpha/2}$ and $\chi^2_{n-1,\alpha/2}$ such that

$$\begin{aligned} 1-\alpha &= P(\chi^2_{n-1,1-\alpha/2} < Q < \chi^2_{n-1,\alpha/2}) \\ &= P\left\{\chi^2_{n-1,1-\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{n-1,\alpha/2}\right\} \\ &= P\left\{\underbrace{\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}}_{\widehat{\sigma}^2_L} < \sigma^2 < \underbrace{\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}}_{\widehat{\sigma}^2_U}\right\}. \end{aligned}$$

In this case,

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right)$$

is a $100(1-\alpha)$ percent confidence interval for σ^2.

CONFIDENCE INTERVAL FOR THE DIFFERENCE OF TWO NORMAL MEANS: Suppose that we have two **independent** samples (with **common variance**):

$$\begin{aligned} Y_{11}, Y_{12}, ..., Y_{1n_1} &\sim \text{iid } \mathcal{N}(\mu_1, \sigma^2) \\ Y_{21}, Y_{21}, ..., Y_{2n_2} &\sim \text{iid } \mathcal{N}(\mu_2, \sigma^2) \end{aligned}$$

The quantity

$$t = \frac{(\overline{Y}_{1+} - \overline{Y}_{2+}) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2},$$

where

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2},$$

and, hence, is a pivot. A $100(1-\alpha)$ percent confidence interval for $\mu_1 - \mu_2$ is given by $(\overline{Y}_{1+} - \overline{Y}_{2+}) \pm t_{n_1+n_2-2,\alpha/2} \times S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$.

A NOTE WORTH MENTIONING: When $\sigma_1^2 \neq \sigma_2^2$, it turns out that

$$(\overline{Y}_{1+} - \overline{Y}_{2+}) \pm t_{\nu,\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}},$$

where

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2-1}},$$

is an **approximate** $100(1-\alpha)$ percent confidence for $\mu_1 - \mu_2$.

REMARK: Note that this two-**independent**-sample setup can be expressed in a different way. Let Y_{ij} denote the jth replicate from a $\mathcal{N}(\mu_i, \sigma^2)$ distribution. It then follows that, for $i = 1, 2$ and $j = 1, 2, ..., n_i$,

$$\begin{aligned} Y_{ij} &= \mu + \tau_i + \epsilon_{ij} \\ &= \mu_i + \epsilon_{ij}, \end{aligned}$$

where $\mu_i = \mu + \tau_i$ and $\epsilon_{ij} \sim$ iid $\mathcal{N}(0, \sigma^2)$.

CONFIDENCE INTERVAL FOR THE RATIO OF TWO NORMAL VARIANCES: Suppose that we have two **independent** samples:

$$\begin{gathered} Y_{11}, Y_{12}, ..., Y_{1n_1} \text{ iid } \mathcal{N}(\mu_1, \sigma_1^2) \\ Y_{21}, Y_{21}, ..., Y_{2n_2} \text{ iid } \mathcal{N}(\mu_2, \sigma_2^2). \end{gathered}$$

A $100(1-\alpha)$ percent confidence interval for σ_2^2/σ_1^2 is

$$\left(\frac{S_2^2}{S_1^2} F_{n_1-1, n_2-1, 1-\alpha/2}, \frac{S_2^2}{S_1^2} F_{n_1-1, n_2-1, \alpha/2}\right).$$

1.9 Hypothesis tests

TERMINOLOGY: A **hypothesis test** is a procedure that enables us to draw conclusions about parameter values. The four parts to any statistical test are (a) the null hypothesis, denoted as H_0, (b) the alternative (or researcher's) hypothesis, denoted as H_1, (c) the test statistic, and (d) the rejection region.

TERMINOLOGY: A **test statistic** is a rule that is used to decide between H_0 and H_1.

The **rejection region** specifies the values of the test statistic for which H_0 is rejected. The rejection region is usually located in tails of a well-known probability distribution.

PREVAILING RESULT: *In a statistical test, if the test statistic falls in rejection region, then we reject* H_0 and say that "the result is **statistically significant**." All tests are constructed in a way so that we know the test statistic's sampling distribution when H_0 is true. This construction will allow us to quantify the probabilities of making certain types of mistakes.

TERMINOLOGY: **Type I Error**: *Rejecting* H_0 *when* H_0 *is true.* This event occurs with probability α (the **significance level** for the test).

TERMINOLOGY: **Type II Error**: *Not rejecting* H_0 *when* H_1 *is true.*

STRATEGY: In any hypothesis testing situation, we fix α at something we can "live with," say $\alpha = 0.01$, $\alpha = 0.05$, $\alpha = 0.10$, etc. Smaller values of α correspond to more conservative tests (i.e., tests where it is harder to reject H_0).

P VALUES: Rather than having formal rules for when to reject H_0, one can report the evidence against H_0 numerically. This is done by reporting a P value. The P value is computed under the assumption that H_0 is true; thus, small values of P are evidence against H_0. Essentially, the P value is the smallest value of α for H_0 is rejected. Thus, if the P value for a hypothesis test is smaller than the α used in the test, H_0 is rejected.

ONE SAMPLE t TEST: Suppose that $Y_1, Y_2, ..., Y_n$ are iid $\mathcal{N}(\mu, \sigma^2)$ observations, where both parameters are unknown. Recall that, when $\mu = \mu_0$,

$$t = \frac{\overline{Y} - \mu_0}{S_{\overline{Y}}} = \frac{\overline{Y} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}.$$

Thus, we reject $H_0 : \mu = \mu_0$ if we observe a test statistic t that doesn't follow the t_{n-1} distribution "as expected." The term $S_{\overline{Y}} = S/\sqrt{n}$ is called the **estimated standard error** of $\overline{Y}$.

Example 1.3 (`crabs.sas`). It is thought that the body temperature of intertidal crabs is less than the ambient temperature. Body temperatures were obtained from a random sample of $n = 8$ crabs exposed to an ambient temperature of 25.4 degrees C. Let Y denote this measurement and $y_1, y_2, ..., y_8$ denote the 8 measurements of Y. The data are $25.8, 24.6, 26.1, 24.9, 25.1, 25.3, 24.0,$ and 24.5. Assume that y_i are iid $\mathcal{N}(\mu, \sigma^2)$ observations, where μ denotes the mean body temperature for the population of intertidal crabs exposed to an ambient temperature of 25.4 degrees C. We wish to test, at $\alpha = 0.05$, say,

$$H_0 : \mu = 25.4$$

$$\text{versus}$$

$$H_1 : \mu < 25.4.$$

This is a **one-sided** test. We have $n = 8$, and $t_{7,0.95} = -1.8946$ (from Table C.2). Simple calculations show that $\overline{y} = 25.04$ and $s^2 = 0.4798$; thus, the one-sample t statistic is

$$t = \frac{\overline{y} - \mu}{s_{\overline{Y}}} = \frac{\overline{y} - \mu}{s/\sqrt{n}} = \frac{25.04 - 25.4}{\sqrt{0.4798/8}} = -1.48.$$

There is not enough evidence in the sample, at the five percent level, to suggest that the mean body temperature of intertidal crabs exposed to air at 25.4 degrees Celcius is, in fact, less than 25.4.

COMPUTING THE P VALUE: Note that from Table C.2, $0.05 < P(t_7 < -1.48) < 0.10$. Since the P value is not smaller than the significance level $\alpha = 0.05$, we do not have enough evidence at the five percent level to refute H_0. The exact P value is 0.0912.

COMPUTING A CONFIDENCE INTERVAL: Simple calculations show that, for the crab data, a 95 percent confidence interval, based on the t_7 distribution, is $(24.58, 25.50)$. Note that 25.4 falls in this interval. What does this suggest?

TWO SAMPLE t TEST: Suppose we have two **independent** random samples:

$$Y_{11}, Y_{12}, ..., Y_{1n_1} \text{ iid } \mathcal{N}(\mu_1, \sigma^2)$$

$$Y_{21}, Y_{22}, ..., Y_{2n_2} \text{ iid } \mathcal{N}(\mu_2, \sigma^2).$$

Recall that, when $\mu_1 - \mu_2 = 0$,

$$t = \frac{\overline{Y}_{1+} - \overline{Y}_{2+}}{S_{\overline{Y}_{1+} - \overline{Y}_{2+}}} = \frac{\overline{Y}_{1+} - \overline{Y}_{2+}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2},$$

where

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}.$$

Thus, we reject $H_0 : \mu_1 - \mu_2 = 0$ if we observe a test statistic t that does not follow the $t_{n_1+n_2-2}$ distribution "as expected." The term $S_{\overline{Y}_{1+}-\overline{Y}_{2+}}$ is called the **estimated standard error** of $\overline{Y}_{1+} - \overline{Y}_{2+}$.

Example 1.4 (`tomato.sas`). In an experiment that compared a standard fertiliser (A) and a modified fertiliser (B) for tomato plants, a gardener randomised 5 plants to treatment A and 6 plants to treatment B. The following data are yields that were observed in the experiment:

Standard A:	29.9	11.4	25.3	16.5	21.1	
Modified B:	26.6	23.7	28.5	14.2	17.9	24.3

The gardener wants to determine whether or not there is a difference in the two fertilisers. Assume that yields arise from two **independent** normal populations, let A = population 1, and let B = population 2. We will take $\alpha = 0.05$ and test

$$H_0 : \mu_1 - \mu_2 = 0$$

versus

$$H_1 : \mu_1 - \mu_2 \neq 0.$$

This is a **two-sided** test. We have $n_1 = 5$, $n_2 = 6$, so that $n_1 + n_2 - 2 = 9$, and $t_{9,0.025} = 2.2622$ (Table C.2). Standard calculations show

$$\overline{y}_{1+} = 20.84 \quad \overline{y}_{2+} = 22.53$$

$$s_1^2 = 52.50 \quad s_2^2 = 29.51$$

$$s_p^2 = \frac{4(52.50) + 5(29.51)}{9} = 39.73.$$

Thus, the two-sample t statistic is given by

$$t = \frac{\overline{y}_{1+} - \overline{y}_{2+}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{20.84 - 22.53}{\sqrt{39.73\left(\frac{1}{5}+\frac{1}{6}\right)}} = -0.44.$$

There is not enough evidence at the five percent significance level to suggest the fertilisers are different. The two-sided P value is 0.6677. Thus, H_0 would not be rejected at any reasonable significance level.

1.10 Matched-pairs experiments

In the last section, when testing for the difference in two normal means, it was necessary that the two samples be **independent**; this is a requirement for the underlying mathematical theory to be valid. The fact that the samples are assumed independent is a consequence of **experimental design** – the individuals in each sample do not overlap and were assigned treatments at random. If we design the experiment differently, then, as one should expect, different methods will be appropriate. We now consider an experimental design where we make comparisons **within** pairs of individuals that may tend to be "more alike" than other pairs.

Example 1.5 (`sbp.sas`). A certain stimulus is thought to produce an increase in mean systolic blood pressure (SBP) in middle-aged men.

DESIGN ONE: Take a random sample of men, then randomly assign each man to receive the stimulus or not (so the group of men that receives the treatment and the group that did not could be thought of as two **independent** samples). In this design, methods of the previous section would be appropriate.

DESIGN TWO: Consider an alternative design, the so-called **matched-pairs design**. Rather than assigning men to receive one treatment or the other (stimulus/no stimulus), obtain a response from each man under both treatments! That is, obtain a random sample of middle-aged men and take two readings on each man, with and without the stimulus. In this design, because readings of each type are taken on the same man, the difference between before and after readings on a given man should be less variable than the difference between a before-response on one man and an after-response on a different man. The man-to-man variation inherent in the latter difference is not present in the

Table 1.1: *Sources of variation present in different designs.*

Design	Type of Difference	Sources of Variation
Independent samples	among men	among men, within men
Matched pairs setup	within men	within men

difference between readings taken on the same subject!

ADVANTAGE OF MATCHED PAIRS: In general, by obtaining a pair of measurements on a single individual (e.g., man, rat, pig, plot, tobacco leaf, etc.), where one of the measurements corresponds to treatment 1 and the other measurement corresponds to treatment 2, you eliminate the variation **among** the individuals. Thus, you may compare the treatments (e.g., stimulus/no stimulus, ration A/ration B, etc.) under more **homogeneous** conditions where only variation within individuals is present (that is, the variation arising from the difference in treatments).

REMARK: In some situations, of course, pairing might be impossible or impractical (e.g., destructive testing in manufacturing, etc.). However, in a matched-pairs experiment, we still may think of **two populations**; e.g., those of all men with and without the stimulus. What changes in this setting is really how we have "sampled" from these populations. The two samples are no longer independent, because they involve the same individual.

A NOTE ON RANDOMISATION: In matched-pairs experiments, it is common practice, when possible, to **randomise** the order in which treatments are assigned. This may eliminate "common patterns" (that may confound our ability to determine a treatment effect) from always following, say, treatment A with treatment B. In practice, the experimenter could flip a fair coin to determine which treatment is applied first. If there are **carry-over effects** that may be present, these would have to be dealt with accordingly. We'll assume that there are no carry-over effects as a result from applying one or the other treatments in our discussion here.

IMPLEMENTATION: Matched-pairs designs are analysed by looking at the difference

Table 1.2: *Systolic blood pressure data.*

Subject j	Before (Y_1)	After (Y_2)	Difference ($D_j = Y_{1j} - Y_{2j}$)
1	120	128	-8
2	124	131	-7
3	130	131	-1
4	118	127	-9
5	140	132	8
6	128	125	3
7	140	141	-1
8	135	137	-2
9	126	118	8
10	130	132	-2
11	126	129	-3
12	127	135	-8

in responses of the two treatments. Specifically, compute

$$D_j = Y_{1j} - Y_{2j},$$

for each subject $j = 1, 2, ..., n$. For the SBP example, the data in Table 1.2 are the before and after SBP readings for the $n = 12$ middle-aged men in the experiment. To remind ourself, we could think of the mean of the population of differences as $\mu_1 - \mu_2$, where μ_1 denotes the mean before-stimulus reading and μ_2 denotes the mean after-stimulus reading. We, thus, want to perform the test

$$H_0 : \mu_1 - \mu_2 = 0$$

versus

$$H_1 : \mu_1 - \mu_2 < 0.$$

Note that by computing the data differences $d_1, d_2, ..., d_{12}$, we have now turned this into a "one-sample problem." That is, we are testing hypotheses concerning the value of a single population mean $\mu_1 - \mu_2$, the mean difference between the two treatments.

ANALYSIS: To perform a matched pairs analysis, we compute the one-sample t test on the observed data differences $D_1, D_2, ..., D_n$. We compute the sample mean and variance of $D_1, D_2, ..., D_n$ just like before; i.e.,

$$\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$$

$$S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2$$

The **estimated standard error** for the sample mean $\overline{D}$ is, by analogy to the one-sample case,

$$S_{\overline{D}} = \sqrt{\frac{S_D^2}{n}} = \frac{S_D}{\sqrt{n}}.$$

We thus have the matched-pairs test statistic:

$$t = \frac{\overline{D}}{S_{\overline{D}}}.$$

For an α level test, we reject H_0 in favour of a two-sided alternative if $t > t_{n-1,\alpha/2}$ or $t < -t_{n-1,\alpha/2}$. One sided rejection regions are formed analogously.

For our SBP data, straightforward calculations show that $\overline{d} = -1.833$, $s_D^2 = 33.97$, and $s_D = 5.83$. Hence,

$$t = \frac{\overline{d}}{s_{\overline{D}}} = \frac{\overline{d}}{s_D/\sqrt{n}} = \frac{-1.833}{5.83/\sqrt{12}} = -1.09.$$

For our one-sided alternative (i.e., examining whether or not the after-readings have higher mean than the before-readings), we would not reject H_0 at the $\alpha = 0.05$ level since -1.09 is not smaller than the critical value $t_{11,0.95} = -1.796$. The evidence in these data is not strong enough to suggest that the stimulus raises SBP in middle-aged men.

REMARK: The concept of pairing observations is a special form of **blocking**. As discussed earlier, this idea is an important foundation of experiments in biology, agriculture, engineering, and other applications. The basic idea is that we are limiting the effect of a potential source of variation so that the real differences, if they exist, are more likely to be detected.

1.11 A brief introduction to linear models

We have already introduced different models for explaining an observation on a random variable Y. All models involved a **deterministic** part; e.g., $\mu + \tau_i$, $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}$, etc. and a **stochastic** (i.e., random) error term ϵ representing the unexplained variability inherent in an observation that makes it different from the deterministic part of the model.

A SIMPLE LINEAR MODEL: Perhaps the easiest linear model representation is

$$Y_i = \mu + \epsilon_i.$$

This simple model is a useful representation that may be used as a framework for understanding **sources of variation**.

RECALL: In the last section, we learned about a design involving the pairing of experimental units as an alternative to the design in which observations are obtained from two independent samples. Such a model formalises the idea that the paired design eliminates the effect of variation across pairs of experimental units on the hypothesis test. To see this, consider a model for the case where an experiment is conducted with two independent samples, one from each of two populations. An observation from such an experiment is Y_{ij}, where as before, we mean the observation from the jth experimental unit from the ith sample. We may think of Y_{ij} in terms of the following **linear model**:

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij},$$

for $i = 1, 2$ and $j = 1, 2, ..., n_i$. In this model, μ may be thought of as the **overall mean**; that is, the mean response we would get before the treatments are applied. The parameter τ_i may be thought of as a **treatment effect**; that is, the change in mean that results from applying treatment i. The **random error** ϵ_{ij} represents everything else–sampling variation, biological differences in individuals–anything else unexplained that makes Y_{ij} different from the mean of its population, $\mu_i = \mu + \tau_i$. Thus, this includes all variation among experimental units, from all possible sources.

Now, consider an appropriate model for the case where an experiment is conducted according to a **matched-pairs design**. If we use this design, we have a legitimate basis for pairing experimental units because they are "alike" in some way. Thus, we may think of two ways in which experimental units may vary−**by pairs** and **within pairs**. An observation from such an experiment is again Y_{ij} where now the subscripts represent the observation for treatment i from the jth pair. The model is

$$Y_{ij} = \mu + \tau_i + \rho_j + \epsilon_{ij}.$$

In this model, the interpretations of μ and τ_i are the same as they were before in the two-independent sample model. The difference is the addition of the term ρ_j. This term may be thought of as the "effect" of being in the jth pair; that is, observations on a particular pair j differ from the mean for the treatment in question by an amount unique to pair j. The unexplained variation (the variation not accounted by treatment i and pair j) is represented by ϵ_{ij}.

COMPARISON: Comparing the two models, we see now the difference. In the two-independent-sample model, there is no term ρ_j, because there is no link between observations in each sample (they are all independent!). By pairing, when appropriate, we are "explaining" more of the variation by something we can identify, leaving less unexplained in the error term ϵ_{ij}.

Finally, note that for the matched-pairs experiment model, if we consider the difference for the jth pair, we get

$$\begin{aligned} D_j &= Y_{1j} - Y_{2j} \\ &= (\mu + \tau_1 + \rho_j + \epsilon_{1j}) - (\mu + \tau_2 + \rho_j + \epsilon_{2j}) \\ &= (\tau_1 - \tau_2) + (\epsilon_{1j} - \epsilon_{2j}). \end{aligned}$$

That is, the effect of ρ_j disappears! This makes sense: the whole idea of pairing (blocking) was to eliminate variation across pairs of experimental units. By controlling this variation in using the matched-pairs set up, this should lead to more precise inferences.

2 One-Way Classification and Analysis of Variance

Complimentary reading from Rao: Chapter 8 (§ 8.1-8.7).

2.1 Introduction

The purpose of an experiment is usually to investigate differences between or among treatments. In a statistical model framework, we may do this by comparing the **population means** of the responses to each treatment. We have already discussed designs for comparing two means; namely, a **two-independent-sample design** and a **matched-pairs design**. In this chapter, we consider the comparison of more than two treatment means in a one-way layout setting.

PREVAILING THEME: In order to detect treatment mean differences, we must try to control the effects of **experimental error** so that any variation we observe can be attributed to the effects of the treatments rather than to differences among the experimental units to which treatments were applied.

RECALL: We discussed the idea that designs involving meaningful grouping of experimental units (i.e., **blocking**) can help reduce the effects of experimental error, by identifying systematic components of variation among experimental units that may be due to something besides inherent biological variation among them. The matched-pairs design for comparing two treatments is an example of such a design. In this situation, experimental units themselves are treated as blocks.

The analysis of data from experiments involving blocking in scenarios with more than two treatments, which are representative of many experiments in practice, will be covered later. We start by discussing a simpler setting; that is, the **one-way classification model**. This is basically just an extension of the two-independent-sample design to more than two treatments.

ONE-WAY CLASSIFICATION: Consider an experiment to compare $t \geq 2$ treatment means, set up as follows:

- We obtain a random sample of experimental units and randomly assign them to treatments. In this situation, samples corresponding to the treatment groups are **independent** (i.e., the experimental units in each treatment sample are unrelated).
- We do not attempt to group experimental units according to some factor (e.g., location, gender, initial weight, variety, etc.).

REMARK: In this design, the only way in which experimental units may be "classified" is with respect to which treatment they received. Hence, such an arrangement is often called a **one-way classification**. When experimental units are thought to be "basically alike" (i.e., no apparent grouping seems appropriate), then the experimental error only consists of the variation among the experimental units themselves (that is, there are no other **systematic** sources of variation). Of course, if we were to group individuals in a given way, when, in reality, there was no grouping necessary, we would not add any precision to the experiment.

Example 2.1. In a lab experiment, we are to compare the viscosities among four different chemical mixtures: A, B, C, and D. The four mixtures are randomly assigned to 48 beakers (the experimental units); each mixture to 12 beakers. All the beakers are pretty much the same, so we would not expect variation from other systematic sources before the mixtures (treatments) are applied. In this situation, grouping beakers would be pointless since there is no identifiable reason for doing so.

COMPLETE RANDOMISATION: If there is no basis for grouping, all experimental units should have an equal chance of receiving any of the treatments. When randomisation is carried out in this way, it is called **complete randomisation**; such an experimental design is called a **completely randomised design** (**CRD**). In situations where grouping is involved, different randomisation schemes will be appropriate, as we will discuss later.

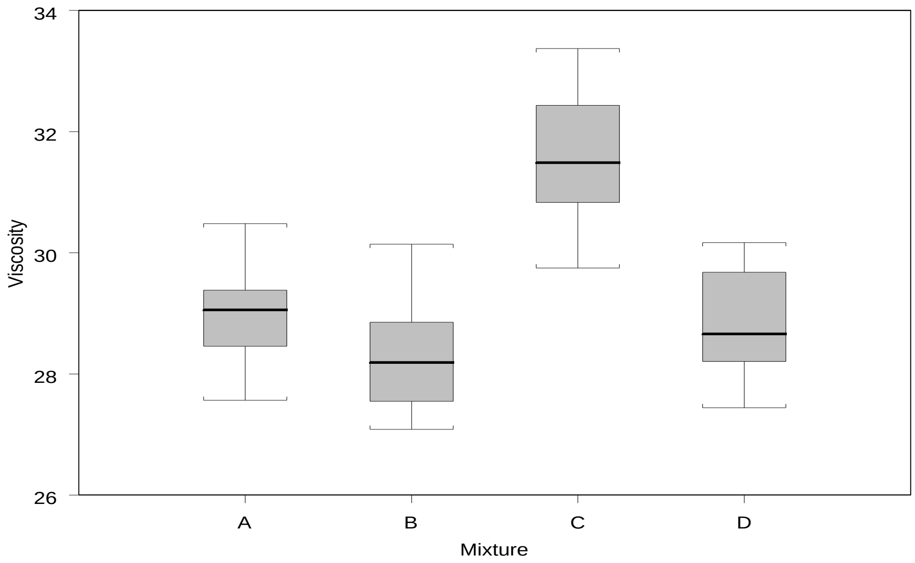

Figure 2.7: *Viscosity measurements for $t = 4$ chemical mixtures.*

ADVANTAGES OF THE ONE-WAY CLASSIFICATION:

- Simplicity of implementation and analysis.
- The size of the experiment is limited only by the availability of experimental units. No special considerations for different types of experimental units are required.

DISADVANTAGES OF THE ONE-WAY CLASSIFICATION:

- **Experimental error**, our assessment of the non-systematic variation believed to be inherent among experimental units, includes **all** sources.
- If it turns out, unexpectedly perhaps, that some of the variation among experimental units is indeed due to a systematic component, it will not be possible to "separate it out" of experimental error, and comparisons may be impossible. Thus, we run the risk of low precision and power if something unexpected arises.

2.2 Analysis of variance for the one-way classification

NOTATION: Let t denote the number of treatments to be compared, and, as before, let

$$Y_{ij} = \text{response on the } j\text{th experimental unit on treatment } i,$$

for $i = 1, 2, ..., t$ and $j = 1, 2, ..., n_i$. Here, n_i is the number of **replications** for treatment i. When $n_1 = n_2 = \cdots = n_t = n$, say, we call this a **balanced design**; otherwise, the design is said to be **unbalanced**. Let $N = n_1 + n_2 + \cdots + n_t$. If the design is balanced, then $N = nt$. Define

$$\begin{aligned}
Y_{i+} = \sum_{j=1}^{n_i} Y_{ij} &= \text{sample total for treatment } i \\
\overline{Y}_{i+} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} &= \text{sample mean for treatment } i \\
S_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i+})^2 &= \text{sample variance for treatment } i \\
Y_{++} = \sum_{i=1}^{t}\sum_{j=1}^{n_i} Y_{ij} &= \text{grand total} \\
\overline{Y}_{++} = \frac{1}{N}\sum_{i=1}^{t}\sum_{j=1}^{n_i} Y_{ij} &= \text{grand mean}
\end{aligned}$$

STATISTICAL HYPOTHESIS: Our first goal is to develop a procedure for testing

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_t$$

versus

$$H_1 : \text{the } \mu_i \text{ are not all equal.}$$

Note that the null hypothesis says that there is "no treatment difference" or "no treatment effect." The alternative hypothesis merely says that a difference among the t means exists somewhere (but does not specify how the means are different). The underlying theory needed to conduct this test is now given.

DATA AND ASSUMPTIONS: We have **independent** random samples from $t \geq 2$ **normal** distributions, each of which has the **same variance** (but possibly different means):

$$\begin{array}{ll}
\text{Sample 1} & Y_{11}, Y_{12}, ..., Y_{1n_1} \text{ iid } \mathcal{N}(\mu_1, \sigma^2) \\
\text{Sample 2} & Y_{21}, Y_{22}, ..., Y_{2n_2} \text{ iid } \mathcal{N}(\mu_2, \sigma^2) \\
\vdots & \vdots \\
\text{Sample } t & Y_{t1}, Y_{t2}, ..., Y_{tn_t} \text{ iid } \mathcal{N}(\mu_t, \sigma^2)
\end{array}$$

MAIN POINT: The ANOVA procedure is formulated by deriving **two** independent estimators the common variance σ^2. These two estimators are formed by (1) looking at the variance of the observations **within** samples, and (2) looking at the variance of the sample means **across** the t samples.

THE "WITHIN" ESTIMATOR FOR σ^2: To estimate the common σ^2 **within** samples, we take a weighted average (weighted by the sample sizes) of the t sample variances; that is, we "pool" all variance estimates together to form one estimate. Define

$$\begin{aligned}
\text{SS[E]} &= (n_1 - 1)S_1^2 + (n_2 - 1)S_2^2 + \cdots + (n_t - 1)S_t^2 \\
&= \sum_{i=1}^{t} \underbrace{\sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i+})^2}_{(n_i-1)S_i^2}.
\end{aligned}$$

We call SS[E] the **error sum of squares**. Under the normality assumption, recall that for each i,

$$\frac{(n_i - 1)S_i^2}{\sigma^2} \sim \chi^2_{n_i-1}.$$

Because the samples are independent, it follows (why?) that

$$\frac{\text{SS[E]}}{\sigma^2} = \sum_{i=1}^{t} \frac{(n_i - 1)S_i^2}{\sigma^2} \sim \chi^2_{N-t}.$$

Since, in general, the mean of a chi-square random variable is its degrees of freedom, we have that

$$E\left(\frac{\text{SS[E]}}{\sigma^2}\right) = N - t;$$

hence, defining

$$\text{MS[E]} = \frac{\text{SS[E]}}{N-t},$$

it follows that $E(\text{MS[E]}) = \sigma^2$. Note that MS[E] is an **unbiased estimator** of σ^2 regardless of whether or not the means $\mu_1, \mu_2, ..., \mu_t$ are equal. MS[E] is our first point estimator for σ^2. We call MS[E] the **mean squared error**.

THE "ACROSS" ESTIMATOR FOR σ^2: To derive the "across-sample" estimator, we will assume a common sample size $n_1 = n_2 = \cdots = n_t = n$ (this just simplifies the mathematics; however, the final result gleaned from this discussion still holds for unbalanced designs). Recall that if a sample arises from a normal population, then the sample mean is also normally distributed; that is, $\overline{Y}_{i+} \sim \mathcal{N}(\mu_i, \sigma^2/n)$.

NOTE: If all the treatment means are equal to a common value, say μ; i.e., H_0 is true, then we know that $\overline{Y}_{i+} \sim \mathcal{N}(\mu, \sigma^2/n)$, for each $i = 1, 2, ..., t$. Thus, *if H_0 is really true*, we may view the t sample means $\overline{Y}_{1+}, \overline{Y}_{2+}, ..., \overline{Y}_{t+}$ as being just an iid sample, of size t, from a normal population with mean μ and variance σ^2/n.

REVELATION: In light of this last remark, consider constructing the **sample variance** of our "random sample," $\overline{Y}_{1+}, \overline{Y}_{2+}, ..., \overline{Y}_{t+}$. This sample variance is given by

$$\frac{1}{t-1}\sum_{i=1}^{t}(\overline{Y}_{i+} - \overline{Y}_{++})^2 \tag{2.1}$$

and has expectation

$$E\left[\frac{1}{t-1}\sum_{i=1}^{t}(\overline{Y}_{i+} - \overline{Y}_{++})^2\right] = \sigma^2/n.$$

Hence, it follows that MS[T], where

$$\text{MS[T]} = \frac{1}{t-1}\underbrace{\sum_{i=1}^{t} n(\overline{Y}_{i+} - \overline{Y}_{++})^2}_{\text{SS[T]}},$$

is an **unbiased estimator** of σ^2; i.e., $E(\text{MS[T]}) = \sigma^2$, when H_0 is true. We call SS[T] the **treatment sums of squares** and MS[T] the **mean squared for treatments**. MS[T] is our second point estimator for σ^2. Recall that MS[T] is an unbiased estimator of σ^2

only when $H_0 : \mu_1 = \mu_2 = \cdots = \mu_t$ is true (this is important!). If we have different sample sizes, we simply adjust MS[T] to

$$\text{MS[T]} = \frac{1}{t-1} \underbrace{\sum_{i=1}^{t} n_i(\overline{Y}_{i+} - \overline{Y}_{++})^2}_{\text{SS[T]}}$$

This is still an unbiased estimator for σ^2 when H_0 is true.

SUMMARY: We have derived two **unbiased estimators** for σ^2:

- the first (*within*), MS[E], is not affected by whether or not the means are different (i.e., it is unbiased for σ^2 regardless of whether or not H_0 is true). This estimate reflects how individual observations differ from their means, regardless of the values of those means. Thus, MS[E] reflects only variation attributable to how individuals differ among themselves.

- the second (*across*), MS[T], is derived assuming that the means are the **same** (i.e., assuming that H_0 is true), and is affected by whether or not the means are different. This estimate MS[T] reflects not only how individual observations differ (through their sample means), but also how the treatment means might differ.

IMPLICATION: We derived MS[T] under the assumption that H_0 was true. Thus, if H_0 really is true (i.e., there are no differences in treatment means), we would expect MS[T] and MS[E] to be "close." On the other hand, if H_0 really is not true, one would expect that MS[T] be **larger** than MS[E]. With this in mind, consider forming the **ratio** $F = \text{MS[T]}/\text{MS[E]}$. We should now see that if

- H_0 is true, F should be close to one.

- H_0 is not true, F should be (perhaps much) larger than one.

CONCLUSION: Large values of the F ratio are evidence against H_0.

THE SAMPLING DISTRIBUTION OF F: Recall that $\text{SS[E]}/\sigma^2 \sim \chi^2_{N-k}$ always. In addition, when H_0 is true (assuming a common n),

$$\frac{\text{SS[T]}}{\sigma^2} = \frac{\sum_{i=1}^t n(\overline{Y}_{i+} - \overline{Y}_{++})^2}{\sigma^2} = \frac{\sum_{i=1}^t (\overline{Y}_{i+} - \overline{Y}_{++})^2}{\sigma^2/n} \sim \chi^2_{t-1}.$$

Thus, $\text{SS[T]}/\sigma^2 \sim \chi^2_{k-1}$ when H_0 is **true** (this statement is also true when different sample sizes are used). Furthermore, SS[T] and SS[E] are independent statistics (why?). Thus, when H_0 is true,

$$F = \frac{\text{MS[T]}}{\text{MS[E]}} = \frac{\frac{\text{SS[T]}}{\sigma^2}/(t-1)}{\frac{\text{SS[E]}}{\sigma^2}/(N-t)} \sim F_{t-1,N-t}.$$

Putting this all together, we see that H_0 is rejected when the test statistic $F = \text{MS[T]}/\text{MS[E]}$ falls in the **upper tail** of the $F_{t-1,N-t}$ distribution. If we want to perform a test at significance level α, the rejection region is given by $RR = \{F : F > F_{t-1,N-t,\alpha}\}$, where $F_{t-1,N-t,\alpha}$ is the $1-\alpha$ quantile of the $F_{t-1,N-t}$ distribution. Note that this is a one-sided, **upper tail** rejection region. As you probably suspect, P values are computed as areas under the $F_{t-1,N-t}$ distribution.

PRESENTATION: We can amalgamate all of this information into an **ANOVA table**. The form of the ANOVA table for the one-way classification model is given in Table 2.3.

Table 2.3: *ANOVA table for the one-way layout.*

Source	df	SS	MS	F
Treatments	$t-1$	SS[T]	$\text{MS[T]} = \frac{\text{SS[T]}}{t-1}$	$F = \frac{\text{MS[T]}}{\text{MS[E]}}$
Error	$N-t$	SS[E]	$\text{MS[E]} = \frac{\text{SS[E]}}{N-t}$	
Total	$N-1$	SS[TOT]		

NOTES ON THE ANOVA TABLE STRUCTURE:

- It is not difficult to show that

$$\text{SS[TOT]} = \text{SS[T]} + \text{SS[E]}.$$

SS[TOT] may be thought of as measuring how observations vary about the overall mean, without regard to treatments; that is, it measures the total variation in all the data. SS[TOT] can be **partitioned** into two **independent** components:

  - SS[T], measuring how much of the total variation is due to the treatments
  - SS[E], measuring the remaining variation, which we attribute to inherent variation among the individuals.

- the degrees of freedom add down.
- in general, mean squares are formed by dividing sums of squares by the corresponding degrees of freedom.

COMPUTING FORMULAE: To summarise, we now provide computing formulae for all the sums of squares. Here, we are assuming that the sample sizes, n_i, are different; if they are all equal, just replace n_i with the common n. First, the **correction term for the overall mean** is given by

$$\text{CM} = \frac{1}{N}Y_{++}^2 = \frac{1}{N}\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}\right)^2 = N\overline{Y}_{++}^2.$$

Sums of squares may be computed as follows

$$\begin{aligned}
\text{SS[TOT]} &= \sum_{i=1}^{t}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_{++})^2 = \sum_{i=1}^{t}\sum_{j=1}^{n_i} Y_{ij}^2 - \text{CM}, \\
\text{SS[T]} &= \sum_{i=1}^{t} n_i(\overline{Y}_{i+}-\overline{Y}_{++})^2 = \sum_{i=1}^{t}\frac{1}{n_i}\Bigg(\underbrace{\sum_{j=1}^{n_i} Y_{ij}}_{Y_{i+}}\Bigg)^2 - \text{CM} \\
\text{SS[E]} &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_{i+})^2 = \text{SS[TOT]} - \text{SS[T]}.
\end{aligned}$$

Example 2.2 (`pea.sas`). The following data record the length of pea sections, in ocular units (×0.114 mm), grown in tissue culture with auxin (a plant hormone) present. The purpose of the experiment was to test the effects of the addition of various sugars on growth as measured by length. Pea plants were randomly assigned to one of $t = 5$

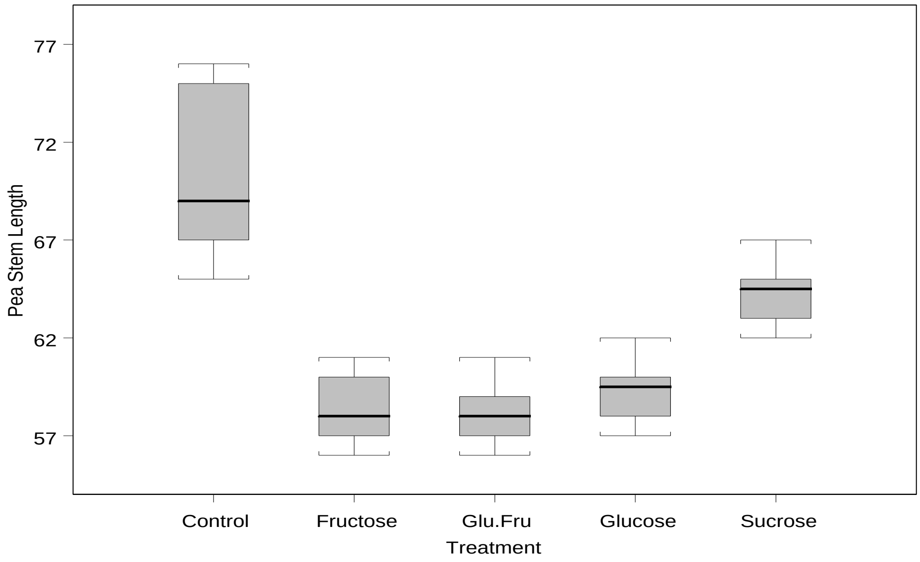

Figure 2.8: *Pea section data for $t = 5$ treatments.*

treatment groups: control (no sugar added), 2% fructose added, 1% glucose and 1% fructose added, 2% glucose added, and 2% sucrose added. Ten observations ($n_i = 10$) were obtained for each group of plants (a balanced design). In all, $N = 50$ pea plants were used; lengths are given in Table 2.4.

NOTE: Here, the individual plants are the **experimental units**, and we are applying the sugars (**treatments**) to the plants using complete randomisation.

RECALL: Our assumptions for the one-way layout are that (a) the samples are **independent** (how can this be achieved), (b) the measurements (i.e., growths) are **normally distributed** and (c) the measurements have **constant variance**. Do you think these are good assumptions here? How can we check this assumptions?

HAND CALCULATIONS: We calculate (following steps on page 287-8 Rao)

$$\text{CM} = \frac{1}{N}Y_{++}^2 = \frac{1}{N}\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}\right)^2 = \frac{1}{50}(3097)^2 = 191828.18$$

Table 2.4: *Pea plant experiment data.*

	Control	2% fru	1%/1% g/f	2% glu	2% suc	
	75	58	58	57	62	
	67	61	59	58	66	
	70	56	58	60	65	
	75	58	61	59	63	
	65	57	57	62	64	
	71	56	56	60	62	
	67	61	58	60	65	
	67	60	57	57	65	
	76	57	57	59	62	
	68	58	59	61	67	
Y_{i+}	701	582	580	593	641	$Y_{++} = 3097$
$\overline{Y}_{i+}$	70.1	58.2	58.0	59.3	64.1	$\overline{Y}_{++} = 61.94$

and

$$\sum_{i=1}^{t}\sum_{j=1}^{n_i} Y_{ij}^2 = 75^2 + 58^2 + 58^2 + \cdots + 61^2 + 67^2 = 193161.00$$

which gives

$$\text{SS[TOT]} = \sum_{i=1}^{t}\sum_{j=1}^{n_i} Y_{ij}^2 - \text{CM} = 193161.00 - 191828.18 = 1332.82.$$

Also, we have the treatment sums of squares

$$\begin{aligned}\text{SS[T]} &= \sum_{i=1}^{t}\frac{1}{n_i}\bigg(\underbrace{\sum_{j=1}^{n_i} Y_{ij}}_{Y_{i+}}\bigg)^2 - \text{CM} \\ &= \frac{1}{10}(701^2 + 582^2 + 580^2 + 593^2 + 641^2) - 191828.18 = 1077.32.\end{aligned}$$

Finally, the error sums of squares is found by subtraction

$$\text{SS[E]} = \text{SS[TOT]} - \text{SS[T]} = 1332.82 - 1077.32 = 245.50.$$

Table 2.5: *Analysis of variance: Pea section data.*

Source	df	SS	MS	F
Treatments	4	1077.32	269.33	49.37
Error	45	245.50	5.46	
Total	49	1322.82		

We also have $t-1=4$ and $N-t=45$, so that

$$\text{MS[T]} = \frac{1077.32}{4} = 269.33, \quad \text{MS[E]} = \frac{245.50}{45} = 5.46, \quad F = \frac{\text{MS[T]}}{\text{MS[E]}} = 49.37.$$

ANALYSIS: The ANOVA table for the pea-section data is given in Table 2.5. To perform the hypothesis test for differences among the treatment means; i.e., to test

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_5$$
$$\text{versus}$$
$$H_1 : \text{the } \mu_i \text{ are not all equal,}$$

where μ_i denotes the mean growth for treatment i, we can compare F to the appropriate critical value from the F table (Table C.4 in Rao). Using a significance level of $\alpha = 0.05$, we have

$$2.58 = F_{4,50,0.05} < F_{4,45,0.05} < F_{4,40,0.05} = 2.61,$$

so that $49.37 > F_{4,45,0.05}$. Thus, we reject H_0 and conclude there is (overwhelming) evidence, at the $\alpha = 0.05$ level, that the mean lengths of pea stems are different depending upon which (if any) sugar was added. The P value for this test, the area to the **right** of 49.37 on the $F_{4,45}$ density curve is < 0.0001.

2.3 Linear models for the one-way classification

In a one-way classification model, recall that we have independent random samples from $t \geq 2$ normal distributions, each of which has the same variance. Schematically, we can envision our data as follows:

$$\begin{array}{ll} \text{Sample 1} & Y_{11}, Y_{12}, ..., Y_{1n_1} \text{ iid } \mathcal{N}(\mu_1, \sigma^2) \\ \text{Sample 2} & Y_{21}, Y_{22}, ..., Y_{2n_2} \text{ iid } \mathcal{N}(\mu_2, \sigma^2) \\ \vdots & \vdots \\ \text{Sample } t & Y_{t1}, Y_{t2}, ..., Y_{tn_t} \text{ iid } \mathcal{N}(\mu_t, \sigma^2) \end{array}$$

MEANS MODEL: In terms of a linear model, we can express this setup as

$$Y_{ij} = \mu_i + \epsilon_{ij},$$

for $i = 1, 2, ..., t$ and $j = 1, 2, ..., n_i$, where μ_i denotes the mean of treatment i and $\epsilon_{ij} \sim$ iid $\mathcal{N}(0, \sigma^2)$. This is sometimes called a **one-way means model**, since the parameters $\mu_1, \mu_2, ..., \mu_t$ are the means of the t population distributions.

LEAST SQUARES ESTIMATION: We now develop estimators for the parameters $\mu_1, \mu_2, ..., \mu_t$ in the one-way means model above using the **method of least squares**. To find the **least squares estimators** of $\mu_1, \mu_2, ..., \mu_t$, we minimise

$$\sum_{i=1}^{t}\sum_{j=1}^{n_i} \epsilon_{ij}^2 = \sum_{i=1}^{t}\sum_{j=1}^{n_i} (Y_{ij} - \mu_i)^2$$

with respect to $\mu_1, \mu_2, ..., \mu_t$. This is a t dimensional minimisation problem, and straightforward differentiable calculus methods will apply here. The appropriate values are solutions to the t simultaneous equations (sometimes called the **normal equations**)

$$\frac{\partial}{\partial \mu_i}\sum_{i=1}^{t}\sum_{j=1}^{n_i} (Y_{ij} - \mu_i)^2 = 0,$$

for $i = 1, 2, ..., t$. It is easy to show (try it!) that these minimisers are given by $\widehat{\mu}_i = \overline{Y}_{i+}$. That is, $\overline{Y}_{i+}$ is the **least squares estimator** of μ_i, for $i = 1, 2, ..., t$.

EFFECTS MODEL: We can also express the one-way classification model as

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij},$$

for $i = 1, 2, ..., t$ and $j = 1, 2, ..., n_i$, where μ denotes the overall mean, τ_i denotes the effect of receiving treatment i, and $\epsilon_{ij} \sim$ iid $\mathcal{N}(0, \sigma^2)$. This is sometimes called a **one-way effects model**, since the parameters $\tau_1, \tau_2, ..., \tau_t$ represent effects rather than means.

COMPARISON: Structurally, there is no difference between the means model and effects model. The only difference is in the interpretation. In fact, starting with the effects model and letting $\mu_i = \mu + \tau_i$, we arrive back at the means model. Put another way, the two models are simply **reparameterisations** of one another.

LEAST SQUARES ESTIMATION: With the one-way effects model, we now derive the least squares estimators of μ and the τ_i's. To find the least squares estimators of $\mu, \tau_1, \tau_2, ..., \tau_t$, we minimise

$$\sum_{i=1}^{t}\sum_{j=1}^{n_i}\epsilon_{ij}^2 = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(Y_{ij} - \mu - \tau_i)^2$$

with respect to $\mu, \tau_1, \tau_2, ..., \tau_t$. This is now a $t+1$ dimensional minimisation problem. We proceed as before; appropriate values solve the $t+1$ equations

$$\begin{aligned} \frac{\partial}{\partial\mu}\sum_{i=1}^{t}\sum_{j=1}^{n_i}(Y_{ij} - \mu - \tau_i)^2 &= 0 \\ \frac{\partial}{\partial\tau_i}\sum_{i=1}^{t}\sum_{j=1}^{n_i}(Y_{ij} - \mu - \tau_i)^2 &= 0 \qquad i = 1, 2, ..., t. \end{aligned}$$

This yields the following **normal equations** (verify!):

$$\begin{aligned} N\mu + n_1\tau_1 + n_2\tau_2 + \cdots n_t\tau_t &= Y_{++} \\ n_1\mu + n_1\tau_1 &= Y_{1+} \\ n_2\mu + n_2\tau_2 &= Y_{2+} \\ &\vdots \\ n_t\mu + n_t\tau_t &= Y_{t+}. \end{aligned}$$

AN APPARENT PROBLEM: Notice that if we add the last t normal equations above, we get the first one. Thus, the normal equations are **not** linearly independent, and, hence, no **unique** solution for $\mu, \tau_1, \tau_2, ..., \tau_t$ exists. In fact, there are infinitely many solutions for $\mu, \tau_1, \tau_2, ..., \tau_t$!

TERMINOLOGY: Mathematically speaking, a model that contains components that can not be estimated uniquely is said to be **overparameterised**. The one-way effects model $Y_{ij} = \mu + \tau_i + \epsilon_{ij}$ is an example of such a model.

SIDE CONDITIONS: One way to obtain a unique solution to the normal equations is to impose certain **side conditions** or **constraints** on the parameters. Basically, these are just extra conditions that are specified so that the normal equations can be solved uniquely. In fact, fundamentally, one choice of side conditions is as good as any other! One common side condition used in the one-way layout is

$$\sum_{i=1}^{t} n_i\tau_i = 0.$$

Using this constraint, we can obtain a unique solution to the normal equations (verify!)

$$\begin{aligned}\widehat{\mu} &= \overline{Y}_{++}\\ \widehat{\tau}_i &= \overline{Y}_{i+} - \overline{Y}_{++}, \qquad i = 1, 2, ..., t.\end{aligned}$$

I want to emphasise that this side condition is not unique; in fact, there are infinitely many side conditions one could impose to "estimate" the parameters. Another commonly-used side condition (which is used by SAS) is to specify $\tau_t = 0$. In this case (verify!),

$$\begin{aligned}\widehat{\mu} &= \overline{Y}_{t+}\\ \widehat{\tau}_i &= \overline{Y}_{i+} - \overline{Y}_{t+}, \qquad i = 1, 2, ..., t-1.\end{aligned}$$

Note that the "estimates" under both constraints are different!

REMARK: At first glance, this may seem unfortunate; namely, that different side conditions lead to different least-squares estimates. All the restriction does is impose a particular interpretation on our linear model. For example, the condition $\sum_{i=1}^{t} n_i\tau_i = 0$ goes along with the interpretation of the τ_i as "deviations" from an overall mean. The treatments "affect" the response in different "directions;" some of the τ_i must be negative and others positive, for them to all sum to zero.

ESTIMABLE FUNCTIONS: There are certain functions of the model parameters that are always uniquely estimated, regardless of the side conditions used; these are called **estimable** functions. Estimability is an important concept in the theory of linear models.

EXAMPLE: The reason we didn't have non-unique solutions with the means model is that the normal equations could be solved uniquely. In addition, in the means model, we saw that μ_i was uniquely estimated by $\widehat{\mu}_i = \overline{Y}_{i+}$. However, in the effects model, with any suitable side condition, it turns out that $\mu_i = \mu + \tau_i$ is also uniquely estimated by $\overline{Y}_{i+}$, even though, individually, μ and τ_i are **not** estimable (verify!). The function $\mu_i \equiv \mu + \tau_i$ is an example of an **estimable function** in the effects model because it is always uniquely estimated. Those parametric functions (e.g., μ, τ_i, $\tau_i + \tau_{i'}$, etc.) whose least-squares estimates change with different side conditions are **not estimable**.

2.4 Model diagnostics

Recall our assumptions in the usual one-way layout; namely, (a) the observed responses are independent random samples from t populations, (b) the populations have normal distributions, and (c) the population variances are equal. We convey these assumptions in the error term ϵ_{ij} by saying "$\epsilon_{ij} \sim$ iid $\mathcal{N}(0, \sigma^2)$." Independence is hopefully conferred by the design of the experiment; namely, randomly assigning our random sample of experimental units to treatments and performing the experiment under identical conditions so that no other systematic sources of variability exist. However, how do we know if the normality and **homoscedastic** (i.e., constant variance) error assumptions are met in practice? These assumptions can be checked by looking at **residuals**.

RESIDUALS: Of course, we never get to actually see the ϵ_{ij}'s (i.e., the errors) because they are unobservable random variables. However, we can observe "proxies" of the errors; namely, the residuals. In the one-way layout, define

$$e_{ij} = y_{ij} - \overline{y}_{i+}$$

to be the **residual** associated with the observation y_{ij}. Recall that $\overline{y}_{i+}$ is the least-squares estimate of $\mu_i = \mu + \tau_i$.

REMARK: With nearly all statistical models, a residual represents the difference between the **observed** and the **predicted** (or **expected**) values. Here, the observed value is Y_{ij}. The predicted value is the least squares estimator for μ_i; namely, $\overline{Y}_{i+}$. This makes sense intuitively for the one-way classification model; we would expect an observation to be "close" to the mean (expected) $\overline{Y}_{i+}$; however, not all will equal the mean necessarily; these deviations in the observed and expected responses are the residuals.

A NOTE ON THE NOTATION: Make sure to recognise the difference between an **error** and a **residual**. They are not the same.

$$\begin{aligned} \epsilon_{ij} &= \text{error; can not be observed} \\ e_{ij} &= \text{residual; can be observed} \end{aligned}$$

Also, Rao uses E_{ij} to denote an error term. While there are good reasons for doing this (e.g., the use of a capital letter to emphasise the random aspect), the fact is I just don't like this notation! I'll use ϵ_{ij} instead.

DIAGNOSING NORMALITY: If we specify that $\epsilon_{ij} \sim$ iid $\mathcal{N}(0, \sigma^2)$ and the normality assumption holds, then the residuals (which, again, can be thought of proxies to the errors) are also normally distributed. In fact, it is not difficult to show that when the model holds, e_{ij}, when viewed as a **random variable**; i.e.,

$$e_{ij} = Y_{ij} - \overline{Y}_{i+} \sim \mathcal{N}\left[0, \sigma^2(1 - n_i^{-1})\right].$$

Thus, if the normality assumption is true, and the model holds, a histogram of the observed $e_{ij} = y_{ij} - \overline{y}_{i+}$ should look normally distributed, centered around zero.

NORMALITY PLOTS: Another way to diagnose normality is to use a **normal probability plot**; these are sometimes also called **quantile-quantile plots** (or **qq plots**). This plot is easy to construct. All you do is order the N observed residuals $e_{ij} = y_{ij} - \overline{y}_{i+}$ from lowest to smallest, say, $e_{(1)} \leq e_{(2)} \leq \cdots \leq e_{(N)}$, then plot the ordered residuals against the associated N ordered quantiles from the standard normal distribution; i.e., the N ordered Z values which delimit $N + 1$ equal areas. If the normality assumption holds, and the model is correct, this plot should look like a **straight line**.

REMARK: If the underlying distribution of the data is not normal, then, in theory, the rationale we used to develop the statistic is affected. However, in practice, ANOVA methods, in general, are pretty **robust** to the normality assumption. That is, as long as there are no "gross departures" from normality, you can feel fairly confident in the validity of the procedure. Of course, if there are gross departures from normality, the hypothesis test may be flawed. That is, the **true** significance level of the procedure may be much larger than α. In these instances, we might conclude that there is a treatment difference, when there really isn't! That is, we may think that we are seeing a difference in means, when actually we are just seeing a **serious** departure from normality.

OUTLIERS: Normal probability plots can also be used to detect **outliers**. Also, another check for outliers may be made by examining the **studentised residuals**

$$r_{ij} = \frac{y_{ij} - \overline{y}_{i+}}{\sqrt{\text{MS[E]}\left(1 - \frac{1}{n_i}\right)}}.$$

When the model holds, $r_{ij} \sim t_{N-t}$. Thus, when N is large relative to t, a studentised residual larger than 3 or 4 (in absolute value) is a potential outlier.

DIAGNOSING NONCONSTANT VARIANCE: If the variances across treatment groups are not the same, then the rationale we used to develop the F statistic is lost. *A violation in this assumption can be much more serious than departures from normality.* Small departures aren't too problematic in balanced designs. However, in unbalanced designs, especially where the variances are very different, the overall F test may have a **seriously** inflated Type I Error rate.

MULTIPLICATIVE MODELS: There are some physical situations where a more plausible model is not additive, but **multiplicative**; that is,

$$Y_{ij} = \mu^* \tau_i^* \epsilon_{ij}^*.$$

Such a model is often appropriate for growth data, or in other situations where the variability in response tends to get larger as the response becomes larger.

Example 2.3. Consider a one-way classification experiment with five different dose levels of a growth hormone used in mice. The higher the level of dose, the higher the concentration of the hormone. In Figure 2.9, we see that the variation levels are not

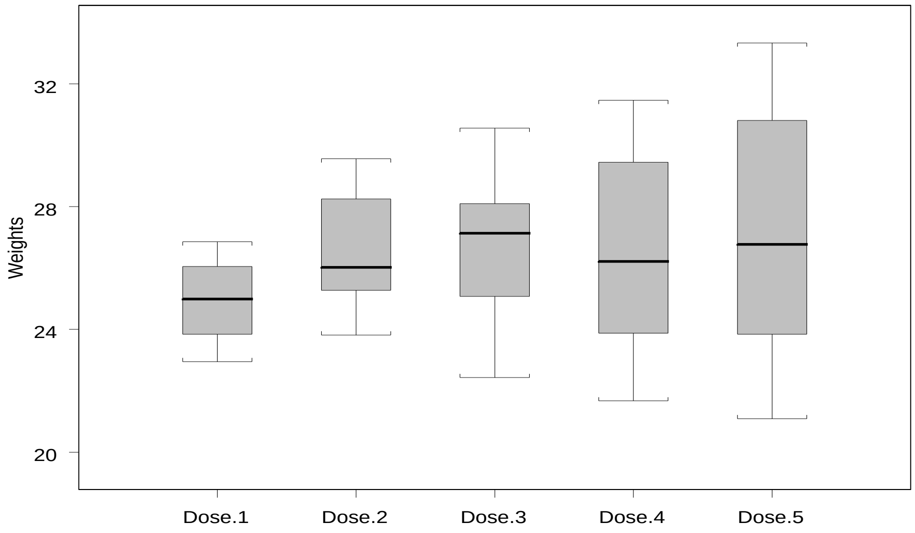

Figure 2.9: *Weight gain data at increasing dose levels.*

constant for the different doses; rather, there seems to be an increasing trend in the variation levels! That is, the **constant variance** assumption appears to be violated.

RESIDUAL PLOTS: A good visual display to use for diagnosing nonconstant variance is the plot of residuals versus predicted values. This is sometimes called a **residual plot**. If the model holds, then one can show that

$$\text{Cov}(e_{ij}, \overline{Y}_{i+}) = 0;$$

i.e., the residuals and predicted values are **uncorrelated**. Thus, residual plots that display nonrandom patterns suggest that there are some problems with our model assumptions.

TESTS FOR EQUAL VARIANCES: There are two procedures for testing

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_t^2$$

versus

$$H_1 : \text{the } \sigma_i^2 \text{ are not all equal.}$$

One procedure is **Barlett's test**; however, it assumes normality and is not robust to departures from it. **Levene's test**, a nonparametric procedure, makes no distributional assumption on the error terms, and is often preferred in practice.

2.5 Analysis of transformed data

The usual approach for dealing with nonconstant variance, when it occurs, is to apply a **variance-stabilising transformation** and then run the ANOVA on the transformed data. We now outline how such transformations are derived. Suppose that Y is a random variable with mean $E(Y) = \mu$ and variance $V(Y) = v(\mu)$, some function of the mean. Of course, as μ changes, so does $V(Y)$, which violates the homoscedastic assumption, and, hence, this is the problem! So, the idea is to find a function of Y, say, $g(Y)$ so that $V[g(Y)]$ is constant. The function g is the variance-stabilising transformation.

Here are some common probability distributions where the variance $V(Y)$ is a function of $E(Y)$.

Model	Mean	Variance, $v(\mu)$	Data description	$g(Y)$
Poisson	$E(Y) = \mu$	$V(Y) = \mu$	count data	$\sqrt{Y}$
Exponential	$E(Y) = \mu$	$V(Y) = \mu^2$	time to failure data	$\log Y$
Bernoulli	$E(Y) = \mu$	$V(Y) = \mu(1-\mu)$	proportion of successes	$\sin^{-1}(\sqrt{Y})$

FINDING THE TRANSFORMATION: Write a one-term Taylor series expansion of $g(Y)$ about the point μ as

$$g(Y) \approx g(\mu) + g'(\mu)(Y - \mu).$$

Then, using this linear approximation, it follows that

$$V[g(Y)] \approx [g'(\mu)]^2 v(\mu),$$

where $V(Y) = v(\mu)$. Setting $[g'(\mu)]^2 v(\mu)$ equal to a constant, say, c, which is free of μ, and solving for g, we get

$$g(\mu) = \int \frac{c_0}{\sqrt{v(\mu)}}\, d\mu,$$

where $c_0 = \sqrt{c}$ (still just a constant). For example, if our response is really Poisson, then

$$g(\mu) = \int \frac{c_0}{\sqrt{\mu}}\, d\mu = 2c_0\sqrt{\mu} + c_1,$$

where c_1 is a constant free of μ. Taking $c_0 = \frac{1}{2}$ and $c_1 = 0$ gives $g(\mu) = \sqrt{\mu}$. Thus, the square root transformation is the appropriate variance-stabilising transformation for Poisson data. Rao gives two good examples (Examples 8.9 and 8.10) on pages 308-311.

BOX-COX TRANSFORMATION: The power transformation

$$g(Y) = \begin{cases} \log Y, & \lambda = 0 \\ Y^{\lambda}, & \lambda > 0 \end{cases}$$

was suggested by Box and Cox (1964). The log and square root transformations are special cases with $\lambda = 0$ and $\lambda = 1/2$, respectively. Approximate $100(1-\alpha)$ percent confidence intervals for λ are available.

DISCLAIMER: It is important to remember that, when analysing data on a different scale, the conclusions of the analysis apply to the transformed scale. Furthermore, one should remember that transforming the data may fix one problem, but it may create other violations of the model.

OTHER PLOTS: Plotting residuals in time order of data collection is helpful to detect correlation between them; having “runs” of positive and negative residuals might challenge the independence assumption. For example, suppose we were sampling plants in contiguous regions of a field. Plants in the same row, say, may be “more alike” than those farther apart, so we might expect for observations to be correlated. In general, strong correlation among the errors can be problematic, so it is important to prevent the problem by using a proper design. In this example, it may have been better to incorporate “row-effects” into the analysis by using rows as blocks beforehand.

NONPARAMETRIC METHODS: As an alternative to model-based inference procedures, nonparametric methods can be useful. In general, such methods require little or no distributional assumptions on the error terms. The nonparametric analogue of the one-way ANOVA F test is the Kruskal-Wallis test (see § 8.8 if you are interested).

3 Comparing Treatment Means and Multiple Comparisons in the One-Way Layout

Complimentary reading from Rao: Chapter 9 (§ 9.1-9.6).

3.1 Introduction

You will recall that, in the last chapter, we developed an overall procedure (i.e., **an overall F test**) for testing

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_t$$
$$\text{versus}$$
$$H_1 : \text{the } \mu_i \text{ are not all equal}$$

in the **one-way classification model** with complete randomisation. We saw that large values of the F statistic led to the rejection of H_0.

REMARK: In the overall test, rejecting H_0 simply tells us that there is a difference somewhere among the t means; it does not say where the difference is. Because of this, the overall F test is largely uninformative. In fact, if H_0 is rejected, have we really learned anything new? Put another way, does anyone think that all t means are ever going to be *identical*? Because of this, some experimenters choose to skip the overall test altogether and restrict attention only to specific comparisons among the means under consideration. Such comparisons generally fall into one of three categories:

1. **pre-planned** comparisons; i.e., those comparisons planned before the experiment,
2. **unplanned** comparisons; those comparisons specified after the experiment has been conducted, and
3. all possible **pairwise** comparisons.

Example 3.1. In the pea experiment from Example 2.2, suppose that the researcher wanted to investigate these particular issues: (a) whether or not the fructose and glucose only treatments (2 and 4) were different, (b) whether or not the treatments involving fructose and/or glucose (2, 3, and 4) were different than with the sucrose treatment (5), (c) whether or not the sugar treatments (2, 3, 4, and 5) differed from the control (1), and (d) whether or not the fructose and glucose only treatments (2 and 4) are different from the glucose/fructose combination treatment (3). If the investigator had decided to make these comparisons before the experiment was conducted (i.e., before the data were collected), then this would fall into the category of **pre-planned comparisons**. If it was decided to make any or all of these comparisons after the data were observed, these would be **unplanned comparisons**.

3.2 Pre-planned comparisons using contrasts

We consider those comparisons which have been specified before the experiment has been conducted. Comparing two or more treatments in this way may be done using **contrasts**. Contrasts are widely-used with ANOVA models because they can be used to explain the differences in treatment means. In fact, it is often the case that specific research questions can be addressed by examining the “right” contrasts.

TERMINOLOGY: Let $\mu_1, \mu_2, ..., \mu_t$ denote the population means in a one-way classification model, and suppose that $c_1, c_2, ..., c_t$ are known constants. The linear combination

$$\theta = \sum_{i=1}^{t} c_i\mu_i = c_1\mu_1 + c_2\mu_2 + \cdots + c_t\mu_t,$$

with the restriction that $\sum_{i=1}^{t} c_i = 0$, is called a **contrast**. In the one-way layout model, all contrasts are estimable. Since $\overline{Y}_{i+}$ is the least-squares estimator of μ_i, $i = 1, 2, ..., t$, it follows that

$$\widehat{\theta} = \sum_{i=1}^{t} c_i\overline{Y}_{i+} = c_1\overline{Y}_{1+} + c_2\overline{Y}_{2+} + \cdots + c_t\overline{Y}_{t+}$$

is the least-squares estimator of θ. The least-squares estimator $\widehat{\theta}$ is called a **contrast of sample means**.

Example 3.1 (continued). Each of the preplanned comparisons in Example 3.1 is now considered. To investigate

- whether or not the fructose and glucose only treatments (2 and 4) were different, we may specify
$$\theta_1 = \mu_2 - \mu_4.$$
Note that this is a contrast with $c_2 = 1$, $c_4 = -1$, and $c_1 = c_3 = c_5 = 0$.

- whether or not the treatments involving fructose and/or glucose (2, 3, and 4) were different than with the sucrose treatment (5), we may specify
$$\theta_2 = \frac{1}{3}(\mu_2 + \mu_3 + \mu_4) - \mu_5.$$
Note that this is a contrast with $c_1 = 0$, $c_2 = c_3 = c_4 = \frac{1}{3}$, and $c_5 = -1$.

- whether or not the sugar treatments (2, 3, 4, and 5) differed from the control (1), we may specify
$$\theta_3 = \mu_1 - \frac{1}{4}(\mu_2 + \mu_3 + \mu_4 + \mu_5).$$
Note that this is a contrast with $c_1 = 1$ and $c_2 = c_3 = c_4 = c_5 = -\frac{1}{4}$.

- whether or not the fructose and glucose only treatments (2 and 4) are different from the glucose/fructose combination treatment (3), we may specify
$$\theta_4 = \frac{1}{2}(\mu_2 + \mu_4) - \mu_3$$
Note that this is a contrast with $c_2 = c_4 = \frac{1}{2}$, $c_3 = -1$, and $c_1 = c_5 = 0$.

SAMPLING DISTRIBUTION OF $\widehat{\theta}$: First, recall that under our usual one-way model assumptions, $\overline{Y}_{i+} \sim \mathcal{N}(\mu_i, \sigma^2/n_i)$. Thus, $\widehat{\theta}$, too, is normally distributed since it is just a linear combination of the (sample) treatment means $\overline{Y}_{1+}, \overline{Y}_{2+}, ..., \overline{Y}_{t+}$. The mean of $\widehat{\theta}$ is given by
$$E(\widehat{\theta}) = E\left(\sum_{i=1}^{t} c_i \overline{Y}_{i+}\right) = \sum_{i=1}^{t} c_i E(\overline{Y}_{i+}) = \sum_{i=1}^{t} c_i \mu_i = \theta,$$

and the variance of $\widehat{\theta}$ is

$$V(\widehat{\theta}) = V\bigg(\sum_{i=1}^{t} c_i\overline{Y}_{i+}\bigg) = \sum_{i=1}^{t} c_i^2 V(\overline{Y}_{i+}) = \sigma^2 \sum_{i=1}^{t} \frac{c_i^2}{n_i}.$$

Thus, the quantity

$$\widehat{\theta} \sim \mathcal{N}\bigg(\theta, \sigma^2 \sum_{i=1}^{t} \frac{c_i^2}{n_i}\bigg).$$

CONFIDENCE INTERVALS AND HYPOTHESIS TESTS: Standardising $\widehat{\theta}$, we get

$$Z = \frac{\widehat{\theta} - \theta}{\sqrt{\sigma^2\Big(\sum_{i=1}^{t} \frac{c_i^2}{n_i}\Big)}} \sim \mathcal{N}(0,1),$$

and

$$t = \frac{\widehat{\theta} - \theta}{\sqrt{\text{MS[E]}\Big(\sum_{i=1}^{t} \frac{c_i^2}{n_i}\Big)}} = \frac{\frac{\widehat{\theta}-\theta}{\sqrt{\sigma^2\big(\sum_{i=1}^{t} \frac{c_i^2}{n_i}\big)}}}{\sqrt{\frac{\text{SS[E]}}{\sigma^2}/(N-t)}} \sim t_{N-t}.$$

Therefore, using t as a pivot, a $100(1-\alpha)$ percent confidence interval for θ is given by

$$\widehat{\theta} \pm t_{N-t,\alpha/2} \underbrace{\sqrt{\text{MS[E]}\bigg(\sum_{i=1}^{t} \frac{c_i^2}{n_i}\bigg)}}_{\widehat{\sigma}_{\widehat{\theta}},\ \text{standard error of } \widehat{\theta}}.$$

In addition, the hypothesis test with $H_0 : \theta = \theta_0$ (usually, $\theta_0 = 0$) can be performed by using

$$t = \frac{\widehat{\theta} - \theta_0}{\sqrt{\text{MS[E]}\Big(\sum_{i=1}^{t} \frac{c_i^2}{n_i}\Big)}}$$

as a test statistic with an appropriate rejection region from the t_{N-t} reference distribution.

Example 3.2 (continuation of Example 3.1). Before the experiment was conducted, it was decided to compare the fructose and glucose only treatments. To be specific, the researcher decided to test $H_0 : \theta_1 = 0$ versus $H_1 : \theta_1 \neq 0$, where the contrast $\theta_1 = \mu_2 - \mu_4$. Its least-squares estimator is $\widehat{\theta}_1 = \overline{Y}_{2+} - \overline{Y}_{4+}$. The t statistic is given by

$$t = \frac{\widehat{\theta}_1 - \theta_0}{\sqrt{\text{MS[E]}\Big(\sum_{i=1}^{t} \frac{c_i^2}{n_i}\Big)}} = \frac{(58.2 - 59.3) - 0}{\sqrt{5.46\left[\frac{1^2}{10} + \frac{(-1)^2}{10}\right]}} = -1.05,$$

which has a two-sided probability value of $P \approx 0.298$. There does not appear to be a significant difference between the fructose and glucose treatments with respect to their affect on growth. With $t_{45,0.025} \approx 2.0141$, a 95 percent confidence interval for θ is given by

$$(58.2 - 59.3) \pm 2.0141\sqrt{5.46\left[\frac{1^2}{10} + \frac{(-1)^2}{10}\right]}, \text{ or } (-3.20, 1.00).$$

SUMS OF SQUARES FOR CONTRASTS: Suppose that θ is a contrast. We have learned that testing $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$ can be performed using a two-sided t test (see Example 3.2). However, recall the relationship between the t and F distributions; namely, that $t_\nu^2 = F_{1,\nu}$. Thus, a **two**-sided t test and a **one**-sided F test are **equivalent** testing procedures. In this light, we can test H_0 versus H_1 by

- rejecting H_0, at level α, when

$$|t| = \left|\frac{\widehat{\theta}}{\sqrt{\text{MS[E]}\left(\sum_{i=1}^t \frac{c_i^2}{n_i}\right)}}\right| \geq t_{N-t,\alpha/2} \iff |\widehat{\theta}| \geq \underbrace{t_{N-t,\alpha/2}\sqrt{\text{MS[E]}\left(\sum_{i=1}^t \frac{c_i^2}{n_i}\right)}}_{\text{Fisher CCV}}$$

- or, equivalently, rejecting H_0, at level α, when

$$F = t^2 = \frac{\widehat{\theta}^2}{\text{MS[E]}\left(\sum_{i=1}^t \frac{c_i^2}{n_i}\right)} \geq F_{1,N-t,\alpha}.$$

TERMINOLOGY: We define the **sum of squares for** $\widehat{\theta}$ as

$$\text{SS}(\widehat{\theta}) = \frac{\widehat{\theta}^2}{\sum_{i=1}^t \frac{c_i^2}{n_i}}.$$

It is not difficult to argue that when $H_0 : \theta = 0$ is true, $\text{SS}(\widehat{\theta})/\sigma^2 \sim \chi_1^2$ (verify!). Also, since $\text{SS}(\widehat{\theta})$ has only one degree of freedom associated with it, $\text{MS}(\widehat{\theta}) = \text{SS}(\widehat{\theta})$. Thus,

$$F = \frac{\text{MS}(\widehat{\theta})}{\text{MS[E]}} = \frac{\text{SS}(\widehat{\theta})/\sigma^2}{\frac{\text{SS[E]}}{\sigma^2}/(N-t)} \sim F_{1,N-t},$$

since $\text{SS}[E]/\sigma^2 \sim \chi_{N-t}^2$ and $\text{SS}(\widehat{\theta})$ and SS[E] are independent statistics. Note that this quantity can be interpreted as a ratio of two mean squares; one for $\widehat{\theta}$ and one for error.

Theorem. Multiplying the coefficients of any contrast $\widehat{\theta}$ by a constant does not change the value of $\text{SS}(\widehat{\theta})$. That is, $\text{SS}(\widehat{\theta}) = \text{SS}(a\widehat{\theta})$, for any nonzero constant a.

Example 3.3 (continuation of Example 3.2). To illustrate the F test for a single degree of freedom contrast (any contrast has one degree of freedom associated with it), we will revisit Example 3.2 and test $H_0 : \theta_1 = 0$ versus $H_1 : \theta_1 \neq 0$ where, recall, $\theta_1 = \mu_2 - \mu_4$ is a contrast. For the pea data, we have $\widehat{\theta}_1^2 = (\overline{y}_{2+} - \overline{y}_{4+})^2 = (58.2 - 59.3)^2 = 1.21$; thus, $\text{SS}(\widehat{\theta}_1)$ is given by

$$\text{SS}(\widehat{\theta}_1) = \frac{\widehat{\theta}^2}{\sum_{i=1}^t \frac{c_i^2}{n_i}} = \frac{1.21}{\frac{1}{10}[1^2 + (-1)^2]} = 6.05.$$

Finally,

$$F = \frac{\text{MS}(\widehat{\theta}_1)}{\text{MS}[E]} = \frac{6.05}{5.46} = 1.11,$$

which is (up to rounding error) the square of $t = -1.05$ in Example 3.2. What is the P value?

ORTHOGONALITY: Two contrasts of sample means

$$\widehat{\theta}_1 = \sum_{i=1}^t c_i\overline{Y}_{i+} = c_1\overline{Y}_{1+} + c_2\overline{Y}_{2+} + \cdots + c_t\overline{Y}_{t+}$$

and

$$\widehat{\theta}_2 = \sum_{i=1}^t d_i\overline{Y}_{i+} = d_1\overline{Y}_{1+} + d_2\overline{Y}_{2+} + \cdots + d_t\overline{Y}_{t+}$$

are said to be **orthogonal** if

$$\sum_{i=1}^t \frac{c_i d_i}{n_i} = 0.$$

When $n_i = n$ for all i; i.e., the design is **balanced**, then the last equation reduces to

$$\boldsymbol{c}'\boldsymbol{d} = \sum_{i=1}^t c_i d_i = 0,$$

where $\boldsymbol{c} = (c_1, c_2, ..., c_t)'$ and $\boldsymbol{d} = (d_1, d_2, ..., d_t)'$. A set of contrasts, say, $\widehat{\theta}_1, \widehat{\theta}_2, ..., \widehat{\theta}_k$, $k < t$, is said to be a **mutually orthogonal set** if $\widehat{\theta}_j$ and $\widehat{\theta}_{j'}$ are orthogonal for all $j \neq j'$.

REMARK: Contrasts are only of practical interest when they define interesting functions of the μ_i's. Orthogonal contrasts are most useful in **balanced** problems because a set of orthogonal contrasts can retain interesting interpretations (like in Example 3.1). In unbalanced designs, orthogonality depends on the unequal n_i's, so there is rarely more than one interpretable contrast in a set of orthogonal contrasts.

Example 3.4. In Example 3.1, we defined a set of four contrasts $\{\theta_1, \theta_2, \theta_3, \theta_4\}$. Table 3.6 displays the contrast coefficients. It follows that $\{\theta_1, \theta_2, \theta_3, \theta_4\}$ is a mutually orthogonal set of contrasts since

$$\sum_{i=1}^{5} c_{ij}c_{ij'} = 0,$$

for all $j \neq j'$; $j, j' = 1, 2, 3, 4$.

Table 3.6: *Table of contrast coefficients for Example* 3.1.

θ_j	c_{1j}	c_{2j}	c_{3j}	c_{4j}	c_{5j}
θ_1	0	1	0	-1	0
θ_2	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	-1
θ_3	1	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$
θ_4	0	$\frac{1}{2}$	-1	$\frac{1}{2}$	0

Theorem. The sums of squares for treatments is always greater or equal to the sums of squares for any contrast; that is, for any contrast, $\widehat{\theta}$, $\text{SS[T]} \geq \text{SS}(\widehat{\theta})$.

Proof. Assume that $n_i = n$ for all i (i.e., a balanced design) and let $\widehat{\theta} = \sum_{i=1}^{t} c_i\overline{Y}_{i+}$ be any contrast. For any vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{R}^t$, recall that $|\boldsymbol{u}'\boldsymbol{v}| \leq ||\boldsymbol{u}||||\boldsymbol{v}||$, where $||\boldsymbol{u}|| = \sqrt{\boldsymbol{u}'\boldsymbol{u}}$ denotes the norm of $\boldsymbol{u}$ (this is just the Cauchy-Schwartz Inequality). Define

$$\boldsymbol{u} = (c_1, c_2, ..., c_t)' \quad \text{and} \quad \boldsymbol{v} = (\overline{Y}_{1+} - \overline{Y}_{++}, \overline{Y}_{2+} - \overline{Y}_{++}, ..., \overline{Y}_{t+} - \overline{Y}_{++})'$$

and note that

$$\text{SS}(\widehat{\theta}) = \frac{\widehat{\theta}^2}{\sum_{i=1}^{t} \frac{c_i^2}{n}} = \frac{n(\boldsymbol{u}'\boldsymbol{v})^2}{\boldsymbol{u}'\boldsymbol{u}} \leq \frac{n\boldsymbol{u}'\boldsymbol{u}\boldsymbol{v}'\boldsymbol{v}}{\boldsymbol{u}'\boldsymbol{u}} = n\boldsymbol{v}'\boldsymbol{v} = n\sum_{i=1}^{t}(\overline{Y}_{i+} - \overline{Y}_{++})^2 = \text{SS[T]},$$

since $\widehat{\theta} = \sum_{i=1}^{t} c_i\overline{Y}_{i+} = \sum_{i=1}^{t} c_i(\overline{Y}_{i+} - \overline{Y}_{++})$. Since $\widehat{\theta}$ is arbitrary, the result holds for balanced designs. In unbalanced designs, the result is still true, but is more difficult to prove. □

Theorem. There is always one contrast that accounts for all of SS[T]; that is, there exists a contrast, say, $\widehat{\theta}^*$ such that $\text{SS[T]} = \text{SS}(\widehat{\theta}^*)$.

REMARK: That there exists such a contrast that accounts for all the of sums of squares for treatments is theoretically interesting and useful (it establishes that Scheffe's method is legitimate, as we'll see later). From a practical standpoint, however, this magic contrast is usually not interpretable.

Theorem. Let $\{\widehat{\theta}_1, \widehat{\theta}_2, ..., \widehat{\theta}_{t-1}\}$ be a set of $t-1$ **mutually orthogonal contrasts** in the one-way layout. Then,

$$\text{SS[T]} = \text{SS}(\widehat{\theta}_1) + \text{SS}(\widehat{\theta}_2) + \cdots + \text{SS}(\widehat{\theta}_{t-1}).$$

That is, the sums of squares for treatments can be broken into components corresponding to sums of squares for individual orthogonal contrasts.

Example 3.5 (continuation of Example 3.4) (`pea-mc.sas`). Consider the set of mutually orthogonal contrasts $\{\theta_1, \theta_2, \theta_3, \theta_4\}$. In Example 3.3, we computed $\text{SS}(\widehat{\theta}_1) = 6.05$. Analogous calculations show (verify yourself!) that $\text{SS}(\widehat{\theta}_2) = 235.20$, $\text{SS}(\widehat{\theta}_3) = 832.32$, and $\text{SS}(\widehat{\theta}_4) = 3.75$. Since the contrasts are all orthogonal, it follows that

$$\text{SS}(\widehat{\theta}_1) + \text{SS}(\widehat{\theta}_2) + \text{SS}(\widehat{\theta}_3) + \text{SS}(\widehat{\theta}_4) = 6.05 + 235.20 + 832.32 + 3.75 = 1077.32 = \text{SS[T]}.$$

Furthermore, the F statistics associated with each contrast,

$$F_j = \frac{\text{SS}(\widehat{\theta}_j)}{\text{MS[E]}} = \frac{\text{MS}(\widehat{\theta}_j)}{\text{MS[E]}},$$

along with right-tail P values are provided in Table 3.7. It looks as though most of the observed variability in the treatment means is due to the second and third contrasts. However, we should make a special note of the following fact: the conclusions that we draw from Table 3.7 have not been adjusted for **multiplicity** in any way. *That is, we have not accounted for the fact that we are making multiple statements here about the variability among the treatment means* (in the form of these contrasts). The upshot of this result is that we do not know what our (overall) Type I Error rate is for the **family** of tests $H_0 : \theta_j = 0$ versus $H_1 : \theta_j \neq 0$; $j = 1, 2, 3, 4$. We could make a conclusion about

Table 3.7: *F statistics for Example* 3.5.

