David Youngberg
BSAD 210—Montgomery College

LECTURE 24: MULTIVARIATE REGRESSIONS I

I. What Determines a House's Price?
   a. Open Data Set 6 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, what happens to the price?"
      ii. We need to ask: "If you buy an ***identical*** home farther from the city, 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.

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

i. Note that the explanatory variables now have a subscript to differentiate them from other explanatory variables.

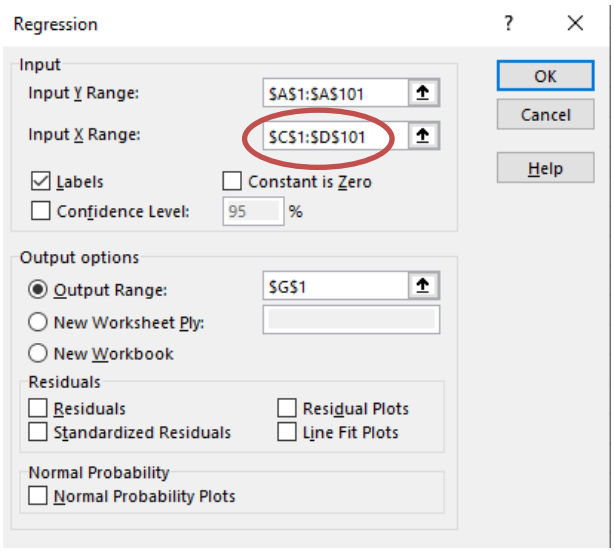
b. You want to do a multiple regression because you think multiple variables matter.

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 value, it is now the change in the dependent variable for every unit change in the corresponding explanatory variable, ***holding 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. Note Excel will always insert to the left of whatever you've selected.

e. Here's the housing regression, 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%	pper 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.

ii. Our estimated line is:

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

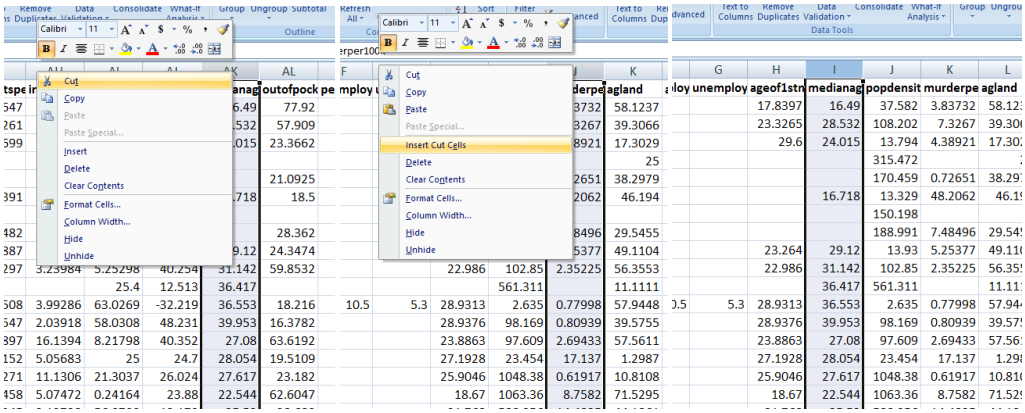
iii. For every additional square foot a house has the price increases by $236.96, holding distance from the city center constant.

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

III. Preparing Your Data

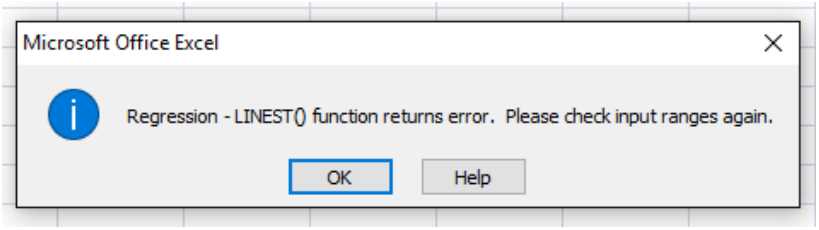
a. Excel requires that all explanatory variables for a regression are next to each other. Suppose, for example, I'm interested in how ageof1st marriage, population density, and median age affect the murder rate.

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.

