David Youngberg
BSAD 210—Montgomery College

LECTURE 17: MULTIVARIABLE REGRESSIONS I

I. What Determines a House's Price?

a. Open **Data Set 5** to help us answer this question. You'll see pricing data for homes based on when they were built, how big each home is, how far it is from the city center, and how many days it was on the market before being sold.

   i. I don't remember where I got this data from. I'm pretty confident it's real but I doubt it's for our area.

b. Suppose you're researching how home prices change as you get closer to a city's downtown area. You'd suspect that homes should get cheaper as you go further from the city.

c. Here's a regression output (n=100) with miles from city center causing housing prices:

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	508457.7719	57163.32439	8.894825	3.02137E-14	395019.02	621896.5288	395019	621896.5
miles	-5517.997542	11504.88918	-0.47962	0.632564886	-28349.08	17313.08059	-28349.1	17313.08

d. While the coefficient is negative (as expected: more miles means a lower price) the result is ***not statistically significant***. Location, location, location…doesn't matter?

e. That can't be right—and it's not. The problem with this analysis is as homes get farther out, they get bigger.

   i. We asked the question, "If you buy a home farther from the city center, what happens to the price?"

   ii. We need to ask: "If you buy an ***identical*** home farther from the city center, what happens to the price?"

f. While it's hard to get data so we can compare "identical" homes, we can get data on one of the big variables here: size. Both size (in square feet) and distance from city center (in miles) matter for housing prices. So we turn to a multivariate regression.

g. Excluding an important variable can distort the regression analysis, resulting in *omitted variable bias*. It's when a variable that's correlated with the dependent variable and at least one independent variable is not included in the regression.

i. In our example, size was correlated both distance and price. Without size, we got a distorted understanding of what was going on. We were missing an important "control."

II. Basics

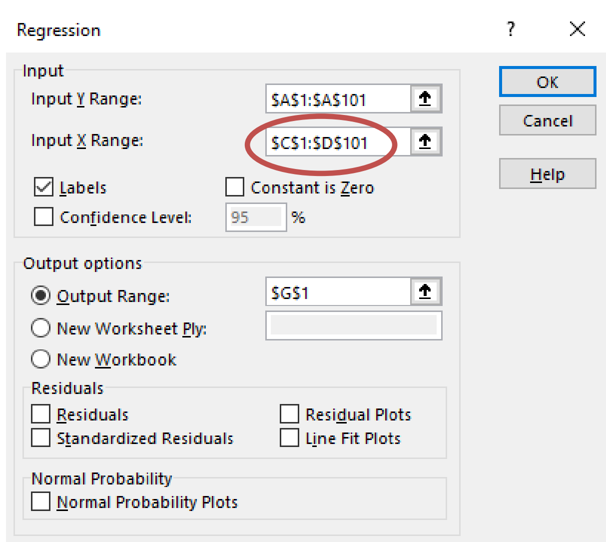
a. A multivariate regression has more than one explanatory variable.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \varepsilon_i$$

b. You want to do a multiple regression because you think multiple variables matter.
   i. Example: Life expectancy depends on both diet and exercise.
   ii. Example: Sales depends on price, the unemployment rate, advertising, and so on.

c. When you interpret a particular beta, it is still the change in the dependent variable for every unit change in the corresponding explanatory variable, but now it also ***holds all other explanatory variables constant***.

d. To do a multivariant regression in Excel, you need to draw a continuous box around multiple X variables for the Input X Range, as so:



   i. Note this means that all your X variables have to be next to each other. Recall that you can move columns of data by right-clicking the column letter you wish to move, selecting Cut, then right clicking the column letter you wish to move the column to and selecting Insert Cut Cells. Recall that Excel will always insert to the left of whatever you've selected.

e. Here's the bottom part of the housing regression results, now with size and location predicting price (remember when I suggested you use labels? This is why):

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	Ipper 95.0%
Intercept	78.70913492	82608.5323	0.000953	0.999241735	-163876.4	164033.7786	-163876	164033.8
sqft	236.9646693	32.00587304	7.403787	4.84964E-11	173.44187	300.4874676	173.4419	300.4875
miles	-23792.47937	9567.375458	-2.48683	0.014596448	-42781.07	-4803.887471	-42781.1	-4803.89

i. Now distance (and size!) are statistically significant (both p-values are less than 0.05).

