

Name: \_\_\_\_\_

## Where Do (Linear) Functions Come From? Mac Excel 2008 Version

**This sheet includes both instruction sections (labeled with letters) and problem sections (labeled with numbers). Please work through the instructions and answer the questions in the problem sections. Turn in your answers to the *italicized prompts*, written in complete sentences. You may write your answers on these instructions or on a separate sheet if you desire.**

Your textbook includes many examples of linear models for different situations. That is, it gives a linear function which it claims gives the value of some dependent variable (for example, marijuana usage among high school seniors) from the value of an independent variable (for example, years since 1990). We call such a function a model because it usually won't produce the exact same values as we start with, but comes close enough that understanding the function tells us something about the actual situation. Once we have a function like this, we can use algebraic techniques to answer a variety of questions about the situation.

But where do such functions come from in the first place? Occasionally we may start with a table of values and be able to deduce the function by guess and check methods, such as used in the previous studio. But in general the data we have will not fit a simple linear function exactly. Instead, we use the technique of "linear regression" to find the line that "best fits" the data and take the function associated to this line. The textbook has a brief description of this method, and if you are interested you are welcome to stop by my office and I'll explain it in more detail. For now, we will just let the spreadsheet handle the work and consider when linear functions are (and aren't) reasonable.

Do be warned that the textbook's suggestion that linear models are reasonable only when you have "nearly constant" first differences is ridiculous. Linear models are accurate and important in many other situations as well. Note that all the data we will work with in this lab is real data, not just made up numbers.

### Baseball

A) Open the spreadsheet studio2datas09.xls which is posted on the class web site. It may be easiest to save it first and then open it in Excel rather than opening it directly in the browser. If you do have to open it within the browser, you will probably want to copy the data into a stand-alone Excel session, rather than working within the browser. The first tab on the spreadsheet has Runs Scored, Runs Allowed, and Wins for the 2008 season for all Major League Baseball teams (data available on mlb.com). We hypothesize the number of wins should be a function of the difference between the runs a team scores and the runs they allow their opponents to score. So we will try to create a function to fit this data.

Name: \_\_\_\_\_

- B) First we compute the required values. We already have the number of wins, so we just need to compute runs scored – runs allowed, which we will call run differential. Enter the formula  $=B2 - C2$  in cell D2. Click on cell D2 and then on the small square that appears in the lower left corner and drag the formula down to cell D31. We now have a list of the run differentials for all 30 teams.
- C) Click and drag to select D2..E31 and then graph the data (use a scatter plot without lines connecting the points). Check that the data seems to fall in a linear pattern, so it is reasonable that we should look for a linear function to model this data. Note that if you use a scatter plot with lines connecting the points, then the lines will crisscross all over your graph since your data is not in any particular order.
- D) Next we will ask the spreadsheet to determine which linear function (what values for slope and intercept) best fits the data. Control-click (right-click if you have a mouse with a right-click) on any one of the data points in the graph, then select “Add Trendline” from the options that appear. This brings up the Format Trendline dialog box. Since we are trying to fit a linear function, make sure the Type (3<sup>rd</sup> choice on the left menu) is set to Linear (which is the default value). Click on Options in the left menu and check the boxes “Display equation on chart” and “Display R-squared value on chart.” Then click OK. The spreadsheet will add a line (the trendline) to the graph, and will also print on the graph the equation of the line, and a value for  $R^2$ , which is a measure of how well the line fits the data. In this case we have  $R^2 = 0.8523$ , which is quite good. This means 85% of the variation (spread) in the data is explained by the model. Note that the spreadsheet often places the equation and  $R^2$  somewhere that it is difficult to read because the curve runs right over them, but you can click and drag the equation to a blank area of the graph to make it easier to read.
1. The equation you should have found is  $y = 0.1042x + 80.9$ . Is this a reasonable formula? In particular, *is 81 a reasonable value for the y-intercept? Explain why or why not.* (Hint: a major league season is 162 games long).

Name: \_\_\_\_\_

2. *Suppose a team gets a chance to add a player who will increase their runs scored by 1 over the course of the season. How many additional wins are they likely to have? How many additional runs will they need to add to increase their expected number of wins by 10?*

E) We can now compute the number of wins our function predicts and compare this to the actual number of wins. Go to cell F2 and enter the formula  $=0.1042*D2+80.9$ , then pull the formula down to cell F31. You use D2 since that is the run differential, which is the input (x) value for the model. Column F now contains the predicted number of wins. Next, go to cell G2 and enter the formula  $=E2 - F2$ , then pull the formula down to cell G31. Since column E has the actual number of wins and column F has the predicted number of wins, this means column G will have the residuals, the differences between the observed and predicted values.

Name: \_\_\_\_\_

3. Another question we can answer deals with whether teams are lucky or good. A team who has won many more games than their average output predicts might be considered lucky (and historically is likely not to do as well the next year). *Which was the luckiest team in the majors in 2008? Which was the unluckiest? Washington had the worst record in baseball in 2008. Were they unlucky or just plain bad?* You may find it easiest to identify the luckiest and unluckiest teams if you graph the residuals to pick out the largest and smallest (most negative) values.