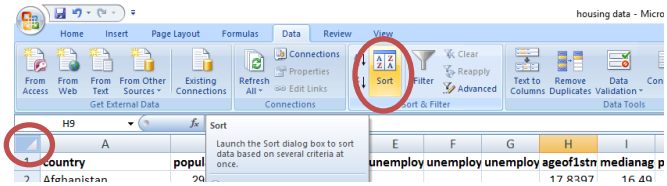
b. Excel requires that all variables have no blank observations. If you get this error message:



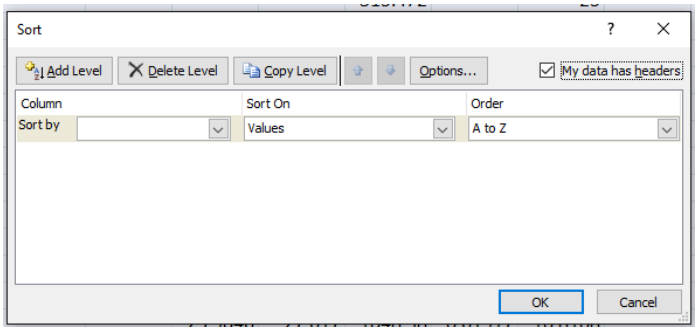
It means you are trying to run a regression using variables with missing values.

i. Normally, a program would just ignore those observations.
ii. But Excel is kind of dumb. You have to delete them. So 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.

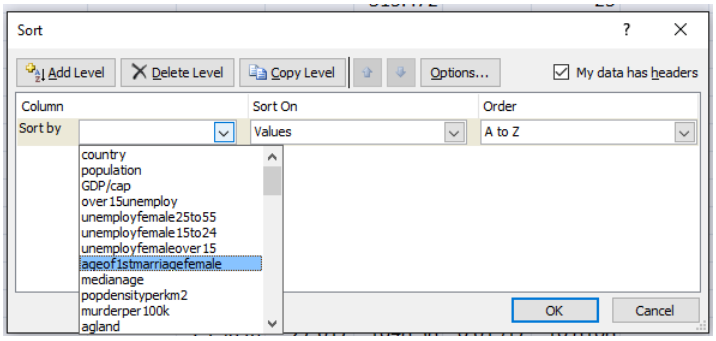


d. You'll get something that looks like this:



i. Make sure to select "My data has headers." It'll make this a lot easier.

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.

