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 level of influence 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 call the independent variable(s) the explanatory variable(s) because it explains what the dependent variable is.
   e. When choosing variables, be sure that your explanatory variable(s) logically matches with your dependent variable. Some common mistakes:
      i. One variable adjusts for population and other doesn't (e.g. population density and total number of crimes).
      ii. Adjusting for population for a variable adjust when that doesn't make sense either because:
         1. It already adjusts for population (e.g. percent of people in a country who are teenagers or average life expectancy)
         2. Population isn't relevant (e.g. average rainfall or if a state voted for a Democrat or Republican in the last presidential election).
      iii. The same can be said for other adjustments, such as land area.

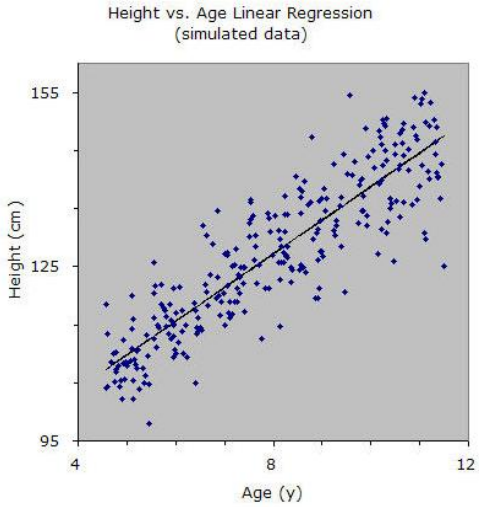
II. Basics

a. *Least Squares Regression*—line which minimizes the sum of squared deviations between the constructed line and the actual data points.

i. It is also referred to as Ordinary Least Squares (OLS).

ii. Here, we're determining the line:

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



The ε is the *residual*, the distance between what's predicted and what's observed. Sometimes it's called the *error term* but that's a bit deceiving. It's not suggesting anyone did anything wrong. Still, many sources (including your book) refer to it as error so mention that here to avoid future confusion.

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. Repeating i means those values are for the same observation: the ith observation.

iv. This line is determined by minimizing the sum of the squared *vertical* distance between the line and a data point. This is built to minimize this value (Residual Sum of Squares):

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

1. Where y_i-hat is the estimated value based on the regression line;
2. y_i is a particular observation; and
3. n is the sample size.

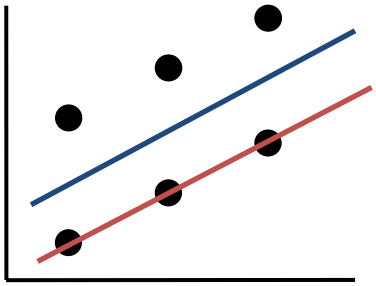
v. Note this line is not a perfect fit. That's because other factors influence height besides age such as genetics, diet, and exercise. These unmeasured factors are captured in ε_i. The i subscript means that ε changes from observation to observation. Its placement is for technical reasons, creating the connection between how much the y value should be and how much the y value is.

b. β_1 is the slope of the line. It tells us *how much* age matters to height. Suppose the line is $HEIGHT_i = 80 + 5.6\ AGE_i + \varepsilon_i$.
   i. We can estimate that someone who is 8 years old is probably 80 + 5.6(8) = 124.8 cm tall.
   ii. For every year someone ages, they get 5.6 cm taller.

III. Why Square the Difference?

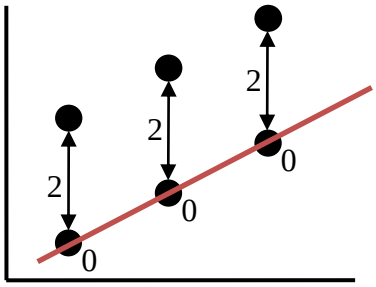
a. The best fit line relies on squaring the vertical difference between what the line predicts a value should be and what it actually is. Why not just take the absolute value.

b. Consider this hypothetical data set and two different regression lines:

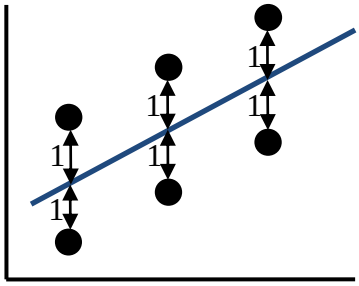


c. At halfway between three pairs of data, the blue line is clearly the better fit. But if you use absolute value, there's no difference in quality. Suppose each vertical pair is a distance of 2 apart with the blue line halfway between each pair.

d. For the red line, there's three vertical distances of 2 and three vertical distances of 0.
   i. Absolute value method: 2+2+2+0+0+0 = 6
   ii. Squared method: $2^2+2^2+2^2+0^2+0^2+0^2 = 12$



e. For the blue line, there's six vertical distances of 1 each.
   i. Absolute value method: 1+1+1+1+1+1 = 6
   ii. Squared method: $1^2+1^2+1^2+1^2+1^2+1^2 = 6$



f. Using absolute value, there's no difference between the two lines. But under the correct method—summing the squares—the blue line is better.

