David Youngberg
BSAD 210—Montgomery College

LECTURE 25: MULTIVARIATE REGRESSIONS II

I. Dummy variables
   a. A common control is a *dummy variable*—a variable that's either zero (for "no") or one (for "yes").
      i. These variables are binomial: gender (male or female), employment (working or not working), immigration status (legal or illegal).

Company	West?	Midwest?	Northeast?
Red Sun	1	0	0
Yellow Sun	0	0	0
Blue Sun	0	0	1
Green Sun	1	0	0
Orange Sun	0	1	0
Purple Sun	0	0	0
Black Sun	0	0	1
White Sun	0	0	0
Grey Sun	0	1	0

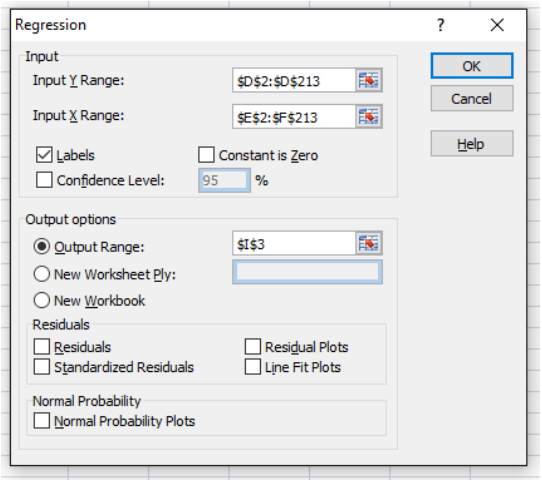
      ii. You can use multiple dummies for a variable with a few categories (White? Black? Asian? Hispanic?). For example, here's hypothetical data where each observation is a U.S. company. The dummy variable is the region of country where the company's headquarters are.
      iii. You typically want to have a number of dummies equal to one minus the number of categories. If the dummy is "Female?" then you know 1=F and 0=M. Adding "Male?" is redundant. Note on the table of the hypothetical firms, there is no dummy variable for the South. That's because if a U.S. firm doesn't have their HQ in any of the other regions, it must have it in the South. That's where Yellow Sun, Purple Sun, and White Sun have their HQs.
      iv. The only time you don't want to have one fewer dummy variables than categories is when the categories aren't mutually exclusive. A firm can't have their HQ in two different regions. But a student can have more than one major, a person can identify as multiple races, a rug can have several different colors in it, etc.
   b. You interpret the variable as you would when there's a single variable: examine the coefficient. Again, you're holding the other variables constant.

II. More Output from Excel
   a. Let's go back to the Rate My Professor ratings data in Data Set 5. Recall we explored how a professor's easiness can predict his or her quality.
   b. Rate My Professor also asks students to indicate if the professor is attractive or not (hot or not). I've set this up as a dummy variable: 1

means the professor is rated as hot and 0 means the professor is rated as not hot.

c. If a professor becomes "hot," is it possible that results in a better quality? We need a plausible causation story (remember: regressions are all about causation). Perhaps students pay more attention and are more likely to attend class if the professor is attractive. That means students learn more and the class is more enjoyable, encouraging students to think the professor is a better educator.

d. To run a regression with multiple explanatory variables, you just highlight multiple columns for the X range rather than just one column. I so below, highlighting the E and F columns:



i. This is why all your dependent variables have to be next to each other: so you can create a continuous box.

e. Here is the full output:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.806814419
R Square	0.650949506
Adjusted R Square	0.647593251
Standard Error	0.518875933
Observations	211

These are the items we will focus on. The rest we've already discussed or don't matter for our purposes. Well, expect observations but it's obvious what that is.

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	104.435809	52.2179	193.9512	2.88655E-48
Residual	208	56.00030473	0.269232		
Total	210	160.4361137			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	5.756302906	0.153848027	37.41551	2.51E-94	5.453001574	6.059604238	5.453001574	6.059604238
DIFFICULTY	-0.754138639	0.049301502	-15.2965	6.84E-36	-0.85133333	-0.656943949	-0.85133333	-0.656943949
Hot?	0.552312714	0.086112722	6.413834	9.39E-10	0.382547109	0.722078319	0.382547109	0.722078319

f. If a professor simply becomes "Hot" (going from a 0 to a 1), his or her rating increases by about 0.55, holding their DIFFICULTY rating constant. Note this is the most a professor could get out of this variable because there're only two values this variable can be.

