\*David Youngberg

Acct/Busa/Econ 222—Bethany College

**Lecture 08: Multivariate Analysis**

1. What determines attraction?
2. Control
	1. A multiple regression has more than one independent variable.

$$Y\_{i}=β\_{1}X\_{1i}+β\_{2}X\_{2i}+β\_{3}X\_{3i}+…+ε\_{i}$$

* 1. Sometimes you want to do a multiple regression because you think multiple variables matter.
		1. Example: It makes sense that life expectancy depends on both diet and exercise.
	2. The nice thing about this is that this technique lets you control for other factors.
		1. Example: Using “distance from city center” doesn’t appear to influence the price of houses. But that’s because homes that are far away tend to be larger. Include square footage—allowing you to ask “How does a home change price if we keep it the same size and move it closer?”—and you’ll see a significant result.
	3. A common control is a *dummy variable*—a variable that’s either zero (for “no”) or one (for “yes”).
		1. These variables are binomial: gender (male or female), employment (working or not working), immigration status (legal or illegal).
		2. You can use multiple dummies for a variable with a few categories (White? Black? Asian? Hispanic?)
		3. You always 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.
	4. You interpret the variable as you would when there’s a single variable: examine the coefficient. But this time, you’re holding other variables constant.
1. Output from Excel
	1. *Explained Sum of Squares (ESS)*—the squared difference between the average and the predicted value of the dependent variable. This difference is taken for each observation and then added together. Excel calls this the “regression sum of squares.”
	2. *Residual Sum of Squares (RSS)*—the squared difference between the observed and predicted value of the dependent variable. This difference is taken for each observation and then added together.
	3. *Total Sum of Squares (TSS)*—ESS + RSS
	4. *R2*—ESS/TSS, or the percent of deviation that our regression explains. There is no threshold for a “good” R2.
		1. This is sometimes also called the “coefficient of determination.”
	5. *Adjusted R2*—The R2 value adjusted for the number of explanatory variables.
		1. A weakness of R2 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.
		2. Adjusted R2 penalizes the researcher for adding explanations, especially if it’s large relative to the number of observations. The equation is:

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

Where n is the number of observations and k is the number of explanatory variables, including the intercept.

* 1. *F-statistic*—The ratio between the explained and unexplained variance. This is the same F-value from last class.
		1. 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.
		2. The *p-value* does this work for you: if it is smaller than, say, 0.05, the model is significant with 95% confidence. If it is smaller than 0.01, the model is significant with 99% confidence.
	2. *p-value*—each variable will give you a p-value, which is a summary of significance (it will also report the t-value but significance depends on that and on the degrees of freedom). The smaller the value, the better the threshold you achieve (if less than 0.05, it’s significant to the 5% level; less than 0.01 and it’s significant to the 1% level).
1. Multicollinearity
	1. If two or more of our explanatory variables are highly correlated, we have multicollinearity.
		1. It doesn’t require perfect correlation—if there’s perfect correlation than the program will drop one of the variables.
		2. Example: The dummy variable trap, when you have a number of dummies equal to the number of categories. Having both “Male?” and “Female?” will lead to perfect correlation.
	2. Multicollinearity is a problem because the regression will try to get two variables to do the same job. It can easily render both variables insignificant by producing large standard errors.
		1. On the right, there is a portion of variation in Y that can be attributed to either X.
	3. You can notice multicollinearity if:
		1. The F-test suggests your model as a whole is strong but none of your variables are significant.
		2. The regression coefficients change a lot when a variable is added or deleted.
	4. The easiest way to correct for multicollinearity is to drop variables.
2. Trend variables
	1. Any two variables with similar trends will appear highly correlated. But this doesn’t mean one causes the other (correlation doesn’t mean causation!). This often occurs in time-series data.
		1. Example: The stock price of Exxon doesn’t influence the stock price of IBM.
	2. So how can you test if two variables are actually connected? Add an additional explanatory variable called a *trend variable*—one that starts at 1 for the first observation, then 2, then 3, etc.
		1. “First observation” means the observation in the first time series; again, this kind of problem is most common in time series.

**Lab Section**

1. Chapter 14
2. Multicollinearity
	1. There are sophisticated tests for multicollinearity but for our purposes, a simple correlation check is fine.
	2. Pull up Data Analysis >> Correlation
	3. Select all independent variables, including the label, for the input. Make sure the correct “Grouped By” option is selected and you’ve told Excel there are labels in the first row.
	4. Once you’ve selected an output cell, touch “OK.”
	5. The output is a correlation table with values on a scale of 0 to 1. Note that the variable will have a correlation coefficient of 1 with itself.
	6. A lower value means less correlation. For this class, we will use an admittedly conservative standard of |0.5| (the absolute value of 0.5). If it is more than that absolute value for any two variables, you have multicollinearity.
	7. If you have multicollinearity, the easiest answer is to drop as few variables as possible and still eliminate the MC problem. Re-run the regression without those variables. Use different ones instead.
3. Homework
	1. Chapter 14: 1-2
	2. In addition to these two questions, check for multicollinearity as described above. Interpret the results and suggest a correction.