4452 Mathematical Modeling Lecture 17: Modeling of Data: Linear Regression

Introduction

In modeling of data, we are given a set of data points, and we want to fit a function with adjustable parameters to the data points. Obviously, we want the function to approximate the data as well as possible, and to do that you have to choose certain values for the parameters in your function. You choose these parameter values by first designing a merit function which you wish to minimize. When the merit function is minimized, the agreement between the function and the data will have close agreement.

You can see that fitting a function to data becomes a problem of minimization in many dimensions (the number of adjustable parameters in your function is the dimension of the problem).

Once we have fit the function to the data, we need to assess how good the fit actually is. There has to be some sort of statistical analysis of the fit.

The bulk of this discussion was based on Reference [1], which is an excellent first resource for a variety of applied numerical analysis.

General Set Up

We have N data points (t_i, y_i), $i = 1, 2, \ldots, N$ which we want to fit to a model function which has M adjustable parameters, $f(t) = f(t; \alpha_1, \ldots, \alpha_M)$.

We actually have a great deal of choice in what type of function we want to minimize. It can be anything that will measure the relation of the data to the model function. The vector which compares the data to the model function at each point is given by

$$\begin{pmatrix} y_1 - f(t_1; \alpha_1, \ldots, \alpha_M) \\ y_2 - f(t_2; \alpha_1, \ldots, \alpha_M) \\ \vdots \\ y_N - f(t_N; \alpha_1, \ldots, \alpha_M) \end{pmatrix}$$

We can minimize this vector based on a variety of different norms:

$$\begin{aligned} l_1 \text{ norm:} \quad & \sum_{i=1}^{N} |y_i - f(t_i; \alpha_1, \ldots, \alpha_M)| \\ l_p \text{ norm:} \quad & \left(\sum_{i=1}^{N} (y_i - f(t_i; \alpha_1, \ldots, \alpha_M))^p \right)^{1/p} \\ l_\infty \text{ norm:} \quad & \max_{i=1}^{N} (y_i - f(t_i; \alpha_1, \ldots, \alpha_M)) \end{aligned}$$

What is typically done is that the l_2 norm is used, since it is the Euclidean space norm, and the *square* of

the norm is minimized (hence the name: least squares fit):

$$\text{minimize } F(\alpha_1, \ldots, \alpha_N) = \sum_{i=1}^{N} (y_i - f(t_i; \alpha_1, \ldots, \alpha_M))^2$$

The above result assumes we know the data points only. However, what if we know the data points and for each data point a standard deviation, σ_i? How could this be incorporated into the function we wish to minimize? We can do the following

$$\text{minimize } F(\alpha_1, \ldots, \alpha_N) = \sum_{i=1}^{N} \left(\frac{y_i - f(t_i; \alpha_1, \ldots, \alpha_M)}{\sigma_i} \right)^2 \tag{1}$$

which assumes that each data point has a measurement error which is independently random and distributed as a normal distribution around the actual model $f(t)$. This result is based on a great deal of statistics, and the idea that random deviations will converge to a normal distribution. Of course, this may not be the case in practice.

Frequently, it is true that $\sigma_i = \sigma$ is the same for all the data points. In that case, σ^2 can be factored out of the sum and the σ does not appear in the solution for α_1 and α_2. Since this is the case, if you are given data which does not have an associated error σ_i which depends on the data point you can simply set $\sigma_i = 1$ and proceed with the analysis.

Minimizing Eq. (1) is just a multivariable unconstrained minimization procedure, which yields the system of equations

$$0 = \sum_{i=1}^{N} \left(\frac{y_i - f(t_i; \alpha_1, \ldots, \alpha_M)}{\sigma_i^2} \right) \left(\frac{\partial}{\partial \alpha_k} f(t_i; \alpha_1, \ldots, \alpha_M) \right), \quad k = 1, \ldots, M \tag{2}$$

which must be solved for the M unknowns α_i.

