David Youngberg
BSAD 210—Montgomery College

LECTURE 28: SIMPLE LINEAR REGRESSION I

I. From Correlation to Regression

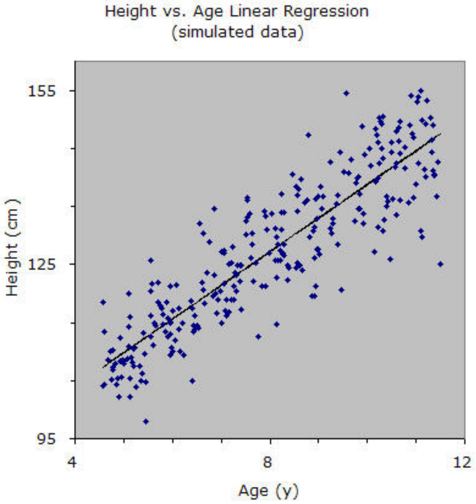
- a. Recall earlier in the semester when we discussed two basic types of correlation (positive and negative).
- b. While it's usually clear from a scatter plot if two variables are correlated (and in which direction they are correlated), we often want more than that. In a world of cost-benefit analysis, 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 the best *estimate*, an estimate because there will always be some things we cannot predict.
- 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. Best Linear Unbiased Estimator

- a. *Least Squares Regression*—line which minimizes the sum of squared deviations between the constructed line and the actual data points.
  - i. This is also known as a line of best fit, or the Best Linear Unbiased Estimator (BLUE). 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 I will use that here to avoid confusion.

iii. 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 (Sum of Squares Error):

$$SSE = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2$$

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

iv. Note this line is not a perfect fit. That's because other factors influence height besides age. What others factors could play a role in height?

b. β_1 is the slope of the line. It tells us *how much* age matters to height. Suppose the line is $\text{HEIGHT}_i = 80 + 5.6\ \text{AGE}_i + \varepsilon_i$.

i. We can estimate that someone who is 8 years old is probably 80 + 5.6(8) = 124.8 cm tall. For every year someone ages, they get 5.6 cm taller.

III. Excel

a. The calculation for the BLUE line is really time-consuming and practically impossible for humans to do if you have a lot of observations. So we turn to computers.

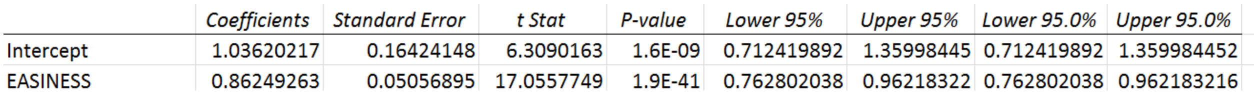
b. Microsoft Excel can do this well so let's focus on understanding Excel's output for a regression. Let's try this out on something we discussed earlier: professor ratings on Rate My Professor.

c. Suppose we want to tell a story that an easy professor will led 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: Easiness causes Quality.

$$\text{QUALITY}_i = \beta_0 + \beta_1 * \text{EASINESS}_i + \varepsilon_i$$

ii. When I run this regression for my 211 observations, Excel outputs the following results (it outputs more than this, but let's start with this).



	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	1.03620217	0.16424148	6.3090163	1.6E-09	0.712419892	1.35998445	0.712419892	1.359984452
EASINESS	0.86249263	0.05056895	17.0557749	1.9E-41	0.762802038	0.96218322	0.762802038	0.962183216

d. 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; I have no idea why.

e. 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—1.03620217—is β_0. Our estimated line is thus:

$$QUALITY_i = 1.036 + 0.862*EASINESS_i + \varepsilon_i$$

   i. Note as well this result is statistically significant. The t-stat is huge and p is functionally zero.

