Linear Models

Ethan D. Bolker
Maura B. Mast

November 20, 2007

Plan

- Constructing linear models by eye and with Excel
- Using linear models for predictions
- How good is the model?
- Correlation versus causation

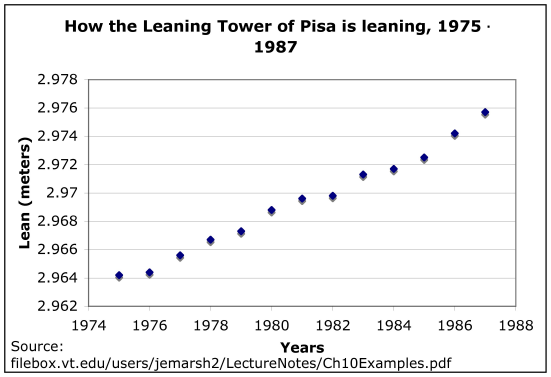
Lecture notes

Constructing linear models by eye and with Excel

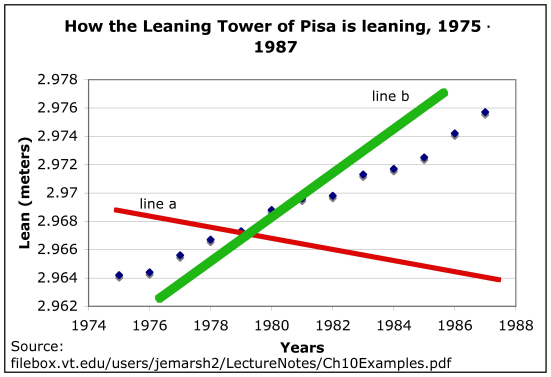
We have seen that linear functions are very easy to work with, especially when we want to do calculations. Data in real life, however, rarely follow a perfect linear pattern. That is, if you graph the data points it is unlikely that they will all lie on one line. But there may be a rough linear pattern to the data, in that the data point show a linear trend. We use this idea to construct a *linear model* that helps us understand the data. A model is a way of making complicated information (or data) easier to understand. In this context, we find a linear function that approximates the trend that we see in the data. This is called a linear model (later we'll build exponential models, using exponential functions). Whenever we make this kind of approximation, we have to be careful. It is important to remember that we are simplifying the data trend, and to ask how well this simplification really captures the data pattern (we'll see a way to measure this).

We can build linear models by hand or using Excel. Excel is very useful in this type of situation as it performs the rather messy calculation needed to construct the line. We will do this first by eye, then see how close our intuition for the linear model matches Excel's calculation.

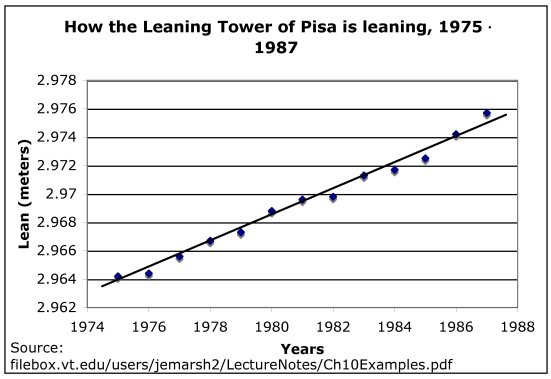
Open the data set `http://www.cs.umb.edu/~eb/m114/lectureNotes/1120/pisaData.xls`.This file gives some measurements for the amount of lean that the Leaning Tower of Pisa experienced between 1975 and 1987. The Leaning Tower of Pisa was constructed in the 1170s in Pisa, Italy and, due to foundation issues, began to lean soon after construction began. In this file, the lean (the distance between where a point on the tower would be if the tower were straight and where it actually is) is measured in meters. A scatterplot of the data shows a clear linear trend:



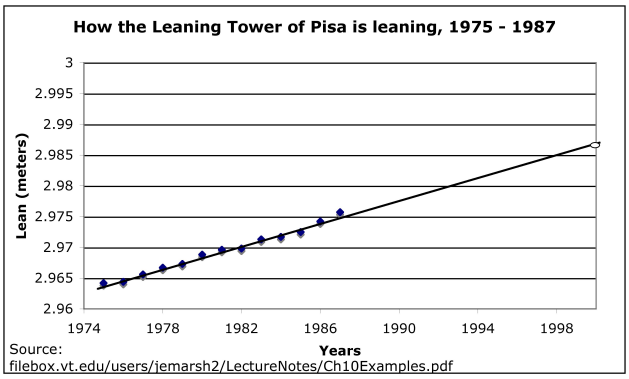
To find a line that captures this trend, you can print the graph and use a ruler to sketch a line, or you can use the Drawing tool in Excel. You want to sketch a line that follows the trend of the data. The following graph shows two lines that are not good candidates:



The red line doesn't capture the upward trend in the data (that is, as the years pass, the amount of lean increases). The green line, line b, does show the upward trend but it is not very close to the data points. We could do better by sketching a line that goes as close to as many points on the graph as possible, as the next graph shows.