θ_j	$\widehat{\theta}_j$	$\text{SS}(\widehat{\theta}_j)$	F_j	P	Decision
θ_1	-1.10	6.05	1.11	0.298	Not significant
θ_2	-5.60	235.20	43.11	<0.0001	Significant
θ_3	10.20	832.32	152.56	<0.0001	Significant
θ_4	0.75	3.75	0.69	0.411	Not significant

a single contrast (of course, provided that it was a preplanned comparison); however, we can not make any joint statements about two or more the contrasts in this problem, and have our overall error rate be controlled appropriately. Methods for adjusting for multiplicity will be handled later.

3.3 Testing single contrasts suggested by the data

When contrasts are **preplanned**, we can use the Fisher critical contrast value (CCV) to test $H_0 : \theta = 0$ versus $H_0 : \theta \neq 0$ as mentioned earlier. Such a test declares $\widehat{\theta}$ significant when

$$|t| = \left| \frac{\widehat{\theta}}{\sqrt{\text{MS[E]}\left(\sum_{i=1}^{t} \frac{c_i^2}{n_i}\right)}} \right| \geq t_{N-t,\alpha/2} \Longleftrightarrow |\widehat{\theta}| \geq \underbrace{t_{N-t,\alpha/2}\sqrt{\text{MS[E]}\left(\sum_{i=1}^{t} \frac{c_i^2}{n_i}\right)}}_{\text{Fisher CCV}}$$

If the decision to test the significance of $\widehat{\theta}$ without regard to the observed outcome of the experiment, then this test guarantees that the probability of falsely declaring significance is α. A $100(1-\alpha)$ percent confidence interval for θ, based on the Fisher's method of testing preplanned contrasts, as derived earlier, is given by

$$\widehat{\theta} \pm t_{N-t,\alpha/2}\sqrt{\text{MS[E]}\left(\sum_{i=1}^{t} \frac{c_i^2}{n_i}\right)}.$$

However, if the decision is made to test $H_0 : \theta = 0$ versus $H_0 : \theta \neq 0$ *after seeing the data*, these methods are no longer appropriate!

THE PROBLEM WITH "DATA-SNOOPING": Suppose we decide to examine the sample means $\overline{Y}_{1+}, \overline{Y}_{2+}, ..., \overline{Y}_{t+}$ and compare only those that appear to be different. That is, we are making those comparisons "suggested" by the data. To see why this is not a valid method of analysis, suppose we decide to conduct a t test at level $\alpha = 0.05$ for a difference in the two treatments observed to have the highest and lowest sample means among all t treatments. Since the data are just random samples from the treatment populations of interest, the sample means $\overline{Y}_{i+}$ could have ended up the way they did because

- there *really is a difference* in the population means, or
- we have "unusual" samples, and there is not a difference in the population means.

Of course, since chance is involved, either of these explanations is possible. It turns out that we will still be more likely to reject the null hypothesis of no difference in the two extreme means, even if they are the really the same! Put another way, it is not legitimate to "find the best and worst treatments, compare them, and say that your error rate is five percent." In fact, with $\alpha = 0.05$, it turns out that the true error rate in this instance (comparing the best treatment to the worst treatment after the experiment has been performed) is actually about 0.13 if $t = 3$, about 0.60 if $t = 10$, and about 0.90 if $t = 20$!

THE SCHEFFE METHOD OF TESTING SUGGESTED CONTRASTS: When a contrast *is suggested by the data*, we can use the Scheffe critical contrast value (CCV) to test $H_0 : \theta = 0$ versus $H_0 : \theta \neq 0$. Such a test declares $\widehat{\theta}$ significant when

$$|\widehat{\theta}| \geq \underbrace{\sqrt{(t-1)F_{t-1,N-t,\alpha}} \times \widehat{\sigma}_{\widehat{\theta}}}_{\text{Scheffe CCV}},$$

where

$$\widehat{\sigma}_{\widehat{\theta}} = \sqrt{\text{MS[E]}\left(\sum_{i=1}^{t} \frac{c_i^2}{n_i}\right)}.$$

Scheff's procedure allows us to estimate *all possible contrasts simultaneously* at level α. Since we are only doing one, the method is always conservative; i.e., the error rate may be much less than the nominal α level. Thus, if the contrast is truly significant, Scheffe's method may have a hard time detecting it. This is the price to pay for controlling the error rate; however, on the flip side, you can be extra confident that, in fact, you are not making a Type I Error!

SCHEFFE CONFIDENCE INTERVALS: A $100(1-\alpha)$ percent confidence interval for θ, based on the Scheffe method of testing suggested contrasts is given by

$$\widehat{\theta} \pm \sqrt{(t-1)F_{t-1,N-t,\alpha}} \times \widehat{\sigma}_{\widehat{\theta}}.$$

Example 3.6 (`pea-mc.sas`). Suppose that after seeing the pea-data, our researcher decides to test whether or not the fructose-only and sucrose treatments are statistically different, using $\alpha = 0.05$. Since this comparison was not preplanned, he must use Scheffe's method of testing contrasts. This comparison is done by using the contrast $\theta = \mu_2 - \mu_5$. We have $\widehat{\theta} = \overline{y}_{2+} - \overline{y}_{5+} = 58.2 - 64.1 = -5.9$. Note that

$$\widehat{\sigma}_{\widehat{\theta}} = \sqrt{\text{MS[E]}\left(\sum_{i=1}^{t}\frac{c_i^2}{n_i}\right)} = \sqrt{5.46\left[\frac{1^2}{10} + \frac{(-1)^2}{10}\right]} = 1.045.$$

With $F_{4,45,0.05} = 2.5787$, Scheffe's CCV is given by

$$\sqrt{(t-1)F_{t-1,N-t,\alpha}} \times \widehat{\sigma}_{\widehat{\theta}} = \sqrt{(5-1)F_{4,45,0.05}} \times 1.045 = 3.355.$$

Since $|\widehat{\theta}| = 5.9 > 3.355$, we would reject $H_0 : \theta = 0$ at the $\alpha = 0.05$ level; that is, there is a statistical difference between the expected growth for pea stems receiving the fructose-only and sucrose treatments.

REMARK: It is interesting to note that in Example 3.6, had the comparison of the fructose-only and sucrose means been **preplanned**, the Fisher CCV could have been used; this value is given by

$$t_{45,0.025} \times \widehat{\sigma}_{\widehat{\theta}} = 2.0141 \times 1.045 = 2.104,$$

which is much smaller than Scheffe's CCV. Indeed, this should make sense intuitively. If comparisons are to be made after seeing the data, we should have overwhelming evidence to reject H_0—much more than had we decided to make the comparison beforehand.

3.4 Multiple comparisons of means in the one-way layout

Whenever we construct a confidence interval or perform a hypothesis test, there is a built-in chance for error. The unfortunate reality is that the more inferences we perform, the more likely we are to commit an error. The purpose of **multiple comparison methods** is to control the probability of making a certain type of error.

THE COMPARISONWISE ERROR RATE: Suppose that we have a set of contrasts in mind, say, $\theta_1, \theta_2, ..., \theta_k$, and that we would like to test $H_0 : \theta_j = 0$ versus $H_1 : \theta_j \neq 0$ for $j = 1, 2, ..., k$. In a multiple-comparisons procedure, the expected **proportion** of contrasts that will be declared as significant when there really are no differences (i.e., all the H_0's are true) is called the **comparisonwise error rate**; i.e.,

$$\text{Comparisonwise error rate } (H_0 \text{ true}) = \frac{\text{number of erroneous inferences}}{\text{number of inferences made}}.$$

A multiple comparisons procedure in which each **preplanned** contrast is tested at level α will have a comparisonwise error rate of α. This is true because if the probability of a Type I Error is α for each contrast, the proportion of Type I Errors committed (in the set) is also α. Thus, the Fisher multiple comparison method controls the comparisonwise error rate for a set of k contrasts.

FISHER'S MULTIPLE COMPARISON METHOD: Let $\widehat{\theta}_1$, $\widehat{\theta}_2$, ..., $\widehat{\theta}_k$ denote a set of k contrasts in a one-way layout. The multiple comparison procedure that declares $\widehat{\theta}_j$ significant if

$$|\widehat{\theta}_j| \geq \underbrace{t_{N-t,\alpha/2} \times \widehat{\sigma}_{\widehat{\theta}_j}}_{\text{Fisher CCV}},$$

where

$$\widehat{\sigma}_{\widehat{\theta}_j} = \sqrt{\text{MS[E]} \left(\sum_{i=1}^{t} \frac{c_i^2}{n_i} \right)},$$

has a **comparisonwise** error rate α. This only applies when the contrasts are **preplanned**. Confidence intervals for θ_j are given by $\widehat{\theta}_j \pm t_{N-t,\alpha/2} \times \widehat{\sigma}_{\widehat{\theta}_j}$, for $j = 1, 2, ..., k$.

Example 3.7. We now compute the Fisher CCVs for the four contrasts in Example 3.1 using $\alpha = 0.05$. Recall that $t_{N-t,\alpha/2} = t_{45,0.025} = 2.0141$. For our four contrasts,

Table 3.8: *Fisher multiple comparisons for the pea growth data.*

θ_j	$\widehat{\theta}_j$	$\widehat{\sigma}_{\widehat{\theta}_j}$	$t_{45,0.025}$	Fisher's CCV	Decision
θ_1	-1.10	1.045	2.0141	2.015	Not significant
θ_2	-5.60	0.853	2.0141	1.718	Significant
θ_3	10.20	0.826	2.0141	1.664	Significant
θ_4	0.75	0.905	2.0141	1.822	Not significant

$\widehat{\sigma}_{\widehat{\theta}_1} \approx 1.045$, $\widehat{\sigma}_{\widehat{\theta}_2} \approx 0.853$, $\widehat{\sigma}_{\widehat{\theta}_3} \approx 0.826$, and $\widehat{\sigma}_{\widehat{\theta}_4} \approx 0.905$ (verify!). Values of least-squares estimates, standard errors, and Fisher's CCVs are given in Table 3.8. Provided that these were preplanned comparisons, we could conclude that the contrasts $\widehat{\theta}_2$ and $\widehat{\theta}_3$ are significantly different from zero at the $\alpha = 0.05$ **comparisonwise** error rate.

DRAWBACK WITH COMPARISONWISE CONTROL: The comparisonwise error rate is rarely quoted in practice because of the following reason: the expected number of false significances (i.e., rejecting H_0 when H_0 is true) depends on k, the number of contrasts tested, and increases as k does.

THE EXPERIMENTWISE ERROR RATE: In practice, one usually prefers to control a different error rate; namely, *one that expresses the probability of making an error in any of the tests* (when all of the H_0's are true). Limiting this probability is referred to as control of the **experimentwise error rate**. The experimentwise error rate of a multiple comparisons procedure is a natural generalisation of the Type I Error rate associated with the significance level of a single contrast; i.e.,

$$\text{Experimentwise error rate } (H_0 \text{ true}) = \frac{\#\text{ of experiments with} \geq 1 \text{ erroneous inference}}{\#\text{ of experiments conducted}}.$$

A multiple comparisons procedure is said to have an experimentwise error rate α if the probability of declaring **at least one** false significance when testing k contrasts that are not truly significant (i.e., all the H_0's are true) is α.

IMPORTANT: Fisher's method does **not** control the experimentwise error rate!!

THE SCHEFFE MULTIPLE COMPARISON METHOD: Let $\widehat{\theta}_1$, $\widehat{\theta}_2$, ..., $\widehat{\theta}_k$ denote a set of k contrasts in a one-way layout. The multiple comparison procedure that declares $\widehat{\theta}_j$ significant if

$$|\widehat{\theta}_j| \geq \underbrace{\sqrt{(t-1)F_{t-1,N-t,\alpha}} \times \widehat{\sigma}_{\widehat{\theta}_j}}_{\text{Scheffe CCV}},$$

where

$$\widehat{\sigma}_{\widehat{\theta}_j} = \sqrt{\text{MS[E]}\left(\sum_{i=1}^{t}\frac{c_i^2}{n_i}\right)},$$

has an **experimentwise** error rate α. **Simultaneous** $100(1-\alpha)$ percent confidence intervals for θ_j are given by $\widehat{\theta}_j \pm \sqrt{(t-1)F_{t-1,N-t,\alpha}} \times \widehat{\sigma}_{\widehat{\theta}_j}$, for $j = 1, 2, ..., k$.

REMARK: Scheffe's method is primarily used with contrasts that are "suggested" by the data and is valid for examining any and all contrasts **simultaneously**. Scheffe's test cannot possibly be rejected unless the overall F test is rejected (verify!). This confers the control of the experimentwise error rate for multiple tests.

Example 3.8. We now compute the Scheffe CCVs for the four contrasts in Example 3.1 using $\alpha = 0.05$. Recall that $F_{4,45,0.05} = 2.5787$, and hence, $\sqrt{(5-1)F_{4,45,0.05}} = 3.212$. Values of least-squares estimates, standard errors, and Scheffe's CCVs are given in Table 3.9. The least-squares estimates and standard errors are the same as in Table 3.8 (the only thing that changes are the critical values!). Regardless of whether or not these were preplanned or unplanned comparisons, we could conclude that the contrasts $\widehat{\theta}_2$ and $\widehat{\theta}_3$ are significantly different from zero at the $\alpha = 0.05$ **experimentwise** error rate.

Table 3.9: *Scheffe multiple comparisons for the pea growth data.*

θ_j	$\widehat{\theta}_j$	$\widehat{\sigma}_{\widehat{\theta}_j}$	$\sqrt{(5-1)F_{4,45,0.05}}$	Scheffe's CCV	Decision
θ_1	-1.10	1.045	3.212	3.357	Not significant
θ_2	-5.60	0.853	3.212	2.740	Significant
θ_3	10.20	0.826	3.212	2.653	Significant
θ_4	0.75	0.905	3.212	2.907	Not significant

THE BONFERRONI MULTIPLE COMPARISON METHOD: Let $\widehat{\theta}_1$, $\widehat{\theta}_2$, ..., $\widehat{\theta}_k$ denote a set of k contrasts in a one-way layout. The multiple comparison procedure that declares $\widehat{\theta}_j$ significant if

$$|\widehat{\theta}_j| \geq \underbrace{t_{N-t,\alpha/2k} \times \widehat{\sigma}_{\widehat{\theta}_j}}_{\text{Bonferroni CCV}},$$

where

$$\widehat{\sigma}_{\widehat{\theta}_j} = \sqrt{\text{MS[E]}\left(\sum_{i=1}^{t} \frac{c_i^2}{n_i}\right)},$$

has an **experimentwise** error rate α. **Simultaneous** $100(1-\alpha)$ percent confidence intervals for θ_j are given by $\widehat{\theta}_j \pm t_{N-t,\alpha/2k} \times \widehat{\sigma}_{\widehat{\theta}_j}$, for $j = 1, 2, ..., k$.

REMARK: Bonferroni's method controls the experimentwise error rate by employing a simple adjustment to the significance level of each individual test. Loosely speaking, if you have planned to do k tests, you just perform each at the α/k level rather than the α level. *This method is not appropriate for unplanned comparisons!*

THEORETICAL JUSTIFICATION: The theory for this method rests on the Bonferroni Inequality from set theory. For any sets $A_1, A_2, ..., A_k$,

$$P\left(\bigcup_{j=1}^{k} A_j\right) \leq \sum_{j=1}^{k} P(A_j);$$

i.e., probability measures are finitely subadditive. If all hypotheses $H_0 : \theta_j = 0$ are true, then the experimentwise error rate is

$$\begin{aligned}
P\left(|\widehat{\theta}_j| \geq t_{N-t,\alpha/2k} \times \widehat{\sigma}_{\widehat{\theta}_j} \text{ for some } j\right) &= P\left(\bigcup_{j=1}^{k}\left\{|\widehat{\theta}_j| \geq t_{N-t,\alpha/2k} \times \widehat{\sigma}_{\widehat{\theta}_j}\right\}\right) \\
&\leq \sum_{j=1}^{k} P\left(\left\{|\widehat{\theta}_j| \geq t_{N-t,\alpha/2k} \times \widehat{\sigma}_{\widehat{\theta}_j}\right\}\right) \\
&= \sum_{j=1}^{k} \frac{\alpha}{k} = \alpha.
\end{aligned}$$

Thus, by using Bonferroni's approach, the experimentwise error rate is controlled at α.

Example 3.9. We now compute the Bonferroni CCVs for the four contrasts in Example 3.1 using $\alpha = 0.05$. From SAS, I computed $t_{N-t,\alpha/2k} = t_{45,0.00625} = 2.6021$. Values

Table 3.10: *Bonferroni multiple comparisons for the pea growth data.*

θ_j	$\widehat{\theta}_j$	$\widehat{\sigma}_{\widehat{\theta}_j}$	$t_{45,0.00625}$	Bonferroni's CCV	Decision
θ_1	-1.10	1.045	2.6021	2.719	Not significant
θ_2	-5.60	0.853	2.6021	2.220	Significant
θ_3	10.20	0.826	2.6021	2.149	Significant
θ_4	0.75	0.905	2.6021	2.355	Not significant

of least-squares estimates, standard errors, and Bonferroni's CCVs are given in Table 3.10. Provided that these were **preplanned comparisons**, we could conclude that the contrasts $\widehat{\theta}_2$ and $\widehat{\theta}_3$ are significantly different from zero at the $\alpha = 0.05$ **experimentwise** error rate.

SUMMARY OF THE PROCEDURES: The most general procedures for making simultaneous inferences are Fisher's, Bonferroni, and Scheffe. These are listed in order from least conservative (most likely to reject an individual H_0) to most conservative (least likely to reject). Scheffe's method can be used for "data-snooping" purposes; that is, to make comparisons after the data have been observed. To decide on a method, you need to decide on how conservative you want to be. If it is very important not to claim differences when there are none, you should be very conservative. If it is most important to identify differences that *may* exist, you should choose less conservative methods.

3.5 Multiple pairwise comparisons of means

Pairwise comparisons are useful when the investigator has no preplanned comparisons specified and simply wants to examine the statistical (and practical) differences among the means. For a set of t means under consideration, there are $\binom{t}{2} = t(t-1)/2$ possible pairwise tests (confidence intervals). The term "pairwise" means that we are looking at pairs of means; i.e., $\theta_{ij} = \mu_i - \mu_j$ for $i \neq j$. Clearly, θ_{ij} is a contrast with $c_i = 1$ and $c_j = -1$ (with all other contrast coefficients equal to zero) with least squares estimator

$\widehat{\theta}_{ij} = \overline{Y}_{i+} - \overline{Y}_{j+}$. Rao calls such pairwise contrasts **simple contrasts**. It is easy to show (verify!) that the (estimated) standard error of any simple contrast is given by

$$\widehat{\sigma}_{\widehat{\theta}_{ij}} = \sqrt{\text{MS[E]}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}.$$

Thus, all of our previous results from the last section apply to the case of simple contrasts.

FISHER'S MULTIPLE PAIRWISE COMPARISON METHOD: In testing simple contrasts $H_0 : \theta_{ij} = \mu_i - \mu_j = 0$ versus $H_1 : \theta_{ij} = \mu_i - \mu_j \neq 0$, the Fisher multiple pairwise comparison method rejects H_0 when

$$|\widehat{\theta}_{ij}| \geq \underbrace{t_{N-t,\alpha/2} \times \sqrt{\text{MS[E]}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}_{\text{LSD}_{ij}(\text{F})}.$$

This procedure has a **comparisonwise** error rate α. Confidence intervals for θ_{ij} are given by $\widehat{\theta}_{ij} \pm t_{N-t,\alpha/2} \times \widehat{\sigma}_{\widehat{\theta}_{ij}}$, for $i \neq j$. LSD stands for **least significant difference**.

SCHEFFE'S MULTIPLE PAIRWISE COMPARISON METHOD: In testing simple contrasts $H_0 : \theta_{ij} = \mu_i - \mu_j = 0$ versus $H_1 : \theta_{ij} = \mu_i - \mu_j \neq 0$, the Scheffe multiple pairwise comparison method rejects H_0 when

$$|\widehat{\theta}_{ij}| \geq \underbrace{\sqrt{(t-1)F_{t-1,N-t,\alpha}} \times \sqrt{\text{MS[E]}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}_{\text{LSD}_{ij}(\text{S})}.$$

This procedure has a **experimentwise** error rate α. **Simultaneous** $100(1-\alpha)$ percent confidence intervals for θ_{ij} are given by $\widehat{\theta}_{ij} \pm \sqrt{(t-1)F_{t-1,N-t,\alpha}} \times \widehat{\sigma}_{\widehat{\theta}_{ij}}$, for $i \neq j$.

BONFERRONI'S MULTIPLE PAIRWISE COMPARISON METHOD: In testing simple contrasts $H_0 : \theta_{ij} = \mu_i - \mu_j = 0$ versus $H_1 : \theta_{ij} = \mu_i - \mu_j \neq 0$, the Bonferroni multiple pairwise comparison method rejects H_0 when

$$|\widehat{\theta}_{ij}| \geq \underbrace{t_{N-t,\alpha/2k} \times \sqrt{\text{MS[E]}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}_{\text{LSD}_{ij}(\text{B})}.$$

This procedure has a **experimentwise** error rate α. **Simultaneous** $100(1-\alpha)$ percent confidence intervals for θ_{ij} are given by $\widehat{\theta}_{ij} \pm t_{N-t,\alpha/2k} \times \widehat{\sigma}_{\widehat{\theta}_{ij}}$, for $i \neq j$.

STUDENTISED RANGE STATISTIC: In balanced designs, when H_0 is true; i.e., $\mu_i = \mu$, for each i, we know that, under our one-way model assumptions, the $\overline{Y}_{i+}$'s form a random sample of size t from a $\mathcal{N}(\mu, \sigma^2/n)$ distribution. Looking at the range of this "sample" and dividing by the natural independent χ^2 estimate of the standard deviation leads to the statistic

$$q = \frac{\max_i \overline{Y}_{i+} - \min_i \overline{Y}_{i+}}{\sqrt{\text{MS[E]}/n}}.$$

This is called the **studentised range statistic**. Furthermore, it is possible to find the distribution q; it has two degree of freedom parameters and is tabled in Table C.11 (Rao) for $\alpha = 0.01$ and $\alpha = 0.05$. Note that $|\overline{Y}_{i+} - \overline{Y}_{j+}|$ never exceeds $\max_i \overline{Y}_{i+} - \min_i \overline{Y}_{i+}$. This simple fact establishes experimentwise control with Tukey's method.

TUKEY'S MULTIPLE PAIRWISE COMPARISON METHOD: In testing simple contrasts $H_0 : \theta_{ij} = \mu_i - \mu_j = 0$ versus $H_1 : \theta_{ij} = \mu_i - \mu_j \neq 0$, the Tukey multiple pairwise comparison method rejects H_0 when

$$|\widehat{\theta}_{ij}| \geq \underbrace{q_{t,N-t,\alpha} \times \sqrt{\frac{\text{MS[E]}}{n}}}_{\text{LSD}_{ij}(\text{T})}.$$

This procedure has a **experimentwise** error rate α and is only appropriate for **balanced designs**; i.e., $n_i = n$ for all i. **Simultaneous** $100(1-\alpha)$ percent confidence intervals for θ_{ij} are given by

$$\widehat{\theta}_{ij} \pm q_{t,N-t,\alpha} \times \sqrt{\frac{\text{MS[E]}}{n}},$$

for $i \neq j$. The **Tukey-Kramer** procedure is an extension of Tukey's procedure to unbalanced designs (it is conservative); $H_0 : \theta_{ij} = \mu_i - \mu_j = 0$ is rejected when

$$|\widehat{\theta}_{ij}| \geq \underbrace{q_{t,N-t,\alpha} \times \sqrt{\text{MS[E]}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}_{\text{LSD}_{ij}(\text{TK})}.$$

Confidence intervals are formed analogously. The Tukey and Tukey-Kramer procedures are explicitly designed for **pairwise comparisons**; they can not be used when dealing with contrasts that are not simple.

NOTE: We will avoid the Student-Newman-Keuls and Duncan methods (p. 356, Rao).

Example 3.10 (`pea-mc.sas`). We now illustrate the four pairwise comparisons methods (Fisher, Scheffe, Bonferroni, and Tukey) with the pea section data from Example 2.2 using $\alpha = 0.05$. Recall that there are $t = 5$ treatments, which gives us $\binom{5}{2} = 10$ pairwise tests (or confidence intervals). Recall that $F_{4,45,0.05} = 2.5787$. We have

$$\begin{aligned}
\text{LSD}_{ij}(\text{F}) &= \underbrace{t_{45,0.025}}_{2.0141} \times \sqrt{5.46\left(\frac{1}{10}+\frac{1}{10}\right)} = 2.104 \\
\text{LSD}_{ij}(\text{S}) &= \sqrt{(5-1)F_{4,45,0.025}} \times \sqrt{5.46\left(\frac{1}{10}+\frac{1}{10}\right)} = 3.355 \\
\text{LSD}_{ij}(\text{B}) &= \underbrace{t_{45,0.0.0025}}_{2.9521} \times \sqrt{5.46\left(\frac{1}{10}+\frac{1}{10}\right)} = 3.084 \\
\text{LSD}_{ij}(\text{T}) &= \underbrace{q_{5,45,0.05}}_{\approx\, 4.02} \times \sqrt{\frac{5.46}{10}} = 2.970.
\end{aligned}$$

Table 3.11: *All pairwise comparisons for the pea growth data.*

	$\widehat{\theta}_{ij}$	Comparison	Methods rejecting H_0
θ_{12}	11.9	C vs. F	FSBT
θ_{13}	12.1	C vs. GF	FSBT
θ_{14}	10.8	C vs. G	FSBT
θ_{15}	6.0	C vs. S	FSBT
θ_{23}	0.2	F vs. GF	none
θ_{24}	-1.1	F vs. G	none
θ_{25}	-5.9	F vs. S	FSBT
θ_{34}	-1.3	FG vs. G	none
θ_{35}	-6.1	FG vs. S	FSBT
θ_{45}	-4.8	G vs. S	FSBT

ANALYSIS: For the pea-section data, we see from Table 3.11 that all four pairwise procedures yield the same conclusions for each of the 10 simple contrasts! Of course, this need not always be the case. Also, recall that Fisher's method doesn't control the experimentwise error rate (whereas the other three methods do).

REMARK: With the pea data, note that for those methods that control the experimentwise error rate,

$$\text{LSD}_{ij}(\text{T}) < \text{LSD}_{ij}(\text{B}) < \text{LSD}_{ij}(\text{S}).$$

This is usually how the ordering goes when comparing these three procedures. Tukey's procedure is **exact**; i.e., the experimentwise error rate is controlled at α. The other two procedures, Bonferroni and Scheffe, are **conservative**; i.e., the experimentwise error rate for the set of comparisons is smaller than α. Thus, if all pairwise comparisons are of interest, Tukey's method is preferred over the Bonferroni and Scheffe methods (in cases of equal replication) since we are more likely to reject the relevant hypotheses. In unbalanced designs, because the Scheffe method considers all possible contrasts in its formulation, not just ones involving pairs, the Bonferroni method may be preferred if the number of contrasts is small. In this case, it is likely that $\text{LSD}_{ij}(\text{B}) < \text{LSD}_{ij}(\text{S})$.

REMARK: It is legitimate to conduct hypothesis tests using all of these methods (where applicable) and then choose the one that rejects most often. This is valid because the "cutoff" values $\text{LSD}_{ij}(\text{T}), \text{LSD}_{ij}(\text{TK}), \text{LSD}_{ij}(\text{B})$, and $\text{LSD}_{ij}(\text{S})$ do not depend on the data.

SOME MULTIPLE COMPARISON CAVEATS: Because the number of comparisons in a multiple-comparisons setting may be large and of our desire to control the **experimentwise** error rate, we are likely to have low power (that is, we may have a difficult time detecting real differences among the comparisons in our set). This is true because we must use critical values (for each comparison) larger than we would if the comparisons were made separately at level α. This problem has tempted some investigators to try to figure out ways "around the issue;" for example, claiming that certain comparisons were of interest in advance when they really were not, so as to salvage "significant" results. This is, of course, not appropriate! The only way to ensure enough power to test all questions of interest is to design the experiment with a large enough sample size. These issues are explored in Section 9.8 (Rao).

4 Simple Linear Regression and Correlation

Complimentary reading from Rao: Chapter 10 (§ 10.1-10.8).

4.1 Introduction

We have largely focused our attention on problems where the main issue is to identify differences among treatment means in a one-way layout. Determining differences among means can be achieved through the ANOVA and the use of contrasts. Another problem that often arises in economics, industrial applications, and biological settings is that of investigating the mathematical relationship between two (or more) variables. Depending on the nature of the variables, and the observations on them, the methods of **regression analysis** or **correlation analysis** are appropriate. Our development of the methods for identifying differences among treatment means, those of analysis of variance, are, in fact, very similar to regression analysis methods. Both sets of methods are predicated on representing the data by a **linear model** which includes components representing both systematic and random sources of variation.

Example 4.1. Many fishes have a lateral line system enabling them to experience mechanoreception (the ability to sense physical contact on the surface of the skin or movement of the surrounding environment, such as sound waves in air or water). The frequency (number per second) of electrical impulses (EI) emitted from one particular fish was measured at several temperatures (measured in Celcius).

Temperature (X):	20	22	23	25	27	28	30
Frequency (Y):	224	252	267	287	301	306	318

Figure 4.10 is a scatterplot of the data pairs. Does the straight line seem to "fit" the data well? That is, does this **straight-line model** seem to be appropriate here? Or, perhaps the true relationship between temperature and frequency is not a straight line, but, rather a quadratic curve. The **quadratic model** fit appears in Figure 4.11.

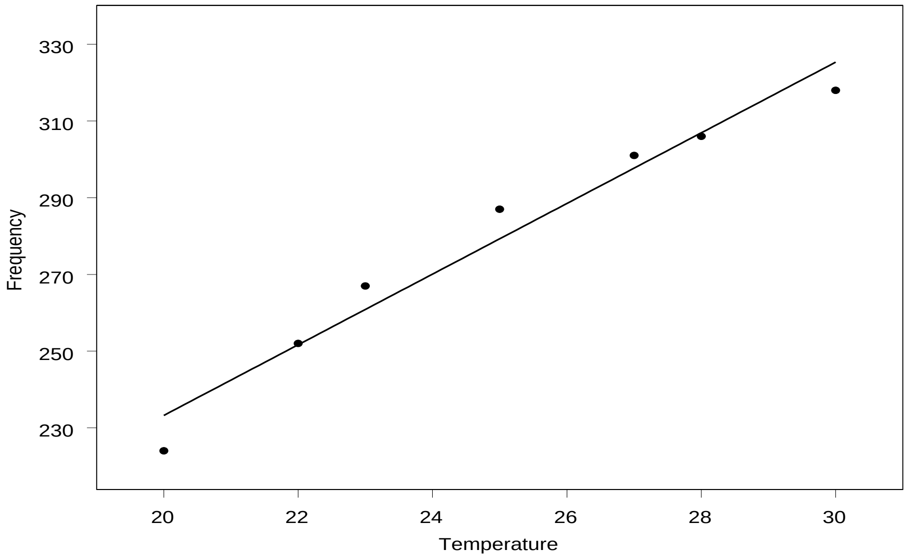

Figure 4.10: *EI frequency at different temperatures with a straight-line fit.*

4.2 Experimental data versus observational data

SCENARIO: We are interested in modelling the relationship between two variables, X and Y. We observe the pair (X, Y) on each of a sample of experimental units, and we wish to use them to say something about the relationship. *How we view the relationship is dictated by whether or not we have experimental data or observational data.*

EXPERIMENTAL DATA: Observations on X and Y are planned as the result of an experiment. That is, we control or fix the values of X, and we observe the resulting Y.

- X = dose of a drug, Y = change in blood pressure for a human subject
- X = concentration of toxic substance, Y = number of mutant offspring observed for a pregnant rat
- X = temperature, Y = frequency of electrical impulses

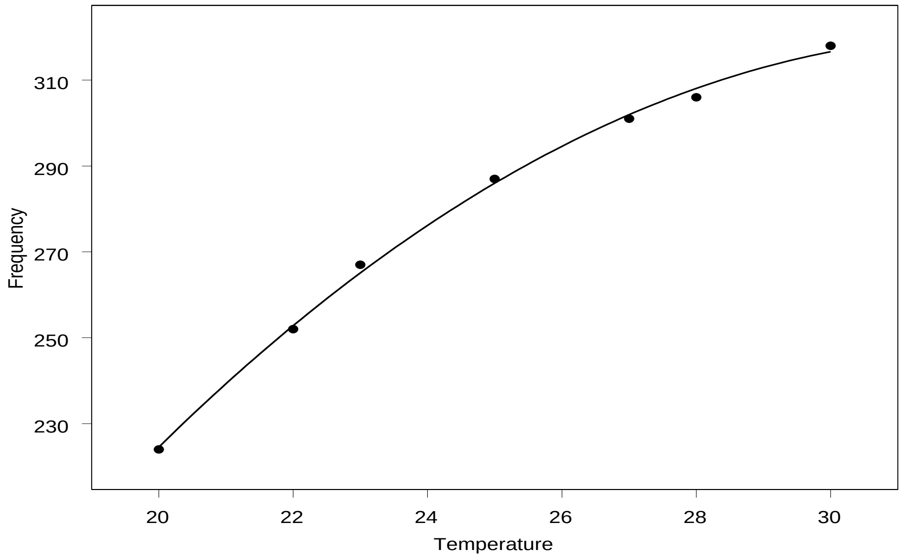

Figure 4.11: *EI frequency at different temperatures with a quadratic fit.*

OBSERVATIONAL DATA: We observe both X and Y values, neither of which is under our control. For example,

- X = weight, Y = height of a human subject
- X = average heights of plants in a plot, Y = yield
- X = SAT score, Y = first year college GPA.

REMARK: In experimental data situations, there is a distinction between what we call X and what we call Y, because the values of X are specified by the investigator (hence, X is **fixed**). In observational data situations, we do not choose the values of X; we merely observe the pair (X, Y). In this setting, the X variable is best regarded as **random**.

RESULT: With experimental data, when X is best regarded as fixed, regression analysis methods are appropriate, whereas, with observational data, when X is best regarded as a random variable, correlation analysis methods are appropriate.

RELATIONSHIPS BETWEEN TWO VARIABLES: In some situations, scientific theory may suggest that two variables, X and Y, are functionally related, e.g., $Y = g(X)$. However, even if there is no suitable theory, we may still suspect that some kind of systematic relationship between X and Y exists, and we may be able to choose a function g that provides a reasonable **empirical** description.

PRACTICAL PROBLEM: In most situations, the values we observe for Y (and sometimes X) are not exact. In particular, due to biological variation among experimental units sampling error, imprecision and/or inaccuracy of measuring devices, etc., we may only observe values of Y (and also possibly X) *with error*. Thus, based on a sample of (X, Y) pairs, our ability to see the *exact* relationship is obscured by this error.

STATISTICAL MODELS: These considerations dictate how we should think of formal statistical models for each situation:

- *Experimental data*: A natural way to think about Y is by $Y = g(x) + \epsilon$. Here, we believe the function g describes the relationship, but values of Y we observe are not exactly equal to $g(x)$ because of the errors mentioned above. The additive error ϵ characterises this, just as in our ANOVA models. In this situation, the following terminology is often used:
  - Y = "response" or "dependent variable," and
  - X = "predictor," "independent variable," or "covariate."
- *Observational data*: In this situation, there is really not much distinction between X and Y, as both are seen as random. Here, the terms "independent" and "dependent" variable may be misleading. For example, if we have observed n randomly selected subjects and record X = weight and Y = systolic blood pressure, we may be just as interested in how Y relates to X as we are in how X relates to Y!

We begin our discussion of these problems by considering regression models appropriate for experimental data. Correlation analysis for observational data is addressed later.

4.3 An introduction to regression models

OBJECTIVE: The goal of regression analysis is to model the relationship between a response variable Y and one or more independent variables, say, $x_1, x_2, ..., x_p$. That is, we want to find a function g that describes the relationship between Y and $x_1, x_2, ..., x_p$. Consider the following **statistical model**

$$Y = g(x_1, x_2, ..., x_p) + \epsilon,$$

where $E(\epsilon) = 0$. You see that this model consists of two parts: (1) the **deterministic** part, $Y = g(x_1, x_2, ..., x_p)$, and (2) the **random** part ϵ. From our previous discussion, we know that, in practice, it is unreasonable to think that the observed values of Y will be *perfectly related* to $x_1, x_2, ..., x_p$ through the g function. The random error ϵ conveys the fact that there will (most likely) not be a perfect relationship.

REMINDER: It is important to remember that in this regression model, the independent variables $x_1, x_2, ..., x_p$ are **fixed**; they are not random, and they are measured without error. Since, $E(\epsilon) = 0$, we can write the model equivalently as

$$E(Y|x_1, x_2, ..., x_p) = g(x_1, x_2, ..., x_p).$$

That is, the expected value of our response Y, conditioned on the k independent variables $x_1, x_2, ..., x_p$, is equal to $g(x_1, x_2, ..., x_p)$.

ASSUMPTIONS ON THE ERROR TERM: It is common to assume that $\epsilon \sim \mathcal{N}(0, \sigma^2)$. That is, the error term is normally distributed with mean zero and variance σ^2. The variance quantifies the amount of dispersion about the **true regression function** $g(x_1, x_2, ..., x_p)$. Note that if $\epsilon \sim \mathcal{N}(0, \sigma^2)$, it follows immediately that

$$Y \sim \mathcal{N}\{g(x_1, x_2, ..., x_p), \sigma^2\}.$$

Thus, $E(Y|x_1, x_2, ..., x_p) = g(x_1, x_2, ..., x_p)$ and $V(Y|x_1, x_2, ..., x_p) = \sigma^2$. The last point about the variance is important; namely, that the variance of Y is *constant* across the values of $x_1, x_2, ..., x_p$. This is analogous to the homoscedastic assumption on the errors in the one-way ANOVA model.

STRAIGHT-LINE SETTING: Suppose that $p = 1$; that is, we only have one independent variable, say x. Oftentimes it is reasonable to assume that the relationship between Y and x, i.e., the form of g, is, in fact, a **straight line**. We may write this as

$$Y = \beta_0 + \beta_1 x + \epsilon \quad \text{or, equivalently} \quad E(Y|x) = \beta_0 + \beta_1 x,$$

for some values β_0 and β_1. Here, $g(x) = \beta_0 + \beta_1 x$ is a straight line with **intercept** β_0 (i.e., the value of Y when $x = 0$) and **slope** β_1. The slope β_1 expresses the rate of change in Y, i.e., the change in the mean of Y brought about by a one-unit change in x.

TERMINOLOGY: We call the model $Y = \beta_0 + \beta_1 x + \epsilon$ a **simple linear regression model**. The modifier "simple" means that we are considering only one predictor x. The term "linear" is not taken to mean that the true regression equation $g(x) = \beta_0 + \beta_1 x$ is a straight line. In fact, many linear regression models are not straight lines. Linear regression models that include more than one x are called **multiple linear regression models**.

TERMINOLOGY: In the simple linear regression model $Y = \beta_0 + \beta_1 x + \epsilon$, the constants β_0 and β_1 are called **regression parameters**. These parameters are fixed, unknown values (they refer to the *true relationship* between Y and x); thus, as you may suspect, they must be **estimated** with data. Recall that in a regression setting, x is fixed and that ϵ is a random variable with mean zero. Clearly, Y, since it is a function of ϵ, is a random quantity, too. Since $E(Y|x) = g(x) = \beta_0 + \beta_1 x$, for a given value of x, we would expect Y to equal $g(x) = \beta_0 + \beta_1 x$; however, due to random variation (e.g., biological, measurement error, etc.), we see Y values dispersed about the line $g(x) = \beta_0 + \beta_1 x$.

OTHER MODELS: In theory, there are many different possible g functions one could use to model the relationship between Y and x. We don't have to restrict ourselves to straight-line models. For example, it could be that the true g function that relates Y to x is given by $g(x) = \beta_0 + \beta_1 x + \beta_2 x^2$ or $g(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$. Even though the quadratic and cubic equations are not straight lines, they still fall into a general class of models called **linear regression models**. We will discuss this in a more mathematical context shortly.

A NONLINEAR MODEL: In some situations, a different type of model may be appropriate. For example, suppose that Y is some measure of growth of a plant and x denotes time, and that we want to study how Y relates to x. Here, we would eventually expect the relationship to "level off" when x gets large, as plants can not continue to get large forever! A popular model for this is the **logistic model** given by

$$Y = \underbrace{\frac{\beta_0}{1+\beta_1 e^{\beta_2 x}}}_{g(x)} + \epsilon,$$

as this g function, when $\beta_2 < 0$, would begin to "flatten out" for large values of x.

NOTE: In the quadratic equation with $g(x) = \beta_0 + \beta_1 x + \beta_2 x^2$, note that, although the function is no longer a straight line, the regression parameters β_0, β_1, and β_2 *enter in a linear fashion.* Contrast this with the logistic growth model. This function is not linear as a function of β_0, β_1, and β_2; rather, it is better described as **nonlinear**.

MATHEMATICAL DESCRIPTION: The model

$$Y = \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p}_{g(x_1, x_2, ..., x_p)} + \epsilon,$$

with regression parameters $\beta_0, \beta_1, ..., \beta_p$, is called a **linear regression model**. As hinted at earlier, this "linear" term refers to how each term enters the regression equation. It does not refer to the shape of the true regression function g. To be precise, when we refer to a model as a "linear model," we mean that the true regression function g is *linear in the parameters.* Mathematically, this means that the $p+1$ partial derivatives

$$\frac{\partial g(\beta_0, \beta_1, ..., \beta_p)}{\partial \beta_i} \qquad i = 0, 1, ..., p,$$

are all **free** of the parameters $\beta_0, \beta_1, ..., \beta_p$. Each of the models is a linear model:

$$\begin{aligned} Y &= \underbrace{\beta_0 + \beta_1 x}_{g(x)} + \epsilon \\ Y &= \underbrace{\beta_0 + \beta_1 x + \beta_2 x^2}_{g(x)} + \epsilon \\ Y &= \underbrace{\beta_0 + \beta_1 \log x_1 + \beta_2 \sqrt{\cos x_2}}_{g(x_1, x_2)} + \epsilon. \end{aligned}$$

It is easy to show that each model above is linear in the parameters. On the other hand, the logistic model is not linear since

$$\frac{\partial}{\partial \beta_0}\left(\frac{\beta_0}{1+\beta_1 e^{\beta_2 x}}\right) = \frac{1}{1+\beta_1 e^{\beta_2 x}},$$

which is not free of the regression parameters β_1 and β_2.

NOTE: The class of **linear models** is enormous! Many models used in statistics fall into this classification. For example, this class includes the ANOVA and regression models, as well as many models used with time series analysis, multivariate data, spatial processes, and others.

GOALS: For the linear statistical model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$, among other things, our main goals in **regression analysis** will be to

- estimate the regression parameters, $\beta_0, \beta_1, ..., \beta_p$
- diagnose the fit (i.e., perform model diagnostics),
- estimate mean responses and make predictions about future values.

We will start discussing these issues in our **simple linear regression** setting where $g(x) = \beta_0 + \beta_1 x$. Later, we'll extend these ideas to multiple regression settings.

4.4 Using least squares to fit a straight line regression

TERMINOLOGY AND ASSUMPTIONS: When we say, "fit a regression model," we basically mean that we are estimating the parameters in the model with observed data. The **method of least squares** provides a way to do this. For observations (x_i, Y_i), $i = 1, 2, ..., n$, we postulate the **simple linear regression model**

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$. We wish to fit this model by estimating the intercept and slope parameters, β_0 and β_1, respectively.

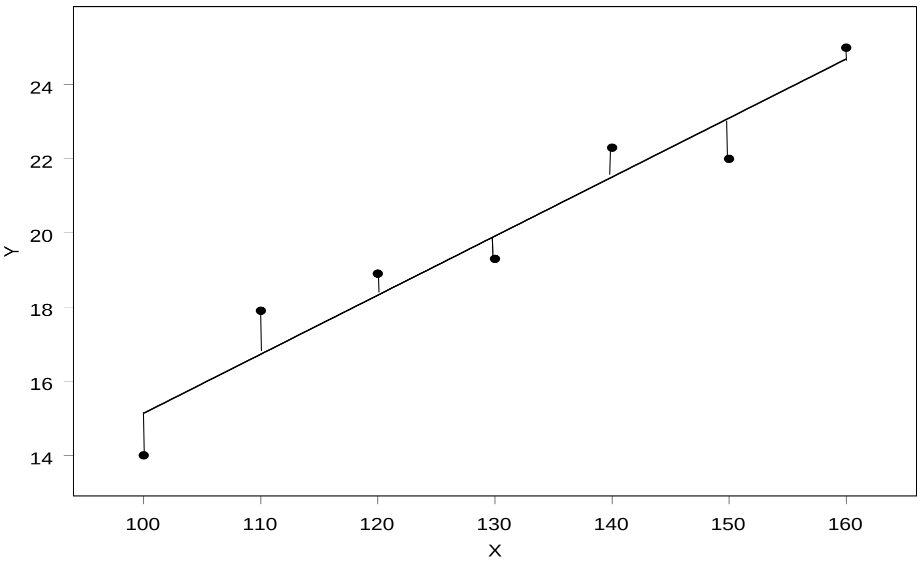

Figure 4.12: *A scatterplot with straight line and residuals associated with the straight line.*

THE METHOD OF LEAST SQUARES: The most widely-accepted method for estimating β_0 and β_1 is using **the method of least squares**. It turns out to be the most appropriate way to estimate β_0 and β_1 under the normality assumption on the error term ϵ. For each Y_i, and *given values* of β_0 and β_1, note that the quantity

$$e_i = Y_i - (\beta_0 + \beta_1 x_i)$$

measures the *vertical distance* from Y_i to the line $\beta_0 + \beta_1 x_i$. This distance is called the ith **residual** (for particular values of β_0 and β_1). If a point falls above the line in the Y direction, the residual is positive. If a point falls below the line in the Y direction, the residual is negative. A natural way to measure the overall deviation of the observed data Y_i from their means, $\beta_0 + \beta_1 x_i$, is with the **residual sum of squares** given by

$$\text{SSE}(\beta_0, \beta_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \{Y_i - (\beta_0 + \beta_1 x_i)\}^2.$$

The method of least squares says to find the estimates of β_0 and β_1, say, $\widehat{\beta}_0$ and $\widehat{\beta}_1$, respectively, that **minimise** the sum of squared residuals; i.e., that makes the function

$\text{SSE}(\beta_0, \beta_1)$ as small as possible. A two-variable calculus argument is used to find the form of the estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$. Taking partial derivatives of $\text{SSE}(\beta_0, \beta_1)$ and setting them equal to zero, we obtain **the normal equations**

$$\begin{aligned}\frac{\partial \text{SSE}(\beta_0, \beta_1)}{\partial \beta_0} &= -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i) \overset{\text{set}}{=} 0\\ \frac{\partial \text{SSE}(\beta_0, \beta_1)}{\partial \beta_1} &= -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i)x_i \overset{\text{set}}{=} 0.\end{aligned}$$

Solving this 2×2 system for β_0 and β_1, one can show (verify!) that

$$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad \widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{x}$$

are the minimisers, where

$$S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y}) = \sum_{i=1}^{n} x_i Y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} Y_i$$

and

$$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2.$$

The values $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are called the **least squares estimators** for β_0 and β_1. The **residual sum of squares for the least-squares line** is given by

$$\begin{aligned}\text{SS[E]} = \text{SSE}(\widehat{\beta}_0, \widehat{\beta}_1) &= \sum_{i=1}^{n}\{Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i)\}^2\\ &= \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2,\end{aligned}$$

where $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ denotes the ith **fitted value**. The value $e_i = Y_i - \widehat{Y}_i$ is called the the ith **least-squares residual**.

Example 4.2 (`oxygen.sas`). The following data are rates of oxygen consumption of birds (Y) measured at different temperatures (x). Here, the temperatures were set by the investigator, and the O_2 rates, Y, were observed for these particular temperatures. Thus, the assumption of a fixed x is justified.

x, (degrees Celcius)	−18	−15	−10	−5	0	5	10	19
y, (ml/g/hr)	5.2	4.7	4.5	3.6	3.4	3.1	2.7	1.8

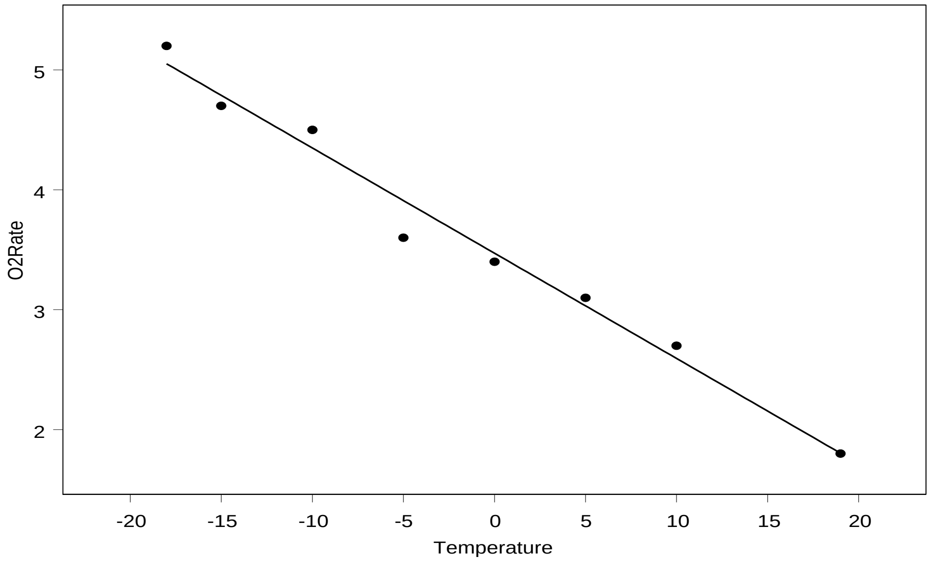

Figure 4.13: *Bird oxygen rate data for different temperatures.*

The scatterplot of the data appears in Figure 4.13. *It is always advisable to plot the data first!* The **least-squares line** is superimposed over the data.

CALCULATIONS: We have $n = 8$.

$$\sum_{i=1}^{8} y_i = 29 \qquad \overline{y} = 3.625 \qquad \sum_{i=1}^{8} y_i^2 = 114.04$$

$$\sum_{i=1}^{8} x_i = -14 \qquad \overline{x} = -1.75 \qquad \sum_{i=1}^{8} x_i^2 = 1160$$

$$\sum_{i=1}^{8} x_i y_i = -150.4$$

$$S_{xy} = -150.4 - \frac{1}{8}(29)(-14) = -99.65 \qquad S_{xx} = 1160 - \frac{1}{8}(-14)^2 = 1135.5.$$

Thus, we obtain

$$\begin{aligned} \widehat{\beta}_1 &= \frac{-99.65}{1135.5} = -0.0878, \quad \text{and} \\ \widehat{\beta}_0 &= 3.625 - (-0.0878)(-1.75) = 3.4714. \end{aligned}$$

The **least-squares regression line** is given by

$$\widehat{Y}_i = 3.4714 - 0.0878x_i; \quad \text{i.e.,} \quad \widehat{\mathrm{O_2Rate}}_i = 3.4714 - 0.0878\mathrm{TEMP}_i.$$

4.5 Properties of least-squares estimators

PROPERTIES OF THE ESTIMATORS: We consider the simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, for $i = 1, 2, ..., n$, where $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$. The following summarise the important sampling distribution results.

(1) $E(\widehat{\beta}_0) = \beta_0$ and $E(\widehat{\beta}_1) = \beta_1$; that is, the least-squares estimators are **unbiased**.

(2) $V(\widehat{\beta}_0) = s_{00}\sigma^2$, where

$$s_{00} = \frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}.$$

(3) $V(\widehat{\beta}_1) = s_{11}\sigma^2$, where

$$s_{11} = \frac{1}{S_{xx}}.$$

(4) $\mathrm{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) = s_{01}\sigma^2$, where

$$s_{01} = \frac{-\overline{x}}{S_{xx}}.$$

(5) $E(\mathrm{MS[E]}) = \sigma^2$, where

$$\mathrm{MS[E]} \equiv \frac{\mathrm{SS[E]}}{n-2}.$$

That is, the statistic $\widehat{\sigma}^2 \equiv \mathrm{MS[E]}$ is an **unbiased estimator** of the error variance σ^2. As before, MS[E] is called the **mean-squared error**.

(6) Both estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are normally distributed.

(7) The random variable

$$\frac{\mathrm{SS[E]}}{\sigma^2} = \frac{(n-2)\mathrm{MS[E]}}{\sigma^2} \sim \chi^2_{n-2}.$$

(8) The mean-squared error MS[E] is independent of both $\widehat{\beta}_0$ and $\widehat{\beta}_1$.

4.6 Confidence intervals and hypothesis tests for β_0, β_1, and σ^2

The statistics $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are **point estimators** of the true population parameters β_0 and β_1, respectively. We now discuss how to obtain **confidence intervals** and **hypothesis tests** for the true regression parameters β_0 and β_1.

INFERENCE FOR β_1: From facts (1), (2), and (6) of the last section, recall that

$$\widehat{\beta}_1 \sim \mathcal{N}(\beta_1, s_{11}\sigma^2).$$

where $s_{11} = 1/S_{xx}$. Thus, by standardising, it follows that

$$Z = \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{s_{11}\sigma^2}} \sim \mathcal{N}(0, 1).$$

Also, recall from Fact (7) that

$$\frac{(n-2)\text{MS[E]}}{\sigma^2} \sim \chi^2_{n-2},$$

and that from Fact (8), $(n-2)\text{MS[E]}/\sigma^2$ is independent of $\widehat{\beta}_1$ (and, thus, is independent of Z). Thus, the quantity

$$t \equiv \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{s_{11}\text{MS[E]}}} = \frac{(\widehat{\beta}_1 - \beta_1)/\sqrt{s_{11}\sigma^2}}{\sqrt{\frac{(n-2)\text{MS[E]}}{\sigma^2}/(n-2)}} \sim t_{n-2}.$$

Since t has a distribution free of all parameters, it is a **pivot**. Thus, we can use t to find a $100(1-\alpha)$ percent confidence interval for β_1. Straightforward calculations show that this interval is given by

$$\widehat{\beta}_1 \pm t_{n-2,\alpha/2}\sqrt{s_{11}\text{MS[E]}},$$

where $t_{n-2,\alpha/2}$ denotes the $1-\alpha/2$ quantile of the t distribution with $n-2$ degrees of freedom. In addition, if we wanted to test, at level α,

$$H_0 : \beta_1 = \beta_{1,0}$$

versus

$$H_1 : \beta_1 \neq \beta_{1,0},$$

for some specified value of $\beta_{1,0}$, we would use

$$t = \frac{\widehat{\beta}_1 - \beta_{1,0}}{\sqrt{s_{11}\text{MS[E]}}}$$

as a test statistic with rejection region $RR = \{t : t > t_{n-2,\alpha/2} \text{ or } t < -t_{n-2,\alpha/2}\}$. If our alternative hypothesis was one sided, we would simply adjust our rejection region accordingly. P values are calculated as appropriate areas under the t_{n-2} distribution.

INFERENCE FOR β_0: A completely analogous argument (try it!) can be used to show that

$$t = \frac{\widehat{\beta}_0 - \beta_0}{\sqrt{s_{00}\text{MS[E]}}} \sim t_{n-2},$$

where $s_{00} = \frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}$, and that a $100(1-\alpha)$ percent confidence interval for β_0 is

$$\widehat{\beta}_0 \pm t_{n-2,\alpha/2}\sqrt{s_{00}\text{MS[E]}}.$$

In addition, a level α test of $H_0 : \beta_0 = \beta_{0,0}$ versus $H_1 : \beta_0 \neq \beta_{0,0}$, for some specified value of $\beta_{0,0}$, can be performed using the test statistic

$$t = \frac{\widehat{\beta}_0 - \beta_{0,0}}{\sqrt{s_{00}\text{MS[E]}}}$$

with rejection region $RR = \{t : t > t_{n-2,\alpha/2} \text{ or } t < -t_{n-2,\alpha/2}\}$. If our alternative hypothesis was one sided, we would simply adjust our rejection region accordingly. P values are calculated as appropriate areas under the t_{n-2} distribution.

INFERENCE FOR σ^2: Since $(n-2)\text{MS[E]}/\sigma^2 \sim \chi^2_{n-2}$, it follows that

$$P\left\{\chi^2_{n-2,1-\alpha/2} \leq \frac{(n-2)\text{MS[E]}}{\sigma^2} \leq \chi^2_{n-2,\alpha/2}\right\} = 1-\alpha,$$

and consequently, a $100(1-\alpha)$ percent confidence interval for σ^2 is given by

$$\left(\frac{(n-2)\text{MS[E]}}{\chi^2_{n-2,\alpha/2}}, \frac{(n-2)\text{MS[E]}}{\chi^2_{n-2,1-\alpha/2}}\right).$$

Example 4.3 (`oxygen.sas`). For the oxygen rate data in Example 4.2, we have $n = 8$, $t_{6,0.025} = 2.447$, $\chi^2_{6,0.975} = 1.2373$, $\chi^2_{6,0.025} = 14.4494$, $\widehat{\beta}_0 = 3.4714$, and $\widehat{\beta}_1 = -0.0878$. We have $\text{MS[E]} \approx 0.028$, $s_{11} = 1/S_{xx} = 1/1135.5 \approx 0.00088$, and

$$s_{00} = \frac{1}{n} + \frac{\overline{x}^2}{S_{xx}} = \frac{1}{8} + \frac{(-1.75)^2}{1135.5} \approx 0.128.$$

- A 95 percent confidence interval for β_1 is given by $-0.0878 \pm 2.447\sqrt{0.00088 \times 0.028}$ or $(-0.0999, -0.0755)$,
- A 95 percent confidence interval for β_0 is given by $3.4714 \pm 2.447\sqrt{0.128 \times 0.028}$ or $(3.3243, 3.6185)$, and
- A 95 percent confidence interval for σ^2 is given by
$$\left(\frac{6 \times 0.028}{14.4494}, \frac{6 \times 0.028}{1.2373}\right), \quad \text{or} \quad (0.012, 0.136).$$

NOTE: In practice, the confidence interval for the slope parameter β_1 is of primary interest because of its connection to the predictor variable x in the regression model $Y = \beta_0 + \beta_1 x + \epsilon$. The confidence interval for β_0 is usually less meaningful (unless you are interested in the mean of Y when $x = 0$). In our example, β_1 has units of the rate of change of oxygen consumption per unit change in temperature; i.e., ml/g/hr per degree Celsius. Hence, given a one-degree increase in temperature, we are 95 percent confident that the change in mean oxygen consumption rate is between -0.0999 and -0.0755.

SIMULTANEOUS (JOINT) CONFIDENCE REGIONS FOR β_0 AND β_1: The goal may be to find a region in the β_0-β_1 plane that contains the vector (β_0, β_1) with probability $1 - \alpha$. This is called a $100(1-\alpha)$ **percent confidence region** for (β_0, β_1). Using the intersection of the two individual confidence intervals, as described above, will have simultaneous coverage less than $1 - \alpha$. In this light, consider two different regions:

1. An exact elliptical region
$$\left\{(\beta_0, \beta_1) : \frac{1}{2\text{MS[E]}} \begin{pmatrix} \widehat{\beta}_0 - \beta_0 \\ \widehat{\beta}_1 - \beta_1 \end{pmatrix}' \begin{pmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix} \begin{pmatrix} \widehat{\beta}_0 - \beta_0 \\ \widehat{\beta}_1 - \beta_1 \end{pmatrix} \leq F_{2,n-2,\alpha}\right\}$$
2. A rectangular region (equivalent to the intersection of two intervals) using a Bonferroni-type correction
$$\widehat{\beta}_0 \pm t_{n-2,\alpha/4}\sqrt{s_{00}\text{MS[E]}} \qquad \text{and} \qquad \widehat{\beta}_1 \pm t_{n-2,\alpha/4}\sqrt{s_{11}\text{MS[E]}}.$$
This pair constitutes a joint confidence region whose probability of joint coverage is at least $1 - \alpha$.

4.7 Confidence intervals for linear functions of β_0 and β_1

Consider our straight-line regression model $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where $\epsilon_i \sim \text{iid } \mathcal{N}(0, \sigma^2)$. In many settings, it is of interest to make inferential statements about a **linear combination** of the regression parameters. That is, we might want to write a confidence interval or conduct a hypothesis test for the parameter $\theta = c_0\beta_0 + c_1\beta_1$, where c_0 and c_1 are constants.

POINT ESTIMATOR FOR θ: Using the least-squares estimators of $\widehat{\beta}_0$ and $\widehat{\beta}_1$, the **least squares estimator** for θ becomes

$$\widehat{\theta} = c_0\widehat{\beta}_0 + c_1\widehat{\beta}_1.$$

THE SAMPLING DISTRIBUTION OF $\widehat{\theta}$: Recall that both $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are both normally distributed. Thus, since $\widehat{\theta}$ is just a **linear combination** of $\widehat{\beta}_0$ and $\widehat{\beta}_1$, it, too, is normally distributed. It is also easy to see that $E(\widehat{\theta}) = \theta$; i.e., $\widehat{\theta}$ is **unbiased** for θ, and that the variance of $\widehat{\theta}$ is given by (verify!)

$$V(\widehat{\theta}) \equiv \sigma^2_{\widehat{\theta}} = \sigma^2 \left(c_0^2 s_{00} + c_1^2 s_{11} + 2s_{01}c_0c_1 \right).$$

Thus, we have that $\widehat{\theta} \sim \mathcal{N}(\theta, \sigma^2_{\widehat{\theta}})$.

NOTE: You will see that the variance of our estimator $\sigma^2_{\widehat{\theta}}$, depends on the unknown parameter σ^2. An estimate of $\sigma^2_{\widehat{\theta}}$ is given by

$$\widehat{\sigma}^2_{\widehat{\theta}} = \text{MS[E]} \left(c_0^2 s_{00} + c_1^2 s_{11} + 2s_{01}c_0c_1 \right).$$

Now, whereas

$$Z = \frac{\widehat{\theta} - \theta}{\sigma_{\widehat{\theta}}} \sim \mathcal{N}(0, 1),$$

it follows that (verify!)

$$t = \frac{\widehat{\theta} - \theta}{\widehat{\sigma}_{\widehat{\theta}}} \sim t_{n-2}.$$

Since t is a pivotal quantity, a $100(1-\alpha)$ percent confidence interval becomes $\widehat{\theta} \pm t_{n-2,\alpha/2}\widehat{\sigma}_{\widehat{\theta}}$, and tests of hypotheses concerning θ use the t_{n-2} reference distribution.

SPECIAL CASE: CONFIDENCE INTERVALS FOR THE MEAN OF Y: One very useful application of the preceding result is to the problem of estimating the **mean value** of Y at a fixed value of x, say, x_0. In our straight-line regression setting,

$$E(Y|x_0) \equiv \mu_{Y|x_0} = \beta_0 + \beta_1 x_0.$$

However, you quickly will note that $E(Y|x_0) = \beta_0 + \beta_1 x_0$ is just a linear combination of β_0 and β_1 with $c_0 = 1$ and $c_1 = x_0$. Thus, the previous result applies to this special situation! With $c_0 = 1$ and $c_1 = x_0$, we have that the variance of $\widehat{\theta} = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$ is given by (verify!)

$$V(\widehat{\theta}) \equiv \sigma^2_{\widehat{\theta}} = \sigma^2 \left(s_{00} + x_0^2 s_{11} + 2s_{01}x_0\right) = \sigma^2 \left\{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right\}.$$

An estimate of this variance is given by

$$\widehat{\sigma}^2_{\widehat{\theta}} = \text{MS[E]} \left\{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right\}.$$

Thus, a $100(1 - \alpha)$ percent confidence interval for $E(Y|x_0) = \beta_0 + \beta_1 x_0$, the **mean response** of Y for a fixed value of x_0, is given by

$$(\widehat{\beta}_0 + \widehat{\beta}_1 x_0) \pm t_{n-2,\alpha/2} \sqrt{\text{MS[E]} \left\{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right\}}.$$

LENGTH OF THIS CONFIDENCE INTERVAL: The confidence interval for $E(Y|x_0) = \beta_0 + \beta_1 x_0$ will be different depending on the value of x_0. In fact, the expression for $\widehat{\sigma}^2_{\widehat{\theta}}$ above will be smallest when $x_0 = \overline{x}$, and will get larger the farther x_0 is from $\overline{x}$ in either direction. This implies that the **precision** with which we estimate $E(Y|x_0) = \beta_0 + \beta_1 x_0$ decreases the farther we get away from $\overline{x}$. This makes intuitive sense—we would expect to have the most "confidence" in our fitted line near the "center" of the observed data. The result is that the confidence intervals for $E(Y|x_0) = \beta_0 + \beta_1 x_0$ will be wider the farther x_0 is from $\overline{x}$. Thus, if the fitted regression line is used to estimate means for values of x besides those used in the experiment, it is important to use a range of x values which contains the future values of interest.

EXTRAPOLATION: It is sometimes desired to estimate $E(Y|x_0) = \beta_0 + \beta_1 x_0$ based on the fit of the straight line for values of x_0 outside the range of x values used in the experiment. This is called **extrapolation**, and can be very dangerous. In order for our inferences to be valid, we must believe that the straight line relationship holds for x values outside the range where we have observed data. In some situations, this may be reasonable; in others, we may have no theoretical basis for making such a claim without data to support it. Thus, it is very important that the investigator have an honest sense of the relevance of the straight line model for values outside those used in the experiment if inferences such as estimating the mean for such x_0 values are to be reliable.

Example 4.4 (`oxygen.sas`). In Example 4.2, suppose that our researcher desires to get a 95 percent confidence interval for the mean O_2 rate of birds in an environment at $x_0 = 2.5$ degrees Celcius. She is thus interested in the linear combination

$$E(Y|x_0 = 2.5) = \beta_0 + 2.5\beta_1.$$

Using our last results, we see the desired confidence interval is

$$[3.4714 - 2.5(0.0878)] \pm 2.447 \times \sqrt{0.028\left\{\frac{1}{8} + \frac{[2.5-(-1.75)]^2}{1135.5}\right\}},$$

or $(3.0975, 3.4066)$ ml/g/hr. Thus, one would expect for the mean O_2 rate of birds living in a 2.5 degree Celcius environment to be between 3.0975 and 3.4066.

CONFIDENCE BANDS: A $100(1-\alpha)$ **percent confidence band** is simply the locus of confidence intervals for $E(Y|x) = \beta_0 + \beta_1 x$ for all x. A 95 percent confidence band for the bird-oxygen data from Example 4.2 is given in Figure 4.14.

UNFORTUNATE REALITY: If I obtain $100(1-\alpha)$ percent confidence intervals for $E(Y|x) = \beta_0 + \beta_1 x$ at many different values of x, the probability that all intervals contain their respective means is less than $1-\alpha$.

SIMULTANEOUS CONFIDENCE BANDS: A $100(1-\alpha)$ **percent simultaneous confidence band** for the true regression function $E(Y|x) = \beta_0 + \beta_1 x$ is given by

$$(\widehat{\beta}_0 + \widehat{\beta}_1 x) \pm \sqrt{2F_{2,n-2,\alpha}}\sqrt{\text{MS[E]}\left\{\frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}\right\}},$$

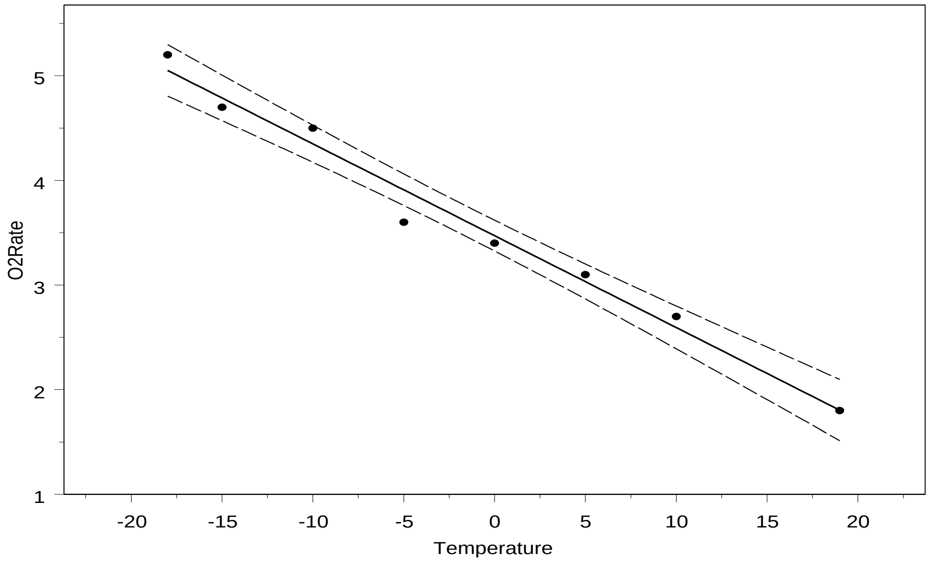

Figure 4.14: *A* 95 *percent confidence band on* $E(Y|x) = \beta_0 + \beta_1 x$ *for the bird oxygen data.*

for all $x \in \mathcal{R}$. Note that this expression is identical to the expressions for the endpoints for the confidence interval for $E(Y|x) = \beta_0 + \beta_1 x$ except that $\sqrt{2F_{2,n-2,\alpha}}$ replaces $t_{n-2,\alpha/2}$. Since $\sqrt{2F_{2,n-2,\alpha}} \geq t_{n-2,\alpha/2}$ for all n and α, the $100(1-\alpha)$ percent simultaneous confidence band is wider than the collection of all $100(1-\alpha)$ percent confidence intervals for $E(Y|x) = \beta_0 + \beta_1 x$. This simultaneous confidence band is sometimes called the **Working-Hotelling simultaneous confidence band**.

4.8 Prediction intervals for a future Y using simple-linear regression

Sometimes, depending on the context, we may not be interested in the **mean** $E(Y|x_0) = \beta_0 + \beta_1 x_0$, but rather the actual value of Y we might observe when $x = x_0$. On the surface, this may sound like the same problem, but they are, indeed, very different. For example, consider a stockbroker who would like to learn about the value of a stock based on previous data. In this setting, the stockbroker would like to **predict** or **forecast** the actual value of the stock, say, Y_0, that might be observed when $x = x_0$. On the other hand, the stockbroker probably does not care about what might happen "on the average" at some future time x_0; that is, she is probably not concerned with estimating $E(Y|x_0) = \beta_0 + \beta_1 x_0$.

REMARK: In this kind of situation, we are interested not in the **population mean** $E(Y|x_0) = \beta_0 + \beta_1 x_0$, but rather the actual value that might be taken on by the **random variable**, Y. In the context of our model, we are interested in the "future" observation

$$Y_0 = \beta_0 + \beta_1 x_0 + \epsilon_0,$$

where ϵ_0 is the "error" associated with Y_0 that makes it differ from the mean $\beta_0 + \beta_1 x_0$. It is important to recognise that Y_0 is not a parameter but rather is a **random variable**; thus, we do not wish to **estimate** a fixed parameter, but, instead, we wish to **predict** a random quantity.

POINT ESTIMATOR: Our point estimator of Y_0 is given by the quantity

$$\widehat{Y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0.$$

This is identical to before when we were estimating the mean $E(Y|x_0) = \beta_0 + \beta_1 x_0$. However, we use a different symbol in this context to remind ourselves that we are interested in predicting Y_0, not estimating $E(Y|x_0)$. We call $\widehat{Y}_0$ a **prediction** or **forecast** rather than an "estimate" to make the distinction clear. Of course, just as we do in the estimation of fixed parameters, we would still like to have some idea of how well we can predict/forecast. To get an idea, we would like to characterise the uncertainty that we have about $\widehat{Y}_0$ as a guess for Y_0. Intuitively, there will be two sources of error:

- part of the error in $\widehat{Y}_0$ arises from the fact that we do not know β_0 and β_1 and that they must **estimated** from the observed data.
- additional error arises from the fact that Y_0 itself is a **random variable**, so its value varies itself! Thus, additional uncertainty is introduced because we are trying to characterise a quantity that itself is uncertain.

THE SAMPLING DISTRIBUTION OF THE ERROR IN PREDICTION: The **error in prediction** $\widehat{Y}_0 - Y_0$ is normally distributed; more precisely,

$$\widehat{Y}_0 - Y_0 \sim \mathcal{N}\left\{0, \sigma^2\left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]\right\}.$$

Comparing the variance $\widehat{Y}_0 - Y_0$ to the variance of $\widehat{E(Y|x_0)}$ in the last section, we see that there is an extra "1" added on. This accounts for the additional variation arising from the fact that Y_0 itself is a random quantity, and we have to **predict** its value. It follows straightforwardly that

$$Z = \frac{\widehat{Y}_0 - Y_0}{\sqrt{\sigma^2\left\{1 + \frac{1}{n} + \frac{(x_0-\overline{x})^2}{S_{xx}}\right\}}} \sim \mathcal{N}(0, 1)$$

and that

$$t = \frac{\widehat{Y}_0 - Y_0}{\sqrt{\text{MS[E]}\left\{1 + \frac{1}{n} + \frac{(x_0-\overline{x})^2}{S_{xx}}\right\}}} \sim t_{n-2}.$$

Thus, using t as a pivot, it follows that

$$\widehat{Y}_0 \pm t_{n-2,\alpha/2}\underbrace{\sqrt{\text{MS[E]}\left\{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right\}}}_{\text{standard error of } \widehat{Y}_0 - Y_0}$$

is a $100(1-\alpha)$ percent **prediction interval** for Y_0.

Example 4.5 (`oxygen.sas`). In Example 4.2, suppose that our researcher desires to get a 95 percent prediction interval for a particular bird in an environment at $x_0 = 2.5$ degrees Celcius (compare this with Example 4.5). This prediction interval is given by

$$[3.4714 - 2.5(0.0878)] \pm 2.447 \times \sqrt{0.028\left\{1 + \frac{1}{8} + \frac{[2.5 - (-1.75)]^2}{1135.5}\right\}},$$

or $(2.8123, 3.6918)$, Thus, we are 95 percent confident that the O_2 rate for this particular bird will be between 2.8123 and 3.6918. One will note that the prediction interval for a single bird at $x_0 = 2.5$ is wider than the confidence interval for $E(Y|x_0 = 2.5)$. This is because of the additional variability arising from having to predict a random variable, Y_0, rather than estimating a mean, $E(Y|x_0)$.

PREDICTION BANDS: A $100(1-\alpha)$ **percent prediction band** is simply the locus of prediction intervals for Y_0 for all possible values of $x = x_0$. This provides a graphical representation of the prediction intervals. Figure 10.10 on p. 412 (Rao) shows a nice comparison between prediction bands for Y_0 and confidence bands for $E(Y|x) = \beta_0 + \beta_1 x$.

SIMULTANEOUS PREDICTION BANDS: A $100(1-\alpha)$ **percent simultaneous prediction band** for Y is given by

$$(\widehat{\beta}_0 + \widehat{\beta}_1 x) \pm \sqrt{2F_{2,n-2,\alpha}}\sqrt{\text{MS[E]}\left\{1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}\right\}},$$

for all $x \in \mathcal{R}$. Note that this expression is identical to the expressions for the endpoints for the prediction interval for Y_0 except that $\sqrt{2F_{2,n-2,\alpha}}$ replaces $t_{n-2,\alpha/2}$. This simultaneous prediction band is sometimes called the **Working-Hotelling simultaneous prediction band**.

4.9 The analysis of variance for simple linear regression

We may also use an **analysis of variance** approach to test the significance of the regression; this approach, as before, is based on a partitioning of total variability in the observed response data Y. To be precise, algebraically, it follows that

$$\underbrace{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}_{\text{SS[TOT]}} = \underbrace{\sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2}_{\text{SS[R]}} + \underbrace{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}_{\text{SS[E]}}.$$

- SS[TOT] denotes the **total sums of squares**. SS[TOT] measures the total variation in the data $Y_1, Y_2, ..., Y_n$. One will also note that this is just the numerator of the sample variance $S^2 = (n-1)^{-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$.

- SS[R] denotes the **regression sums of squares**. SS[R] measures the variation in the data $Y_1, Y_2, ..., Y_n$ explained by the straight-line model.

- SS[E] denotes the **error sums of squares**. SS[E] measures the variation in the data $Y_1, Y_2, ..., Y_n$ **not** explained by the straight-line model.

COMPUTING TIP: Note that an alternative expression for SS[R] is given by

$$\begin{aligned}
\text{SS[R]} &= \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2 \\
&= \sum_{i=1}^{n}\left[(\widehat{\beta}_0 + \widehat{\beta}_1 x_i) - (\widehat{\beta}_0 + \widehat{\beta}_1 \overline{x})\right]^2 \\
&= \widehat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \overline{x})^2 \\
&= \widehat{\beta}_1^2 S_{xx} = \frac{S_{xy}^2}{S_{xx}}.
\end{aligned}$$

THE ANOVA TABLE FOR STRAIGHT-LINE REGRESSION: Just as we did with the one-way layout, we can combine all of this information into a tabular display.

Table 4.12: *The general form of an analysis of variance table for straight-line regression.*

Source	*df*	SS	MS	*F*
Regression	1	SS[R]	MS[R]	$F = \frac{\text{MS[R]}}{\text{MS[E]}}$
Error	$n-2$	SS[E]	MS[E]	
Total	$n-1$	SS[TOT]		

NOTES ON THE GENERAL ANOVA TABLE STRUCTURE:

- The degrees of freedom *add down*. There are $n-1$ degrees of freedom associated with SS[TOT]; this can be viewed as a statistic that has "lost" a degree of freedom for having to estimate the overall mean of Y (β_0) with $\overline{Y}$. There is only one degree of freedom associated with SS[R] since there is only one predictor variable. The degrees of freedom for SS[E] can be thought of as the divisor needed to create an unbiased estimator of σ^2. Recall that SS[E]$/(n-2)$ is an unbiased estimator of σ^2.

- The sum of squares also *add down*; this follows from the algebraic identity presented at the beginning of this section; namely, SS[TOT] $=$ SS[R] $+$ SS[E].

- Mean squares are the sums of squares divided by their degrees of freedom.

USING THE F STATISTIC: The F statistic tests whether or not at least one of the independent variables add to the model; i.e., whether or not at least one of the β's associated with the predictor variables is nonzero. In the straight-line regression setting, we only have one x variable! Thus, in the straight-line setting, the F statistic tests

$$H_0 : \beta_1 = 0$$
$$\text{versus}$$
$$H_1 : \beta_1 \neq 0.$$

The interpretation of the test is as follows: Under the assumption that a straight line relationship exists, we are testing whether or not the slope of this relationship is, in fact, zero. A zero slope means that there is no systematic change in mean along with change in x; that is, there is **no linear association** between Y and x.

MATHEMATICAL FORMULATION OF THE F STATISTIC: More advanced linear model arguments (that we will not discuss) show that when H_0 is true, $\text{SS[R]}/\sigma^2 = \text{MS[R]}/\sigma^2 \sim \chi^2_1$ and that $\text{SS[E]}/\sigma^2 \sim \chi^2_{n-2}$. Furthermore, it follows that SS[R] and SS[E] are **independent**. Thus, when H_0 is true,

$$F = \frac{\text{MS[R]}}{\text{MS[E]}} = \frac{\text{SS[R]}/\sigma^2}{\frac{\text{SS[E]}}{\sigma^2}/(n-2)} \sim F_{1,n-2}.$$

Thus, we can use the $F_{1,n-2}$ distribution as a **reference distribution** to test $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$. Even though this is a two-sided H_1, we place the entire rejection region in the upper tail. Thus, the rejection region is given by $RR = \{F : F > F_{1,n-2,\alpha}\}$, where $F_{1,n-2,\alpha}$ denotes the $1-\alpha$ quantile of the $F_{1,n-2}$ distribution. P-values are computed as right tail areas on the $F_{1,n-2}$ distribution as well.