ii. Our estimated line is:

$$PRICE = 78.7 + 237 * (SQFT) + -23{,}792 * (MILES)$$

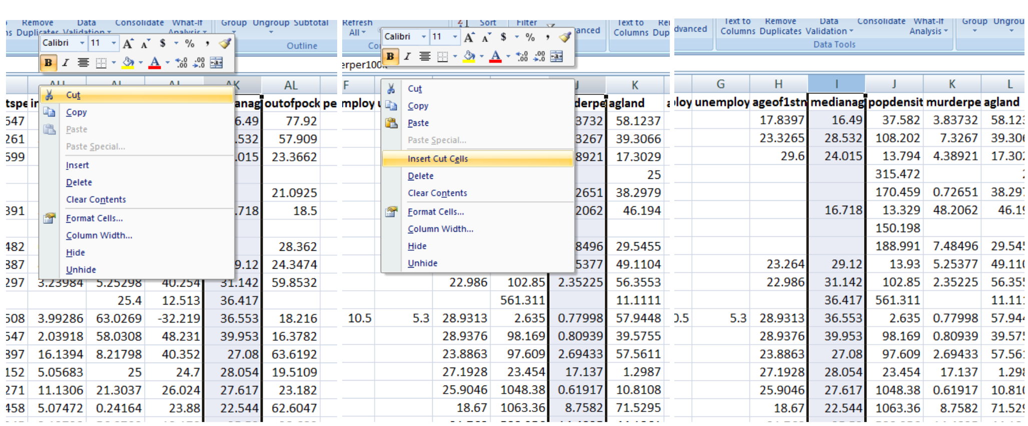
iii. For every additional square foot a house has the price increases by $236.96, controlling for distance from city center.

iv. For every additional mile the house is from the city center the price decreases by $23,792.48, controlling for size.

III. Preparing Your Data

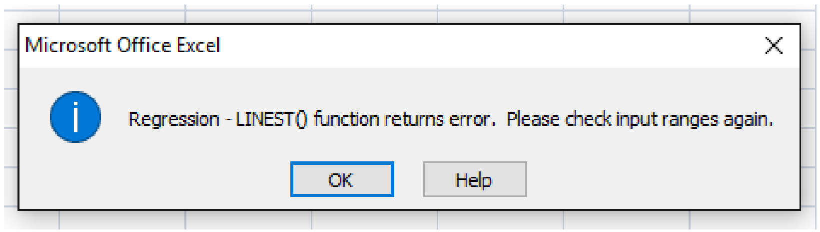
a. Excel requires that all explanatory variables for a regression are next to each other. Suppose, for example, I'm interested in how ageof1st marriage, population density, and median age affect the murder rate.

i. The easiest way to do this is to right click the column with the variable you're interested in, select "Cut", right click the column of another variable you're interested in, and select "Insert Cut Cells." Like this:



ii. Now all my explanatory variables are next to each other.

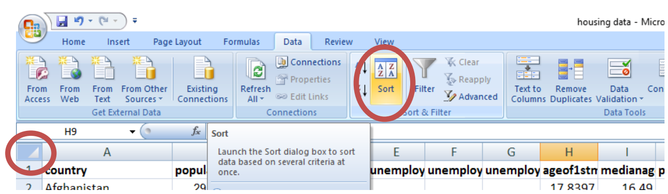
b. Excel requires that all variables have no blank observations. If you get this error message:



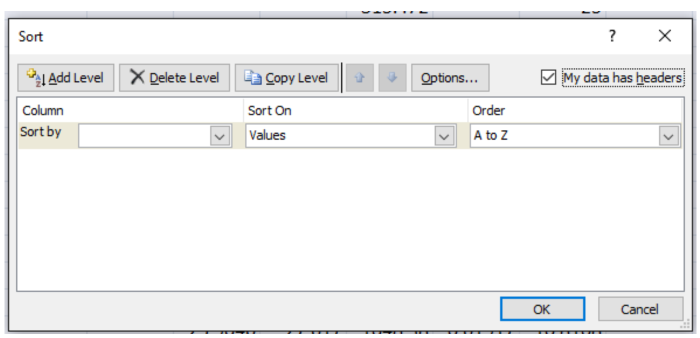
It means you are trying to run a regression using variables with missing values.

