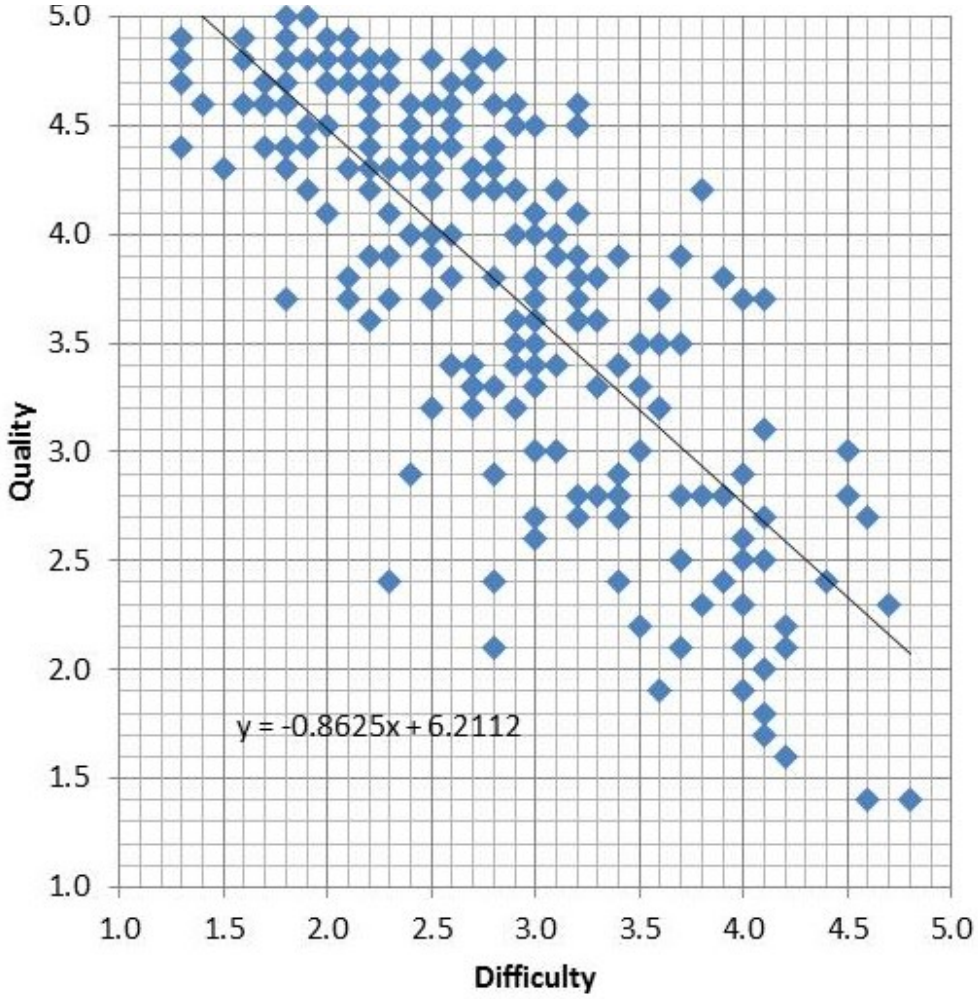
David Youngberg
BSAD 210—Montgomery College

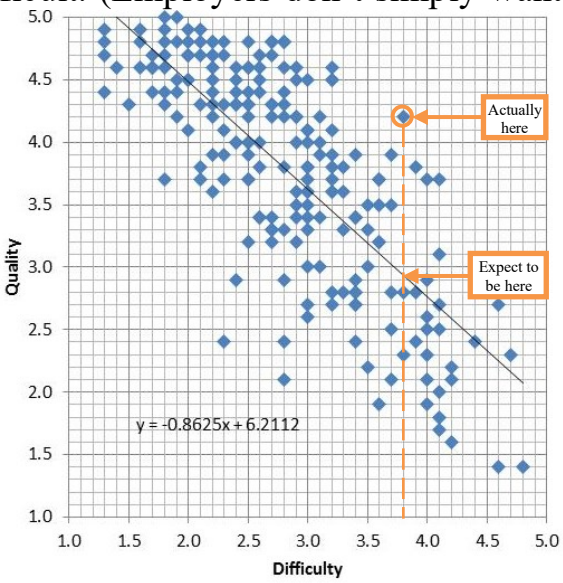
LECTURE 16: SIMPLE LINEAR REGRESSION II

I. On Causation and Evaluation.
   a. Let's revisit our regression from last class.
   b. Here is the graph with DIFFICULTY causing QUALITY:

