David Youngberg
BSAD 210—Montgomery College

LECTURE 15: SIMPLE LINEAR REGRESSION I

I. From Correlation to Regression
   a. Recall last class when we discussed two basic types of correlation (positive and negative).
   b. While it's usually clear from a scatter plot if two variables are correlated (and in which direction they are correlated), we often want more than that. In a world of cost-benefit analysis, correlation is not enough. The magnitude is needed as well.
   c. This is why we do regressions: they let us know *how much* one variable influences another.
      i. These are the best *estimate*, an estimate because there will always be some things we cannot predict. At the very least, no sample is perfectly precise.
   d. It is thus important to remember that when you construct a regression, ***you are making a causal claim***. You are claiming one thing (x) causes another thing (y). If x increases, y will change. Y cannot change without x changing; y cannot change independently.
      i. This is why we call y a *dependent* variable (it depends on x), and x an *independent* variable (changes to it happen independent of the model).
      ii. We also refer to an independent variable as an explanatory variable because it explains the dependent variable.
   e. Be sure that your explanatory variable(s) logically matches with your dependent variable. Something that comes up a lot is adjusting for population—recall we discussed this at the end of Unit 1.

II. Basics
   a. Regression involves creating a "best fit" line. Let's begin with something familiar:

$$Y = mX + b$$

      i. Recall this equation from earlier math, where X is the independent variable, Y is the dependent variable, m is the slope, and b is the y-intercept.

ii. As we'll discuss later in the unit, regression can have more than one slope so when it comes to regression, we're going to change the notation and addition order a bit:

$$Y = \beta_0 + \beta_1 * X_1$$

iii. The y-intercept is now first, and we call it β_0 (beta zero).
iv. The slope is now second, and we call it β_1 (beta one).
    1. Note that β_1 is also called a "coefficient," or a value that's multiplied by a variable. Excel refers to all betas (including the intercept) as "coefficients."
v. Note that there's a subscript 1 next to X. Again, that's because we can have more than one independent variable (and thus more than one slope).
vi. Keep in mind that the mathematical rules for this equation are the same for y=mx+b.
    1. If you want to know *a value for Y*, you insert a value for X, multiply by β_1, and then add β_0.
    2. If you want to know *how much Y will change*, you multiply the change in X (usually 1) by its coefficient. Keeping in mind that "Δ" means "change in" and "k" refers to a particular β and X pair (e.g. β_1 and X_1):

$$\Delta Y = \beta_k * (\Delta X_k)$$

    3. **BURN THIS EQUATION INTO YOUR BRAIN.**

b. At this point, you might wonder where these actual beta values come from. They are the result of a series of equations that you don't have to know because nobody finds the best fit line by hand. Excel will do this for you and I'll show you how. But you need to know what defines "best fit," and for that we need to understand the residual.
c. The *residual* (ε) is the distance between what's predicted (according to the best fit line) and what's observed (according to the data).
    i. Every single observation has its own residual—the vertical distance between the best fit line and the observation.
    ii. You can think of the residual as the stuff the regression line can't explain.
    iii. Sometimes the residual is called the "error term" but that's a bit deceiving because it implies someone did something wrong.

Still, many resources refer to it as the error term so I mention that here to avoid future confusion.

d. Minimizing the sum of squared residuals is what defines the line of best fit. Hence, this technique is referred to as *least squares regression*.[1]
    i. In other words, square each observation's residual and add them together. You'll get something called Residual Sum of Squares, or RSS:

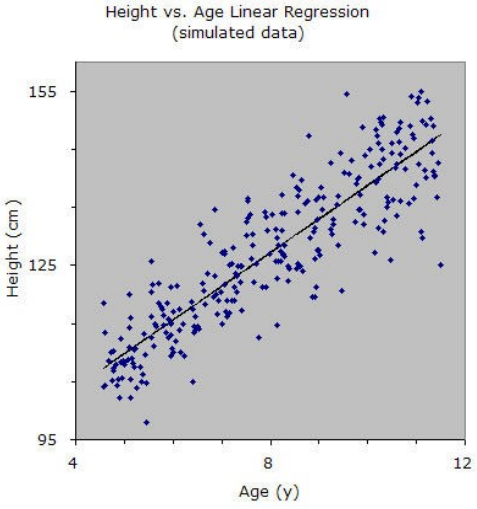
$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