Linear Regression

Linear regression does *not* mean fitting data to a straight line! The "linear" refers to the models dependence on the parameters α_k. However, for now we are interested in fitting to a straight line. In this case, our fitting function becomes

$$f(t; \alpha_1, \alpha_2) = \alpha_1 + \alpha_2 t.$$

The system of equations in Eq. (2) becomes

$$\begin{aligned} \alpha_1 \sum_{i=1}^{N} \frac{1}{\sigma_i^2} + \alpha_2 \sum_{i=1}^{N} \frac{t_i}{\sigma_i^2} &= \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2} \\ \alpha_1 \sum_{i=1}^{N} \frac{t_i}{\sigma_i^2} + \alpha_2 \sum_{i=1}^{N} \frac{t_i^2}{\sigma_i^2} &= \sum_{i=1}^{N} \frac{y_i t_i}{\sigma_i^2} \end{aligned} \tag{3}$$

We can simplify the notation if we use the following:

$$S = \sum_{i=1}^{N} \frac{1}{\sigma_i^2}, \quad S_t = \sum_{i=1}^{N} \frac{t_i}{\sigma_i^2}, \quad S_y = \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2}, \quad S_{tt} = \sum_{i=1}^{N} \frac{t_i^2}{\sigma_i^2}, \quad S_{ty} = \sum_{i=1}^{N} \frac{t_i y_i}{\sigma_i^2}, \quad \Delta = S S_{tt} - S_t^2.$$

The solution to Eq. (3) is given by

$$\begin{aligned} \alpha_1 &= \frac{S_{tt} S_y - S_t S_{ty}}{\Delta} \\ \alpha_2 &= \frac{S S_{ty} - S_t S_y}{\Delta} \end{aligned}$$

The Correlation Coefficient–How Good is Our Model Function?

All we need now is an estimate of how good our linear fit is. Reference [1] has a significantly expanded discussion on determining how good your linear regression model is.

We will consider the case where $\sigma_i \sim \sigma$ for all i. This is frequently not a restrictive assumption, since the sources of error in measuring the data that lead to σ_i are frequently the same for all measurements.

We calculate what is called the *correlation coefficient*, R^2, which is a ratio of the model sum of the squares to the total sum of the squares

$$R^2 = \frac{\sum_{i=1}^{N} (f(t_i) - \bar{y})^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2}, \quad \bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i.$$

Thus, if $R^2 \sim 1$ the model is a good representation of the data, and the data are representative of a linear function. If $R^2 \sim 0$ the data are essentially random, and a linear function cannot represent the data well.

The results of this analysis for a particular data set (contained in the corresponding *Mathematica* file) are shown in Fig. 1. The data set is given for completeness in the Appendix.

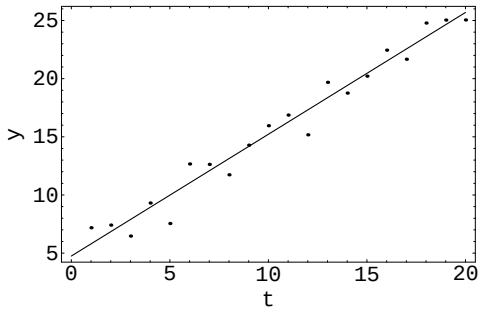

Figure 1: Linear fit, $f(t) = 4.74964 + 1.04711t$, to a data set. For this fit, the correlation coefficient was found to be $R^2 = 0.96$, which indicates that the data does represent a linear function, and the linear function we found represents the data well.

Extrapolation

Once we have found the linear model of the data, what is it good for? Many times, what is of interest is the slope of the curve, or the y-intercept. Or, the model can be used for extrapolation. In any case, we would like an estimate of the standard deviation of the model from the data. We can get an estimate by computing

$$\sigma = \sqrt{\sum_{i=1}^{N}(y_i - f(t_i))^2} \tag{4}$$

If our data is normally distributed about the model function (which it may very well not be!), we would expect measurements will be within $\pm\sigma$ of the model function 68% of the time, and within $\pm 2\sigma$ 95% of the time. Figure 2 shows this result.

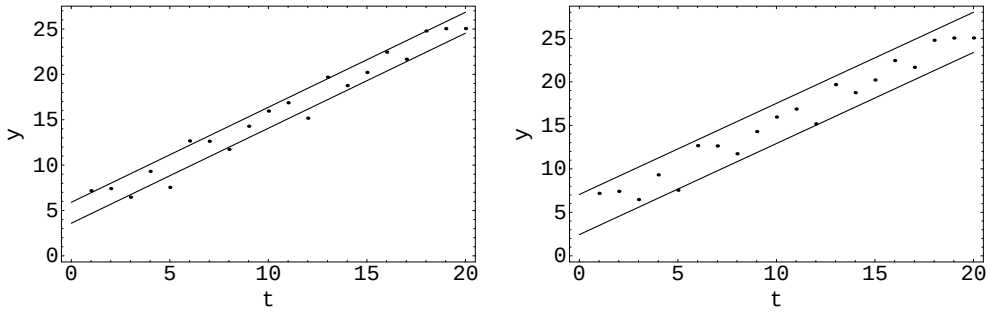

Figure 2: An example of the number of data points contained within $\pm\sigma$ (left) and $\pm 2\sigma$ (right) of the model function, with $\sigma = 1.15$ from Eq. (4). We get 60% within $\pm\sigma$ and 95% within $\pm 2\sigma$.