NOTE: The F test of $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ is equivalent to the **two-sided** t test of $H_0 : \beta_1 = \beta_{1,0}$ versus $H_1 : \beta_1 \neq \beta_{1,0}$, where $\beta_{1,0} = 0$. This follows (verify!) because

$$t^2 = \left(\frac{\widehat{\beta}_1 - 0}{\sqrt{s_{11}\text{MS[E]}}}\right)^2 = \frac{\text{MS[R]}}{\text{MS[E]}} = F$$

from the ANOVA table. The t test is a more flexible procedure than the F test if interest lies in drawing inference about β_1. With a t procedure, one can specify a nonzero value of $\beta_{1,0}$ and use one-sided alternatives. The F test, on the other hand, only allows a two-sided alternative with $\beta_{1,0} = 0$.

EXPECTED MEAN SQUARES: Theorem B.2 (Appendix B, Rao) can be used to show that the **expected mean squares**, that is, the values estimated by MS[R] and MS[E], are given by $E(\text{MS[R]}) = \sigma^2 + \beta_1^2 S_{xx}$ and $E(\text{MS[E]}) = \sigma^2$, respectively. Hence, if $\beta_1 = 0$, i.e., there is no linear relationship between Y and x, the two mean square statistics, MS[R] and MS[E], should be about the same since they are estimating the same quantity. This would yield an F statistic that was close to one. On the other hand, if $\beta_1 \neq 0$, then we would expect the F ratio to be larger than 1. This gives another reason of why the F ratio will get large when H_0 is not true.

THE COEFFICIENT OF DETERMINATION: Since SS[TOT] = SS[R] + SS[E], it follows that the proportion of the total variation in the data explained by the model is

$$R^2 = \frac{\text{SS[R]}}{\text{SS[TOT]}}.$$

The statistic R^2 is called the **coefficient of determination**. Clearly, $0 \leq R^2 \leq 1$. The larger the R^2, the better the deterministic part of the straight-line model $\beta_0 + \beta_1 x$ explains the variability in the data. Thus, an R^2 value "close" to 1 is taken as evidence that the regression model does "a good job" at describing the variability in the data.

IMPORTANT: It is critical to understand what R^2 does and does not measure. Its value is computed under the assumption that the simple linear regression model **is correct**; i.e., that it is a good description of the underlying relationship between Y and x. Thus, it assesses, if the relationship between x and Y really is a straight line, how much of the variation in the data may actually be attributed to that relationship rather than just to inherent variation. If R^2 is small, it may be that there is a lot of random inherent variation in the data, so that, although the straight line is a reasonable model, it can only explain so much of the observed overall variation. Alternatively, R^2 may be close to 1, but the straight-line model may not be the most appropriate model! In fact, R^2 may be quite "high," but, in a sense, is irrelevant, because it assumes the straight line model is correct. In reality, a better model may exist (e.g., a quadratic model, etc.).

Table 4.13: *The ANOVA table for the bird oxygen rate data from Example* 4.2.

Source	df	SS	MS	F
Regression	1	8.745	8.745	308.927
Error	6	0.170	0.028	
Total	7	8.915		

Example 4.6 (`oxygen.sas`). With the bird-oxygen rate data of Example 4.2, we now present the ANOVA table. The regression sum of squares is given by

$$\text{SS[R]} = \frac{S_{xy}^2}{S_{xx}} = \frac{(-99.65)^2}{1135.5} = 8.745.$$

The total sum of squares is given by

$$\text{SS[TOT]} = \sum_{i=1}^{8}(y_i - \overline{y})^2 = 8.915.$$

Thus, the error sum of squares (obtained by subtraction) is

$$\text{SS[E]} = \text{SS[TOT]} - \text{SS[R]} = 8.915 - 8.745 = 0.170.$$

The complete ANOVA table appears in Table 4.13.

ANALYSIS: If the researcher wants to test $H_0 : \beta_1 = 0$ (no linear trend between oxygen rate and temperature) versus $H_1 : \beta_1 \neq 0$, she would strongly reject H_0 since $F = 308.927$ is much larger than $F_{1,6,0.05} = 5.99$. There is overwhelming evidence to support the contention that the oxygen rate and temperature are linearly related. You will recall that this was the same conclusion reached by examining the confidence interval for β_1. The coefficient of determination is given by

$$R^2 = \frac{\text{SS[R]}}{\text{SS[TOT]}} = \frac{8.745}{8.915} = 0.981.$$

Thus, 98.1 percent of the variability in the oxygen rates is explained by the independent variable (temperature).

4.10 Checking the assumptions for straight-line regression

We have considered the **simple linear regression model** $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, for $i = 1, 2, ..., n$, where $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$. Of course, like all statistical models, there are assumptions that go along with this one; these assumptions are the following:

1. the errors are normally distributed (with mean zero),

2. the errors have constant variance,

3. the errors are independent, and

4. the true form of the regression function g, is, in fact, a straight line $g(x) = \beta_0 + \beta_1 x$.

RESIDUALS: As in the one-way layout, we never get to see the ϵ_i's (i.e., the errors) because they are unobservables. However, we can observe the residuals. Recall that

$$e_i = y_i - \widehat{y}_i$$

is the residual associated with y_i. Also, recall that $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$. Note how the model's residual takes the familiar form of "observed" y minus the "predicted" $\widehat{y}$.

DIAGNOSING NORMALITY: If we specify that $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$ and the normality assumption holds, then the residuals are also normally distributed. Mathematics can show that when the model holds, e_i, when viewed as a **random variable**; i.e.,

$$e_i = Y_i - \widehat{Y}_i \sim \mathcal{N}\left[0, \sigma^2\left\{1 - \left(\frac{1}{n} + \frac{(x_i - \overline{x})^2}{S_{xx}}\right)\right\}\right].$$

Thus, if the normality assumption is true, and the model holds, a histogram of the observed $e_i = y_i - \widehat{y}_i$ should look normally distributed, centered around zero.

NORMALITY PLOTS: As in the one-way layout, a **normal probability plot** is constructed by plotting the n ordered residuals against the n ordered quantiles from the standard normal distribution. If the normality assumption holds, and the model is correct, this plot should look like a straight line. Small departures from normality are usually of little concern in regression analysis. However, large departures could drastically affect the validity of our confidence intervals and hypothesis tests.

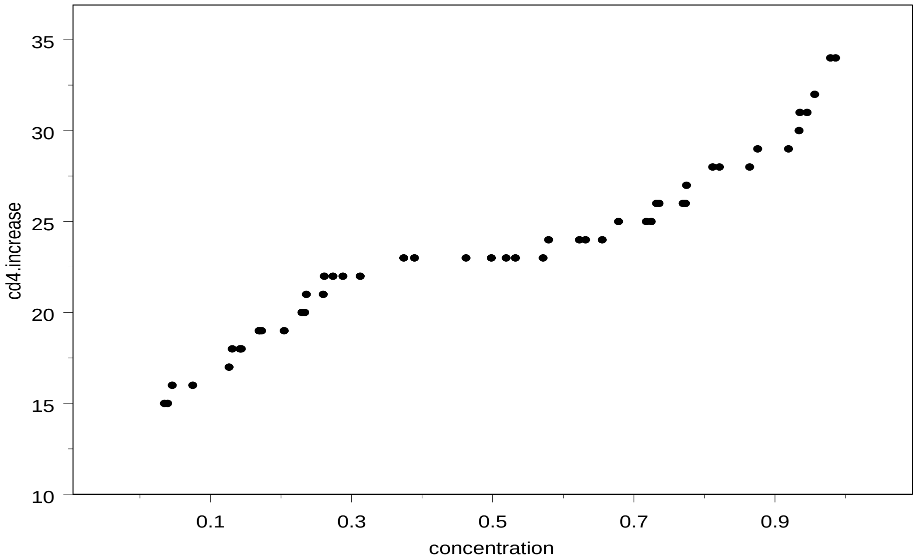

Figure 4.15: HIV *study:* CD4 *increase versus drug concentration.*

Example 4.7. In order to assess the effects of drug concentration on the resulting increase in CD4 counts, physicians used a sample of $n = 50$ advanced HIV patients with different drug concentrations and observed the resulting CD4 count increase. (Fictitious) data appear in Figure 4.15. There looks to be a significant linear trend between drug concentration and CD4 increase. In fact, the test for $H_0 : \beta_1 = 0$ is highly significant ($P < 0.0001$) and $R^2 = 0.927$. A normal probability plots looks fine.

DIAGNOSING NONCONSTANT VARIANCE AND OTHER MODEL INADEQUACIES: A good visual display to use for diagnosing nonconstant variance and model misspecification is the plot of residuals versus predicted values; i.e., a plot of e_i versus $\widehat{y}_i$. This is sometimes called a **residual plot**. If the model holds, it follows that

$$\text{Cov}(e_i, \widehat{Y}_i) = 0;$$

i.e., the residuals and predicted values are **uncorrelated**. *Thus, residual plots that display nonrandom patterns suggest that there are some problems with our model assumptions.* In particular, if a nonconstant variance problem exists, the residual plot will

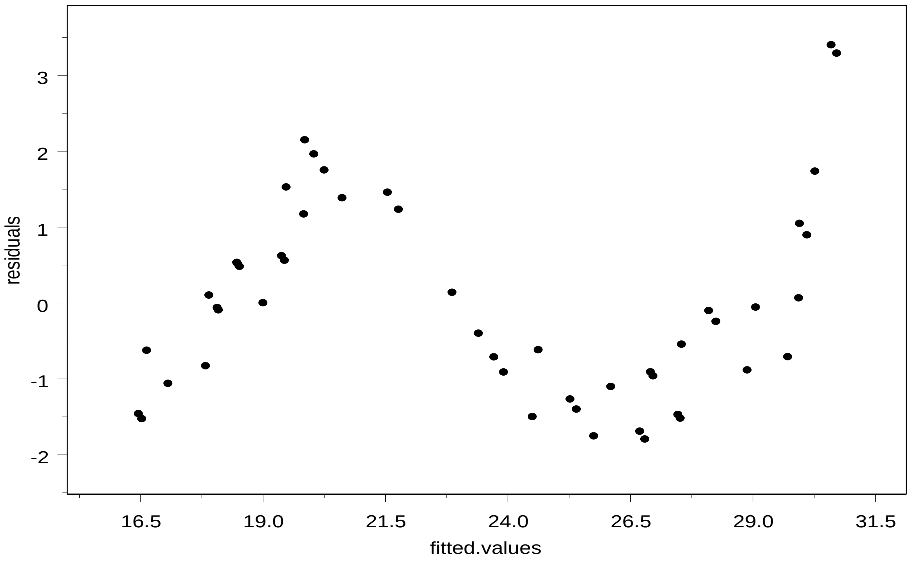

Figure 4.16: HIV *study: Residual plot from straight-line fit.*

likely display a "fanning out" shape. Also, if the true form of the regression function is misspecified, the residual plot should show the exact nature of the misspecification.

Example 4.7 (continued). From the residual plot in Figure 4.16, we see clear evidence that the straight line model $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ is *not* appropriate. The residual plot reveals that the straight-line fit just does not capture all the structure that is present in the data. The plot suggests that perhaps a curvlinear g function should be used, say, $g(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$.

Example 4.8. An entomological experiment was conducted to study the survivability of stalk borer larvae. It was of interest to develop a model relating the mean size of larvae (cm) as a function of the stalk head diameter (cm). Data from the experiment appear in Figure 4.17. There looks to be a moderate linear trend between larvae size and head diameter size. In fact, the test for $H_0 : \beta_1 = 0$ is highly significant ($P < 0.001$) and $R^2 = 0.625$. A normal probability plots looks fine. However, from the residual plot in Figure 4.18, we see clear evidence of a violation in the constant variance assumption.

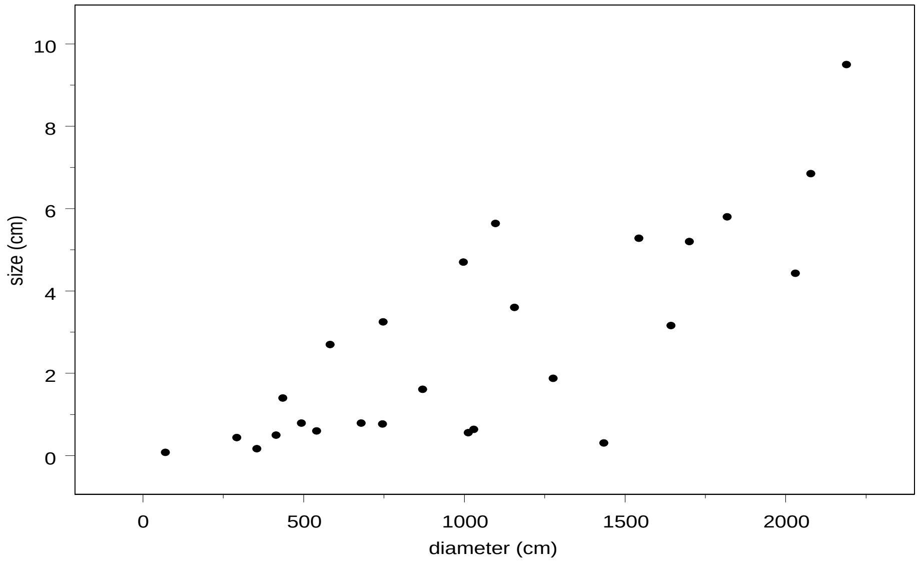

Figure 4.17: *Entomology experiment: Larvae size versus head diameter size.*

TRANSFORMATIONS: We have already discussed the notion of transforming the data as a way of handling violations of the usual assumptions. In the regression context, this may be done in a number of ways. One way is to invoke an appropriate transformation, and then postulate a regression model on the transformed scale. Sometimes, in fact, it may be that, although the data do exhibit constant variance on the original scale, they may on some transformed scale. *However, it is important to remember that if a transformation is used, the resulting inferences apply to this transformed scale* (and no longer to the original scale). Another approach is to proceed with a regression method known as **weighted-least squares**. In a weighted regression analysis, different responses are given different weights depending on their variances; see § 10.10 in Rao.

OUTLIERS: Another problem is that of **outliers**; i.e., data points that do not fit well with the pattern of the rest of the data. In straight-line regression, an outlier might be an observation that falls far off the apparent approximate straight line trajectory followed by the remaining observations. Practitioners often "toss out" such anomalous points, which may or may not be a good idea. If it is clear that an "outlier" is the result of a

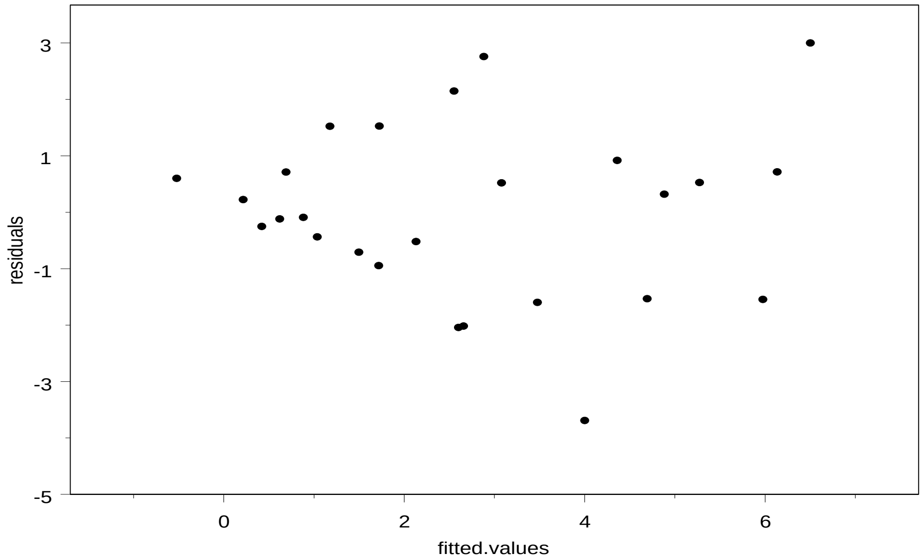

Figure 4.18: *Entomology experiment: Residual plot from straight-line fit.*

mishap or a gross recording error, then this may be acceptable. On the other hand, if no such basis may be identified, the outlier may, in fact, be a genuine response; in this case, it contains information about the process under study, and may be reflecting a legitimate phenomenon. In this case, "throwing out" an outlier may lead to misleading conclusions, because a legitimate feature is being ignored.

LEVERAGES: To identify outliers, we should consider first looking at the residual plot of e_i versus $\widehat{y}_i$. However, when interpreting residual plots, recall that

$$e_i = Y_i - \widehat{Y}_i \sim \mathcal{N}\left\{0, \sigma^2(1 - h_{ii})\right\},$$

where

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{S_{xx}}$$

is called the **leverage** for the ith case. So, in general, the residuals (unlike the errors) do not have constant variance! Also, the residuals (unlike the errors) are slightly correlated! Observations where x_i is far away from $\overline{x}$ will have large values of h_{ii}. However, not all observations with large leverages are necessarily outliers.

STUDENTISED RESIDUALS: To account for the different variances among residuals, we consider "studentising" the residuals (i.e., dividing by an estimate of their standard deviation). There are two ways to do this.

1. **Internally** studentised residuals:

   $$r_i = \frac{e_i}{\sqrt{s^2(1-h_{ii})}},$$

   where $s^2 = \text{MS[E]}$ is computed from all of the data. Calculations can show that $E(r_i) = 0$ and $V(r_i) = 1$; that is, internally studentised residuals have a constant variance regardless of the location of the x's. Values of $|r_i|$ larger than 3 or so should cause concern.

2. **Externally** studentised residuals:

   $$t_i = \frac{e_i}{\sqrt{s^2_{-i}(1-h_{ii})}},$$

   where $s^2_{-i} = \text{MS[E]}$ computed from all of the data **except** the ith case. It can be shown algebraically that, in the straight-line regression case,

   $$s^2_{-i} = \frac{(n-1)s^2 - e_i^2/(1-h_{ii})}{n-2}.$$

DETECTING OUTLIERS: If the ϵ_i's are normally distributed, then it turns out that, in the straight-line regression case, $t_i \sim t_{n-2}$. However, r_i does not follow a well-known distribution. Consequently, many feel that the externally studentised are more useful for outlier detection. How do we use the t_i's to formerly detect outliers? If the investigator suspects that case i may be an outlier, *prior to examining the data*, then since $t_i \sim t_{n-2}$, he could compare t_i to the t_{n-2} distribution and classify case i as an outlier if $|t_i| \geq t_{n-2,\alpha/2}$. In practice, however, often the investigator has no predetermined case to investigate as an outlier. So, he examines each case (computes the t_i's for all cases) and then chooses the largest one in magnitude to test as an outlier (so, actually, n tests are being carried out simultaneously). To correct for multiplicity, we can use the Bonferroni approach to multiple testing. To test for outliers in n observations at the α level of significance, we need to use $t_{n-2,\alpha/2n}$ as a critical value for each $|t_i|$.

STRATEGIES TO DEAL WITH OUTLIERS: What should we do if an outlier (or outliers) are identified? Unfortunately, there is no clear-cut answer! However, here are some possibilities:

1. Delete outliers and redo the analysis (new outliers may surface).
2. Sometimes the purpose of the experiment is just to identify the outliers. In this case, there is no need to redo the analysis.
3. Check the experimental circumstances surrounding the data collection for the outlying cases.
4. Report the analysis both with and without the analysis and let the reader decide.

4.11 Correlation analyses

In most of our discussions up until now, the variable x has been best regarded as **fixed**. In this setting, the methods of regression are appropriate in relating a response Y to x. In **observational data** situations, however, we do not choose the values of x; rather, we merely observe the pair (X, Y). Thus, the X variable is best regarded as **random**. In this setting, we, thus, must think about the **bivariate distribution** of the random vector (X, Y).

BIVARIATE DISTRIBUTIONS: Suppose that the random vector (X, Y) has the continuous joint pdf $f_{X,Y}(x, y)$. The **correlation coefficient**, $\rho_{X,Y} \equiv \rho$, is a parameter associated with the bivariate model $f_{X,Y}(x, y)$ that has the following properties:

- The linear relationship between X and Y is characterised by the parameter ρ, where $-1 \leq \rho \leq 1$. If $\rho = 1$, then all possible values of X and Y lie on a straight line with **positive** slope. If $\rho = -1$, then all possible values of X and Y lie on a straight line with **negative** slope. If $\rho = 0$, then there is **no linear relationship** between X and Y.

- When $0 < \rho < 1$, there is a tendency for the values to vary together in a positive way. When $-1 < \rho < 0$ there is a tendency for the values to vary together in a negative way.
- Recall that
$$\rho = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y},$$
where the covariance $\text{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$.

CORRELATION: The correlation coefficient ρ is a measure of the degree of **linear association** between two random variables. It is very important to understand what correlation does **not** measure. Investigators sometimes confuse the value of the correlation coefficient and the slope of an apparent underlying straight line relationship. Actually, these do not have anything to do with each other.

- The correlation coefficient may be very close to 1, implying an almost perfect association, but the slope may be very small. Although there is indeed an almost perfect association, the rate of change of Y values with X values may be very slow.
- The correlation coefficient may be very small, but the apparent "slope" of the relationship could be very steep. In this situation, it may be that, although the rate of change of Y values with X values is fast, there is large inherent variation in the data.

THE BIVARIATE NORMAL DISTRIBUTION: The random vector (X,Y) is said to have a **bivariate normal distribution** if its joint pdf is given by

$$f_{X,Y}(x,y|\mu_X,\mu_Y,\sigma_X^2,\sigma_Y^2,\rho) = \begin{cases} \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}e^{-Q/2}, & (x,y) \in \mathcal{R}^2 \\ 0, & \text{otherwise} \end{cases}$$

where

$$Q = \frac{1}{1-\rho^2}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right)+\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right].$$

FACTS ABOUT THE BIVARIATE NORMAL DISTRIBUTION:

- Marginally, $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$.
- If $\rho = 0$, then X and Y are **independent**. This is only true in the bivariate normal setting.

ESTIMATION: Suppose that $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ is an iid sample from a bivariate normal distribution with parameters $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2$, and ρ. The **likelihood equation** is given by

$$\begin{aligned} L \equiv L(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho|\boldsymbol{x}, \boldsymbol{y}) &= \prod_{i=1}^{n} f_{X,Y}(x_i, y_i|\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho) \\ &= \left(\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\right)^n e^{-\sum_{i=1}^{n} Q_i/2}, \end{aligned}$$

where

$$Q_i = \frac{1}{1-\rho^2}\left[\left(\frac{x_i-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x_i-\mu_X}{\sigma_X}\right)\left(\frac{y_i-\mu_Y}{\sigma_Y}\right)+\left(\frac{y_i-\mu_Y}{\sigma_Y}\right)^2\right].$$

Maximum likelihood estimators for the parameters $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2$, and ρ are obtained by maximising L with respect to $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2$, and ρ. These estimators are given by $\widehat{\mu}_X = \overline{X}$, $\widehat{\mu}_Y = \overline{Y}$, $\widehat{\sigma}_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$, $\widehat{\sigma}_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$, and

$$r_{XY} \equiv r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}},$$

where $S_{yy} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \text{SS[TOT]}$. We sometimes call r the **sample correlation coefficient**.

FACT: Further intuition is gained from that, using $\widehat{\beta}_1 = S_{xy}/S_{xx}$ as an estimator for β_1, we have

$$r^2 = \frac{S_{xy}^2/S_{xx}}{S_{yy}} = \frac{\text{SS[R]}}{\text{SS[TOT]}} = R^2.$$

Thus, the square of the sample correlation equals the coefficient of determination for straight-line regression. This illustrates a nice computational link between regression and correlation analyses. The quantity r^2 is often called the **coefficient of determination** (like R^2) in this setting, where correlation analysis is appropriate. However, it is important to recognise that the interpretation is very different! Here, we are not acknowledging a straight line relationship; rather, we are just modelling the data in terms of a bivariate normal distribution with correlation ρ. Thus, the former interpretation for the quantity r^2 has no meaning here. Likewise, the idea of correlation really only has meaning when both variables X and Y are random variables.

HYPOTHESIS TESTS CONCERNING ρ: Suppose that the random vector (X, Y) has a bivariate normal distribution with parameters $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2$, and ρ. A conditioning argument can be used to show that

$$E(Y|X = x) = \beta_0 + \beta_1 x,$$

where $\beta_1 = \rho \times \sigma_Y/\sigma_X$ and $\beta_0 = \mu_Y - \beta_1\mu_X$. How is this helpful? Suppose that in an observational data setting, it is desired to test $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$, say. Since σ_X and σ_Y are both positive, it follows that $\rho = 0 \Leftrightarrow \beta_1 = 0$ and that $\rho \neq 0 \Leftrightarrow \beta_1 \neq 0$ (similar relations hold if $H_1 : \rho < 0$ or $H_1 : \rho > 0$). Thus, we can perform a hypothesis test involving ρ by using the test statistic

$$t = \frac{\widehat{\beta}_1 - 0}{\sqrt{s_{11}\text{MS[E]}}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}},$$

and comparing t to the t_{n-2} distribution.

REMARK: It may also be of interest to test $H_0 : \rho = \rho_0$ versus an alternative where $\rho_0 \neq 0$. In this case, one is forced to appeal to an approximate result based on the **Fisher** Z **transformation**. The method is based on the result that

$$W_r = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \sim \mathcal{AN}\left[\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right), \frac{1}{n-3}\right],$$

when n is large. The test is carried out by computing

$$Z_r = \sqrt{n-3}\left[\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) - \frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right)\right],$$

and comparing Z_r to the standard normal distribution.

WARNING: This procedure is only **approximate**, even under our bivariate normal assumption. It is an example of the type of approximation that is often made in difficult problems, that of approximating the behaviour of a statistic under the condition that the sample size, n, is large. If n is small, the procedure may be unreliable. Thus, testing aside, one should be very wary of "over-interpreting" the estimate of ρ when n is small; after all, one "outlying" or "unusual" observation could be enough to affect the computed value substantially! It may be very difficult to detect when ρ is significantly different from zero with a small sample size.

CONFIDENCE INTERVALS: Because r is an estimator of the population parameter ρ, it would be desirable to report, along with the estimate itself, a **confidence interval** for ρ. Since

$$W_r = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \sim \mathcal{AN}\left[\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right), \frac{1}{n-3}\right],$$

when n is large, we can construct an approximate $100(1-\alpha)$ confidence interval for

$$W_\rho = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right);$$

this interval is given by (W_L', W_U'), where W_L' and W_U' are given by

$$W_L' = W_r - z_{\alpha/2}\sqrt{\frac{1}{n-3}}$$

and

$$W_U' = W_r + z_{\alpha/2}\sqrt{\frac{1}{n-3}},$$

respectively (verify!). Now, once we have this confidence interval, we can **transform** the endpoints W_L' and W_U' to obtain the endpoints for the confidence interval for ρ. Applying the necessary transformation (which as turns out is the hyperbolic tangent function, tanh), we obtain

$$(\tanh W_L', \tanh W_U') \quad \text{or} \quad \left[\frac{\exp(2W_L')-1}{\exp(2W_L')+1}, \frac{\exp(2W_U')-1}{\exp(2W_U')+1}\right]$$

as an approximate $100(1-\alpha)$ percent confidence interval for ρ.

Example 4.9 (`wingtail.sas`). The following data are measurements on wing length (X) and tail length (Y) for a sample of $n = 12$ birds. Both measurements are in centimeters.

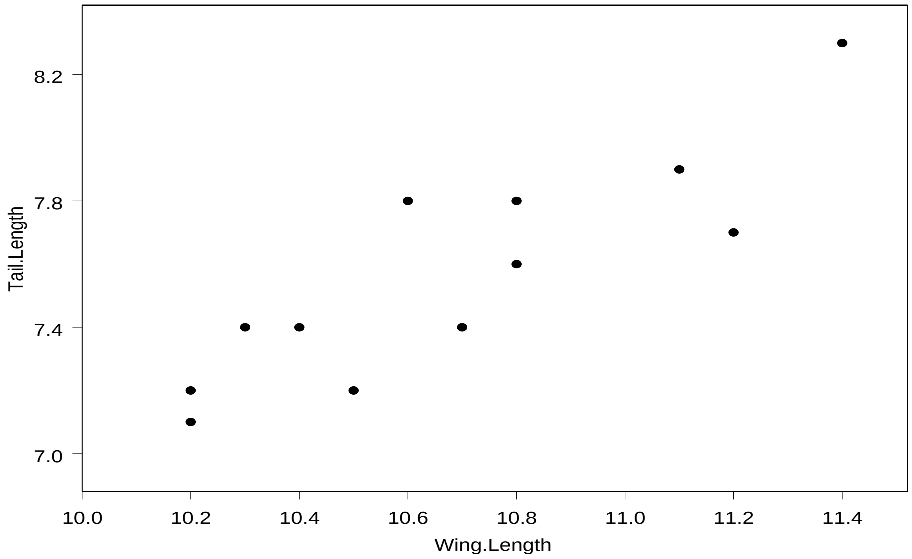

Figure 4.19: *Tail length and wing length for twelve birds.*

Wing X:	10.4	10.8	11.1	10.2	10.3	10.2	10.7	10.5	10.8	11.2	10.6	11.4
Tail Y:	7.4	7.6	7.9	7.2	7.4	7.1	7.4	7.2	7.8	7.7	7.8	8.3

The scatterplot of the data in Figure 4.19 reveals a positive association. It will be of interest to determine if the correlation between X and Y, ρ, is significantly different from zero. Straightforward calculations show that $S_{xx} = 1.717$, $S_{yy} = 1.347$, and $S_{xy} = 1.323$. Thus, our estimate of ρ is given by

$$r = \frac{1.323}{\sqrt{(1.717)(1.347)}} = 0.8704.$$

We can also compute an approximate confidence interval for ρ using the Fisher Z transformation method we discussed. We first compute

$$W_r = \frac{1}{2}\ln\left(\frac{1+0.8704}{1-0.8704}\right) = 1.335.$$

We have $z_{0.025} = 1.96$ and $\sqrt{1/(n-3)} = 1/3$, so that the confidence interval for W_ρ is given by $1.335 \pm 1.96 \times (1/3)$ or $(0.681, 1.988)$. Finally, we transform the interval for ρ

itself and obtain

$$\left[\frac{\exp(2\times 0.681)-1}{\exp(2\times 0.681)+1}, \frac{\exp(2\times 1.988)-1}{\exp(2\times 1.988)+1}\right],$$

or $(0.592, 0.963)$. The interval does not contain zero. Thus, if we were test $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$, using $\alpha = 0.05$, we would reject H_0 and say that there is an association between wing and tail lengths (under the assumption that our bivariate normal distribution model is correct).

4.12 Comparison of regression and correlation models

We have identified two appropriate statistical models for thinking about the problem of assessing **association** between two variables X and Y. These may be thought of as

- **Fixed** x: Postulate a model for the mean of the random variable Y as a function of x (in particular, we focused on the straight line function $g(x) = \beta_0 + \beta_1 x$) and then estimate the parameters in the model to characterise the relationship.
- **Random** X: Characterise the **linear relationship** between X and Y by the **correlation** between them (in a bivariate normal probability model) and estimate the correlation parameter, ρ.

PARALLELS: Arithmetic operations for regression analysis under the first scenario and correlation analysis under the second are the same! That is, to **fit** the regression model by estimating the intercept and slope parameters and to **estimate** the correlation coefficient, we use the same arithmetic operations. Of course, as always, the important issue is in the **interpretation** of the results.

SUBTLETY: In settings where X is best regarded as a **random variable**, many investigators still want to fit regression models treating X as fixed. This is because, although correlation does describe the "degree of association" between X and Y, it does not characterise the relationship in a way suitable for some purposes. For example, an investigator may desire to **predict** the yield based on observing the average height of plants on a plot. The correlation coefficient does not allow this. He would rather fit a regression model, even though X is random. Is this legitimate? If we are careful about the interpretation, it may be. If X and Y are really both observed random variables, and we fit a regression to characterise the relationship, technically, any subsequent analyses are regarded as "conditional on the values of X involved." This means that we essentially regard X as "fixed," even though it is not. However, this may be adequate for the prediction problem for our experimenter. **Conditional** on having seen a particular average height, he wants to get a "best guess" for yield. He is not saying that he could **control** heights and thereby influence yields; only that, given he sees a certain height, he might be able to say something about the associated yield. This subtlety is an important one, but important. Inappropriate use of statistical techniques could lead one to erroneous or irrelevant inferences.

CORRELATION VERSUS CAUSATION: Investigators are often tempted to infer a **causal relationship** between X and Y when they fit a regression model or perform a correlation analysis. However, a significant association between X and Y in either situation does not necessarily imply a causal relationship!

Example 4.10. A Chicago newspaper reported that "there is a strong correlation between the numbers of fire fighters (X) at a fire and the amount of damage (Y, measured in \$1000's) that the fire does." Data from 20 recent fires in the Chicago area appear in Figure 4.20. From the plot, there appears to be a strong linear association between X and Y. Few people, however, would infer that the increase in the number of fire trucks *causes* the observed increase in damages! Often, when two variables X and Y have a strong association, it is because both X and Y are, in fact, each associated with a third variable, say W. In the example, both X and Y are probably strongly linked to W, the severity of the fire, so it is understandable that X and Y would increase together.

MORAL: This phenomenon is the basis of the remark "Correlation does not necessarily imply causation." An investigator should be aware of the temptation to infer causation

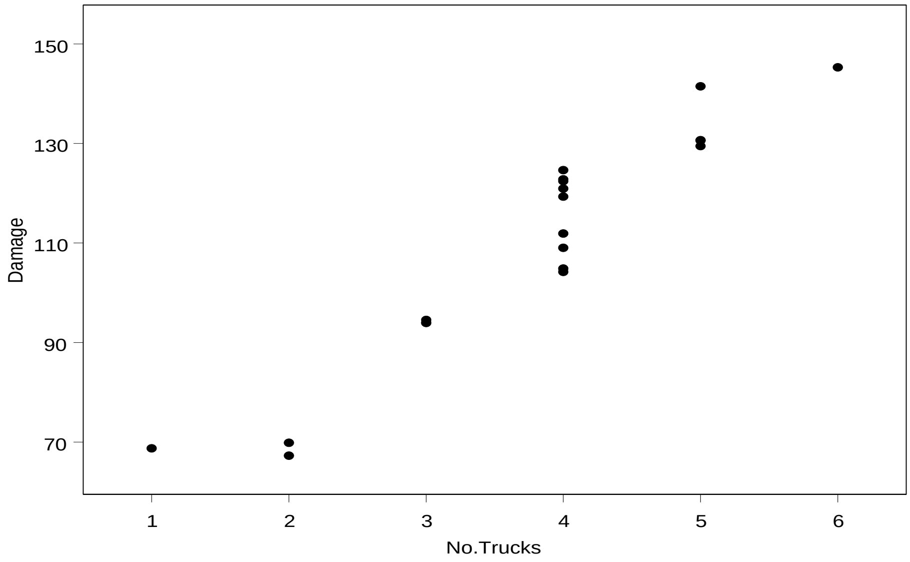

Figure 4.20: *Chicago fire damages* ($\$1000's$) *and the number of fire trucks.*

in setting up a study, and be on the lookout for "lurking" variables like W above that are actually the driving force behind observed results. In general, the best way to **control** the effects of "lurking" variables is to use a carefully designed experiment. In observational studies, it is very difficult to make causal statements. Oftentimes, the best we can do is make statements documenting the observed association, and nothing more.

5 Matrix Algebra Results for Linear Models

5.1 Basic definitions and results

TERMINOLOGY: A **matrix** $\boldsymbol{A}$ is a rectangular array of elements; e.g.,

$$\boldsymbol{A} = \begin{pmatrix} 3 & 5 & 4 \\ 1 & 2 & 8 \end{pmatrix}.$$

Elements of $\boldsymbol{A}$ are denoted by a_{ij}; e.g., $a_{11} = 3$, $a_{12} = 5$, ..., $a_{23} = 8$. In some instances, we may write $\boldsymbol{A} = (a_{ij})$. The i refers to the row; the j refers to the column. In general, the **dimensions** of $\boldsymbol{A}$ are n (the number of rows) by m (the number of columns). If we want to emphasise the dimension of $\boldsymbol{A}$, we may write $\boldsymbol{A}_{n\times m}$. If $n = m$, we call $\boldsymbol{A}$ a **square** matrix.

TERMINOLOGY: If $\boldsymbol{A} = (a_{ij})$ is an $n \times m$ matrix, the **transpose** of $\boldsymbol{A}$, denoted $\boldsymbol{A}'$, is the $m \times n$ matrix (a_{ji}); e.g.,

$$\boldsymbol{A}' = \begin{pmatrix} 3 & 1 \\ 5 & 2 \\ 4 & 8 \end{pmatrix}.$$

TERMINOLOGY: For any square matrix $\boldsymbol{A}$, if $\boldsymbol{A}' = \boldsymbol{A}$, we say that $\boldsymbol{A}$ is **symmetric**; that is, $a_{ij} = a_{ji}$ for all values of i and j.

TERMINOLOGY: A **vector** is a matrix consisting of one column or one row. A **column vector** is denoted by $\boldsymbol{a}_{n\times 1}$, and a **row vector** is denoted by $\boldsymbol{a}_{1\times m}$. We will assume that all vectors written as $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$, etc., are column vectors. All vectors written as $\boldsymbol{a}'$, $\boldsymbol{b}'$, $\boldsymbol{c}'$, etc., are row vectors.

$$\boldsymbol{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} \qquad \boldsymbol{a}' = \begin{pmatrix} a_1 & a_2 & \cdots & a_n \end{pmatrix}.$$

TERMINOLOGY: Two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ are said to be **orthogonal** if $\boldsymbol{a}'\boldsymbol{b} = 0$.

MATRIX MULTIPLICATION: If $\boldsymbol{A}_{n\times m}$ and $\boldsymbol{B}_{m\times p}$, then $\boldsymbol{AB}$ is an $n\times p$ matrix.

RESULT: If the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are of conformable dimensions,

(a) $(\boldsymbol{A}')' = \boldsymbol{A}$

(b) $(\boldsymbol{AB})' = \boldsymbol{B}'\boldsymbol{A}'$

(c) $(\boldsymbol{A}+\boldsymbol{B})' = \boldsymbol{A}'+\boldsymbol{B}'$.

(d) $\boldsymbol{A}'\boldsymbol{A}$ and $\boldsymbol{AA}'$ are symmetric.

NOTATION: The **determinant** of an $n\times n$ matrix $\boldsymbol{A}$ is denoted by $|\boldsymbol{A}|$ or $\det(\boldsymbol{A})$.

SPECIAL CASE: The determinant of the 2×2 matrix

$$\boldsymbol{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

is given by $|\boldsymbol{A}| = ad - bc$.

RESULT: If the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are of conformable dimensions,

(a) for any square matrix $\boldsymbol{A}$, $|\boldsymbol{A}'| = |\boldsymbol{A}|$.

(b) for any $n\times n$ upper (lower) triangular matrix (this includes diagonal matrices),

$$|\boldsymbol{A}| = \prod_{i=1}^{n} a_{ii}.$$

TERMINOLOGY: A popular matrix in linear models is the $n\times n$ **identity matrix** $\boldsymbol{I}$ given by

$$\boldsymbol{I} = \boldsymbol{I}_n = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix};$$

that is, $a_{ij} = 1$ for $i = j$, and $a_{ij} = 0$ when $i \neq j$.

TERMINOLOGY: Another popular matrix is the $n \times n$ **matrix of ones $\boldsymbol{J}$** given by

$$\boldsymbol{J} = \boldsymbol{J}_n = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix};$$

that is, $a_{ij} = 1$ for all i and j. Note that $\boldsymbol{J} = \boldsymbol{1}\boldsymbol{1}'$, where $\boldsymbol{1}$ is an $n \times 1$ vector of 1s.

TERMINOLOGY: The $n \times n$ matrix where $a_{ij} = 0$ for all i and j, is called the **zero matrix**, and is denoted $\boldsymbol{0}$.

TERMINOLOGY: If $\boldsymbol{A}$ is an $n \times n$ matrix, and there exists a matrix $\boldsymbol{C}$ such that $\boldsymbol{AC} = \boldsymbol{CA} = \boldsymbol{I}$, then $\boldsymbol{A}$ is said to be **nonsingular**, and $\boldsymbol{C}$ is called the **inverse** of $\boldsymbol{A}$, henceforth denoted as $\boldsymbol{A}^{-1}$. If $\boldsymbol{A}$ is nonsingular, the inverse matrix $\boldsymbol{A}^{-1}$ is unique. If $\boldsymbol{A}$ is not nonsingular, we say that $\boldsymbol{A}$ is **singular**, in which case $\boldsymbol{A}^{-1}$ does not exist.

SPECIAL CASE: The inverse of the 2×2 matrix

$$\boldsymbol{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \text{is given by} \quad \boldsymbol{A}^{-1} = \frac{1}{ad - bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

SPECIAL CASE: The inverse of the $n \times n$ diagonal matrix

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{pmatrix} \quad \text{is given by} \quad \boldsymbol{A}^{-1} = \begin{pmatrix} a_{11}^{-1} & 0 & \cdots & 0 \\ 0 & a_{22}^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{nn}^{-1} \end{pmatrix}.$$

RESULT: If $\boldsymbol{A}$ and $\boldsymbol{B}$ are $n \times n$ matrices,

(a) $\boldsymbol{A}$ is nonsingular if and only if $|\boldsymbol{A}| \neq 0$.

(b) then if $\boldsymbol{A}$ and $\boldsymbol{B}$ are nonsingular, $\boldsymbol{(AB)}^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$.

(c) then if $\boldsymbol{A}$ is nonsingular, $(\boldsymbol{A}')^{-1} = (\boldsymbol{A}^{-1})'$.

TERMINOLOGY: An $n \times n$ matrix $\boldsymbol{A}$ is said to be **idempotent** if $\boldsymbol{A}^2 = \boldsymbol{A}$.

TERMINOLOGY: If $\boldsymbol{A}$ is an $n \times n$ matrix, the **trace** of $\boldsymbol{A}$ is defined as follows:

$$\text{tr}(\boldsymbol{A}) = \sum_{i=1}^{n} a_{ii};$$

that is, $\text{tr}(\boldsymbol{A})$ is the sum of the diagonal elements of $\boldsymbol{A}$. Also, it follows that for any conformable matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, $\text{tr}(a\boldsymbol{A} + b\boldsymbol{B}) = a\text{tr}(\boldsymbol{A}) + b\text{tr}(\boldsymbol{B})$, for any constants a and b, and $\text{tr}(\boldsymbol{AB}) = \text{tr}(\boldsymbol{BA})$.

5.2 Linear independence and rank

TERMINOLOGY: The $n \times 1$ vectors $\boldsymbol{a}_1$, $\boldsymbol{a}_2$, ..., $\boldsymbol{a}_m$ are said to be **linearly dependent** if and only if there exist scalars $c_1, c_2, ..., c_m$ such that

$$\sum_{i=1}^{m} c_i \boldsymbol{a}_i = \mathbf{0}$$

and at least one of the c_i's is not zero. On the other hand, if

$$\sum_{i=1}^{m} c_i \boldsymbol{a}_i = \mathbf{0} \Longrightarrow c_1 = c_2 = \cdots = c_m = 0,$$

then we say that $\boldsymbol{a}_1$, $\boldsymbol{a}_2$, ..., $\boldsymbol{a}_m$ are **linearly independent**.

RESULT: Suppose that $\boldsymbol{a}_1$, $\boldsymbol{a}_2$, ..., $\boldsymbol{a}_m$ is a set of $n \times 1$ vectors. Then,

(a) the vectors are linearly dependent if and only if it is possible to express at least one vector as a linear combination of the others.

(b) the vectors are linearly independent if and only if it is **not** possible to express one vector as a linear combination of the others.

TERMINOLOGY: The **rank** of any matrix $\boldsymbol{A}$ is defined as

$$\begin{aligned} r(\boldsymbol{A}) &= \text{number of linearly independent columns of } \boldsymbol{A} \\ &= \text{number of linearly independent rows of } \boldsymbol{A}. \end{aligned}$$

RESULT: The number of linearly independent rows of any matrix is equal to the number of linearly independent columns.

TERMINOLOGY: Suppose that $\boldsymbol{A}$ is an $n \times p$ matrix. Then $r(\boldsymbol{A}) \leq \min\{n, p\}$. If $r(\boldsymbol{A}) = \min\{n, p\}$, then $\boldsymbol{A}$ is said to have **full rank**. If $r(\boldsymbol{A}) = n$, we say that $\boldsymbol{A}$ is of full *row* rank. If $r(\boldsymbol{A}) = p$, we say that $\boldsymbol{A}$ is of full *column* rank. If $r(\boldsymbol{A}) < \min\{n, p\}$, we say that $\boldsymbol{A}$ is **rank deficient** or **less than full rank**.

REALISATION: Since the maximum possible rank of an $n \times p$ matrix is the minimum of n and p, for any **rectangular** matrix, either the rows or columns (or both) must be linearly dependent!

Example 5.1. The rank of

$$\boldsymbol{A} = \begin{pmatrix} 1 & -2 & 1 \\ 5 & 2 & 17 \end{pmatrix}$$

is 2 because the two rows are linearly independent (neither is a multiple of the other). Thus, $\boldsymbol{A}$ has full row rank. Furthermore, by the definition of rank, the number of linearly independent columns is also 2. Therefore, the columns are linearly dependent; that is, there exist constants c_1, c_2, and c_3 such that

$$c_1 \begin{pmatrix} 1 \\ 5 \end{pmatrix} + c_2 \begin{pmatrix} -2 \\ 2 \end{pmatrix} + c_3 \begin{pmatrix} 1 \\ 17 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Taking $c_1 = 3$, $c_2 = 1$ and $c_3 = -1$ are examples of such constants; that is, $\boldsymbol{a}_3$, the third of column of $\boldsymbol{A}$, is equal to $3\boldsymbol{a}_1 + \boldsymbol{a}_2$, where $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ are the first and second columns of $\boldsymbol{A}$, respectively.

CONNECTION: For any $n \times n$ matrix $\boldsymbol{A}$, the following statements are equivalent:

$$r(\boldsymbol{A}) = n \iff \boldsymbol{A}^{-1} \text{ exists} \iff |\boldsymbol{A}| \neq 0.$$

IMPORTANT FACT: For any matrix $\boldsymbol{A}$, $r(\boldsymbol{A}'\boldsymbol{A}) = r(\boldsymbol{A})$.

USEFUL FACT: For any $n \times n$ idempotent matrix $\boldsymbol{A}$, $r(\boldsymbol{A}) = \text{tr}(\boldsymbol{A})$.

5.3 Generalised inverses and systems of equations

TERMINOLOGY: A **generalised inverse** of a matrix $\boldsymbol{A}$ is any matrix $\boldsymbol{G}$ which satisfies $\boldsymbol{AGA} = \boldsymbol{A}$. We usually denote the generalised inverse of $\boldsymbol{A}$ by $\boldsymbol{A}^{-}$. Although any matrix $\boldsymbol{A}$ has a generalised inverse, we'll pay special attention to the case when $\boldsymbol{A}$ is symmetric.

REMARK: If $\boldsymbol{A}$ is nonsingular, then $\boldsymbol{A}^{-} = \boldsymbol{A}^{-1}$.

SYSTEMS OF EQUATIONS: Consider the (linear) system of equations

$$\boldsymbol{A}_{p\times p}\boldsymbol{x}_{p\times 1} = \boldsymbol{c}_{p\times 1}.$$

Of course, if $r(\boldsymbol{A}) = p$, then $\boldsymbol{A}^{-1}$ exists, and $\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{c}$; i.e., we get a unique solution. If $r(\boldsymbol{A}) < p$, however, $\boldsymbol{A}^{-1}$ does not exist uniquely. In this case, the system $\boldsymbol{Ax} = \boldsymbol{c}$ might have (a) no solution, (b) finitely many solutions, or (c) infinitely many solutions!

TERMINOLOGY: A linear system $\boldsymbol{Ax} = \boldsymbol{c}$ is said to be **consistent** if it has a solution; i.e., there exists an $\boldsymbol{x}^*$ such that $\boldsymbol{Ax}^* = \boldsymbol{c}$.

IMPORTANT FACT: If $\boldsymbol{Ax} = \boldsymbol{c}$ is a consistent system, then $\boldsymbol{x} = \boldsymbol{A}^{-}\boldsymbol{c}$ is a solution.
Proof. Suppose that $\boldsymbol{Ax} = \boldsymbol{c}$ is consistent; that is, there exists an $\boldsymbol{x}^*$ such that $\boldsymbol{Ax}^* = \boldsymbol{c}$. It follows that $\boldsymbol{x} = \boldsymbol{A}^{-}\boldsymbol{c}$ is a solution since $\boldsymbol{AA}^{-}\boldsymbol{c} = \boldsymbol{AA}^{-}\boldsymbol{Ax}^* = \boldsymbol{Ax}^* = \boldsymbol{c}$. □

COMPUTING GENERALISED INVERSES: Consider the following algorithm for finding a generalised inverse $\boldsymbol{A}^{-}$ for any $n \times p$ matrix of rank r.

1. Find any $r \times r$ nonsingular submatrix $\boldsymbol{C}$. It is not necessary that the elements of $\boldsymbol{C}$ occupy adjacent rows and columns in $\boldsymbol{A}$.
2. Find $\boldsymbol{C}^{-1}$ and $(\boldsymbol{C}^{-1})'$.
3. Replace the elements of $\boldsymbol{C}$ by the elements of $(\boldsymbol{C}^{-1})'$.
4. Replace all other elements of $\boldsymbol{A}$ by zeros.
5. Transpose the resulting matrix.

5.4 Column and row spaces

TERMINOLOGY: For the matrix $\boldsymbol{A}_{n\times m} = [\boldsymbol{a}_1\ \boldsymbol{a}_2\ \cdots\ \boldsymbol{a}_m]$, where $\boldsymbol{a}_j$ is $n\times 1$, the **column space** of $\boldsymbol{A}$,

$$\begin{aligned}\mathcal{C}(\boldsymbol{A}) &= \left\{\boldsymbol{v}\in\mathcal{R}^n : \boldsymbol{v} = \sum_{j=1}^m c_j\boldsymbol{a}_j;\ c_j\in\mathcal{R}\right\}\\ &= \{\boldsymbol{v}\in\mathcal{R}^n : \boldsymbol{v} = \boldsymbol{A}\boldsymbol{c};\ \boldsymbol{c}\in\mathcal{R}^m\},\end{aligned}$$

is the set of all $n\times 1$ vectors *spanned by the columns* of $\boldsymbol{A}$; that is, $\mathcal{C}(\boldsymbol{A})$ is the set of all vectors that can be written as a **linear combination** of the columns of $\boldsymbol{A}$.

TERMINOLOGY: Similarly, the **row space** of $\boldsymbol{A}$, denoted $\mathcal{R}(\boldsymbol{A})$, is the set of all vectors that can be written as a linear combination of the rows of $\boldsymbol{A}$.

RESULT: For any matrix $\boldsymbol{A}$,

(a) $\mathcal{C}(\boldsymbol{A}'\boldsymbol{A}) = \mathcal{C}(\boldsymbol{A})$ and $\mathcal{R}(\boldsymbol{A}'\boldsymbol{A}) = \mathcal{R}(\boldsymbol{A})$

(b) $\mathcal{R}(\boldsymbol{A}') = \mathcal{C}(\boldsymbol{A})$ and $\mathcal{C}(\boldsymbol{A}') = \mathcal{R}(\boldsymbol{A})$.

5.5 Quadratic forms

TERMINOLOGY: Let $\boldsymbol{x} = (x_1, x_2, ..., x_n)'$ denote an $n\times 1$ vector, and suppose that $\boldsymbol{A}$ is an $n\times n$ symmetric matrix. A **quadratic form** is a function $f : \mathcal{R}^n \to \mathcal{R}$ of the form

$$f(\boldsymbol{x}) = \sum_{i=1}^n\sum_{j=1}^n a_{ij}x_ix_j = \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}.$$

The matrix $\boldsymbol{A}$ is sometimes called the **matrix of the quadratic form**. Note that this sum will involve both squared terms x_i^2 and the cross-product terms x_ix_j. With $\boldsymbol{x}\neq\boldsymbol{0}$,

- if $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}\geq 0$, the quadratic form and the matrix $\boldsymbol{A}$ are said to be **nonnegative definite**.
- if $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} > 0$, the quadratic form and the matrix $\boldsymbol{A}$ are said to be **positive definite**.

5.6 Eigenvalues and eigenvectors

TERMINOLOGY: Let $\boldsymbol{A}$ be an $n \times n$ matrix. The **characteristic equation** of $\boldsymbol{A}$ is given by

$$|\boldsymbol{A} - \lambda \boldsymbol{I}_n| = 0.$$

This equation in λ has exactly n roots $\lambda_1, \lambda_2, ..., \lambda_n$ which are called **eigenvalues**. These values are not necessarily distinct or even real. The solutions to the characteristic equation are identical to the solutions to the equation

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x},$$

for some $\boldsymbol{x} \neq \boldsymbol{0}$. An $\boldsymbol{x}$ corresponding to λ_i, denoted as $\boldsymbol{x}_i$ is called an **eigenvector** associated with λ_i.

Example 5.2. Let

$$\boldsymbol{A} = \begin{pmatrix} 3 & 0 \\ 8 & -1 \end{pmatrix}.$$

It follows that

$$\boldsymbol{A} - \lambda \boldsymbol{I}_2 = \begin{pmatrix} 3 & 0 \\ 8 & -1 \end{pmatrix} - \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix} = \begin{pmatrix} 3-\lambda & 0 \\ 8 & -1-\lambda \end{pmatrix},$$

which has determinant equal to $(3-\lambda)(-1-\lambda)$. Thus, the eigenvalues are 3 and -1. The eigenvector associated with $\lambda = 3$ is $\boldsymbol{x}_i = (1, 2)'$ since

$$\begin{pmatrix} 3 & 0 \\ 8 & -1 \end{pmatrix}\begin{pmatrix} 1 \\ 2 \end{pmatrix} = 3\begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 3 \\ 6 \end{pmatrix}.$$

RESULT: For any **symmetric** matrix $\boldsymbol{A}$,

(a) $\boldsymbol{A}$ is positive definite $\Leftrightarrow$ all eigenvalues of $\boldsymbol{A}$ are positive $\Leftrightarrow$ $|\boldsymbol{A}| > 0$

(b) $\boldsymbol{A}$ is nonnegative definite $\Leftrightarrow$ all eigenvalues of $\boldsymbol{A}$ are nonnegative $\Leftrightarrow$ $|\boldsymbol{A}| \geq 0$.

(c) all of $\boldsymbol{A}$'s eigenvalues are real.

5.7 Means and variances of random vectors

TERMINOLOGY: Suppose that $Z_1, Z_2, ..., Z_n$ are random variables. We call

$$\boldsymbol{Z} = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{pmatrix}$$

a **random vector**. The (multivariate) pdf of $\boldsymbol{Z}$ is denoted by $f_{\boldsymbol{Z}}(\boldsymbol{z})$. It describes mathematically how $Z_1, Z_2, ..., Z_n$ are distributed jointly (and, hence, is sometimes called a **joint distribution**). If $Z_1, Z_2, ..., Z_n$ are iid from $f_Z(z)$, the joint pdf of $\boldsymbol{Z}$ is given by

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = \prod_{i=1}^{n} f_Z(z_i).$$

THE MEAN AND VARIANCE OF A RANDOM VECTOR: Suppose that $E(Z_i) = \mu_i$, $V(Z_i) = \sigma_i^2$, for $i = 1, 2, ..., n$, and $\text{Cov}(Z_i, Z_j) = \sigma_{ij}$ for $i \neq j$. We define the **mean** of a random vector to be

$$E(\boldsymbol{Z}) = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} = \boldsymbol{\mu}$$

and the **variance** of a random vector to be the $n \times n$ matrix

$$V(\boldsymbol{Z}) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{pmatrix} = \boldsymbol{V}.$$

We may also refer to $\boldsymbol{V}$ as the **variance-covariance matrix** of $\boldsymbol{Z}$, because it contains the variances $\sigma_1^2, \sigma_2^2, ..., \sigma_n^2$ on the diagonal and the $\binom{n}{2}$ covariance terms $\text{Cov}(Z_i, Z_j)$, for $i < j$. Since $\text{Cov}(Z_i, Z_j) = \text{Cov}(Z_j, Z_i)$, it follows that the variance-covariance matrix $\boldsymbol{V}$ is **symmetric**.

THE COVARIANCE OF TWO RANDOM VECTORS: Suppose that $\boldsymbol{Y}_{n\times 1}$ and $\boldsymbol{Z}_{m\times 1}$ are random vectors. The covariance between $\boldsymbol{Y}$ and $\boldsymbol{Z}$ is given by

$$\text{Cov}(\boldsymbol{Y},\boldsymbol{Z}) = \begin{pmatrix} \text{Cov}(Y_1,Z_1) & \text{Cov}(Y_1,Z_2) & \cdots & \text{Cov}(Y_1,Z_m) \\ \text{Cov}(Y_2,Z_1) & \text{Cov}(Y_2,Z_2) & \cdots & \text{Cov}(Y_2,Z_m) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(Y_n,Z_1) & \text{Cov}(Y_n,Z_2) & \cdots & \text{Cov}(Y_n,Z_m) \end{pmatrix}_{n\times m}.$$

MEAN AND VARIANCE OF LINEAR FUNCTIONS OF A RANDOM VECTOR: Suppose that $\boldsymbol{Z}$, $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ are all $n\times 1$ random vectors and that $\boldsymbol{a}_{m\times 1}$, $\boldsymbol{A}_{m\times n}$, and $\boldsymbol{B}_{m\times n}$ are nonrandom (i.e., they contain only non-varying constants). The following facts are easily shown:

1. $E(\boldsymbol{a}+\boldsymbol{B}\boldsymbol{Z}) = \boldsymbol{a}+\boldsymbol{B}E(\boldsymbol{Z}) = \boldsymbol{a}+\boldsymbol{B}\boldsymbol{\mu}$
2. $V(\boldsymbol{a}+\boldsymbol{B}\boldsymbol{Z}) = \boldsymbol{B}V(\boldsymbol{Z})\boldsymbol{B}' = \boldsymbol{B}\boldsymbol{V}\boldsymbol{B}'$
3. $\text{Cov}(\boldsymbol{A}\boldsymbol{Z}_1,\boldsymbol{B}\boldsymbol{Z}_2) = \boldsymbol{A}\text{Cov}(\boldsymbol{Z}_1,\boldsymbol{Z}_2)\boldsymbol{B}'$.

MEAN OF A QUADRATIC FORM: Let $\boldsymbol{Y}$ be an n-dimensional random vector with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{V}$. Then, $\boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y}$ is a **quadratic form** and

$$E(\boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y}) = \boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu} + \text{tr}(\boldsymbol{A}\boldsymbol{V}).$$

5.8 The multivariate normal distribution

We have talked about the (univariate) normal distribution for a random variable Y and the bivariate normal distribution for a two-dimensional random vector $(Y_1,Y_2)'$. It turns out that we can quite naturally extend the notion of joint normality of $\boldsymbol{Y} = (Y_1,Y_2,...,Y_n)'$ to n dimensions.

TERMINOLOGY: The random vector $\boldsymbol{Y} = (Y_1,Y_2,...,Y_n)'$ is said to have a **multivariate normal distribution** with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{V}$ if its (joint) pdf is given by

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{V}|^{1/2}} \exp\left\{-(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right\},$$

for all $\boldsymbol{y} \in \mathcal{R}^n$. Shorthand notation for this statement is $\boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{V})$.

FACT: If $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_n)' \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{V})$, then, marginally, each $Y_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$.

FACT: Suppose that $\boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{V})$ and that $\boldsymbol{a}_{m\times 1}$ and $\boldsymbol{B}_{m\times n}$ are nonrandom. Then, $\boldsymbol{a} + \boldsymbol{BY} \sim \mathcal{N}_m(\boldsymbol{a} + \boldsymbol{B\mu}, \boldsymbol{BVB}')$. Of course, if $m = 1$, then $\boldsymbol{a}_{m\times 1}$ is a scalar, $\boldsymbol{B}_{m\times n}$ is a row vector, and $\boldsymbol{a} + \boldsymbol{BY}$ has a univariate normal distribution.

Example 5.3. With $\boldsymbol{Y} = (Y_1, Y_2, Y_3)'$, suppose that

$$\boldsymbol{Y} \sim \mathcal{N}_3\left\{\begin{pmatrix} 4 \\ 6 \\ 10 \end{pmatrix}, \begin{pmatrix} 8 & 5 & 0 \\ 5 & 12 & 4 \\ 0 & 4 & 9 \end{pmatrix}\right\}.$$

(a) Find the distribution of $X_1 = Y_1 - Y_2 + Y_3$.

(b) Let $X_2 = Y_1 - 3Y_2 + 2Y_3$. Find the joint distribution of X_1 and X_2.

(c) Find ρ_{X_1,X_2}, the correlation between X_1 and X_2.

SOLUTION. (a) Writing

$$X_1 = \begin{pmatrix} 1 & -1 & 1 \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix},$$

we identify $a = 0$ and $\boldsymbol{B}_{1\times 3} = (1, -1, 1)$. Straightforward calculations show that

$$\boldsymbol{B\mu} = \begin{pmatrix} 1 & -1 & 1 \end{pmatrix} \begin{pmatrix} 4 \\ 6 \\ 10 \end{pmatrix} = 8$$

and

$$\boldsymbol{BVB}' = \begin{pmatrix} 1 & -1 & 1 \end{pmatrix} \begin{pmatrix} 8 & 5 & 0 \\ 5 & 12 & 4 \\ 0 & 4 & 9 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} = 11.$$

Thus, $X_1 \sim \mathcal{N}(8, 11)$.

(b) Writing

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} 1 & -1 & 1 \\ 1 & -3 & 2 \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix},$$

we identify $\boldsymbol{a} = \boldsymbol{0}$ and

$$\boldsymbol{B} = \begin{pmatrix} 1 & -1 & 1 \\ 1 & -3 & 2 \end{pmatrix}.$$

Straightforward calculations show that

$$\boldsymbol{B}\boldsymbol{\mu} = \begin{pmatrix} 1 & -1 & 1 \\ 1 & -3 & 2 \end{pmatrix} \begin{pmatrix} 4 \\ 6 \\ 10 \end{pmatrix} = \begin{pmatrix} 8 \\ 6 \end{pmatrix}$$

and

$$\boldsymbol{B}\boldsymbol{V}\boldsymbol{B}' = \begin{pmatrix} 1 & -1 & 1 \\ 1 & -3 & 2 \end{pmatrix} \begin{pmatrix} 8 & 5 & 0 \\ 5 & 12 & 4 \\ 0 & 4 & 9 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ -1 & -3 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 11 & 22 \\ 22 & 74 \end{pmatrix}.$$

Thus,

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}_2 \left\{ \begin{pmatrix} 8 \\ 6 \end{pmatrix}, \begin{pmatrix} 11 & 22 \\ 22 & 74 \end{pmatrix} \right\}.$$

(c) The correlation between X_1 and X_2 is given by

$$\rho_{X_1,X_2} = \frac{\sigma_{X_1,X_2}}{\sigma_{X_1}\sigma_{X_2}} = \frac{22}{\sqrt{(11)(74)}} \approx 0.771.$$

NOTE: To do the $\boldsymbol{B}\boldsymbol{V}\boldsymbol{B}'$ calculation, I used the MAPLE commands

```
> with(linalg);
> B:=array(1..2,1..3,[[1,-1,1],[1,-3,2]]);
> V:=array(1..3,1..3,[[8,5,0],[5,12,4],[0,4,9]]);
> variance:=multiply(B,V,transpose(B));
```

NOTE: MAPLE is great for matrix calculations!

6 Introduction to Multiple Regression Models

Complimentary reading from Rao: Chapter 11 (§ 11.1-11.6).

6.1 Introduction

In Chapter 4, we considered the simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, for $i = 1, 2, ..., n$, where $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$. We talked about three main issues: (a) least-squares estimation and inference for β_0 and β_1, (b) estimating mean values of Y and predicting future values of Y, and (c) model diagnostics. While the straight-line model serves as an adequate description for many situations, more often than not, researchers engaged in model building consider more than just one predictor variable x. In fact, it is often the case that the researcher has a set of p **candidate** predictor variables, say, $x_1, x_2, ..., x_p$, and desires to model Y as a function of one or more of these p variables. To accommodate this situation, we must extend our linear regression model to handle more than one predictor variable.

Example 6.1. The taste of matured cheese is related to the concentration of several chemicals in the final product. In a study of cheddar cheese from the LaTrobe Valley of Victoria, Australia, samples of cheese were analysed for their chemical composition and were subjected to taste tests. Overall taste scores were obtained by combining the scores from several tasters. Data were collected on the following variables:

$Y =$ taste test score (`TASTE`)
$x_1 =$ concentration of acetic acid (`ACETIC`)
$x_2 =$ concentration of hydrogen sulfide (`H2S`)
$x_3 =$ concentration of lactic acid (`LACTIC`).

Variables `ACETIC` and `H2S` are both on the (natural) log scale. The variable `LACTIC` has not been transformed. Table 6.14 contains concentrations of the various chemicals in $n = 30$ specimens of mature cheddar cheese and the observed taste score.

Table 6.14: *Cheese data.*

TASTE	ACETIC	H2S	LACTIC	TASTE	ACETIC	H2S	LACTIC
12.3	4.543	3.135	0.86	40.9	6.365	9.588	1.74
20.9	5.159	5.043	1.53	15.9	4.787	3.912	1.16
39.0	5.366	5.438	1.57	6.4	5.412	4.700	1.49
47.9	5.759	7.496	1.81	18.0	5.247	6.174	1.63
5.6	4.663	3.807	0.99	38.9	5.438	9.064	1.99
25.9	5.697	7.601	1.09	14.0	4.564	4.949	1.15
37.3	5.892	8.726	1.29	15.2	5.298	5.220	1.33
21.9	6.078	7.966	1.78	32.0	5.455	9.242	1.44
18.1	4.898	3.850	1.29	56.7	5.855	10.20	2.01
21.0	5.242	4.174	1.58	16.8	5.366	3.664	1.31
34.9	5.740	6.142	1.68	11.6	6.043	3.219	1.46
57.2	6.446	7.908	1.90	26.5	6.458	6.962	1.72
0.7	4.477	2.996	1.06	0.7	5.328	3.912	1.25
25.9	5.236	4.942	1.30	13.4	5.802	6.685	1.08
54.9	6.151	6.752	1.52	5.5	6.176	4.787	1.25

Suppose that the researchers postulate that each of the three chemical composition covariates x_1, x_2, and x_3 are important in describing the taste. In this case, they might initially consider the following regression model

$$Y_i = \underbrace{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}}_{g(x_1,x_2,x_3)} + \epsilon_i,$$

for $i = 1, 2, ..., 30$. Are there other predictor variables that influence taste not considered here? Alternatively, what if not all of x_1, x_2, and x_3 are needed in the model? After all, it may be that one or more of x_1, x_2, and x_3 are not helpful in describing taste. For example, if the acetic acid concentration (x_1) is not helpful in describing taste, then we might consider a smaller model which excludes it; i.e.,

$$Y_i = \underbrace{\beta_0 + \beta_2 x_{i2} + \beta_3 x_{i3}}_{g(x_2,x_3)} + \epsilon_i,$$

for $i = 1, 2, ..., 30$. *The goal of any regression modelling problem should be to identify each of the important predictors, and then find the smallest model that does the best job.*

MULTIPLE REGRESSION SETTING: Consider an experiment in which n observations are collected on the response variable Y and p predictor variables $x_1, x_2, ..., x_p$. Schematically, we can envision data from such an experiment as in Table 6.15.

Table 6.15: *Schematic representation of data collected in a multiple regression setting.*

Individual	$\boldsymbol{Y}$	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\cdots$	$\boldsymbol{x}_p$
1	Y_1	x_{11}	x_{12}	$\cdots$	x_{1p}
2	Y_2	x_{21}	x_{22}	$\cdots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	Y_n	x_{n1}	x_{n2}	$\cdots$	x_{np}

To describe Y as a function of the p independent variables $x_1, x_2, ..., x_p$, we posit the **multiple linear regression model**

$$Y_i = \underbrace{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}}_{g(x_1, x_2, ..., x_p)} + \epsilon_i,$$

for $i = 1, 2, ..., n$, where $n > p + 1$ and $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$. The values β_0, β_1, ..., β_p, as before, are still called regression coefficients, and, since we are talking about regression models, we assume that $x_1, x_2, ..., x_p$ are all **fixed**, measured without error. Here, the random errors ϵ_i are still assumed to be independent and have a normal distribution with mean zero and a **common** variance σ^2. Just as before in the straight-line case, our main goals in regression analysis will be to

- estimate the regression parameters, $\beta_0, \beta_1, ..., \beta_p$,
- diagnose the fit (i.e., perform model diagnostics), and
- estimate mean responses and make predictions about future values.

PREVIEW: To estimate the regression parameters β_0, β_1, ..., β_p, we will still use the method of **least squares**. Simple computing formulae for parameter estimators are no longer available (this is the price to pay for making the leap to a multiple-regression setting!). We can, however, find closed-form solutions in terms of matrices and vectors.

6.2 Multiple linear regression models using matrices

Repeatedly writing the multiple linear regression model as

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i,$$

for $i = 1, 2, ..., n$, where $\epsilon_i \sim \text{iid } \mathcal{N}(0, \sigma^2)$, can get tiring. It turns out that we can more succinctly express the model, and, hence, greatly streamline our presentation, with the use of matrices and vectors. In particular, defining

$$\boldsymbol{Y}_{n\times 1} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \boldsymbol{X}_{n\times(p+1)} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta}_{(p+1)\times 1} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix},$$

and

$$\boldsymbol{\epsilon}_{n\times 1} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix},$$

the model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i,$$

for $i = 1, 2, ..., n$, where $\epsilon_i \sim \text{iid } \mathcal{N}(0, \sigma^2)$, can be equivalently expressed as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$.

NOTE: The matrix $\boldsymbol{X}$ is sometimes called the **design matrix**. It contains all of the predictor variable information and is assumed to be fixed (i.e., not random). The vector $\boldsymbol{\beta}$ contains all the true regression coefficients (which are also assumed to be fixed quantities). The only random quantities in the model are $\boldsymbol{\epsilon}$ and $\boldsymbol{Y}$. Since $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $V(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_n$, it follows that $E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$ and that $V(\boldsymbol{Y}) = \sigma^2 \boldsymbol{I}_n$. Since $\boldsymbol{\epsilon}$ has a normal distribution, so does $\boldsymbol{Y}$; thus, $\boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I}_n)$. Summarising, $\boldsymbol{Y}$ has a multivariate normal distribution with mean

$$E(\boldsymbol{Y}) = \boldsymbol{X\beta} = \begin{pmatrix} \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} \\ \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} \\ \vdots \\ \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} \end{pmatrix}$$

and variance

$$V(\boldsymbol{Y}) = \sigma^2\boldsymbol{I}_n = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix}.$$

ESTIMATING $\boldsymbol{\beta}$ *USING LEAST SQUARES*: The notion of least-squares estimation here is the same as it was in the straight-line model. In a multiple-regression setting, we want to find the values of $\beta_0, \beta_1, ..., \beta_p$ that **minimise** the sum or squared deviations

$$\text{SSE}(\beta_0, \beta_1, ..., \beta_p) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_p x_{ip})^2,$$

or, in matrix notation, the value of $\boldsymbol{\beta}$ that minimises

$$\text{SSE}(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X\beta})'(\boldsymbol{Y} - \boldsymbol{X\beta}).$$

Because $(\boldsymbol{Y} - \boldsymbol{X\beta})'(\boldsymbol{Y} - \boldsymbol{X\beta}) = (\boldsymbol{Y} - \boldsymbol{X\beta})'\boldsymbol{I}_n(\boldsymbol{Y} - \boldsymbol{X\beta})$ is a quadratic form; i.e., it is just a scalar function of the $p+1$ elements of $\boldsymbol{\beta}$, it is possible to use calculus to determine the values of the $p+1$ elements that minimise it. Formally, one would take the $p+1$ partial derivatives, with respect to each of $\beta_0, \beta_1, ..., \beta_p$, respectively, and set these expressions equal to zero; i.e.,

$$\frac{\partial \text{SSE}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \begin{pmatrix} \frac{\partial \text{SSE}}{\partial \beta_0} \\ \frac{\partial \text{SSE}}{\partial \beta_1} \\ \vdots \\ \frac{\partial \text{SSE}}{\partial \beta_p} \end{pmatrix} \overset{\text{set}}{=} \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

These are called the **normal equations**. These rather intimidating expressions, in non-matrix form, appear on p. 485 (Rao). Solving the normal equations for $\beta_0, \beta_1, ..., \beta_p$ gives the least squares estimators $\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_p$.

NORMAL EQUATIONS: Using the calculus of matrices (see Searle, 1982) makes this *much* easier; in particular, we can show the normal equations can be expressed as

$$-2\boldsymbol{X}'\boldsymbol{Y}+2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}=\mathbf{0} \quad \text{or, equivalently} \quad \boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}=\boldsymbol{X}'\boldsymbol{Y}.$$

Provided that $\boldsymbol{X}'\boldsymbol{X}$ is full rank, the (unique) value of $\boldsymbol{\beta}$ that solves this minimisation problem is given by

$$\widehat{\boldsymbol{\beta}}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}.$$

This is called the **least-squares estimator** of $\boldsymbol{\beta}$. Each entry in the vector is the least-squares estimator of the corresponding value of β_i, for $i=0,1,...,p$; i.e.,

$$\widehat{\boldsymbol{\beta}}=\begin{pmatrix}\widehat{\beta}_0\\ \widehat{\beta}_1\\ \widehat{\beta}_2\\ \vdots\\ \widehat{\beta}_p\end{pmatrix}.$$

NOTE: For the least-squares estimator $\widehat{\boldsymbol{\beta}}$ to be unique, we need $\boldsymbol{X}$ to be of **full column rank**; i.e., $r(\boldsymbol{X})=p+1$. That is, there are no linear dependencies among the columns of $\boldsymbol{X}$. If $r(\boldsymbol{X})<p+1$, then, since $r(\boldsymbol{X})=r(\boldsymbol{X}'\boldsymbol{X})$, $\boldsymbol{X}'\boldsymbol{X}$ does not have a unique inverse. In this case, the normal equations can not be solved uniquely.

WORKING ASSUMPTION: To avoid the more technical details of working with non full rank matrices (for now) we will assume, unless otherwise stated, that $\boldsymbol{X}$ *is* full rank. In this case, we know that $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ exists uniquely.

NOTE: Computing $\widehat{\boldsymbol{\beta}}$ for general p is not feasible by hand, of course, particularly nasty is the inversion of $\boldsymbol{X}'\boldsymbol{X}$ when p is large. Software for multiple regression analysis includes routines for inverting a matrix of any dimension; thus, estimation of $\boldsymbol{\beta}$ for a general multiple regression model is best carried out in this fashion.

RAO'S NOTATION: For some reason, Rao calls the $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ matrix $\boldsymbol{S}$. I will not use this notation.

ERROR SUM OF SQUARES: We define the **error sum of squares** by

$$\text{SS[E]} \equiv \text{SSE}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = (\boldsymbol{Y} - \widehat{\boldsymbol{Y}})'(\boldsymbol{Y} - \widehat{\boldsymbol{Y}}) = \boldsymbol{e}'\boldsymbol{e}.$$

The vector $\widehat{\boldsymbol{Y}} \equiv \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ contains the n fitted values. The vector $\boldsymbol{e} \equiv \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ contains the n least-squares residuals.

ESTIMATION OF σ^2: An unbiased estimator of σ^2 is given by

$$\widehat{\sigma}^2 = \text{MS[E]} \equiv \frac{\text{SS[E]}}{n-p-1} = \frac{(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})}{n-p-1}.$$

Proof. It may be demonstrated (verify!) that

$$\text{SS[E]} = (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = \boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{M})\boldsymbol{Y},$$

where $\boldsymbol{M} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. Since $E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$ and $V(\boldsymbol{Y}) = \sigma^2\boldsymbol{I}_n$, it follows that

$$E[\boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{M})\boldsymbol{Y}] = \underbrace{(\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{I}_n - \boldsymbol{M})\boldsymbol{X}\boldsymbol{\beta}}_{=\,0} + \text{tr}[(\boldsymbol{I}_n - \boldsymbol{M})\sigma^2\boldsymbol{I}_n]$$

(you might recall means of quadratic forms from Chapter 5). However, it is easy to see that $\boldsymbol{M}\boldsymbol{X} = \boldsymbol{X}$, in which case the first term is zero. As for the second term, note that,

$$\begin{aligned} \text{tr}[(\boldsymbol{I}_n - \boldsymbol{M})\sigma^2\boldsymbol{I}_n] &= \sigma^2\{\text{tr}(\boldsymbol{I}_n) - \text{tr}(\boldsymbol{M})\} \\ &= \sigma^2\{n - \text{tr}[\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']\} \end{aligned}$$

Now, since $\text{tr}(\boldsymbol{A}\boldsymbol{B}) = \text{tr}(\boldsymbol{B}\boldsymbol{A})$ for any conformable matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, we can write the last expression as

$$\begin{aligned} \sigma^2\{n - \text{tr}[\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']\} &= \sigma^2\{n - \text{tr}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}]\} \\ &= \sigma^2\{n - \text{tr}(\boldsymbol{I}_{p+1})\} \\ &= \sigma^2(n-p-1). \end{aligned}$$

We have shown that $E(\text{SS[E]}) = \sigma^2(n-p-1)$. Thus,

$$E(\text{MS[E]}) = E\left(\frac{\text{SS[E]}}{n-p-1}\right) = \sigma^2,$$

showing that, indeed, MS[E] is an **unbiased estimator** of σ^2. □

Example 6.2 (`cheese.sas`). With the cheese data in Example 6.1, consider the full model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i,$$

for $i = 1, 2, ..., 30$, or in matrix notation $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$. From Table 6.14, we have

$$\boldsymbol{y}_{30\times 1} = \begin{pmatrix} 12.3 \\ 20.9 \\ 39.0 \\ \vdots \\ 5.5 \end{pmatrix}, \quad \text{and} \quad \boldsymbol{X}_{30\times 4} = \begin{pmatrix} 1 & 4.543 & 3.135 & 0.86 \\ 1 & 5.159 & 5.043 & 1.53 \\ 1 & 5.366 & 5.438 & 1.57 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 6.176 & 4.787 & 1.25 \end{pmatrix}.$$

Our parameter vector (to be estimated) is

$$\boldsymbol{\beta}_{4\times 1} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}.$$

We compute

$$\boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} 30 & 164.941 & 178.254 & 43.260 \\ 164.941 & 916.302 & 1001.806 & 240.879 \\ 178.254 & 1001.806 & 1190.343 & 269.113 \\ 43.26 & 240.879 & 269.113 & 65.052 \end{pmatrix},$$

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{pmatrix} 3.795 & -0.760 & 0.087 & -0.071 \\ -0.760 & 0.194 & -0.020 & -0.128 \\ 0.087 & -0.020 & 0.015 & -0.046 \\ -0.071 & -0.128 & -0.046 & 0.726 \end{pmatrix},$$

and

$$\boldsymbol{X}'\boldsymbol{y} = \begin{pmatrix} 736.000 \\ 4194.442 \\ 5130.932 \\ 1162.065 \end{pmatrix}.$$

Thus, the least squares estimate of $\boldsymbol{\beta}$ for these data is given by

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \\
&= \begin{pmatrix} 3.795 & -0.760 & 0.087 & -0.071 \\ -0.760 & 0.194 & -0.020 & -0.128 \\ 0.087 & -0.020 & 0.015 & -0.046 \\ -0.071 & -0.128 & -0.046 & 0.726 \end{pmatrix} \begin{pmatrix} 736.000 \\ 4194.442 \\ 5130.932 \\ 1162.065 \end{pmatrix} \\
&= \begin{pmatrix} -28.877 \\ 0.328 \\ 3.912 \\ 19.670 \end{pmatrix} = \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \widehat{\beta}_3 \end{pmatrix}.
\end{aligned}$$

Our least-squares regression equation becomes

$$\widehat{Y}_i = -28.877 + 0.328x_{i1} + 3.912x_{i2} + 19.670x_{i3},$$

or, in terms of the variable names,

$$\widehat{\texttt{TASTE}}_i = -28.877 + 0.328\texttt{ACETIC}_i + 3.912\texttt{H2S}_i + 19.670\texttt{LACTIC}_i.$$

An unbiased estimate of the error variance σ^2 is given by

$$\widehat{\sigma}^2 = \text{MS[E]} = \frac{(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})}{26} \approx 102.630.$$

6.3 Sampling distributions

Consider our multiple linear regression model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$. We now investigate the **sampling distribution** of the least-squares estimator $\widehat{\boldsymbol{\beta}}$. This will help us construct confidence intervals for individual elements of $\boldsymbol{\beta}$ and perform hypothesis tests which compare different regression models.