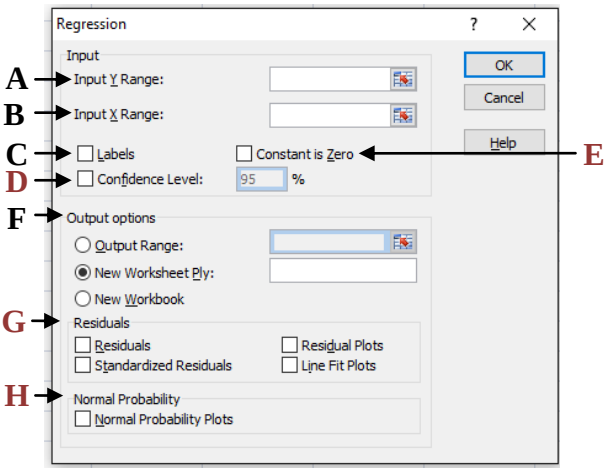
IV. Basic Regression

a. Open Data Set 5; 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, the level of difficulty, and if users rate that professor "hot" or not.
   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.

$$QUALITY_i = \beta_0 + \beta_1 * DIFFICULTY_i + \varepsilon_i$$

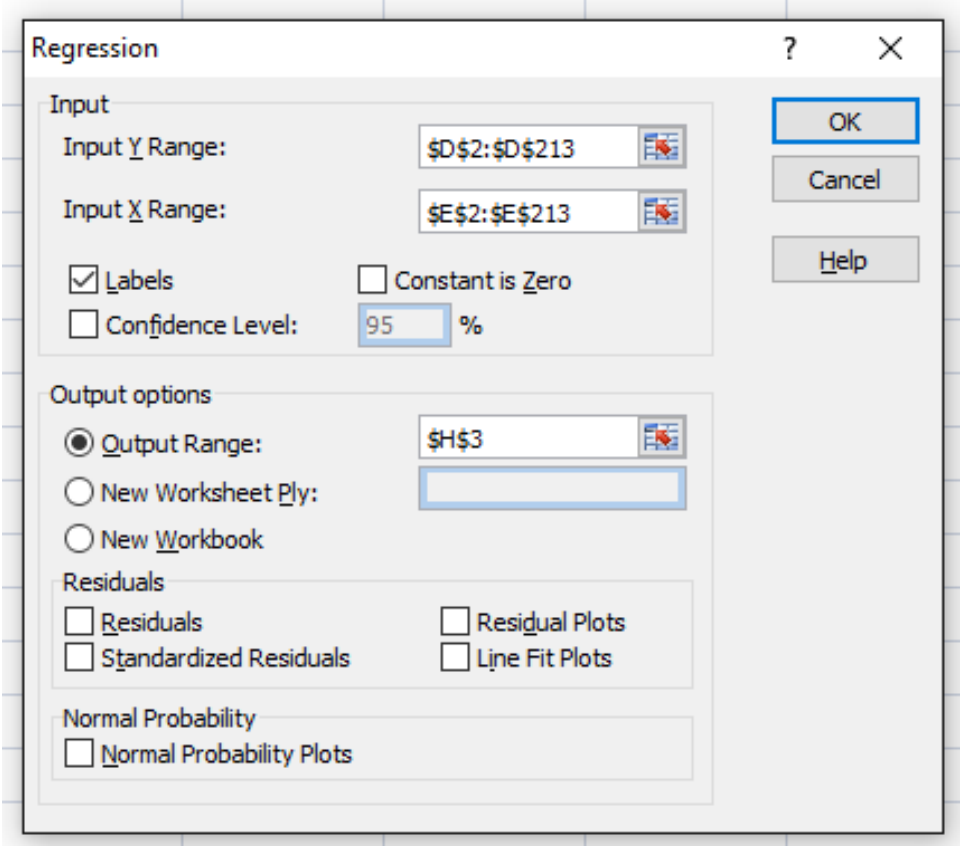
c. To run a regression in Excel, go to Data >>> Data Analysis >>> Regression. You'll get a window that looks like this:



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

d. I filled the box as so:

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.



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 (and it's possible to have many, which we will discuss later), 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*—tells you the threshold of significance you achieve for a particular t-statistic. (Remember critical t values changes based on degrees of freedom.) If the p-value is below 0.05, it's significant to the 5% (95% confidence) level. If below 0.01, it's significant to the 1% level, etc. It's basically the α.
   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 is β_0; it's not really a variable and the t-stat other information doesn't matter too much. But the coefficient does. That number (6.2112) is β_0. Our estimated line is thus:

$$QUALITY_i = 6.2112 - 0.8625*DIFFICULTY_i + \varepsilon_i$$

   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.