We can use this σ to estimate the error in extrapolation. Since we are assuming the model function is accurately representing the underlying linear dependence of the data, the random fluctuations in measuring the data would still account for error in future measurement. We would expect, for example, since $f(25) = 30.9237$ and $\sigma = 1.15$ that if we measured the system at $t = 25$ we would find

$$29.7693 \leq y_{25} \leq 32.0781$$

95% of the time.

Outliers

The average of all the data points is the point $(\bar{t}, \bar{y})$ where

$$\bar{t} = \frac{1}{N}\sum_{i=1}^{N} t_i, \quad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

The model function will always go through this point if we have $\sigma_i = 1$. In a sense, this represents the centroid of the data set.

An *outlier* is a data point which lies far off the regression line. The effect of this outlier is to inordinately affect the underlying model function. Figure 3 shows a data set with an outlier, and the fit which was obtained.

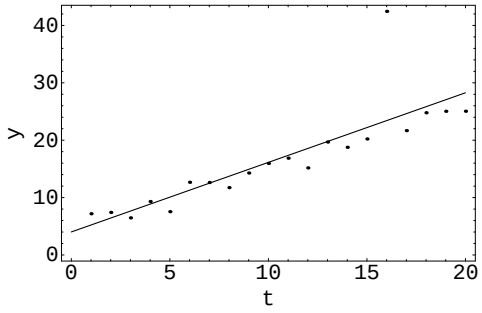

Figure 3: Linear fit to a data set containing an outlier. For this fit, the correlation coefficient was found to be $R^2 = 0.69$, which should make us question whether or not the fit is accurately representing the data. The outlier is strong enough to make all the other points from 10 to 20 lie beneath the model function.

It is imperative that any analysis of data sets which uses linear regression includes some analysis of outliers, especially for small data sets, like the one pictured, for which the effect of one outlier can be enormous. If an outlier can be identified, it should be deleted from the data set and the linear regression redone.

You can detect outliers either visually (if they are obviously outliers, like in my example), or by a more systematic analysis of the quantity $|y_i - f(t_i)|$. If this quantity is extremely large for a few data points, they may be outliers. A more statistical way to find outliers it is to search the data for points for which $|y_i - f(t_i)| > 2\sigma$.

Obviously, deleting data from your data set should be done with extreme caution, and must be reported fully in any use of the data set or model function which is created after the removal of outliers.

Other Forms of Linear Regression

As mentioned before, the linear in *linear regression* comes not from the fact that we are fitting a straight line to the data points, but that we are fitting a model function which depends linearly on the fit parameters α_k. We could be sines, cosines, or other powers of t.

Let's look at an example of using a different fit model. Consider the data set shown in Fig. 4.

From looking at the data, we think that it looks quadratic. So our model function should be chosen to be

$$f(t;\alpha) = \alpha t^2.$$

Note that if we wanted to chose a model function that looked like $\alpha_1 t^{\alpha_2 2}$ this would be a nonlinear regression problem, since the parameters α_k no longer appear in a linear manner.