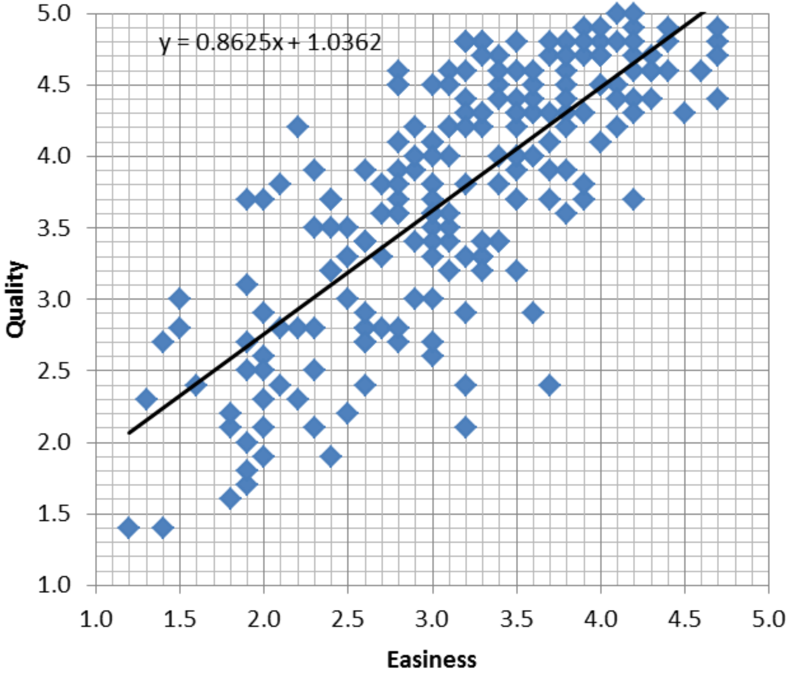
f. Increasing EASINESS by one point increases QUALITY by 0.862.

g. A professor with an EASINESS of 3 has a quality of about 3.622.

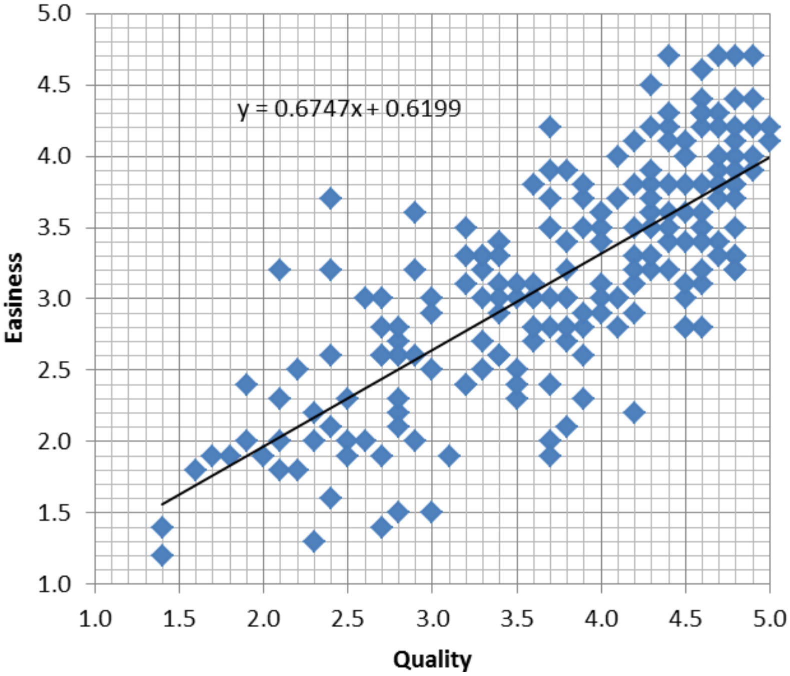
   i. If the professor is actually above or below that predicted value, you can infer that there is something special (good or bad) about his or her teaching.

h. Causation matters!

i. Here is the graph with EASINESS causing QUALITY:



ii. Here is QUALITY causing EASINESS:



iii. Because the regression is minimizing a vertical distance of a completely different variable, we get a totally different line.

iv. The professor with a QUALITY of 2.4 and EASINESS of 1.6 is right on the predicted line in the first graph. But reversing the causation moves that professor below the line.