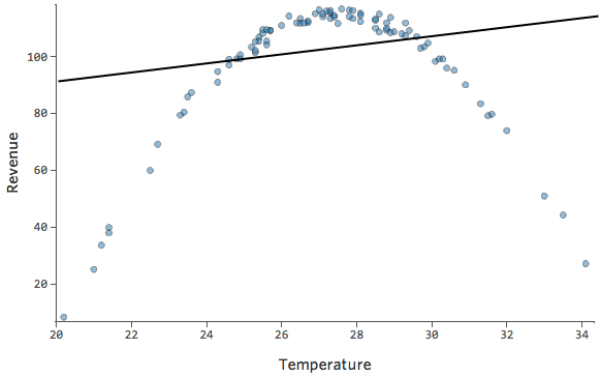
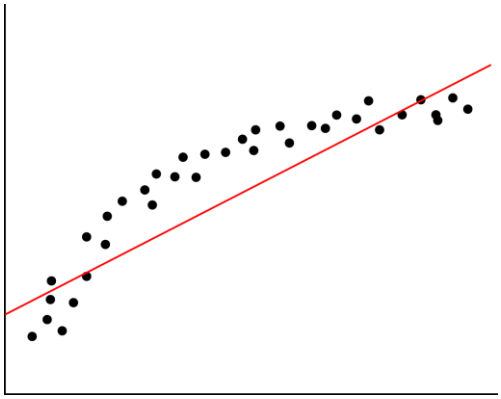
David Youngberg
BSAD 210—Montgomery College

# LECTURE 23: SIMPLE LINEAR REGRESSIONS III

I. 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 dataset.
   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:

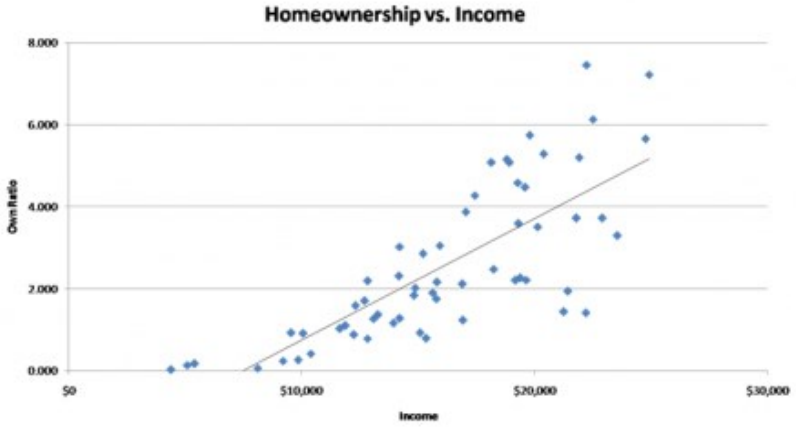


      iii. *In other words, it should make sense that slope is constant.*
   c. **The residuals follow a normal distribution**.
      i. If you plot all the residuals in a histogram, you should get something that looks like a normal distribution.
      ii. Note the connection between this and the CLT: the 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 from one another**. There should be statistical independence of errors. If there's not, it's called autocorrelation.
    i. In other words, you shouldn't be able to predict the residual of the one observation with the residual of the previous observation.
    ii. Example: Stock prices often have this issue; the price now is not independent from the previous price. In fact, time series often have autocorrelation issues.
e. **There is homoscedasticity**; this requires some explanation.

II. 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. Why should we care? Heteroscedasticity biases our error which means our t-statistic is higher (or lower) than it should be.
    i. In practice, it is not much higher so if you're significant to the 1% level, you're probably fine.
    ii. But if your values are barely significant (close to 5%), then you have a problem. If you'd adjust for heteroscedasticity, your significant result might cease being so!
d. 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).