Of course, there are people who spend a *lot* of time finding appropriate models for sporting events so they can answer questions about which player is more valuable and when different plays should and shouldn't be tried. In real life (to the extent this is real life) the models (functions) they use are often much more complicated than a simple linear function. For example, mlb.com reports expected winning percentage with the

formula  $Winning\ Percentage = \frac{Runs\ Scored^{1.82}}{Runs\ Scored^{1.82} + Runs\ Allowed^{1.82}}$ . Also note that in this

case we can't even compute first differences the way the text suggests, since our data is not taken at evenly spaced values. And if we computed successive slopes, we would find they varied significantly. Despite this, the linear model actually works quite well.

The general approach used here isn't just for baseball. In business and social science you will often have an output (wins) that you think is a function of an input (run differential) where you want to know how much a change in the input will affect the output, say because you want to know how valuable the input is to you. In business-speak, question 1 is asking for the marginal value of a run for example. Baseball happens to be a simple example where it is easy to get actual data.

Name: \_\_\_\_\_

### Population Growth

Another situation where it is easy to get data and build models is in population growth. If you click on the U.S. Pop tab, you will find the population of the U.S. as measured by the decennial census from 1790-2000.

- H) Select the range A2..B23 and click on the Chart Tab to create an XY (Scatter) plot of population vs. year. Control-click on one of the data points and choose Add Trendline from the options available. Select a linear trendline and on the Options tab, check the boxes to display the equation and the  $R^2$  value on the chart, and press OK to graph a linear model of the data. You should get  $y = 1E+06x - 2E+09$  and  $R^2 = 0.9201$
- I) Notice that the  $R^2$  is closer to 1 (which is better) than it was last time, even though the line doesn't look like a very good model (and of course it isn't a good model). There is no single statistic that tells you if you have a good model and there are situations (like this) when  $R^2$  can be misleading. Also note that the equation is difficult to read. The E+06 and E+09 is called scientific notation and is used to represent very large (or small) numbers. 2E+09 means 2,000,000,000 for example. Numbers like these are difficult to work with, and roundoff error becomes more troubling when you just get a single digit to work with. Therefore, we will rescale our numbers and try again.
- J) Enter the labels **Since 1790** and **Pop in Millions** in cells C1 and D1. Then enter the formula **=A2-1790** into cell C2 and copy it down to cell C23. Enter the formula **=B2/1000000** into cell D2. The denominator is 1 million (there are 6 zeros) and the value in D2 should be 3.929... Then copy the formula from cell D2 down to cell D23. Now repeat your steps from part H with the range C2..D23 to find the best linear model for population of the U.S. (in millions)  $y$  as a function of  $x$ , years since 1790.
4. *What is the equation of the best linear function for the rescaled data? How does it compare to the original model?* You should observe that the numbers are much easier to work with (and the spreadsheet can show more digits) when we have rescaled the values to run from 0 to 210 rather than 1790 to 2000 and from 3.9 to 281.4 instead of being in the millions. You should also note that the exercises in the book typically do similar rescalings.

Name: \_\_\_\_\_

5. Repeat the process of finding a linear function to model the population data, only this time just model the population during the 20<sup>th</sup> century, from 1900-1990. As in H above, click and drag to select the range C13..D23, graph it, and then add a trendline. *How well does a linear function work for this restricted range of values? Is it better or worse than for the whole range from 1790-2000?*

Notice that even though the data values are clearly curved, a linear function works pretty well over a limited range of the data. This is a general property (one can argue in fact that it is the fundamental idea underlying calculus). This is one reason why linear models are so popular; as long as the range you are dealing with is short enough, you are likely to get a pretty good approximation using a linear function, and linear functions are also particularly easy to work with. On the other hand, if you apply this model over a larger range, you may run into trouble if the actual data is curving away from the linear model, as we have already encountered in computing marijuana usage in 1978.

### **Bonus – Tornadoes**

The Tornadoes tab has data for the number of tornadoes observed in the U.S. between 1950 and 1994. Build a linear model for both the number of tornadoes as a function of the year. Note that while tornado data (even more so than most weather data) varies greatly from year to year, the linear model for tornadoes actually explains nearly 60% of the variance in the data and is highly statistically significant.

Name: \_\_\_\_\_

6. *What does the slope of the linear function that models the tornado data suggest about how the number of tornados varies over time? What could possibly explain this? (Hint: if a tree falls in the forest and no one is there to hear it, does it make a sound?)*

7. *How many tornadoes does the model predict will be reported during 2007? How much confidence do you have that this prediction will be accurate? Is there reason to think you could find a better fit than a linear function, or is it just that the data is naturally quite variable?*

While it is nice when we get a function that can be used for prediction, frequently (as in this case), the data is too variable to allow us to accurately predict what will happen for a particular input (in this case, a particular year). On the other hand, the model may still provide useful and accurate information about general trends (e.g. number of reported tornados has tended to increase over the years), even when the data is sufficiently variable that predictions for individual years are not valid.