SAMPLING DISTRIBUTION OF $\widehat{\boldsymbol{\beta}}$: Recall that the least-squares estimator for $\boldsymbol{\beta}$ is given by

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}.$$

The mean of $\widehat{\boldsymbol{\beta}}$ is given by

$$E(\widehat{\boldsymbol{\beta}}) = E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}] = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{Y}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}.$$

Thus, $\widehat{\boldsymbol{\beta}}$ is an **unbiased estimator** of $\boldsymbol{\beta}$. The variance of $\widehat{\boldsymbol{\beta}}$ is given by

$$\begin{aligned} V(\widehat{\boldsymbol{\beta}}) &= V[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}] \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'V(\boldsymbol{Y})[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']' \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\sigma^2\boldsymbol{I}_n\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}. \end{aligned}$$

Now $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ is a $(p+1)\times(p+1)$ matrix. Thus, $V(\widehat{\boldsymbol{\beta}})$ is a $(p+1)\times(p+1)$ matrix of the **true** variances and covariances; it has the following structure:

$$V(\widehat{\boldsymbol{\beta}}) = \begin{pmatrix} V(\widehat{\beta}_0) & \text{Cov}(\widehat{\beta}_0,\widehat{\beta}_1) & \text{Cov}(\widehat{\beta}_0,\widehat{\beta}_2) & \cdots & \text{Cov}(\widehat{\beta}_0,\widehat{\beta}_p) \\ & V(\widehat{\beta}_1) & \text{Cov}(\widehat{\beta}_1,\widehat{\beta}_2) & \cdots & \text{Cov}(\widehat{\beta}_1,\widehat{\beta}_p) \\ & & V(\widehat{\beta}_2) & \vdots & \text{Cov}(\widehat{\beta}_2,\widehat{\beta}_p) \\ & & & \ddots & \vdots \\ & & & & V(\widehat{\beta}_p) \end{pmatrix}_{(p+1)\times(p+1)}.$$

Notice that I only gave the upper triangle of the $V(\widehat{\boldsymbol{\beta}})$ matrix since it is **symmetric**. Of course, rarely will anyone ever tell us the value of σ^2, so $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ really is not that useful. For practical use, we must estimate σ^2 with its unbiased estimate MS[E]. Thus, the **estimated variance-covariance matrix** is given by $\widehat{V(\widehat{\boldsymbol{\beta}})} = \text{MS[E]}(\boldsymbol{X}'\boldsymbol{X})^{-1}$. This is a $(p+1)\times(p+1)$ matrix of the **estimated** variances and covariances; it has the following structure:

$$\widehat{V(\widehat{\boldsymbol{\beta}})} = \begin{pmatrix} \widehat{V(\widehat{\beta}_0)} & \widehat{\text{Cov}(\widehat{\beta}_0,\widehat{\beta}_1)} & \widehat{\text{Cov}(\widehat{\beta}_0,\widehat{\beta}_2)} & \cdots & \widehat{\text{Cov}(\widehat{\beta}_0,\widehat{\beta}_p)} \\ & \widehat{V(\widehat{\beta}_1)} & \widehat{\text{Cov}(\widehat{\beta}_1,\widehat{\beta}_2)} & \cdots & \widehat{\text{Cov}(\widehat{\beta}_1,\widehat{\beta}_p)} \\ & & \widehat{V(\widehat{\beta}_2)} & \vdots & \widehat{\text{Cov}(\widehat{\beta}_2,\widehat{\beta}_p)} \\ & & & \ddots & \vdots \\ & & & & \widehat{V(\widehat{\beta}_p)} \end{pmatrix}_{(p+1)\times(p+1)}$$

Usually, computing packages will give us this estimated variance-covariance matrix. This matrix will be very helpful when we want to get confidence intervals for $E(Y|\boldsymbol{x}_0)$ or prediction intervals for a new Y_0 (coming up!).

Example 6.3 (`cheese.sas`). With the cheese data in Example 6.1, consider the full model $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$, for $i = 1, 2, ..., 30$, or in matrix notation $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. The estimated variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$ is given by

$$\widehat{V(\widehat{\boldsymbol{\beta}})} = \text{MS[E]}(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{pmatrix} 398.480 & -77.977 & 8.960 & -7.333 \\ -77.977 & 19.889 & -2.089 & -13.148 \\ 8.960 & -2.089 & 1.558 & -4.670 \\ -7.333 & -13.148 & -4.670 & 74.461 \end{pmatrix}.$$

NORMALITY: Since $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ is just a linear combination of $Y_1, Y_2, ..., Y_n$ (all of which are normally distributed under our model assumptions), it follows that $\widehat{\boldsymbol{\beta}}$ is normally distributed as well. More precisely, it follows a $(p+1)$-dimensional **multivariate normal** distribution. Summarising,

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}_{p+1}[\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}].$$

IMPLICATIONS: In our multiple linear regression setting, we would like to do many of the things that we did in the simple-linear case; e.g., obtain confidence intervals for regression coefficients, prediction intervals, etc. The following facts provide the basis for many of these objectives and are direct consequences of our previous discussion:

(1) $E(\widehat{\beta}_j) = \beta_j$, for $j = 0, 1, ..., p$; that is, the least-squares estimators are **unbiased**.

(2) $V(\widehat{\beta}_j) = s_{jj}\sigma^2$, where
$$s_{jj} = (\boldsymbol{X}'\boldsymbol{X})^{-1}_{j,j},$$
for $j = 0, 1, ..., p$. The value $(\boldsymbol{X}'\boldsymbol{X})^{-1}_{j,j}$ represents the jth **diagonal element** of the $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ matrix. An estimate of $V(\widehat{\beta}_j)$ is given by $\widehat{V(\widehat{\beta}_j)} = s_{jj}\text{MS[E]}$.

(3) $\text{Cov}(\widehat{\beta}_j, \widehat{\beta}_{j'}) = s_{jj'}\sigma^2$, where
$$s_{jj'} = (\boldsymbol{X}'\boldsymbol{X})^{-1}_{j,j'},$$
for $j \neq j'$. The value $(\boldsymbol{X}'\boldsymbol{X})^{-1}_{j,j'}$ is the entry in the jth row and j'th column of the $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ matrix. An estimate of $\text{Cov}(\widehat{\beta}_j, \widehat{\beta}_{j'})$ is given by $\widehat{\text{Cov}(\widehat{\beta}_j, \widehat{\beta}_{j'})} = s_{jj'}\text{MS[E]}$.

(4) Marginally, $\widehat{\beta}_j \sim \mathcal{N}(\beta_j, s_{jj}\sigma^2)$, for $j = 0, 1, ..., p$.

SAMPLING DISTRIBUTION OF MS[E]: Under our model assumptions, the random variable

$$\frac{\text{SS[E]}}{\sigma^2} = \frac{(n-p-1)\text{MS[E]}}{\sigma^2} \sim \chi^2_{n-p-1}.$$

This fact is also useful in deriving many of the inferential procedures that we will discuss.

6.4 Inference for parameters in multiple regression

INFERENCE FOR INDIVIDUAL REGRESSION PARAMETERS: Consider our multiple linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i,$$

for $i = 1, 2, ..., n$, where $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$. Writing a confidence interval or performing a hypothesis test for single regression parameter β_j can help us assess the importance of using the independent variable x_j in the full model.

CONFIDENCE INTERVALS: Since $\widehat{\beta}_j \sim \mathcal{N}(\beta_j, s_{jj}\sigma^2)$, where $s_{jj} = (\boldsymbol{X}'\boldsymbol{X})^{-1}_{j,j}$, for $j = 0, 1, 2, ..., p$, it follows that

$$Z = \frac{\widehat{\beta}_j - \beta_j}{\sqrt{s_{jj}\sigma^2}} \sim \mathcal{N}(0, 1)$$

and that

$$t = \frac{\widehat{\beta}_j - \beta_j}{\sqrt{s_{jj}\text{MS[E]}}} \sim t_{n-p-1},$$

Confidence intervals and hypothesis tests are based on this pivotal quantity. Specifically, a $100(1-\alpha)$ percent confidence interval for β_i is given by

$$\widehat{\beta}_j \pm t_{n-p-1,\alpha/2}\sqrt{s_{jj}\text{MS[E]}}.$$

HYPOTHESIS TESTS: To test $H_0 : \beta_j = \beta_{j,0}$ versus a one or two-sided alternative, we use the test statistic

$$t = \frac{\widehat{\beta}_j - \beta_{j,0}}{\sqrt{s_{jj}\text{MS[E]}}}.$$

The rejection region is located in the appropriate tail(s) on the t_{n-p-1} reference distribution. P values are the appropriate areas under the t_{n-p-1} distribution.

NOTE: To assess whether or not x_j is useful in describing Y, *with the inclusion of the other predictor variables in the model*, we can test

$$H_0 : \beta_j = 0$$
$$\text{versus}$$
$$H_1 : \beta_j \neq 0.$$

It is important to recognise that tests of this form are "conditional" on there being the other predictors present in the model.

Example 6.4 (`cheese.sas`). With the cheese data in Example 6.1, consider the full model $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$, for $i = 1, 2, ..., 30$. To assess the importance of the hydrogen sulfide concentration and its influence on taste, we can test $H_0 : \beta_2 = 0$ versus $H_1 : \beta_2 \neq 0$. With $\widehat{\beta}_2 = 3.912$, $s_{22} = 0.015$, and $\text{MS[E]} = 102.630$, our test statistic becomes

$$t = \frac{\widehat{\beta}_2 - 0}{\sqrt{s_{22}\text{MS[E]}}} = \frac{3.912}{\sqrt{0.015 \times 102.630}} = 3.134,$$

which is larger than $t_{26,0.025} = 2.056$. Thus, at the $\alpha = 0.05$ significance level, we have significant evidence that the hydrogen sulfide concentration, *after adjusting for the effects of acetic and lactic concentrations*, is important in describing taste.

SIMULTANEOUS CONFIDENCE REGIONS FOR $\boldsymbol{\beta}$: The goal may be to find a region that contains $\boldsymbol{\beta}$ with probability $1 - \alpha$. This is called a $100(1 - \alpha)$ **percent confidence region** for $\boldsymbol{\beta}$. Consider two different regions:

1. An exact elliptical region
$$\left\{ \boldsymbol{\beta} : \frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{(p + 1)\text{MS[E]}} \leq F_{p,n-p-1,\alpha} \right\}$$

2. A rectangular region (equivalent to the intersection of $p + 1$ intervals) using a Bonferroni-type correction; more precisely, $\widehat{\beta}_i \pm t_{n-p-1,\alpha/2(p+1)}\sqrt{s_{ii}\text{MS[E]}}$, for $i = 0, 1, 2, ..., p$. This collection of intervals constitutes a joint confidence region whose probability of joint coverage is at least $1 - \alpha$.

6.5 Simple linear regression in matrix notation

The simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, for $i = 1, 2, ..., n$, is just a special case of the multiple linear regression model when $p = 1$. Thus, if we define

$$\boldsymbol{Y}_{n\times 1} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \boldsymbol{X}_{n\times 2} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \boldsymbol{\beta}_{2\times 1} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix},$$

and

$$\boldsymbol{\epsilon}_{n\times 1} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Then, the model $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, for $i = 1, 2, ..., n$, where $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$, can also be equivalently expressed as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$. Now, straightforward calculations (verify!) show that

$$\boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix} \quad \text{and} \quad (\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{pmatrix} \frac{1}{n} + \frac{\overline{x}^2}{S_{xx}} & -\frac{\overline{x}}{S_{xx}} \\ -\frac{\overline{x}}{S_{xx}} & \frac{1}{S_{xx}}. \end{pmatrix}$$

Thus, it can also be shown (verify!) that

$$\boldsymbol{X}'\boldsymbol{Y} = \begin{pmatrix} \sum_i Y_i \\ \sum_i x_i Y_i \end{pmatrix} \quad \text{and} \quad \widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \begin{pmatrix} \overline{Y} - \widehat{\beta}_1 \overline{x} \\ \frac{S_{xy}}{S_{xx}} \end{pmatrix} = \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{pmatrix}.$$

NOTE: Variances and covariances for $\widehat{\beta}_0$ and $\widehat{\beta}_1$ found in Chapter 4 are identical to those given by $V(\widehat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$. We can also compute an estimate of this variance-covariance matrix by using $\widehat{V(\widehat{\boldsymbol{\beta}})} = \text{MS[E]}(\boldsymbol{X}'\boldsymbol{X})^{-1}$. The elements of this matrix are useful in constructing confidence intervals and performing hypothesis tests for β_0 and β_1.

REMARK: Everything we have shown (and will show) for the multiple regression setting also holds in the simple-linear regression setting (i.e., when $p = 1$).

6.6 The hat matrix and geometric considerations

We now discuss some geometric considerations involved with fitting linear models. Although this discussion isn't imperative to your being able to fit models to data (e.g., run SAS), it can greatly enhance your appreciation for the theory that underpins it.

TERMINOLOGY: We call $\boldsymbol{M} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ the **hat matrix**. Observe that

$$\boldsymbol{M}\boldsymbol{Y} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{Y}},$$

so that, in a sense, the hat matrix $\boldsymbol{M}$ "puts the hat on" $\boldsymbol{Y}$; i.e., $\boldsymbol{M}$ turns the observed responses in $\boldsymbol{Y}$ into the fitted values in $\widehat{\boldsymbol{Y}}$. Here are some facts regarding the hat matrix:

- The hat matrix $\boldsymbol{M}$ is symmetric and idempotent; i.e., $\boldsymbol{M}' = \boldsymbol{M}$ and $\boldsymbol{M}^2 = \boldsymbol{M}$, and $\boldsymbol{M}\boldsymbol{X} = \boldsymbol{X}$.
- $\mathcal{C}(\boldsymbol{M}) = \mathcal{C}(\boldsymbol{X})$ and $r(\boldsymbol{M}) = r(\boldsymbol{X})$.
- Since $\boldsymbol{M}\boldsymbol{Y} = \widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} \in \mathcal{C}(\boldsymbol{X})$, geometrically, $\boldsymbol{M}$ projects $\boldsymbol{Y}$ onto $\mathcal{C}(\boldsymbol{X})$. Furthermore, since $\widehat{\boldsymbol{\beta}}$ is the value that minimises $\text{SSE}(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = ||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||^2$, which is simply the squared distance between $\boldsymbol{Y}$ and $\boldsymbol{X}\boldsymbol{\beta}$, we sometimes call $\boldsymbol{M}$ a **perpendicular projection matrix**.
- Let h_{ij} denote the (i,j)th element of $\boldsymbol{M}$. The diagonal elements of $\boldsymbol{M}$; i.e.,

  $$h_{ii} = \boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i,$$

  where $\boldsymbol{x}_i'$ denotes the ith row of the design matrix $\boldsymbol{X}$, are called the **leverages**. As we have already seen, leverages can be important in detecting outliers.

ANOTHER IMPORTANT MATRIX: The matrix $\boldsymbol{I} - \boldsymbol{M}$ is also important. Like $\boldsymbol{M}$, the matrix $\boldsymbol{I} - \boldsymbol{M}$ is also symmetric and idempotent (verify!). We know what $\boldsymbol{M}$ "does" to $\boldsymbol{Y}$. What does $\boldsymbol{I} - \boldsymbol{M}$ "do" to $\boldsymbol{Y}$? Note that

$$(\boldsymbol{I} - \boldsymbol{M})\boldsymbol{Y} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{e},$$

so the matrix $\boldsymbol{I}-\boldsymbol{M}$ turns $\boldsymbol{Y}$ into $\boldsymbol{e}$, the vector of least-squares residuals. Geometrically, the matrix $\boldsymbol{I}-\boldsymbol{M}$ projects $\boldsymbol{Y}$ into a space called the **error space**. This space is denoted by $\mathcal{C}(\boldsymbol{I}-\boldsymbol{M})$.

ORTHOGONALITY: It is easy to show that $(\boldsymbol{I}-\boldsymbol{M})\boldsymbol{X}=\boldsymbol{0}$. This results from the fact that $\mathcal{C}(\boldsymbol{X})$ and $\mathcal{C}(\boldsymbol{I}-\boldsymbol{M})$ are **orthogonal** spaces; i.e., any vector in $\mathcal{C}(\boldsymbol{X})$ is orthogonal to any vector in $\mathcal{C}(\boldsymbol{I}-\boldsymbol{M})$. Also, note that

$$\begin{aligned}\boldsymbol{Y} &= \boldsymbol{M}\boldsymbol{Y}+(\boldsymbol{I}-\boldsymbol{M})\boldsymbol{Y}\\ &= \widehat{\boldsymbol{Y}}+(\boldsymbol{Y}-\widehat{\boldsymbol{Y}})\\ &= \widehat{\boldsymbol{Y}}+\boldsymbol{e}.\end{aligned}$$

RESULT: Any data vector $\boldsymbol{Y}$ can be uniquely **decomposed** into two parts; namely, $\widehat{\boldsymbol{Y}}\in\mathcal{C}(\boldsymbol{X})$ and $\boldsymbol{e}\in\mathcal{C}(\boldsymbol{I}-\boldsymbol{M})$. Since $\widehat{\boldsymbol{Y}}$ and $\boldsymbol{e}$ are orthogonal, it follows that $\widehat{\boldsymbol{Y}}'\boldsymbol{e}=0$.

6.7 The analysis of variance for multiple linear regression

Just as we did in the simple linear case, we aim to summarise, in tabular form, the amount of variability due to the **regression** (model) and the amount of variability due to the **error**.

RESULT: Consider the quadratic form $\boldsymbol{Y}'\boldsymbol{Y}$, and note that we can write it as

$$\begin{aligned}\boldsymbol{Y}'\boldsymbol{Y} &= \boldsymbol{Y}'(\boldsymbol{M}+\boldsymbol{I}-\boldsymbol{M})\boldsymbol{Y}\\ &= \boldsymbol{Y}'\boldsymbol{M}\boldsymbol{Y}+\boldsymbol{Y}'(\boldsymbol{I}-\boldsymbol{M})\boldsymbol{Y}\\ &= \boldsymbol{Y}'\boldsymbol{M}\boldsymbol{M}\boldsymbol{Y}+\boldsymbol{Y}'(\boldsymbol{I}-\boldsymbol{M})(\boldsymbol{I}-\boldsymbol{M})\boldsymbol{Y}\\ &= \widehat{\boldsymbol{Y}}'\widehat{\boldsymbol{Y}}+\boldsymbol{e}'\boldsymbol{e}.\end{aligned}$$

Algebraically, this means that

$$\sum_{i=1}^{n}Y_i^2=\sum_{i=1}^{n}\widehat{Y}_i^2+\sum_{i=1}^{n}(Y_i-\widehat{Y}_i)^2.$$

In words, the **uncorrected** total sum of squares $\sum_{i=1}^{n}Y_i^2$ equals the **uncorrected** regression (model) sum of squares $\sum_{i=1}^{n}\widehat{Y}_i^2$ plus the error sum of squares $\sum_{i=1}^{n}(Y_i-\widehat{Y}_i)^2$.

CORRECTED SUMS OF SQUARES: Since interest often lies in those regression coefficients attached to predictor variables; i.e., $\beta_1, \beta_2, ..., \beta_p$, it is common to "remove" the effects of fitting the intercept term β_0, the overall mean of Y (ignoring the predictor variables). This removal is accomplished by subtracting $n\overline{Y}^2$ from both sides of the last equation to get

$$\sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2 = \sum_{i=1}^{n} \widehat{Y}_i^2 - n\overline{Y}^2 + \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2.$$

or, equivalently,

$$\underbrace{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}_{\text{SS[TOT]}} = \underbrace{\sum_{i=1}^{n} (\widehat{Y}_i - \overline{Y})^2}_{\text{SS[R]}} + \underbrace{\sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2}_{\text{SS[E]}}.$$

The terms SS[TOT] and SS[R] here are called the **corrected** total and **corrected** regression sums of squares, respectively, for the reasons just described. *This is the partitioning of the sums of squares that SAS uses!!*

SUMS OF SQUARES AS QUADRATIC FORMS: To enhance the understanding of the partitioning of sums of squares, we express the SS[TOT] = SS[R] + SS[E] partition in terms of **quadratic forms**. The basic (uncorrected) partition is given by

$$\boldsymbol{Y}'\boldsymbol{Y} = \boldsymbol{Y}'\boldsymbol{M}\boldsymbol{Y} + \boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{M})\boldsymbol{Y}$$

In words, the **uncorrected** total sum of squares $\boldsymbol{Y}'\boldsymbol{Y}$ equals the **uncorrected** regression (model) sum of squares $\boldsymbol{Y}'\boldsymbol{M}\boldsymbol{Y}$ plus the error sum of squares $\boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{M})\boldsymbol{Y}$.

CORRECTED SUMS OF SQUARES: As before, we can "remove" the effects of fitting the intercept term β_0. This removal is accomplished by subtracting $n\overline{Y}^2 = \boldsymbol{Y}'n^{-1}\boldsymbol{J}\boldsymbol{Y}$ from both sides of the last equation to get

$$\boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{Y}'n^{-1}\boldsymbol{J}\boldsymbol{Y} = \boldsymbol{Y}'\boldsymbol{M}\boldsymbol{Y} - \boldsymbol{Y}'n^{-1}\boldsymbol{J}\boldsymbol{Y} + \boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{M})\boldsymbol{Y}$$

or, equivalently,

$$\underbrace{\boldsymbol{Y}'(\boldsymbol{I} - n^{-1}\boldsymbol{J})\boldsymbol{Y}}_{\text{SS[TOT]}} = \underbrace{\boldsymbol{Y}'(\boldsymbol{M} - n^{-1}\boldsymbol{J})\boldsymbol{Y}}_{\text{SS[R]}} + \underbrace{\boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{M})\boldsymbol{Y}}_{\text{SS[E]}}.$$

We combine all of this information into an ANOVA table.

Table 6.16: *The general form of an analysis of variance table for multiple linear regression. All sums of squares are assumed to be of the corrected type.*

Source	df	SS	MS	F
Regression	p	SS[R]	MS[R]	$F = \frac{\text{MS[R]}}{\text{MS[E]}}$
Error	$n-p-1$	SS[E]	MS[E]	
Total	$n-1$	SS[TOT]		

INTERPRETING THE DEGREES OF FREEDOM: The ANOVA table here has the same structure as the ANOVA table for straight-line model. The only difference is the degrees of freedom.

- The **corrected** SS[TOT] has $n-1$ degrees of freedom. Again, as before, we can view this as a statistic that has "lost" a degree of freedom for having to estimate the overall mean of Y, β_0, with $\overline{Y}$. Also, one will note that $r(\boldsymbol{I} - n^{-1}\boldsymbol{J}) = n-1$.
- The **corrected** SS[R] has p degrees of freedom, the number of predictor variables in the model. Since $r(\boldsymbol{M}) = r(\boldsymbol{X}) = p+1$, one will note that $r(\boldsymbol{M} - n^{-1}\boldsymbol{J}) = p$.
- SS[E] has $n-p-1$ degrees of freedom, obtained by subtraction. Also, one will note that $r(\boldsymbol{I} - \boldsymbol{M}) = n-p-1$.

USING THE F STATISTIC: The F statistic in the ANOVA table above is used to test whether or not **at least one** of the independent variables $x_1, x_2, ..., x_p$ adds to the model; i.e., it is used to test

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

versus

$$H_1 : \text{at least one of the } \beta_i \text{ is nonzero}$$

in the full model $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$. Note that if H_0 is rejected, we do not know which one or how many of the β_i's are nonzero; we only that at least one is.

MATHEMATICAL JUSTIFICATION: When H_0 is true, it follows that $\text{SS[R]}/\sigma^2 \sim \chi^2_p$, $\text{SS[E]}/\sigma^2 \sim \chi^2_{n-p-1}$, and that SS[R] and SS[E] are independent. These facts would be proven in a more advanced linear models course. Thus, under H_0,

$$F = \frac{\text{SS[R]}/p\sigma^2}{\text{SS[E]}/(n-p-1)\sigma^2} = \frac{\text{MS[R]}}{\text{MS[E]}} \sim F_{p,n-p-1}.$$

As in the straight-line case, H_0 is rejected whenever the F statistic gets too large; that is, this is a **one-sided**, **upper-tail** test with rejection region $RR = \{F : F > F_{p,n-p-1,\alpha}\}$ where $F_{p,n-p-1,\alpha}$ denotes the $1-\alpha$ quantile of the F distribution with p (numerator) and $n-p-1$ (denominator) degrees of freedom. P values are computed as areas to the right of F on the $F_{p,n-p-1}$ distribution.

EXPECTED MEAN SQUARES: We have already shown that $E(\text{MS[E]}) = \sigma^2$. This holds whether or not H_0 is true. We now investigate $E(\text{MS[R]})$. Since $E(\boldsymbol{Y}) = \boldsymbol{X\beta}$ and $V(\boldsymbol{Y}) = \sigma^2\boldsymbol{I}$, we can compute

$$\begin{aligned} E(\text{SS[R]}) &= E[\boldsymbol{Y}'(\boldsymbol{M} - n^{-1}\boldsymbol{J})\boldsymbol{Y}] \\ &= \underbrace{(\boldsymbol{X\beta})'(\boldsymbol{M} - n^{-1}\boldsymbol{J})\boldsymbol{X\beta}}_{=\ 0,\ \text{under } H_0} + \text{tr}[(\boldsymbol{M} - n^{-1}\boldsymbol{J})\sigma^2\boldsymbol{I}]. \end{aligned}$$

When $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$ is true, the first term equals zero (verify!). To handle the second term, note that

$$\begin{aligned} \text{tr}[(\boldsymbol{M} - n^{-1}\boldsymbol{J})\sigma^2\boldsymbol{I}] &= \sigma^2\text{tr}[(\boldsymbol{M} - n^{-1}\boldsymbol{J})] \\ &= \sigma^2[\text{tr}(\boldsymbol{M}) - \text{tr}(n^{-1}\boldsymbol{J})] \\ &= \sigma^2[r(\boldsymbol{M}) - \text{tr}(n^{-1}\boldsymbol{J})] \\ &= \sigma^2[r(\boldsymbol{M}) - 1] \\ &= \sigma^2[r(\boldsymbol{X}) - 1] \\ &= \sigma^2[p + 1 - 1] \\ &= p\sigma^2. \end{aligned}$$

We have shown that, when H_0 is true, $E(\text{SS[R]}) = p\sigma^2$. Thus, only when H_0 is **true**,

$$E(\text{MS[R]}) = E\left(\frac{\text{SS[R]}}{p}\right) = \sigma^2.$$

REALISATION: When H_0 is true, both MS[R] and MS[E] are estimating the same quantity, and, thus, the F statistic should be close to one. When H_0 is not true, the term $(\boldsymbol{X\beta})'(\boldsymbol{M} - n^{-1}\boldsymbol{J})\boldsymbol{X\beta} > 0$, and, hence, MS[R] is estimating something (perhaps much) larger than σ^2. In this case, we would expect F to be (perhaps much) larger than one. This gives an intuitive explanation of why F should be large when H_0 is not true. The term $(\boldsymbol{X\beta})'(\boldsymbol{M} - n^{-1}\boldsymbol{J})\boldsymbol{X\beta}$ is sometimes called a **non-centrality parameter**.

THE COEFFICIENT OF DETERMINATION: Since SS[TOT] = SS[R] + SS[E], it follows that the proportion of the total variation in the data explained by the model is

$$R^2 = \frac{\text{SS[R]}}{\text{SS[TOT]}}.$$

As in the straight-line case, the statistic R^2 is called the **coefficient of determination**. It has the analogous interpretation in multiple linear regression settings as it did in the straight-line case.

Example 6.5 (`cheese.sas`). With the cheese data in Example 6.1, consider the full model $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$, for $i = 1, 2, ..., 30$. The ANOVA table for these data, obtained using SAS, is shown below.

Source	DF	SS	MS	F	Pr > F
Model	3	4994.509	1664.836	16.22	<0.0001
Error	26	2668.378	102.629		
Corrected Total	29	7662.887			

The F statistic is used to test

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

versus

$$H_1 : \text{not } H_0.$$

Since the P value for the test is so small, we would conclude that at least one of x_1, x_2, or x_3 is important in describing taste. The coefficient of determination is $R^2 \approx 0.652$. Thus, about 65 percent of the variability in the taste data is explained by x_1, x_2, and x_3, assuming that $E(Y|x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ is the true regression function.

6.8 Sequential and partial sums of squares

There are two types of sums of squares that can be helpful in hypothesis testing with regression models; namely, **sequential** sums of squares and **partial** sums of squares. We now examine them both.

SEQUENTIAL SUMS OF SQUARES: **Sequential sums of squares** build up the sums of squares due to the regression (model). Their values depend on the particular **order** in which the x's are "added to the model." Consider the breakdown of SS[R] in Table 6.17.

Table 6.17: *Sequential sums of squares and their partitioning of the corrected* SS[R].

Sequential ANOVA	df	SS
Regression on x_1 (after β_0)	1	$R(\beta_1\vert\beta_0)$
Regression on x_2 (after β_0 and x_1)	1	$R(\beta_2\vert\beta_0,\beta_1)$
Regression on x_3 (after β_0, x_1, and x_2)	1	$R(\beta_3\vert\beta_0,\beta_1,\beta_2)$
$\vdots$	$\vdots$	$\vdots$
Regression on x_p (after β_0, x_1, x_2, ..., x_{p-1})	1	$R(\beta_p\vert\beta_0,\beta_1,...,\beta_{p-1})$

FACT: The sequential sums of squares in Table 6.17 add up to the **corrected** SS[R] from the overall ANOVA; that is,

$$\text{SS[R]} = R(\beta_1|\beta_0) + R(\beta_2|\beta_0,\beta_1) + R(\beta_3|\beta_0,\beta_1,\beta_2) + \cdots + R(\beta_p|\beta_0,\beta_1,...,\beta_{p-1}).$$

REMARK: Sequential sums of squares help us assess the merit of adding an individual x to the model in a stepwise manner. There are cases where this way of thinking is particularly helpful. For example, consider the cubic model

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_i,$$

and suppose that we were interested in whether or not the quadratic and cubic terms were really needed (i.e., whether or not $\beta_2 = \beta_3 = 0$). We can address this question by examining the sizes of $R(\beta_2|\beta_0,\beta_1)$ and $R(\beta_3|\beta_0,\beta_1,\beta_2)$. If these are both large, then both x^2 and x^3 should be kept in the model. If not, then neither should be kept. If $R(\beta_2|\beta_0, \beta_1)$ is large, but $R(\beta_3|\beta_0, \beta_1, \beta_2)$ is not, then we might consider keeping x^2, but not x^3. We'll formalise this notion of "large" momentarily....

NON-UNIQUENESS: The order of the sequential ANOVA is not unique!! *If you change the order in which the variables are added to the model, the sequential sums of squares will also change* (of course, they will still sum to the uncorrected SS[R]). In SAS, the order depends upon the left-to-right specification in the `MODEL` statement. Sequential sums of squares are referred to as **Type I SS** in SAS.

Example 6.6 (`cheese.sas`). With the cheese data in Example 6.1, consider the full model

$$\texttt{TASTE} = \beta_0 + \beta_1\texttt{ACETIC} + \beta_2\texttt{H2S} + \beta_3\texttt{LACTIC} + \epsilon.$$

We now compute the sequential sums of squares for the cheese data. The model statement

```
model taste = acetic h2s lactic;
```

in `glm` produces the following sequential sums of squares breakdown:

Source	DF	Type I SS	MS	F Value	Pr>F
acetic	1	2314.142	2314.142	22.55	<.0001
h2s	1	2147.108	2147.108	20.92	0.0001
lactic	1	533.259	533.259	5.20	0.0311

Thus, $R(\beta_1|\beta_0) = 2314.142$, $R(\beta_2|\beta_0, \beta_1) = 2147.108$, and $R(\beta_3|\beta_0, \beta_1, \beta_2) = 533.259$. One will note that these sum to SS[R] = 4994.509 from the overall ANOVA. The F statistics test, **sequentially**, whether or not each variable should be added to the model. For example, $F = 22.55$ tests

$$\begin{aligned} H_0 &: \quad \texttt{TASTE} = \beta_0 + \epsilon \\ H_1 &: \quad \texttt{TASTE} = \beta_0 + \beta_1\texttt{ACETIC} + \epsilon. \end{aligned}$$

It looks like `ACETIC` should be added to the model that includes only the intercept β_0.

Next, $F = 20.92$ tests

$$\begin{aligned} H_0: &\quad \texttt{TASTE} = \beta_0 + \beta_1\texttt{ACETIC} + \epsilon \\ H_1: &\quad \texttt{TASTE} = \beta_0 + \beta_1\texttt{ACETIC} + \beta_2\texttt{H2S} + \epsilon. \end{aligned}$$

It looks like `H2S` should be added to the model that already includes `ACETIC` (and β_0). Finally, $F = 5.20$ tests

$$\begin{aligned} H_0: &\quad \texttt{TASTE} = \beta_0 + \beta_1\texttt{ACETIC} + \beta_2\texttt{H2S} + \epsilon \\ H_1: &\quad \texttt{TASTE} = \beta_0 + \beta_1\texttt{ACETIC} + \beta_2\texttt{H2S} + \beta_3\texttt{LACTIC} + \epsilon. \end{aligned}$$

Thus, `LACTIC` should be added to the model that includes `H2S` and `ACETIC` (and β_0).

A DIFFERENT ORDERING: Now, suppose that I had used the model statement

```
model taste = h2s lactic acetic;
```

for the model

$$\texttt{TASTE} = \beta_0 + \beta_1\texttt{H2S} + \beta_2\texttt{LACTIC} + \beta_3\texttt{ACETIC} + \epsilon$$

(note that I permuted the order of the independent variables in the model statement). This produces the following sequential sums of squares breakdown:

Source	DF	Type I SS	MS	F Value	Pr>F
h2s	1	4376.833	4376.833	42.65	<.0001
lactic	1	617.120	617.120	6.01	0.0212
acetic	1	0.555	0.555	0.01	0.9479

Using this ordering of the x's produces a different breakdown; here, $R(\beta_1|\beta_0) = 4376.833$, $R(\beta_2|\beta_0, \beta_1) = 617.120$, and $R(\beta_3|\beta_0, \beta_1, \beta_2) = 0.555$; however, one will note that the sequential sums of squares for this new ordering still sum to SS[R] = 4994.509 from the overall ANOVA! It is interesting to note that, for this ordering, we would not add `ACETIC` to a model that already includes `H2S` and `LACTIC`. The reader will remember that we added `ACETIC` under the last ordering; however, in that case, we were adding `ACETIC` to a model that only included β_0.

F STATISTICS BASED ON SEQUENTIAL SUMS OF SQUARES: The F statistics based on sequential sums of squares are used to test whether or not x_k should be added to a model that already includes $x_1, x_2, ..., x_{k-1}$ (and β_0), for $k = 1, 2, ..., p$; i.e., the statistic F_k tests

$$\begin{aligned} H_0: &\quad Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{k-1} x_{i(k-1)} + \epsilon_i \\ H_1: &\quad Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{k-1} x_{i(k-1)} + \beta_k x_{ik} + \epsilon_i \end{aligned}$$

and is formed by taking the ratio of $R(\beta_k|\beta_0, \beta_1, ..., \beta_{k-1})$ to the MS[E] from the overall ANOVA; i.e.,

$$F_k = \frac{R(\beta_k|\beta_0, \beta_1, ..., \beta_{k-1})}{\text{MS[E]}}.$$

When H_0 is true (i.e., the model which excludes x_k is more appropriate; i.e., $\beta_k = 0$), then $F_k \sim F_{1,n-p-1}$. Large values of F_k are evidence against H_0.

PARTIAL SUMS OF SQUARES: **Partial sums of squares** help us to assess the value of adding a predictor variable x to a model that already contains all other $p-1$ covariates (and β_0). That is, for $k = 1, 2, ..., p$,

$$\text{Partial SS for } x_k = R(\beta_k|\beta_0, \beta_1, ..., \beta_{k-1}, \beta_{k+1}, ..., \beta_p)$$

Unlike the sequential sums of squares, *partial sums of squares do not necessarily sum to* SS[R]. Partial sums of squares are referred to as **Type II SS** or **Type III SS** in SAS.

INTERESTING NOTE: When the columns of $\boldsymbol{X}$ are **orthogonal**, it turns out that the sequential and partial sums of squares will always be equal. In this situation, and only in this situation, the partial sums of squares *will* sum to SS[R]. This fact turns out to be very useful with ANOVA models (not so much with regression models).

PARTIAL F TESTS: In a sense, the partial sums of squares allow us to study the effect of putting a particular x_k into the model **last**; that is, they allow us to test, for any $k = 1, 2, ..., p$,

$$\begin{aligned} H_0: &\quad Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{k-1} x_{i(k-1)} + \beta_{k+1} x_{i(k+1)} + \cdots + \beta_p x_{ip} + \epsilon_i \\ H_1: &\quad Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i \quad \text{(the full model).} \end{aligned}$$

The F statistic used to test H_0 versus H_1 is given by

$$F_k = \frac{R(\beta_k|\beta_0, \beta_1, ..., \beta_{k-1}, \beta_{k+1}, ..., \beta_p)}{\text{MS[E]}}.$$

When H_0 is true (i.e., the model which excludes x_k is more appropriate; i.e., $\beta_k = 0$), then $F_k \sim F_{1,n-p-1}$. Large values of F_k are evidence against H_0.

REALISATION: You will recall from Section 6.4, we used the statistic

$$t = \frac{\widehat{\beta}_k}{\sqrt{s_{kk}\text{MS[E]}}}$$

to test $H_0 : \beta_k = 0$ versus $H_1 : \beta_k \neq 0$. This is essentially the same situation expressed in the last set of hypotheses! Our interpretation is the same here using partial sums of squares as it was using the t statistic; namely, we are assessing whether or not x_k is useful in describing Y, *with the inclusion of the other predictor variables in the model.*

MAIN POINT: Testing

$$\begin{aligned} H_0 : &\quad Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{k-1} x_{i(k-1)} + \beta_{k+1} x_{i(k+1)} + \cdots + \beta_p x_{ip} + \epsilon_i \\ H_1 : &\quad Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i \quad \text{(the full model).} \end{aligned}$$

using the F statistic based on the partial sums of squares and testing $H_0 : \beta_k = 0$ versus $H_1 : \beta_k \neq 0$ using the t statistic mentioned above are, in fact, the **same test**! Actually, the F statistic from the partial F test is the **square** the t statistic from the t test; that is, $F = t^2$.

REMARK: Some (but not all) prefer the more flexible t procedure in this situation. The partial F test is rather restrictive; it only tests $H_0 : \beta_k = 0$ versus $H_1 : \beta_k \neq 0$. The t procedure can handle one-sided alternatives and nonzero values under H_0.

ALGEBRAIC LINK: Equating F and t^2 above, we find that

$$\text{Partial SS for } x_k = R(\beta_k|\beta_0, \beta_1, ..., \beta_{k-1}, \beta_{k+1}, ..., \beta_p) = \frac{\widehat{\beta}_k^2}{s_{kk}},$$

where $s_{kk} = (\boldsymbol{X}'\boldsymbol{X})^{-1}_{k,k}$, for $k = 1, 2, ..., p$.

Example 6.7 (`cheese.sas`). With the cheese data in Example 6.1, consider the full model $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$, for $i = 1, 2, ..., 30$. In Example, 6.4, we tested whether or not the hydrogen sulfide concentration was important in describing taste by testing $H_0 : \beta_2 = 0$ versus $H_1 : \beta_2 \neq 0$. Recall, that we computed $t = 3.134$ and rejected H_0. Essentially, the test of $H_0 : \beta_2 = 0$ versus $H_1 : \beta_2 \neq 0$ is equivalent to testing the smaller model that excludes x_2 versus the larger model which includes it; i.e.,

$$\begin{aligned} H_0 : &\quad Y_i = \beta_0 + \beta_1 x_{i1} + \beta_3 x_{i3} + \epsilon_i \\ H_1 : &\quad Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i \end{aligned}$$

so we can also use a partial F test to test this same hypothesis. From SAS, I computed the partial **(Type III)** sums of squares table:

Source	DF	Type III SS	MS	F Value	Pr>F
acetic	1	0.555	0.555	0.01	0.9419
h2s	1	1007.691	1007.691	9.82	0.0042
lactic	1	533.259	533.259	5.20	0.0311

The F statistic, based on the partial sums of squares, is given by

$$F_2 = \frac{R(\beta_2|\beta_0, \beta_1, \beta_3)}{\text{MS[E]}} = \frac{1007.691}{102.630} = 9.82,$$

which is highly significant. Thus, at the $\alpha = 0.05$ level, we have significant evidence that the hydrogen sulfide concentration, *after adjusting for the effects of acetic and lactic concentrations*, is important in describing taste. This is the same conclusion we reached with the t test (as it should be). In fact, you will note that $t^2 = (3.134)^2 \approx 9.82 = F$ (up to rounding error). Finally, note that

$$\frac{\widehat{\beta}_2^2}{s_{22}} = \frac{(3.91178179)^2}{0.015185245} \approx 1007.691 = R(\beta_2|\beta_0, \beta_1, \beta_3) = \text{Partial SS for } x_2.$$

Note that I retained many decimal places in this calculation so I didn't incur too large of a rounding error. Recall that SAS gives $\widehat{\beta}_2$ and s_{22} to many decimal places.

6.9 Confidence intervals for $E(Y|\boldsymbol{x})$ in multiple regression settings

In Chapter 4, with our simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, for $i = 1, 2, ..., n$, where $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$, we learned how to obtain confidence intervals for the mean response $E(Y|x_0) = \beta_0 + \beta_1 x_0$ and prediction intervals for a new value Y_0. Extending these ideas to a multiple-regression setting is straightforward.

SETTING AND GOALS: Consider our multiple regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i,$$

for $i = 1, 2, ..., n$, where $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$, or, equivalently,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. Our goal is to obtain a $100(1-\alpha)$ percent confidence interval for

$$E(Y|\boldsymbol{x}_0) \equiv \mu(\boldsymbol{x}_0) = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \cdots + \beta_p x_{0p},$$

the **mean** of the response Y when $\boldsymbol{x}' = \boldsymbol{x}_0' \equiv (x_{01}, x_{02}, ..., x_{0p})$.

AN OBVIOUS POINT ESTIMATOR: Define $\boldsymbol{a}' = (1, \boldsymbol{x}_0') = (1, x_{01}, x_{02}, ..., x_{0p})$. Then, we can write

$$E(Y|\boldsymbol{x}_0) = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \cdots + \beta_p x_{0p} = \boldsymbol{a}'\boldsymbol{\beta} \equiv \theta.$$

Note that θ is a scalar since $\boldsymbol{a}'$ is $1 \times (p+1)$ and $\boldsymbol{\beta}$ is $(p+1) \times 1$. To estimate $\boldsymbol{a}'\boldsymbol{\beta}$, we will use $\widehat{\theta} = \boldsymbol{a}'\widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}}$ is the least-squares estimator of $\boldsymbol{\beta}$. To find the confidence interval, we need to obtain the sampling distribution of $\widehat{\theta} = \boldsymbol{a}'\widehat{\boldsymbol{\beta}}$. The mean of $\widehat{\theta}$ is given by

$$E(\widehat{\theta}) = E(\boldsymbol{a}'\widehat{\boldsymbol{\beta}}) = \boldsymbol{a}'E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{a}'\boldsymbol{\beta} = \theta;$$

thus, $\widehat{\theta}$ is an **unbiased estimator** of θ. The variance of $\widehat{\theta}$ is given by

$$\begin{aligned} V(\widehat{\theta}) &= V(\boldsymbol{a}'\widehat{\boldsymbol{\beta}}) \\ &= \boldsymbol{a}'V(\widehat{\boldsymbol{\beta}})\boldsymbol{a} \\ &= \boldsymbol{a}'\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{a} \\ &= \sigma^2\boldsymbol{a}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{a}. \end{aligned}$$

Furthermore, $\widehat{\theta} = \boldsymbol{a}'\widehat{\boldsymbol{\beta}}$ is just a **linear combination** of $\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_p$, all of which are normally distributed. Thus, $\widehat{\theta} = \boldsymbol{a}'\widehat{\boldsymbol{\beta}}$ is normally distributed as well! Summarising, we have that

$$\widehat{\theta} = \boldsymbol{a}'\widehat{\boldsymbol{\beta}} \sim \mathcal{N}\big\{\theta, \sigma^2\boldsymbol{a}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{a}\big\},$$

where $\theta = \boldsymbol{a}'\boldsymbol{\beta}$. Standardising, it follows that

$$Z = \frac{\widehat{\theta} - \theta}{\sqrt{\sigma^2\boldsymbol{a}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{a}}} \sim \mathcal{N}(0,1)$$

and that

$$t = \frac{\widehat{\theta} - \theta}{\sqrt{\text{MS[E]}\boldsymbol{a}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{a}}} \sim t_{n-p-1}.$$

Thus, we can use t as a pivotal quantity to derive confidence intervals and hypothesis tests involving $\theta = \boldsymbol{a}'\boldsymbol{\beta}$. A $100(1-\alpha)$ percent confidence interval for $\theta = \boldsymbol{a}'\boldsymbol{\beta} = E(Y|\boldsymbol{x}_0)$ is given by

$$\widehat{\theta} \pm t_{n-p-1,\alpha/2}\sqrt{\text{MS[E]}\boldsymbol{a}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{a}},$$

and a level α test of $H_0 : \theta = \theta_0$ versus a one or two-sided alternative uses

$$t = \frac{\widehat{\theta} - \theta_0}{\sqrt{\text{MS[E]}\boldsymbol{a}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{a}}}$$

as a test statistic. The rejection region is located in the appropriate tails of the t_{n-p-1} reference distribution. P-values come from this distribution as well.

Example 6.8 (`cheese.sas`). With our cheese data from Example 6.1, we want to find a 95 percent confidence interval for $E(Y|\boldsymbol{x}_0) = \beta_0 + 5\beta_1 + 6\beta_2 + \beta_3$. Here, $\boldsymbol{x}_0 = (5,6,1)$, so that $\boldsymbol{a}' = (1,5,6,1)$. We have

$$\widehat{E(Y|\boldsymbol{x}_0)} = \widehat{\theta} = \boldsymbol{a}'\widehat{\boldsymbol{\beta}} = (1,5,6,1)\begin{pmatrix} -28.877 \\ 0.328 \\ 3.912 \\ 19.670 \end{pmatrix} \approx 15.905$$

and

$$\boldsymbol{a}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{a} = (1,5,6,1)\begin{pmatrix} 3.795 & -0.760 & 0.087 & -0.071 \\ -0.760 & 0.194 & -0.020 & -0.128 \\ 0.087 & -0.020 & 0.015 & -0.046 \\ -0.071 & -0.128 & -0.046 & 0.726 \end{pmatrix}\begin{pmatrix} 1 \\ 5 \\ 6 \\ 1 \end{pmatrix} \approx 0.181.$$

Thus, a 95 percent confidence interval for $E(Y|\boldsymbol{x}_0) = \theta = \beta_0 + 5\beta_1 + 6\beta_2 + \beta_3$ is given by

$$15.905 \pm 2.056\sqrt{102.630 \times 0.181},$$

or $(7.310, 24.497)$. Thus, when $x_{01} = 5$, $x_{02} = 6$, and $x_{03} = 1$, we are 95 percent confident that the **mean** taste rating $E(Y|\boldsymbol{x}_0)$ is between 7.310 and 24.497.

6.10 Prediction intervals for a future value of Y in multiple linear regression

Analogous to our derivation for prediction intervals in the simple linear regression setting, it is not overly difficult to show that, with $\widehat{\theta} = \boldsymbol{a}'\widehat{\boldsymbol{\beta}}$,

$$\widehat{\theta} \pm t_{n-p-1,\alpha/2}\sqrt{\text{MS[E]}\,\{1 + \boldsymbol{a}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{a}\}},$$

is a $100(1-\alpha)$ percent **prediction interval** for Y_0 when $\boldsymbol{x}' = \boldsymbol{x}_0' \equiv (x_{01}, x_{02}, ..., x_{0p})$. Here, $\boldsymbol{a}_0 = (1, \boldsymbol{x}_0) = (1, x_{01}, x_{02}, ..., x_{0p})$, as before. Comparing this to the $100(1-\alpha)$ percent confidence interval for $E(Y|\boldsymbol{x}_0)$, we see, again, that there is an extra "1" in the estimated standard error. As in the straight-line case, this results from the extra variability which arises from having to **predict** the random quantity Y_0 instead of having to **estimate** the mean $E(Y|\boldsymbol{x}_0)$.

Example 6.9 (`cheese.sas`). With our cheese data from Example 6.1, we want to find a 95 percent prediction interval for a particular taste rating score Y when $\boldsymbol{x}_0 = (5, 6, 1)$. Here, $\boldsymbol{x}_0 = (5, 6, 1)$, so that $\boldsymbol{a}' = (1, 5, 6, 1)$. The prediction interval is given by

$$15.905 \pm 2.056\sqrt{102.630(1 + 0.181)},$$

or $(-6.624, 38.431)$. Thus, when $x_{01} = 5$, $x_{02} = 6$, and $x_{03} = 1$, we are 95 percent confident that the taste rating for a new value of Y will be between -6.624 and 38.431. One will immediately note that, again, this prediction interval is wider than the corresponding confidence interval for $E(Y|\boldsymbol{x}_0)$. Also, note that the lower limit for the prediction interval is negative, which may not even make sense in this application.

6.11 Testing full versus reduced models in multiple linear regression

SETTING: Consider our (full) multiple regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i,$$

for $i = 1, 2, ..., n$, where $\epsilon_i \sim \text{iid } \mathcal{N}(0, \sigma^2)$, or, equivalently,

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. Sometimes, it is of interest to determine whether or not a **smaller model** is adequate for the data. That is, can we "throw out" some of the independent variables, write a smaller model, and have that smaller model do "just as well" as the full model? In terms of the regression parameters themselves, we basically are asking, "are some of the parameters $\beta_0, \beta_1, \beta_2, ..., \beta_p$ essentially no different from zero?"

TERMINOLOGY: A well-known principle in science is the **Parsimony Principle**, which states, loosely speaking, that the simplest of two competing theories is to be preferred. Applying this principle to regression modelling leads us to choose models that are as simple as possible, but, yet, do an adequate job of describing the response. Seldom (i.e., never) will there be a model that is exactly correct, but, hopefully, we can find a model that is reasonable and does a good job summarising the true relationship between the response and the available predictor variables.

REMARK: Besides their ease of interpretation, smaller models confer statistical benefits, too. Remember that for each additional predictor variable we add to the model, there is an associated regression parameter which has to be estimated. For each additional regression parameter that we have to estimate, we lose a degree of freedom for error. Why is this important? Remember that MS[E], our estimator for the error variance σ^2 uses the degrees of freedom for error in its computation! Thus, the smaller this degree of freedom value is, the fewer observations we are using to estimate σ^2. With a poor estimate of σ^2, hypothesis tests, confidence intervals, and prediction intervals are likely to be poor as well.

Example 6.10 (`cheese.sas`). With our cheese data from Example 6.1, the **full** model is given by

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i.$$

Suppose that we consider the smaller model (i.e., the **reduced** model)

$$Y_i = \gamma_0 + \gamma_1 x_{i3} + \epsilon_i,$$

for $i = 1, 2, ..., 30$. In the reduced model, we are excluding the independent variables x_1 (`ACETIC`) and x_2 (`H2S`). Does the smaller model do just as well at describing the data as the full model? How can we **test** this? One will note that, in this example, we are essentially asking ourselves whether or not $H_0 : \beta_1 = \beta_2 = 0$ is supported by the data.

FULL-VERSUS-REDUCED MODEL TESTING SETUP: Consider testing

$$\boldsymbol{Y} = \boldsymbol{X}_0\boldsymbol{\gamma} + \boldsymbol{\epsilon} \quad \text{(a reduced model)}$$

versus

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \text{(the full model)}$$

where $\boldsymbol{X}_0$ is a submatrix of $\boldsymbol{X}$. Define the hat matrices for the full and reduced models as $\boldsymbol{M} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ and $\boldsymbol{M}_0 = \boldsymbol{X}_0(\boldsymbol{X}_0'\boldsymbol{X}_0)^{-1}\boldsymbol{X}_0'$, respectively, so that $\boldsymbol{M}$ projects $\boldsymbol{Y}$ onto $\mathcal{C}(\boldsymbol{X})$ and that $\boldsymbol{M}_0$ projects $\boldsymbol{Y}$ onto $\mathcal{C}(\boldsymbol{X}_0)$. Clearly, $\mathcal{C}(\boldsymbol{X}_0) \subseteq \mathcal{C}(\boldsymbol{X})$.

REGRESSION SUMS OF SQUARES FOR THE TWO COMPETING MODELS: From Section 6.7, we know that $\text{SS[R]}_F = \boldsymbol{Y}'(\boldsymbol{M} - \frac{1}{n}\boldsymbol{J})\boldsymbol{Y}$ is the regression sum of squares for the **full** model (correcting for β_0). Similarly, $\text{SS[R]}_R = \boldsymbol{Y}'(\boldsymbol{M}_0 - \frac{1}{n}\boldsymbol{J})\boldsymbol{Y}$ is the regression sum of squares for the **reduced** model (again, correcting for β_0). *Since the regression sums of squares* SS[R] *can never decrease by adding predictor variables*, it follows that

$$\text{SS[R]}_F = \boldsymbol{Y}'(\boldsymbol{M} - n^{-1}\boldsymbol{J})\boldsymbol{Y} \geq \boldsymbol{Y}'(\boldsymbol{M}_0 - n^{-1}\boldsymbol{J})\boldsymbol{Y} = \text{SS[R]}_R$$

whenever $\mathcal{C}(\boldsymbol{X}_0) \subseteq \mathcal{C}(\boldsymbol{X})$; i.e., the sums of squares for regression in the smaller model is always less than (or equal to) than the sums of squares for regression in the full model. In the light of this, our intuition should suggest the following:

- if $\text{SS[R]}_F = \boldsymbol{Y}'(\boldsymbol{M} - \frac{1}{n}\boldsymbol{J})\boldsymbol{Y}$ and $\text{SS[R]}_R = \boldsymbol{Y}'(\boldsymbol{M}_0 - \frac{1}{n}\boldsymbol{J})\boldsymbol{Y}$ are "close," then the additional predictor variables are not adding too much (in the form of an increase in sums of squares for the regression), and the reduced model probably does just as well at describing the data as the full model does.

- if $\text{SS[R]}_F = \boldsymbol{Y}'(\boldsymbol{M} - \frac{1}{n}\boldsymbol{J})\boldsymbol{Y}$ and $\text{SS[R]}_R = \boldsymbol{Y}'(\boldsymbol{M}_0 - \frac{1}{n}\boldsymbol{J})\boldsymbol{Y}$ are not "close," then the additional predictor variables are adding a lot (in the form of an increase in sums of squares for the regression). This suggests that the reduced model does an insufficient job of describing the data when compared to the full model.

REMARK: As you might suspect, we base our decision on whether or not the reduced model is adequate by examining the size of

$$\begin{aligned}\text{SS[R]}_F - \text{SS[R]}_R &= \boldsymbol{Y}'(\boldsymbol{M} - n^{-1}\boldsymbol{J})\boldsymbol{Y} - \boldsymbol{Y}'(\boldsymbol{M}_0 - n^{-1}\boldsymbol{J})\boldsymbol{Y} \\ &= \boldsymbol{Y}'(\boldsymbol{M} - \boldsymbol{M}_0)\boldsymbol{Y}.\end{aligned}$$

If this difference is large, this suggests that the reduced model does not do a good job of describing the data (when compared to the full model). *You should be aware that in this presentation, we are assuming that the full model already does a good job of describing the data; we are trying to find a smaller model that does just as well.*

REALISATION: Note that $\text{SS[R]}_F - \text{SS[R]}_R = \text{SS[E]}_R - \text{SS[E]}_F$. That is, the difference in the regression sums of squares is always equal to the difference in the error sums of squares (in absolute value).

F STATISTIC: Theoretical arguments in linear models show that when H_0 is true,

$$F = \frac{\boldsymbol{Y}'(\boldsymbol{M} - \boldsymbol{M}_0)\boldsymbol{Y}/(dfe_R - dfe_F)}{\text{MS[E]}_F} \sim F_{dfe_R - dfe_F, dfe_F},$$

where dfe_R and dfe_F denote the **error degrees of freedom** associated with the reduced and full models, respectively, and MS[E]_F denotes the MS[E] from the full model. Thus, to conduct an α level test, we reject the reduced model $\boldsymbol{Y} = \boldsymbol{X}_0\boldsymbol{\gamma} + \boldsymbol{\epsilon}$ in favour of the full model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ when F gets large. That is, this is one-sided upper-tail test with rejection region $RR = \{F : F > F_{dfe_R - dfe_F, dfe_F, \alpha}\}$.

Example 6.11 (`cheese.sas`). In Example 6.10, we want to test, using $\alpha = 0.05$,

$$\begin{aligned} H_0: &\quad Y_i = \gamma_0 + \gamma_1 x_{i3} + \epsilon_i \\ H_1: &\quad Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i. \end{aligned}$$

That is, we want to know whether or not the variables x_1 and x_2 should be added to the model. Here, the submatrix $\boldsymbol{X}_0$ corresponding to the reduced model is the full model matrix $\boldsymbol{X}$ with the second and third columns removed; i.e., the columns for x_1 and x_2. Note that we can fit the reduced model by changing the model statement in SAS to `MODEL TASTE = LACTIC`. Here are the ANOVA tables for the reduced and full models.

Analysis of Variance: Reduced Model

Source	DF	SS	MS	F	Pr > F
Model	1	3800.398	3800.398	27.55	<0.0001
Error	28	3862.489	137.946		
Corrected Total	29	7662.887			

Analysis of Variance: Full Model

Source	DF	SS	MS	F	Pr > F
Model	3	4994.509	1664.836	16.22	<0.0001
Error	26	2668.378	102.629		
Corrected Total	29	7662.887			

Thus, with $\boldsymbol{Y}'(\boldsymbol{M} - \boldsymbol{M}_0)\boldsymbol{Y} = \text{SS[R]}_F - \text{SS[R]}_R = 4994.509 - 3800.398 = 1194.111$, $dfe_R = 28$, $dfe_F = 26$, and $\text{MS[E]}_F = 102.630$, we compute the F statistic to be

$$F = \frac{\boldsymbol{Y}'(\boldsymbol{M} - \boldsymbol{M}_0)\boldsymbol{Y}/(dfe_R - dfe_F)}{\text{MS[E]}_F} = \frac{1194.111/2}{102.629} = 5.82.$$

Since $F = 5.82 > F_{2,26,0.05} = 3.369$, we would reject H_0 and conclude that the reduced model is not as good as the full model at describing these data. That is, `ACETIC` and `H2S` significantly add to a model that already includes `LACTIC`.

Question: In this example, how could you compute $\boldsymbol{Y}'(\boldsymbol{M} - \boldsymbol{M}_0)\boldsymbol{Y}$ using **sequential** sums of squares?

7 Additional Topics in Multiple Linear Regression

Complimentary reading from Rao: Chapter 11 (§ 11.8-11.10).

7.1 Introduction

In Chapter 6, we discussed many of the important concepts involving multiple linear regression models of the form $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. In particular, we focused on (a) least-squares estimation and inference for $\boldsymbol{\beta}$, (b) estimating mean values of Y and predicting future values of Y (at a given value of $\boldsymbol{x}_0$), (c) the different types of sums of squares, and (d) the general notion of reduced-versus-full model testing. In this chapter, we tie up some loose ends. In particular, I will focus on multiple-regression model diagnostics, outlier detection, influential observations, multicollinearity, criteria for choosing the best model, and sequential variable selection procedures.

Regression analysis consists of two different areas. So far, we have largely focused on **aggregate analysis** (e.g., inference for regression coefficients, testing models, predictions, etc.). This chapter starts off by discussing **case analysis** (e.g., checking the model assumptions and model-improvement strategies). The two types of analysis are not disjoint; rather, they are interrelated and complimentary. The following steps, however, outline generally what I envision to be a thorough regression analysis:

1. Formulate the problem, adopt a model, make assumptions
2. Estimation procedure
3. Estimates, confidence intervals, hypothesis tests, tentative conclusions
4. Check model assumptions (using the observed residuals), compute diagnostic statistics, transformations, determination of influential cases and outliers
5. Update model as necessary.

7.2 Least-squares residuals and residual plots

Recall our multiple linear regression model $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$, for $i = 1, 2, ..., n$, where $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$, or, in matrix notation, $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$. We can check many of the usual assumptions (e.g., normality, constant variance, etc.) by analysing the residuals. Recall that the ith least-squares **residual** is given by

$$e_i = Y_i - \widehat{Y}_i,$$

where Y_i is ith **observed** value and $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \cdots + \widehat{\beta}_p x_{ip}$ is the ith **fitted** value. We can also express the residuals using vector notation. Recall that the vector of least-squares residuals

$$\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}},$$

where $\boldsymbol{Y}$ is the vector of observed responses, and $\widehat{\boldsymbol{Y}}$ is the vector of fitted values. Also, recall that

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{M}\boldsymbol{Y},$$

where the hat matrix $\boldsymbol{M} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. The matrix $\boldsymbol{M}$ has diagonal elements $h_{ii} = \boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i$, where $\boldsymbol{x}_i'$ denotes the ith row of the design matrix $\boldsymbol{X}$. The value h_{ii} is called the **leverage** for the ith case; $i = 1, 2, ..., n$. Note that since $\widehat{\boldsymbol{Y}} = \boldsymbol{M}\boldsymbol{Y}$, we can write

$$\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{M}\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{M})\boldsymbol{Y}.$$

From this, we can show (verify!) that $E(\boldsymbol{e}) = \boldsymbol{0}$ and that

$$V(\boldsymbol{e}) = \sigma^2(\boldsymbol{I} - \boldsymbol{M}) = \begin{pmatrix} \sigma^2(1-h_{11}) & -\sigma^2 h_{12} & \cdots & -\sigma^2 h_{1n} \\ -\sigma^2 h_{21} & \sigma^2(1-h_{22}) & \cdots & -\sigma^2 h_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -\sigma^2 h_{n1} & -\sigma^2 h_{n2} & \cdots & \sigma^2(1-h_{nn}) \end{pmatrix}.$$

Thus, the variance of the ith residual is given by $V(e_i) = \sigma^2(1 - h_{ii})$, for $i = 1, 2, ..., n$. Finally, recall that under our model assumptions, $e_i = Y_i - \widehat{Y}_i$ (when viewed as a random variable) is a linear combination of the normal random variables Y_i and $\widehat{Y}_i$; thus, it, too,

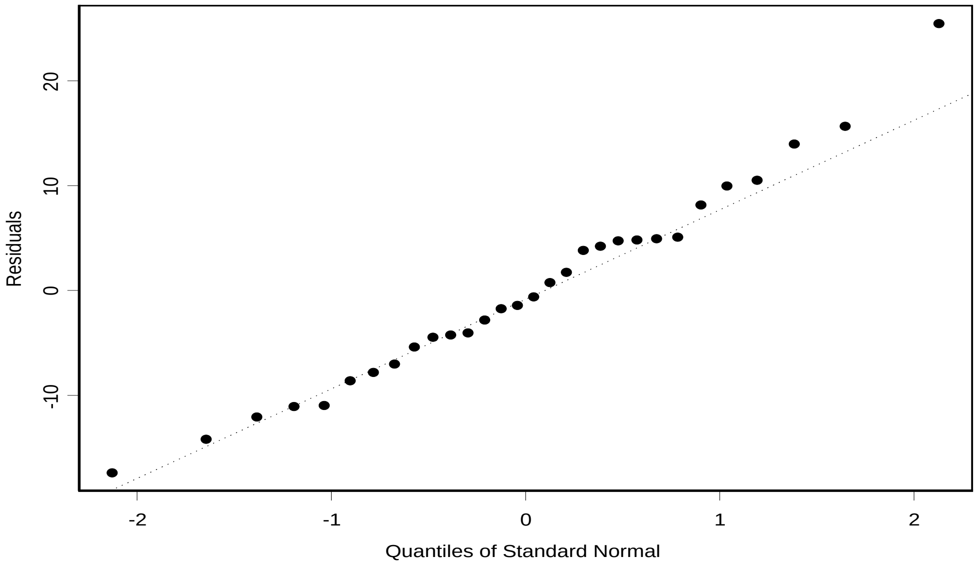

Figure 7.21: *QQ plot of the residuals from the full model fit to the cheese data.*

is normally distributed! Putting this all together, we have that, when e_i is viewed as a **random variable**,

$$e_i \sim \mathcal{N}\{0, \sigma^2(1 - h_{ii})\},$$

when all of our model assumptions hold.

NORMALITY: As in the straight-line regression setting, we can assess the plausibility of the normality assumption by plotting the residuals. Histograms of the observed residuals and normality plots are quite common.

DIAGNOSING NONCONSTANT VARIANCE AND OTHER MODEL INADEQUACIES: As in straight-line regression, a good visual display to use for diagnosing nonconstant variance and model misspecification is the plot of (observed) residuals versus predicted values; i.e., a plot of e_i versus $\widehat{y}_i$. If the model holds, it follows that $\text{Cov}(\boldsymbol{e}, \widehat{\boldsymbol{Y}}) = \mathbf{0}$, (verify!) in which case, it follows that $\text{Cov}(e_i, \widehat{Y}_i) = 0$, for each $i = 1, 2, ..., n$; i.e., the residuals and predicted values are **uncorrelated**. *Thus, again, residual plots that display nonrandom patterns suggest that there are some problems with our model assumptions.*

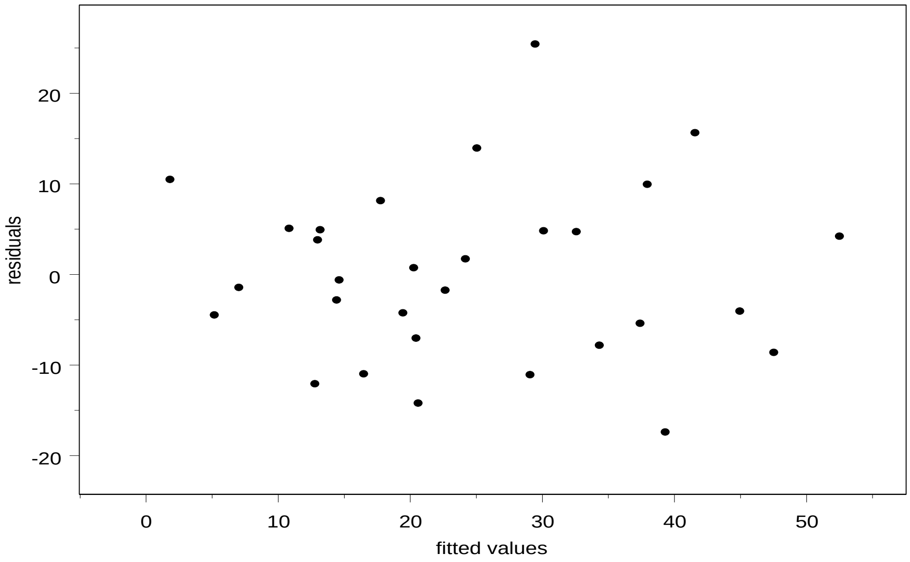

Figure 7.22: *Residual plot from the full model fit to the cheese data.*

Example 7.1 (`cheese-2.sas`). We now analyse the residuals from the full model fit to the cheese data from Example 6.1. Recall that our fitted model was given by

$$\widehat{\texttt{TASTE}}_i = -28.877 + 0.328\texttt{ACETIC}_i + 3.912\texttt{H2S}_i + 19.670\texttt{LACTIC}_i.$$

Figure 7.21 displays the normal probability plot (qq-plot); the normality assumption doesn't seem to be a large concern. Figure 7.22 displays the residual plot; this plot doesn't suggest any obvious model misspecification or a problem with constant variance.

7.3 Outlier detection and influence analysis

Just as we did in the simple-linear regression case, we can use residuals to help us assess whether or not a particular case is an **outlier**. Recall that, under our model assumptions, $e_i \sim \mathcal{N}\{0, \sigma^2(1 - h_{ii})\}$, where h_{ii} is the ith diagonal element of the hat matrix $\boldsymbol{M}$. To overcome the problem of unequal variances, we can, as in the simple linear case, construct studentised residuals that do have equal variances.

1. **Internally** studentised residuals:
$$r_i = \frac{e_i}{\sqrt{s^2(1-h_{ii})}},$$
where $s^2 = \text{MS[E]}$ is computed from all of the data. Just as in the straight-line case, $E(r_i) = 0$ and $V(r_i) = 1$. Values of $|r_i|$ larger than 3 or so should cause concern.

2. **Externally** studentised residuals:
$$t_i = \frac{e_i}{\sqrt{s^2_{-i}(1-h_{ii})}},$$
where $s^2_{-i} = \text{MS[E]}$ computed from all of the data **except** the ith case. It can be shown algebraically that
$$s^2_{-i} = \frac{(n-p)s^2 - e_i^2/(1-h_{ii})}{n-p-1}.$$

DETECTING OUTLIERS: Under our multiple-regression model assumptions, it turns out that $t_i \sim t_{n-p-1}$. Thus, at the α significance level, we may classify observation i as an outlier, after seeing the data, if $|t_i| \geq t_{n-p-1,\alpha/2n}$.

Example 7.2 (`cheese-2.sas`). With the cheese data from Example 6.1, with $n = 30$, we could classify an observation as an outlier, at the $\alpha = 0.05$ significance level, if we observe $|t_i| \geq t_{26,0.000833} = 3.51$. However, the largest $|t_i|$ is only 3.02, so we could not single out any observation as an outlier using this criterion.

TERMINOLOGY: In regression analysis, a case is said to be **influential** if its removal from consideration causes a large change in the analysis (e.g., large change in the estimated regression coefficients, ANOVA table, R^2, etc.). An influential observation need not be an outlier (or vice versa). However, most of the time, observations that outliers are usually influential (and vice versa).

COOK'S DISTANCE: To measure the influence of the ith case, Cook (1997) proposed the following statistic:
$$D_i = \frac{(\widehat{\boldsymbol{\beta}}_{-i} - \widehat{\boldsymbol{\beta}})'(\boldsymbol{X}'\boldsymbol{X})^{-1}(\widehat{\boldsymbol{\beta}}_{-i} - \widehat{\boldsymbol{\beta}})}{(p+1)\text{MS[E]}} = \frac{r_i^2}{p+1}\left(\frac{h_{ii}}{1-h_{ii}}\right),$$

where $\widehat{\boldsymbol{\beta}}_{-i}$ denotes the least-squares estimator for $\boldsymbol{\beta}$ with the ith case removed and MS[E] is the mean-squared error using all the data. Large values of D_i correspond to observations that have larger influence. An **ad-hoc guideline** is to categorise case i as an influential observation if $D_i \geq 4/n$.

OTHER INFLUENCE MEASURES: There are other influence diagnostic statistics; e.g., DFITS, DFBETAS, Hadi's influence measure, etc. Cook's D is the most-widely used.

7.4 Partial regression plots

In simple-linear regression, the relationship between the response Y and the predictor x is displayed by a scatterplot. In multiple linear regression, the situation is complicated by the relationship *between* the several predictors, so a scatterplot between Y and one of the x's may not reflect the relationship when *adjusted* for the other x's. The **partial regression plot** (or **added-variable plot**) is a graphical device that allows one to display this relationship. The following steps will produce such a plot for the variable x_j:

1. Fit the model regressing Y on all of the x's except x_j and save the residuals. Denote these residuals by $e_{Y|x_{-j}}$. Note that the variability in Y which can be explained by the modelled relationship between Y and all of the x's *except* x_j has been **removed** from these residuals.

2. Fit the model regressing x_j on all the other x's and save the residuals. Denote these residuals as $e_{x_j|x_{-j}}$. Note that these residuals contain information on the variability in x_j which cannot be explained by the modelled relationship between x_j and the other x's.

3. Plot $e_{Y|x_{-j}}$ on the vertical axis versus $e_{x_j|x_{-j}}$. This is the **partial regression plot** for x_j. It shows the relationship between the variability in Y and in x_j which is not explained by their respective linear relationships with all the other x's.

4. Interpret the plot as you would a scatterplot in the simple linear regression model.

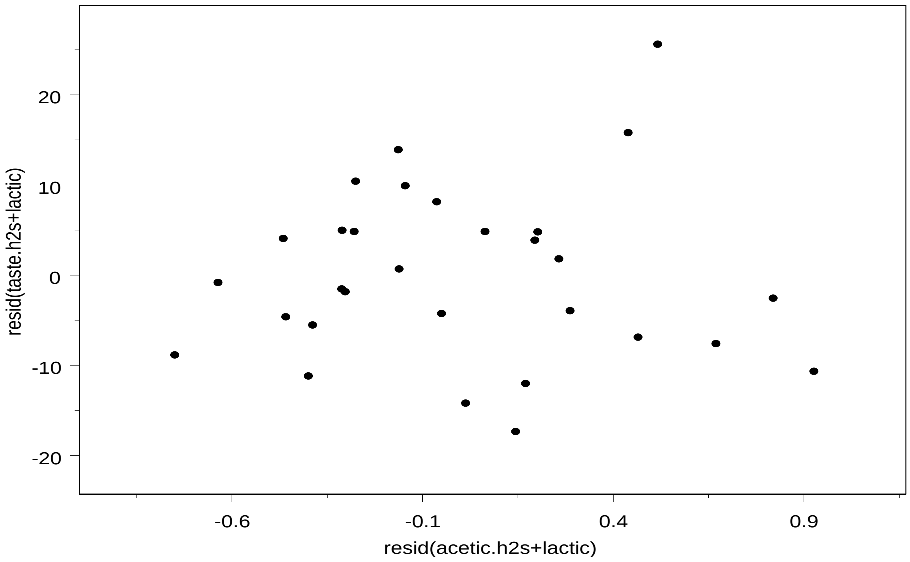

Figure 7.23: *Partial regression plot for the* `ACETIC` *predictor with the cheese data.*

- A strong linear trend indicates a strong linear relationship between x_j and Y, after having adjusted for all of the other x's.

- If this plot shows a **quadratic trend**, say, this suggests that x_j and perhaps x_j^2 should be included in the model as predictors.

- Random scatter indicates that x_j will give you little or no explanatory power in Y over that which has already been obtained with the other x's.

- If you create partial regression plots for all of the x's, do not simply eliminate all x's for which there is random scatter in the added variable plot. **Multicollinearity** between two x's could make both of their partial regression plots look random, when individually, they both were excellent predictors! If you decide to eliminate one of the x's from consideration for the final model, it is better to redo the remaining plots removing that discarded variable from the calculations of the residuals before deciding which variable to eliminate next.

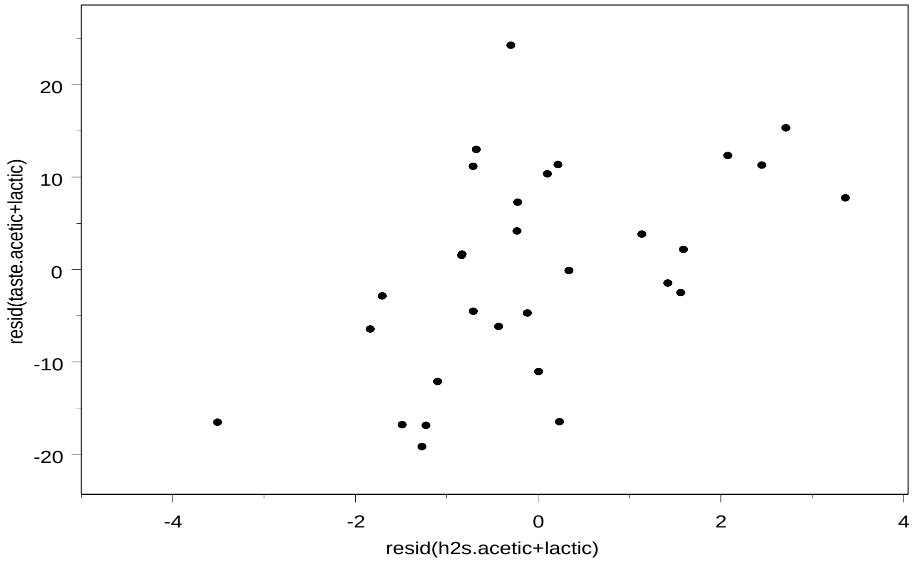

Figure 7.24: *Partial regression plot for the* H2S *predictor with the cheese data.*

Example 7.3 (cheese-2.sas). Figures 7.23, 7.24, and 7.25 display the partial regression plots for the ACETIC, H2S, and LACTIC variables, respectively. Of these, it looks like H2S has the strongest linear relationship with TASTE (when adjusting for the other predictors). This should not be surprising since H2S had the largest (partial) Type III sums of squares contribution (followed by LACTIC and ACETIC, in that order). One needs to be careful not to over-analyse these plots. Remember, that relationships among the predictors themselves might mask each predictor's linear relationship with TASTE. But, clearly, H2S's looks to be the most important predictor when adjusting for the other two because of the strong linear trend in the partial regression plot and because of its large (partial) Type III SS. The following table is reproduced from Chapter 6 for reference. No partial regression plot shows any evidence of the need to include higher order terms in the model.

Source	DF	Type III SS	MS	F Value	Pr>F
acetic	1	0.555	0.555	0.01	0.9419
h2s	1	1007.691	1007.691	9.82	0.0042
lactic	1	533.259	533.259	5.20	0.0311

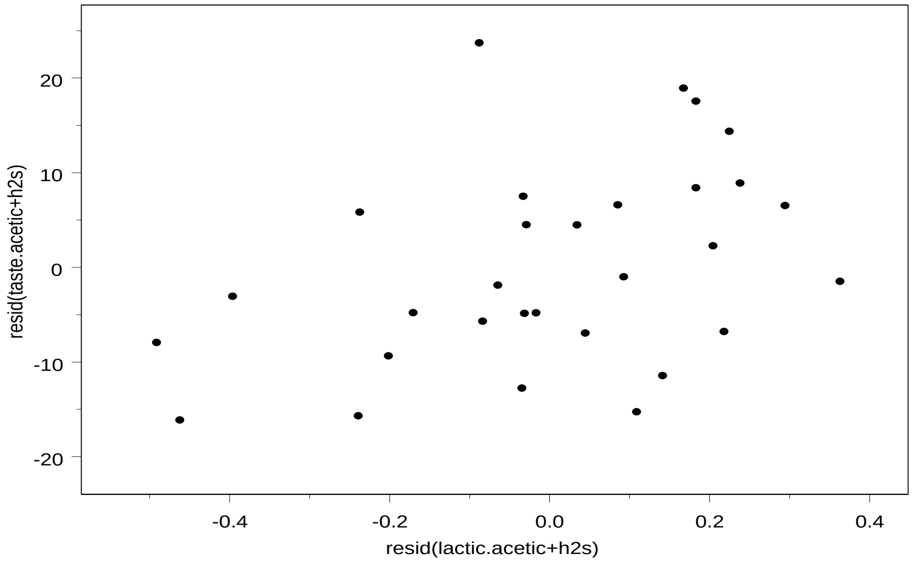

Figure 7.25: *Partial regression plot for the* `LACTIC` *predictor with the cheese data.*

7.5 Multicollinearity

Multicollinearity (or **collinearity**) exists when the predictors $x_1, x_2, ..., x_p$ are correlated. We use the term "correlated" loosely here, because in a regression analysis, the predictors $x_1, x_2, ..., x_p$ are assumed to be **fixed**.