        1. Where y_i is a particular observation;
        2. y_i-hat is the estimated value based on the regression line (we use the "hat" notation to distinguish between the estimated value and the actual value; and
        3. n is the sample size.
    ii. For the line of best fit, RSS is the lowest it can possibly be. The RSS for any other line would be higher.

e. Example—height
    i. Here, we're determining the line:

$$HEIGHT_i = \beta_0 + \beta_1 * AGE_i + \varepsilon_i$$

    ii. ε is the residual.
    iii. The subscript, i, refers to a particular observation. For example, i=1 is the first observation, i=2 is the second, and so on. Note the order of observations doesn't matter; it just for differentiating one observation from another.
        1. Note that i repeats for both variables. That's because for any observation, we know that observation's height and age.



[1] It is also called ordinary least squares, or OLS.

iv. Note this line is not a perfect fit. That's because other factors influence height besides age such as genetics, diet, and exercise. These unmeasured factors are captured in the residual—the stuff we can't explain.

f. β_1 is the slope of the line. It tells us *how much* age matters to height. Recall that this is the entire point of the regression.

g. Suppose the line is $HEIGHT_i = 80 + 6*AGE_i + \varepsilon_i$.

i. We can estimate that someone who is 8 years old is probably 80 + 6*8 = 128 cm tall.

ii. For every year someone ages, they get 6 cm taller.

III. Statistical significance

a. It's not enough to find the line that minimizes RSS and reference beta to determine magnitude because beta might not be statistically significant.

b. In regression, the null hypothesis is that beta equals zero (it is always a two-tailed test).

c. Don't worry about calculating anything like you had to do last unit—Excel will find the p-value for each beta (including the intercept, but no one really cares about the p-value for the intercept).

IV. Your First Regression

a. Open **Data Set 4**; you'll find data on Montgomery College professors from Rate My Professor.

i. Like before, this data includes every MC professor with at least 25 ratings, gathered in July 2014. We have data on their department, the number of ratings, the overall quality, and the level of difficulty.

ii. There are 211 observations (professors).

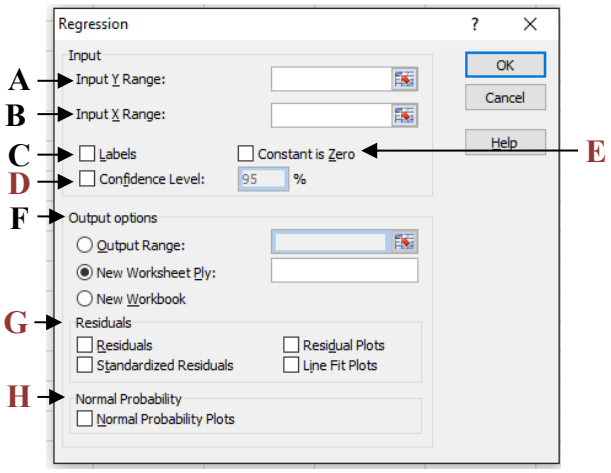
b. Suppose we want to tell a story that an easy professor will lead a student to rate that professor well on overall teaching. (Perhaps, because the professor is easy, students think they've learned a lot and thus rate the professor as quite skilled in pedagogy.)

i. Thus our causal claim: Difficulty causes Quality.[2]

$$QUALITY = \beta_0 + \beta_1*DIFFICULTY$$

[2] Astute observers will notice that I dropped the I subscript and I should technically have a "hat" symbol over QUALITY because there's no residual in the equation. I removed these notational bits to reduce clutter.