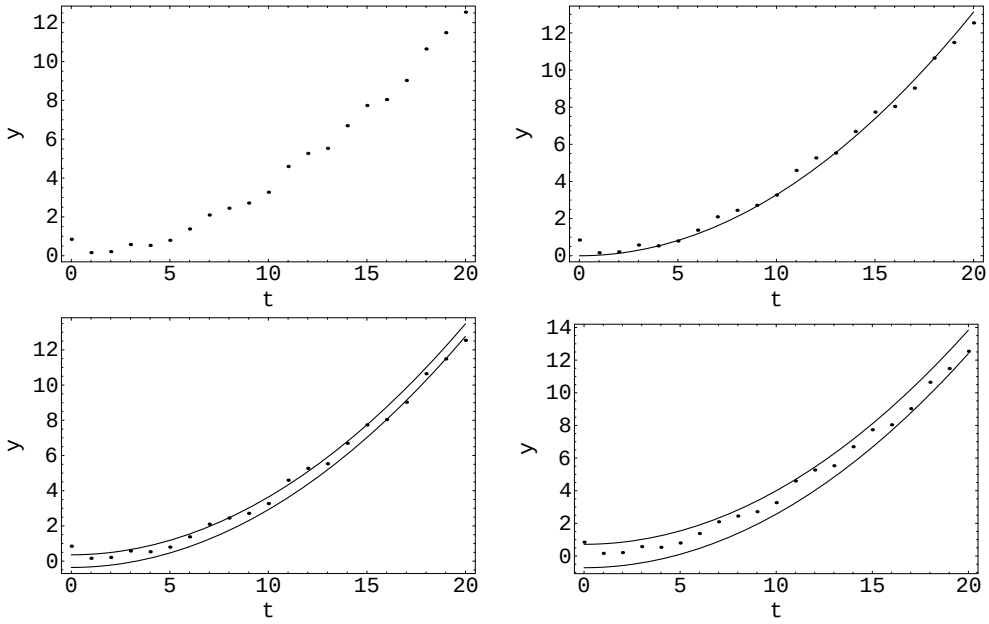

Figure 4: Data set which is not modeled well by a straight line (top left); linear regression model function $f(t) = 0.0327963\,t^2$, $\sigma = 0.36$ (top right); $\pm\sigma$ interval contains 70% of the points (bottom left); $\pm 2\sigma$ interval contains 95% of the points (bottom right).

We can use the linear regression ideas to determine α. The system in Eq. (2) is now simply the equation

$$0 = \sum_{i=1}^{N} (y_i - \alpha t_i^2) t_i^2$$

which means the parameter α is given by

$$\alpha = \frac{\sum_{i=1}^{N} y_i t_i^2}{\sum_{i=1}^{N} t_i^4}.$$

I again estimated σ as

$$\sigma = \sqrt{\sum_{i=1}^{N} (y_i - f(t_i))^2}.$$

Final Thought

My analysis in this lecture is satisfactory for most applications of linear regression you will run into. However, if you get into serious linear regression applications, you should understand the deeper statistical theory behind it. You will need to have a deeper understanding if you use statistics packages to do your regression for you, since you should always understand what it is that the computer is telling you. *Mathematica* produces

a tremendous amount of output with its linear regression package, and I have to say that I understand only a small portion of it. If I use it, I only use the packages I understand.

Appendix

Data set used in Figs. 1 and 2:

```
{{1, 7.18437}, {2, 7.41171}, {3, 6.46758}, {4, 9.30980}, {5, 7.54707},
 {6,12.6665}, {7, 12.63065}, {8, 11.73464}, {9, 14.28273}, {10, 15.95493},
 {11, 16.86929}, {12, 15.17158}, {13, 19.68491}, {14, 18.75987}, {15, 20.21733},
 {16, 22.45007}, {17, 21.67026}, {18, 24.78354}, {19, 25.03960}, {20, 25.04889}}
```

Data set with outlier, used in Fig. 3:

```
{{1, 7.18437}, {2, 7.41171}, {3, 6.46758}, {4, 9.30980}, {5, 7.54707},
 {6, 12.6665}, {7, 12.63065}, {8, 11.73464}, {9, 14.28273}, {10, 15.95493},
 {11, 16.86929}, {12, 15.17158}, {13, 19.68491}, {14, 18.75987}, {15, 20.21733},
 {16, 42.45007}, {17, 21.67026}, {18, 24.78354}, {19, 25.03960}, {20, 25.04889}}
```

Data set used in Fig. 4:

```
{{0, 0.847361}, {1, 0.161051}, {2, 0.209979}, {3, 0.576722}, {4, 0.533068},
 {5, 0.795271}, {6, 1.37877}, {7, 2.09677}, {8, 2.44694}, {9, 2.71329},
 {10, 3.26994}, {11, 4.59521}, {12, 5.26873}, {13, 5.52998}, {14, 6.69377},
 {15, 7.73503}, {16, 8.03944}, {17, 9.02567}, {18, 10.6446}, {19, 11.4847},
 {20, 12.5403}}
```

References

[1] W. H. Press, S. A. Teukolsky, W. T. Vetterling & B. P. Flannery, Numerical Recipes in Fortran 77, 2nd ed, Cambridge University Press (New York) 1992.

[2] M. Meerschaert, Mathematical Modelling, 2nd ed., Academic Press (San Diego) 1999.