TERMINOLOGY: A set of regression variables $x_1, x_2, ..., x_p$ is said to be **exactly collinear** if there exist constants c_0, c_1, ..., c_p (not all zero) such that

$$\sum_{j=1}^{p} c_j \boldsymbol{x}_j = c_0.$$

Example 7.4. Consider the following data:

x_1	x_2	x_3
1	8	7
2	5	3
6	10	4

These predictors are exactly collinear since $\sum_{j=1}^{3} c_j\boldsymbol{x}_j = c_0$, with $c_1 = -1$, $c_2 = 1$, $c_3 = -1$, and $c_0 = 0$; i.e., $\boldsymbol{x}_3 = \boldsymbol{x}_2 - \boldsymbol{x}_1$.

A CLOSER LOOK: If a set of variables $x_1, x_2, ..., x_p$ is **exactly collinear**, this means that there exists an x_j that can be written as a linear combination of the other x's; i.e.,

$$\begin{aligned} c_j\boldsymbol{x}_j &= c_0 - c_1\boldsymbol{x}_1 - \cdots - c_{j-1}\boldsymbol{x}_{j-1} - c_{j+1}\boldsymbol{x}_{j+1} - \cdots - c_p\boldsymbol{x}_p \\ \Longrightarrow \quad \boldsymbol{x}_j &= \frac{c_0}{c_j} - \frac{c_1}{c_j}\boldsymbol{x}_1 - \cdots - \frac{c_{j-1}}{c_j}\boldsymbol{x}_{j-1} - \frac{c_{j+1}}{c_j}\boldsymbol{x}_{j+1} - \cdots - \frac{c_p}{c_j}\boldsymbol{x}_p. \end{aligned}$$

Thus, in this situation, x_j does not add any information to the regression that is not already there in the other x's. If we have **exact collinearity**, $\boldsymbol{X}'\boldsymbol{X}$ is singular; i.e., $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ does not exist. In this extreme situation, we can not compute $\widehat{\boldsymbol{\beta}}$ uniquely.

NOTE: When the collinearity condition holds approximately; i.e., there exist constants c_0, c_1, ..., c_p (not all zero) such that

$$\sum_{j=1}^{p} c_j\boldsymbol{x}_j \approx c_0,$$

the regression variables are said to be **approximately collinear**.

REMARK: In regression, it is almost unheard of to have x's that display no collinearity (correlation). That is, the predictors are almost never **orthogonal**. However, even the presence of approximate collinearity can be rather problematic. If the x's are approximately collinear (but not perfectly collinear), $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ does exist, and we can still compute $\widehat{\boldsymbol{\beta}}$ uniquely; however, $V(\widehat{\beta}_j)$ will be very large, which makes the estimator $\widehat{\beta}_j$ practically worthless. Why? Recall that $V(\widehat{\beta}_j) = s_{jj}\sigma^2$, where $s_{jj} = (\boldsymbol{X}'\boldsymbol{X})^{-1}_{j,j}$. If we have approximate collinearity, the s_{jj} term gets large; this, in turn, inflates $V(\widehat{\beta}_j)$. In fact, it can be shown that

$$V(\widehat{\beta}_j) = \sigma^2 \underbrace{\left(\frac{1}{1-R_j^2}\right)}_{\text{VIF}_j} \left\{\sum_{i=1}^{n}(x_{ij} - \overline{x}_{+j})^2\right\}^{-1},$$

where R_j^2 is the coefficient of determination from the regression of x_j on the other predictors. The larger the degree of collinearity between x_j and the other predictors, the larger

R_j^2 becomes. As R_j^2 approaches one, $V(\widehat{\beta}_j)$ grows without bound. Thus, multicollinearity, in general, can greatly inflate the variance of our least-squares regression estimators. This, in turn, can have a large impact on the quality of the least-squares fit, and, hence, will affect the precision of confidence intervals, hypothesis tests, predictions, etc.

WAYS TO MEASURE COLLINEARITY:

1. The simplest way to diagnose collinearity is to compute all pairwise sample correlations among the predictors; that is, compute $r_{jj'}$, for all $j \neq j'$; $j = 1, 2, ..., p$. This is not foolproof though, since some collinearities may involve three or more predictors and any subset of two of them may not show collinearity.

2. Compute the **variance inflation factor** (**VIF**) defined as
$$\text{VIF}_j = \frac{1}{1 - R_j^2},$$
for $j = 1, 2, ..., p$. Values of VIF_j larger than 10 or so indicate strong collinearity. This measure, however, does not work well with categorical variables; in addition, R_j^2 is sensitive to outliers.

3. Look at the eigenvalues of (a centered and scaled version of) $\boldsymbol{X}'\boldsymbol{X}$ matrix and form the **condition index**. Belsley, Kuh, and Welsch (1980) argue that large values of
$$\kappa = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}},$$
where λ_{max} and λ_{min} denote the largest and smallest eigenvalues of (a centered and scaled version of) $\boldsymbol{X}'\boldsymbol{X}$, respectively, suggest the presence of multicollinearity. Rules declaring collinearity as a problem when $\kappa \geq 30$ or $\kappa \geq 100$ have been proposed but have little theoretical justification. The κ statistic is motivated from **principle components analysis**, a topic covered in our multivariate course.

Example 7.5 (`cheese-2.sas`). The variance inflation factors for the three predictors `ACETIC`, `H2S`, and `LACTIC`, are 1.83, 1.99, and 1.94, respectively. These are not large enough to cause great concern. The condition index is $\kappa = 3.9154/0.0036 = 33.1342$, which does suggest a mild amount of collinearity.

7.6 Criteria for choice of best model

In Chapter 6, I made the following statement:

> *The goal of any regression modelling problem should be to identify each of the important predictors, and then find the smallest model that does the best job.*

It is important to remember that we never get to know the "right" model for the data. In fact, George Box, a famous statistician, once said, "All models are wrong; some are useful." However, there is nothing that says we can't try to find the *best* model! In this section, we investigate some commonly-used statistics used for model selection. No one statistic should be used as a litmus test; instead, we should examine them all in making an informed decision. In what follows, I will assume that we have a set of **candidate models** that we are considering; again, our goal is to choose the best one.

THE COEFFICIENT OF DETERMINATION: As we have already seen, the **coefficient of determination** is proportion of the total variation in the data explained by the model; it is given by

$$R^2 = \frac{\text{SS[R]}}{\text{SS[TOT]}} = 1 - \frac{\text{SS[E]}}{\text{SS[TOT]}}.$$

Intuitively, larger values of R^2 suggest that the model is capturing more of the variability in the data explained by the regression.

THE MEAN-SQUARED ERROR: The mean-squared error, MS[E], our unbiased estimator of σ^2, can also be used as a model-selection tool. A reasonable, and certainly simple, plan is to choose the candidate model with the smallest MS[E].

NOTES ON R^2 AND MS[E]:

- R^2 can never decrease (and usually increases) by adding additional predictor variables−even if these additional predictors, in no way, provide any additional explanatory power! For example, the R^2 for the larger model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ is always at least as large as the R^2 for the smaller model $Y = \beta_0 + \beta_1 x_1 + \epsilon$.

- MS[E], on the other hand, is not always smaller for a larger model. It is true that SS[E] is always smaller for a larger model, but so is the degrees of freedom value $n-p-1$. Thus, larger models don't necessarily decrease MS[E].
- Even if R^2 is higher and MS[E] is smaller for a particular candidate model, problems caused by multicollinearity may lead us to favour the smaller model. Also, remember the **Parsimony Principle**; i.e., other things being equal, choose the smaller model for simplicity's sake.

ADJUSTED R^2: The reason that R^2 will always be larger for a larger model is that, unlike MS[E], it is not penalised for decreasing $n-p-1$ while increasing SS[R]. Thus, for this reason, the use of R^2 only as a statistic for discriminating among competing models can be extremely hazardous! A modified version, the **adjusted** R^2 addresses the deficiency. Its value is given by

$$\overline{R}^2 = 1 - \frac{\text{SS[E]}/(n-p-1)}{\text{SS[TOT]}/(n-1)} = 1 - \frac{\text{MS[E]}}{S_Y^2},$$

where $S_Y^2 = \text{SS[TOT]}/(n-1)$ is just the sample variance of $Y_1, Y_2, ..., Y_n$. Note that, holding all other things constant, as p increases, $n-p-1$ decreases, and, thus, $\overline{R}^2$ decreases. So, adding additional predictor variables does not necessarily increase $\overline{R}^2$.

NOTE: $\overline{R}^2$ and MS[E] are not necessarily "prediction-oriented;" i.e., they are not directly related to how good the model is for prediction purposes. Two well-known prediction-oriented statistics are Press and Mallows' C_p.

PRESS STATISTIC: Let $\widehat{Y}_{i,-i}$ denote the fitted value for the regression which excludes the ith case; i.e., $\widehat{Y}_{i,-i} = \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{-i}$, where $\boldsymbol{x}_i'$ is the ith row of $\boldsymbol{X}$ and $\widehat{\boldsymbol{\beta}}_{-i}$ is the least-squares estimator for $\boldsymbol{\beta}$ with the ith case removed. The quantity $e_{i,-i} = Y_i - \widehat{Y}_{i,-i}$ is called the ith Press residual. Since we remove each observation, one at a time, the Press residuals are *true prediction errors* with $\widehat{Y}_{i,-i}$ being independent of Y_i. Ideally, one would want these residuals to be small. The Press statistic amalgamates all n Press residuals together in the following way:

$$\text{Press} = \sum_{i=1}^{n} e_{i,-i}^2.$$

Models with relatively low values of the PRESS *statistic should be better than models with high* PRESS *statistics.* At first glance, your impression may be that we have to actually fit all n regressions (one time for removing each observation). However, in the early 1970's, it was discovered that the sum of squared PRESS residuals could be calculated without having to fit n regressions; in particular,

$$\text{PRESS} = \sum_{i=1}^{n} e_{i,-i}^2 = \sum_{i=1}^{n} \left(\frac{e_i}{1 - h_{ii}} \right)^2,$$

where h_{ii} is the corresponding diagonal element of the hat matrix $\boldsymbol{M}$.

MALLOWS' C_p: The basic principle behind this statistic is that it penalises the researcher for overfitting (i.e., including too many unnecessary predictor variables) and underfitting (not including the important predictors). To illustrate, suppose that we have 6 predictor variables $x_1, x_2, ..., x_6$, and that Y, in truth, follows the regression model $Y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \beta_5 x_5 + \epsilon$. Of course, we do not get to know the true model in practice. So, say that we fit the model $Y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \epsilon$. In this situation, we are **underfitting**. If we fit the full model with all 6 predictors, we are **overfitting**.

- If you underfit the model, the estimated regression coefficients and MS[E] will be biased. In particular, MS[E] overestimates the true σ^2 since we are ignoring important predictors.
- If you overfit the model, it turns out that regression coefficients and MS[E] are still unbiased estimators, but, you run the risk of inflated variances in the regression coefficients due to collinearity.

Mallows' C_p incorporates penalties for both the bias and inflated variances into one statistic. Let MS[E]_p denote the mean-squared error for a candidate model with $p \leq m$ available predictors. Then, this candidate model's C_p statistic is given by

$$C_p = (p + 1) + \frac{(\text{MS[E]}_p - \text{MS[E]}_m)(n - p - 1)}{\text{MS[E]}_m}.$$

What are "good" values of C_p? Well, if the candidate model with p predictors is truly correct, then both MS[E]_m and MS[E]_p estimate the same quantity; i.e., σ^2. In this case, one can show that

$$E(C_p) = (p+1) + \underbrace{E\left\{\frac{(\text{MS[E]}_p - \text{MS[E]}_m)(n-p-1)}{\text{MS[E]}_m}\right\}}_{\approx\, 0} \approx p+1;$$

thus, values of $C_p \approx p+1$ are preferred. Those models with C_p much greater than $p+1$ probably do not include the right combination of predictor variables (we are underfitting). Of course, the full model with $p = m$ predictors always has $C_p = p+1$.

Example 7.6 (`cheese-2.sas`). We now compute the values of PRESS, R^2, $\overline{R}^2$, MS[E], and C_p for each of the seven possible models with the cheese data in Example 6.1. It looks like the best model, based on these criteria, is the one that includes `H2S` and `LACTIC` as predictors.

Number in			Adjusted			
Model	PRESS	R-Square	R-Square	MS[E]	C(p)	Variables in Model
1	3687.95	0.5712	0.5559	117.359	6.018	h2s
1	4375.64	0.4959	0.4779	137.946	11.635	lactic
1	6111.25	0.3020	0.2771	191.027	26.116	acetic
2	3135.38	0.6517	0.6259	98.849	2.005	h2s lactic
2	3877.49	0.5822	0.5512	118.579	7.195	acetic h2s
2	4535.47	0.5203	0.4847	136.150	11.818	acetic lactic
3	3402.17	0.6518	0.6116	102.630	4.0000	acetic h2s lactic

7.7 Sequential variable selection procedures

Best subset selection methods evaluate all the possible subsets of variables from a full model and identify the best reduced regression models based on some criterion. Evaluating all possible models is the most reasonable way to proceed in variable selection, but the computational demands of evaluating every model may not be practical. For example, if we have 8 predictor variables $x_1, x_2, ..., x_8$, there are $2^8 = 256$ models to consider! There are three selection strategies that we will focus on: forward selection, backward selection, and stepwise selection.

FORWARD SELECTION:

- Start with a model that only includes an intercept.
- Consider all one-variable models. Select the one-variable model whose 1st variable has the largest t ratio, as long as it is larger than some prespecified cutoff, say t_c.
- Consider all the two-variable models obtained by adding a regression variable to the selected one-variable model. Select the two-variable model whose 2nd variable has the largest t ratio, as long as $t \geq t_c$.
- Continue this process (the next step would be to consider all three-variable models). The process stops when no other variables can be added that have $t \geq t_c$.

Using this algorithm, the number of models to be fit is at most $p(p+1)/2$, which is much less that 2^p if p is large. A common cutoff point is $t_c = 2$, because it provides an overall rule that gives an approximate size $\alpha = 0.05$ test for the importance of that regression variable. Once a variable is added using this algorithm, it must stay in the model!

BACKWARD SELECTION:

- Start with the full regression model; i.e., the model that includes all p predictors.
- Consider eliminating one variable. Delete the variable with the smallest t ratio, provided that it is smaller than some prespecified cutoff, say t_c.
- Next, refit the $p-1$ predictor variable model with this variable deleted. Then, delete the variable with the smallest t ratio if $t < t_c$.
- Continue this process. The process stops when no other variables can be deleted that have $t < t_c$.

Using this algorithm, the number of models to be fit is at most $p-1$, which is much less that 2^p if p is large. Again, a common cutoff point is $t_c = 2$ for the same reason mentioned above.

STEPWISE SELECTION:

- Same as the first three steps using forward selection.
- Now, consider deleting any variable with $t < t_c$ (it has to be the first variable added, if any here). This is what makes stepwise selection different from forward selection (a variable can actually be removed from the model).
- With the model from the last step (which may contain one or two variables at this point), add the variable with the largest t ratio, as long as $t \geq t_c$.
- With the model from the last step (which may contain two or three variables at this point), delete any variable with $t < t_c$.
- Continue this process.

Example 7.7 (`cheese-2.sas`). Applying all three sequential strategies to the cheese data from Example 6.1, we find that the two-variable model including `H2S` and `LACTIC` is selected using all three stepwise procedures.

SOME COMMENTS ON STEPWISE PROCEDURES: There are some problems with stepwise methods. First of all, it is not necessarily the case that you will arrive at the same model using all three methods! Furthermore, stepwise methods can actually give models that contain none of the variables that are in the best regressions! This is because they handle *one variable at a time.* Also, influential observations can be problematic in stepwise selections. Some statisticians argue that the models arrived at from stepwise techniques depend almost exclusively on the individual observations and have little to do with real-world effects. Variable selection should be viewed as an exploratory technique. John Tukey, among others, has emphasised the difference between **exploratory** and **confirmatory** analyses. Briefly, exploratory data analysis deals with situations in which you trying to find out what is going on in a set of data. Confirmatory data analysis looks at things like tests and confidence intervals. *If you know what variables are important beforehand, there is no need for any sequential procedures.*

8 Introduction to the General Linear Model

Complimentary reading from Rao: Chapter 12.

8.1 Introduction

We have talked a great deal about the one-way ANOVA and linear regression models. I view this chapter as the "bridge" from our discussion of regression back to a discussion of ANOVA. However, part of crossing this bridge will be realising that ANOVA and regression models can be placed under a common umbrella. This is an important point to remember for the remainder of the course.

THE GENERAL LINEAR MODEL: The model $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$, for $i = 1, 2, ..., n$, where $\epsilon_i \sim \text{iid } \mathcal{N}(0, \sigma^2)$, or, in matrix notation, $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$, is called a **general linear model**.

NOTE: The normality assumption really has nothing to do with $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ being classified as a general linear model. The important characteristics are that $E(\boldsymbol{\epsilon}) = \boldsymbol{0}$, so that $E(\boldsymbol{Y}) \in \mathcal{C}(\boldsymbol{X})$, and that $V(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_n$. The normality assumption is often appended so that we can construct confidence intervals and tests for (estimable) parameters.

8.2 One-way ANOVA as a special case of regression

Recall that we can express our one-way ANOVA model using two parameterisations:

$$\begin{aligned} \text{Means model:} \quad & Y_{ij} = \mu_i + \epsilon_{ij} \\ \text{Effects model:} \quad & Y_{ij} = \mu + \tau_i + \epsilon_{ij}, \end{aligned}$$

for $i = 1, 2, ..., t$, $j = 1, 2, ..., n_i$, where $\epsilon_{ij} \sim \text{iid } \mathcal{N}(0, \sigma^2)$. Each of these models may be expressed in the form of $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$. To help simplify the exposition, we will consider the following example.

Example 8.1 (`fert.sas`). Suppose we record wheat yields for 20 different plots, in a one-way layout, with $n_i = 5$, and $t = 4$ fertilizers (A, B, C, D). The data are:

Fertilizer	Yields, Y_{ij}	$\overline{Y}_{i+}$
A	60, 61, 59, 60, 60	$\overline{Y}_{1+} = 60$
B	62, 61, 60, 62, 60	$\overline{Y}_{2+} = 61$
C	63, 61, 61, 64, 66	$\overline{Y}_{3+} = 63$
D	62, 61, 63, 60, 64	$\overline{Y}_{4+} = 62$

The ANOVA table, from `PROC GLM`, is given by

Source	df	SS	MS	F
Fertilizer	3	25	9.333	3.92
Error	16	34	2.125	
Total	19	59		

MEANS MODEL: Consider our **means model** representation $Y_{ij} = \mu_i + \epsilon_{ij}$, and note that we can write

$$\boldsymbol{Y}_{20\times 1} = \begin{pmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{45} \end{pmatrix}, \quad \boldsymbol{X}_{20\times 4} = \begin{pmatrix} \mathbf{1}_5 & \mathbf{0}_5 & \mathbf{0}_5 & \mathbf{0}_5 \\ \mathbf{0}_5 & \mathbf{1}_5 & \mathbf{0}_5 & \mathbf{0}_5 \\ \mathbf{0}_5 & \mathbf{0}_5 & \mathbf{1}_5 & \mathbf{0}_5 \\ \mathbf{0}_5 & \mathbf{0}_5 & \mathbf{0}_5 & \mathbf{1}_5 \end{pmatrix}, \quad \boldsymbol{\beta}_{4\times 1} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{pmatrix},$$

and

$$\boldsymbol{\epsilon}_{20\times 1} = \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \vdots \\ \epsilon_{45} \end{pmatrix},$$

where $\mathbf{1}_5$ is a 5×1 column vector of ones and $\mathbf{0}_5$ is a 5×1 column vector of zeros. Thus, we have expressed the means model $Y_{ij} = \mu_i + \epsilon_{ij}$ in the form $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. As usual, the least-squares estimator $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$. Note that $r(\boldsymbol{X}) = 4$; i.e., $\boldsymbol{X}$ is full rank; thus, $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ exists, and $\widehat{\boldsymbol{\beta}}$ can be computed uniquely.

For the means model, straightforward calculations show that

$$\boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} 5 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 5 \end{pmatrix}, \quad (\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{pmatrix} \frac{1}{5} & 0 & 0 & 0 \\ 0 & \frac{1}{5} & 0 & 0 \\ 0 & 0 & \frac{1}{5} & 0 \\ 0 & 0 & 0 & \frac{1}{5} \end{pmatrix}, \quad \text{and} \quad \boldsymbol{X}'\boldsymbol{Y} = \begin{pmatrix} Y_{1+} \\ Y_{2+} \\ Y_{3+} \\ Y_{4+} \end{pmatrix}.$$

Thus, the least squares estimate of $\boldsymbol{\beta}$ for the fertilizer data is given by

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \begin{pmatrix} \frac{1}{5} & 0 & 0 & 0 \\ 0 & \frac{1}{5} & 0 & 0 \\ 0 & 0 & \frac{1}{5} & 0 \\ 0 & 0 & 0 & \frac{1}{5} \end{pmatrix} \begin{pmatrix} Y_{1+} \\ Y_{2+} \\ Y_{3+} \\ Y_{4+} \end{pmatrix} = \begin{pmatrix} \overline{Y}_{1+} \\ \overline{Y}_{2+} \\ \overline{Y}_{3+} \\ \overline{Y}_{4+} \end{pmatrix} = \begin{pmatrix} 60 \\ 61 \\ 63 \\ 62 \end{pmatrix}.$$

This agrees with our conclusion from Chapter 2; namely, that $\widehat{\mu}_i = \overline{Y}_{i+}$, for $i = 1, 2, 3, 4$.

EFFECTS MODEL: Consider our **effects model** representation $Y_{ij} = \mu + \tau_i + \epsilon_{ij}$, and note that we can write

$$\boldsymbol{Y}_{20\times 1} = \begin{pmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{45} \end{pmatrix}, \quad \boldsymbol{X}_{20\times 5} = \begin{pmatrix} \mathbf{1}_5 & \mathbf{1}_5 & \mathbf{0}_5 & \mathbf{0}_5 & \mathbf{0}_5 \\ \mathbf{1}_5 & \mathbf{0}_5 & \mathbf{1}_5 & \mathbf{0}_5 & \mathbf{0}_5 \\ \mathbf{1}_5 & \mathbf{0}_5 & \mathbf{0}_5 & \mathbf{1}_5 & \mathbf{0}_5 \\ \mathbf{1}_5 & \mathbf{0}_5 & \mathbf{0}_5 & \mathbf{0}_5 & \mathbf{1}_5 \end{pmatrix}, \quad \boldsymbol{\beta}_{5\times 1} = \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \end{pmatrix},$$

and

$$\boldsymbol{\epsilon}_{20\times 1} = \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \vdots \\ \epsilon_{45} \end{pmatrix},$$

where $\mathbf{1}_5$ is a 5×1 column vector of ones and $\mathbf{0}_5$ is a 5×1 column vector of zeros. Thus, we have expressed the effects model $Y_{ij} = \mu + \tau_i + \epsilon_{ij}$ in the form $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$. As before with the means model, we would like to estimate $\boldsymbol{\beta}$. However, note that $r(\boldsymbol{X}) = 4 < 5 = p+1$ (the sum of the last four columns of $\boldsymbol{X}$ equals the first column). That is, $\boldsymbol{X}$ is not of full rank which means that $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ does not exist. Our estimator $\widehat{\boldsymbol{\beta}}$ can not be computed uniquely using the effects model parameterisation.

SOLVING THE NORMAL EQUATIONS USING GENERALISED INVERSES: Let's go back and recall the normal equations

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}.$$

The problem with the effects model is that $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ does not exist. On the surface, this may sounds disastrous; however, we still can find a solution (there are potentially infinitely many) for $\widehat{\boldsymbol{\beta}}$ using a **generalised inverse** of $\boldsymbol{X}'\boldsymbol{X}$. We have to remember, though, that this solution is not unique (so the solution is, in a sense, rather arbitrary—but, remember how we appended those arbitrary side conditions in the one-way layout?). To be specific, a solution is given by

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{Y},$$

where $(\boldsymbol{X}'\boldsymbol{X})^{-}$ is any generalised inverse of $\boldsymbol{X}'\boldsymbol{X}$. Straightforward calculations show that (for the generalised inverse that I chose)

$$\boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} 20 & 5 & 5 & 5 & 5 \\ 5 & 5 & 0 & 0 & 0 \\ 5 & 0 & 5 & 0 & 0 \\ 5 & 0 & 0 & 5 & 0 \\ 5 & 0 & 0 & 0 & 5 \end{pmatrix}, \ (\boldsymbol{X}'\boldsymbol{X})^{-} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{5} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{5} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{5} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{5} \end{pmatrix}, \text{ and } \boldsymbol{X}'\boldsymbol{Y} = \begin{pmatrix} Y_{++} \\ Y_{1+} \\ Y_{2+} \\ Y_{3+} \\ Y_{4+} \end{pmatrix},$$

so that

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{Y} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{5} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{5} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{5} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{5} \end{pmatrix} \begin{pmatrix} Y_{++} \\ Y_{1+} \\ Y_{2+} \\ Y_{3+} \\ Y_{4+} \end{pmatrix} = \begin{pmatrix} 0 \\ \overline{Y}_{1+} \\ \overline{Y}_{2+} \\ \overline{Y}_{3+} \\ \overline{Y}_{4+} \end{pmatrix}$$

We see that this solution $\widehat{\boldsymbol{\beta}}$ is the solution that corresponds to the one that uses the side condition $\mu = 0$. If I had used another generalised inverse of $\boldsymbol{X}'\boldsymbol{X}$, I would have gotten a different solution (that corresponds to some other suitable side condition). To illustrate, my solution with the fertilizer data from Example 8.1 is

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} 0 \\ \overline{Y}_{1+} \\ \overline{Y}_{2+} \\ \overline{Y}_{3+} \\ \overline{Y}_{4+} \end{pmatrix} = \begin{pmatrix} 0 \\ 60 \\ 61 \\ 63 \\ 62 \end{pmatrix}.$$

However, SAS's solution, which uses the side condition $\tau_4 = 0$ (so SAS is computing a different generalised inverse of $\boldsymbol{X}'\boldsymbol{X}$) is given by

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \overline{Y}_{4+} \\ \overline{Y}_{1+} - \overline{Y}_{4+} \\ \overline{Y}_{2+} - \overline{Y}_{4+} \\ \overline{Y}_{3+} - \overline{Y}_{4+} \\ 0 \end{pmatrix} = \begin{pmatrix} 62 \\ -2 \\ -1 \\ 1 \\ 0 \end{pmatrix}.$$

KEY POINT: That we obtain different solutions by using different generalised inverses $(\boldsymbol{X}'\boldsymbol{X})^{-}$ is completely analogous to us obtaining different solutions using different side conditions (recall Chapter 2). In fact, we are seeing the same phenomenon; it is just being presented in a different way.

THE GOOD NEWS: In the light of our recent revisiting to the one-way layout model, and the distracting computational anomalies that have resurfaced, the good news here is that all **estimable** functions involving μ, τ_1, τ_2, τ_3, and τ_4 are uniquely estimated (as they were back in Chapter 2), even though these parameters, when considered by themselves, are not! Examples of estimable functions include $\mu_i = \mu + \tau_i$ and contrasts involving μ_i. For example, using the fertilizer data from Example 8.1, what is the least squares estimate of $\mu + \tau_1$? $\tau_3 - \tau_2$? $\tau_1 + \tau_4$? (this last function is not estimable).

MORE GOOD NEWS: Regardless of which generalised inverse $(\boldsymbol{X}'\boldsymbol{X})^{-}$ is used (or, equivalently, which side condition is used) to solve the normal equations, the ANOVA table is not affected! In addition, the ANOVA table is the same for the means model as it is for the effects model! Theoretically, this last note follows since $\mathcal{C}(\boldsymbol{X})$ is the same for both the means and effects models in the one-way layout (verify!).

A SAS SIDE NOTE: To impose the side condition $\tau_4 = 0$, essentially what SAS does to fit the regression (i.e., the ANOVA) is to take the effects model $\boldsymbol{X}$ matrix and drop the last column; i.e.,

$$\boldsymbol{X}_{20\times 4} = \begin{pmatrix} \mathbf{1}_5 & \mathbf{1}_5 & \mathbf{0}_5 & \mathbf{0}_5 \\ \mathbf{1}_5 & \mathbf{0}_5 & \mathbf{1}_5 & \mathbf{0}_5 \\ \mathbf{1}_5 & \mathbf{0}_5 & \mathbf{0}_5 & \mathbf{1}_5 \\ \mathbf{1}_5 & \mathbf{0}_5 & \mathbf{0}_5 & \mathbf{0}_5 \end{pmatrix}.$$

As you can see, this design matrix is now full rank, and thus $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ can now be computed uniquely again. In fact, for the fertilizer data, the solution is given by $\widehat{\boldsymbol{\beta}} = (62, -2, -1, 1, 0)'$, as above. By writing out the model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ in non-matrix notation, with the design matrix $\boldsymbol{X}$ as given above, we get the "regression-like" model

$$Y_i = \mu + \tau_1 z_{i1} + \tau_2 z_{i2} + \tau_3 z_{i3} + \epsilon_i,$$

for $i = 1, 2, ..., 20$, where

$$z_{i1} = \begin{cases} 1, & \text{if the response is from treatment A} \\ 0, & \text{otherwise} \end{cases}$$

$$z_{i2} = \begin{cases} 1, & \text{if the response is from treatment B} \\ 0, & \text{otherwise} \end{cases}$$

$$z_{i3} = \begin{cases} 1, & \text{if the response is from treatment C} \\ 0, & \text{otherwise,} \end{cases}$$

which is just a multiple linear regression model with μ, τ_1, τ_2, and τ_3 playing the roles of $\beta_0, \beta_1, \beta_2$, and β_3, respectively! The variables z_1, z_2, and z_3 are sometimes called **dummy variables** or **indicator variables**. They simply take on values of 0 and 1 to indicate treatment group membership.

NOTE: In general, for a t-treatment one-way layout experiment, there is no need to include any more than $t-1$ indicator variables in the regression model. Including t indicator variables leads to the overparameterised effects model, and we are left having to deal with the computational anomalies that result (which are really not that prohibitive).

REALISATION: *ANOVA is just a special case of regression with indicator variables!*

8.3 Regression models with indicator variables

In discussing multiple linear regression in Chapters 6 and 7, it was implicitly assumed that our independent variables $x_1, x_2, ..., x_p$ were **quantitative** in nature. However, as we have just seen, the multiple linear regression model is flexible enough to handle non-quantitative (i.e., **categorical**) regression variables. This is accomplished through the use of indicator variables. For example, suppose that Y = IQ score and z = gender. We can study the effect of gender on IQ via the regression of Y on z; i.e., $Y_i = \beta_0 + \beta_1 z_i + \epsilon_i$, where

$$z_i = \begin{cases} 1, & \text{if the } i\text{th subject is female} \\ 0, & \text{if the } i\text{th subject is male.} \end{cases}$$

In this example, $E(Y_i) = \beta_0 + \beta_1$ for the female subjects and $E(Y_i) = \beta_0$ for the male subjects. Actually, fitting this regression and doing the t-test for $H_0 : \beta_1 = 0$ is equivalent to the two-sample t test for the equality of two means (verify!). A slightly more elaborate model is $Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i$, where x_i = years of college. In this model, we see that there is a quantitative variable (years of college) and a categorical variable (gender).

PREVIEW: In this section, we will cover regression with categorical variables; in particular, we will discuss

1. one categorical variable with two levels (dichotomous)
2. one categorical variable with more than two levels (polytomous)
3. testing for equality of slopes and regression lines (using reduced-versus-full model tests).

ONE DICHOTOMOUS VARIABLE: As in the gender-IQ example above, we have one quantitative variable and one indicator variable. Testing $H_0 : \beta_2 = 0$ in the model $Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i$ can help us assess whether or not there is a "gender effect." If H_0 is rejected, this means that the straight-line relationship is different for the different genders (i.e., there are really two parallel lines—one for each gender). If H_0 is not rejected, we have no reason to believe that there are really two different regression lines.

Table 8.18: *Teacher effectiveness data.*

Males					Females				
Y	x_1	x_2	x_3	x_4	Y	x_1	x_2	x_3	x_4
489	81	151	45.50	43.61	410	69	125	59.00	55.66
423	68	156	46.45	44.69	569	57	131	31.75	63.97
507	80	165	76.50	54.57	425	77	141	80.50	45.32
467	107	149	55.50	43.27	344	81	122	75.50	46.67
340	43	134	49.40	49.21	324	19	141	49.00	41.21
524	129	163	72.00	49.96	505	53	152	49.35	43.83
488	139	159	86.20	53.05	235	77	141	60.75	41.61
445	88	135	64.00	49.51	501	76	132	41.25	64.57
388	99	141	44.15	39.57	600	65	157	50.75	42.41

Example 8.2 (`teacher.sas`). Eighteen student-teachers took part in an evaluation program designed to measure teacher effectiveness. Nine male and nine female instructors took part in the program. The response Y was a quantitative evaluation made on the cooperating teacher. The (quantitative) regressor variables x_1, x_2, x_3, and x_4 were the results of four different standardised tests given to each subject. The data are given in Table 8.18. Since it was believed that all of x_1, x_2, x_3, and x_4 and gender were important in describing Y (the evaluation score), the researchers decided to fit

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 z_i + \epsilon_i,$$

where

$$z_i = \begin{cases} 1, & \text{if the } i\text{th subject is female} \\ 0, & \text{if the } i\text{th subject is male.} \end{cases}$$

The role of the categorical variable, gender, can be determined by testing $H_0 : \beta_5 = 0$ versus $H_1 : \beta_5 \neq 0$. This test can be carried out using (a) a t test for β_5, or (b) a partial F test for β_5 (recall the equivalence between t tests and partial F tests). Since $\boldsymbol{X}'\boldsymbol{X}$ is large, let's use SAS. Using the t test, we see that $t = 1.11$ ($P = 0.2867$), so there is little evidence suggesting a difference between genders (of course, after adjusting for the other predictors, not all of which look to be significant). Alternatively, the partial F test

Table 8.19: *Cleansing data for different polymers.*

Polymer A		Polymer B		Polymer C	
Y	x	Y	x	Y	x
292	6.5	410	9.2	167	6.5
329	6.9	198	6.7	225	7.0
352	7.8	227	6.9	247	7.2
378	8.4	277	7.5	268	7.6
392	8.8	297	7.9	288	8.7

statistic is 1.24, which, of course, is not significant either. Recall that $t^2 = F$ (up to rounding error) and how each statistic is computed. The t statistic is given by

$$t = \frac{\widehat{\beta}_5}{\sqrt{s_{55}\text{MS[E]}}} = \frac{47.4865}{\sqrt{0.3758 \times 4827.294}} = 1.11498,$$

and the partial F statistic is given by

$$F_5 = \frac{R(\beta_5|\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)}{\text{MS[E]}} = \frac{6001.2}{4827.294} = 1.24318.$$

Also, recall that

$$6001.2 = \text{Partial SS for } x_5 = R(\beta_5|\beta_0, \beta_1, \beta_2, \beta_3, \beta_4) = \frac{\widehat{\beta}_5^2}{s_{55}} = \frac{(47.4865)^2}{0.3758}$$

(up to rounding error). Here, recall that s_{55} is the corresponding diagonal entry of the $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ matrix.

ONE POLYTOMOUS VARIABLE: Frequently, there is the need to incorporate a categorical variable with more than two levels. The extension from dichotomous to polytomous variables is straightforward. As we defined a single indicator variable for a dichotomous predictor (like gender), we can create $l - 1$ indicator variables $z_1, z_2, ..., z_{l-1}$ for a polytomous predictor with $l > 1$ levels.

Example 8.3 (`polymer.sas`). An engineer is studying the effects of the pH for a cleansing tank and polymer type on the amount of suspended solids in a coal cleansing system. Data from the experiment are given in Table 8.19. Since the engineer believed that pH

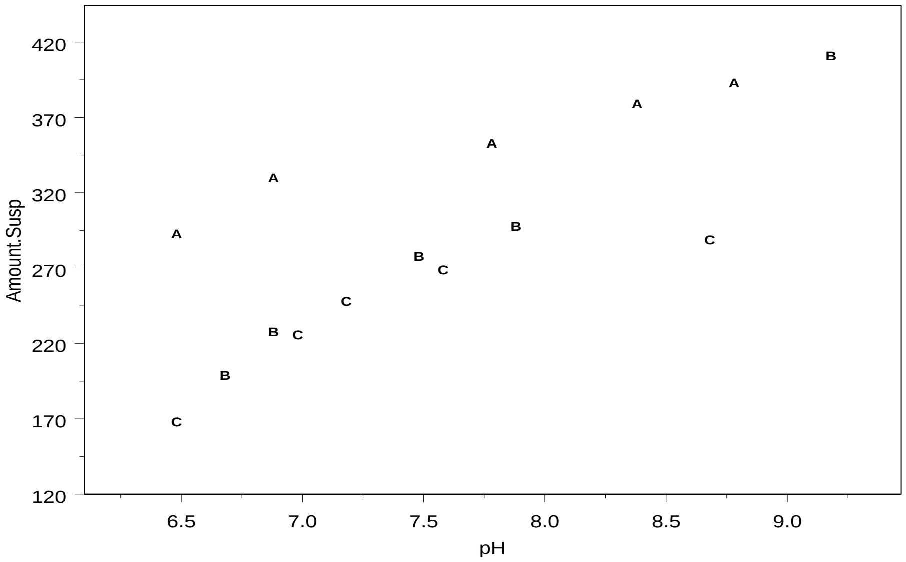

Figure 8.26: *Amount of suspended material, as a function of pH, for three polymers.*

and polymer type were both important in describing Y, the amount of suspended solids, she initially considered the **no-interaction** model (this model assumes **parallelism**):

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_{i1} + \beta_3 z_{i2} + \epsilon_i,$$

where

$$z_{i1} = \begin{cases} 1, & \text{if polymer A is used} \\ 0, & \text{otherwise,} \end{cases}$$

and

$$z_{i2} = \begin{cases} 1, & \text{if polymer B is used} \\ 0, & \text{otherwise.} \end{cases}$$

From the model $Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_{i1} + \beta_3 z_{i2} + \epsilon_i$, we can see that

$$E(Y_i) = \begin{cases} (\beta_0 + \beta_2) + \beta_1 x_i, & \text{if polymer A is used} \\ (\beta_0 + \beta_3) + \beta_1 x_i, & \text{if polymer B is used} \\ \beta_0 + \beta_1 x_i, & \text{if polymer C is used,} \end{cases}$$

so that the true regression function really is three **parallel lines**; one for each polymer. Furthermore, the test of $H_0 : \beta_2 = \beta_3 = 0$ allows us to assess whether or not there is a difference in the polymers (after adjusting for the pH level). Note that the test of $H_0 : \beta_2 = \beta_3 = 0$ is essentially the same as testing the **reduced** and **full** models:

$$\begin{aligned} H_0: &\quad Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{(reduced model)} \\ H_1: &\quad Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_{i1} + \beta_3 z_{i2} + \epsilon_i \quad \text{(full model)}. \end{aligned}$$

The full model $\boldsymbol{X}$ matrix and $\boldsymbol{\beta}$ are given below. The submatrix $\boldsymbol{X}_0$, corresponding to the reduced model, simply takes $\boldsymbol{X}$ and removes the last two columns; i.e.,

$$\boldsymbol{X} = \begin{pmatrix} 1 & 6.5 & 1 & 0 \\ 1 & 6.9 & 1 & 0 \\ 1 & 7.8 & 1 & 0 \\ 1 & 8.4 & 1 & 0 \\ 1 & 8.8 & 1 & 0 \\ 1 & 9.2 & 0 & 1 \\ 1 & 6.7 & 0 & 1 \\ 1 & 6.9 & 0 & 1 \\ 1 & 7.5 & 0 & 1 \\ 1 & 7.9 & 0 & 1 \\ 1 & 6.5 & 0 & 0 \\ 1 & 7.0 & 0 & 0 \\ 1 & 7.2 & 0 & 0 \\ 1 & 7.6 & 0 & 0 \\ 1 & 8.7 & 0 & 0 \end{pmatrix} \quad \boldsymbol{X}_0 = \begin{pmatrix} 1 & 6.5 \\ 1 & 6.9 \\ 1 & 7.8 \\ 1 & 8.4 \\ 1 & 8.8 \\ 1 & 9.2 \\ 1 & 6.7 \\ 1 & 6.9 \\ 1 & 7.5 \\ 1 & 7.9 \\ 1 & 6.5 \\ 1 & 7.0 \\ 1 & 7.2 \\ 1 & 7.6 \\ 1 & 8.7 \end{pmatrix}, \quad \text{and} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}.$$

We have a reduced-versus-full model testing situation here, and we know to base our decision on whether or not the reduced model is adequate by examining the size of

$$\text{SS[R]}_F - \text{SS[R]}_R = \boldsymbol{Y}'(\boldsymbol{M} - \boldsymbol{M}_0)\boldsymbol{Y} = R(\beta_2, \beta_3|\beta_0, \beta_1),$$

where $\boldsymbol{M} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ and $\boldsymbol{M}_0 = \boldsymbol{X}_0(\boldsymbol{X}_0'\boldsymbol{X}_0)^{-1}\boldsymbol{X}_0'$. Recall that $R(\beta_2, \beta_3|\beta_0, \beta_1)$ is the *additional sums of squares* from the regression of Y on z_1 and z_2 (after regressing on β_0 and x). Also, $R(\beta_2, \beta_3|\beta_0, \beta_1) = R(\beta_2|\beta_0, \beta_1) + R(\beta_3|\beta_0, \beta_1, \beta_2)$ so we can compute $R(\beta_2, \beta_3|\beta_0, \beta_1)$ by looking at the **sequential** SS for z_1 and z_2 (after fitting β_0 and x).

ANALYSIS: From SAS, we compute $R(\beta_2|\beta_0,\beta_1) = 20407.666$ and $R(\beta_3|\beta_0,\beta_1,\beta_2) = 2035.206$ so that $R(\beta_2,\beta_3|\beta_0,\beta_1) = 20407.666 + 2035.206 = 22442.8723$. Thus, the reduced-versus-full model F test statistic for H_0 versus H_1 is given by

$$F = \frac{R(\beta_2,\beta_3|\beta_0,\beta_1)/(dfe_R - dfe_F)}{\text{MS[E]}_F} = \frac{22442.8723/2}{519.681} = 21.593,$$

which is much larger than $F_{2,11,0.05} = 3.982$. Thus, the reduced model does not do as well at describing these data as does the full model; that is, the relationship is better described by 3 parallel lines (one for each polymer) rather than a single line. *Note that we have not concluded that a parallel-lines model is appropriate for these data; only that it does a better job than a single regression line.*

REGRESSION MODELS WITH INTERACTION: Instead of requiring equal slopes, a more flexible model is one that allows for individual lines (planes) to have different slopes. Regression lines (planes) with different slopes occurs when there is an **interaction** between the quantitative and categorical predictors. Two variables (e.g., pH and polymer type) are said to have an **interaction effect** on the response if the change in the expected response that results from changing the level of one of variable depends on the level of another variable.

Example 8.4 (`polymer.sas`). For the polymer data, instead of assuming the parallel-lines regression model $Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_{i1} + \beta_3 z_{i2} + \epsilon_i$, suppose that we consider the **interaction-regression model**

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_{i1} + \beta_3 z_{i2} + \beta_4 x_i z_{i1} + \beta_5 x_i z_{i2} + \epsilon_i,$$

where, as before,

$$z_{i1} = \begin{cases} 1, & \text{if polymer A is used} \\ 0, & \text{otherwise,} \end{cases}$$

and

$$z_{i2} = \begin{cases} 1, & \text{if polymer B is used} \\ 0, & \text{otherwise.} \end{cases}$$

Interaction can be modelled nicely by adding the predictors xz_1 and xz_2, which are formed by multiplication. From the interaction model, we can see that

$$E(Y_i) = \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_4)x_i, & \text{if polymer A is used} \\ (\beta_0 + \beta_3) + (\beta_1 + \beta_5)x_i, & \text{if polymer B is used} \\ \beta_0 + \beta_1 x_i, & \text{if polymer C is used.} \end{cases}$$

RECALL: We saw that that the parallel-lines model is better at describing the data than a single-regression line. However, does the interaction model (which allows for lines with different slopes and intercepts) do a better job than the parallel-lines model?

TESTING FOR COMMON SLOPES: To test for **common slopes** among the three regression lines, we could test $H_0 : \beta_4 = \beta_5 = 0$, or, equivalently, in a reduced-versus-full model setup, test

$$\begin{aligned} H_0 : &\quad Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_{i1} + \beta_3 z_{i2} + \epsilon_i \quad \text{(reduced model)} \\ H_1 : &\quad Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_{i1} + \beta_3 z_{i2} + \beta_4 x_i z_{i1} + \beta_5 x_i z_{i2} + \epsilon_i \quad \text{(full model).} \end{aligned}$$

The interaction model is the full model and the parallel-lines model is the reduced model. Our full and reduced model design matrices are

$$\boldsymbol{X} = \begin{pmatrix} 1 & 6.5 & 1 & 0 & 6.5 & 0 \\ 1 & 6.9 & 1 & 0 & 6.9 & 0 \\ 1 & 7.8 & 1 & 0 & 7.8 & 0 \\ 1 & 8.4 & 1 & 0 & 8.4 & 0 \\ 1 & 8.8 & 1 & 0 & 8.8 & 0 \\ 1 & 9.2 & 0 & 1 & 0 & 9.2 \\ 1 & 6.7 & 0 & 1 & 0 & 6.7 \\ 1 & 6.9 & 0 & 1 & 0 & 6.9 \\ 1 & 7.5 & 0 & 1 & 0 & 7.5 \\ 1 & 7.9 & 0 & 1 & 0 & 7.9 \\ 1 & 6.5 & 0 & 0 & 0 & 0 \\ 1 & 7.0 & 0 & 0 & 0 & 0 \\ 1 & 7.2 & 0 & 0 & 0 & 0 \\ 1 & 7.6 & 0 & 0 & 0 & 0 \\ 1 & 8.7 & 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \boldsymbol{X}_0 = \begin{pmatrix} 1 & 6.5 & 1 & 0 \\ 1 & 6.9 & 1 & 0 \\ 1 & 7.8 & 1 & 0 \\ 1 & 8.4 & 1 & 0 \\ 1 & 8.8 & 1 & 0 \\ 1 & 9.2 & 0 & 1 \\ 1 & 6.7 & 0 & 1 \\ 1 & 6.9 & 0 & 1 \\ 1 & 7.5 & 0 & 1 \\ 1 & 7.9 & 0 & 1 \\ 1 & 6.5 & 0 & 0 \\ 1 & 7.0 & 0 & 0 \\ 1 & 7.2 & 0 & 0 \\ 1 & 7.6 & 0 & 0 \\ 1 & 8.7 & 0 & 0 \end{pmatrix}.$$

To determine whether parallelism (reduced model) or interaction (full model) is more supported by the data, all we need to do is examine the size of

$$\text{SS[R]}_F - \text{SS[R]}_R = \boldsymbol{Y}'(\boldsymbol{M} - \boldsymbol{M}_0)\boldsymbol{Y} = R(\beta_4, \beta_5|\beta_0, \beta_1, \beta_2, \beta_3),$$

where $\boldsymbol{M} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ and $\boldsymbol{M}_0 = \boldsymbol{X}_0(\boldsymbol{X}_0'\boldsymbol{X}_0)^{-1}\boldsymbol{X}_0'$. For this test, recall that $R(\beta_4, \beta_5|\beta_0, \beta_1, \beta_2, \beta_3)$ is the *additional sums of squares* from the regression of Y on xz_1 and xz_2 (after regressing on all other variables). If this quantity is large, this suggests that the full model is better than the reduced model at describing the data.

ANALYSIS: We make this more concrete by constructing the reduced-versus-full model F statistic. From SAS, we compute

$$F = \frac{R(\beta_4, \beta_5|\beta_0, \beta_1, \beta_2, \beta_3)/(dfe_R - dfe_F)}{\text{MS[E]}_F} = \frac{(2017.17633 + 1604.98851)/2}{232.703} = 7.783,$$

which is larger than $F_{2,9,0.05} = 4.256$. Thus, we have enough evidence at the $\alpha = 0.05$ level to conclude that the interaction model is better at describing these data than the model which assumes parallelism. The P value for this test 0.0109.

TESTING FOR COMMON INTERCEPTS: Instead of testing for common slopes, we might want to test whether or not there are **common intercepts** in the three regression lines (this is sometimes called **concurrent regression**). From the interaction model $Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_{i1} + \beta_3 z_{i2} + \beta_4 x_i z_{i1} + \beta_5 x_i z_{i2} + \epsilon_i$, we can see, again, that

$$E(Y_i) = \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_4)x_i, & \text{if polymer A is used} \\ (\beta_0 + \beta_3) + (\beta_1 + \beta_5)x_i, & \text{if polymer B is used} \\ \beta_0 + \beta_1 x_i, & \text{if polymer C is used.} \end{cases}$$

Thus, to test for common intercepts, we are basically assessing whether or not the hypothesis $H_0 : \beta_2 = \beta_3 = 0$ is supported by the data. Note that this test still allows the slopes to be different. As you may be guessing by now, we can conduct this test by posing it as a reduced-versus-full model test! Here, we can consider

$$\begin{aligned} H_0 :&\quad Y_i = \beta_0 + \beta_1 x_i + \beta_4 x_i z_{i1} + \beta_5 x_i z_{i2} + \epsilon_i \quad \text{(reduced model)} \\ H_1 :&\quad Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_{i1} + \beta_3 z_{i2} + \beta_4 x_i z_{i1} + \beta_5 x_i z_{i2} + \epsilon_i \quad \text{(full model),} \end{aligned}$$

where the reduced model forces all intercepts to be the same and the full model does not. The reduced and full model matrices $\boldsymbol{X}_0$ and $\boldsymbol{X}$, respectively, for this test are given by

$$\boldsymbol{X} = \begin{pmatrix} 1 & 6.5 & 1 & 0 & 6.5 & 0 \\ 1 & 6.9 & 1 & 0 & 6.9 & 0 \\ 1 & 7.8 & 1 & 0 & 7.8 & 0 \\ 1 & 8.4 & 1 & 0 & 8.4 & 0 \\ 1 & 8.8 & 1 & 0 & 8.8 & 0 \\ 1 & 9.2 & 0 & 1 & 0 & 9.2 \\ 1 & 6.7 & 0 & 1 & 0 & 6.7 \\ 1 & 6.9 & 0 & 1 & 0 & 6.9 \\ 1 & 7.5 & 0 & 1 & 0 & 7.5 \\ 1 & 7.9 & 0 & 1 & 0 & 7.9 \\ 1 & 6.5 & 0 & 0 & 0 & 0 \\ 1 & 7.0 & 0 & 0 & 0 & 0 \\ 1 & 7.2 & 0 & 0 & 0 & 0 \\ 1 & 7.6 & 0 & 0 & 0 & 0 \\ 1 & 8.7 & 0 & 0 & 0 & 0 \end{pmatrix} \quad \boldsymbol{X}_0 = \begin{pmatrix} 1 & 6.5 & 6.5 & 0 \\ 1 & 6.9 & 6.9 & 0 \\ 1 & 7.8 & 7.8 & 0 \\ 1 & 8.4 & 8.4 & 0 \\ 1 & 8.8 & 8.8 & 0 \\ 1 & 9.2 & 0 & 9.2 \\ 1 & 6.7 & 0 & 6.7 \\ 1 & 6.9 & 0 & 6.9 \\ 1 & 7.5 & 0 & 7.5 \\ 1 & 7.9 & 0 & 7.9 \\ 1 & 6.5 & 0 & 0 \\ 1 & 7.0 & 0 & 0 \\ 1 & 7.2 & 0 & 0 \\ 1 & 7.6 & 0 & 0 \\ 1 & 8.7 & 0 & 0 \end{pmatrix}.$$

To determine whether common slopes (i.e., the reduced model) is supported by the data, we need to examine the size of

$$\text{SS[R]}_F - \text{SS[R]}_R = \boldsymbol{Y}'(\boldsymbol{M} - \boldsymbol{M}_0)\boldsymbol{Y}$$

where $\boldsymbol{M} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ and $\boldsymbol{M}_0 = \boldsymbol{X}_0(\boldsymbol{X}_0'\boldsymbol{X}_0)^{-1}\boldsymbol{X}_0'$. Unfortunately, we can not get $\boldsymbol{Y}'(\boldsymbol{M}-\boldsymbol{M}_0)\boldsymbol{Y}$ here using sequential SS (unless we change the order the predictors were entered). This is not problematic because I can fit the reduced and full model regressions and retrieve SS[R]_R and SS[R]_F from the ANOVA tables (on the next page).

ANALYSIS: From SAS, we compute $\boldsymbol{Y}'(\boldsymbol{M}-\boldsymbol{M}_0)\boldsymbol{Y} = \text{SS[R]}_F - \text{SS[R]}_R = 70164.07724 - 65340.17632 = 4823.90104$ so that

$$F = \frac{\boldsymbol{Y}'(\boldsymbol{M} - \boldsymbol{M}_0)\boldsymbol{Y}/(dfe_R - dfe_F)}{\text{MS[E]}_F} = \frac{4823.90104/2}{232.703} = 10.365,$$

which is larger than $F_{2,9,0.05} = 4.256$. Thus, we have enough evidence at the $\alpha = 0.05$ level to conclude that the common intercept model is not appropriate for these data.

Reduced Model: Common intercepts

Source	DF	SS	MS	F	Pr > F
Model	3	65340.17632	21780.05877	34.63	<0.0001
Error	11	6918.22368	628.92943		
Corrected Total	14	72258.40000			

Full Model

Source	DF	SS	MS	F	Pr > F
Model	5	70164.07724	14032.81545	60.30	<0.0001
Error	9	2094.32276	232.70253		
Corrected Total	14	72258.40000			

OVERALL CONCLUSION: We need the **full model** to completely describe the polymer data; that is, the slopes and intercepts all vary. We have arrived at this conclusion by systematically ruling out a straight-line model (for all three polymers combined), a parallel-lines regression model, and a concurrent (common intercepts) model.

SUMMARY: We have discussed four different models for the polymer data:

$$\begin{aligned}
\text{Coincident:} \quad & Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \\
\text{Common slope:} \quad & Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_{i1} + \beta_3 z_{i2} + \epsilon_i \\
\text{Common intercept:} \quad & Y_i = \beta_0 + \beta_1 x_i + \beta_4 x_i z_{i1} + \beta_5 x_i z_{i2} + \epsilon_i \\
\text{Interaction (Full):} \quad & Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_{i1} + \beta_3 z_{i2} + \beta_4 x_i z_{i1} + \beta_5 x_i z_{i2} + \epsilon_i,
\end{aligned}$$

It is interesting to observe the **hierarchical structure** among these four models; in particular,

- the first three are reduced models when compared to the full model.
- the first model is a reduced model when compared to the second and third models.
- the second and third models are not related.

8.4 Analysis of covariance (ANCOVA)

Analysis of covariance incorporates one or more regression variables (i.e., covariates) into an analysis of variance. For now, we will just focus on the one-way ANOVA model. Recall that this model can be expressed as $Y_{ij} = \mu + \tau_i + \epsilon_{ij}$, for $i = 1, 2, ..., t$ and $j = 1, 2, ..., n_i$, where $\epsilon_{ij} \sim$ iid $\mathcal{N}(0, \sigma^2)$. Before we proceed with ANCOVA models, I want to step back and talk about the design assumptions which underpin the one-way ANOVA model.

RECALL: In the one-way ANOVA model, we are basically assuming a one-way classification; i.e., experimental units are thought to be "basically alike," and the only way units may be "classified" is with respect to which treatment they receive. In addition, we require that experimental units are randomly assigned to treatments. We called this **complete randomisation**, and we called this experimental design a **completely randomised design**. When experimental units are thought to be "basically alike," then the experimental error only consists of the variation among the experimental units themselves (that is, there are no other systematic sources of variation).

USING COVARIATES: In some applications, we may have access to one or more variables that provides extra information about each experimental unit (that is, information beyond knowing only which treatment the unit received). These variables are sometimes called **covariates** or **concomitant variables**. Clearly, if these variables are important in describing Y, then we would be foolish not to use them. After all, including these variables in our model could reduce the experimental error variance σ^2. This, in turn, could greatly sharpen the analysis because we are explaining more of the variation by including the covariates. Of course, if the covariates are not important in describing Y, then there is no need to use them.

Example 8.5. Suppose that in Example 8.1, we recorded x, the nitrogen content (in ppm) for each plot. In this case, we have access to extra information regarding each experimental unit (plot) beyond knowing only which fertilizer the plot received. If nitrogen is important in describing the mean yield, then we should incorporate it somehow.

USING COVARIATES VERSUS BLOCKING: In general, if, beforehand, we know that experimental units (e.g., plots, rats, plants, etc.) are systematically different in some way, then we should **not** use a completely randomised design. In this situation, it is more appropriate to group the individuals into **blocks**, where, in each block, individuals are "basically alike." Treatments should then be assigned at random within each block (this is sometimes called **restricted randomisation**). However, in many experimental situations, covariate information is observed (or given to the statistician) after a complete randomisation scheme has been set and the data have already been collected. So, in a sense, ANCOVA, which incorporates this covariate information, can be thought of as a "salvage effort." Had we known about this extra information beforehand, we could have designed the experiment to include it (in the form of blocking). If this information is given after the fact, we are left asking ourselves simply if this information is helpful in determining whether or not there are really differences among the treatments.

ANCOVA LINEAR MODEL: A model for the analysis of covariance is given by

$$\begin{aligned} Y_{ij} &= \mu_i + \gamma(x_{ij} - \overline{x}_{++}) + \epsilon_{ij} \\ &= \mu + \tau_i + \gamma(x_{ij} - \overline{x}_{++}) + \epsilon_{ij}, \end{aligned}$$

for $i = 1, 2, ..., t$, $j = 1, 2, ..., n_i$, where $\epsilon_{ij} \sim$ iid $\mathcal{N}(0, \sigma^2)$. Here, $\mu_i = \mu + \tau_i$ and $\overline{x}_{++}$ is the overall mean of the x_{ij}'s. In this analysis of covariance model, the x_{ij}'s are assumed to be fixed quantities whose values are not influenced by the treatments (since γ is the same for each treatment). For this model, it can be shown (verify!) that

$$E(\overline{Y}_{i+}) = \mu + \tau_i + \gamma(\overline{x}_{i+} - \overline{x}_{++}),$$

where $\overline{x}_{i+} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$, and that

$$E(\overline{Y}_{++}) = \mu.$$

This is an advantage to parameterising the ANCOVA model as above; namely, using the centred value $(x_{ij} - \overline{x}_{++})$ in the model formulation ensures that $E(\overline{Y}_{++}) = \mu$.

THE GENERAL LINEAR MODEL: Either ANCOVA model (the means or effects version) can be written in the general form $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. To see this, suppose that $t = 3$ and that $n_1 = n_2 = n_3 = 2$. For the means model formulation $Y_{ij} = \mu_i + \gamma(x_{ij} - \overline{x}_{++}) + \epsilon_{ij}$, we have

$$\boldsymbol{Y} = \begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 1 & 0 & 0 & x_{11} - \overline{x}_{++} \\ 1 & 0 & 0 & x_{12} - \overline{x}_{++} \\ 0 & 1 & 0 & x_{21} - \overline{x}_{++} \\ 0 & 1 & 0 & x_{22} - \overline{x}_{++} \\ 0 & 0 & 1 & x_{31} - \overline{x}_{++} \\ 0 & 0 & 1 & x_{32} - \overline{x}_{++} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \gamma \end{pmatrix}, \quad \text{and} \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{31} \\ \epsilon_{32} \end{pmatrix}.$$

Since $\boldsymbol{X}$ is full rank (when at least one x_{ij} is different), we can compute the least squares estimate $\widehat{\boldsymbol{\beta}}$ uniquely. For the effects model formulation $Y_{ij} = \mu + \tau_i + \gamma(x_{ij} - \overline{x}_{++}) + \epsilon_{ij}$, we have

$$\boldsymbol{Y} = \begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 1 & 1 & 0 & 0 & x_{11} - \overline{x}_{++} \\ 1 & 1 & 0 & 0 & x_{12} - \overline{x}_{++} \\ 1 & 0 & 1 & 0 & x_{21} - \overline{x}_{++} \\ 1 & 0 & 1 & 0 & x_{22} - \overline{x}_{++} \\ 1 & 0 & 0 & 1 & x_{31} - \overline{x}_{++} \\ 1 & 0 & 0 & 1 & x_{32} - \overline{x}_{++} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \\ \gamma \end{pmatrix}, \quad \text{and} \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{31} \\ \epsilon_{32} \end{pmatrix}.$$

Here, as we have seen with effects models, the design matrix $\boldsymbol{X}$ is not full rank. The least squares estimate $\widehat{\boldsymbol{\beta}}$ can not be computed without imposing a suitable side condition (e.g., $\sum_i \tau_i = 0$, etc.).

IMPORTANT REALISATION: For now, it is important to note that the ANCOVA model is just a special case of the general linear model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

Example 8.6 (`pigs.sas`). Each of 19 randomly selected pigs is assigned, at random, to one of four diet regimes (A, B, C, and D). Note that complete randomisation is used. Here, the individual pigs are the experimental units and the treatments are the diets. The response variable, Y, is pig weight (lbs) after having been raised on the diets. In addition to treatment assignment, we also have x, the initial weight of each pig, prior to the experiment. The data are given in Table 8.20.

Table 8.20: *Pig weight gain data.*

Diet A		Diet B		Diet C		Diet D	
Y	x	Y	x	Y	x	Y	x
60.8	663.3	68.7	517.2	102.6	538.3	87.9	555.4
50.7	623.3	67.7	589.3	102.1	512.8	84.2	469.2
65.0	689.5	74.0	596.0	100.2	558.2	83.1	454.8
58.6	665.7	66.3	589.3	96.5	499.3	85.7	505.3
61.7	605.9	69.8	521.1			90.3	615.5

ANALYSIS: To start off, let's ignore the covariate (initial weight) and just fit a one-way ANOVA model $Y_{ij} = \mu + \tau_i + \epsilon_{ij}$. The ANOVA table and parameter estimates, from `PROC GLM`, are given by

```
                  One-way ANOVA

   Source  df     SS          MS          F
   ------------------------------------------
    Diet    3  4433.035   1477.678   104.77
   Error   15   211.554     14.104
   ------------------------------------------
   Total   18  4644.589
```

```
      Parameter Estimates

    Intercept   Estimate
    ----------------------
    Intercept    86.240 B
      Diet A    -26.780 B
      Diet B    -16.940 B
      Diet C     14.110 B
      Diet D      0.000 B
```

Based on this analysis (which ignores the initial weights), there looks to be a significant difference among the four diets. The least squares estimates of $\mu + \tau_1$, $\mu + \tau_2$, $\mu + \tau_3$, and $\mu + \tau_4$ are given by $\overline{y}_{1+} = 59.46$ (i.e., 86.24—26.78), $\overline{y}_{2+} = 69.30$, $\overline{y}_{3+} = 100.35$, and $\overline{y}_{4+} = 86.24$, respectively. These are sometimes called the **unadjusted treatment means** because they do not account for the covariate (initial weight). Recall that in the one-way ANOVA model, $E(\overline{Y}_{i+}) = \mu + \tau_i = \mu_i$.

Example 8.7 (`pigs.sas`). For the pig data in Example 8.6, let's use the initial weight information in an ANCOVA model. From SAS, we compute

`ANCOVA model`

Source	df	SS	MS	F
Model	4	4487.167	1121.792	99.76
Error	14	157.423	11.244	
Total	18	4644.589		

Source	df	Type III SS	MS	F	Pr>F
Diet	3	2825.215	941.738	83.75	<0.0001
Init	1	54.131	54.131	4.81	0.0456

PRELIMINARY OBSERVATIONS: First, note that the MS[E] from the ANCOVA model is smaller than that for the ANOVA analysis. This suggests that initial weight describes some of the experimental error from the ANOVA model which ignores initial weight. Is the covariate (initial weight) statistically important in describing weight gain? We can answer this question by noting that the F statistic associated with weight gain is large enough at the $\alpha = 0.05$ level ($P = 0.0456$). You should also note that the F ratio for `DIET` is large. This suggests that, *after adjusting for the initial weight covariate*, there is a significant difference among the diets; that is, one would reject $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$.

TESTING THE SIGNIFICANCE OF THE COVARIATE: We discuss briefly the theory for testing $H_0 : \gamma = 0$ in the ANCOVA model. It may come as no shock to you that we can test this hypothesis using a reduced-versus-full model testing setup! Consider the reduced and full models

$$\begin{aligned} H_0: &\quad Y_{ij} = \mu_i + \epsilon_{ij} \quad \text{(reduced model)} \\ H_1: &\quad Y_{ij} = \mu_i + \gamma(x_{ij} - \overline{x}_{++}) + \epsilon_{ij} \quad \text{(full model),} \end{aligned}$$

and let $\boldsymbol{M}_0$ and $\boldsymbol{M}$ denote the reduced and full model hat matrices. As usual, we base our decision on the size of $\text{SS[R]}_F - \text{SS[R]}_R = \boldsymbol{Y}'(\boldsymbol{M} - \boldsymbol{M}_0)\boldsymbol{Y}$, or equivalently, on the size of

$$F = \frac{\boldsymbol{Y}'(\boldsymbol{M} - \boldsymbol{M}_0)\boldsymbol{Y}/1}{\text{MS[E]}_F}$$

Note that $\boldsymbol{Y}'(\boldsymbol{M} - \boldsymbol{M}_0)\boldsymbol{Y}$ can also be computed by using the partial sums of squares for γ. To illustrate using the pig weight-gain data, SAS gives $\boldsymbol{Y}'(\boldsymbol{M} - \boldsymbol{M}_0)\boldsymbol{Y} = R(\gamma|\mu_1, \mu_2, \mu_3, \mu_4) = 54.131257$ and $\text{MS[E]}_F = 11.244$. The ratio of these two quantities is $F = 4.81$, which is given in the SAS output.

ADJUSTED TREATMENT MEANS: Recall that in the ANCOVA model,

$$E(\overline{Y}_{i+}) = \mu + \tau_i + \gamma(\overline{x}_{i+} - \overline{x}_{++}).$$

Thus, in the ANCOVA model, the sample mean $\overline{Y}_{i+}$ estimates $\mu_i = \mu + \tau_i$ plus the extra term $\gamma(\overline{x}_{i+} - \overline{x}_{++})$. The extra term accounts for the fact that we are using the covariate x in the model. In the light of this, we define the **adjusted treatment mean** to be

$$\text{Adj. } \overline{Y}_{i+} = \overline{Y}_{i+} - \widehat{\gamma}(\overline{x}_{i+} - \overline{x}_{++}),$$

where $\widehat{\gamma}$ is the least-squares estimator of γ. As the unadjusted treatment mean $\overline{Y}_{i+}$ is the least squares estimator of $\mu + \tau_i = \mu_i$ in the ANOVA model, the adjusted treatment mean Adj. $\overline{Y}_{i+}$ is the least squares estimator of $\mu + \tau_i = \mu_i$ in the ANCOVA model.

Example 8.8 (`pigs.sas`). For the pig data in Example 8.6, we have $\overline{x}_{1+} = 649.54$, $\overline{x}_{2+} = 562.58$, $\overline{x}_{3+} = 527.15$, $\overline{x}_{4+} = 520.04$, $\overline{x}_{++} = 566.81$, and $\widehat{\gamma} = 0.0422$ (from SAS). Thus, the **adjusted treatment means** are given by

$$\begin{aligned}
\text{Adj. } \overline{y}_{1+} &= 59.46 - 0.0422(649.54 - 566.81) = 55.97 \\
\text{Adj. } \overline{y}_{2+} &= 69.30 - 0.0422(562.58 - 566.81) = 69.48 \\
\text{Adj. } \overline{y}_{3+} &= 100.35 - 0.0422(527.15 - 566.81) = 102.02 \\
\text{Adj. } \overline{y}_{4+} &= 86.24 - 0.0422(520.04 - 566.81) = 88.21.
\end{aligned}$$

As one can see, these are somewhat different from the unadjusted treatment means. This suggests that the covariate (initial weight) does have an impact on the analysis. SAS calls the adjusted treatment means "least-squares means," or "LSMEANS," for short.

CONFIDENCE INTERVALS FOR μ_i: Denoting $N = \sum_i n_i$, one can show that

$$\sigma^2_{\text{Adj.}\overline{Y}_{i+}} \equiv V(\text{Adj. } \overline{Y}_{i+}) = \sigma^2 \left[\frac{1}{n_i} + \frac{(\overline{x}_{i+} - \overline{x}_{++})^2}{E_{xx}} \right],$$

where $E_{xx} = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_{i+})^2$. An estimate of this variance is given by

$$\widehat{\sigma}^2_{\text{Adj.}\overline{Y}_{i+}} = \text{MS[E]} \left[\frac{1}{n_i} + \frac{(\overline{x}_{i+} - \overline{x}_{++})^2}{E_{xx}} \right],$$

where MS[E] is the mean-squared error from the ANCOVA model. A $100(1-\alpha)$ percent confidence interval for μ_i, based on the adjusted treatment mean Adj. $\overline{Y}_{i+}$, is given by

$$\text{Adj. } \overline{Y}_{i+} \pm t_{N-p-1,\alpha/2} \times \widehat{\sigma}_{\text{Adj.}\overline{Y}_{i+}},$$

where $N-p-1$ denotes the error degrees of freedom from the ANCOVA model fit. SAS provides these confidence intervals on request.

Example 8.9 (`pigs.sas`). For the pig data in Example 8.6, confidence intervals for the treatment means (using the adjusted estimates) are given in the following table:

Diet	Gain LSMEAN	95% Confidence Limits	
A	55.972	51.285	60.659
B	69.478	66.257	72.699
C	102.022	98.071	105.972
D	88.211	84.462	91.961

CONFIDENCE INTERVALS FOR PAIRWISE DIFFERENCES: Obtaining a confidence interval for $\mu_i - \mu_{i'}$ can help us assess whether or not the treatment means μ_i and $\mu_{i'}$ are different. A natural point estimator for $\mu_i - \mu_{i'}$, based on the ANCOVA model, is given by

$$\text{Adj. } \overline{Y}_{i+} - \text{Adj. } \overline{Y}_{i'+} = (\overline{Y}_{i+} - \overline{Y}_{i'+}) - \widehat{\gamma}(\overline{x}_{i+} - \overline{x}_{i'+}).$$

It is not overly difficult to show that

$$\sigma^2_{\text{Adj.}\overline{Y}_{i+} - \text{Adj.}\overline{Y}_{i'+}} \equiv V(\text{Adj. } \overline{Y}_{i+} - \text{Adj. } \overline{Y}_{i'+}) = \sigma^2 \left[\frac{1}{n_i} + \frac{1}{n_{i'}} + \frac{(\overline{x}_{i+} - \overline{x}_{i'+})^2}{E_{xx}} \right],$$

where $E_{xx} = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_{i+})^2$. An estimate of this variance is given by

$$\widehat{\sigma}^2_{\text{Adj.}\overline{Y}_{i+}-\text{Adj.}\overline{Y}_{i'+}} = \text{MS[E]}\left[\frac{1}{n_i} + \frac{1}{n_{i'}} + \frac{(\overline{x}_{i+} - \overline{x}_{i'+})^2}{E_{xx}}\right],$$

where MS[E] is the mean-squared error from the ANCOVA model. Thus, a $100(1-\alpha)$ percent confidence interval for $\mu_i - \mu_{i'}$, based on the adjusted treatment means, is given by

$$(\text{Adj. }\overline{Y}_{i+} - \text{Adj. }\overline{Y}_{i'+}) \pm t_{N-p-1,\alpha/2} \times \widehat{\sigma}_{\text{Adj.}\overline{Y}_{i+}-\text{Adj.}\overline{Y}_{i'+}},$$

where $N-p-1$ denotes the error degrees of freedom from the ANCOVA model fit. SAS provides these confidence intervals on request.

Example 8.10 (`pigs.sas`). For the pig data in Example 8.6, confidence intervals for the pairwise differences of treatment means (using the adjusted estimates) are given in the following table:

```
                       95% Confidence Limits for
i  j  Differences        LSMean(i) - LSMean(j)
-----------------------------------------------
1  2      -13.506     -19.296        -7.715
1  3      -46.049     -53.029       -39.070
1  4      -32.239     -39.251       -25.227
2  3      -32.543     -37.584       -27.502
2  4      -18.733     -23.608       -13.858
3  4       13.810       8.976        18.643
```

REMARK: Just because we are discussing a different model (ANCOVA) doesn't mean that we still don't have to worry about multiple comparisons! If we want to form **simultaneous** confidence intervals for **all pairwise differences**, then we need to adjust for multiplicity by using an adjusted critical value instead of using $t_{N-p-1,\alpha/2}$. For pairwise differences, I would use Tukey's procedure (exactly how one would use it in the one-way layout without covariates).

ANCOVA AS A SPECIAL CASE OF REGRESSION: It turns out that we can reparameterise ANCOVA models to make them "look like" regression models. Consider the means version of our ANCOVA model $Y_{ij} = \mu_i + \gamma(x_{ij} - \overline{x}_{++}) + \epsilon_{ij}$, for $i = 1, 2, ..., t$ and $j = 1, 2, ..., n_i$ and define

$$z_{ij} = \begin{cases} 1, & \text{if } i\text{th individual receives treatment } j \\ 0, & \text{otherwise,} \end{cases}$$

for $j = 1, 2, ..., t$ and $i = 1, 2, ..., N$, where $N = \sum_i n_i$ denotes the total number of observations. We can write the **means** version of our ANCOVA model as

$$Y_i = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \cdots + \beta_{t-1} z_{i(t-1)} + \gamma(x_i - \overline{x}_+) + \epsilon_i,$$

for $i = 1, 2, ..., N$, where x_i is the covariate value for individual i and $\overline{x}_+$ is the mean of all of the covariate values. Likewise, the **effects** version of our ANCOVA model $Y_{ij} = \mu + \tau_i + \gamma(x_{ij} - \overline{x}_{++}) + \epsilon_{ij}$, for $i = 1, 2, ..., t$ and $j = 1, 2, ..., n_i$, can be expressed as

$$Y_i = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \cdots + \beta_t z_{it} + \gamma(x_i - \overline{x}_+) + \epsilon_i,$$

for $i = 1, 2, ..., N$, where x_i and $\overline{x}_+$ are defined as above.

REPARAMETERISED MODELS: As you can see, in the means model reparameterisation $Y_i = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \cdots + \beta_{t-1} z_{i(t-1)} + \gamma(x_i - \overline{x}_+) + \epsilon_i$, we have left off the indicator variable z_t (for the last treatment) because it is not needed. This leads to a full-rank design matrix. In the effects model reparameterisation $Y_i = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \cdots + \beta_t z_{it} + \gamma(x_i - \overline{x}_+) + \epsilon_i$, the inclusion of the indicator variable z_t leads to a design matrix that is not full rank (verify!). All this means is that $\widehat{\boldsymbol{\beta}}$ can not be computed uniquely.

IMPORTANT NOTE: Both the means and effects ANCOVA models, and their reparameterisations all have the same ANOVA tables! This follows because $C(\boldsymbol{X})$ for all four design matrices is the same. Thus, it doesn't matter which parametrisation you use; the analysis will always be the same.

REALISATION: We should see now that ANCOVA (with a single covariate) is really just a special case of parallel-lines regression, where each line corresponds to the relationship between Y and x for treatment i. Thus, all of the theory we developed for parallel-lines regression holds here in the ANCOVA setting as well. There is a subtle difference in the interpretation. In a parallel-lines regression setting, we are primarily focused on the relationship between Y and x, and the categorical variable z is used because we believe the relationship between Y and x is different for different levels of z. In an ANCOVA setting, we are interested in comparing the means of Y at the different levels of the treatment indicator z; however, the covariate x is used because we believe that it sharpens the analysis (i.e., it makes the comparison of the treatment means more precise).

ANCOVA MODEL: UNEQUAL SLOPES: A slight variant of our previous ANCOVA regression model is one that allows for different slopes; i.e.,

$$Y_{ij} = \mu + \tau_i + \gamma_i(x_{ij} - \overline{x}_{++}) + \epsilon_{ij},$$

for $i = 1, 2, ..., t$ and $j = 1, 2, ..., n_i$. This model may be appropriate when the treatments and the covariates **interact** with each other. Even though we are allowing for different slopes, this model still falls in the $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ formulation! To illustrate this, suppose that $t = 3$ and that $n_1 = n_2 = n_3 = 2$. The design matrix $\boldsymbol{X}$ and parameter vector $\boldsymbol{\beta}$, for the unequal-slopes ANCOVA model, are given by

$$\boldsymbol{X} = \begin{pmatrix} 1 & 1 & 0 & 0 & x_{11} - \overline{x}_{++} & 0 & 0 \\ 1 & 1 & 0 & 0 & x_{12} - \overline{x}_{++} & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & x_{21} - \overline{x}_{++} & 0 \\ 1 & 0 & 1 & 0 & 0 & x_{22} - \overline{x}_{++} & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & x_{31} - \overline{x}_{++} \\ 1 & 0 & 0 & 1 & 0 & 0 & x_{32} - \overline{x}_{++} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \\ \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{pmatrix},$$

respectively. For the unequal-slopes ANCOVA model, the **adjusted treatment mean** is given by

$$\text{Adj. } \overline{Y}_{i+} = \overline{Y}_{i+} - \widehat{\gamma}_i(\overline{x}_{i+} - \overline{x}_{++}),$$

where $\widehat{\gamma}_i$ is the least-squares estimator of γ_i. You will note that this is the same expression for Adj. $\overline{Y}_{i+}$ as before, except that $\widehat{\gamma}_i$ replaces $\widehat{\gamma}$ (this change accommodates for the unequal slopes).

TESTING FOR HOMOGENEOUS SLOPES: In an ANCOVA setting, it is often of interest to determine whether or not the slope coefficients are equal; that is, we would like to test $H_0 : \gamma_1 = \gamma_2 = \cdots = \gamma_t = \gamma$. Here, H_0 corresponds to the situation wherein the treatments and covariate do not interact. To answer this question, we can perform a reduced-versus-full model test! In this setting, our smaller model is the equal-slopes version and the full model is the non-equal-slopes version; i.e.,

$$\begin{aligned} H_0 : &\quad Y_{ij} = \mu + \tau_i + \gamma(x_{ij} - \overline{x}_{++}) + \epsilon_{ij} \quad \text{(reduced model)} \\ H_1 : &\quad Y_{ij} = \mu + \tau_i + \gamma_i(x_{ij} - \overline{x}_{++}) + \epsilon_{ij} \quad \text{(full model).} \end{aligned}$$

If we let $\boldsymbol{M}_0$ and $\boldsymbol{M}$ denote the reduced and full model hat matrices, we know to base our decision on the size of $\text{SS[R]}_F - \text{SS[R]}_R = \boldsymbol{Y}'(\boldsymbol{M} - \boldsymbol{M}_0)\boldsymbol{Y}$, or equivalently, on the size of

$$F = \frac{\boldsymbol{Y}'(\boldsymbol{M} - \boldsymbol{M}_0)\boldsymbol{Y}/(t-1)}{\text{MS[E]}_F}$$

Fortunately, fitting both models (i.e., the parallel-lines model and the non-parallel-lines model) is easy using SAS so we can easily compute F.