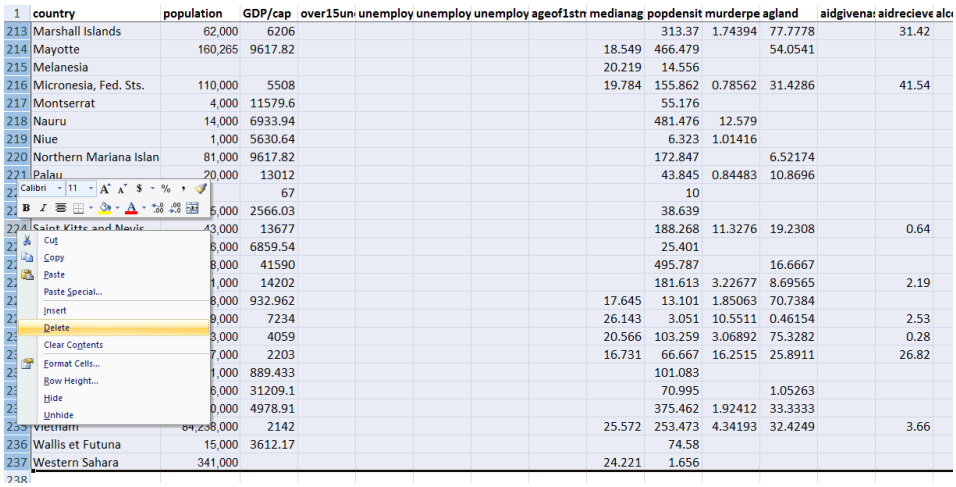
1	country	population	GDP/cap	over15un	unemploy	unemploy	unemploy	ageof1stn	medianag	popdensit	murderpe	agland	aidgivena	aidrecieve	alc
173	Sweden	9,041,000	31995	7.8	6.3	21.4	7.6	32.405	40.103	20.149	0.89648	7.8374	0.94		
174	Jamaica	2,651,000	7132					33.2029	25.467	242.715	42.5942	43.121		0.44	
175	Martinique	396,000	14627.1					33.2695	36.709	361.279	4.5				
176	American Samoa	185,000	9617.82							315.472		25			
177	Andorra	67,000	39002.4							170.459	0.72651	38.2979			
178	Angola	15,941,000	3533						16.718	13.329	48.2062	46.194		1.59	
179	Anguilla	12,000	19478.9							150.198					
180	Antigua and Barbuda	81,000	14579							188.991	7.48496	29.5455		0.95	
181	Aruba	99,000	26762.7						36.417	561.311		11.1111			
182	Bermuda	64,000	69916.8							1210.83		20			
183	Bosnia and Herzegovina	3,907,000	6506						37.332	73.857	1.82565	42.2941		4.7	
184	British Virgin Islands	22,000	44961.2							145.781					
185	Caribbean								28.033	173.12					
186	Cayman Islands	45,000	48632.3							199.182		8.33333			
187	Channel Islands	149,000							40.05	762.303		36.8421			
188	Congo, Dem. Rep.	57,549,000	330						16.079	25.194	45.129	9.90274		26.35	
189	Congo, Rep.	3,999,000	3621						18.943	9.99	23.4758	30.8697		31.69	
190	Cook Islands		18482.1							80.864	0.84027				
191	Cuba	11,269,000	7407.24						35.488	100.966	5.78205	62.3544			
192	Djibouti		1964						20.077	34.696	3.77898	73.3822		9.81	
193	Dominica	79,000	8576							89.822	9.84239	30.6667		7.64	
194	Equatorial Guinea	504,000	11999						18.583	21.704	28.2999	11.5508		1.09	
195	Faeroe Islands	47,000	39495.7							34.738		2.15827			
196	Falkland Islands (Malvir	3,000	26101.6							0.244					
197	Gibraltar	28,000	40734.2							5119					

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.

ii. Remember when we used RMP data and I had you analyze ratings for different disciplines? I got all the disciplines next to each other using the Sort function.

iii. You can also add "levels" (see window on the previous page). This will tell Excel to sort within categories. For example, I could have it sort by discipline and then sort by number of ratings. Within each discipline, the professor with the most (or, if I choose, fewest) ratings would be listed first.

g. Highlight rows starting in 176 all the way down to 237. Right click and select Delete.



1	country	population	GDP/cap	over15un	unemploy	unemploy	unemploy	ageof1stn	medianag	popdensit	murderpe	agland	aidgivena	aidrecieve	alc
213	Marshall Islands	62,000	6206							313.37	1.74394	77.7778		31.42	
214	Mayotte	160,265	9617.82						18.549	466.479		54.0541			
215	Melanesia								20.219	14.556					
216	Micronesia, Fed. Sts.	110,000	5508						19.784	155.862	0.78562	31.4286		41.54	
217	Montserrat	4,000	11579.6							55.176					
218	Nauru	14,000	6933.94							481.476	12.579				
219	Niue	1,000	5630.64							6.323	1.01416				
220	Northern Mariana Islan	81,000	9617.82							172.847		6.52174			
221	Palau	20,000	13012							43.845	0.84483	10.8696			
222			67							10					
223		5,000	2566.03							38.639					
224	Saint Kitts and Nevis	43,000	13677							188.268	11.3276	19.2308		0.64	
225		8,000	6859.54							25.401					
226		8,000	41590							495.787		16.6667			
227		1,000	14202							181.613	3.22677	8.69565		2.19	
228		8,000	932.962						17.645	13.101	1.85063	70.7384			
229		9,000	7234						26.143	3.051	10.5511	0.46154		2.53	
230		8,000	4059						20.566	103.259	3.06892	75.3282		0.28	
231		7,000	2203						16.731	66.667	16.2515	25.8911		26.82	
232		1,000	889.433							101.083					
233		8,000	31209.1							70.995		1.05263			
234		0,000	4978.91							375.462	1.92412	33.3333			
235	Vietnam	84,238,000	2142						25.572	253.473	4.34193	32.4249		3.66	
236	Wallis et Futuna	15,000	3612.17							74.58					
237	Western Sahara	341,000							24.221	1.656					
238															

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