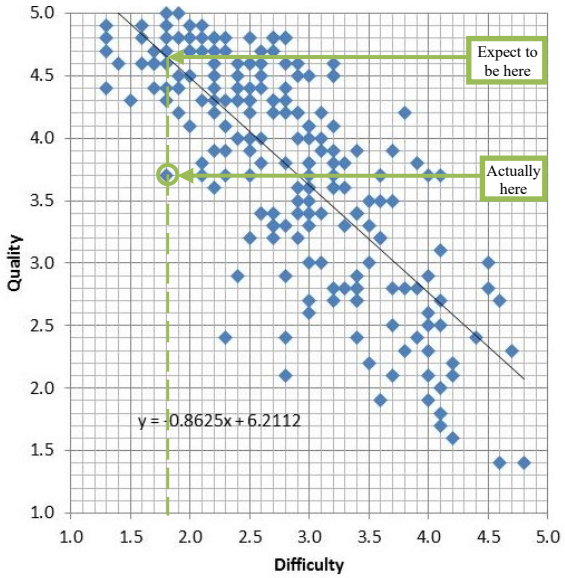


   i. I made this with the Add Trendline… option found after you right-click the data on a scatter plot. By opting to Display Equation, it will show you the line's equation.
   ii. Note (a) it's in a slightly different format (y=mx+b) and (b) it doesn't give you statistical significance. It's better to use Data Analysis to run the regression but this option is useful for visualization purposes.

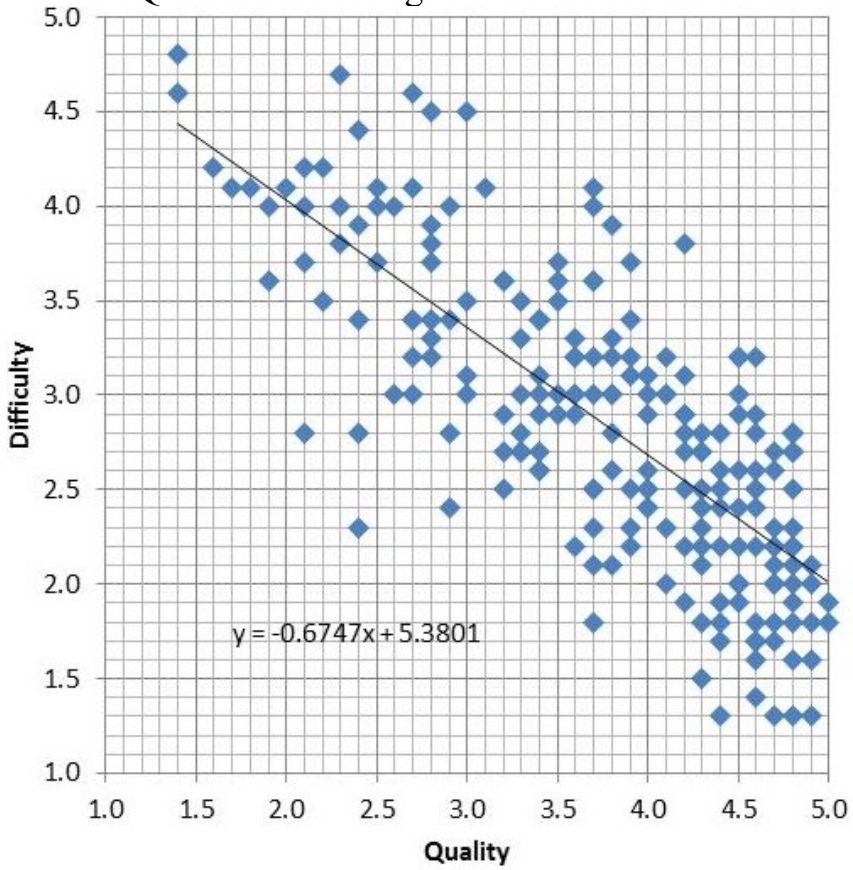
c. Interestingly, we could use this to evaluate professors. A good professor gets high ratings while being difficult. (Employers don't simply want "A" students. They want "A" students who had to work really hard for the grade.)