Example 8.11 (`pigs.sas`). For the pig data in Example 8.6, we would like to determine whether or not the treatments (diets) and the covariate (initial weight) interact. As noted earlier, this corresponds to testing $H_0 : \gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = \gamma$ using a reduced-versus-full model approach. Here are the two ANOVA tables:

ANCOVA: Full model

Source	df	SS	MS	F
Model	7	4502.569	643.22	99.76
Error	11	142.021	12.911	
Total	18	4644.589		

ANCOVA: Reduced Model

Source	df	SS	MS	F
Model	4	4487.167	1121.792	99.76
Error	14	157.423	11.244	
Total	18	4644.589		

It is easy to compute $\boldsymbol{Y}'(\boldsymbol{M}-\boldsymbol{M}_0)\boldsymbol{Y} = \text{SS[R]}_F - \text{SS[R]}_R = 4502.569 - 4487.167 = 15.402$. Thus, the F statistic for this reduced-versus-full model test is given by

$$F = \frac{\boldsymbol{Y}'(\boldsymbol{M}-\boldsymbol{M}_0)\boldsymbol{Y}/(t-1)}{\text{MS[E]}_F} = \frac{15.402/(4-1)}{12.911} = 0.398.$$

This F is not large, so we would not reject H_0. The smaller model which uses equal slopes is appropriate for these data; that is, it doesn't appear as though the diets and initial weight interact.

REMARK: We have talked about the adjusted treatment means in the unequal slopes ANCOVA model; i.e., Adj. $\overline{Y}_{i+} = \overline{Y}_{i+} - \widehat{\gamma}_i(\overline{x}_{i+} - \overline{x}_{++})$. Just as we did in the equal-slopes ANCOVA model, it is possible to derive confidence intervals for μ_i and confidence intervals for pairwise differences using Adj. $\overline{Y}_{i+}$ (the formulae change slightly from the equal slopes case).

REMARK: It is also possible to include more than one covariate! For example, with the pig experiment from Example 8.6, we could have used x_1 = initial weight, x_2 = average body temperature, x_3 = average heart rate, etc. The extension of the ANCOVA formulation to handle multiple covariates is straightforward. Also, the use of covariates is not limited to a one-way layout with the usual one-way ANOVA analysis. Covariates may also be used in with other designs and treatment structures (e.g., randomised complete block designs, factorial treatment structures, Latin square designs, etc.).

REMARK: As we have seen with the pig data in Example 8.6, the use of the covariate (initial weight) can help sharpen the comparison among the different treatments (diets). I can not overemphasise, however, the importance of thinking about covariates like initial weight, and other possible covariates, before the experiment is designed and the data are collected. This is an important way of thinking about the **design of experiments**; namely, to identify all possible sources of variation *beforehand*, and then design the experiment to incorporate these sources. As mentioned earlier, ANCOVA can be thought of a "salvage effort." I say this because, with the pig data, for example, had we identified initial weight as a possible source of variation beforehand, we could have incorporated that into the design by using initial weight as a blocking factor.

9 Factorial Treatment Structures: Part I

Complimentary reading from Rao: Chapter 13.1-13.5.

Factorial treatment structures are simply an efficient way of defining treatments in an experiment. They can be used in any of the standard experimental designs, such as completely randomised designs (CRD), randomised complete block designs (RCBD), Latin square designs, split-plot designs, etc. Up until now, we have discussed the notion of **complete randomisation** (i.e., individuals are randomly assigned to treatments under no restriction). Any design that uses complete randomisation is called a **completely randomised design**. RCBD's and split plot designs will be discussed later.

9.1 Introduction

To motivate the use of factorial treatment structures in experimental design, we will consider the following example.

Example 9.1. The effect of alcohol and sleeping pills taken together is much greater than one would suspect based on examining the effects of alcohol and sleeping pills separately. Suppose that we did two experiments:

- one experiment involves 20 subjects to establish the effect of a normal dose of alcohol.
- the other experiment involves 20 subjects to establish the effect of a normal dose of sleeping pills.

Note that there are 40 subjects needed for the two experiments. When considering both, a common fallacy is that the effect of taking a normal dose of both alcohol and sleeping pills would be just the sum of the individual effects. Unfortunately, these two separate experiments provide no basis for either accepting or rejecting such a conclusion.

USING A FACTORIAL TREATMENT STRUCTURE: We can redesign the investigation to be more *informative* and *efficient* by using a **factorial treatment structure**.

- The alcohol experiment would involve 10 people getting no alcohol (a_1) and 10 people getting a normal dose of alcohol (a_2).
- The sleeping pill experiment would involve 10 people getting no sleeping pills (b_1) and 10 people getting a normal dose of sleeping pills (b_2).

The two **factors** in this study are alcohol (A) and sleeping pills (B). Each factor has two **levels**, no drug (a_1 and b_1, respectively) and normal dose (a_2 and b_2, respectively). *A factorial treatment structure uses treatments that are all combinations of the different levels of the factors.* A factorial experiment to study alcohol and sleeping pills may have

- 5 people given no alcohol and no sleeping pills (a_1b_1)
- 5 people given no alcohol but a normal dose of sleeping pills (a_1b_2)
- 5 people given a normal dose of alcohol but no sleeping pills (a_2b_1)
- 5 people given a normal dose of alcohol and a normal dose of sleeping pills (a_2b_2).

ADVANTAGES OF USING A FACTORIAL TREATMENT STRUCTURE: Assigning treatments in this way has two major advantages:

1. A factorial treatment structure is more **informative** in that it provides evidence about the effect of taking alcohol **and** sleeping pills *together*; that is, it informs us as to whether or not alcohol and sleeping pills **interact**. If the factors interact,
   - the effect of alcohol depends on whether the person has taken sleeping pills
   - the effect of sleeping pills depends on whether the person has taken alcohol.

   Note that if the two factors interact, the separate experiments described earlier have very little value.

2. If the factors, in fact, do **not** interact, a factorial treatment structure is more **efficient**. The two experiments use a total of 40 subjects, and the factorial experiment uses 20 subjects. Yet, *the factorial experiment contains the same amount of information as the two experiments* since:

    - the effect of alcohol can be studied by contrasting the 5 a_1b_1 people with the 5 a_2b_1 people, and also by comparing the 5 a_1b_2 people with the 5 a_2b_2 people. Thus, we have a total of 10 no alcohol people and 10 alcohol people, just as we did in the separate experiment for alcohol.
    - the effect of sleeping pills can be studied by contrasting the 5 a_1b_1 people with the 5 a_1b_2 people, and also by comparing 5 a_2b_1 people with the 5 a_2b_2 people. Thus, we have a total of 10 no sleeping pills people and 10 sleeping pills people, just as we did in the separate experiment for sleeping pills.

NOTATION: A useful notation for factorial experiments identifies the number of factors and the number of levels of each factor. For example, the alcohol-sleeping pill study is best described as a 2 × 2 factorial experiment, for a total of 4 treatments. If we had three levels of alcohol and 4 doses (levels) of sleeping pills, we would have a 3 × 4 factorial, for a total of 12 treatments. If we had another factor, say, diet (Factor C), with 3 levels (and alcohol and sleeping pills had their original 2 levels), we would have a 2 × 2 × 3 factorial, for a total of 12 treatments.

IN GENERAL: The number of treatments in an $a \times b$ factorial experiment is ab, where a is the number of levels of factor A, and b is the number of levels of factor B. The number of treatments in an $a \times b \times c$ factorial experiment is abc, and so on.

QUESTIONS OF INTEREST: Instead of just determining whether or not there are differences among treatments, we now have the capability of investigating the **effects** of the different factors (which make up the treatments). For example, we may want to investigate whether or not (a) there is an effect due to alcohol, (b) there is an effect due to sleeping pills, and/or (c) alcohol and sleeping pills interact.

9.2 Factorial effects in 2×2 experiments

Example 9.2 (corn.sas) Corn must have adequate, yet efficient, amounts of nitrogen (A) and phosphorus (B) for profitable production and environmental concerns. In a 2×2 factorial experiment, two levels of nitrogen ($a_1 = 10$ and $a_2 = 15$) and two levels of phospherous were used ($b_1 = 2$ and $b_2 = 4$). Applications of nitrogen and fertilizer were measured in pounds per plot. Twenty small (quarter acre) plots were available for experimentation, and the four **treatment combinations** a_1b_1, a_1b_2, a_2b_1, and a_2b_2 were randomly assigned to plots. *This is a completely randomised design with a 2×2 factorial treatment structure.* The response is Y_{ijk}, the yield (in pounds), after harvest, for the kth plot receiving the ith level of nitrogen and the jth level of phosphorous. Here, $i = 1, 2$, $j = 1, 2$, and $k = 1, 2, ..., 5$. Data from the experiment are given in Table 9.21.

Table 9.21: *Corn yield data for nitrogen and phosphorous applications.*

Treatment Combination	Y_{ijk}	$\overline{Y}_{ij+}$	$\sum_{k=1}^{5}(Y_{ijk} - \overline{Y}_{ij+})^2$
a_1b_1	35, 26, 25, 33, 31	30	76
a_1b_2	39, 33, 41, 31, 36	36	68
a_2b_1	37, 27, 35, 27, 34	32	88
a_2b_2	49, 39, 39, 47, 46	44	88

PRELIMINARY ANOVA: Consider the table of **treatment totals**:

	b_1	b_2
a_1	150	180
a_2	160	220

NOTE: In a 2×2 factorial experiment, we really have 4 treatments; i.e., a_1b_1, a_1b_2, a_2b_1, and a_2b_2. Thus, we can construct the ANOVA table treating this as a one-way layout with four treatment means $\mu_{11}, \mu_{12}, \mu_{21}$, and μ_{22} (recall computing formulae from Chapter 2). First, the correction term is given by

$$\text{CM} = \frac{1}{20}Y_{+++}^2 = 710^2/20 = 25205.$$

This helps us get the (corrected) total sum of squares; i.e.,

$$\begin{aligned} \text{SS[TOT]} &= \boldsymbol{Y}'\boldsymbol{Y} - \text{CM} \\ &= \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{5} Y_{ijk}^2 - \text{CM} \\ &= 26100 - 25205 = 895. \end{aligned}$$

The treatment sums of squares is given by

$$\begin{aligned} \text{SS[T]} &= \frac{1}{5}\sum_{i=1}^{2}\sum_{j=1}^{2} Y_{ij+}^2 - \text{CM} \\ &= \frac{1}{5}(150^2 + 160^2 + 180^2 + 210^2) - 25205 = 575. \end{aligned}$$

Thus, we get SS[E] by subtraction; i.e., SS[E] = SS[TOT] − SS[T] = 320. Also, note that $\text{SS[E]} = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{5}(Y_{ijk} - \overline{Y}_{ij+})^2 = 76 + 68 + 88 + 88 = 320$, from Table 9.21. Thus, viewing this as a one-way layout experiment with four treatments, our ANOVA table becomes

Table 9.22: *Analysis of variance: Corn data.*

Source	df	SS	MS	F
Treatments	3	575	191.7	9.6
Error	16	320	20	
Total	19	895		

TEMPORARY CONCLUSION: Since our F statistic is large, e.g., $F_{3,16,0.05} = 3.239$, we would conclude that at least one of the treatment means is different.

REMARK: As before, the overall F test provides very little information. However, with a factorial treatment structure, it is possible to explore the data a little more; in particular, we can assess whether or not there are **effects** due to nitrogen (Factor A), phosphorous (Factor B) or an **interaction** between nitrogen and phosphorous. It turns out that, in factorial experiments, these effects may be represented by **contrasts**.

DIFFERENT TYPES OF EFFECTS: There are three types of effects in a factorial experiment: (a) simple effects, (b) main effects, and (c) interaction effects. We now examine each in the context of a 2×2 experiment. In what follows, we will assume factor A has two levels a_1 and a_2 and factor B has two levels b_1 and b_2. Thus, there are really four **treatments**, and, hence, four **treatment means** $\mu_{11}, \mu_{12}, \mu_{21}$, and μ_{22} (sometimes these are called the **cell means**). As usual, we will assume that our response variable Y follows a normal distribution with constant variance σ^2 (so that we can construct F statistics to test for the presence of main and interaction effects).

Table 9.23: *Population means and simple effects in a 2×2 factorial experiment.*

	Factor B		Simple effect of B
Factor A	b_1	b_2	$\mu[A_iB]$
a_1	μ_{11}	μ_{12}	$\mu[A_1B] = \mu_{12} - \mu_{11}$
a_2	μ_{21}	μ_{22}	$\mu[A_2B] = \mu_{22} - \mu_{21}$
Simple effect of A	$\mu[AB_1] = \mu_{21} - \mu_{11}$	$\mu[AB_2] = \mu_{22} - \mu_{12}$	

SIMPLE EFFECTS: In a 2×2 factorial experiment, there are four **simple effects**. The simple effect of A, at level b_1 of B, is defined as

$$\mu[AB_1] = \mu_{21} - \mu_{11}.$$

In words, this simple effect is the **change** in mean (i.e., the expected response) when A changes level and the level of B is held fixed at b_1. Similarly, the simple effect of A, at level b_2 of B, is defined as

$$\mu[AB_2] = \mu_{22} - \mu_{12}.$$

Simple effects $\mu[A_1B]$ and $\mu[A_2B]$ are defined analogously; i.e., $\mu[A_1B] = \mu_{12} - \mu_{11}$ and $\mu[A_2B] = \mu_{22} - \mu_{21}$.

ESTIMATING SIMPLE EFFECTS: Estimates of the simple effects are obtained by replacing μ_{ij} with $\overline{Y}_{ij+}$; that is,

$$\widehat{\mu[AB_1]} = \overline{Y}_{21+} - \overline{Y}_{11+} \qquad \widehat{\mu[AB_2]} = \overline{Y}_{22+} - \overline{Y}_{12+}$$
$$\widehat{\mu[A_1B]} = \overline{Y}_{12+} - \overline{Y}_{11+} \qquad \widehat{\mu[A_2B]} = \overline{Y}_{22+} - \overline{Y}_{21+}.$$

Table 9.24: *Table of treatment means and simple effects estimates for the corn data.*

	b_1	b_2	Difference	Factor B Means
a_1	30	36	$\overline{Y}_{12+} - \overline{Y}_{11+} = 6$	$\overline{Y}_{1++} = 33$
a_2	32	44	$\overline{Y}_{22+} - \overline{Y}_{21+} = 12$	$\overline{Y}_{2++} = 38$
Difference	$\overline{Y}_{21+} - \overline{Y}_{11+} = 2$	$\overline{Y}_{22+} - \overline{Y}_{12+} = 8$		
Factor A Means	$\overline{Y}_{+1+} = 31$	$\overline{Y}_{+2+} = 40$		

Example 9.2 (*continued*). With the corn data from Example 9.2, the simple effects estimates are $\widehat{\mu[AB_1]} = 2$, $\widehat{\mu[AB_2]} = 8$, $\widehat{\mu[A_1B]} = 6$, and $\widehat{\mu[A_2B]} = 12$.

MAIN EFFECTS: In a 2 × 2 factorial experiment, there are two **main effects**.

$$\begin{aligned}\mu[A] &= \frac{1}{2}\left(\mu[AB_1] + \mu[AB_2]\right) = \frac{1}{2}(\mu_{21} - \mu_{11} + \mu_{22} - \mu_{12}) \\ \mu[B] &= \frac{1}{2}\left(\mu[A_1B] + \mu[A_2B]\right) = \frac{1}{2}(\mu_{12} - \mu_{11} + \mu_{22} - \mu_{21}).\end{aligned}$$

The main effect of A is the average change in the expected response when the level of A is changed from a_1 to a_2. Likewise, the main effect of B is the average change in the expected response when the level of B is changed from b_1 to b_2. Note that $\mu[A]$ and $\mu[B]$ are just **contrasts** in $\mu_{11}, \mu_{12}, \mu_{21}$, and μ_{22}.

ESTIMATING MAIN EFFECTS: Estimates of the main effects are obtained by replacing μ_{ij} with $\overline{Y}_{ij+}$; that is,

$$\begin{aligned}\widehat{\mu[A]} = \frac{1}{2}\left(\widehat{\mu[AB_1]} + \widehat{\mu[AB_2]}\right) &= \frac{1}{2}(\overline{Y}_{21+} - \overline{Y}_{11+} + \overline{Y}_{22+} - \overline{Y}_{12+}) \\ &= \overline{Y}_{2++} - \overline{Y}_{1++}\end{aligned}$$

$$\begin{aligned}\widehat{\mu[B]} = \frac{1}{2}\left(\widehat{\mu[A_1B]} + \widehat{\mu[A_2B]}\right) &= \frac{1}{2}(\overline{Y}_{12+} - \overline{Y}_{11+} + \overline{Y}_{22+} - \overline{Y}_{21+}) \\ &= \overline{Y}_{+2+} - \overline{Y}_{+1+}.\end{aligned}$$

Example 9.2 (*continued*). With the corn data from Example 9.2, the main effects estimates are $\widehat{\mu[A]} = 5$ and $\widehat{\mu[B]} = 9$. What would you expect these estimates to be "close to" if neither nitrogen nor phosphorous had an effect on yield?

INTERACTION EFFECTS: If the difference in response between the levels of one factor is not the same at all levels of the other factor, then we say the factors **interact**; thus, in a 2 × 2 factorial experiment, there is one (pairwise) **interaction effect**; i.e.,

$$\begin{aligned}\mu[AB] &= \frac{1}{2}(\mu[AB_2] - \mu[AB_1]) = \frac{1}{2}(\mu[A_2B] - \mu[A_1B]) \\ &= \frac{1}{2}(\mu_{22} - \mu_{12} - \mu_{21} + \mu_{11}).\end{aligned}$$

This effect measures differences between the simple effects of one factor at different levels of another factor. Note that $\mu[AB]$ is a **contrast** in $\mu_{11}, \mu_{12}, \mu_{21}$, and μ_{22}.

ESTIMATING THE INTERACTION EFFECT: The **estimate** of the interaction effect is obtained by replacing μ_{ij} with $\overline{Y}_{ij+}$; that is,

$$\widehat{\mu[AB]} = \frac{1}{2}\left(\overline{Y}_{22+} - \overline{Y}_{12+} - \overline{Y}_{21+} + \overline{Y}_{11+}\right).$$

Example 9.2 (*continued*). With the corn data from Example 9.2, the interaction effect estimate is $\widehat{\mu[AB]} = 3$. What would you expect this estimate to be "close to" if nitrogen and phosphorous did not interact?

INTERACTION PLOTS: A nice graphical display to help us assess whether or not two factors interact is an **interaction plot**. In such a plot, the levels of Factor A (say) are marked on the horizontal axis. The sample means $\overline{Y}_{11+}$, $\overline{Y}_{12+}$, $\overline{Y}_{21+}$, and $\overline{Y}_{22+}$ are plotted against the levels of A, and the points corresponding to the same level of Factor B are joined by straight lines. The interaction plot for the corn data in Example 9.2 appears in Figure 9.27.

INTERPRETING INTERACTION PLOTS: When the interaction plot displays **parallel lines**; i.e., the difference between the sample means is the same at a_1 as the difference between the sample means at a_2, this suggests that there is **no interaction** between the factors. Of course, perfect parallelism is a rarity in real problems. There is likely to be some interaction suggested by the data, even if A and B truly do not interact. The key question is whether or not the interaction term $\mu[AB]$ is significantly different from zero! The interaction plot can be very helpful in *visually assessing* the degree of interaction. See Figure 13.1 (p. 593, Rao) for some examples of interaction plots.

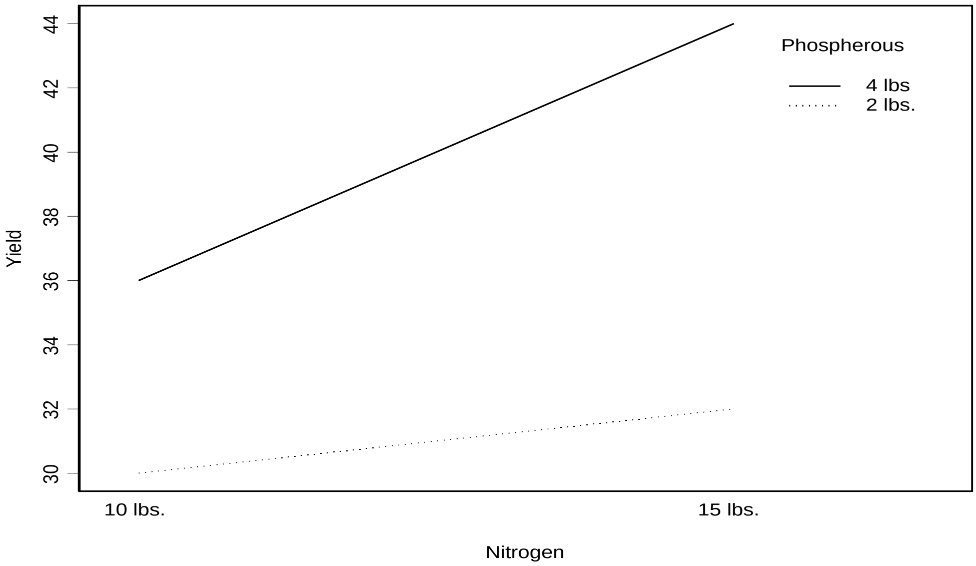

Figure 9.27: *Interaction plot for the corn data from Example* 9.2.

MAIN AND INTERACTION EFFECTS AS CONTRASTS: In a 2×2 factorial experiment, we have four treatment means μ_{11}, μ_{12}, μ_{21}, and μ_{22}. *If the experiment is balanced*, the main effects and interaction effect can be written as **contrasts**; that is,

$$\begin{aligned}
\mu[A] &= \frac{1}{2}(\mu_{21} - \mu_{11} + \mu_{22} - \mu_{12}) \\
\mu[B] &= \frac{1}{2}(\mu_{12} - \mu_{11} + \mu_{22} - \mu_{21}). \\
\mu[AB] &= \frac{1}{2}(\mu_{22} - \mu_{12} - \mu_{21} + \mu_{11}).
\end{aligned}$$

In addition, these are mutually **orthogonal** contrasts! See Table 9.25.

Table 9.25: *Table of contrast coefficients for a* 2×2 *factorial.*

Effect	c_{11}	c_{12}	c_{21}	c_{22}
$\mu[A]$	$-\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
$\mu[B]$	$-\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$
$\mu[AB]$	$\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$

PUNCHLINE: In a 2×2 (balanced) factorial experiment, we can "break down" the treatment sum of squares SS[T] into the sums of squares for **three** orthogonal components; namely, the sum of squares for the contrast $\mu[A]$, the sum of squares for the contrast $\mu[B]$, and the sum of squares for the contrast $\mu[AB]$. Let n_{ij} denote the number of replications at the $a_i b_j$ treatment combination. By "balanced," I mean that $n_{ij} = n$ for all i and j. Let $N = n_{11} + n_{12} + n_{21} + n_{22}$ denote the number of observations observed. The general form of the ANOVA in a 2×2 factorial experiment looks like

Table 9.26: *ANOVA table for the* 2×2 *factorial experiment.*

Source	df	SS	MS	F
A	1	SS[A]	MS[A]	$F_A = \frac{\text{MS[A]}}{\text{MS[E]}}$
B	1	SS[B]	MS[B]	$F_B = \frac{\text{MS[B]}}{\text{MS[E]}}$
AB	1	SS[AB]	MS[AB]	$F_{AB} = \frac{\text{MS[AB]}}{\text{MS[E]}}$
Error	$N-4$	SS[E]	MS[E]	
Total	$N-1$	SS[TOT]		

COMPUTING THE SUMS OF SQUARES: From Chapter 3, recall that the **sum of squares** for a general contrast estimate $\widehat{\theta}$ was given by

$$\text{SS}(\widehat{\theta}) = \frac{\widehat{\theta}^2}{\sum_{i=1}^{t} \frac{c_i^2}{n_i}}.$$

We can use this definition to formulate the sums of squares for main effects and the interaction effect (these formulae are appropriate for balanced or unbalanced experiments):

$$\begin{aligned}
\text{SS[A]} &= \frac{4(\widehat{\mu[A]})^2}{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} \\
\text{SS[B]} &= \frac{4(\widehat{\mu[B]})^2}{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} \\
\text{SS[AB]} &= \frac{4(\widehat{\mu[AB]})^2}{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}.
\end{aligned}$$

Example 9.2. (*continued*). We now compute the sums of squares for the main effects and interaction effect. Recall that $\widehat{\mu[A]} = 5$, $\widehat{\mu[B]} = 9$, and $\widehat{\mu[AB]} = 3$. Also, $n_{ij} = 5$ for all i and j. Thus,

$$\begin{aligned}
\text{SS[A]} &= \frac{4(5)^2}{\frac{1}{5}+\frac{1}{5}+\frac{1}{5}+\frac{1}{5}} = 125\\
\text{SS[B]} &= \frac{4(9)^2}{\frac{1}{5}+\frac{1}{5}+\frac{1}{5}+\frac{1}{5}} = 405\\
\text{SS[AB]} &= \frac{4(3)^2}{\frac{1}{5}+\frac{1}{5}+\frac{1}{5}+\frac{1}{5}} = 45.
\end{aligned}$$

REALISATION: Our ANOVA table that we obtained previously in Table 9.22, from viewing this as a one-way layout with four treatments, can be expressed as

Source	df	SS	MS	F
A	1	125	125	6.25
B	1	405	405	20.25
AB	1	45	45	2.25
Error	16	320	20	
Total	19	895		

You will note that all we have done is broken down the treatment sums of squares SS[T] from the one-way layout into **orthogonal components**; namely, the sums of squares for A, B, and the interaction AB. In particular, note that $\text{SS[T]} = \text{SS[A]} + \text{SS[B]} + \text{SS[AB]}$. When the design is unbalanced, we lose orthogonality, but this general sum of squares breakdown still holds (we'll talk more about this later).

TESTING MAIN AND INTERACTION EFFECTS: For a general contrast θ, recall that when $H_0 : \theta = 0$ is true, $\text{SS}(\widehat{\theta})/\sigma^2 \sim \chi^2_1$. Thus,

$$F_A = \frac{\text{MS[A]}}{\text{MS[E]}} = \frac{\text{SS[A]}/\sigma^2}{\frac{\text{SS[E]}}{\sigma^2}/(N-4)} \sim F_{1,N-4},$$

$$F_B = \frac{\text{MS[B]}}{\text{MS[E]}} = \frac{\text{SS[B]}/\sigma^2}{\frac{\text{SS[E]}}{\sigma^2}/(N-4)} \sim F_{1,N-4},$$

$$F_{AB} = \frac{\text{MS[AB]}}{\text{MS[E]}} = \frac{\text{SS[AB]}/\sigma^2}{\frac{\text{SS[E]}}{\sigma^2}/(N-4)} \sim F_{1,N-4}.$$

The F statistics F_A, F_B, and F_{AB} can be used to test for main effects and an interaction effect, respectively.

GENERAL STRATEGY FOR A FACTORIAL ANALYSIS: The following strategies are common when analysing data from 2 × 2 factorial experiments. In fact, more generally, these are common strategies when analysing data from $a \times b$ factorial experiments (I'll take this general approach, since the 2 × 2 is just a special case).

- Start by looking at whether or not the interaction term $\mu[AB]$ is significantly different from zero. This is done by using F_{AB}.
- **If $\mu[AB]$ is significantly different from zero**, then tests for main effects are less meaningful because their interpretations depend on the interaction. In this situation, the easiest approach is to just do the entire analysis as a one-way ANOVA (recall Chapters 2 and 3) with ab treatments.
  - In this case, you could get pairwise confidence intervals for all ab means. These intervals could help you formulate an ordering among the ab treatment means. In many experiments, it is of interest to find the "best treatment" (e.g., largest yield, smallest time to recovery, etc).
  - You could also look at various contrasts of the ab means that are of interest to the researcher. In either situation (pairwise contrasts or other contrasts) the methods of Chapter 3 apply.
- **If $\mu[AB]$ is not significantly different from zero**, then it is safe to test for main effects (if you want to).
  - Some recommend forming pairwise confidence intervals for the different means of each factor. For example, in a 3 × 4 factorial experiment, suppose that the AB interaction was not significant. In this case, we could construct pairwise intervals for the 3 A means and pairwise intervals for the 4 B means. This may give us insight about where the differences truly are within each factor. Of course, if $a = b = 2$, i.e., in a 2 × 2 factorial experiment, then there is only one pairwise interval for A and one for B.

  - Significant differences can also be found by using contrasts within each factor. For example, in our 3×4 experiment, I could compare the first two B means to the last two B means (if this is of interest to the researcher). Again, non-simple contrasts are only possible when $a > 2$ or $b > 2$.

- *If the interaction term is not significant, should I go ahead and formally test for main effects?*

  - Some statisticians wouldn't even bother doing formal tests for main effects. They would argue that forming pairwise intervals (or contrasts) are much more informative than categorically saying "A is significant" or "A is not significant." If the pairwise intervals or contrasts show "no difference," this suggests that the factor is not significant (although it is theoretically possible to not reject $H_0 : A$ not significant, but still find significant differences among the means of levels of Factor A).
  - In many applied areas, testing for main effects may be encouraged (for publication purposes, say). In this situation, one usually tests for main effects, and then forms pairwise intervals (or contrasts) for means of the levels of main effects which are significant.
  - if the interaction is not significant, I usually glance at the main effects tests, but I base the analysis largely on contrasts. It is often the case that the researcher wants to know *how* the means within a factor are different.

REMARK: If the levels of a factor are **quantitatively ordered**, e.g., increasing doses of a drug, increases concentrations of a chemical treatment, row spacings, times of applications, temperatures, etc., some statisticians argue that using means comparisons ignores the logic of the treatment structure. That is, some feel that it is much more important to look at the "dose-response" relationship. To do this, one could plot the response Y versus the levels of a quantitatively-ordered factor and look for an equation that describes the relationship (e.g., straight line, quadratic curve, etc). See Swallow (1984). **Orthogonal polynomial contrasts** may also be useful here.

Example 9.2. (*continued*). The interaction contrast $\mu[AB]$ is not significantly different from zero since $F_{AB} = 2.25 < F_{1,16,0.05} = 4.453$. Thus, it is safe to test for the main effects due to nitrogen (A) and phosphorous (B). We see that both of the contrasts $\mu[A]$ and $\mu[B]$ are significantly different from zero since $F_A = 6.25$ and $F_B = 20.25$ are both large. Thus, nitrogen and phosphorous are both important in describing yield. Since there are only 2 levels of each factor, there is one pairwise interval for $\mu_{A_2} - \mu_{A_1}$ and one pairwise interval for $\mu_{B_2} - \mu_{B_1}$ (both of these intervals will not include zero since A and B were significant). The intervals are given by

$$(\overline{Y}_{2++} - \overline{Y}_{1++}) \pm t_{N-4,\alpha/2}\sqrt{\text{MS[E]}\left(\frac{1}{10} + \frac{1}{10}\right)}$$

and

$$(\overline{Y}_{+2+} - \overline{Y}_{+1+}) \pm t_{N-4,\alpha/2}\sqrt{\text{MS[E]}\left(\frac{1}{10} + \frac{1}{10}\right)},$$

respectively. Note that there 10 observations for each treatment mean $\overline{Y}_{1++}$, $\overline{Y}_{2++}$, $\overline{Y}_{+1+}$, and $\overline{Y}_{+2+}$. Here, we have $\overline{y}_{1++} = 33$, $\overline{y}_{2++} = 38$, $\overline{y}_{+1+} = 31$, $\overline{y}_{+2+} = 40$, MS[E] $= 20$, and $t_{16,0.025} = 2.1199$. Thus, our intervals become

$$(38 - 33) \pm 2.1199 \times \sqrt{20 \times \left(\frac{1}{10} + \frac{1}{10}\right)} \iff (0.76, 9.24)$$

and

$$(40 - 31) \pm 2.1199 \times \sqrt{20 \times \left(\frac{1}{10} + \frac{1}{10}\right)} \iff (4.76, 13.24).$$

These intervals are much more informative than simply saying "A is significant" or "B is significant." It appears as though the high levels of each factor (nitrogen and phosphorous) correspond to a larger mean yield.

REMARK: For the corn data in Example 9.2, it does not appear as though there is a significant interaction between nitrogen and phosphorous. However, *had there been a significant interaction*, the above pairwise intervals are less meaningful because they ignore the interaction effect. If a significant interaction was present, our follow-up analysis could have involved constructing **pairwise intervals** for the four treatment means μ_{11}, μ_{12}, μ_{21}, and μ_{22}.

9.3 Analysing $a \times b$ factorial experiments

The method of analysing data from a 2×2 experiment can be easily extended to an $a \times b$ experiment. Rao carries out this extension in Section 13.4 by generalising the notion of simple, main, and interaction effects using the notation of Section 13.3. While this extension is informative, I find it somewhat unnecessary. Instead, I prefer to take a model-based approach, as in Section 13.5.

NOTATION: Let Y_{ijk} denote the kth response when Factor A is at the ith level; $i = 1, 2, ..., a$, and Factor B is at the jth level; $j = 1, 2, ..., b$. In general, let n_{ij} denote the number of observations made on treatment $a_i b_j$. We will assume, unless otherwise stated, that the design is **balanced**; i.e., $n_{ij} = n$, for all i and j.

TWO-WAY ANOVA MODEL WITH INTERACTION: Data from $a \times b$ factorial experiments can be modelled by a **two-factor ANOVA model with interaction**; i.e.,

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk},$$

for $i = 1, 2, ..., a$, $j = 1, 2, ..., b$, and $k = 1, 2, ..., n$, where $\epsilon_{ijk} \sim \text{iid}\, \mathcal{N}(0, \sigma^2)$. Here, μ denotes the overall mean (in the absence of any treatments), α_i represents the effect due to the ith level of A, β_j represents the effect due to the jth level of B, and $(\alpha\beta)_{ij}$ represents the interaction effect due to the ith and jth levels of A and B, respectively.

TESTS FOR MAIN EFFECTS AND INTERACTIONS: Taking a model-based approach lends nicely to writing hypotheses for factorial effects. For example, to determine whether or not there is an **interaction effect**, we can test

$$H_0 : (\alpha\beta)_{ij} = 0 \text{ for all } i \text{ and } j$$

versus

$$H_1 : \text{not } H_0.$$

Similarly, to determine whether or not there are **main effects**, one can test the hypotheses $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$ and $H_0 : \beta_1 = \beta_2 = \cdots = \beta_b = 0$.

REVELATION: The two-factor model with interaction can be expressed in the form $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. To help see this, suppose that $a = 3$, $b = 2$, and $n_{ij} = n = 3$. In matrix terms, we write

$$\boldsymbol{Y} = \begin{pmatrix} Y_{111} \\ Y_{112} \\ Y_{113} \\ Y_{121} \\ Y_{122} \\ Y_{123} \\ Y_{211} \\ Y_{212} \\ Y_{213} \\ Y_{221} \\ Y_{222} \\ Y_{223} \\ Y_{311} \\ Y_{312} \\ Y_{313} \\ Y_{321} \\ Y_{322} \\ Y_{323} \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \\ (\alpha\beta)_{11} \\ (\alpha\beta)_{12} \\ (\alpha\beta)_{21} \\ (\alpha\beta)_{12} \\ (\alpha\beta)_{31} \\ (\alpha\beta)_{32} \end{pmatrix},$$

and $\boldsymbol{\epsilon} = (\epsilon_{111}, \epsilon_{112}, ..., \epsilon_{323})'$. It is easy to see that $\boldsymbol{X}$ is not a full rank matrix. In fact, $r(\boldsymbol{X}) = 6 < p = 12$ (p is the number of parameters), so there is no hope in estimating $\boldsymbol{\beta}$ uniquely. However, as before in one-way ANOVA model, we can always find a solution to the normal equations; i.e.,

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{Y},$$

where $(\boldsymbol{X}'\boldsymbol{X})^{-}$ is any generalised inverse of $\boldsymbol{X}'\boldsymbol{X}$. This solution is not unique (so the solution is rather arbitrary). As you might suspect, we can find a unique solution if we impose certain **side conditions**. Recall that in the one-way effects model, our design matrix $\boldsymbol{X}$ was rank deficient by one (see Chapter 8); in this situation, appending one side condition did the trick; namely, this led to a unique solution to the normal equations. Here, we are rank deficient by 6. Thus, to solve the normal equations uniquely with the two-factor interaction model, we would need to append 6 side conditions! For commonly-used side conditions in the two-factor model, see Equation 13.16, p. 621, Rao.

ARE SIDE CONDITIONS REALLY ALL THAT IMPORTANT?: Actually, I'm not too big a fan of side conditions. Part of the reason why is that they are arbitrarily chosen to uniquely solve the normal equations $\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}$. *In a sense, they help us solve a mathematical problem that isn't relevant.* The unimportant fact here is that all 12 parameters in the last model can not be uniquely estimated. What *is* important is that there are certain functions of those 12 parameters that can be uniquely estimated, regardless of which side conditions are used. These are the **estimable functions**, and they are the only functions that we should be concerned with.

ANOVA TABLE FOR THE TWO WAY MODEL WITH INTERACTION: In the model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk},$$

for $i = 1, 2, ..., a$, $j = 1, 2, ..., b$, and $k = 1, 2, ..., n$, if we let $\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$, we can rewrite the two-factor model with interaction as a one-way ANOVA model

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk},$$

with ab means. The ANOVA table for this one-way model (i.e., viewing this as a one-way layout with ab **cell means**) is the same as it was from Chapter 2; i.e.,

Source	df	SS	MS	F
Treatments	$ab - 1$	SS[T]	MS[T]	$F = \frac{\text{MS[T]}}{\text{MS[E]}}$
Error	$N - ab$	SS[E]	MS[E]	
Total	$N - 1$	SS[TOT]		

Here, $N = abn$, the total number of observations. The key point to realise is that we can break up the treatment sums of squares SS[T] into components for the main effect of A,

Table 9.27: *ANOVA table for the $a \times b$ factorial experiment.*

Source	df	SS	MS	F
A	$a-1$	SS[A]	MS[A]	$F_A = \frac{\text{MS[A]}}{\text{MS[E]}}$
B	$b-1$	SS[B]	MS[B]	$F_B = \frac{\text{MS[B]}}{\text{MS[E]}}$
AB	$(a-1)(b-1)$	SS[AB]	MS[AB]	$F_{AB} = \frac{\text{MS[AB]}}{\text{MS[E]}}$
Error	$N-ab$	SS[E]	MS[E]	
Total	$N-1$	SS[TOT]		

the main effect of B, and the interaction term AB. In general, this breakdown gives rise to the ANOVA table for the two-way model with interaction; this table is Table 9.27.

BREAKING UP THE TREATMENT SUMS OF SQUARES: The first key equality is

$$\underbrace{\sum_{i=1}^a\sum_{j=1}^b\sum_{k=1}^n(Y_{ijk}-\overline{Y}_{+++})^2}_{\text{SS[TOT]}} = \underbrace{n\sum_{i=1}^a\sum_{j=1}^b(\overline{Y}_{ij+}-\overline{Y}_{+++})^2}_{\text{SS[T]}} + \underbrace{\sum_{i=1}^a\sum_{j=1}^b\sum_{k=1}^n(Y_{ijk}-\overline{Y}_{ij+})^2}_{\text{SS[E]}},$$

which really is nothing new; it is a straightforward extension of the ubiquitous equality SS[TOT] = SS[T] + SS[E], from Chapter 2. The second key equality is that

$$\begin{aligned}\underbrace{n\sum_{i=1}^a\sum_{j=1}^b(\overline{Y}_{ij+}-\overline{Y}_{+++})^2}_{\text{SS[T]}} &= \underbrace{bn\sum_{i=1}^a(\overline{Y}_{i++}-\overline{Y}_{+++})^2}_{\text{SS[A]}} + \underbrace{an\sum_{j=1}^b(\overline{Y}_{+j+}-\overline{Y}_{+++})^2}_{\text{SS[B]}} \\ &+ \underbrace{n\sum_{i=1}^a\sum_{j=1}^b(\overline{Y}_{ij+}-\overline{Y}_{i++}-\overline{Y}_{+j+}+\overline{Y}_{+++})^2}_{\text{SS[AB]}}.\end{aligned}$$

This equality looks rather daunting, but it is really nothing to get too worked up over. The important fact illustrated here is that we are breaking up the treatment sums of squares SS[T] into components for the main effect of A, the main effect of B, and the interaction term AB. You should also note that the degrees of freedom for treatments $ab-1=(a-1)+(b-1)+(a-1)(b-1)$, which is the breakdown for degrees of freedom for A, B, and AB in Table 9.27.

TESTING MAIN AND INTERACTION EFFECTS: As in the 2×2 factorial experiment, we can test for main and interaction effects in $a \times b$ experiments by using F statistics. In particular, the statistic

$$F_A = \frac{\text{MS[A]}}{\text{MS[E]}} \sim F_{a-1,N-ab}.$$

is used to test $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$ (no A main effect) versus H_1 : not H_0. The statistic

$$F_B = \frac{\text{MS[B]}}{\text{MS[E]}} \sim F_{b-1,N-ab},$$

is used to test $H_0 : \beta_1 = \beta_2 = \cdots = \beta_b = 0$ (no B main effect) versus H_1 : not H_0. Finally, the statistic

$$F_{AB} = \frac{\text{MS[AB]}}{\text{MS[E]}} \sim F_{(a-1)(b-1),N-ab}.$$

is used to test $H_0 : (\alpha\beta)_{ij} = 0$ for all i and j (no interaction) versus H_1 : not H_0.

HAND COMPUTATIONS FOR TWO-WAY ANOVA WITH INTERACTION: These formulae make it easier to do computations by hand. First, find the **correction term** for the mean (i.e., for fitting the overall mean μ); this is given by

$$\text{CM} = \frac{1}{N} Y_{+++}^2,$$

where $N = abn$. The total sum of squares, along with those for the main effects and interaction A, B, and AB, respectively, are given by

$$\begin{aligned}
\text{SS[TOT]} &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} Y_{ijk}^2 - \text{CM} \\
\text{SS[A]} &= \frac{1}{bn}\sum_{i=1}^{a} Y_{i++}^2 - \text{CM} \\
\text{SS[B]} &= \frac{1}{an}\sum_{j=1}^{b} Y_{+j+}^2 - \text{CM} \\
\text{SS[AB]} &= \left(\frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{b} Y_{ij+}^2 - \text{CM}\right) - \text{SS[A]} - \text{SS[B]} \\
\text{SS[E]} &= \text{SS[TOT]} - \text{SS[A]} - \text{SS[B]} - \text{SS[AB]}.
\end{aligned}$$

The error sum of squares can be found by subtraction. As you can see, hand computation becomes simple if we just have **totals** like Y_{i++}, Y_{+j+}, Y_{ij+}, and Y_{+++}.

Example 9.3 (`battery.sas`). An engineer is designing a battery for use in a device that will be subjected to extreme variations in temperature. He has three material types (Factor A) for the battery and plans to set the levels of temperature (Factor B) at 15, 70, and 125 degrees Celcius (note that these levels are quantitatively ordered). Four batteries are randomly assigned to each combination of material and temperature, and all 36 observations ($3 \times 3 \times 4$) are run in a random order. Here, $a = 3$, $b = 3$, and $n = 4$. *This is a completely randomised design with a* 3×3 *factorial treatment structure.* The data from the experiment are in Table 9.28. The response is Y, the effective life (in hours) observed for each battery.

Table 9.28: *Life data (in hours) for the battery design experiment.*

	Temperature (B)						
Material Type (A)	15		70		125		y_{i++}
1	130	155	34	40	20	70	998
	74	180	80	75	82	58	
2	150	188	136	122	25	70	1300
	159	126	106	115	58	45	
3	138	110	174	120	96	104	1501
	168	160	150	139	82	60	
y_{+j+}	1738		1291		770		$y_{+++} = 3799$

CALCULATIONS: First, the correction term for the mean is given by

$$\text{CM} = \frac{1}{N}Y_{+++}^2 = \frac{1}{36}(3799)^2.$$

The total sum of squares is given by

$$\text{SS[TOT]} = (130)^2 + (155)^2 + (74)^2 + \cdots + (60)^2 - \frac{1}{36}(3799)^2 = 77646.97.$$

The sum of squares for the main effects are

$$\text{SS[A]} = \frac{1}{3(4)}\left[(998)^2 + (1300)^2 + (1501)^2\right] - \frac{1}{36}(3799)^2 = 10683.72$$

and

$$\text{SS[B]} = \frac{1}{3(4)}\left[(1738)^2 + (1291)^2 + (770)^2\right] - \frac{1}{36}(3799)^2 = 39118.72.$$

To compute the interaction sum of squares, note that

$$\frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{b} Y_{ij+}^2 - \text{CM} = \frac{1}{4}\left[(539)^2 + (229)^2 + \cdots + (342)^2\right] - \frac{1}{36}(3799)^2 = 59416.22.$$

Thus,

$$\text{SS[AB]} = 59416.22 - \underbrace{10683.72}_{\text{SS[A]}} - \underbrace{39118.72}_{\text{SS[B]}} = 9613.78.$$

Finally, the error sum of squares is obtained by subtraction; i.e.,

$$\text{SS[E]} = 77646.97 - 10683.72 - 39118.72 - 9613.78 = 18230.75.$$

ANALYSIS: Table 9.29 shows the ANOVA table for the battery life data. As in 2 × 2 factorial experiments, the first thing to check (i.e., test) is the interaction effect; that is, we want to first test $H_0 : (\alpha\beta)_{ij} = 0$ for all i and j (no interaction). Here, we would reject H_0 since $F_{AB} = 3.56$ is large enough ($F_{4,27,0.05} \approx 2.73$). Thus, these data display a significant interaction between material type and temperature. This is not surprising because the interaction plot in Figure 9.28 shows a large departure from parallelism. Since significant interaction is present, we should be careful about interpreting the main effects tests (some recommend not to even perform these tests).

Table 9.29: *ANOVA table for the battery life data in Example* 9.3.

Source	df	SS	MS	F
Material	2	10683.72	5341.86	7.91
Temperature	2	39118.72	19558.36	28.97
M*T	4	9613.78	2403.44	3.56
Error	27	18230.75	675.21	
Total	35	77646.97		

ANALYSIS: Since we have a significant interaction effect, comparisons among the means of one factor (e.g., material type) are likely to be obscured by the interaction. In this situation, it would be perfectly acceptable to (this is not an exhaustive list of possibilities):

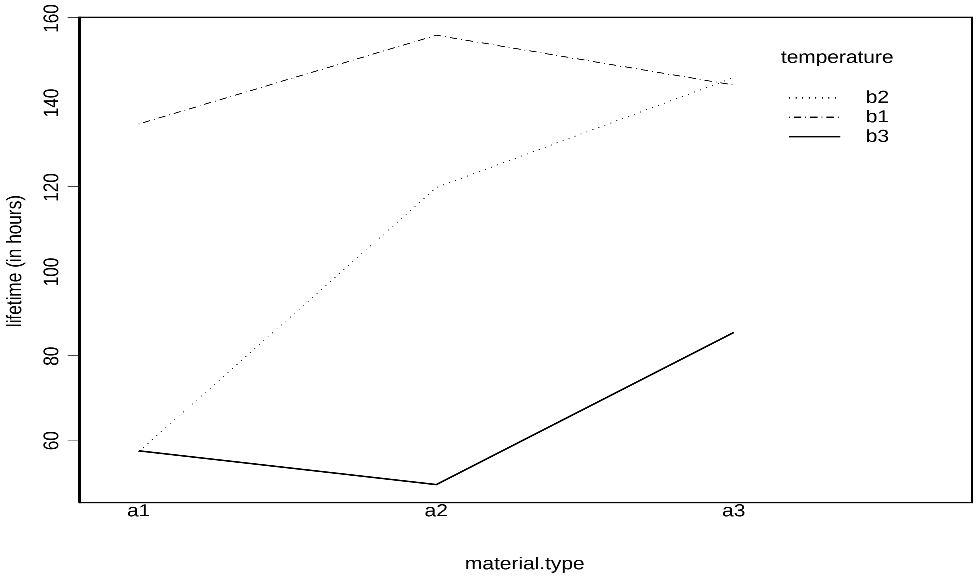

Figure 9.28: *Interaction plot for the battery data from Example* 9.3.

1. treat this as a one-way ANOVA with $3 \times 3 = 9$ treatments, and proceed as we did in Chapter 3 with contrasts among these 9 means (e.g., form all pairwise intervals).

2. compare the means of material type at a **fixed** level of temperature.

3. use orthogonal contrasts to test for *linear* and *quadratic* effects of temperature at a **fixed** level of material type.

4. fit a regression equation that describes effective life (Y) as a function of temperature at a **fixed** level of material type.

FORMING ALL PAIRWISE INTERVALS: A perfectly acceptable approach to analysing treatment differences in the presence of interaction is to form pairwise intervals for all treatment means. Of course, with $t = 9$ treatments, we have $\binom{9}{2} = 36$ pairwise intervals! Since we are exploring the data for differences, we can construct Tukey confidence intervals (remember that we must adjust for multiplicity since there are many confidence intervals). For the battery data, these intervals are in Table 9.30.

Table 9.30: *Battery data: Tukey confidence intervals for all* 36 *pairwise means.*

Difference	Estimate	Std.Error	Lower Bound	Upper Bound	
a1.b1-a1.b2	77.50	18.4	15.70	139.00	****
a1.b1-a1.b3	77.30	18.4	15.40	139.00	****
a1.b1-a2.b1	-21.00	18.4	-82.80	40.80	
a1.b1-a2.b2	15.00	18.4	-46.80	76.80	
a1.b1-a3.b2	-11.00	18.4	-72.80	50.80	
a1.b1-a2.b3	85.30	18.4	23.40	147.00	****
a1.b1-a3.b1	-9.25	18.4	-71.10	52.60	
a1.b1-a3.b3	49.30	18.4	-12.60	111.00	
a1.b2-a1.b3	-0.25	18.4	-62.10	61.60	
a1.b2-a2.b1	-98.50	18.4	-160.00	-36.70	****
a1.b2-a2.b2	-62.50	18.4	-124.00	-0.67	****
a1.b2-a3.b2	-88.50	18.4	-150.00	-26.70	****
a1.b2-a2.b3	7.75	18.4	-54.10	69.60	
a1.b2-a3.b1	-86.80	18.4	-149.00	-24.90	****
a1.b2-a3.b3	-28.20	18.4	-90.10	33.60	
a1.b3-a2.b1	-98.30	18.4	-160.00	-36.40	****
a1.b3-a2.b2	-62.30	18.4	-124.00	-0.42	****
a1.b3-a3.b2	-88.30	18.4	-150.00	-26.40	****
a1.b3-a2.b3	8.00	18.4	-53.80	69.80	
a1.b3-a3.b1	-86.50	18.4	-148.00	-24.70	****
a1.b3-a3.b3	-28.00	18.4	-89.80	33.80	
a2.b1-a2.b2	36.00	18.4	-25.80	97.80	
a2.b1-a3.b2	10.00	18.4	-51.80	71.80	
a2.b1-a2.b3	106.00	18.4	44.40	168.00	****
a2.b1-a3.b1	11.80	18.4	-50.10	73.60	
a2.b1-a3.b3	70.30	18.4	8.43	132.00	****
a2.b2-a3.b2	-26.00	18.4	-87.80	35.80	
a2.b2-a2.b3	70.20	18.4	8.43	132.00	****
a2.b2-a3.b1	-24.30	18.4	-86.10	37.60	
a2.b2-a3.b3	34.30	18.4	-27.60	96.10	
a3.b2-a2.b3	96.20	18.4	34.40	158.00	****
a3.b2-a3.b1	1.75	18.4	-60.10	63.60	
a3.b2-a3.b3	60.30	18.4	-1.57	122.00	
a2.b3-a3.b1	-94.50	18.4	-156.00	-32.70	****
a2.b3-a3.b3	-36.00	18.4	-97.80	25.80	
a3.b1-a3.b3	58.50	18.4	-3.32	120.00	

For example, the confidence interval for $\mu_{11} - \mu_{12}$ (i.e., the first one in Table 9.30) is given by

$$(\overline{Y}_{11+} - \overline{Y}_{12+}) \pm \underbrace{q_{9,27,0.05}}_{\approx 4.76} \sqrt{\frac{\text{MS[E]}}{4}} \iff (15.70, 139.00).$$

The table of **treatment means** $\overline{Y}_{ij+}$ for the battery life data is given by

	b_1	b_2	b_3
a_1	134.75	57.25	57.50
a_2	155.75	113.75	49.50
a_3	144.00	145.75	85.50

From this table, the highest sample treatment means (in descending order) correspond to the treatments $a_2b_1 > a_3b_2 > a_3b_1 > a_1b_1 > a_2b_2 > a_3b_3$. From Table 9.30, the top five treatment means are not significantly different. In addition, the bottom five treatment means are not significantly different. However, the a_2b_1 treatment mean is significantly different from a_3b_3. Overall, it appears that the mean lifetime is largest at the lowest level of temperature (i.e., $b_1 = 15$ degrees). In addition, battery lifetimes seem to decline in higher temperatures. The only violation to this last point comes with material type 3; however, the a_3b_1 and a_3b_2 means are not significantly different.

REMARK: You should remember that, with a large number of treatments (9), the Tukey procedure is going to be very **conservative** since it controls the experimentwise error. It offers too much protection against Type I Errors (claiming that differences are not real) that it is difficult to find any treatment differences, and Type II Errors (failing to detect real differences) become too likely.

COMPARING MEANS AMONG MATERIAL TYPES AT A FIXED TEMPERATURE LEVEL: Since we have a significant interaction between material type (A) and temperature (B), comparing the means of one factor should be done *separately* for each level of the other factor (not doing this would produce results that are obscured by the significant interaction). We now illustrate how to compare the material type means at fixed temperature levels.

DETAILS: To make pairwise comparisons among material type means for a fixed level of temperature, we have to first do a little algebra. Fortunately, the calculations are not too difficult. Recall our two-way ANOVA model for the battery data; i.e.,

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk},$$

for $i = 1, 2, 3$, $j = 1, 2, 3$, and $k = 1, 2, 3, 4$, where $\epsilon_{ijk} \sim \text{iid } \mathcal{N}(0, \sigma^2)$. First, let's fix the temperature at the lowest level; i.e., $b_1 = 15$, and compare the means of a_1b_1, a_2b_1, and a_3b_1. To compare the a_1b_1 and a_2b_1 means, we can construct a confidence interval for $\mu_{11} - \mu_{21}$. Using $\overline{Y}_{11+} - \overline{Y}_{21+}$ as a point estimator for $\mu_{11} - \mu_{21}$, we see that (verify!)

$$E(\overline{Y}_{11+} - \overline{Y}_{21+}) = \underbrace{(\alpha_1 - \alpha_2) + [(\alpha\beta)_{11} - (\alpha\beta)_{21}]}_{\mu_{11}-\mu_{21}}$$

and

$$V(\overline{Y}_{11+} - \overline{Y}_{21+}) = \frac{2\sigma^2}{4}.$$

Furthermore, $\overline{Y}_{11+} - \overline{Y}_{21+}$ is normally distributed since it is a linear combination of the Y_{ijk}s. Thus, a 95 percent confidence interval for $\mu_{11} - \mu_{21}$ would be

$$(\overline{Y}_{11+} - \overline{Y}_{21+}) \pm t_{27,0.025}\sqrt{\frac{2\text{MS[E]}}{4}}.$$

Confidence intervals for $\mu_{11} - \mu_{31}$ and $\mu_{21} - \mu_{31}$ are formed similarly (verify!); however, these three intervals, when viewed **jointly**, would not be adjusted for multiplicity. **Simultaneous** 95 percent Tukey confidence intervals for $\mu_{11} - \mu_{21}$, $\mu_{11} - \mu_{31}$, and $\mu_{21} - \mu_{31}$ are given by

$$(\overline{Y}_{11+} - \overline{Y}_{21+}) \pm q_{3,27,0.05}\sqrt{\frac{\text{MS[E]}}{4}},$$

$$(\overline{Y}_{11+} - \overline{Y}_{31+}) \pm q_{3,27,0.05}\sqrt{\frac{\text{MS[E]}}{4}},$$

and

$$(\overline{Y}_{21+} - \overline{Y}_{31+}) \pm q_{3,27,0.05}\sqrt{\frac{\text{MS[E]}}{4}}$$

(the $\sqrt{2}$ term is absorbed into $q_{3,27,0.05}$). Pairwise intervals for material type means at fixed temperature levels b_2 and b_3 could be formed analogously.

CONTRASTS FOR TRENDS: Recall that the sum of squares for treatments can be broken up into the sum of squares for orthogonal contrasts. There are special contrasts called **orthogonal polynomial contrasts**, because, in balanced designs, they reproduce the sum of squares for treatments for comparing different polynomial regression models. The use of such contrasts is only appropriate when the factor of interest is **quantitative** with **equally spaced** levels; e.g. doses, temperatures, percentages, etc. If levels are not equally spaced, then these polynomial contrasts are not too useful.

APPLICATION: For the battery data, we have a quantitative factor; i.e., temperature; furthermore, it is at equally spaced levels ($b_1 = 15$, $b_2 = 70$, and $b_3 = 125$ degrees). There are three levels of temperature, so, for each material type (we should consider each material type separately since there is significant interaction; i.e., the battery life response to temperature depends on which material type is used), we can construct **linear** and **quadratic** contrasts; these contrasts examine whether or not there are significant linear and quadratic trends across the levels of temperature. For material type 1 (see Figure 9.29), the following contrasts can be used to test for linear and quadratic trends:

$$\begin{aligned}\text{Linear effect:} &\quad -\overline{Y}_{11+} + \overline{Y}_{13+}\\ \text{Quadratic effect:} &\quad \overline{Y}_{11+} - 2\overline{Y}_{12+} + \overline{Y}_{13+}.\end{aligned}$$

One will note that these contrasts have contrast coefficients

Effect	c_1	c_2	c_3
Linear	-1	0	1
Quadratic	1	-2	1

Tables giving linear, quadratic, cubic, and higher order orthogonal polynomial contrasts are widely available (surprisingly not in Rao). Note that (verify!)

$$\begin{aligned}E(-\overline{Y}_{11+} + \overline{Y}_{13+}) &= (-\beta_1 + \beta_3) + [-(\alpha\beta)_{11} + (\alpha\beta)_{13}]\\ E(\overline{Y}_{11+} - 2\overline{Y}_{12+} + \overline{Y}_{13+}) &= (\beta_1 - 2\beta_2 + \beta_3) + [(\alpha\beta)_{11} - 2(\alpha\beta)_{12} + (\alpha\beta)_{13}].\end{aligned}$$

The calculations not only show us exactly what we are estimating, but they also help us with coding these effects so that we can test them in SAS.

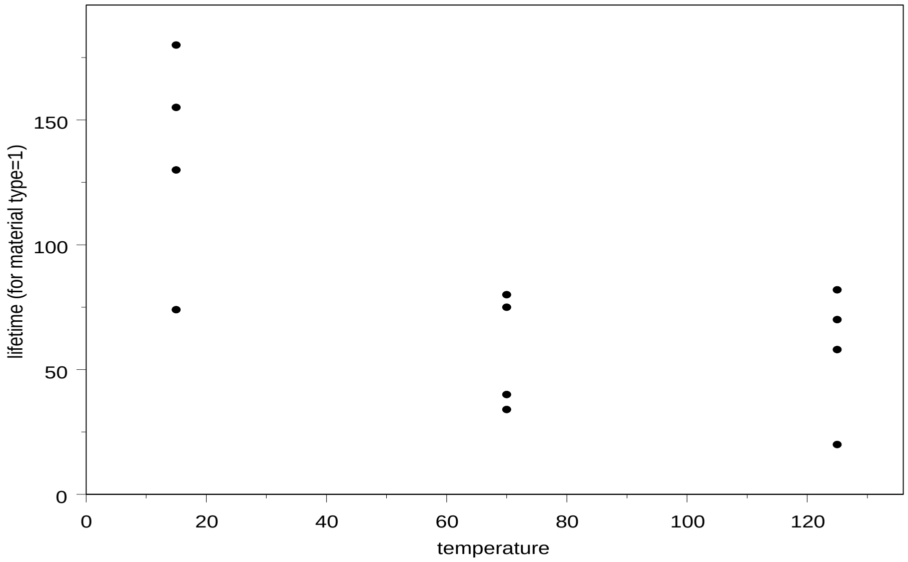

Figure 9.29: *Lifetime versus temperature for Material Type* 1.

SAS COMMANDS: Here is the SAS code to analyse linear and quadratic effects in temperature for material type 1 (see our last calculations):

```
model life = mat temp mat*temp;
contrast 'linear-for-mat1' temp -1 0 1 mat*temp -1 0 1 0 0 0 0 0 0;
contrast 'quadratic-for-mat1' temp 1 -2 1 mat*temp 1 -2 1 0 0 0 0 0 0;
```

ANALYSIS: The SAS output for testing linear and quadratic effects in temperature, for material type 1 only, is shown:

```
     Contrast        DF  Contrast SS  Mean Square     F     Pr > F
  linear-for-mat1     1   11935.125    11935.125    17.68   0.0003
quadratic-for-mat1    1    4030.042     4030.042     5.97   0.0214
```

There are significant linear and quadratic trends in temperature for material type 1. Had these contrasts been preplanned, we can make these conclusions **jointly** at the $\alpha = 0.05$ level, using a Bonferroni correction, since each P value is smaller than $0.05/2 = 0.025$.

Example 9.4 (`tomato2.sas`). Three different tomato varieties (Harvester, Pusa Early Dwarf, and Ife No. 1) and four different plant densities (10, 20, 30 and 40 thousand plants per hectare) were considered by a food scientist for planting. The goal of the experiment was to determine whether variety and plant density affect yield. Thirty-six plots were used with 3 replicates for each of 12 treatment combinations, which were assigned at random to the 36 plots. Here, $a = 3$, $b = 4$, and $n = 3$. *This is a completely randomised design with a* 3×4 *factorial treatment structure.* The data from the experiment are in Table 9.31. The response is Y, the yield, measured in tons/hectacre.

Table 9.31: *Tomato experiment yield data.*

	Density (B)			
Variety (A)	10	20	30	40
1	7.9	11.2	12.1	9.1
	9.2	12.8	12.6	10.8
	10.5	13.3	14.0	12.5
2	8.1	11.5	13.7	11.3
	8.6	12.7	14.4	12.5
	10.1	13.7	15.4	14.5
3	15.3	16.6	18.0	17.2
	16.1	18.5	20.8	18.4
	17.5	19.2	21.0	18.9

ANALYSIS: The ANOVA table is given in Table 9.32. The interaction term is not significant since $F_{VD} = 0.84 < F_{6,24,0.05} = 2.508$; i.e., variety and plant density do not interact. Also, note that the interaction plot in Figure 9.30 does not show a large departure from parallelism. There are main effects due to variety and density since both F_V and F_D are large. Since variety and density do not interact, we can treat the factors **separately** and compare means within each factor. Since variety is qualitative, we will do this using pairwise comparisons. Since density is quantitative (with equally-spaced levels), we will do this using orthogonal polynomial contrasts.

Table 9.32: *ANOVA table for the tomato yield data in Example* 9.4.

Source	df	SS	MS	F
Variety	2	327.597	163.799	103.34
Density	3	86.687	28.896	18.23
V*D	6	8.032	1.339	0.84
Error	24	38.040	1.585	
Total	35	460.356		

PAIRWISE COMPARISONS FOR VARIETY MEANS: Our initial two-way interaction model for the tomato data was $Y_{ijk} = \mu+\alpha_i+\beta_j+(\alpha\beta)_{ij}+\epsilon_{ijk}$, for $i = 1, 2, 3$, $j = 1, 2, 3, 4$, and $k = 1, 2, 3$, where $\epsilon_{ijk} \sim$ iid $\mathcal{N}(0, \sigma^2)$. However, you'll recall that we did not reject $H_0 : (\alpha\beta)_{ij} = 0$ for all i and j. Because of this, we have, thus, basically acknowledged that the **two-factor no-interaction model**

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk},$$

for $i = 1, 2, 3$, $j = 1, 2, 3, 4$, and $k = 1, 2, 3$, is appropriate for these data. To compare means among different varieties, we can form confidence intervals for $\mu_1 - \mu_2$, $\mu_1 - \mu_3$, and $\mu_2 - \mu_3$, where $\mu_i = \mu + \alpha_i$. The confidence interval for $\mu_1 - \mu_2$ is based off the point estimator $\overline{Y}_{1++} - \overline{Y}_{2++}$. The mean and variance of this estimator are given by (verify!)

$$E(\overline{Y}_{1++} - \overline{Y}_{2++}) = \alpha_1 - \alpha_2$$

and

$$V(\overline{Y}_{1++} - \overline{Y}_{2++}) = \frac{2\sigma^2}{12},$$

respectively. Furthermore, $\overline{Y}_{1++} - \overline{Y}_{2++}$ is normally distributed since it is a linear combination of the Y_{ijk}s. Thus, a 95 percent confidence interval for $\mu_1 - \mu_2$ would be

$$(\overline{Y}_{1++} - \overline{Y}_{2++}) \pm t_{24,0.025}\sqrt{\frac{2\text{MS[E]}}{12}}.$$

Individual confidence intervals for $\mu_1-\mu_3$ and $\mu_2-\mu_3$ are formed similarly. **Simultaneous** 95 percent Tukey confidence intervals for $\mu_1 - \mu_2$, $\mu_1 - \mu_3$, and $\mu_2 - \mu_3$ are given by

$$(\overline{Y}_{1++} - \overline{Y}_{2++}) \pm q_{3,24,0.05}\sqrt{\frac{\text{MS[E]}}{12}},$$

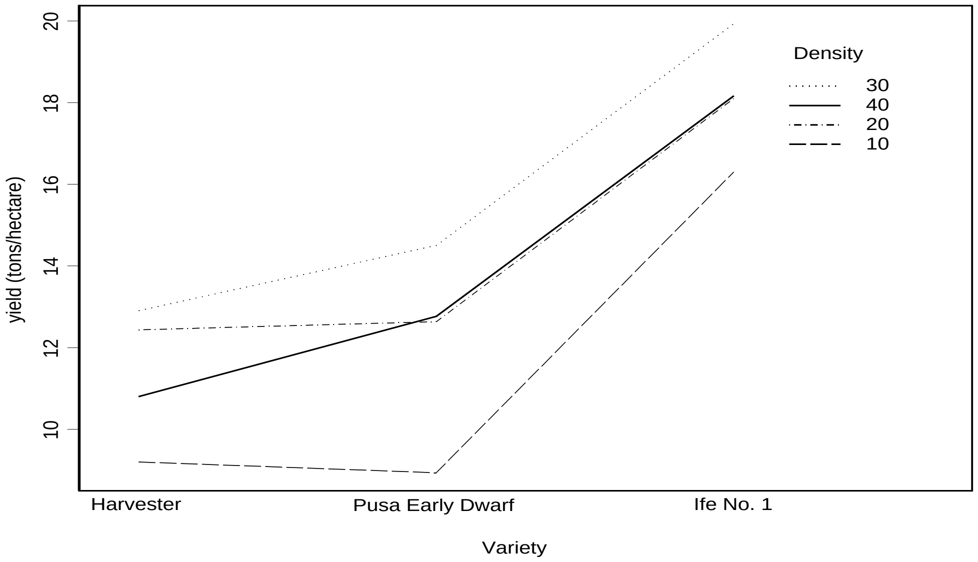

Figure 9.30: *Interaction plot for the tomato yield data in Example* 9.4.

$$(\overline{Y}_{1++} - \overline{Y}_{3++}) \pm q_{3,24,0.05}\sqrt{\frac{\text{MS[E]}}{12}},$$

and

$$(\overline{Y}_{2++} - \overline{Y}_{3++}) \pm q_{3,24,0.05}\sqrt{\frac{\text{MS[E]}}{12}}$$

(the $\sqrt{2}$ term is absorbed into $q_{3,24,0.05}$). These intervals can be easily computed in SAS; here are the results for the different varieties:

Difference	Estimate	Simultaneous 95% Confidence Limits		
a1−a2	-0.8750	-2.1585	0.4085	
a1−a3	-6.7917	-8.0752	-5.5081	***
a2−a3	-5.9167	-7.2002	-4.6331	***

Thus, it looks as though variety 3 produces the highest mean yield; it is significantly different from the variety 1 and 2 means. The variety 1 and 2 means are not significantly different. Note that we can make statements about the **orderings** among these three

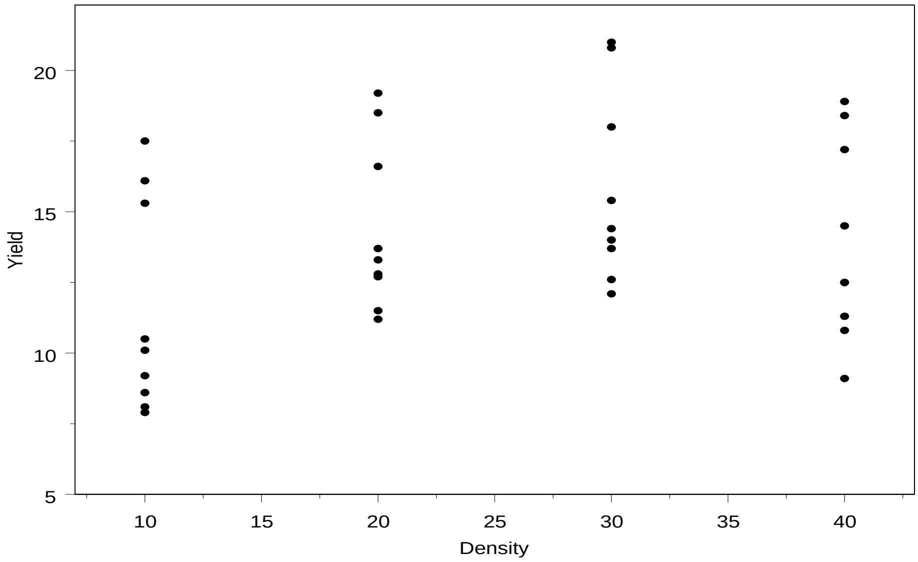

Figure 9.31: *Tomato yields versus plant density.*

means, because we have adjusted for multiplicity in the set of three intervals $\mu_1 - \mu_2$, $\mu_1 - \mu_3$, and $\mu_2 - \mu_3$.

ORTHOGONAL POLYNOMIAL CONTRASTS FOR DENSITY: For the tomato yield data, we have a quantitative factor; i.e., plant density; furthermore, it is at equally spaced levels ($b_1 = 10$, $b_2 = 20$, $b_3 = 30$, and $b_4 = 40$). There are four levels of density, so we can construct **linear**, **quadratic**, and **cubic** contrasts; these contrasts examine whether or not there are significant linear, quadratic, and cubic trends across the levels of plant density (see Figure 9.31). The following contrasts can be used to test for linear, quadratic, and cubic effects:

$$\begin{aligned}
\text{Linear effect:} &\quad -3\overline{Y}_{+1+} - \overline{Y}_{+2+} + \overline{Y}_{+3+} + 3\overline{Y}_{+4+} \\
\text{Quadratic effect:} &\quad \overline{Y}_{+1+} - \overline{Y}_{+2+} - \overline{Y}_{+3+} + \overline{Y}_{+4+} \\
\text{Cubic effect:} &\quad -\overline{Y}_{+1+} + 3\overline{Y}_{+2+} - 3\overline{Y}_{+3+} + \overline{Y}_{+4+}.
\end{aligned}$$

One will note that these contrasts have contrast coefficients

Effect	c_1	c_2	c_3	c_4
Linear	-3	-1	1	3
Quadratic	1	-1	-1	1
Cubic	-1	3	-3	1

Note that in the two-way no-interaction model (verify!)

$$\begin{aligned}
E(-3\overline{Y}_{+1+} - \overline{Y}_{+2+} + \overline{Y}_{+3+} + 3\overline{Y}_{+4+}) &= -3\beta_1 - \beta_2 + \beta_3 + 3\beta_4 \\
E(\overline{Y}_{+1+} - \overline{Y}_{+2+} - \overline{Y}_{+3+} + \overline{Y}_{+4+}) &= \beta_1 - \beta_2 - \beta_3 + \beta_4 \\
E(-\overline{Y}_{+1+} + 3\overline{Y}_{+2+} - 3\overline{Y}_{+3+} + \overline{Y}_{+4+}) &= -\beta_1 + 3\beta_2 - 3\beta_3 + \beta_4
\end{aligned}$$

The calculations not only show us exactly what we are estimating, but they also help us with coding these effects in SAS.

SAS COMMANDS: Here is the SAS code to analyse linear, quadratic, and cubic effects in plant density (see our last calculations):

```
model yield = variety density variety*density;
contrast "linear component of density effect" density -3 -1 1 3;
contrast "quadratic component of density effect" density 1 -1 -1 1;
contrast "cubic component of density effect" density -1 3 -3 1;
```

ANALYSIS: The SAS output for testing these effects is

```
 Contrast   DF  Contrast SS  Mean Square     F     Pr > F
 linear      1      33.974       33.974    21.43    0.0001
quadratic    1      51.361       51.361    32.40   <0.0001
 cubic       1       1.352        1.352     0.85    0.3649
```

Thus, there looks to be a significant linear and a significant quadratic trend in density. Had these contrasts been preplanned, we could also make these conclusions jointly at the $\alpha = 0.05$ level using a Bonferroni correction. The cubic effect is not significant.

10 Factorial Treatment Structures: Part II

Complementary reading from Rao: Chapter 13.5-13.7.

10.1 More on $a \times b$ factorial experiments

In the last chapter, we were introduced to the notation, philosophy, and analysis of $a \times b$ factorial experiments. In particular, we learned that data from balanced $a \times b$ factorial experiments can be modelled by a **two-factor ANOVA model with interaction**; i.e.,

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk},$$

for $i = 1, 2, ..., a$, $j = 1, 2, ..., b$, and $k = 1, 2, ..., n$, where $\epsilon_{ijk} \sim$ iid $\mathcal{N}(0, \sigma^2)$. In this model, μ denotes the overall mean, α_i represents the effect due to the ith level of A, β_j represents the effect due to the jth level of B, and $(\alpha\beta)_{ij}$ represents the interaction effect due to the ith and jth levels of A and B, respectively.

INTERACTION: Data analysis in factorial experiments (involving two factors) should begin by checking whether or not the interaction term is significant; we do this by testing $H_0 : (\alpha\beta)_{ij} = 0$ (no interaction) versus H_1 : not H_0. As we have seen in Examples 9.3 and 9.4, the significance of the interaction term plays a major role in determining how we perform a follow-up analysis (e.g., constructing confidence intervals, performing tests for orthogonal polynomial contrasts, etc.). When we do not reject H_0, we are acknowledging that the $(\alpha\beta)_{ij}$ term is not significantly different from zero. In this situation, we might consider a new model that excludes the interaction term.

THE TWO-FACTOR NO-INTERACTION MODEL: The two-factor no-interaction model is given by

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk},$$

for $i = 1, 2, ..., a$, $j = 1, 2, ..., b$, and $k = 1, 2, ..., n$, where $\epsilon_{ijk} \sim$ iid $\mathcal{N}(0, \sigma^2)$. The terms in the no-interaction model have the same interpretation they have in the interaction model. This model, however, assumes that Factors A and B do not interact.

REVELATION: We have already seen that the two-factor interaction ANOVA model can be expressed in the general $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ notation. The two-factor no-interaction model can also put expressed in this form. To help see this, suppose that $a = 3$, $b = 2$, and $n_{ij} = n = 3$. In matrix terms, we write

$$\boldsymbol{Y} = \begin{pmatrix} Y_{111} \\ Y_{112} \\ Y_{113} \\ Y_{121} \\ Y_{122} \\ Y_{123} \\ Y_{211} \\ Y_{212} \\ Y_{213} \\ Y_{221} \\ Y_{222} \\ Y_{223} \\ Y_{311} \\ Y_{312} \\ Y_{313} \\ Y_{321} \\ Y_{322} \\ Y_{323} \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \end{pmatrix},$$

and $\boldsymbol{\epsilon} = (\epsilon_{111}, \epsilon_{112}, ..., \epsilon_{323})'$. It is easy to see that $\boldsymbol{X}$ is not a full rank matrix. In fact, $r(\boldsymbol{X}) = 4 < p = 6$ (p is the number of parameters), so we can't estimate $\boldsymbol{\beta}$ uniquely. However, the normal equations are still **consistent**; a solution is given by $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{Y}$, where $(\boldsymbol{X}'\boldsymbol{X})^{-}$ is any generalised inverse of $\boldsymbol{X}'\boldsymbol{X}$. Equivalently, if we want to force a particular solution to the normal equations $\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}$ for the two-factor no-interaction model, we could use side conditions. Since $\boldsymbol{X}$ is rank deficient by 2, we would need two side conditions; $\alpha_+ = \beta_+ = 0$ is commonly used.

TESTING THE INTERACTION TERM: In the two-factor interaction model, we know that testing $H_0 : (\alpha\beta)_{ij} = 0$ is performed by using F_{AB}, the F statistic for interaction. However, we can also test $H_0 : (\alpha\beta)_{ij} = 0$ by pitting the no-interaction model against the interaction model; i.e.,

$$\begin{aligned} H_0: &\quad Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk} \quad \text{(reduced model)} \\ H_1: &\quad Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \quad \text{(full model)}, \end{aligned}$$

and performing a reduced-versus-full model test. To decide which model is more appropriate, all we have to do is examine the size of $\text{SS[R]}_F - \text{SS[R]}_R = \boldsymbol{Y}'(\boldsymbol{M} - \boldsymbol{M}_0)\boldsymbol{Y}$, where $\boldsymbol{M}_0$ and $\boldsymbol{M}$ are the reduced and full model hat matrices, respectively, or, equivalently, examine the size of

$$F = \frac{\boldsymbol{Y}'(\boldsymbol{M} - \boldsymbol{M}_0)\boldsymbol{Y}/r(\boldsymbol{M} - \boldsymbol{M}_0)}{\text{MS[E]}_F}.$$

The term $r(\boldsymbol{M} - \boldsymbol{M}_0)$, the **rank** of $\boldsymbol{M} - \boldsymbol{M}_0$, is the difference in degrees of freedom from the full and reduced model fits. Note that $\boldsymbol{Y}'(\boldsymbol{M} - \boldsymbol{M}_0)\boldsymbol{Y}$ can also be computed by using the **partial** sums of squares for the interaction term.

Example 10.1 (`corn.sas`). Consider the corn yield data from Example 9.2, and recall that we did not reject $H_0 : (\alpha\beta)_{ij} = 0$ since $F_{AB} = 2.25$ was not large. We could have also arrived at this conclusion from testing the no-interaction model versus the interaction model. From SAS, here are the ANOVA tables from both models:

`No interaction: Reduced Model`

Source	df	SS	MS	F
Treatments	2	530	265	12.34
Error	14	365	21.5	
Total	19	895		

`Interaction: Full model`

Source	df	SS	MS	F
Treatments	3	575	191.7	9.6
Error	16	320	20	
Total	19	895		

The F statistic to test the no-interaction model versus the interaction model is given by

$$F = \frac{(\text{SS[R]}_F - \text{SS[R]}_R)/(3-2)}{\text{MS[E]}_F} = \frac{575 - 530}{20} = 2.25,$$

which is the same as F_{AB} from the full model (this should not be surprising). Note that $\text{SS[R]}_F - \text{SS[R]}_R = 45$ is also the partial sum of squares for the interaction term.