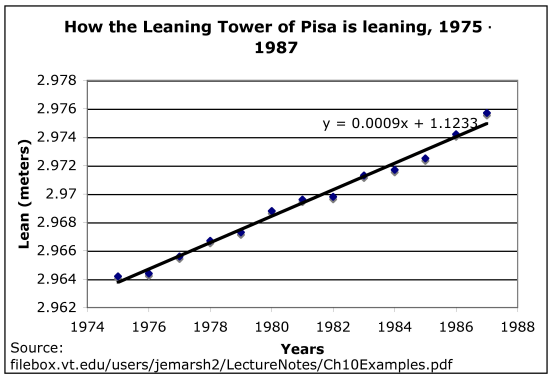


Now we can use this line as a model. That is, we can use the line to estimate information about the lean in the tower. For example, we could use our line to predict the amount of lean in the year 2000. To do this, we would have to extend our line and estimate the amount of lean from the graph. We need to change the scale on the graph to extend it to 2000, then extend our line. To change the scale, double-click on one of the years and set the maximum year to be 2000. Once you do this, you'll see that you also need to adjust the vertical scale. By extending the line, we estimate that in the year 2000, the tower will be leaning 2.986 meters. The circle in the picture below shows the point (2000,2.986).



The work that we just did is fine for a rough estimate. We can be more precise by drawing a *regression line*. This line, also called a *line of best fit* or *trendline*,

is calculated using a least square formula. This formula produces a line that is as close as possible (in terms of vertical distance) to as many of the points as possible. It is not a difficult formula, but it is tedious. Fortunately, Excel will do the calculations for us. To use Excel for this, click once on any of the points on the scatter plot (the points should all be highlighted after you do this. Go to the Chart menu and click on "Add Trendline". In the window that appears, click on "Linear", then go to the Options tab at the top and click on that. Check the box that reads "Display equation on chart" and then click OK. Excel will put the trendline on the graph and display the equation for the line, which is $y = 0.0009x + 1.1233$. Here is the picture:



To predict the lean in the year 2000, we can either extend this line or insert $x = 2000$ into the equation. Remember that x represents the year and y represents the lean in meters. Doing the calculation, we obtain

$$y = 0.0009 * 2000 + 1.1233 = 2.9233.$$

In other words, we use this trendline to predict that in the year 2000, the lean will be 2.9233 meters.

As always, when technology gives us an answer it is a good idea to step back and think about the answer. Is it really correct? Should we believe it just because the software program told us it is the answer? Sometimes we can, but in this case we should be a bit concerned. When we estimated the lean in the year 2000 by hand, we came up with about 2.986 meters. Using the trendline for Excel, we calculated 2.9233 meters. This number is smaller than the numbers in our data set. If we believe this number, then that means that the Tower began to reverse its lean. The data (and the graph) don't support this, so we need to look a bit deeper to see what's going on.

It turns out that Excel is rounding the slope and the y-intercept values in the trendline that it calculates. It writes these numbers to four decimal places, then rounds. Often this is fine, as we aren't interested in four decimal places of accuracy. But in this example it turns out to be important enough to give us a confusing answer. If we ask Excel to re-do the calculation, with more decimal places, we will get a more accurate answer. Here's one way to do this. We can ask Excel to directly calculate the slope and intercept of a line through a set of data points. Instead of finding the trendline, we ask Excel to first find the slope, and then the intercept. The commands are straightforward: SLOPE and INTERCEPT are the functions that Excel uses. You need to first highlight the y-values (the leans) and then the x-values (the years). Here is what you would write in two different cells

```
=SLOPE(B9:B21,A9:A21)
=INTERCEPT(B9:B21,A9:A21)
```

We get that the slope is 0.000932 and the intercept is 1.123339, so that the trendline is $y = 0.000932x + 1.123339$. We let $x = 2000$ and calculate that the predicted lean in the year 2000 is 2.9873 meters. This is consistent with the trend we see in the data, and it is fairly close to our estimate of 2.986 meters for the lean.

When to round?

When you do a calculation, when should you round? How many decimal places should you use in your answer? Do more decimal places mean more accuracy? These are good questions to think about and to explore using some examples. Generally speaking, you should not round until you are absolutely done with your calculation. If your calculation contains several steps, that means that you should carry through all of the decimals (here is where a calculator is helpful) and not round until you are done. Why not? Each time you round, you approximate the answer. As you then use your approximation for the next step of the calculation, and approximate that answer, the errors in the approximations can build and can make your final answer fairly far off from the true answer. Here is a rather exaggerated example.

Suppose you have been working at your job for 4.25 years (that is, 4 years and 3 months). The owner decides to reward employees by giving them a $1.00 per hour raise for every year that they have worked. If the Payroll Office rounds the number of years, then you would receive a $4.00 per hour raise instead of a $4.25 per hour raise. Now suppose that your supervisor decides to double the raise for some employees, including you. If the supervisor uses the $4.00 per hour

figure, you would receive an $8.00 per hour raise. If the supervisor started from the beginning and did the following calculation: $4.25 \times 1.00 \times 2 = 8.5$ and then rounded, then your raise would be $9.00. By rounding first, we have lost some information. If we wait until the end of the calculation to round, then we obtain a more accurate answer.

Once you have done your calculation, how many decimal places should you use? Does it make a difference if you write your answer out to 10 or more decimals? The simple answer is that you should look at the number of digits in your original numbers, and use those as a guide. Write your answer using at least as many digits, perhaps adding one more digit (or decimal place) but not more than that. For example, we would round $\frac{2}{3}$ to 0.67, rather than writing 0.666666666666667. To be more careful about this, you could pay attention to the number of *significant digits*. We won't go into the details here; as with many of the topics that we are covering, the best approach is to think about what you are doing. Use your common sense and write an answer that looks reasonable given the information you are using.

How good is the linear model?

Correlation versus Causation