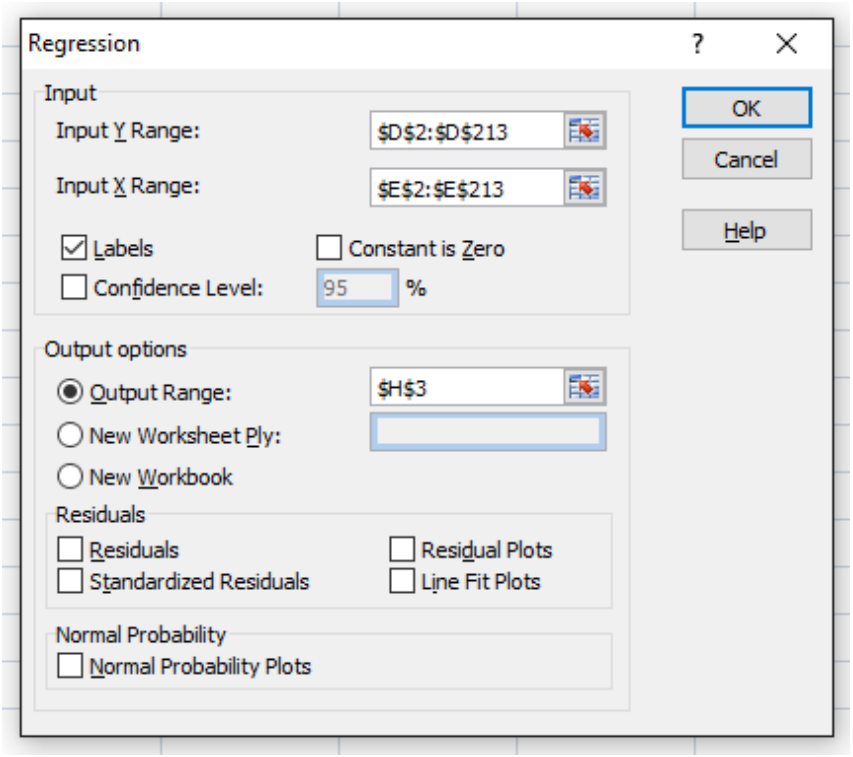
c. To run a regression in Excel, go to Data >>> Data Analysis >>> Regression. You'll get a window that looks like this:



A: Where the range for your dependent variable goes.
B: Where the range for your explanatory variable goes.
C: If you check this box, Excel will assume the first row of your data is the label for that column. It is useful to use this option, as we'll see soon.
D: Excel will output the confidence interval of your dependent variable's coefficient, defaulting to 95% confidence. Check this box to change the confidence level.
E: Check this box if you want to force the intercept (β_0) to be zero. You won't need this option for this class.
F: As before, this how you tell Excel where you want the results. I usually select the first option and select an out-of-the-way cell.
G: Excel can give you information on the residuals for each observation.
H: Used to analyze the data to see how it deviates from a normal distribution.

   i. A **red** letter means this option can be ignored for purposes of this class.

d. I filled the box as so:



e. And got this result:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.762834144
R Square	0.581915931
Adjusted R Square	0.579915528
Standard Error	0.566512809
Observations	211

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	93.36033044	93.36033	290.8995	1.90326E-41
Residual	209	67.0757833	0.320937		
Total	210	160.4361137			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	6.211157933	0.149062283	41.66821	3.3E-103	5.917299611	6.505016255	5.917299611	6.505016255
DIFFICULTY	-0.862492627	0.05056895	-17.0558	1.9E-41	-0.962183216	-0.762802038	-0.962183216	-0.762802038

V. Interpretation

a. For now, focus only on the end of the regression (we'll take about the rest of it later).

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	6.211157933	0.149062283	41.66821	3.3E-103	5.917299611	6.505016255	5.917299611	6.505016255
DIFFICULTY	-0.862492627	0.05056895	-17.0558	1.9E-41	-0.962183216	-0.762802038	-0.962183216	-0.762802038

b. For each variable in the regression, Excel will tell you the following:
   i. *Coefficient*—this is the beta-value for the variable; the slope.
   ii. *Standard Error*—this is the dispersion of the coefficients. If you draw multiple unbiased samples, this gives an idea of how much the coefficients would change.
   iii. *t-statistic*—ratio of the estimated coefficient to the standard error of the estimated coefficient (coefficient divided by error).
   iv. *p-value*—as discussed previously.
   v. *Confidence interval*—describes the range that the true value of the parameter could fall with a certain level of certainty (usually 95%). It outputs this result twice, the second one for whatever you customized Excel to do (e.g. 97% rather than 95%).

c. The intercept refers to β_0; the "coefficient" for the intercept is 6.2112. Our estimated line is thus:

$$QUALITY = 6.2112 - 0.8625*DIFFICULTY$$

    i. Note as well this result is statistically significant. The t-stat is huge and p is functionally zero.
d. Increasing DIFFICULTY by one point decreases QUALITY by 0.8625 points.
e. A professor with a DIFFCULTY of 3 is expected to have a QUALITY of about 3.624.
    i. If the professor is actually above or below that predicted value, you can infer that there is something special (good or bad) about his or her teaching.