III. Interpretation

a. *Explained (Regression) Sum of Squares (ESS)*—the squared vertical difference between the average and the predicted value of the dependent variable. This difference is taken for each observation and then added together.
b. *Residual Sum of Squares (RSS)*—The squared vertical difference between the observed value and the predicted value. This difference is taken for each observation and then added together.
c. *Total Sum of Squares (TSS)*—ESS + RSS
d. R^2—ESS/TSS, or the percent of deviation that our regression explains. There is no threshold for a "good" R^2.
    i. We are explaining 65% of the distance between a rating's observed value and the average rating.
    ii. R^2 is sometimes also called the "coefficient of determination."
e. *Adjusted R^2*—The R^2 value adjusted for the number of explanatory variables.
    i. A weakness of R^2 is that it adding additional explanatory variables causes it to increase, regardless of the quality of explanatory variables. This is a problem because having many explanations for something is the same as having few.
    ii. Adjusted R^2 penalizes the researcher for adding explanations, especially if it's large relative to the number of observations. The equation is:

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

    Where n is the number of observations and k is the number of explanatory variables, excluding the intercept.
f. *F*—The ratio between the explained and unexplained variance. Like R^2, it's used for evaluating the model as a whole. And like the t distribution, the F distribution is a family of distributions. Significance level depends on degrees of freedom.
    i. Higher values of F indicate a model with more explanatory power. Because the shape of the F distribution is known (its exact shape changes based on the number of observations and number

of explanatory variables), it is possible to determine critical values.

g. *Significance F*—this is the p-value for the F stat and uses the same criteria. If the value is very small, the model is quite good.

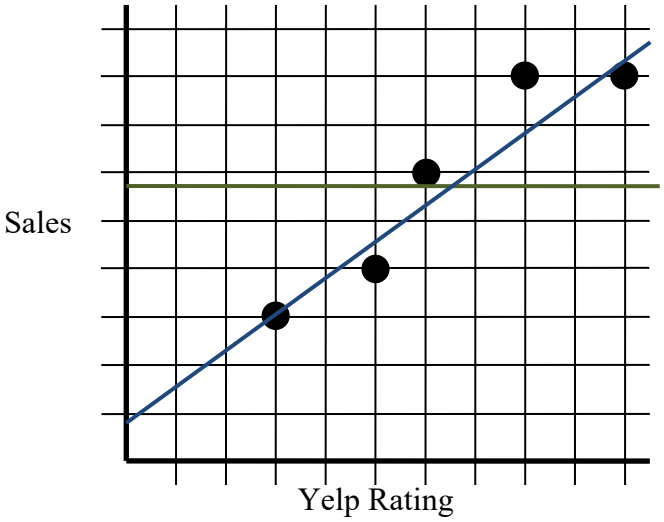
IV. Thinking About Regressions

a. Suppose you have sales data on various Chinese restaurants. If you pick a restaurant at random, what do you suppose that restaurant's sales are?

b. Your best guess would be average sales. Obviously, your guess probably won't be right but based on how little information you have, there's no better guess.

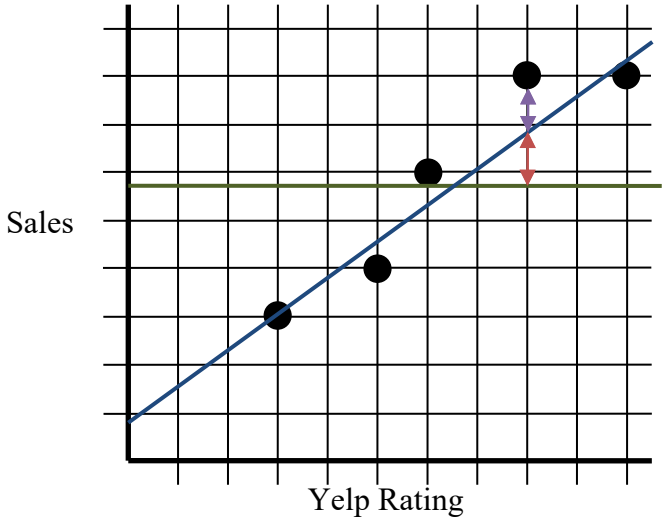
c. Now suppose you know that restaurant you chose has 4 out of 5 stars on Yelp, the popular review site. How do you adjust your expected sales? It should go up, right?

d. Regressions are about how you can explain why an observation's value is different from the average (that's why causation is so important).



e. The green line is the average sales. The blue line is the regression line. Note that we get a much better estimation of sales if we employ something we know that has predictive power (Yelp ratings) than if we just guessed based on the average.

   i. Indeed, of the five observations, four give us a much better estimate of sales than the average (one is spot on!). Only one observation—the middle one—does using the line rather than the average worsen the guess. And it's not that much worse.



f. The red line is that observation's contribution to ESS; it's the part of the deviation the regression line can explain.
g. The purple line is that observation's contribution to RSS; it's the part of the deviation the regression line can't explain.
h. I write "contribution' in each of these cases because ESS and RSS are the ***sum*** of squares. It's the result (after squaring it) from all the observations.

V.