   i. Normally, a program would just ignore those observations. But Excel is kind of dumb. You have to delete observations that don't have values for every variable. Let me first show you a quick way to do that.

c. The Sort function is in the Data tab. Highlight the whole Excel sheet (by clicking in the upper right-hand corner of sheet) and select Data.



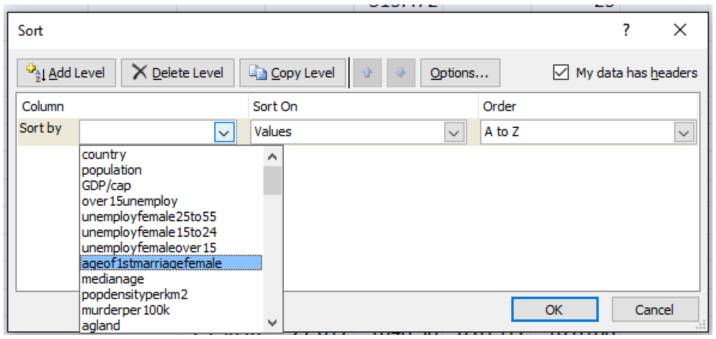
d. You'll get something that looks like this:



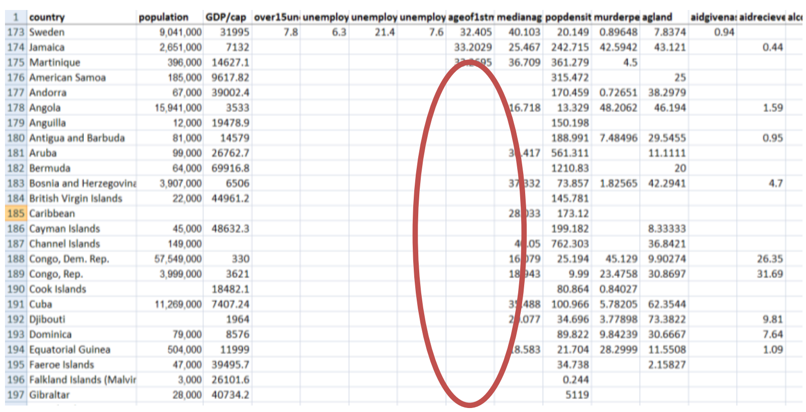
   i. Make sure to select "My data has headers." It'll make this a lot easier. Also make sure you selected all the variables.

      1. Sometimes students will only select the variable they wish to sort. When they do that, they render the data set worthless because all the data get unpaired from their element. Don't do that—select the whole data set.

e. In the dropdown menu, select ageof1stmarriagefemale. Then press OK.

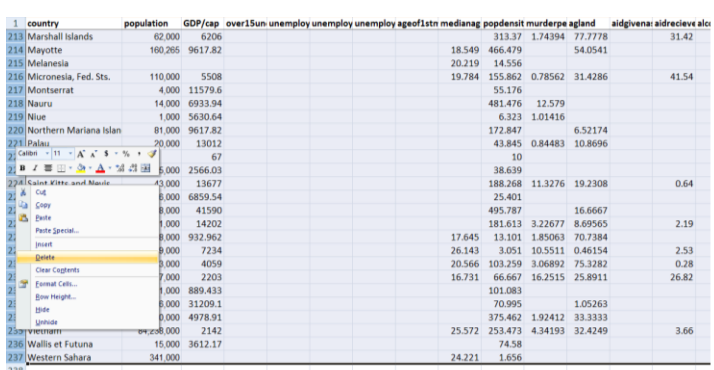


f. Excel will reorder the data based on that variable. This means all the blank values end up in the same place: at the end. This makes it a lot easier to find and delete all the observations with blank values.



i. ***This is an incredibly useful function for your everyday understanding of data.*** It makes it easy to, for example, find the largest values or put all observations of the same category next to each other.

g. Highlight rows (selecting the numbers so you get the whole row) starting in 176 all the way down to 237. Right click and select Delete.



h. Repeat this process for each variable that you care about (including your dependent variable) and you're ready to run the regression.

