

Lab 6: Exercises on simple linear regressionDue October 20th

Computing regression coefficients and standard error:

I know that most of you have figured out how to ask Excel to add trendlines to your plots, but we are going to take a step back and compute the regression ourselves. Its always a good idea to know what is going on inside the “Black Box” Excel formulations and we need to calculate some additional parameters that Excel doesn’t offer up for you.

Several people seemed confused in class about how to pick which variable is X and which is Y. In the statistics we have done so far X and Y were commutative, but in regression the distinction is important. We assume that X causes Y (for instance, we assume that El Nino causes changes in Temperature and not the other way around). We also assume that there is **no error** in X which allows us to minimize only the errors in Y when fitting a line. The “no error in X” clause isn’t usually fulfilled with real data unless the variable is time or some construct like that, but we do regressions anyway. It’s probably a good idea to keep this in mind as we assess results, especially if our X variable has large error associated with it. If you have a choice between multiple X variables for a simple regression all else being equal you should pick the X with the lowest error. In short, the nature of the variables tells you which will be X and which will be Y because there should be a casual relationship between them.

On to actually calculating a regression. The first thing to do is make a scatter plot of the data with X as plotted on the X axis and Y plotted on the Y axis. Next calculate the mean and standard deviation of both X and Y and the correlation between them. These are the tools you need to make a prediction and assess the success of it using the equations that follow.

As we saw in lecture we can calculate each value of Y-predicted by:

$$y_{pred} = \bar{y} + (corr_{xy}) \frac{SD_y}{SD_x} (x - \bar{x}) \quad (1)$$

This equation arises from minimizing the squared errors in Y between each point and a regression line. If you want to know more about how this is derived come chat with me and Ill do my best to show you.

To analyze the success of our regression we can calculate the Standard Error of our prediction as follows:

$$SE_{pred} = SD_y \sqrt{1 - corr_{xy}^2} \quad (2)$$

In addition it is useful to look at the slope of the regression line to assess the relationship between X and Y. If the relationship is weak, it is possible that the slope is not statistically different from zero. This means that we can not say if the relationship shown between the two variables did not arise by chance alone.

$$y = a + bx \quad (3)$$

is a generic equation for a line in which “a” is the intercept and ”b” is the slope of the line. We can re-arrange equation 1 to read:

$$y_{pred} = (\bar{y} - b\bar{x}) + bx, \text{ where } b = corr_{xy} * \frac{SD_y}{SD_x} \quad (4, 5)$$

Now we can see more clearly that “b” is the slope and (ybar-b*xbar) is the intercept. Notice that the intercept is set by requiring that the regression line pass through the mean of X and Y. Now that we have an equation to calculate the slope directly (5) we need to know how certain we are of that estimate. To do this we need to consider the strength of the correlation between the two variables as well as how many samples we have. The following equation gives the standard error of “b” (from equation 5)

$$SE_b = \frac{b}{corr_{xy}} \frac{\sqrt{1 - corr_{xy}^2}}{\sqrt{n - 2}} \quad (6)$$

where n is the number of samples you used in the regression. Now that we have the standard error we need to test and see if it is significantly different than zero. Recall from lab 4 that to test the significance of the difference between two numbers we converted that difference to units of standard deviation by dividing by the standard error and looked up our value on the z or t table. If we want to test our slope value “b” to see if it is different from zero so we simply need to divide it by its standard error and check our t value on the table for the confidence that we choose.

$$t_b = \frac{b}{SE_b} \quad (7)$$

It turns out that not only does this seem like an important constraint (the regression slope statistically different from zero) but the t value for “b” is equivalent to the t value for the correlation between X and Y, so it is THE important constraint for testing significance.

Excel’s built in stuff:

TREND(known y’s, known x’s, new x’s, TRUE) ***This is an array function*** that means you need to highlight an array for output and execute it with command-enter (control-shift-enter on a pc).

The last term refers to setting the intercept to zero, and the new x’s are the values you wish to plot your predicted line over. This gives you the same time series as equation 1 or 4, but you still need to calculate standard errors and parameters all on your own.

“Add a Trendline” can be found by right clicking (or control-click for Mac) on a datapoint on your scatterplot. In the options tab of the trendline window you can also ask Excel to print the equation of the regression line on the graph. Once again I highly recommend calculating this at least once without the Excel functions.

Project 1: Global Warming in Siberia

One of the regions of the globe that is thought to be showing the clearest signs of global warming is Siberia during the cold half of the year. To help provide evidence that this is indeed the case, compute the correlation between carbon dioxide concentration at Mauna Loa Observatory and the 5-month average temperature at Jenisejsk, Siberia for 1958 to 2003. (For the carbon dioxide, use the year corresponding to Jan-Feb-March, since more of the five months are in the later year than in the earlier year.) Based on the correlation, how much of the variance of Jenisejsk temperature is "explained" by the CO2 concentration?

Make a scatterplot of Jenisejsk winter temperature as a function of CO2 concentration. Develop the regression equation that predicts the temperature from CO2 concentration. Compute the standard error of prediction for the regression, the standard error of the slope and test to see if the slope is significantly different from zero at 95% confidence.

Project 2: El Nino’s effect on precipitation

Most of Florida is generally known to receive above normal winter rainfall when there is an El Nino and below normal rainfall when there is a La Nina. Using the data for SST in the tropical Pacific region called Nino3.4, examine this relationship with precipitation at two Florida stations that are fairly close to one another: Tampa and Tarpon Springs. Use the period of 1951-1996 since Tarpon Springs only goes up to 1996. Look at correlations, and then derive a regression equation and assess the standard error of estimate, the slope and test the slope against zero as done for problem 1 above. Assess in your mind whether ENSO is an important factor governing winter precipitation along the central Gulf (west) coast of Florida. Here are some hints:

-The precipitation data for Tampa are given in millimeters, while at Tarpon Springs it is in inches. This is

because the data were downloaded from differing data sets, as one station had more missing months in one data set and the other had more in the other data set. The conversion factor is $25.4 \text{ mm} = 1 \text{ inch}$. The units should be made equal before beginning if both stations are to be correlated with SST. (To correlate them with one another, or even find a regression equation between them, the units could remain different without affecting the result.)

-Correlations with single month precipitation are not be the best way to distinguish a climate relationship, since 1-month precipitation data are "noisy". It would be best to average several winter months together.

-Correlations with single station precipitation, even for several months averaged, are still degraded by the "noisiness" of a single rain gauge at a single location, such as that caused by the luck of getting some thunderstorms head-on and missing other large nearby storms. The second station is given to help balance this luck somewhat. It is suggested that the total (or average) precipitation for both stations be used in this analysis. First it should be verified that there is a fairly strong positive correlation between the precipitation between the two stations. (If there is not, especially in winter, there could be a "bug", such as one station not being what it was thought to be.)

-To help evaluate the strength of the relationship after the analyses are finished, determine if the correlation coefficient is significant at the 5% level. Determine what percentage of the rainfall variance is explained by the SST variations. Keep in mind also that correlations of 0.3 or higher are considered usable (better than nothing) if they are statistically significant, and that correlations of 0.5 or higher are considered useful.