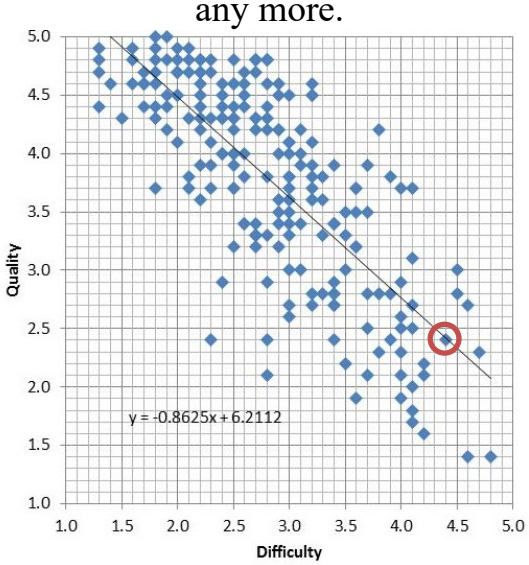
   i. Professors above the line having a higher quality than you'd expect given their difficulty rating.
   ii. Professors below the line have a lower quality than you'd expect given their difficulty rating.
   iii. The professor highlighted with the orange circle has a good quality rating (4.2) but when you consider how hard s/he is (difficulty is 3.8), it's much more impressive. You'd expect the quality rating to be only about 2.9 with a course that difficult. Despite being hard, the students like the professor. That difference is the error term, ε, mentioned earlier.
   iv. The professor highlighted with the green circle seems to be pretty good (quality of 3.7) but with a difficulty of 1.8, you'd expect a rating of about 4.7. The quality rating is quite low for how difficulty the professor is. Again, that difference is the error term, ε, mentioned earlier.

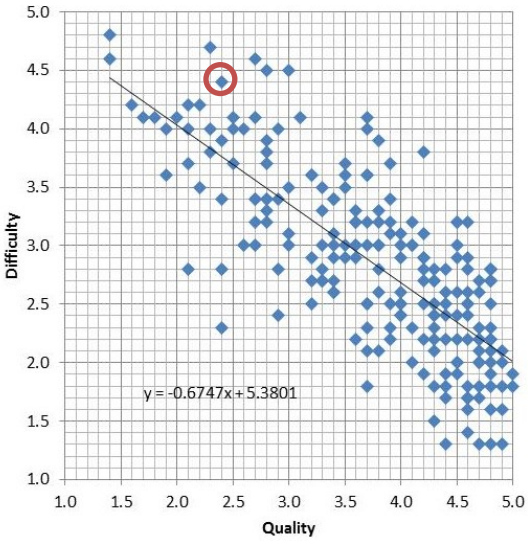


d. Consider QUALITY causing DIFFICULTY:



i. The professor with a QUALITY of 2.4 and DIFFICULTY of 4.4 is right on the predicted line in the first graph. But reversing the causation moves that professor above the line. S/he is not average any more.





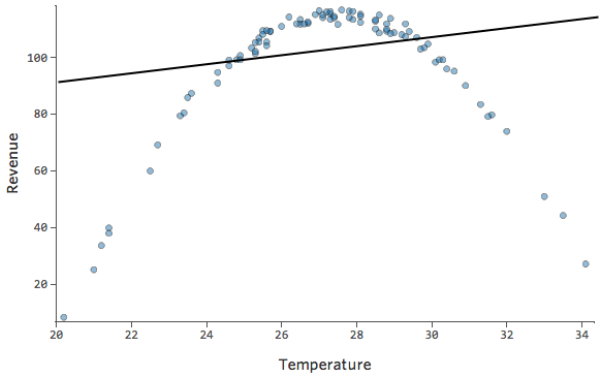
e. Why does it change? Linear regressions minimize the summed and squared *vertical* distance. What's set as the y variable and what's set as the x variable determines the line. Swapping the two variables results in a fundamentally different line.