IV. 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.

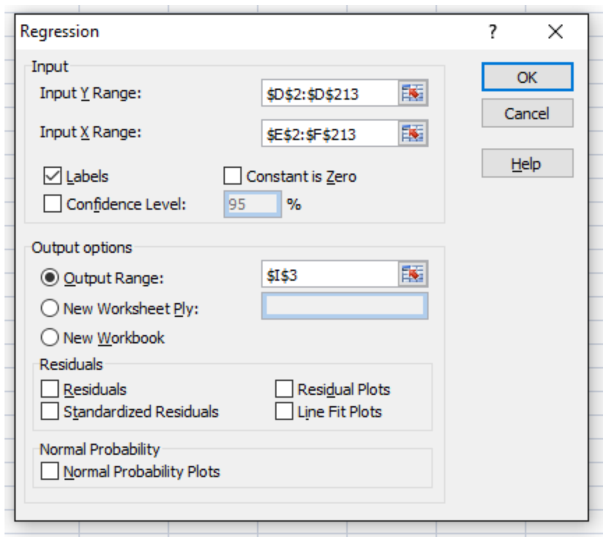
V. More Output from Excel

a. **Data Set 5** also has our RMP data, but now with a new variable: HOT?

b. Rate My Professor once asked 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

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

> 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.

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.

VI. 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.

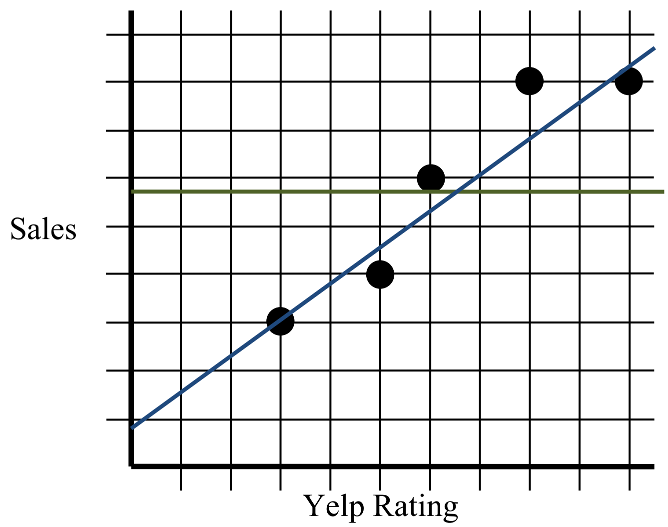
VII. Bonus: Understanding ESS and RSS

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.

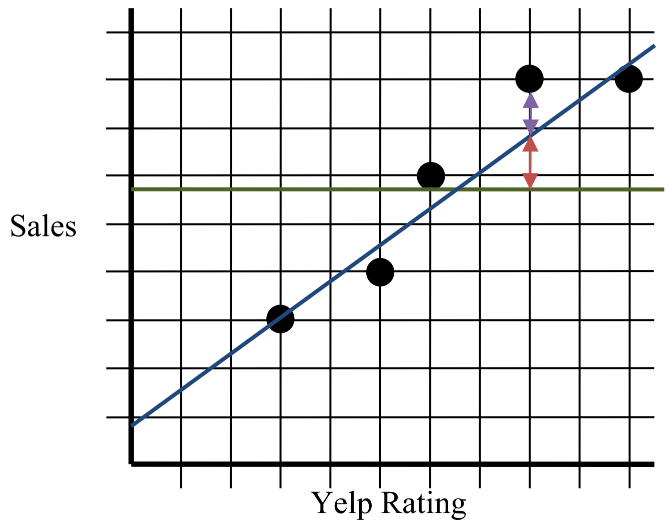
c. Now suppose you know that restaurant you chose has 4 out of 5 stars on Yelp, the popular review site. How do you adjust your expected sales? It should go up, right?

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.