MODEL SELECTION: Consider the two-way interaction model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk},$$

for $i = 1, 2, ..., a$, $j = 1, 2, ..., b$, and $k = 1, 2, ..., n$, where $\epsilon_{ijk} \sim$ iid $\mathcal{N}(0, \sigma^2)$. We have already seen that the no-interaction model $Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$ is a reduced version of the interaction model. However, the one-way ANOVA models $Y_{ijk} = \mu + \alpha_i + \epsilon_{ijk}$ and $Y_{ijk} = \mu + \beta_j + \epsilon_{ijk}$ are also reduced models when compared to the interaction model! Thus, from a modelling perspective, there is nothing to prevent us from testing each reduced model (separately) against the full model; that is, test

$$\begin{aligned} H_0 &: \quad Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk} \quad \text{(reduced model)} \\ H_1 &: \quad Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \quad \text{(full model)}, \end{aligned}$$

$$\begin{aligned} H_0 &: \quad Y_{ijk} = \mu + \alpha_i + \epsilon_{ijk} \quad \text{(reduced model)} \\ H_1 &: \quad Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \quad \text{(full model)}, \end{aligned}$$

and

$$\begin{aligned} H_0 &: \quad Y_{ijk} = \mu + \beta_j + \epsilon_{ijk} \quad \text{(reduced model)} \\ H_1 &: \quad Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \quad \text{(full model)}. \end{aligned}$$

NOTATION: There are four models of interest here (the interaction model, the no-interaction model, and the two one-way models), so it is helpful to define some notation to help us keep track which model is which. We'll define [AB] to represent the two-factor interaction model, [A][B] to represent the two-factor no-interaction model, and [A] and [B] to represent the models which include only Factor A and Factor B, respectively. The idea is that each model is identified by the highest order terms in it. Thus, we are interested in testing [A][B] versus [AB], [A] versus [AB], and [B] versus [AB]. We now illustrate this with an example.

Example 10.2 (`sheep.sas`). A biologist is studying glucose levels (Y, measured in mg/dl) in sheep. He randomly assigns 18 sheep to treatments in a 2×3 factorial experiment. The first factor (A) is drug (control and slaframine) and the second factor

Table 10.33: *Glucose levels in sheep.*

Drug	Control	AH	CM
Control	52, 58, 53	67, 67, 77	63, 55, 54
Slaframine	59, 84, 85	74, 74, 72	63, 79, 66

(B) is diet (control, AH = alfalfa hay, and CM = cottonseed meal). *This is a completely randomised design with a* 2×3 *factorial treatment structure.* The data are in Table 10.33. The mean-squared error from model [AB] (i.e., the full model) is MS[E] = 59.78. From SAS, here are the results from all four model fits: [AB], [A][B], [A], and [B]:

Predictors	Model	SS[T]	df	F
`drug diet drug*diet`	[AB]	1173.78	5	
`drug diet`	[A][B]	912.33	3	2.19
`drug`	[A]	672.22	1	2.10
`diet`	[B]	240.11	2	5.21

The F statistics are formed by pitting each of the smaller models versus the full model [AB]; for example,

$$F = \frac{(1173.78 - 912.33)/(5-3)}{59.78} = 2.19.$$

It looks as though models [A] and [A][B] both fit the data as well as the interaction model (neither F statistic is significant at the $\alpha = 0.05$ level). However, there is no reason to choose the [A][B] model if [A] fits just as well. In fact, in the [A][B] model, the test for $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ is not rejected ($P = 0.2153$). The analysis might continue by writing a confidence interval for the difference in means from the control and slaframine subjects. In the model $Y_{ijk} = \mu + \alpha_i + \epsilon_{ijk}$, it is easy to show (verify!) that

$$E(\overline{Y}_{1++} - \overline{Y}_{2++}) = \alpha_1 - \alpha_2$$

and that

$$V(\overline{Y}_{1++} - \overline{Y}_{2++}) = \frac{2\sigma^2}{9}.$$

Furthermore, $\overline{Y}_{1++} - \overline{Y}_{2++}$ is normally distributed since it is a linear combination of the Y_{ijk}s. Thus, a 95 percent confidence interval for $\alpha_1 - \alpha_2$ would be

$$(\overline{Y}_{1++} - \overline{Y}_{2++}) \pm t_{12,0.025}\sqrt{\frac{2\text{MS[E]}}{9}},$$

where MS[E] is the mean-squared error from the full model fit. With $\overline{y}_{1++} = 60.67$, $\overline{y}_{2++} = 72.89$, MS[E] $= 59.78$ (full), and $t_{12,0.025} = 2.1788$, a 95 percent confidence interval for $\alpha_1 - \alpha_2$, the difference in means for the control and slaframine treatments, is

$$(60.67 - 72.89) \pm 2.1788\sqrt{\frac{2(59.78)}{9}} \iff (-20.16, -4.28).$$

It looks as though slaframine significantly increases the mean glucose level in these sheep.

10.2 Factorial experiments with three or more factors

The extension of factorial treatment structures to more than two factors, and the analysis of data from such experiments, is straightforward. To illustrate the extension, we focus on three-factor experiments. Suppose that Factor A has a levels, Factor B has b levels, and Factor C has c levels. The **three-factor full-interaction model** is given by

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijkl},$$

for $i = 1, 2, ..., a$, $j = 1, 2, ..., b$, $k = 1, 2, ..., c$, and $l = 1, 2, ..., n$, where $\epsilon_{ijk} \sim$ iid $\mathcal{N}(0, \sigma^2)$. We'll continue to assume that our design is **balanced**.

TYPES OF EFFECTS: In the three-factor model, there are three types of effects:

- **Second-order interaction effect**: ABC. This is the effect of interaction between the three factors.
- **First-order interaction effects**: AB, AC, and BC. These are the effects of interactions between any two factors.
- **Main effects**: A, B, and C. These are the three main effects.

NOTE: First order-interaction effects are the interactions between two factors averaged over the other factor. When there is a second-order interaction, the first-order interactions behave differently across the levels of the other factor. See p. 623, Rao.

PARTITIONING SUMS OF SQUARES: In an $a \times b \times c$ factorial experiment, we could ignore the treatment structure and just pretend that we have a one-way layout with abc means. Doing so leads to the usual breakdown

$$\begin{aligned}\underbrace{\sum_{i=1}^a\sum_{j=1}^b\sum_{k=1}^c\sum_{l=1}^n(Y_{ijkl}-\overline{Y}_{++++})^2}_{\text{SS[TOT]}} &= \underbrace{n\sum_{i=1}^a\sum_{j=1}^b\sum_{k=1}^c(\overline{Y}_{ijk+}-\overline{Y}_{++++})^2}_{\text{SS[T]}} \\ &\quad + \underbrace{\sum_{i=1}^a\sum_{j=1}^b\sum_{k=1}^c\sum_{l=1}^n(Y_{ijkl}-\overline{Y}_{ijk+})^2}_{\text{SS[E]}},\end{aligned}$$

which corresponds to the one-way ANOVA table:

Source	df	SS	MS	F
Treatments	$abc-1$	SS[T]	MS[T]	$F=\frac{\text{MS[T]}}{\text{MS[E]}}$
Error	$N-abc$	SS[E]	MS[E]	
Total	$N-1$	SS[TOT]		

Here, $N = abcn$, the total number of observations. The key point to realise is that, like before with the two-factor model, we can break up the treatment sums of squares SS[T] into components for the main effects and the interaction effects. That is,

$$\text{SS[T]} = \text{SS[A]} + \text{SS[B]} + \text{SS[C]} + \text{SS[AB]} + \text{SS[AC]} + \text{SS[BC]} + \text{SS[ABC]}.$$

In general, this breakdown gives rise to the ANOVA table for the three-way full-interaction model; this table is Table 10.34. For what they are worth, I will now provide the computing formulae for the sums of squares. The **correction term** for the mean is $\text{CM} = Y_{++++}^2/N$. The total sum of squares is given by

$$\text{SS[TOT]} = \sum_{i=1}^a\sum_{j=1}^b\sum_{k=1}^c\sum_{l=1}^n Y_{ijkl}^2 - \text{CM}.$$

Table 10.34: *ANOVA table for the $a \times b \times c$ factorial experiment.*

Source	df	SS	MS	F
A	$a-1$	SS[A]	MS[A]	$F_A = \frac{\text{MS[A]}}{\text{MS[E]}}$
B	$b-1$	SS[B]	MS[B]	$F_B = \frac{\text{MS[B]}}{\text{MS[E]}}$
C	$c-1$	SS[C]	MS[C]	$F_C = \frac{\text{MS[C]}}{\text{MS[E]}}$
AB	$(a-1)(b-1)$	SS[AB]	MS[AB]	$F_{AB} = \frac{\text{MS[AB]}}{\text{MS[E]}}$
AC	$(a-1)(c-1)$	SS[AC]	MS[AC]	$F_{AC} = \frac{\text{MS[AC]}}{\text{MS[E]}}$
BC	$(b-1)(c-1)$	SS[BC]	MS[BC]	$F_{BC} = \frac{\text{MS[BC]}}{\text{MS[E]}}$
ABC	$(a-1)(b-1)(c-1)$	SS[ABC]	MS[ABC]	$F_{ABC} = \frac{\text{MS[ABC]}}{\text{MS[E]}}$
Error	$N-abc$	SS[E]	MS[E]	
Total	$N-1$	SS[TOT]		

The sum of squares for the **main effects** are

$$\begin{aligned}
\text{SS[A]} &= \frac{1}{bcn}\sum_{i=1}^{a} Y_{i+++}^2 - \text{CM} \\
\text{SS[B]} &= \frac{1}{acn}\sum_{j=1}^{b} Y_{+j++}^2 - \text{CM} \\
\text{SS[C]} &= \frac{1}{abn}\sum_{k=1}^{c} Y_{++k+}^2 - \text{CM}.
\end{aligned}$$

The sum of squares for the **interaction effects** are

$$\begin{aligned}
\text{SS[AB]} &= \left(\frac{1}{cn}\sum_{i=1}^{a}\sum_{j=1}^{b} Y_{ij++}^2 - \text{CM}\right) - \text{SS[A]} - \text{SS[B]} \\
\text{SS[AC]} &= \left(\frac{1}{bn}\sum_{i=1}^{a}\sum_{k=1}^{c} Y_{i+k+}^2 - \text{CM}\right) - \text{SS[A]} - \text{SS[C]} \\
\text{SS[BC]} &= \left(\frac{1}{an}\sum_{j=1}^{b}\sum_{k=1}^{c} Y_{+jk+}^2 - \text{CM}\right) - \text{SS[B]} - \text{SS[C]} \\
\text{SS[ABC]} &= \left(\frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c} Y_{ijk+}^2 - \text{CM}\right) - \text{SS[A]} - \text{SS[B]} - \text{SS[C]} \\
&\qquad -\text{SS[AB]} - \text{SS[AC]} - \text{SS[BC]}.
\end{aligned}$$

F STATISTICS: As in the two-factor setting, we can test for main and interaction effects in $a \times b \times c$ experiments by using F statistics. For example, the statistic

$$F_A = \frac{\text{MS[A]}}{\text{MS[E]}} \sim F_{a-1,N-abc}.$$

is used to test $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$ (no A main effect) versus H_1 : not H_0. The statistic

$$F_{BC} = \frac{\text{MS[BC]}}{\text{MS[E]}} \sim F_{(b-1)(c-1),N-abc}.$$

is used to test $H_0 : (\beta\gamma)_{jk} = 0$ for all j and k (no BC interaction) versus H_1 : not H_0. The statistic

$$F_{ABC} = \frac{\text{MS[ABC]}}{\text{MS[E]}} \sim F_{(a-1)(b-1)(c-1),N-abc}.$$

is used to test $H_0 : (\alpha\beta\gamma)_{ijk} = 0$ for all i, j, and k (no ABC interaction) versus H_1 : not H_0. The other F statistics are defined analogously.

THE GENERAL LINEAR MODEL: The three-factor full interaction model can be expressed in the form $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. As an exercise, you might try to show this in a simple case, say, when $a = b = c = n = 2$. However, as you might suspect, the $\boldsymbol{X}$ matrix is not full rank, and, hence, the normal equations $\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}$ can not be solved uniquely. We could impose side conditions to solve them, but we know from a practical standpoint that this is not a big issue.

GENERAL STRATEGY FOR A FACTORIAL ANALYSIS: The following strategies are common when analysing data from $a \times b \times c$ factorial experiments.

- Start by looking at whether or not the second-order interaction term ABC is significant. This is done by using F_{ABC}. As we saw in two-factor experiments, the highest order interaction term dictates the entire follow-up analysis!
  - **If there is a second-order interaction**, then tests for main effects and first-order interaction effects are less meaningful because their interpretations depend on the second-order interaction. In this situation, the easiest approach is to just do the entire analysis as a one-way ANOVA with abc treatments.

Table 10.35: *Legume data for different soils.*

A (Species)	Alfalfa			Red clover			Sweet clover		
B (Soil)	loam	sand	clay	loam	sand	clay	loam	sand	clay
treated (c_1)	8	17	22	7	26	34	10	24	39
	17	13	20	10	24	32	9	24	36
not treated (c_2)	5	11	16	3	17	32	5	16	33
	4	10	15	5	19	29	4	16	34

  - **If the second-order interaction term is not significant**, look at the first-order interaction terms. If they are all significant, I would probably do three separate analyses: one for the *ab* means, one for the *ac* means, and one for the *bc* means. If two of the first-order interaction effects are significant, I would do two interaction analyses. If only one is significant, I would do one interaction analysis. If none of the first-order interaction terms are significant, I would move on to analysing main effects analogously to how we did this in two-factor experiments.

- In practice, instead of formally looking at F statistics like F_A and F_{BC}, I usually just fit all possible reduced models and find the smallest one that fit the data well. We now illustrate this approach with an example.

Example 10.3 (`seed.sas`). In a greenhouse experiment, a plant pathologist wanted to determine the rate of emergences of seed for three species of legumes (alfalfa, red clover, and sweet clover) and three soil types (silt loam, sand, and clay). Legumes were either treated with a fungicide or not. So, we have three factors here. For notational purposes, we will take legume as Factor A (3 levels), soil type as Factor B (3 levels), and fungicide as Factor C (2 levels). The response is Y, the number of plants emerged. Each of the 18 treatment combinations were randomly assigned 36 pots. *This is a completely randomised design with a* $3 \times 3 \times 2$ *factorial treatment structure.* The data are given in Table 10.35.

ANALYSIS: Our first priority is to figure out which interactions are important. The mean-squared error from model [ABC] (i.e., the full model) is MS[E] = 4.167. The model notation here mirrors that as before. For example, model [AB][AC] corresponds to $Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + \epsilon_{ijkl}$, model [A][B][C] corresponds to $Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \epsilon_{ijkl}$, etc. As you will note, the main effects are always all included. From SAS, here are the results:

Model	SS[T]	df	F
[ABC]	3848.89	17	
[AB][AC][BC]	3833.61	13	0.92
[AB][AC]	3823.55	11	1.01
[AB][BC]	3830.72	11	0.73
[AC][BC]	3406.33	9	13.28
[AB][C]	3820.67	9	0.85
[AC][B]	3396.28	7	10.89
[BC][A]	3403.44	7	10.69
[A][B][C]	3393.39	5	9.11

All of the models have been compared to the full model using F statistics. The F statistics are formed by pitting each of the smaller models versus the full model [ABC]; for example, to test model [A][B][C] versus the full model, we have

$$F = \frac{(3848.89 - 3393.39)/(17-5)}{4.167} = 9.11.$$

Other F statistics are formed similarly. It takes neither a genius nor an F table to see that the only models that fit the data are those that include the AB (legume-soil) interaction. Among these models, there is no reason not to choose model [AB][C] since it is the smallest. In fact, the extra interaction terms in models [AB][AC], [AB][BC], and [AB][AC][BC] are all not significant. Writing out the model [AB][C], it is

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + \epsilon_{ijkl}.$$

The interaction plot for soil and legume is in Figure 10.32. Continuing with the analysis,

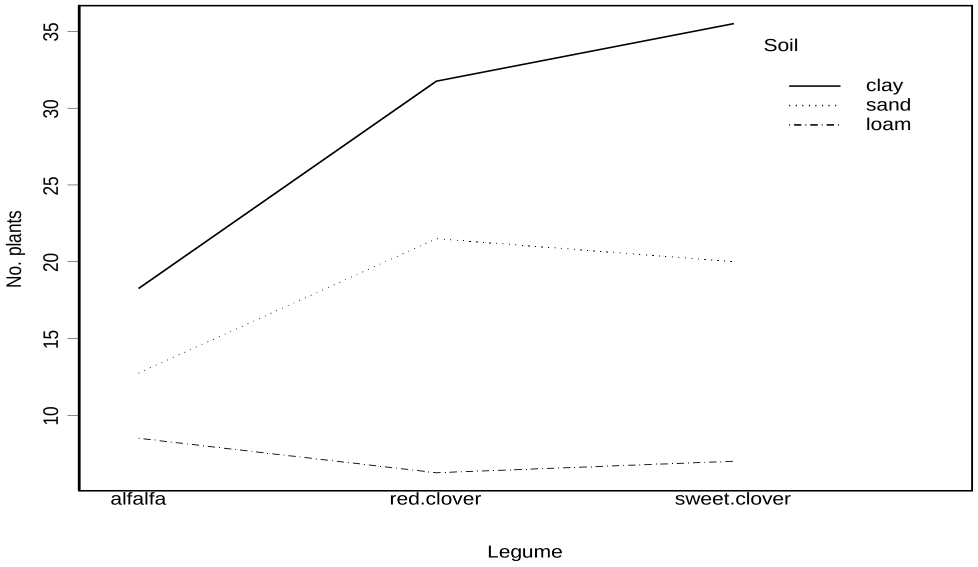

Figure 10.32: *Legume-soil interaction plot for the seed data from Example* 10.3.

we should investigate the AB (legume-soil) interaction. The two levels of C (fungicide) can be examined individually. Starting with the latter, we can construct a confidence interval for the difference of means for treated and untreated seeds. In the reduced model $Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + \epsilon_{ijkl}$, it is easy to show (verify!) that

$$E(\overline{Y}_{++1+} - \overline{Y}_{++2+}) = \gamma_1 - \gamma_2$$

and that

$$V(\overline{Y}_{++1+} - \overline{Y}_{++2+}) = \frac{2\sigma^2}{18}.$$

Furthermore, $\overline{Y}_{++1+} - \overline{Y}_{++2+}$ is normally distributed since it is a linear combination of the Y_{ijkl}s. Thus, a 95 percent confidence interval for $\gamma_1 - \gamma_2$ would be

$$(\overline{Y}_{++1+} - \overline{Y}_{++2+}) \pm t_{18,0.025}\sqrt{\frac{2\text{MS[E]}}{18}},$$

where MS[E] is the mean-squared error from the full model fit. With $\overline{y}_{++1+} = 20.67$, $\overline{y}_{++2+} = 15.22$, MS[E] $= 4.167$ (full), and $t_{18,0.025} = 2.1009$, a 95 percent confidence interval for $\gamma_1 - \gamma_2$, the difference in means for the treated and untreated seeds, is given

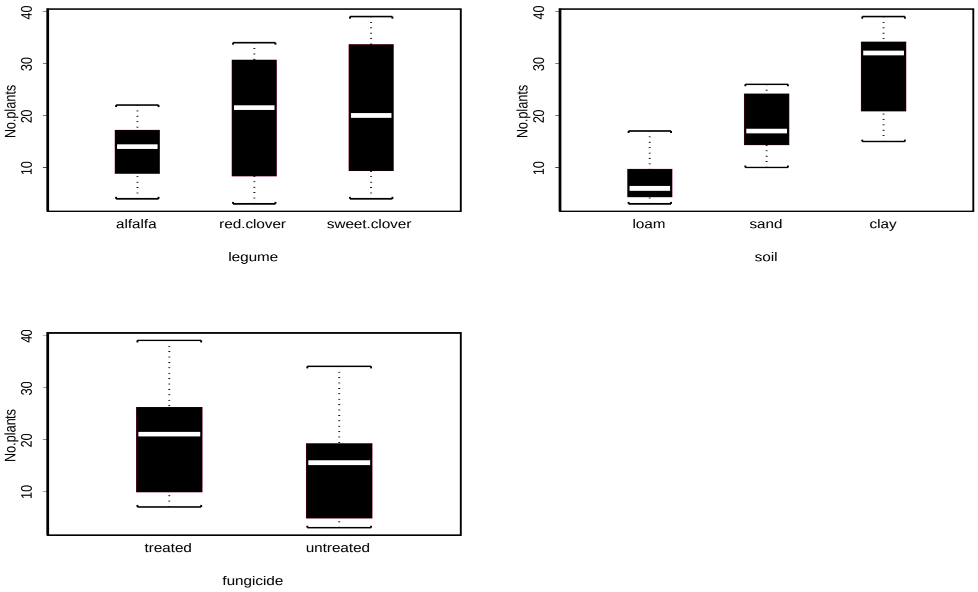

Figure 10.33: *Factor plot for the seed data from Example* 10.3.

by $(4.01, 6.87)$. Thus, it looks as though the treated seeds produce a larger mean number or emerged plants than those seeds untreated. To conduct the legume-soil analysis, I would construct pairwise intervals for the means of the different legumes for each soil (separately). For example, to compare the legumes (a_1, a_2 and a_3) for the sand (b_3) group only, note that in the reduced model $Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + \epsilon_{ijkl}$, we have the following (verify!):

$$
\begin{aligned}
E(\overline{Y}_{23++} - \overline{Y}_{13++}) &= (\alpha_2 - \alpha_1) + [(\alpha\beta)_{23} - (\alpha\beta)_{13}] \\
E(\overline{Y}_{33++} - \overline{Y}_{13++}) &= (\alpha_3 - \alpha_1) + [(\alpha\beta)_{33} - (\alpha\beta)_{13}] \\
E(\overline{Y}_{33++} - \overline{Y}_{23++}) &= (\alpha_3 - \alpha_2) + [(\alpha\beta)_{33} - (\alpha\beta)_{23}].
\end{aligned}
$$

These calculations form the basis for the SAS commands:

```
estimate 'alf-red sand'   legume -1 1 0 legume*soil 0 0 -1 0 0 1 0 0 0;
estimate 'alf-sweet sand' legume -1 0 1 legume*soil 0 0 -1 0 0 0 0 0 1;
estimate 'red-sweet sand' legume 0 -1 1 legume*soil 0 0 0 0 0 -1 0 0 1;
```

DIATRIBE: As illustrated in Examples 10.2 and 10.3, you might see that I find testing reduced models (versus a full model) to be more sensible than testing hypotheses about parameters in the full model. This is why I resist looking at statistics like F_A and F_{AB} in full three-factor ANOVA models. These statistics test hypotheses about **parameters**. While parameters are an integral part of most models, I don't believe they are an end in themselves. I believe that good models are the end product. Of course, in a three-factor full-interaction ANOVA model, the statistic F_{ABC} tests a hypothesis about parameters in the full model, but it also tests a reduced model versus the full model (which reduced model?), so I don't mind looking at it.

UNBALANCED DATA: When the number of replications is different for different treatment combinations (e.g., see Example 13.16, p. 630-1, Rao), contrasts that measure main and interaction effects are no longer orthogonal (recall the 2×2 case). This is not prohibitive. All we lose is having nice computing formulae that we probably will never use anyway (unless we are stranded somewhere without SAS). *You will note that this approach of examining all reduced models does not require that the design be balanced. It only relies on the general theory of reduced-model testing and our usual model assumptions.* When $n_{ijk} = 0$; i.e., we have no measurements of Factors A, B, and C, at levels i, j, and k, respectively, then we have estimability issues. In this case, certain factorial effects can not be estimated. However, we can still take the approach of testing reduced and full models.

10.3 Expected mean squares

Mean squares can be viewed as **random** quantities because they depend on observations of Y, which is random itself. That is, one can think of a mean square statistic as a **random variable**. Like any random variable, it has, among other things, its own expected value, its own variance, and its own probability distribution! The expected value of a mean square, or, **expected mean square**, for short, is an important quantity. These quantities help us construct F statistics. You'll recall that we have already examined

Table 10.36: *ANOVA table for the $a \times b$ factorial experiment.*

Source	df	SS	MS	F
A	$a-1$	SS[A]	MS[A]	$F_A = \frac{\text{MS[A]}}{\text{MS[E]}}$
B	$b-1$	SS[B]	MS[B]	$F_B = \frac{\text{MS[B]}}{\text{MS[E]}}$
AB	$(a-1)(b-1)$	SS[AB]	MS[AB]	$F_{AB} = \frac{\text{MS[AB]}}{\text{MS[E]}}$
Error	$N-ab$	SS[E]	MS[E]	
Total	$N-1$	SS[TOT]		

expected mean squares in the one-way layout (see Chapter 2) and in regression models (see Chapters 4 and 6). Doing so aided our understanding of why one should reject null hypotheses when F statistics become large.

EXPECTED MEAN SQUARES IN THE TWO-FACTOR MODEL: Consider our two-factor interaction model; i.e.,

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk},$$

for $i = 1, 2, ..., a$, $j = 1, 2, ..., b$, and $k = 1, 2, ..., n$, where $\epsilon_{ijk} \sim$ iid $\mathcal{N}(0, \sigma^2)$. If we impose the "usual" side conditions $\alpha_+ = \beta_+ = (\alpha\beta)_{i+} = (\alpha\beta)_{+j} = 0$ (see Equation 13.16, p. 621, Rao), then it follows that

$$\begin{aligned} E(\text{MS[A]}) &= \sigma^2 + bn(a-1)^{-1}\sum_{i=1}^{a}\alpha_i^2 \\ E(\text{MS[B]}) &= \sigma^2 + an(b-1)^{-1}\sum_{j=1}^{b}\beta_j^2 \\ E(\text{MS[AB]}) &= \sigma^2 + n[(a-1)(b-1)]^{-1}\sum_{i=1}^{a}\sum_{j=1}^{b}(\alpha\beta)_{ij}^2 \\ E(\text{MS[E]}) &= \sigma^2. \end{aligned}$$

USING THE EXPECTED MEAN SQUARES: These equations can be helpful. For example, suppose that we wanted to test for the presence of an AB interaction; that is, test $H_0 : (\alpha\beta)_{ij} = 0$ for all i and j (no interaction) versus H_1 : not H_0. When H_0 is true, then both MS[AB] and MS[E] estimate the same quantity; namely, σ^2. In this case, we would expect $F_{AB} = \text{MS[AB]}/\text{MS[E]}$ to be close to one. When H_0 is not true, we would expect MS[AB] to estimate something larger than σ^2. This will cause F_{AB} to get large. We could apply the same reasoning to see why $F_A = \text{MS[A]}/\text{MS[E]}$ and $F_B = \text{MS[B]}/\text{MS[E]}$ get large when the hypotheses $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$ and $H_0 : \beta_1 = \beta_2 = \cdots = \beta_b = 0$, respectively, are not true. You will note that in all three cases, we are looking for a denominator mean square that has the same expectation as the numerator mean square when a specific H_0 is true. In all three cases, this denominator mean square is MS[E]. In more complicated models (e.g., models with random and/or nested factors), MS[E] is not always the "right" denominator.

EXPECTED MEAN SQUARES IN THE THREE-FACTOR MODEL: Consider our three-factor full-interaction model; i.e.,

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijkl},$$

for $i = 1, 2, ..., a$, $j = 1, 2, ..., b$, $k = 1, 2, ..., c$, and $l = 1, 2, ..., n$, where $\epsilon_{ijk} \sim \text{iid } \mathcal{N}(0, \sigma^2)$. If we impose the "usual" side conditions analogous to the two-factor interaction model, then it follows that

$$\begin{aligned}
E(\text{MS[A]}) &= \sigma^2 + bcn(a-1)^{-1}\sum_{i=1}^{a}\alpha_i^2 \\
E(\text{MS[B]}) &= \sigma^2 + abn(b-1)^{-1}\sum_{j=1}^{b}\beta_j^2 \\
E(\text{MS[C]}) &= \sigma^2 + bcn(c-1)^{-1}\sum_{k=1}^{c}\gamma_k^2 \\
E(\text{MS[AB]}) &= \sigma^2 + cn[(a-1)(b-1)]^{-1}\sum_{i=1}^{a}\sum_{j=1}^{b}(\alpha\beta)_{ij}^2 \\
E(\text{MS[AC]}) &= \sigma^2 + bn[(a-1)(c-1)]^{-1}\sum_{i=1}^{a}\sum_{j=1}^{c}(\alpha\gamma)_{ik}^2 \\
E(\text{MS[BC]}) &= \sigma^2 + an[(b-1)(c-1)]^{-1}\sum_{j=1}^{b}\sum_{k=1}^{c}(\beta\gamma)_{jk}^2 \\
E(\text{MS[ABC]}) &= \sigma^2 + n[(a-1)(b-1)(c-1)]^{-1}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(\alpha\beta\gamma)_{ijk}^2 \\
E(\text{MS[E]}) &= \sigma^2.
\end{aligned}$$

Table 10.37: *Yield data for different alcohols and bases.*

Alcohol					
a_1		a_2		a_3	
Base 1	Base 2	Base 3	Base 4	Base 5	Base 6
91.3	90.7	89.9	88.1	79.5	87.6
91.4	94.7	92.3	93.1	70.1	89.8
89.9	89.5	92.6	90.8	79.7	88.2
94.4	85.2	94.4	98.1	76.1	87.7

USING THE EXPECTED MEAN SQUARES: Again, we can see the usefulness of these expressions to construct and appreciate F statistics. For example, if $H_0 : (\alpha\beta\gamma)_{ijk} = 0$ for all i, j, and k (no ABC interaction) is true, then $E(\text{MS[ABC]}) = E(\text{MS[E]}) = \sigma^2$, in which case the ratio $F_{ABC} = \text{MS[ABC]}/\text{MS[E]}$ should be close to one. If H_0 is not true, then we would expect F_{ABC} to be larger than one. For all the available F tests, we see that MS[E] is the “right” denominator.

10.4 Experiments with nested factors

CROSSED FACTORS: Up until now, in our discussion of factorial experiments, we have assumed that factors are crossed. In general, we say that factors A and B are **crossed** if every level of A occurs in combination with every level of B. Examples 9.1-9.4, as well as Examples 10.2-3 all are examples with a crossed factorial treatment structure. In some experiments, however, the levels of factor B (say) do not appear with all levels of factor A. Consider the following (fictitious) example.

Example 10.4 (`chemical.sas`). A chemical production process consists of a first reaction with an alcohol (A) and a second reaction with a base (B). We have three alcohols (a_1, a_2, and a_3) and six bases (b_1, b_2, b_3, b_4, b_5, and b_6). The data are in Table 10.37. The response is Y, the percent yield. Clearly, alcohol and base are not crossed factors. For example, there is no treatment combination a_1b_3. We only have the following (six) treatment combinations: $a_1b_1, a_1b_2, a_2b_3, a_2b_4, a_3b_5$, and a_3b_6. This is a completely randomised design with a 3×2 **nested** factorial treatment structure. There are 3 levels of alcohol and 2 levels of base within each level of alcohol.

NESTED FACTORS: When the levels of one factor (e.g., factor B) are similar but not identical for different levels of another factor (e.g., factor A), we say that the levels of B are **nested** within A and write $B(A)$. In other words, if you tell me which base was used in Example 10.4, I can tell you, with certainty, which level of alcohol was used (there is only one). *Thus, base is nested within alcohol.*

NOTE: If, in Example 10.4, we only had two bases, say, b_1 and b_2, and each base appeared under each alcohol, then we would have a completely randomised design with a 3×2 **crossed** factorial treatment structure.

LINEAR STATISTICAL MODEL: The linear statistical model for the **two-factor nested design** is

$$Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \epsilon_{ijk},$$

for $i = 1, 2, ..., a$, $j = 1, 2, ..., b$, and $k = 1, 2, ..., n_{ij}$, where $\epsilon_{ijk} \sim$ iid $\mathcal{N}(0, \sigma^2)$. In this model, μ denotes the overall mean, α_i represents the effect due to the ith level of A, $\beta_{j(i)}$ represents the effect of the jth level of B, nested within the ith level of A. In Example 10.4, we have $a = 3$ and $b = 2$. We'll continue to assume that our design is **balanced**.

REVELATION: It should come as no surprise to you that the two-factor nested model can be expressed in the form $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$. To help see this, suppose that $a = 3$, $b = 2$, and $n_{ij} = n = 4$, as in Example 10.4. The matrix formulation of the nested model is on the next page. As you might suspect, the $\boldsymbol{X}$ matrix is not full rank. Here, you will note that the last six columns of $\boldsymbol{X}$ are linearly independent, and the first four columns are each a linear combination of the last six columns; thus, $r(\boldsymbol{X}) = 6 < 10 = p$, so the normal equations $\boldsymbol{X}'\boldsymbol{X\beta} = \boldsymbol{X}'\boldsymbol{Y}$ can not be solved uniquely. We can impose side conditions to solve them. In balanced designs, as in Example 10.4, commonly-used side conditions in the two-factor nested model are $\alpha_+ = \beta_{+(i)} = 0$, for $i = 1, 2, ..., a$. You will note that there are four side conditions here (since $a = 3$), which is the difference between the number of parameters and $r(\boldsymbol{X})$. Alternatively, we could just use a generalised inverse of $\boldsymbol{X}'\boldsymbol{X}$ to solve the normal equations; i.e., $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{Y}$; this solution simply corresponds to using a certain side condition (which one depends on which generalised inverse was used).

$$\boldsymbol{Y} = \begin{pmatrix} Y_{111} \\ Y_{112} \\ Y_{113} \\ Y_{114} \\ Y_{121} \\ Y_{122} \\ Y_{123} \\ Y_{124} \\ Y_{211} \\ Y_{212} \\ Y_{213} \\ Y_{214} \\ Y_{221} \\ Y_{222} \\ Y_{223} \\ Y_{224} \\ Y_{311} \\ Y_{312} \\ Y_{313} \\ Y_{314} \\ Y_{321} \\ Y_{322} \\ Y_{323} \\ Y_{324} \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ (\alpha\beta)_{1(1)} \\ (\alpha\beta)_{2(1)} \\ (\alpha\beta)_{1(2)} \\ (\alpha\beta)_{2(2)} \\ (\alpha\beta)_{1(3)} \\ (\alpha\beta)_{2(3)} \end{pmatrix},$$

and $\boldsymbol{\epsilon} = (\epsilon_{111}, \epsilon_{112}, ..., \epsilon_{324})'$.

Table 10.38: *ANOVA table for the balanced $a \times b$ nested experiment.*

Source	df	SS	MS	F
A	$a-1$	SS[A]	MS[A]	$F_A = \frac{\text{MS[A]}}{\text{MS[E]}}$
B(A)	$a(b-1)$	SS[B(A)]	MS[B(A)]	$F_{B(A)} = \frac{\text{MS[B(A)]}}{\text{MS[E]}}$
Error	$N-ab$	SS[E]	MS[E]	
Total	$N-1$	SS[TOT]		

ANOVA TABLE FOR TWO-FACTOR NESTED EXPERIMENTS: The breakdown for the sum of squares in (balanced) two-factor nested experiments is based on the identity

$$\underbrace{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk}-\overline{Y}_{+++})^2}_{\text{SS[TOT]}} = \underbrace{nb\sum_{i=1}^{a}(\overline{Y}_{i++}-\overline{Y}_{+++})^2}_{\text{SS[A]}} + \underbrace{n\sum_{i=1}^{a}\sum_{j=1}^{b}(\overline{Y}_{ij+}-\overline{Y}_{i++})^2}_{\text{SS[B(A)]}}$$
$$+\underbrace{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk}-\overline{Y}_{ij+})^2}_{\text{SS[E]}};$$

i.e., SS[TOT] = SS[A] + SS[B(A)] + SS[E]. Computing formulae for the sums of squares are given on p. 639, Rao. The ANOVA table for a balanced two-factor nested experiment is given in Table 10.38. Here, $N = abn$. Since there are a levels of Factor A, there are $a-1$ degrees of freedom. For each level of A, there are $b-1$ degrees of freedom for Factor B. Thus, in all, there are $a(b-1)$ degrees of freedom for $B(A)$, across all levels of A. *Since every level of B does not appear with every level of A, we can not explicitly compute an interaction between A and B.*

F STATISTICS: We can test for effects in $a \times b$ nested experiments by using F statistics. In particular, the statistic

$$F_A = \frac{\text{MS[A]}}{\text{MS[E]}} \sim F_{a-1,N-ab}.$$

is used to test $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$ (no A main effect) versus H_1 : not H_0. The statistic

$$F_{B(A)} = \frac{\text{MS[B(A)]}}{\text{MS[E]}} \sim F_{a(b-1),N-ab},$$

is used to test $H_0 : \beta_{1(i)} = \beta_{2(i)} = \cdots = \beta_{b(i)} = 0$ (no B main effect within ith level of A), for all $i = 1, 2, ..., a$, versus H_1 : not H_0. Rejecting H_0 here means that, within at least one level of A, different levels of B do not have the same effect.

PLAN OF ATTACK: In nested experiments, it is suggested to start by looking at $F_{B(A)}$; that is, first test whether or not there is a B main effect nested within the levels of A.

- if the B main effect is significant, a test for the main effect of A would be less meaningful because it ignores the fact that different levels of A use different levels of B. In this situation, one could treat the problem as a one-way ANOVA with ab treatments and analyse the data accordingly; e.g., form pairwise intervals, contrasts, etc. I would tend to examine the levels of B, separately, within each level of A.
- if the B main effect is not significant, it is safe to test for the main effect of A. In this situation, I would immediately proceed to comparing the levels of A using pairwise intervals or contrasts.

ANALYSIS OF YIELD DATA: Here is the ANOVA table for the yield data (for the different alcohols and bases) in Example 10.4:

Table 10.39: *ANOVA table for the yield data in Example* 10.4.

Source	df	SS	MS	F
Alcohol	2	471.823	235.912	23.05
Base(Alcohol)	3	292.854	97.618	9.54
Error	18	184.243	10.236	
Total	23	948.920		

Since $F_{AB} = 9.54$ is large (e.g., $F_{3,18,0.05} = 3.160$), this suggests that bases affect the mean yield differently within at least one level of alcohol; that is, $H_0 : \beta_{1(i)} = \beta_{2(i)} = 0$, for all $i = 1, 2, 3$, is rejected. With this information, few would be bold enough to try to understand what $F_A = 23.05$ is saying. *Are the alcohols really different, or, is F_A large because different bases are used with each alcohol?* Unfortunately, since each alcohol uses different bases, it is impossible to tell.

CONTINUING THE ANALYSIS: To further understand where the differences are, let's compare the levels of base (B) within each level of alcohol (A). For example, to estimate $\mu_{12}-\mu_{11}$, the difference of the a_1b_2 and a_1b_1 treatment means, we would use $\overline{Y}_{12+}-\overline{Y}_{11+}$. Straightforward calculations show that (verify!)

$$E(\overline{Y}_{12+}-\overline{Y}_{11+}) = \beta_{2(1)}-\beta_{1(1)}$$

and that

$$V(\overline{Y}_{12+}-\overline{Y}_{11+}) = \frac{2\sigma^2}{4}.$$

Furthermore, $\overline{Y}_{12+}-\overline{Y}_{11+}$ is normally distributed since it is a linear combination of the Y_{ijk}s. Thus, a 95 percent confidence interval for $\beta_{2(1)}-\beta_{1(1)}$ would be

$$(\overline{Y}_{12+}-\overline{Y}_{11+}) \pm t_{18,0.025}\sqrt{\frac{2\text{MS[E]}}{4}}.$$

With $\overline{y}_{11+}=91.75$, $\overline{y}_{12+}=90.025$, $t_{18,0.025}=2.1009$, and $\text{MS[E]}=10.235$, the confidence interval for $\beta_{2(1)}-\beta_{1(1)}$ is $(-3.03, 6.48)$. Thus, there is not a significant difference between the two bases within the first level of alcohol. We could construct intervals for $\beta_{2(2)}-\beta_{1(2)}$ and $\beta_{2(3)}-\beta_{1(3)}$ in an analogous manner (worrying about multiplicity if we want to make joint statements). Alternatively, we could code the following statements in SAS:

```
contrast 'b2-b1 within a1' base(alcohol) -1 1 0 0 0 0;
contrast 'b4-b3 within a2' base(alcohol) 0 0 -1 1 0 0;
contrast 'b6-b5 within a3' base(alcohol) 0 0 0 0 -1 1;
```

These statements allow us to test $H_0: \beta_{2(1)}-\beta_{1(1)}=0$, $H_0: \beta_{2(2)}-\beta_{1(2)}=0$, and $H_0: \beta_{2(3)}-\beta_{1(3)}=0$, respectively. The output is given below:

```
    Contrast       DF  Contrast SS  Mean Square     F    Pr > F
---------------------------------------------------------------
b2-b1 within a1    1       5.951        5.951     0.58   0.4556
b4-b3 within a2    1       0.101        0.101     0.01   0.9219
b6-b5 within a3    1     286.801      286.801    28.02   <.0001
```

ANALYSIS: Thus, it looks as though bases 5 and 6 (those within the third level of alcohol) are significantly different. We have already seen that bases 1 and 2 are not different; neither are bases 3 and 4. Notice what we have done here; namely, we have broken up SS[B(A)] into the sums of squares for three **orthogonal** contrasts. Note that SS[B(A)] $= 292.854 = 5.951 + 0.101 + 286.801$ (up to rounding error).

NOTE: Had there been no main effect for base (within alcohol), it would have been acceptable to examine the main effect of alcohol. Had this been the reality, $F_A = 23.05$ suddenly becomes more meaningful; it tells us that the alcohol means are different (but, of course, it doesn't tell us *how* they are different!). Constructing pairwise intervals for the alcohol means would tell us where the differences are. Doing this in the presence of a base (B) main effect would give you little information about the levels of alcohol.

EXPECTED MEAN SQUARES IN THE TWO-FACTOR NESTED MODEL: Consider our two-factor nested model; i.e.,

$$Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \epsilon_{ijk},$$

for $i = 1, 2, ..., a$, $j = 1, 2, ..., b$, and $k = 1, 2, ..., n_{ij}$, where $\epsilon_{ijk} \sim$ iid $\mathcal{N}(0, \sigma^2)$. If we impose the "usual" side conditions $\alpha_+ = \beta_{+(i)} = 0$ (see p. 637, Rao), then it follows that

$$\begin{aligned} E(\text{MS[A]}) &= \sigma^2 + bn(a-1)^{-1}\sum_{i=1}^{a}\alpha_i^2 \\ E(\text{MS[B(A)]}) &= \sigma^2 + an(b-1)^{-1}\sum_{i=1}^{a}\sum_{j=1}^{b}\beta_{j(i)}^2 \\ E(\text{MS[E]}) &= \sigma^2. \end{aligned}$$

USING THE EXPECTED MEAN SQUARES: Again, we can see the usefulness of these expressions. For example, if $H_0 : \beta_{j(i)} = 0$, for all $i = 1, 2, ..., a$, (no B main effect) is true, then $E(\text{MS[B(A)]}) = E(\text{MS[E]}) = \sigma^2$, in which case the ratio $F_{ABC} = \text{MS[B(A)]}/\text{MS[E]}$ should be close to one. If H_0 is not true, then we would expect $F_{B(A)}$ to be larger than one. For each F test, we see that MS[E] is the "right" denominator. Here, we are assuming that both A and B are **fixed** factors. If B was **random** (coming up!), then F_A would not use MS[E] as a denominator; F_A would use MS[B(A)] as a denominator!

11 Some Comments on Designing Experiments

11.1 Introduction

We now revisit some of the key issues involved with designing experiments.

TERMINOLOGY: An **experiment** is an investigation set up to provide answers to a research question (or questions) of interest. In our context, an experiment is most likely to involve a comparison of **treatments** (e.g., fertilizers, rations, drugs, methods, varieties, etc.). The outcome of an experiment is information in the form of observations on a **response**, Y (e.g., yield, number of insects, weight gain, length of life, etc.).

PREVAILING ISSUE: Because of uncertainty in the responses due to sampling and biological variation, we can **never** provide definitive answers to the question(s) of interest based on such observations. However, we can make inferences that incorporate and quantify inherent uncertainty.

SOME IMPORTANT THEMES: We have mentioned some of these ideas already, but their importance cannot be overemphasised.

- Before an experiment may be designed, the question(s) of interest must be well-formulated. Nothing should start (including data collection) until this happens.
- The investigator and statistician should work together, before data collection begins, to identify important features and the appropriate design.
- The design of the experiment lends naturally to the analysis of the data collected from the experiment.
- If the design changes, the analysis will as well. If experiments are designed appropriately, the analysis should be routine. If not, researchers run the risk of not being able to address any of their main questions.

11.2 Roles of investigator and statistician

The investigator and statistician may have different perspectives on an experiment. If they work together from the initial planning stage, the investigation can provide the information necessary to give insight into important question(s) of scientific interest.

THE INVESTIGATOR:

- *Formulate broad questions of interest.* The first step in experimental design is to decide what the problem is! In almost all situations, this is the domain of the investigator. The statistician is generally not in a position to judge this.
- *Decide which comparisons are relevant.* Once the general questions of interest have been determined, the investigator must decide what issues are to be the focus of the experiment. Often, one experiment can not answer all the questions of interest.
- *Decide on a meaningful difference.* The investigator, based on scientific aspects, should have a general idea of what kind of differences among treatments are important. This knowledge is usually required to determine sample sizes.
- *Identify the resources available.* The investigator should have an idea of limitations that may be imposed (e.g., money, time, space, personnel, etc.). This almost always will have an effect on how the experiment is designed.
- *Identify any peculiarities associated with the situation.* The investigator should think of anything that might have an effect on how the experiment may be carried out. For example, if an experiment is to be conducted in a greenhouse, and there is an air conditioner at one end dripping condensation that might systematically affect the outcome for plants near it, the statistician needs to know this peculiarity! If informed beforehand, the statistician can design the experiment around this peculiarity (i.e., design the experiment to control this source of variability).
- *Decide on the scope of interest.* The investigator should have some sense of the desired applicability of the results from the experiment.

THE STATISTICIAN:

- *Identify the relevant population.* The ultimate objective for the statistician is to cast the problem in a formal statistical model framework. The first step is to identify the population(s) from which sampling will be required.
- *Identify an appropriate probability model.* Based on the type of response (e.g., discrete, continuous, etc.) to be observed, the statistician must determine how to represent the populations in terms of probability models.
- *Cast the question of interest as statistical hypotheses.* With the chosen probability model, express the scientific questions of interest in terms of population parameters.
- *Design the experiment.* Taking into account the limitations on resources, peculiarities, and meaningful scientific differences to be identified, determine an appropriate plan to sample from the population(s). This involves assigning the treatments to experimental units in such a way that (a) samples are representative of the population, (b) no confounding or bias is possible, (c) the relevant comparisons may be addressed, (d) meaningful differences can be detected with sufficient power.
- *Recognise possible shortcomings.* It may very well be the case that some or all research questions can not be suitably addressed with the available resources. In this situation, the statistician must be up front and honest about this issue.

A COMMON OCCURRENCE: In many situations, the statistician may come up with a design that the investigator realises is impossible to carry out (perhaps due to unmentioned peculiarities, constrained resources, unreasonable randomisation protocols, etc.). Taking this into consideration, the statistician may have to design the experiment differently. Also, the statistician may determine that the available resources are not sufficient to detect differences with the desired precision. Based on this consideration, the investigator may decide to scale back the scope of inference or seek additional resources. *It should be clear that to maximize the usefulness of an experiment, the investigator and statistician should work together from the outset.*

11.3 Statistical issues in experimental design

TERMINOLOGY: A **treatment** is a procedure whose effect is to be measured and compared with other procedures. The **experimental unit** is the unit of experimental material to which one application of the treatment is applied. The **experimental design** is a plan for applying the treatments to experimental units in such a way that experimental units are alike except for the treatments.

A SIMPLE SETTING: Suppose that we are comparing two treatments (e.g., drugs) in a balanced one-way layout. A statistical model for this situation is $Y_{ij} = \mu + \tau_i + \epsilon_{ij}$, for $i = 1, 2$ and $j = 1, 2, ..., n$, where $\epsilon_{ijk} \sim$ iid $\mathcal{N}(0, \sigma^2)$. Recall that our test statistic is

$$t = \frac{\overline{Y}_{1+} - \overline{Y}_{2+}}{\sqrt{2\text{MS[E]}/n}},$$

where MS[E] is the pooled variance estimator of σ^2. The denominator of t; i.e., the estimated standard error of $\overline{Y}_{1+} - \overline{Y}_{2+}$, measures how precisely we can estimate the difference between the two treatment means. It depends on the two components; namely, the number of replicates, n, and σ^2, the experimental error variance. *These are the two key aspects that must be considered when designing any experiment.* Here we are only considering the simple case of two treatments in a one-way layout; however, these two issues arise generally in more elaborate designs.

REPLICATION: In the two-sample case, the sample size n represents the number of experimental units (i.e., the number of replicates) seen on each treatment.

- Increasing the number of replicates, in general, decreases the standard error. This increases precision and our ability to detect treatment differences. Any experimental design must have a sufficient number of experimental units on each treatment. This is well under our control and limited only by the resources available.
- The number of replicates per treatment is a key factor in determining precision and power. If we have a fixed number of experimental units available, part of the design is how to make the best use of them for detecting treatment differences.

- It should be clear that, under these conditions, we would be better off with a few treatments and many replicates instead of many treatments and only a few replicates on each. The same total number of experimental units can lead to two very different designs: one that is powerful in detecting differences, and one that is not. If limited resources are available, it is better to reduce the number of treatments to be considered or postpone the experiment rather than proceed with too few replicates.

EXPERIMENTAL ERROR: Consider the model for our balanced one-way layout; i.e., $Y_{ij} = \mu + \tau_i + \epsilon_{ij}$, for $i = 1, 2$ and $j = 1, 2, ..., n$, where $\epsilon_{ijk} \sim$ iid $\mathcal{N}(0, \sigma^2)$.

- The term μ represents the mean of responses for all experimental units before treatment application. The term τ_i represents the effect of treatment i. Together, these components characterise the two treatment means μ_1 and μ_2. The treatment mean is fixed for each treatment and does not vary. What does vary are the observations Y_{ij} because of inherent biological differences in the sample experimental units to which the treatments are applied.
- The error term, ϵ_{ij} characterises the inherent variation in experimental units that makes them yield different responses. The (unknown) value σ^2 characterises the variation in the ϵ_{ij} values in the population of experimental units.

REALISATION: We should see that precision of estimation and power for testing differences depends on the inherent variation in the experimental units. If this variation is large, our ability to provide good estimates and detect differences may be limited. Inspection of $2\sigma^2/n$ (the standard error) shows that one way we might try to overcome this variability is increase the number of experimental units per treatment to make $2\sigma^2/n$ small. Unlike replication, we cannot control all variation. There is always variation attributable to the experimental units used in the experiment. However, by being “clever,” paying careful attention to the nature of the experimental material, we may be able to reduce the effects of inherent variation by designing the experiment appropriately.

EXPERIMENTAL ERROR: If we wish to compare treatments, the experimental unit is the relevant unit of experimental material when assessing the available information. The experimental error measures mainly inherent variation among the experimental units. To "be clever," consider the following key points:

- If we wish to detect differences among treatments, then we hope that most of the variability in results is due to the systematic effect of treatments rather than the experimental error.
- *If we could reduce the magnitude of experimental error somehow, we would be in a better position to detect differences among the treatments.* As we have just discussed, we can not eliminate variation completely. But we can think of how it arises and then design the experiment in light of this information. For example, if we are trying to compare two drugs, the experimental units would be subjects to whom we administer the drugs. Subjects may vary in their responses to the drugs because they are just inherently different, but they may also differ in their responses for systematic reasons, such as gender, age, condition, etc. Thus, part of the variation in the experimental units may be due to **systematic** causes we can identify (and incorporate into the design).
- If we can attribute some of the variation in experimental units to systematic sources, we could reduce the effect of the inherent variation among them. That is, we could reduce our assessment of experimental error.

BLOCKING: The obvious strategy is to set up the experiment so that the systematic variation among experimental units may be separated from the inherent variation. If we group experimental units according to systematic features they share (i.e., **block**), we can hopefully explain part of the variation that we observe across groups. The remaining variation will be that arising within groups of experimental units. This variation would comprise experimental error. Because the units in groups are apt to be "more alike," the hope is that experimental error will be smaller.

RESULT: Experimental design is founded on the principle of reducing experimental error by meaningful grouping of experimental units. Although we can not eliminate inherent variability completely, we can try to be careful about what we consider to be inherent variability by identifying all the possible systematic components of it.

Example 11.1. Current nitrogen fertilization recommendations for wheat include applications of specified amounts at specified states of plant growth. An experiment was carried out which included $t = 6$ different nitrogen application timing and rate schedules. The experiment was conducted in an irrigated field with a water gradient along one direction of the experimental plot area as a result of irrigation. Since plant responses are affected by variability in the amount of available moisture, the field plots were grouped into four blocks, each consisting of the six plots (one plot per treatment), such that each block occurred in the same part of the water gradient. In this example, the experimenter recognises the fact that the water gradient in the field is a **systematic** source of variation. Thus, by making sure each block of land occurs in the same part of the gradient, responses within a given block should be "more alike," than, say, responses from other blocks. The researcher has designed the experiment to **control** this source of variation. Had she not, then, resulting differences in the treatments (rates/schedules) would be confounded by the effect of the water gradient.

STATISTICAL MODELS: In Example 12.1, if we had ignored the effect of the water gradient, we could have adopted the usual one-way layout model to compare the six nitrogen fertilizers; i.e.,

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij},$$

for $i = 1, 2, ..., 6$ and $j = 1, 2, 3, 4$, where $\epsilon_{ijk} \sim \text{iid}\, \mathcal{N}(0, \sigma^2)$. Here, the experimental error, σ^2, includes all sources of variability not explained by the treatments, including the source corresponding to the water gradient. An alternative model, which incorporates water gradient variability, is the two-factor no-interaction ANOVA model

$$Y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij},$$

for $i = 1, 2, ..., 6$ and $j = 1, 2, 3, 4$, where $\epsilon_{ijk} \sim \text{iid}\, \mathcal{N}(0, \sigma^2)$. The term β_j describes the effect of the jth block. Insight is gained by examining these two models. In the one-way model, the variability due to the different blocks is absorbed into σ^2, the experimental error variance. If there is variation among the four blocks, the two-way model explains more of the total variability, and, thus, the experimental error variance σ^2 is reduced.

RANDOMISATION: In terms of experimental design, **randomisation** involves the assignment of treatments to experimental units, based on the chosen design, by a chance mechanism. The purpose is to ensure that no treatment is somehow favoured, or handicapped, so that the replicates receiving each treatment are representative of the population except for the treatments. Non-random assignment usually leads to a biased design plagued with confounding effects. As we have also discussed, randomisation ensures that observations represent random samples from populations of interest. This ensures the validity of statistical methods.

COMPLETE RANDOMISATION VERSUS RESTRICTED RANDOMISATION: In the one-way layout model, **complete randomisation** is assumed; i.e., each fertilizer is to be randomly assigned to the experimental units (this is a **completely randomised design**). In this case, it is possible that four adjacent plots occurring in the same part of the water gradient all receive the same fertilizer! In the two-way ANOVA model, which is used to incorporate the blocks, experimental units are assumed to be randomly assigned to plots *within each block*; this is called **restricted randomisation**. In each block, each treatment is represented only once (this is called a **randomised complete block design**).

NOTE: Remember, the goal of any experiment is to compare the treatments (not the blocks). By ensuring that treatments are represented in each block, where the experimental units are more homogeneous, we enhance our ability to detect differences among the treatments. Restricted randomisation protocols confer this benefit. Complete randomisation does not confer this; i.e., treatment differences could be "masked" by the true differences that exist among the blocks. Of course, if there is no true variation among the blocks, then using a restricted randomisation protocol does not deliver any benefits.

11.4 Experimental unit versus sampling unit

In the preceding discussion, we have seen how important the notions of experimental units and replication are to experimental design. Correctly identifying the relevant experimental unit is one of the most important aspects of design. In many experiments, confusion may arise. In our work so far, we have considered only cases where a single observation of the response is made on each experimental unit; however, it is not uncommon to take more than one observation on an experimental unit (this is called **subsampling**).

TERMINOLOGY: The **sampling unit** is the fraction of the experimental unit upon which a single observation is made. To understand the distinction between experimental units and sampling units, and its implications for design, consider the following examples:

	Treatment	Experimental Unit	Sampling Unit	Response
(i)	Food rations	20 swine in a pen	a single pig	weight gain
(ii)	Insecticides	50 plants on a plot	a single plant	# of insects
(iii)	Drugs	a single patient	a single patient	survival time
(iv)	Variety	3 row plots of plants	a single row	yield

In (iii), the experimental unit and the sampling unit are the same. In the others:

(i) It is common to confine animals to the same pen or cage. Thus, it is simpler to feed the entire pen the same ration rather than to feed them individually. However, it is logical to observe a separate weight gain for each animal. Because the whole pen receives the same ration, it constitutes an experimental unit. Weight gains are measures on each pig within the pen, thus, they constitute the sampling units.

(ii) Similarly, it is easier to spray a whole plot with insecticide rather than individual plants, but logical to count insects on each plant.

(iv) It is logistically simpler to plant large areas with the same variety; here, a 3-row plot. However, the rows (within plot) may be sufficiently separated so that it is possible to measure yield on each row separately.

KEY PRINCIPLE: It is the **experimental unit** that is the relevant unit of experimental material to consider when assessing available information, not the sampling unit. *Many of the disasters that occur in planning and analysing studies occur because people misunderstand this difference.* Consider the following example.

Example 11.2. A graduate student wanted to study the effects of two drugs on mice. He collected 200 observations in the following way. Two mice were randomly assigned to each drug. From each mouse, tissue samples were collected at 50 sites. The **experimental units** were the mice because drugs were applied to the mice, not the tissue sites. There are two sources of variation: mouse-to-mouse variation and within-mouse variation.

- The 50 observations (subsamples) on each mouse greatly reduce the within-variation but do nothing to reduce the mouse-to-mouse variation. Relative to the mouse-to-mouse variation, there are only two observations that have the same treatments. Thus, each of the two treatment groups provides only one degree of freedom for estimating the variance that applies to treatment comparisons; that is, the experiment provides two degrees of freedom for the experimental error!
- In this example, we may know quite a bit about each mouse, with 50 observations on each. But, we know almost nothing about how the treatments compare in the population of such mice. We have only seen two mice on each treatment, so we have only two observations per treatment.
- It is a common mistake for investigators to use too few replicates per treatment, but take many subsamples on each. This is likely due to confusion about what is meant by **replication**. The number of replicates (and not the number of subsamples per replicate), is the important feature for determining precision and power. Thus, it is better to have a large number of experimental units and a small number of sampling units on each, rather than vice versa.

REMEMBER: Statisticians can not, in general, save an experiment that was performed poorly or that was inundated with faulty technique.

12 Random Effects Models and Subsampling

Complementary reading from Rao: Chapter 14 (§ 14.1-14.3 and § 14.6).

12.1 Introduction

Up until now, in our discussion of ANOVA models, we have assumed that factor levels are **fixed**; that is, the levels used are the only levels of interest, and if we were to repeat the experiment again, we would use the same levels. However, in some situations, it may be more reasonable to assume that the levels of a factor, in fact, constitute a random sample from a larger population of levels. In this case, we say that the factor is **random**.

12.2 Random effects in the one-way layout

ONE-WAY LAYOUT: Recall our effects model for the one-layout (with no subsampling)

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij},$$

for $i = 1, 2, ..., t$ and $j = 1, 2, ..., n_i$, where $\epsilon_{ij} \sim$ iid $\mathcal{N}(0, \sigma^2)$. In this model, recall that μ denotes the overall mean, and that τ_i denotes the effect of receiving treatment i. Consider each of the following one-way layout experiments.

Example 12.1. In a potato-breeding experiment, we are interested in $t = 3$ different fertilizer mixtures for which mean yields are to be compared (for one specific variety of potatoes). Mixture 1 contains 80 percent sodium nitrate ($\mathrm{NaNO_3}$), Mixture 2 contains 90 percent $\mathrm{NaNO_3}$, and Mixture 3 contains 100 percent $\mathrm{NaNO_3}$. Here, we are interested in comparing three specific treatments. If we repeated the experiment again, we would still be interested in these three mixtures. In this example, there is a particular (or fixed) set of treatments of interest; namely, the three mixtures of fertilizer. These levels of fertilizer are best regarded as **fixed**.

Example 12.2. Suppose that in our potato-breeding experiment, interest lies not in comparing different levels of fertilizer, but rather comparing the yields for multiple varieties for a fixed amount of $NaNO_3$. There are thousands of varieties that are available for study, so we take a random sample of $t = 5$ varieties and use those in the experiment. The hope is that the results for the five varieties involved in the experiment may be generalised to gain insight into the behaviour of all varieties. In this example, the varieties are best regarded as **random**.

DISTINCTION: The model $Y_{ij} = \mu + \tau_i + \epsilon_{ij}$ may be appropriate for each situation in Examples 12.1 and 12.2. However, there is an important distinction to be made:

- In Example 12.1, the particular treatments (mixtures) are the only ones of interest, so there is no uncertainty involved with their selection. Thus, the τ_i are regarded as **fixed** quantities, as they describe a particular set of fertilizers. In this situation, the τ_i are referred to as **fixed effects**.

- In Example 12.2, there is additional uncertainty involved, because the treatments (varieties) are no longer fixed; they are chosen at random from a large population of treatments. The τ_i are regarded as **random** and are referred to as **random effects**. Thus, it makes sense to think of the τ_i as random variables with variance, say, σ_τ^2. This variance σ_τ^2 characterises the variability in the population of all possible treatments; in our example, the variability across all possible potato varieties. If varieties are quite different in terms of yield, σ_τ^2 will be large. If yields are consistent across varieties, σ_τ^2 will be small.

TERMINOLOGY: In **fixed effects models** (i.e., models where factors are treated as fixed), we are interested in making statements about treatment means. In **random effects models** (i.e., models where factors are treated as random), we do not care about the means of the particular treatments. We are trying to make a statement about the entire population of treatments based only on those we use in the experiment. **Mixed-effects models** are models that include both fixed and random factors.

Table 12.40: *ANOVA table for the balanced one-way layout with fixed or random effects.*

Source	df	SS	MS	F
Treatments	$t-1$	SS[T]	MS[T]	$F = \frac{\text{MS[T]}}{\text{MS[E]}}$
Error	$N-t$	SS[E]	MS[E]	
Total	$N-1$	SS[TOT]		

IMPORTANT RESULT: In assessing treatment differences for the one-way layout, the ANOVA computations are the same regardless of whether we are dealing with fixed or random effects. *The interpretation, however, will be different!!* Since the computations are the same, the calculation formulae from Chapter 2 apply; the ANOVA table for the one-way random effects model is shown in Table 12.40.

HYPOTHESIS TESTS IN THE ONE-WAY LAYOUT: Consider the one-way effects model $Y_{ij} = \mu + \tau_i + \epsilon_{ij}$, for $i = 1, 2, ..., t$ and $j = 1, 2, ..., n_i$, where $\epsilon_{ij} \sim$ iid $\mathcal{N}(0, \sigma^2)$. With **fixed** effects, we are interested in comparing the treatment means. Since the τ_is are regarded as fixed quantities, we are interested in testing

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_t = 0$$

$$\text{versus}$$

$$H_1 : \text{not } H_0.$$

In the **random** effects case, we are interested in the entire population of treatments. In this situation, we assume that $\tau_i \sim$ iid $\mathcal{N}(0, \sigma_\tau^2)$ and that τ_i and ϵ_{ij} are independent random variables. If there are no differences among treatments, then $\sigma_\tau^2 = 0$. Thus, with random effects, we write the hypotheses as

$$H_0 : \sigma_\tau^2 = 0$$

$$\text{versus}$$

$$H_1 : \sigma_\tau^2 > 0.$$

Table 12.41: *EMS for the balanced one-way layout with fixed or random effects.*

		Expected mean squares (EMS)	
Source of variation	df	Fixed effects	Random effects
Treatments	$t-1$	$\sigma^2 + \frac{n\sum_{i=1}^{t}\tau_i^2}{t-1}$	$\sigma^2 + n\sigma_\tau^2$
Error	$t(n-1)$	σ^2	σ^2

NOTE: In either case, we judge the amount of evidence against H_0 by comparing F to a $F_{t-1,N-t}$ distribution. As always, large values of F are not consistent with H_0.

EXPECTED MEAN SQUARES: Table 12.41 shows the expected mean squares for both models (the fixed-effects case assumes the $\tau_+ = 0$ side condition) when the design is balanced; i.e., $n_i = n$, for all i. In the **fixed** effects case, when $H_0 : \tau_1 = \tau_2 = \cdots = \tau_t = 0$ is true, or, in the **random** effects case, when $H_0 : \sigma_\tau^2 = 0$ is true, MS[T] and MS[E] both estimate the same quantity, and, hence, $F =$ MS[T]/MS[E] should be close to one. In both situations, when H_0 is not true, we would expect F to get large. In unbalanced designs, the formulae for expected mean squares changes to incorporate the unbalance (see p. 651-2 in Rao), but this general notion of F statistics getting large (or staying close to one) does not.

Example 12.3 (`coal.sas`). Core soil specimens are taken in each of six locations within a territory being investigated for surface mining of bituminous coal. A random sample of $n = 4$ specimens is taken from each of $t = 6$ randomly-selected locations and the sulfur content (Y) is observed for each specimen. *Researchers want to know if different locations are associated with different sulfur contents.* The data are in Table 12.42. The ANOVA table for these data, computed with the help of SAS, is below.

Source	df	SS	MS	F
Location	5	91.398	18.280	56.20
Error	18	5.856	0.325	
Total	23	97.254		

Table 12.42: *Sulfur content data.*

	Specimen			
Location	1	2	3	4
A	15.2	16.8	17.5	16.2
B	13.1	13.8	12.6	12.9
C	17.5	17.1	16.7	16.5
D	18.3	18.4	18.6	17.9
E	12.8	13.6	14.2	14.0
F	13.5	13.9	13.6	14.1

ANALYSIS: In this example, researchers are most likely to be interested in a larger population of locations, from which these 6 were randomly selected. Thus, in the context of our one-way layout model, $Y_{ij} = \mu + \tau_i + \epsilon_{ij}$, the τ_i's (i.e., the effects due to the differences among locations) are best regarded as **random**. The test of $H_0 : \sigma^2_\tau = 0$ would be rejected since F is too large ($P < 0.0001$). It looks like there is significant variability among the locations in terms of their sulfur contents.

12.3 Variance components in the one-way layout

In this subsection, we introduce variance components in the one-way layout.

VARIANCE COMPONENTS: Consider the one-way balanced random-effects model $Y_{ij} = \mu + \tau_i + \epsilon_{ij}$, for $i = 1, 2, ..., t$ and $j = 1, 2, ..., n$, where $\tau_i \sim \text{iid}\, \mathcal{N}(0, \sigma^2_\tau)$, $\epsilon_{ij} \sim \text{iid}\, \mathcal{N}(0, \sigma^2)$, and τ_i and ϵ_{ij} are independent. In this model, since both τ_i and ϵ_{ij} are random variables, the variance of any observation Y_{ij} is

$$\sigma^2_Y = V(Y_{ij}) = \sigma^2_\tau + \sigma^2.$$

The variances σ^2_τ and σ^2 are called **variance components**. As we see, there are two components of variability; one that is due to the differences among possible treatments (σ^2_τ) and one that is due to differences among experimental units (σ^2).

ESTIMATION: We would like to estimate the variance components in our one-way random-effects model. In the last subsection, we learned that, in balanced designs,

$$\begin{aligned} E(\text{MS[T]}) &= \sigma^2 + n\sigma_\tau^2 \\ E(\text{MS[E]}) &= \sigma^2. \end{aligned}$$

If we use MS[E] as an estimate of E(MS[E]) and MS[T] as an estimate of E(MS[T]) (this is basically the **method-of-moments** approach to estimation), we can see that

$$\begin{aligned} \widehat{\sigma}_\tau^2 &= \frac{\text{MS[T]} - \widehat{\sigma}^2}{n} \\ \widehat{\sigma}^2 &= \text{MS[E]}. \end{aligned}$$

The values $\widehat{\sigma}_\tau^2$ and $\widehat{\sigma}^2$ are called **variance component estimators**.

REMARK: Occasionally, this method of estimation can produce a negative estimate for σ_τ^2; i.e., whenever MS[T] is smaller than MS[E]. Clearly, any population variance must be nonnegative, so a negative estimate of it is viewed with some concern. One course of action is to accept the estimate and use it as evidence that the true value of σ_τ^2 is close to zero, assuming that inherent sampling variability led to the negative estimate. This has an intuitive appeal, but it suffers from some theoretical difficulties. Personally, I might consider this as evidence that the one-way model is, perhaps, not appropriate for the data and reexamine the problem.

UNBALANCED DATA: Formulae for expected mean squares and variance component estimators, in unbalanced experiments, appears on p. 651-3 Rao. The only thing that changes in the estimates is that the n in the denominator of $\widehat{\sigma}_\tau^2$ is replaced by

$$n_0 = (t-1)^{-1}\left(N - \frac{\sum_i n_i^2}{N}\right).$$

COEFFICIENT OF VARIATION: In some applications, it is of interest to get an idea of variability on a **relative** basis; that is, we would like to have a unitless measure that describes the variation in the data. The use of the **coefficient of variation** lies partly in the fact that *the mean and standard deviation tend to change together in many* *experiments.* The coefficient of variation is a relative measure that expresses variability as a percentage of the mean. In the one-way random effects model, the population coefficient of variation (CV) is defined to be

$$\text{CV} = \frac{\sigma_Y}{|\mu_Y|} = \frac{\sqrt{\sigma_\tau^2 + \sigma^2}}{|\mu_Y|}.$$

An estimate of population CV is given by the sample CV; i.e.,

$$\widehat{\text{CV}} = \frac{\sqrt{\widehat{\sigma}_\tau^2 + \widehat{\sigma}^2}}{|\overline{Y}_{++}|}.$$

REMARK: CV is a useful quantity for comparing the results from different experiments, but it can also be useful when attention is focused on a single experiment. It provides an impression of the amount of variation in the data relative to the mean size of the characteristic being measured. If CV is large, it is an indication that we will have a difficult time learning about the "signal" (μ_Y) relative to the "noise" in the data (σ_Y). Confidence interval estimation for the population CV is possible, but deriving a formula for the interval is quite difficult.

INTRACLASS CORRELATION COEFFICIENT: In the one-way random effects model, the **intraclass correlation coefficient** is defined as

$$\rho_I = \frac{\sigma_\tau^2}{\sigma^2 + \sigma_\tau^2}.$$

The parameter ρ_I is the proportion of variability of the observed responses that can be attributed to the variation in the treatment effects. For example, if $\rho_I = 0.8$, then 80 percent of the variability in the data is due to the differences among treatments. An estimate of ρ_I is given by

$$\widehat{\rho}_I = \frac{\widehat{\sigma}_\tau^2}{\widehat{\sigma}^2 + \widehat{\sigma}_\tau^2}.$$

CONFIDENCE INTERVAL: Unlike the CV, finding a confidence interval for ρ_I is not too difficult. In the balanced one-way random effects model, it follows that

$$\frac{(t-1)\text{MS[T]}}{\sigma^2 + n\sigma_\tau^2} \sim \chi^2_{t-1}.$$

It is also not difficult to show that

$$\frac{(N-t)\text{MS[E]}}{\sigma^2} \sim \chi^2_{N-t},$$

and that MS[T] and MS[E] are independent statistics. These facts enable us to create a pivotal quantity; namely,

$$\frac{\text{MS[T]}/(\sigma^2 + n\sigma^2_\tau)}{\text{MS[E]}/\sigma^2} \sim F_{t-1,N-t}.$$

Hence, it follows that

$$P\left\{F_{t-1,N-t,1-\alpha/2} \le \frac{\text{MS[T]}/(\sigma^2 + n\sigma^2_\tau)}{\text{MS[E]}/\sigma^2} \le F_{t-1,N-t,\alpha/2}\right\} = 1 - \alpha.$$

Straightforward algebra shows that the last probability statement is equivalent to

$$P\left(\frac{L}{1+L} \le \rho_I \le \frac{U}{1+U}\right) = 1 - \alpha,$$

where

$$\begin{aligned} L &= \frac{1}{n}\left(\frac{\text{MS[T]}}{F_{t-1,N-t,\alpha/2} \times \text{MS[E]}} - 1\right) \\ U &= \frac{1}{n}\left(\frac{\text{MS[T]}}{F_{t-1,N-t,1-\alpha/2} \times \text{MS[E]}} - 1\right). \end{aligned}$$

Thus, $(L/(1+L), U/(1+U))$ is a $100(1-\alpha)$ percent confidence interval for the intraclass correlation coefficient ρ_I.

ESTIMATING THE OVERALL MEAN: Consider the one-way balanced random-effects model $Y_{ij} = \mu + \tau_i + \epsilon_{ij}$, for $i = 1, 2, ..., t$ and $j = 1, 2, ..., n$. In some situations, it may be of interest to write a confidence interval for μ, the overall mean. Using $\overline{Y}_{++}$ as a point estimator for μ, it is easy to see that, under our model assumptions, $E(\overline{Y}_{++}) = \mu$ and that $V(\overline{Y}_{++}) = (\sigma^2 + n\sigma^2_\tau)/nt$. Thus, standardising, we have

$$Z = \frac{\overline{Y}_{++} - \mu}{(\sigma^2 + n\sigma^2_\tau)/nt} \sim \mathcal{N}(0, 1).$$

Recalling that

$$\frac{(t-1)\text{MS[T]}}{\sigma^2 + n\sigma^2_\tau} \sim \chi^2_{t-1},$$

it follows that

$$t = \frac{\overline{Y}_{++} - \mu}{\text{MS[T]}/nt} = \frac{\frac{\overline{Y}_{++} - \mu}{(\sigma^2 + n\sigma^2_\tau)/nt}}{\sqrt{\frac{(t-1)\text{MS[T]}}{\sigma^2 + n\sigma^2_\tau}/(t-1)}} \sim t_{t-1}.$$

Thus, a $100(1-\alpha)$ percent confidence interval for the overall mean μ is given by

$$\overline{Y}_{++} \pm t_{t-1,\alpha/2}\sqrt{\text{MS[T]}/nt}.$$

Example 12.4. With the sulfur data from Example 12.3, we compute MS[T] $= 18.280$, MS[E] $= 0.325$, and $\overline{y}_{++} = 15.367$. The variance component estimates are given by

$$\begin{aligned} \widehat{\sigma}^2_\tau &= \frac{18.280 - 0.325}{4} = 4.489 \\ \widehat{\sigma}^2 &= 0.325. \end{aligned}$$

The coefficient of variation for these data is given by

$$\widehat{\text{CV}} = \frac{\sqrt{\widehat{\sigma}^2_\tau + \widehat{\sigma}^2}}{|\overline{y}_{++}|} = \frac{\sqrt{4.489 + 0.325}}{15.367} = 0.143.$$

Thus, the total variation in the data is roughly 14.3 percent of the overall mean. An estimate of the intraclass correlation coefficient ρ_I is given by

$$\widehat{\rho}_I = \frac{\widehat{\sigma}^2_\tau}{\widehat{\sigma}^2 + \widehat{\sigma}^2_\tau} = \frac{4.489}{0.325 + 4.489} = 0.932.$$

Thus, for these data, 93.2 percent of the total variation is due to the differences among the treatments. Next we aim to construct a 95 percent confidence interval for the population intraclass correlation ρ_I. To get the confidence interval, first, we compute

$$L = \frac{1}{n}\left(\frac{\text{MS[T]}}{F_{t-1,N-t,\alpha/2} \times \text{MS[E]}} - 1\right) = \frac{1}{4}\left(\frac{18.280}{3.3820 \times 0.325} - 1\right) = 3.908$$

$$U = \frac{1}{n}\left(\frac{\text{MS[T]}}{F_{t-1,N-t,1-\alpha/2} \times \text{MS[E]}} - 1\right) = \frac{1}{4}\left(\frac{18.280}{0.1572 \times 0.325} - 1\right) = 89.200.$$

Thus, the 95 percent confidence interval for ρ_I is

$$\left(\frac{3.908}{1+3.908}, \frac{89.200}{1+89.200}\right) \Longleftrightarrow (0.796, 0.989).$$

Thus, we are 95 percent confident that ρ_I is between 0.796 and 0.989. Finally, a 95 percent confidence interval for the overall mean sulfur content μ is given by

$$15.367 \pm t_{5,0.025}\sqrt{18.280/4(6)} \Longleftrightarrow (13.124, 17.610).$$

Thus, we are 95 percent confident that the true mean sulfur content μ is between 13.124 and 17.610.

12.4 One-way ANOVA models with subsampling

In the last chapter, we discussed the difference between an **experimental unit** and a **sampling unit**. The experimental unit is the element of experimental material that receives an application of the treatment; thus, it is the entity of interest for assessing experimental error. However, as we have already seen in the last chapter, it is not uncommon to have several sampling units (i.e., subsamples) on each experimental unit.

Example 12.5. In an agricultural experiment, suppose that we are administering different rations (treatments) to pigs. The location of the experiment is set up so that swine are kept in pens, each consisting of 20 animals. The pigs are fed by introducing a trough of ration into the pen. Here, the experimental unit is the pen, as the ration is applied to it. The weight gain for each pig is recorded at the end of the experiment. The sampling units are the individual pigs.

IMPORTANCE: In order to assess the effects of the treatments, we must be assured that treatments are applied in a way, so that we may attribute differences observed to the treatments. If, in Example 12.5, we had treated individual pigs as the experimental unit, we see that this assurance would not be fulfilled−once the ration is introduced to the pen, we have no control over how it is applied to individual pigs. Larger animals might “squeeze out” smaller ones, so that the ration is not applied equally to all pigs. Thus, different animals in the pen might exhibit different weight gains simply because they did not receive the same application of the treatment. Treating the pigs as experimental units would incorrectly assume that they all did receive the same application of the ration.

SUBSAMPLING: So far, we have discussed the analysis of variance procedure within the context of one observation per experimental unit; that is, the experimental and sampling units were the same. We now consider how the procedure might be extended to the case of more than one sampling unit per experimental unit. This situation is referred to as **subsampling**. To facilitate our discussion, we will assume, for simplicity, that a one-way classification is appropriate.

ONE-WAY ANOVA MODELS WITH SUBSAMPLING: When we have subsampling, we may classify an individual observation (on a sampling unit now) as the kth subsample, on the jth experimental unit, receiving treatment i; i.e.,

$$Y_{ijk} = \mu + \tau_i + \epsilon_{ij} + \delta_{ijk},$$

for $i = 1, 2, ..., t$, $j = 1, 2, ..., n_i$, and $k = 1, 2, ..., s_{ij}$, where τ_i is the effect associated with treatment i (τ_i could be fixed or random), $\epsilon_{ij} \sim$ iid $\mathcal{N}(0, \sigma^2_\epsilon)$, $\delta_{ijk} \sim$ iid $\mathcal{N}(0, \sigma^2)$, and the ϵ_{ij} and δ_{ijk} are independent.

NOTATION: In the one-way model with subsampling, t is the number of treatments (as before), n_i is the number of replicates (experimental units) on treatment i, and s_{ij} is the number of subsampling units (subsamples) on the jth experimental unit receiving treatment i. Note that there are two error terms! The term ϵ_{ij} is the error associated with the jth experimental unit receiving treatment i. The error ϵ_{ij} quantifies how this experimental unit varies in the population of all experimental units. The term δ_{ijk} is the additional error associated with the kth sampling unit, which we often refer to as **sampling error**. For simplicity, we will assume that the design is totally **balanced**; i.e., $n_i = n$ for all i, and $s_{ij} = s$, for all i and j.

FIXED VERSUS RANDOM TREATMENTS: As mentioned before, the τ_i can be treated as fixed or random effects. If they are random, it is customary to assume that $\tau_i \sim$ iid $\mathcal{N}(0, \sigma^2_\tau)$ and that τ_i, ϵ_{ij}, and δ_{ijk} are mutually independent random variables.

REMARK: Even though our model is more complex now, as we have altered it to incorporate data from experiments which use subsampling, it is important to remember that our main goal has not changed; namely, we are interested in comparing the treatments. If the τ_i are fixed, we want to compare the treatment means. If the τ_i are random, we want to examine the variation among treatments.