II. Assumptions of a Linear Regression

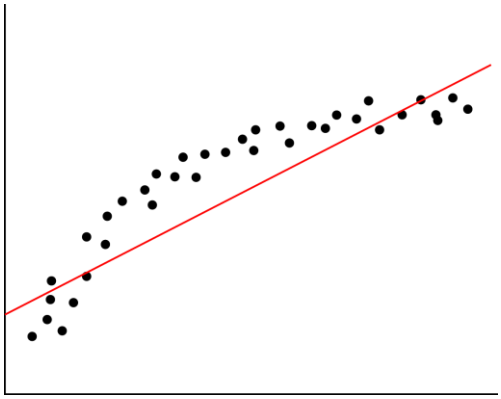
a. For purposes of this class, we will assume these assumptions hold for the regressions we run but you should be aware that they may not actually be true for a particular data set.

b. **The regression is linear**. In other words as the independent variable increases, the other dependent variable increases or decreases. The dependent variable:

i. Sometimes increases and sometimes decreases. Like this:



ii. Increases or decreases at a variable rate. Like this:



iii. *In other words, it should make sense that slope is constant.*

c. **The residuals follow a normal distribution**.

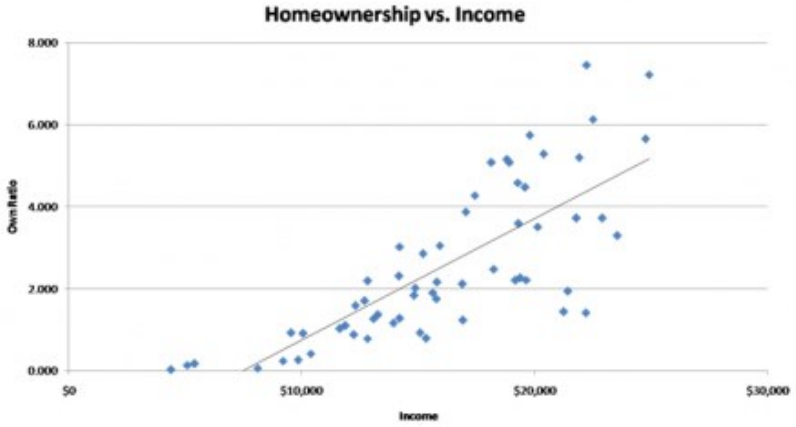
i. If you plot all the residuals in a histogram, you should get something that looks like a normal distribution.

ii. Note the connection between this and the CLT: the Central Limit Theorem notes that, by chance, many residuals will be close to zero and a few will be very low (large negative) or very high (large positive).

d. **The residuals are independent of one another (no autocorrelation)**. If you plot the residuals in data order, it should look like a random scattering around zero. If there is a pattern, you have "autocorrelation."
    i. Broadly speaking, autocorrelation is when a variable is correlated with itself.[1]
    ii. Autocorrelation is a particular concern in a time series, when the order of the data matters. But it doesn't have to occur in a time series—in a cross-sectional data, order can still matter.
    iii. Here's a nice example of autocorrelation in cross-sectional data.
e. **There is no multicollinearity.** We'll talk about this later.
f. **There is homoscedasticity**; this requires some explanation.

III. Homoscedasticity

a. *Homoscedasticity* is that the variance (or the deviation) from the regression line is the same, regardless the value of the independent variable(s).
b. When we lack homoscedasticity we have heteroscedasticity, or the variance is not the same for all values of our independent variable.
    i. Heteroscedasticity can show up in different ways. Here we see how variance increases as income increases. But if variance decreased, or increased and then decreased, or decreased and then increased, etc. we'd still have a problem.



c. The simplest way to detect heteroscedasticity is to make a scatter plot and add a regression line. This visualization test is intuitive (but not precise).

[1] Autocorrelation occurs outside of regressions. As the linked article shows, the famous Dunning-Kruger effect (that the less-skilled tend to be more confident in their ability) is a myth—its statistical evidence is just autocorrelation.