Example 12.6. Suppose an experiment is conducted to compare the effect of 5 different doses of a toxic agent (treatments) on the birth weights of rats. For each dose, several pregnant female rats are given the particular dose. Thus, the female rats are the **experi-**

mental unit, as they receive an application of the treatment. For each mother, suppose that birth weight is recorded for each rat pup. Thus, the rat pups are the **sampling units**. For a given mother rat, all of her pups are, of course, not identical; rather, they exhibit variability among themselves. Furthermore, mother rats vary inherently across themselves. In the model, ϵ_{ij} characterises the mother-to-mother variation. The δ_{ijk} characterises the additional variation that might be present because all rat pups on a given mother are not exactly alike.

- Recall that we think of experimental error as measuring the inherent variation in responses on experimental units; that is, the variation in the data that we attribute to things other than the systematic effects of the treatments. If we think about this variation in the current context, it is clear that there are now **two** sources of inherent variation that may make responses on experimental units differ:
  - variation among experimental units, and
  - variation among sampling units within experimental units.
- Thus, if we wish to assess how much of the overall variation in the data is due to systematic effects of the treatments, we must weigh this against the variation in the data due to inherent, unexplained sources. Both variation among experimental units (e.g., mother rats) and among sampling units (e.g., rat pups within mother rats) contribute to this latter source of variation.
- The result is that our assessment of error must measure the variation both **among** and **within** experimental units. In our linear model, then, it must measure the total variability associated with the error terms ϵ_{ij} and δ_{ijk}.

TERMINOLOGY: A model such as $Y_{ijk} = \mu+\tau_i+\epsilon_{ij}+\delta_{ijk}$, may be referred to as a **nested model**. The data may be classified according to a **hierarchical structure**; experimental units within treatment groups, and then sampling units within experimental units. The units at the "inner" level are entirely contained within a unit at the "outer" level of

Table 12.43: *ANOVA table for the balanced one-way classification with subsampling.*

Source of variation	df	SS	MS	F
Treatments	$t-1$	SS[T]	MS[T]	$F_T = \frac{MST}{MSE}$
Experimental Error	$t(n-1)$	SS[E]	MS[E]	$F_S = \frac{MSE}{MSS}$
Sampling Error	$tn(s-1)$	SS[S]	MS[S]	
Total	$N-1$	SS[TOT]		

this hierarchy; hence, the term "nested." In the rat example, rat pups are nested within mother rats (which, in turn, are nested within treatment dose groups); in Example 12.5, the pigs are nested within pens (which, in turn, are nested within rations). To emphasise the nested structure in the subsampling model notation, it would not be inappropriate to write $Y_{ijk} = \mu + \tau_i + \epsilon_{j(i)} + \delta_{k(ij)}$.

ANOVA FOR THE ONE-WAY LAYOUT WITH SUBSAMPLING: Our goal is to construct the ANOVA table for the balanced one-way model with subsampling; i.e.,

$$Y_{ijk} = \mu + \tau_i + \epsilon_{ij} + \delta_{ijk},$$

for $i = 1, 2, ..., t$, $j = 1, 2, ..., n$, and $k = 1, 2, ..., s$. The ANOVA table will be the same regardless of whether the τ_i are treated as fixed or random. The breakdown of SS[TOT] into its different sources of variation is based on the identity

$$\begin{aligned}\underbrace{\sum_{i=1}^{t}\sum_{j=1}^{n}\sum_{k=1}^{s}(Y_{ijk} - \overline{Y}_{+++})^2}_{\text{SS[TOT]}} &= \underbrace{ns\sum_{i=1}^{t}(\overline{Y}_{i++} - \overline{Y}_{+++})^2}_{\text{SS[T]}} + \underbrace{s\sum_{i=1}^{t}\sum_{j=1}^{n}(\overline{Y}_{ij+} - \overline{Y}_{i++})^2}_{\text{SS[E]}} \\ &\quad + \underbrace{\sum_{i=1}^{t}\sum_{j=1}^{n}\sum_{k=1}^{s}(Y_{ijk} - \overline{Y}_{ij+})^2}_{\text{SS[S]}};\end{aligned}$$

The ANOVA table appears in Table 12.43. As usual, $N = tns$ denotes the total number of observations. The degree of freedom values are analogous to those obtained from the nested factorial ANOVA table from Chapter 10 (this should not be surprising, given the hierarchical structure that we just alluded to).

HAND COMPUTATION FOR THE ONE-WAY MODEL WITH SUBSAMPLING: If you want to compute the ANOVA table by hand, it is easiest to use the following steps.

1. Calculate the **correction term** for the mean.
$$\text{CM} = \frac{1}{tns}\left(\sum_{i=1}^{t}\sum_{j=1}^{n}\sum_{k=1}^{s} Y_{ijk}\right)^2 = Y_{+++}^2/N.$$
2. Calculate SS[TOT], the total sum of squares
$$\text{SS[TOT]} = \sum_{i=1}^{t}\sum_{j=1}^{n}\sum_{k=1}^{s} Y_{ijk}^2 - \text{CM}.$$
3. Calculate SS[T], the treatment sum of squares
$$\text{SS[T]} = \frac{1}{ns}\sum_{i=1}^{t}\left(\sum_{j=1}^{n}\sum_{k=1}^{s} Y_{ijk}\right)^2 - \text{CM}.$$
4. Calculate the **Among Experimental Units SS** (an intermediate calculation−this does not appear in the ANOVA table; loosely speaking, it measures the deviation of individual experimental units about the overall mean; thus, it does not take into account the differences among treatment means).
$$\text{Among Experimental Units SS} = \frac{1}{s}\sum_{i=1}^{t}\sum_{j=1}^{n}\left(\sum_{k=1}^{s} Y_{ijk}\right)^2 - \text{CM}.$$
5. Find SS[E], the experimental error sum of squares, by subtraction.
$$\text{SS[E]} = \text{Among Experimental Units SS} - \text{SS[T]}.$$
6. Find SS[S], the sampling error sum of squares, by subtraction.
$$\text{SS[S]} = \text{SS[TOT]} - \text{Among Experimental Units SS}.$$

Although **Among Experimental Units SS** is not an interesting quantity for our tests, and does not explicitly appear in a row of the ANOVA table, it is useful for computation. As you might suspect, SAS has little difficulty with the computations.

EXPECTED MEAN SQUARES FOR THE ONE-WAY LAYOUT WITH SUBSAMPLING: To gain insight into the nature of the different hypothesis tests and the suitability of the F ratios for testing them, we construct a table of expected mean squares, just as we did when there was no subsampling. Consider the balanced one-way classification model with subsampling; i.e.,

$$Y_{ijk} = \mu + \tau_i + \epsilon_{ij} + \delta_{ijk},$$

for $i = 1, 2, ..., t$, $j = 1, 2, ..., n$, and $k = 1, 2, ..., s$, where τ_i is the effect associated with treatment i (τ_i could be fixed or random), $\epsilon_{ij} \sim$ iid $\mathcal{N}(0, \sigma_\epsilon^2)$, $\delta_{ijk} \sim$ iid $\mathcal{N}(0, \sigma^2)$, and the ϵ_{ij} and δ_{ijk} are independent. In our model, we have the following random quantities:

- ϵ_{ij}, representing errors due to experimental units. We let σ_ϵ^2 denote the variance in the population of experimental units.
- δ_{ijk}, representing errors due to sampling units. We let σ^2 denote the variance in the population of sampling units.
- τ_i, representing the effects of the treatments, *only if the treatment levels are considered random.* In this situation, we assume that $\tau_i \sim$ iid $\mathcal{N}(0, \sigma_\tau^2)$ and that τ_i, ϵ_{ij}, and δ_{ijk} are mutually independent random variables.

CONVENTION: It is convention in most texts on analysis of variance to use the symbol σ^2 to denote the variance associated with the "smallest" unit of measurement; that is, the unit on which individual observations arise. Here, this is the sampling unit (Rao violates this convention and writes σ_S^2 for the sampling unit). In the case of one sampling unit per experimental unit, we used σ^2 to denote the variance in the population of experimental units (the "smallest" unit of measurement in that setting); here, however, the usage is different. So, keep in mind that here σ_ϵ^2 is the variance of interest with regard to **experimental units**, and σ^2 denotes the variance of the **sampling units**.

HYPOTHESIS TESTS: In our one-way model with subsampling, there are two tests that are of interest. Theoretical justification for the appropriateness of each test can be gleaned from the expected mean squares in Table 12.44.

Table 12.44: *EMS for the balanced one-way layout with subsampling.*

		Expected mean square (EMS)	
Source of variation	df	Fixed effects	Random effects
Treatments	$t-1$	$\sigma^2 + s\sigma_\epsilon^2 + \frac{ns\sum_{i=1}^t \tau_i^2}{t-1}$	$\sigma^2 + s\sigma_\epsilon^2 + ns\sigma_\tau^2$
Experimental Error	$t(n-1)$	$\sigma^2 + s\sigma_\epsilon^2$	$\sigma^2 + s\sigma_\epsilon^2$
Sampling Error	$tn(s-1)$	σ^2	σ^2

- First, we focus on testing for differences among treatments. Regardless of whether the treatment levels are best regarded as fixed or random, the appropriate statistic to use is
$$F_T = \frac{\text{MS[T]}}{\text{MS[E]}}.$$
From Table 12.44, we see that when $H_0 : \tau_1 = \tau_2 = \cdots = \tau_t = 0$ (fixed effects hypothesis) or $H_0 : \sigma_\tau^2 = 0$ (random effects hypothesis) is true, both MS[T] and MS[E] estimate the same quantity; namely $\sigma^2 + s\sigma_\epsilon^2$; thus, their ratio should be close to one under either hypothesis. When either is not true, one would expect F_T to be large. Values of F_T are compared to a $F_{t-1,t(n-1)}$ probability distribution.

- Also available is a test to diagnose variation among experimental units. Specifically, we can test $H_0 : \sigma_\epsilon^2 = 0$ versus $H_1 : \sigma_\epsilon^2 > 0$ using the statistic
$$F_S = \frac{\text{MS[E]}}{\text{MS[S]}}.$$
This is an appropriate test regardless of whether the treatments are fixed or random, since when $H_0 : \sigma_\epsilon^2 = 0$ is true, both MS[E] and MS[S] estimate the same quantity. Values of F_S larger than $F_{t(n-1),tn(s-1),\alpha}$ are deemed significant at the α level.

VARIANCE COMPONENTS: As in the random-effects model without subsampling, we have different variance components corresponding to the subsampling model. Under our assumptions associated with the model $Y_{ijk} = \mu + \tau_i + \epsilon_{ij} + \delta_{ijk}$, it is easy to show that

$$V(Y_{ijk}) = \begin{cases} \sigma_\epsilon^2 + \sigma^2, & \text{if } \tau_i \text{ are } \textbf{fixed} \\ \sigma_\tau^2 + \sigma_\epsilon^2 + \sigma^2, & \text{if } \tau_i \text{ are } \textbf{random}. \end{cases}$$

The variances σ_τ^2, σ_ϵ^2, and σ^2 are called **variance components**. There are two or three components of variability, depending on whether the treatment levels are regarded as **fixed** or **random**, respectively. Like before, it is of interest to estimate these components. Doing so helps us quantify the amount of variability due to the different sources.

FIXED EFFECTS: When the τ_i are best regarded as fixed, we see from Table 12.44,

$$\begin{aligned} E(\text{MS[E]}) &= \sigma^2 + s\sigma_\epsilon^2 \\ E(\text{MS[S]}) &= \sigma^2. \end{aligned}$$

Using MS[E] as an estimate of E(MS[E]) and MS[S] as an estimate of E(MS[S]), we have

$$\begin{aligned} \widehat{\sigma}_\epsilon^2 &= \frac{\text{MS[E]} - \widehat{\sigma}^2}{s} \\ \widehat{\sigma}^2 &= \text{MS[S]}. \end{aligned}$$

RANDOM EFFECTS: When the τ_i are regarded as random, we see from Table 12.44,

$$\begin{aligned} E(\text{MS[T]}) &= \sigma^2 + s\sigma_\epsilon^2 + ns\sigma_\tau^2 \\ E(\text{MS[E]}) &= \sigma^2 + s\sigma_\epsilon^2 \\ E(\text{MS[S]}) &= \sigma^2. \end{aligned}$$

Using MS[T] as an estimate of E(MS[T]), MS[E] as an estimate of E(MS[E]), and MS[S] as an estimate of E(MS[S]), we see that

$$\begin{aligned} \widehat{\sigma}_\tau^2 &= \frac{\text{MS[T]} - \widehat{\sigma}^2 - s\widehat{\sigma}_\epsilon^2}{ns} \\ \widehat{\sigma}_\epsilon^2 &= \frac{\text{MS[E]} - \widehat{\sigma}^2}{s} \\ \widehat{\sigma}^2 &= \text{MS[S]}. \end{aligned}$$

Thus, we can estimate the variance components in either the fixed effects or random effects models that incorporate subsampling.

Example 12.7 (`cholesterol.sas`). Three different drugs for the treatment of high cholesterol are produced by three different manufacturers, each of which produces its drug at one of 2 different plants, as shown in Table 12.45. The measurements given are

Table 12.45: *Cholesterol concentration data.*

	Drug 1		Drug 2		Drug 3	
	Plant 1	Plant 2	Plant 1	Plant 2	Plant 1	Plant 2
Female 1	102	103	108	109	104	105
Female 2	104	104	110	108	106	107

cholesterol concentrations (mg/100 ml of plasma) for human females treated with the drugs. A question of interest is whether cholesterol levels for female subjects differ among the three treatments (drugs). These are the only three drugs of interest in the study; hence, the levels of drug are best regarded as fixed effects. The two plants are different for each of the different drugs (manufacturers). Furthermore, for each plant, 2 different females are used, so there are $N = 12$ total measurements. The only classification scheme for these data that makes sense is $Y_{ijk} = \mu + \tau_i + \epsilon_{ij} + \delta_{ijk}$, with

Treatments	Drugs	$t = 3$
Experimental Units	Plants	$n = 2$
Sampling Units	Female subjects	$s = 2$

HAND COMPUTATIONS: The **correction term**, for fitting the overall mean, is

$$\text{CM} = \frac{1}{N}\left(\sum_{i=1}^{3}\sum_{j=1}^{2}\sum_{k=1}^{2} Y_{ijk}\right)^2 = \frac{(1270)^2}{12} = 134408.33.$$

The total sum of squares is given by

$$\text{SS[TOT]} = \sum_{i=1}^{3}\sum_{j=1}^{2}\sum_{k=1}^{2} Y_{ijk}^2 - \text{CM} = 134480.00 - 134408.33 = 71.67,$$

and the treatment sum of squares is given by

$$\text{SS[T]} = \frac{1}{4}\sum_{i=1}^{3}\left(\sum_{j=1}^{2}\sum_{k=1}^{2} Y_{ijk}\right)^2 - \text{CM} = 134469.50 - 134408.33 = 61.17.$$

The intermediate measure of variability calculation

$$\text{Among EUs SS} = \frac{1}{2}\sum_{i=1}^{3}\sum_{j=1}^{2}\left(\sum_{k=1}^{2} Y_{ijk}\right)^2 - \text{CM} = 134471.00 - 134408.33 = 62.67$$

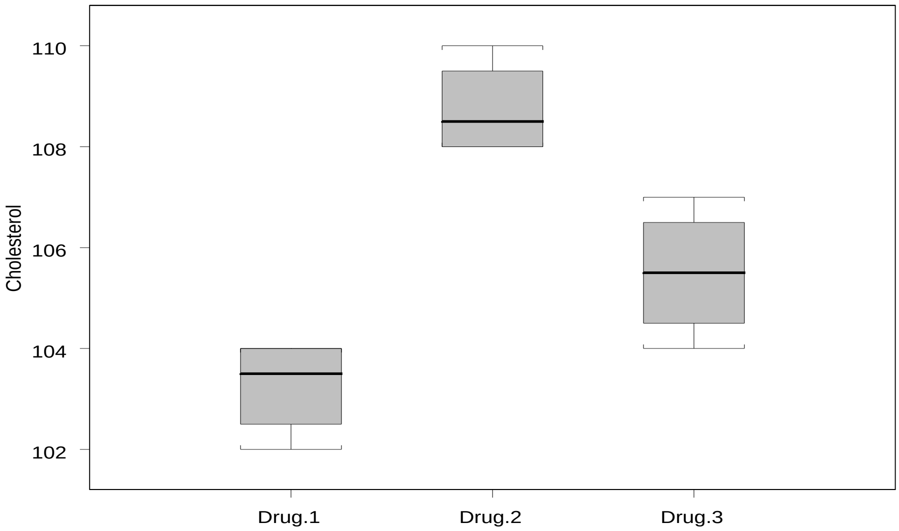

Figure 12.34: *Cholesterol levels for three different drugs.*

helps us get the remaining sums of squares by subtraction; i.e.,

$$\begin{aligned}\text{SS[E]} = \text{Among Experimental Units SS} - \text{SS[T]} &= 62.67 - 61.17 = 1.50 \\ \text{SS[S]} = \text{SS[TOT]} - \text{Among Experimental Units SS} &= 71.67 - 62.67 = 9.00.\end{aligned}$$

Putting all the above calculations together, we can construct the ANOVA table:

Table 12.46: *ANOVA table for cholesterol data in Example* 12.7.

Source of variation	df	SS	MS	F
Treatments	2	61.17	30.58	61.16
Experimental Error	3	1.50	0.50	0.33
Sampling Error	6	9.00	1.50	
Total	11	71.67		

ANALYSIS: The statistic $F_T = 61.16$ says that there is a clear difference in mean cholesterol levels among the three drugs (e.g., $F_{2,3,0.05} = 9.55$); that is, $H_0 : \tau_1 = \tau_2 = \tau_3 = 0$

is rejected. To test for variation among plants; i.e., $H_0 : \sigma_\epsilon^2 = 0$ versus $H_1 : \sigma_\epsilon^2 > 0$, we see that $F_S = 0.33$ is not large (e.g., $F_{3,6,0.05} = 4.76$). There is not enough evidence to suggest that there is significant variation among plants.

PAIRWISE INTERVALS: Since there are significant differences among drugs, let's see where they are; we'll do this by writing simultaneous pairwise intervals for $\tau_2 - \tau_1$, $\tau_3 - \tau_1$, and $\tau_3 - \tau_2$. To construct the confidence interval for $\tau_2 - \tau_1$, we can use the point estimator $\overline{Y}_{2++} - \overline{Y}_{1++}$, the difference in the sample means. Under our model assumptions, straightforward calculations show that (verify!)

$$E(\overline{Y}_{2++} - \overline{Y}_{1++}) = \tau_2 - \tau_1$$

and

$$V(\overline{Y}_{2++} - \overline{Y}_{1++}) = \frac{2(\sigma^2 + 2\sigma_\epsilon^2)}{4}.$$

However, note that MS[E] estimates $\sigma^2 + 2\sigma_\epsilon^2$ (see Table 12.44). Since $\overline{Y}_{2++} - \overline{Y}_{1++}$ is a linear combination of the normally distributed Y_{ijk}'s, then $\overline{Y}_{2++} - \overline{Y}_{1++}$, too, is normally distributed. A 95 percent confidence interval for $\tau_2 - \tau_1$ is, thus, given by

$$(\overline{Y}_{2++} - \overline{Y}_{1++}) \pm t_{3,0.025} \times \sqrt{\frac{2\text{MS[E]}}{4}}.$$

The other confidence intervals are formed similarly. To adjust for multiplicity, we can use a Tukey correction. The simultaneous intervals then take the form

$$(\overline{Y}_{2++} - \overline{Y}_{1++}) \pm q_{3,3,0.05} \times \sqrt{\frac{\text{MS[E]}}{4}}$$

$$(\overline{Y}_{3++} - \overline{Y}_{1++}) \pm q_{3,3,0.05} \times \sqrt{\frac{\text{MS[E]}}{4}}$$

$$(\overline{Y}_{3++} - \overline{Y}_{2++}) \pm q_{3,3,0.05} \times \sqrt{\frac{\text{MS[E]}}{4}}$$

(the $\sqrt{2}$ term is absorbed in $q_{3,3,0.05}$). These are **simultaneous** 95 percent confidence intervals for the differences. For these data, $\overline{y}_{1++} = 103.25$, $\overline{y}_{2++} = 108.75$, $\overline{y}_{3++} = 105.50$, $q_{3,3,0.05} = 5.91$ (not in Rao), and MS[E] $= 0.50$. The simultaneous intervals are given by (3.41,7.59), (0.16, 4.34), and $(-1.16, -5.34)$, respectively. Based on these intervals, it looks as though the Drug 1 is associated with the smallest mean cholesterol level, followed by Drug 3 and Drug 2, in that order.

UNEQUAL REPLICATION AND NUMBERS OF SUBSAMPLES: When the numbers of replicates and/or subsamples are not the same, the same general procedure we used may be used to construct an analysis of variance table; i.e., partitioning SS[TOT] into its components. However, the imbalance leads to some difficulties.

- For example, degrees of freedom become more difficult to calculate. Recall that in Table 12.44, each expected mean square was equal to the one below it plus a term representing extra variation. This nice property no longer holds exactly when the numbers of subsamples are not the same.
- Intuitively, if we have different numbers of subsamples on each experimental unit, the quality of information on each experimental unit is different. We would, thus, expect trying to sort out the different sources of variation to be much harder.
- The result is that exact tests of hypotheses may no longer be carried out. Rather, approximate tests must be conducted. This is analogous to the situation where one is testing the difference between two normal means with unequal variances. The basic problem here is the same; that is, we no longer have the same quality of information (as measured by variance) on all experimental units under study.
- Just keep in mind that care must be taken under these circumstances.

12.5 Using variance components in the one-way layout with subsampling

We have seen in our linear models for the one-way classification, both with and without subsampling, that we construct mean squares to estimate the variability due to different sources, e.g., treatments, experimental units, sampling units. When we have subsampling, recall, from Table 12.44, that this involves variances of various error terms. For example, in the case of random treatment effects and subsampling with equal replication and numbers of subsamples, MS[T] estimates

$$\sigma^2 + s\sigma_\epsilon^2 + ns\sigma_\tau^2.$$

As noted earlier, quantities such σ^2, σ_ϵ^2, and σ_τ^2 are called **variance components**. Here, the variability we might observe across treatments is associated with three sources: individual sampling units (σ^2), experimental units (σ_ϵ^2), and treatments (σ_τ^2).

DESIGN CONSIDERATIONS: In the planning stages of any experiment, the researcher often has desired statistical objectives. In an experiment involving subsamples, we may, in the planning stages, wonder how best to allocate the available resources. We might

- have many experimental units with few sampling units on each, or
- have less experimental units with more sampling units on each.

Generally, as we have discussed, we would like to have enough experimental units to get a good idea of the population; however, we may "fine tune" how we do this if we understand the relative sizes of experimental and sampling error, as well as considerations such as whether subsamples are expensive or difficult to obtain or whether it is impractical to have too many experimental units. Previous experimental data may be used to plan future experiments and give the investigator information on these questions. Estimates of the relative magnitudes of the variance components, σ^2 and σ_ϵ^2, may be used to help guide our planning. In particular, suppose we wish to conduct a future experiment with t treatments and are considering different numbers of replicates, n^*, say, and different numbers of sampling units on each, say, s^*. Given estimates of σ^2 and σ_ϵ^2 (from previous studies, perhaps), we can use this information to help us decide on sensible values of the number of replicates and the number of sampling units. For example, if our estimate of σ_ϵ^2 is large and our estimate of σ^2 is small, then, intuitively, it would be best to use more experimental units with fewer sampling units. Thus, as you can see, this information, along with associated costs of doing the experiment different ways, can be used to decide how best to use resources in the future.

13 ANOVA Models with Block Effects

Complementary reading from Rao: Chapter 15 (§ 15.1-15.2).

13.1 Introduction

As we have already discussed, a common theme in experimental design is the reduction of experimental error by the **meaningful grouping** of experimental units. If we can attribute some of the variation in experimental units to systematic sources, we can reduce the effect of the inherent variation among them; that is, we could reduce our assessment of experimental error. Proper use of meaningful grouping (i.e., **blocking**) allows us to achieve this goal.

WHEN BLOCKING MIGHT BE USED: When no sources of variation, other than treatments, are anticipated, blocking will probably not add very much precision; i.e., it will not reduce our assessment of experimental error. If the experimental units are expected to be fairly uniform, then a completely randomised design (CRD) will probably be sufficient. In many situations, however, other systematic sources of variation are anticipated.

- In a field experiment, adjacent plots will tend to be more alike than those far apart.
- Observations made with a particular measuring device, or, by a particular individual, may be more alike than those made by different devices or individuals.
- Plants kept in one greenhouse may be more alike than those from other greenhouses.
- Patients treated at the same hospital may be more alike than those treated at different hospitals.

REALISATION: In such instances, there is a potential source of **systematic variation** that we may identify in advance. This suggests that we may wish to group experimental units in a meaningful way on this basis.

BLOCKING: In any experiment, we seek to investigate differences among the treatments (whether they be fixed or random). When experimental units are considered in meaningful groups, they may be thought of as being classified not only according to treatment assignment, but also according to which group (i.e., **block**) they belong (e.g., position in field, device or observer, greenhouse, hospital, etc.). By accounting for differences among experimental units through the use of meaningful grouping, we can increase our ability to detect treatment differences, if they exist.

13.2 The randomised complete block design

When experimental units may be meaningfully grouped, a completely randomised design (CRD) is suboptimal. In this situation, an alternative strategy for assigning treatments to experimental units, which takes advantage of the grouping, should be used; namely, a **randomised complete block design** (**RBCD**). The meaningful groups of experimental units are called **blocks**.

- We will assume, for our discussion, that each treatment appears the same number of times in each block; hence, the term "complete" block design. The simplest case is that where each treatment appears exactly once in each block.
- The number of blocks used is denoted by r. The number of treatments used is denoted by t.
- To set up such a design, randomisation is carried out in the following way:
  - assign experimental units to blocks on the basis of the meaningful grouping factor (e.g., greenhouse, hospital, etc.).
  - randomly assign treatments to experimental units within each block.

  This randomisation protocol is sometimes called **restricted randomisation**, because randomisation is carried out only within each block.

RATIONALE: Experimental units within blocks are "as alike as possible," so observed differences among them should be mainly attributable to the treatments. To ensure this, all experimental units within a block should be treated as uniformly as possible; e.g.,

- in a field, all plots should be harvested at the same time of day,
- all measurements using a single device should be made by the same individual if different people use it in a different way,
- all plants in a greenhouse should be watered at the same time of day or by the same amount,
- treatments should be administered to patients following the same protocol.

POTENTIAL ADVANTAGES: If blocking is really warranted, then a RCBD offers **greater precision** than is possible with a CRD (which ignores the need for blocking). Also, we will have an **increased scope** of inference, because more experimental conditions may be included.

13.3 Incorporating block effects in the two-way layout

13.3.1 No replication, no subsampling

We assume here that one observation is taken on each experimental unit (i.e., that there is no subsampling). We will also assume that a RCBD is used with exactly one experimental unit per treatment per block (i.e., there is no replication). For this situation, we may classify an individual observation as being from the jth block on the ith treatment as

$$Y_{ij} = \mu + \tau_i + \rho_j + \epsilon_{ij},$$

for $i = 1, 2, ..., t$ and $j = 1, 2, ..., r$, where $\epsilon_{ij} \sim \text{iid } \mathcal{N}(0, \sigma^2)$. Here, t denotes the number of treatments, r denotes the number of blocks, μ denotes the overall mean, τ_i is the effect of receiving treatment i, and ρ_j is the effect of being in block j. As one can see, this is simply a two-factor ANOVA model without interaction. It is assumed that the block-treatment interaction is error so that an estimate of σ^2 is available.

FIXED VERSUS RANDOM EFFECTS: As in the one-way classification, the treatment and block effects, τ_i and ρ_j, respectively, may be best regarded as fixed or random. We have already discussed the notion of regarding treatments as having fixed or random effects. We may also apply the same reasoning to blocks. Consider the following scenarios:

1. Both τ_i and ρ_j are best regarded as having **fixed** effects. In this case, both describe a particular set of conditions that will not vary across experiments. For example, we might be comparing three drugs (treatments) for 2 breeds of cattle (blocks). If these are the only drugs and breeds that are of interest, then both factors are fixed.

2. Both τ_i and ρ_j are best regarded as having **random** effects; that is, τ_i and ρ_j are treated as random variables drawn from the populations of all possible treatments and blocks, respectively, with variances σ^2_τ and σ^2_ρ. For example, suppose that four machines (treatments) are chosen at random from all machines at a company, and three machine operators (blocks) are chosen at random from all operators employed at the company. In this case, both factors are random. The usual additional assumptions are that $\tau_i \sim$ iid $\mathcal{N}(0, \sigma^2_\tau)$, $\rho_j \sim$ iid $\mathcal{N}(0, \sigma^2_\rho)$, and that τ_i, ρ_j, and ϵ_{ij} are independent random variables.

3. We may also have the situation of a **mixed** model; i.e., one that contains both fixed and random effects. Most often, in this situation, it is the treatment effects τ_i that are best regarded as **fixed** and the block effects ρ_j are treated as **random**. For example, say that we want to compare three fertilizers (treatments) in two different greenhouses (blocks). The three fertilizers are the only ones of interest, so the treatments are fixed. However, if we hope that our inferences apply to all possible greenhouses (not just the two we used!) then we regard the greenhouses as random. The usual additional assumptions are that $\rho_j \sim$ iid $\mathcal{N}(0, \sigma^2_\rho)$, and that ρ_j and ϵ_{ij} are independent random variables.

Table 13.47: *ANOVA table for the RCBD; one observation per treatment per block and no subsampling.*

Source	df	SS	MS	F
Treatments	$t-1$	SS[T]	MS[T]	$F_T = \frac{\text{MS[T]}}{\text{MS[E]}}$
Blocks	$r-1$	SS[B]	MS[B]	$F_B = \frac{\text{MS[B]}}{\text{MS[E]}}$
Error	$(t-1)(r-1)$	SS[E]	MS[E]	
Total	$N-1$	SS[TOT]		

ANOVA TABLE: Our goal is to write the ANOVA table for the two-way classification model; i.e., $Y_{ij} = \mu + \tau_i + \rho_j + \epsilon_{ij}$, for $i = 1, 2, ..., t$ and $j = 1, 2, ..., r$, which is our linear model for incorporating treatment and block effects with no subsampling and with only one observation per treatment per block. The table is based on the identity

$$\begin{aligned}\underbrace{\sum_{i=1}^{t}\sum_{j=1}^{r}(Y_{ij}-\overline{Y}_{++})^2}_{\text{SS[TOT]}} &= \underbrace{r\sum_{i=1}^{t}(\overline{Y}_{i+}-\overline{Y}_{++})^2}_{\text{SS[T]}} + \underbrace{t\sum_{j=1}^{r}(\overline{Y}_{+j}-\overline{Y}_{++})^2}_{\text{SS[B]}} \\ &+ \underbrace{\sum_{i=1}^{t}\sum_{j=1}^{r}(Y_{ij}-\overline{Y}_{i+}-\overline{Y}_{+j}+\overline{Y}_{++})^2}_{\text{SS[E]}}.\end{aligned}$$

The total number of observations is $N = tr$. Degrees of freedom follow their usual patterns. All computations, even the F statistics, are the same regardless of whether the treatments (or blocks) are fixed or random. The interpretations, however, will be different depending on which situation we are in.

HAND COMPUTATION FOR SUMS OF SQUARES: To construct the ANOVA table for the RCBD with one observation per treatment per block and no subsampling, it is easiest to use SAS. However, the following computing formulae are also helpful if you don't like computers. First, the **correction term** for fitting the overall mean is

$$\text{CM} = \frac{1}{tr}\left(\sum_{i=1}^{t}\sum_{j=1}^{r}Y_{ij}\right)^2 = Y_{++}^2/N.$$

The total sum of squares is given by

$$\text{SS[TOT]} = \sum_{i=1}^{t}\sum_{j=1}^{r} Y_{ij}^2 - \text{CM}.$$

The treatment and block sums of squares are given by

$$\begin{aligned}
\text{SS[T]} &= r\sum_{i=1}^{t}(\overline{Y}_{i+} - \overline{Y}_{++})^2 = \frac{1}{r}\sum_{i=1}^{t} Y_{i+}^2 - \text{CM} \\
\text{SS[B]} &= t\sum_{j=1}^{r}(\overline{Y}_{+j} - \overline{Y}_{++})^2 = \frac{1}{t}\sum_{j=1}^{r} Y_{+j}^2 - \text{CM}.
\end{aligned}$$

The error sum of squares is then computed by subtraction; i.e.,

$$\text{SS[E]} = \text{SS[TOT]} - \text{SS[T]} - \text{SS[B]}.$$

TESTING FOR TREATMENT DIFFERENCES: With **fixed** treatments effects, we are, as before, interested in comparing the treatment means; that is, we would like to test $H_0 : \tau_1 = \tau_2 = \cdots = \tau_t = 0$ versus H_1 : not H_0. With **random** effects, we are interested in the entire population of treatments. If there are no differences among treatments, then $\sigma_\tau^2 = 0$. Thus, with random effects, we write the hypotheses as $H_0 : \sigma_\tau^2 = 0$ versus $H_1 : \sigma_\tau^2 > 0$. In either case, we judge the amount of evidence against H_0 by comparing F_T to a $F_{t-1,(t-1)(r-1)}$ distribution. Large values of F_T are not consistent with H_0.

TESTING FOR BLOCK EFFECTS: In most experiments, whether or not there are block differences is not really a main concern, because, by considering blocks up front, we have acknowledged them as a possible nontrivial source of variation. This notwithstanding, there is nothing in the two-way ANOVA model to keep one from testing block effects (see note below). With **fixed** block effects, we would like to test $H_0 : \rho_1 = \rho_2 = \cdots = \rho_r = 0$ versus H_1 : not H_0. With **random** block effects, we are interested in the entire population of blocks, so we test $H_0 : \sigma_\rho^2 = 0$ versus $H_1 : \sigma_\rho^2 > 0$. In either case, we judge the amount of evidence against H_0 by comparing F_B to a $F_{r-1,(t-1)(r-1)}$ distribution. Large values of F_B are not consistent with H_0.

NOTE: Many statisticians have argued that tests for block effects, in fact, should not be conducted. The reason stems from the restricted randomisation protocol used to assign

Table 13.48: *EMS for the RCBD with one observation per treatment per block and no subsampling.*

	Expected mean squares (EMS)		
Source of variation	Both fixed	Both random	Mixed
Treatments	$\sigma^2 + \frac{r\sum_{i=1}^t \tau_i^2}{t-1}$	$\sigma^2 + r\sigma_\tau^2$	$\sigma^2 + \frac{r\sum_{i=1}^t \tau_i^2}{t-1}$
Blocks	$\sigma^2 + \frac{t\sum_{j=1}^r \rho_j^2}{r-1}$	$\sigma^2 + t\sigma_\rho^2$	$\sigma^2 + t\sigma_\rho^2$
Error	σ^2	σ^2	σ^2

treatments to the experimental units. What affect does this have on F_B? Some authors have argued that the key justification for the usual ANOVA F tests is the use of complete randomisation (instead of the normality assumption). Of course, complete randomisation is not used in RCBDs. Others have argued that the two-way additive ANOVA model $Y_{ij} = \mu + \tau_i + \rho_j + \epsilon_{ij}$, being a very simplistic model, is not that appropriate for the RCBD, citing that it isn't clear how the existence of additive block effects reduces variability in a two-way ANOVA. However, as an **approximate** procedure to investigate the effect of the blocking factor, examining F_B is certainly not unreasonable. If it is large, it implies that the blocking factor has a large effect and is probably helpful in improving the precision of the comparison of treatment means. Of course, this is the exact reason why one might use blocking in the first place, so one might expect F_B to be large, if, in fact, there was solid evidence that blocking should have been used to begin with. On the other hand, small values of F_B might suggest that blocking doesn't really add too much precision.

EXPECTED MEAN SQUARES: As we have seen before, it is instructive to examine the expected mean squares under our assumptions for the two-factor ANOVA model without interaction; i.e., $Y_{ij} = \mu + \tau_i + \rho_j + \epsilon_{ij}$. Formulae for expected mean squares depend on whether the treatments (and blocks) are fixed or random. See Table 13.48. Careful examination of the expected mean squares allows insight into the suitability of the tests just described. For the fixed block and/or treatment effects, the usual "sum-to-zero" side conditions are assumed; i.e., $\tau_+ = 0$ and $\rho_+ = 0$.

Table 13.49: *Wheat yield data.*

	Variety							
Block	A	B	C	D	E	F	G	Y_{+j}
I	10	9	11	15	10	12	11	78
II	11	10	12	12	10	11	12	78
III	12	13	10	14	15	13	13	90
IV	14	15	13	17	14	16	15	104
V	13	14	16	19	17	15	18	112
Y_{i+}	60	61	62	77	66	67	69	$Y_{++} = 462$

Example 13.1 (`wheat.sas`). The data in Table 13.49 are yields (Y, measured in bushels/acre) from an agricultural experiment set up in a RCBD. The experiment was designed to investigate the differences in yield for $t = 7$ varieties of wheat, labelled A-G (these are the only varieties of interest). A field was divided into $r = 5$ blocks, each containing seven plots. In each block, the seven plots were assigned at random to be planted with the seven varieties, one plot for each variety.

CALCULATIONS: The correction term for fitting the overall mean is

$$\text{CM} = Y_{++}^2/N = (462)^2/35 = 6098.4$$

The total sum of squares is given by

$$\text{SS[TOT]} = \sum_{i=1}^{t}\sum_{j=1}^{r} Y_{ij}^2 - \text{CM} = (10^2 + 9^2 + \cdots + 18^2) - 6098.4 = 215.6.$$

The treatment and block sums of squares are given by

$$\begin{aligned}\text{SS[T]} &= \frac{1}{r}\sum_{i=1}^{t} Y_{i+}^2 - \text{CM} = \frac{1}{5}(60^2 + 61^2 + \cdots + 69^2) - 6098.4 = 41.6.\\ \text{SS[B]} &= \frac{1}{t}\sum_{j=1}^{r} Y_{+j}^2 - \text{CM} = \frac{1}{7}(78^2 + 78^2 + \cdots + 112^2) - 6098.4 = 134.2.\end{aligned}$$

The error sum of squares is then computed by subtraction; i.e.,

$$\text{SS[E]} = 215.6 - 41.6 - 134.2 = 39.8.$$

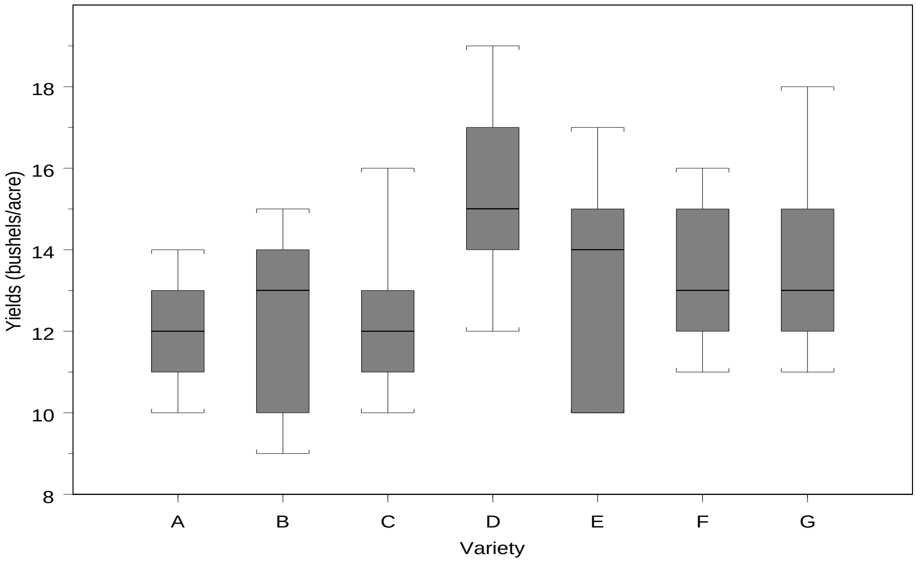

Figure 13.35: *Wheat yield data for different varieties.*

Table 13.50: *RCBD ANOVA for the wheat yield data in Example* 13.1.

Source	df	SS	MS	F
Treatments	6	41.6	6.93	4.18
Blocks	4	134.2	33.54	20.21
Error	24	39.8	1.66	
Total	34	215.6		

ANALYSIS: The ANOVA table for the wheat data appears in Table 13.50. There looks to be significant differences in mean yields among the seven varieties since the statistic $F_T = 4.18$ is large (e.g., $F_{6,24,0.05} = 2.508$). Also, the statistic $F_B = 20.21$ provides strong evidence that blocking, was, in fact, useful. Of course, as usual F_T doesn't provide insight as to where the differences among mean yields actually are located, so let's construct pairwise Tukey intervals to find them. For example, to construct a confidence interval for $\tau_2 - \tau_1$, we would use the point estimator $\overline{Y}_{2+} - \overline{Y}_{1+}$. It is not difficult to show that

$E(\overline{Y}_{2+} - \overline{Y}_{1+}) = \tau_2 - \tau_1$ and that

$$V(\overline{Y}_{2+} - \overline{Y}_{1+}) = \frac{2\sigma^2}{5}.$$

Replacing σ^2 with its unbiased estimate MS[E], and noting that $\overline{Y}_{2+} - \overline{Y}_{1+}$ is normally distributed (under our model assumptions), it follows that

$$(\overline{Y}_{2+} - \overline{Y}_{1+}) \pm t_{24,0.025} \times \sqrt{\frac{2\text{MS[E]}}{5}}.$$

is a 95 percent confidence interval for $\tau_2 - \tau_1$. To adjust for multiplicity, our $\binom{7}{2}$ **simultaneous** 95 percent confidence intervals for $\tau_i - \tau_{i'}$, for $i < i'$, take the form

$$(\overline{Y}_{i'+} - \overline{Y}_{i+}) \pm q_{7,24,0.05} \times \sqrt{\frac{\text{MS[E]}}{5}}$$

(the $\sqrt{2}$ term is absorbed in $q_{7,24,0.05}$). These intervals are provided in the accompanying SAS output; it looks as though the yields for varieties 1, 2, and 3 are significantly lower than the yield for variety 4, but that no other significant differences exist.

USEFULNESS OF BLOCKING: From these results, note that blocking served to explain much of the overall variation. To appreciate this further, suppose that we had not analysed the experiment in Example 13.1 as a RCBD, but, instead, had just analysed the experiment according to a CRD (and that we ended up with the same data). In this case, the ANOVA table would be

Source	df	SS	MS	F
Treatments	6	41.6	6.93	1.12
Error	28	174.0	6.21	
Total	34	215.6		

As you can see, the hypothesis for no treatment difference $H_0 : \tau_1 = \tau_2 = \cdots = \tau_7 = 0$ would not be rejected at any reasonable α level!!

MORAL: In the one-way classification experiment and analysis, there is no accounting for the variation in the data that is attributable to systematic sources; e.g., position in field.

The one-way analysis has no choice but to attribute this variation to experimental error; that is, it regards this variation as just part of the inherent variation among experimental units that we can not explain. The result is that SS[E] in the one-way analysis contains both variation due to the position in field (a systematic source) and inherent variation.

NOTE: In the CRD analysis, note that SS[E] = 174.0, and that in the RBCD analysis, SS[B]+SS[E] = 134.2+39.8 = 174.0; that is, SS[E] in the one-way analysis, which may be regarded as ignoring the blocks, is equal to SS[B]+SS[E] in the RCBD analysis. Thus, the use of blocking allows us to partition SS[E] in the one-way analysis into two orthogonal components−one that is due to the systematic source of variation in the blocks, SS[B], and one that is attributable to the unexplainable variation, SS[E] in the RBCD analysis. In the CRD analysis, MS[E] is too large (so we couldn't detect any differences in the treatments). By blocking, and explicitly acknowledging that field position was a potential source of variation, MS[E] was sufficiently reduced so that we could identify variety differences.

MAIN POINTS: First, blocking can be an effective means of explaining variation (increasing precision) so that differences among treatments (that truly exist) are more likely to be detected. Second, the data from an experiment should be analysed in a way that reflects how the randomisation (design) was carried out. Example 13.1 shows that if we analyse the data according to an incorrect randomisation protocol, then erroneous inferences may likely result. *The design of the experiment always dictates the analysis!*

13.3.2 No replication, subsampling

In the last section, we assumed that exactly one observation was taken on each of the $t \times r$ experimental units in a RCBD. We now consider the extension of this to the case wherein each of the $t \times r$ experimental units has more than one sampling unit (i.e., subsampling). For simplicity, we will consider only the case of an equal number s of sampling units per experimental unit. In this situation, we may identify an observation Y_{ijk} as being from the kth subsampling unit from the jth block receiving the ith treatment. The linear

Table 13.51: *ANOVA table for the RCBD; one observation per treatment per block with subsampling.*

Source	df	SS	MS	F
Treatments	$t-1$	SS[T]	MS[T]	$F_T = \frac{\text{MS[T]}}{\text{MS[E]}}$
Blocks	$r-1$	SS[B]	MS[B]	$F_B = \frac{\text{MS[B]}}{\text{MS[E]}}$
Experimental Error	$(t-1)(r-1)$	SS[E]	MS[E]	$F_S = \frac{\text{MS[E]}}{\text{MS[S]}}$
Sampling Error	$tr(s-1)$	SS[S]	MS[S]	
Total	$N-1$	SS[TOT]		

model for this situation is given by

$$Y_{ijk} = \mu + \tau_i + \rho_j + \epsilon_{ij} + \delta_{ijk},$$

for $i = 1, 2, ..., t$, $j = 1, 2, ..., r$, and $k = 1, 2, ..., s$, where $\epsilon_{ij} \sim$ iid $\mathcal{N}(0, \sigma_\epsilon^2)$, $\delta_{ijk} \sim$ iid $\mathcal{N}(0, \sigma^2)$, and the ϵ_{ij} and δ_{ijk} are independent. As in the last section, either τ_i or ρ_j (or both) may be best regarded as having fixed or random effects.

RECALL: In the last chapter, when we were considering the use of subsampling in the one-way layout (without blocking), our breakdown of the total sum of squares was

$$\text{SS[TOT]} = \text{SS[T]} + \text{SS[E]} + \text{SS[S]}.$$

In our situation now, with the use of blocking, we will take SS[E] from the one-way analysis (with subsampling) and break it into two orthogonal components; namely, SS[B] and a new SS[E] piece which is our experimental error after accounting for the blocking. This orthogonalisation of SS[E] in the one-way analysis is precisely what we did in the last section in the absence of subsampling.

ANOVA TABLE: Our goal is to write the ANOVA table for the two-way classification model; i.e., $Y_{ijk} = \mu + \tau_i + \rho_j + \epsilon_{ij} + \delta_{ijk}$, for $i = 1, 2, ..., t$, $j = 1, 2, ..., r$, and $k = 1, 2, ..., s$, which is our linear model for incorporating treatment and block effects with subsampling (but with only one experimental unit per treatment per block). The ANOVA table will be the same regardless of whether the τ_i and/or ρ_j are treated as fixed or random. The breakdown of SS[TOT] into its different sources of variation is based on the identity

$$\begin{aligned}\underbrace{\sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{k=1}^{s}(Y_{ijk}-\overline{Y}_{+++})^2}_{\text{SS[TOT]}} &= \underbrace{rs\sum_{i=1}^{t}(\overline{Y}_{i++}-\overline{Y}_{+++})^2}_{\text{SS[T]}}+\underbrace{ts\sum_{j=1}^{r}(\overline{Y}_{+j+}-\overline{Y}_{+++})^2}_{\text{SS[B]}}\\ &\quad+\underbrace{s\sum_{i=1}^{t}\sum_{j=1}^{r}(\overline{Y}_{ij+}-\overline{Y}_{+++})^2}_{\text{SS[E]}}+\underbrace{\sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{k=1}^{s}(Y_{ijk}-\overline{Y}_{ij+})^2}_{\text{SS[S]}}.\end{aligned}$$

HAND COMPUTATION: For this situation, performing hand computation of the sums of squares is easiest using the following formulae. First, compute the **correction term** for fitting the overall mean

$$\text{CM}=\frac{1}{trs}\left(\sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{k=1}^{s}Y_{ijk}\right)^2=Y_{+++}^2/N,$$

where $N=trs$. The total sum of squares is given by

$$\text{SS[TOT]}=\sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{k=1}^{s}Y_{ijk}^2-\text{CM}.$$

The treatment and block sums of squares are given by

$$\begin{aligned}\text{SS[T]} &= rs\sum_{i=1}^{t}(\overline{Y}_{i++}-\overline{Y}_{+++})^2=\frac{1}{rs}\sum_{i=1}^{t}Y_{i++}^2-\text{CM}\\ \text{SS[B]} &= ts\sum_{j=1}^{r}(\overline{Y}_{+j+}-\overline{Y}_{+++})^2=\frac{1}{ts}\sum_{j=1}^{r}Y_{+j+}^2-\text{CM}.\end{aligned}$$

The experimental error sum of squares is computed as

$$\text{SS[E]}=\sum_{i=1}^{t}\sum_{j=1}^{r}Y_{ij+}^2-\text{CM}-\text{SS[T]}-\text{SS[B]}.$$

The sampling error sum of squares is then computed by subtraction; i.e.,

$$\text{SS[S]}=\text{SS[TOT]}-\text{SS[T]}-\text{SS[B]}-\text{SS[E]}.$$

TESTING FOR TREATMENT EFFECTS: The major test of interest is the one that concerns the treatments. If the τ_i are best regarded as **fixed**, then we are interested in

Table 13.52: *EMS for the RCBD with one observation per treatment per block and subsampling.*

	Expected mean squares (EMS)		
Source of variation	Both fixed	Both random	Mixed
Treatments	$\sigma^2 + s\sigma^2_\epsilon + \frac{rs\sum_{i=1}^t \tau_i^2}{t-1}$	$\sigma^2 + s\sigma^2_\epsilon + rs\sigma^2_\tau$	$\sigma^2 + s\sigma^2_\epsilon + \frac{rs\sum_{i=1}^t \tau_i^2}{t-1}$
Blocks	$\sigma^2 + s\sigma^2_\epsilon + \frac{ts\sum_{j=1}^r \rho_j^2}{r-1}$	$\sigma^2 + s\sigma^2_\epsilon + ts\sigma^2_\rho$	$\sigma^2 + s\sigma^2_\epsilon + ts\sigma^2_\rho$
Experimental Error	$\sigma^2 + s\sigma^2_\epsilon$	$\sigma^2 + s\sigma^2_\epsilon$	$\sigma^2 + s\sigma^2_\epsilon$
Sampling Error	σ^2	σ^2	σ^2

testing $H_0 : \tau_1 = \tau_2 = \cdots = \tau_t = 0$ versus H_1 : not H_0. With **random** effects, we are interested in testing $H_0 : \sigma^2_\tau = 0$ versus $H_1 : \sigma^2_\tau > 0$. In either case, we judge the amount of evidence against H_0 by comparing F_T to a $F_{t-1,(t-1)(r-1)}$ distribution. Large values of F_T are not consistent with H_0.

TESTING FOR BLOCK EFFECTS: With **fixed** block effects, we would like to test $H_0 : \rho_1 = \rho_2 = \cdots = \rho_r = 0$ versus H_1 : not H_0. With **random** block effects, we are interested in the entire population of blocks, so we test $H_0 : \sigma^2_\rho = 0$ versus $H_1 : \sigma^2_\rho > 0$. In either case, we judge the amount of evidence against H_0 by comparing F_B to a $F_{r-1,(t-1)(r-1)}$ distribution. Large values of F_B are not consistent with H_0. Recall that tests for block effects are sometimes viewed as suspect (see last section).

TEST FOR VARIATION AMONG EXPERIMENTAL UNITS: Also available, as in the one-way layout setting, is a test to diagnose variation among experimental units. Specifically, we can test $H_0 : \sigma^2_\epsilon = 0$ versus $H_1 : \sigma^2_\epsilon > 0$ using the statistic F_S. This is an appropriate test regardless of whether the treatments (or blocks) are fixed or random, since when $H_0 : \sigma^2_\epsilon = 0$ is true, both MS[E] and MS[S] estimate the same quantity. Values of F_S larger than $F_{(t-1)(r-1),tr(s-1),\alpha}$ are deemed significant at the α level.

EXPECTED MEAN SQUARES: The appropriateness of the tests that we just described can be seen by looking at the expected mean squares in Table 13.52.

Table 13.53: *Seedlings height data from Example* 13.2.

	Fertilizer 1		Fertilizer 2		Fertilizer 3	
Greenhouse	I	II	I	II	I	II
Seedling 1	47	46	62	67	41	42
Seedling 2	43	40	68	71	39	46

Example 13.2 (`seedlings.sas`). The operators of a nursery with $r = 2$ greenhouses would like to investigate differences among $t = 3$ fertilizers they might use on plants they are growing for commercial sale. To set up the experiment, they randomly select 12 similar seedlings and randomly allocate them to 6 trays (experimental units), $s = 2$ per tray. The trays are then randomly allocated to be placed in the two greenhouses, three trays per greenhouse. Within each greenhouse, the three fertilizers are assigned to the trays at random (restricted randomisation) so that each tray receives a different fertilizer. At the end of six weeks, the heights of each seedlings, Y (measured in mm), are collected. The data from the experiment are in Table 13.53. To summarise, we have $N = 12$ observations total with

Treatments	Fertilizers	$t = 3$
Blocks	Greenhouses	$r = 2$
Experimental Units	Trays	$tr = 6$
Sampling Units	Seedlings	$s = 2$

In this problem, we will regard both treatments (fertilizers) and blocks (greenhouses) as **fixed** effects.

CALCULATIONS: We have $Y_{1++} = 176$, $Y_{2++} = 268$, and $Y_{3++} = 168$ (these are the treatment totals), $Y_{+1+} = 300$ and $Y_{+2+} = 312$ (these are the block totals), and $Y_{+++} = 612$ (grand total). The correction term for fitting the overall mean

$$\text{CM} = \frac{1}{trs}\left(\sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{k=1}^{s} Y_{ijk}\right)^2 = (612)^2/12 = 31212.$$

The total sum of squares is given by

$$\text{SS[TOT]} = \sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{k=1}^{s} Y_{ijk}^2 - \text{CM} = (47^2 + 43^2 + \cdots + 46^2) - 31212 = 1642.$$

The treatment and block sums of squares are given by

$$\begin{aligned} \text{SS[T]} &= \frac{1}{rs}\sum_{i=1}^{t} Y_{i++}^2 - \text{CM} = \frac{1}{4}(176^2 + 268^2 + 168^2) - 31212 = 1544 \\ \text{SS[B]} &= \frac{1}{ts}\sum_{j=1}^{r} Y_{+j+}^2 - \text{CM} = \frac{1}{6}(300^2 + 312^2) - 31212 = 12. \end{aligned}$$

The experimental error sum of squares is computed as

$$\begin{aligned} \text{SS[E]} &= \sum_{i=1}^{t}\sum_{j=1}^{r} Y_{ij+}^2 - \text{CM} - \text{SS[T]} - \text{SS[B]} \\ &= (90^2 + 86^2 + \cdots + 88^2) - 31212 - 1544 - 12 = 24. \end{aligned}$$

The sampling error sum of squares is then computed by subtraction; i.e.,

$$\text{SS[S]} = 1642 - 1544 - 12 - 24 = 62.$$

Putting all of these calculations together, we can write the ANOVA table for these data; this is given below.

Source	df	SS	MS	F
Fertilizers (Treatments)	2	1544.00	772.00	$F_T = 64.33$
Greenhouses (Blocks)	1	12.00	12.00	$F_B = 1.00$
Experimental Error	2	24.00	12.00	$F_S = 1.16$
Sampling Error	6	62.00	10.33	
Total	11	1642.00		

ANALYSIS: To test for differences among fertilizers, we compare $F_T = 64.33$ to $F_{2,2,0.05} = 19.00$. Since F_T is large, we have evidence to conclude that the mean heights are different for different fertilizers. Note that the statistics F_B and F_S are both small. Thus, there doesn't look to be a greenhouse effect (was blocking really needed?) and there doesn't look to be a significant amount of variability among the trays (experimental units).

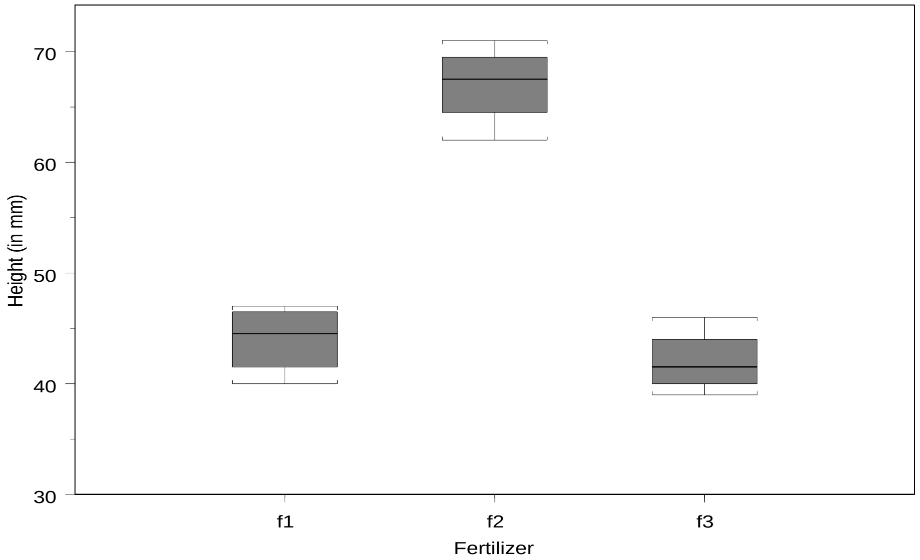

Figure 13.36: *Seedlings height data for three fertilizers.*

PAIRWISE INTERVALS: Since there are significant differences among fertilizers, let's see where they are; we'll do this by writing simultaneous pairwise intervals for $\tau_2 - \tau_1$, $\tau_3 - \tau_1$, and $\tau_3 - \tau_2$. To construct the confidence interval for $\tau_2 - \tau_1$, we can use the point estimator $\overline{Y}_{2++} - \overline{Y}_{1++}$, the difference in the sample means. Under our model assumptions, straightforward calculations show that (verify!)

$$E(\overline{Y}_{2++} - \overline{Y}_{1++}) = \tau_2 - \tau_1$$

and

$$V(\overline{Y}_{2++} - \overline{Y}_{1++}) = \frac{2(\sigma^2 + 2\sigma_\epsilon^2)}{4}.$$

However, note that MS[E] estimates $\sigma^2 + 2\sigma_\epsilon^2$ (see Table 13.52). Since $\overline{Y}_{2++} - \overline{Y}_{1++}$ is a linear combination of the normally distributed Y_{ijk}'s, then $\overline{Y}_{2++} - \overline{Y}_{1++}$, too, is normally distributed. A 95 percent confidence interval for $\tau_2 - \tau_1$ is, thus, given by

$$(\overline{Y}_{2++} - \overline{Y}_{1++}) \pm t_{2,0.025} \times \sqrt{\frac{2\text{MS[E]}}{4}}.$$

The other confidence intervals are formed similarly. To adjust for multiplicity, we can use a Tukey correction. The simultaneous intervals then take the form

$$(\overline{Y}_{2++} - \overline{Y}_{1++}) \pm q_{3,2,0.05} \times \sqrt{\frac{\text{MS[E]}}{4}}$$

$$(\overline{Y}_{3++} - \overline{Y}_{1++}) \pm q_{3,2,0.05} \times \sqrt{\frac{\text{MS[E]}}{4}}$$

$$(\overline{Y}_{3++} - \overline{Y}_{2++}) \pm q_{3,2,0.05} \times \sqrt{\frac{\text{MS[E]}}{4}}$$

(the $\sqrt{2}$ term is absorbed in $q_{3,3,0.05}$). These are **simultaneous** 95 percent confidence intervals for the differences. For these data, $\overline{y}_{1++} = 44$, $\overline{y}_{2++} = 67$, $\overline{y}_{3++} = 42$, and MS[E] $= 12.00$. Unfortunately, I couldn't locate $q_{3,2,0.05}$ anywhere, so I'll refer you to the Tukey intervals on the accompanying SAS program. Based on these intervals, it looks as though fertilizer 2 produces a significantly higher mean height than fertilizers 1 and 3 and that these latter two fertilizers are not significantly different from each other.

NOTE: For those of you that are interested (see me if you want a reference), the value of $q \equiv q_{3,2,0.05}$ solves

$$\int_0^\infty \int_{-\infty}^\infty \left[\Phi(z) - \Phi(z - \sqrt{2}qs)\right]^2 f_Z(z) f_S(s) dz\, ds = 0.95/3,$$

where $\Phi(z)$ is the standard normal cdf, $f_Z(z)$ is the standard normal pdf, and $f_S(s)$ is the pdf of $S \equiv \sqrt{\text{MS[E]}}/\sigma$. Programs are available that solve this double integral numerically.

13.3.3 Replication, no subsampling

So far, we have been concerned with the situation where each treatment is seen only once within each block. There is an intuitively apparent drawback to this type of design; namely, because we only see each treatment once in each block, we do not have sufficient information to understand the interaction between the treatments and the blocks.

Example 13.3. A farmer would like to determine the effect of yield of three cultivars, one of which is known to be drought-resistant. The farmer is interested in how the cultivars differ in the area of his farm where he might plant them. Thus, he divides this area into three blocks, in order to conduct the experiment, basing the blocks on what he best knows about the wetness in various parts of this area. One of the blocks is on a hillside and is quite dry. The farmer expects that the drought-resistant cultivar might give higher yields in this area. However, if the farmer conducted the experiment so that each block was divided into 3 plots, each fertilizer then randomised to a plot, he would only have one experimental unit (plot) per treatment/block combination. From this information, he would not be able to assess the interaction between cultivars and blocks. *Thus, he would not be able to determine whether a high yield for the drought-resistant cultivar in the dry block really was a results of this suspected systematic effect, or just a chance result for the particular plot.*

MAIN POINT: If we only have one experimental unit per treatment/block combination, we can not understand how the treatments and blocks **interact**. To measure this interaction, we must have more than one experimental unit per treatment/block combination; that is, we must have **replicates** of the experimental units.

LINEAR MODEL: We now formally write a linear model for this situation (assuming no subsampling). Here, we may identify an observation as being on the kth experimental unit in the jth block receiving the ith treatment. The model is given by

$$Y_{ijk} = \mu + \tau_i + \rho_j + (\tau\rho)_{ij} + \epsilon_{ijk},$$

for $i = 1, 2, ..., t$, $j = 1, 2, ..., r$, and $k = 1, 2, ..., n_{ij}$, where $\epsilon_{ijk} \sim$ iid $\mathcal{N}(0, \sigma^2)$. Here, n_{ij} denotes the number of experimental units for the ith treatment and jth block. For simplicity, we'll assume that $n_{ij} = n$ (a balanced design) for all i and j. As in the last section, either τ_i or ρ_j (or both) may be regarded as having fixed or random effects (more on this momentarily). The total number of observations (assuming balance) is $N = trn$.

INTERPRETATION: Compare this model to the model for the subsampling model given in the last section. In this model, the components, $(\tau\rho)_{ij}$ and ϵ_{ijk} replace ϵ_{ij} and δ_{ijk} in the subsampling model, respectively (and I use n for the number of experimental unit replicates instead of s for the number of sampling units). Other than this change

Table 13.54: *ANOVA table for the RCBD; more than one experimental unit, with no subsampling.*

Source	df	SS	MS	F
Treatments	$t-1$	SS[T]	MS[T]	F_T
Blocks	$r-1$	SS[B]	MS[B]	F_B
Interaction	$(t-1)(r-1)$	SS[I]	MS[I]	F_I
Experimental Error	$tr(n-1)$	SS[E]	MS[E]	
Total	$N-1$	SS[TOT]		

in symbols, the models look similar. However, the interpretation is quite different!! It is important that you feel comfortable with the fact that, although the models have a similar form, they represent experiments that are very different.

PUNCHLINE: Algebraically, it turns out that the same computations apply in constructing the ANOVA in this setting as did in the subsampling model from the last section. The ANOVA table will be the same regardless of whether the τ_i and/or ρ_j are treated as fixed or random. The breakdown of SS[TOT] into its different sources of variation is based on the identity

$$\begin{aligned}\underbrace{\sum_{i=1}^t\sum_{j=1}^r\sum_{k=1}^n(Y_{ijk}-\overline{Y}_{+++})^2}_{\text{SS[TOT]}} &= \underbrace{rn\sum_{i=1}^t(\overline{Y}_{i++}-\overline{Y}_{+++})^2}_{\text{SS[T]}}+\underbrace{tn\sum_{j=1}^r(\overline{Y}_{+j+}-\overline{Y}_{+++})^2}_{\text{SS[B]}}\\ &\quad +\underbrace{n\sum_{i=1}^t\sum_{j=1}^r(\overline{Y}_{ij+}-\overline{Y}_{+++})^2}_{\text{SS[I]}}+\underbrace{\sum_{i=1}^t\sum_{j=1}^r\sum_{k=1}^n(Y_{ijk}-\overline{Y}_{ij+})^2}_{\text{SS[E]}}.\end{aligned}$$

HAND COMPUTATION: For this situation, performing hand computation of the sums of squares is easiest using the following formulae. First, compute the **correction term** for fitting the overall mean

$$\text{CM}=\frac{1}{trn}\left(\sum_{i=1}^t\sum_{j=1}^r\sum_{k=1}^n Y_{ijk}\right)^2=Y_{+++}^2/N,$$

where $N = trn$. The total sum of squares is given by

$$\text{SS[TOT]} = \sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{k=1}^{n} Y_{ijk}^2 - \text{CM}.$$

The treatment and block sums of squares are given by

$$\begin{aligned}\text{SS[T]} &= rn\sum_{i=1}^{t}(\overline{Y}_{i++} - \overline{Y}_{+++})^2 = \frac{1}{rn}\sum_{i=1}^{t} Y_{i++}^2 - \text{CM} \\ \text{SS[B]} &= tn\sum_{j=1}^{r}(\overline{Y}_{+j+} - \overline{Y}_{+++})^2 = \frac{1}{tn}\sum_{j=1}^{r} Y_{+j+}^2 - \text{CM}.\end{aligned}$$

The treatment/block **interaction** error sum of squares is computed as

$$\text{SS[I]} = \sum_{i=1}^{t}\sum_{j=1}^{r} Y_{ij+}^2 - \text{CM} - \text{SS[T]} - \text{SS[B]}.$$

The experimental error sum of squares is then computed by subtraction; i.e.,

$$\text{SS[E]} = \text{SS[TOT]} - \text{SS[T]} - \text{SS[B]} - \text{SS[I]}.$$

FIXED VERSUS RANDOM EFFECTS: Consider our linear model for analysing balanced data in a RCBD with experimental unit replication; i.e.,

$$Y_{ijk} = \mu + \tau_i + \rho_j + (\tau\rho)_{ij} + \epsilon_{ijk},$$

for $i = 1, 2, ..., t$, $j = 1, 2, ..., r$, and $k = 1, 2, ..., n$, where $\epsilon_{ijk} \sim \text{iid } \mathcal{N}(0, \sigma^2)$. As usual, tests of hypotheses are conducted by using F statistics; however, with this model, we have to be very careful about our interpretations of the treatment and block effects. In particular, the forms of the tests depend greatly on whether or not these factors are best regarded as fixed or random. You'll note that I didn't give the formulae for the F tests in Table 13.54 just yet (for good reason). Let's look at the different possible scenarios.

- **Fixed treatments, fixed blocks.** If the τ_i and ρ_j are best regarded as fixed effects, the systematic interaction effect $(\tau\rho)_{ij}$ is also regarded as fixed. To compute expected mean squares in this situation, the standard side conditions $\tau_+ = \rho_+ = (\tau\rho)_{i+} = (\tau\rho)_{+j} = 0$ are assumed. In this situation, there are no additional model assumptions needed.

- **Random treatments, random blocks.** If the τ_i and ρ_j are best regarded as random effects, the systematic interaction effect $(\tau\rho)_{ij}$ is also regarded as random. In this situation, it is common to assume that $\tau_i \sim$ iid $\mathcal{N}(0, \sigma_\tau^2)$, $\rho_j \sim$ iid $\mathcal{N}(0, \sigma_\rho^2)$, $(\tau\rho)_{ij} \sim$ iid $\mathcal{N}(0, \sigma_{\tau\rho}^2)$, and that τ_i, ρ_j, $(\tau\rho)_{ij}$, and ϵ_{ijk} are independent random variables.

- **Fixed treatments, random blocks.** If the τ_i are fixed effects and the ρ_j are random, how should we think of the interaction term $(\tau\rho)_{ij}$? When blocks are random, we may think of each block in the population of all possible blocks as having a $(\tau\rho)_{ij}$ value associated with it; i.e., if treatment i were applied to a block chosen from this population, the associated deviation would be this value. From this perspective, it seems sensible to think of the $(\tau\rho)_{ij}$ as being random as well. One can envision a population of $(\tau\rho)_{ij}$ values for each treatment i containing all the possible deviations that arise for each possible block. If we think of the $(\tau\rho)_{ij}$ in this way, then, analogous to our last situation, we would assume that $\rho_j \sim$ iid $\mathcal{N}(0, \sigma_\rho^2)$, $(\tau\rho)_{ij} \sim$ iid $\mathcal{N}(0, \sigma_{\tau\rho}^2)$, and that ρ_j, $(\tau\rho)_{ij}$, and ϵ_{ijk} are independent random variables. To compute expected mean squares in this situation, we will assume the restriction $(\tau\rho)_{+j} = 0$; that is, summing the interaction component over the fixed treatment factor equals zero. This restriction implies that certain interaction elements at different levels of the fixed factor are not independent. This version of the mixed model is often called the **restricted** model, because we are imposing the "sum to zero" restriction previously mentioned. There are some authors that do not advocate the use of this restriction! If this restriction is not assumed, the mixed model is often called the **unrestricted** mixed model.

THE MIXED MODEL CONTROVERSY: In light of this multiplicity of mixed models (one that incorporates the restriction and one that does not), which one should we use? I have found that more statisticians prefer the restricted model, and it is the most widely encountered in the literature. The restricted model is actually slightly more general than the unrestricted model, because in the restricted model, the covariance between two observations from the same level of the random factor can be either positive or negative,

Table 13.55: *EMS for the RCBD with replication and no subsampling.*

	Expected mean squares (EMS)		
Source of variation	Both fixed	Both random	(Restricted) Mixed
Treatments	$\sigma^2 + \frac{rn\sum_{i=1}^{t}\tau_i^2}{t-1}$	$\sigma^2 + n\sigma^2_{\tau\rho} + rn\sigma^2_{\tau}$	$\sigma^2 + n\sigma^2_{\tau\rho} + \frac{rn\sum_{i=1}^{t}\tau_i^2}{t-1}$
Blocks	$\sigma^2 + \frac{tn\sum_{j=1}^{r}\rho_j^2}{r-1}$	$\sigma^2 + n\sigma^2_{\tau\rho} + tn\sigma^2_{\rho}$	$\sigma^2 + tn\sigma^2_{\rho}$
Interaction	$\sigma^2 + \frac{n\sum_{i=1}^{t}\sum_{j=1}^{r}(\tau\rho)^2_{ij}}{(t-1)(r-1)}$	$\sigma^2 + n\sigma^2_{\tau\rho}$	$\sigma^2 + n\sigma^2_{\tau\rho}$
Experimental Error	σ^2	σ^2	σ^2

while this covariance can only be positive in the unrestricted model. Voss (1999) discusses the controversy over these two models. At this point, I would like for you to be aware of what the controversy is all about; not necessarily the mathematics that underlies the controversy. *The main point is that the form of tests for random effects can change, depending on which mixed model interpretation you adopt!*

HYPOTHESIS TESTS: If the fixed or random effects models, there is no controversy. From Table 13.55, we see that in the **fixed** case,

$$\begin{aligned} F_T &= \frac{\text{MS[T]}}{\text{MS[E]}} \quad \text{tests} \quad H_0: \tau_1 = \tau_2 = \cdots = \tau_t = 0 \\ F_B &= \frac{\text{MS[B]}}{\text{MS[E]}} \quad \text{tests} \quad H_0: \rho_1 = \rho_2 = \cdots = \rho_r = 0 \\ F_I &= \frac{\text{MS[I]}}{\text{MS[E]}} \quad \text{tests} \quad H_0: (\tau\rho)_{ij} = 0, \text{ for all } i \text{ and } j. \end{aligned}$$

Likewise, in the **random** case, we see that

$$\begin{aligned} F_T &= \frac{\text{MS[T]}}{\text{MS[I]}} \quad \text{tests} \quad H_0: \sigma^2_\tau = 0 \\ F_B &= \frac{\text{MS[B]}}{\text{MS[I]}} \quad \text{tests} \quad H_0: \sigma^2_\rho = 0 \\ F_I &= \frac{\text{MS[I]}}{\text{MS[E]}} \quad \text{tests} \quad H_0: \sigma^2_{\tau\rho} = 0. \end{aligned}$$

When we are in a mixed-model situation, there are some ambiguities that surface. In the **restricted mixed** model case; i.e., the model where we assume $(\tau\rho)_{+j} = 0$ for all j, we see that

$$F_T = \frac{\text{MS[T]}}{\text{MS[I]}} \quad \text{tests} \quad H_0 : \tau_1 = \tau_2 = \cdots = \tau_t = 0$$
$$F_B = \frac{\text{MS[B]}}{\text{MS[E]}} \quad \text{tests} \quad H_0 : \sigma^2_\rho = 0$$
$$F_I = \frac{\text{MS[I]}}{\text{MS[E]}} \quad \text{tests} \quad H_0 : \sigma^2_{\tau\rho} = 0.$$

The appropriateness of these tests can be seen from Table 13.55. However, in the **unrestricted mixed** model case, it follows that $E(\text{MS[B]}) = \sigma^2 + tn\sigma^2_\rho + n\sigma^2_{\tau\rho}$ (all other expected mean squares remain unchanged); this fact suggests that

$$F_B = \frac{\text{MS[B]}}{\text{MS[I]}} \quad \text{tests} \quad H_0 : \sigma^2_\rho = 0.$$

Hence, the controversy! *Tests for block effects are different depending on whether the restricted or unrestricted model is assumed!*

Example 13.4 (`calcium.sas`). The following experiment was set up to determine the effect of a certain hormone treatment on the plasma calcium level, Y, (measured in mg/100 ml) for a certain type of bird. It was thought that the gender of the birds might play a role. Random samples of 10 male birds and 10 female birds were obtained. For each gender, 5 birds were randomly assigned to receive the hormone treatment, and the remaining 5 did not. Thus, we may view this as a RCBD where the blocks are the genders ($r = 2$), which are obviously fixed. Within each block, experimental units (i.e., the birds) were randomised to receive the treatments (hormone/no hormone; $t = 2$). There are $n = 5$ replicates for each treatment/block combination. A single plasma calcium level was recorded for each bird. The treatments are obviously fixed as well; thus, for this situation, we might consider the fixed effects model

$$Y_{ijk} = \mu + \tau_i + \rho_j + (\tau\rho)_{ij} + \epsilon_{ijk},$$

for $i = 1, 2$, $j = 1, 2$, and $k = 1, 2, ..., 5$, where $\epsilon_{ijk} \sim \text{iid } \mathcal{N}(0, \sigma^2)$. It was suspected that the difference between having the hormone treatment or not might be different depending on whether the bird was male or female; that is, a hormone-gender interaction was suspected. The data from the experiment are in Table 13.56.

Table 13.56: *Calcium level data.*

Males		Females	
Hormone	No hormone	Hormone	No hormone
32.0	14.5	39.1	16.5
23.8	11.0	26.2	18.4
28.8	10.8	21.3	12.7
25.0	14.3	35.8	14.0
29.3	10.0	40.2	12.8

ANALYSIS: From SAS, I have computed the ANOVA table (hand computations would mirror those from the last subsection, so I have omitted them).

Source	df	SS	MS	F
Hormone (Treatments)	1	1386.11	1386.11	60.53
Gender (Blocks)	1	70.31	70.31	3.07
Interaction	1	4.90	4.90	0.21
Experimental Error	16	366.37	22.90	
Total	19	1827.69		

Since the τ_i and ρ_j are both best regarded as fixed effects, there is no controversy in the formulation of tests. All three tests (for a treatment effect, block effect, and interaction effect) use $F_{1,16,0.05} = 4.49$ as a critical value. Clearly, there is a significant difference between the hormone/no hormone treatment. There doesn't seem to be a large difference between the genders (was blocking really useful?), nor does there appear to be a hormone-gender interaction. From Figure 13.37, it is easy to see that the hormone application **significantly** increases the mean calcium level (in light of our significant F test for hormone application). To assess whether or not this difference is practical, we can write a confidence interval for $\tau_2 - \tau_1$, the difference in means for the levels of hormone. In the fixed effects model $Y_{ijk} = \mu + \tau_i + \rho_j + (\tau\rho)_{ij} + \epsilon_{ijk}$, for $i = 1, 2$, $j = 1, 2$, and $k = 1, 2, ..., 5$,

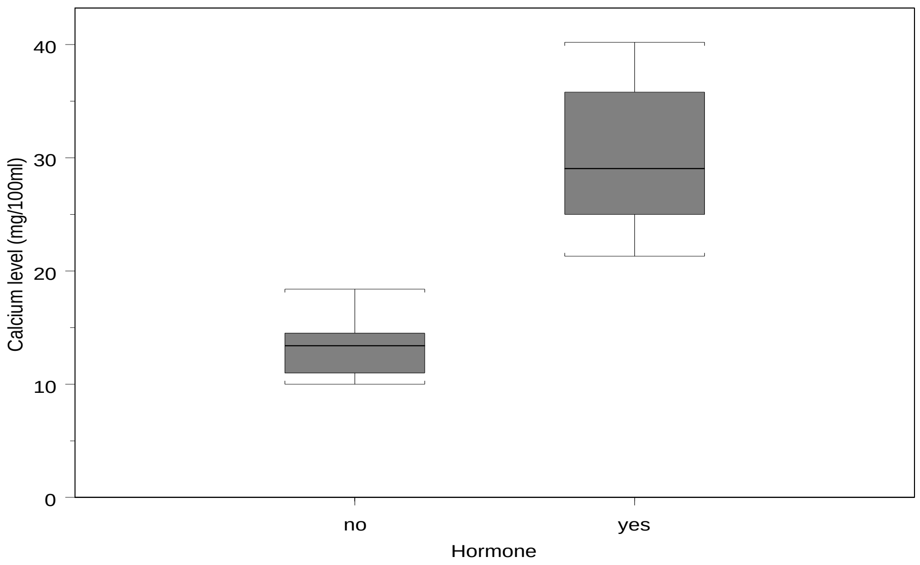

Figure 13.37: *Calcium levels for different levels of hormone application.*

where $\epsilon_{ijk} \sim$ iid $\mathcal{N}(0, \sigma^2)$, it is easy to show that (verify!)

$$E(\overline{Y}_{2++} - \overline{Y}_{1++}) = \tau_2 - \tau_1$$

and

$$V(\overline{Y}_{2++} - \overline{Y}_{1++}) = \frac{2\sigma^2}{10}.$$

A 95 percent confidence interval for $\tau_2 - \tau_1$ becomes

$$(\overline{Y}_{2++} - \overline{Y}_{1++}) \pm t_{16,0.025} \times \sqrt{\frac{2\text{MS[E]}}{10}}.$$

Here, I'll refer to level 2 as hormone application and level 1 as no hormone application. With these data, $\overline{y}_{1++} = 13.50$, $\overline{y}_{2++} = 30.15$, $t_{16,0.025} = 2.1199$, and MS[E] $= 22.90$. A 95 percent confidence interval for $\tau_2 - \tau_1$ is $(12.11, 21.19)$. Again, note the additional information that the confidence interval confers that the F test does not. We have a **statistically significant** result here (i.e., F test rejects; CI doesn't include 0). However, if the experimenter was hoping for a 25 mg mean increase in calcium levels, say, from the application of the hormone, this would not be a **practically significant** result.