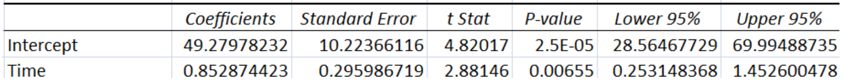
David Youngberg
BSAD 210—Montgomery College

LECTURE 26: MULTIVARIATE REGRESSIONS III

I. Time and Total Example

a. Open Data Set 7; you'll see exam data I gathered for an economics class I taught in the fall of 2011.

b. How does exam score relate to how long a student took to complete the exam? A while back I recorded the time, in minutes, each student took to complete an exam. There were 39 students in all. Then I ran a regression with Time predicting Total:



	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	49.27978232	10.22366116	4.82017	2.5E-05	28.56467729	69.99488735
Time	0.852874423	0.295986719	2.88146	0.00655	0.253148368	1.452600478

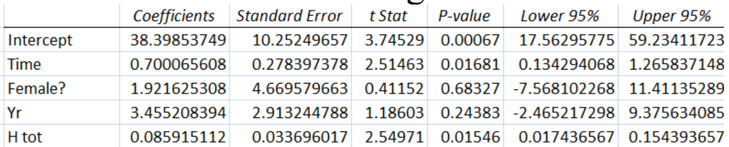
i. I'm claiming spending more time on your exam should increase your total, all other things equal. The story is that you'll complete work you didn't get to, you'll check your answers and fix mistakes, you'll have more time to remember things, etc.

ii. More time means more points, with each additional minute adding about 0.85 points to the exam. It appears we have statistical significance.

II. Time and Total, Advanced

a. Still, my analysis seems incomplete. Shouldn't more knowledgeable students take more time? What's the effect of how much experience they have as a student? Does gender play a role?

b. Let's throw all these into the original model:



	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	38.39853749	10.25249657	3.74529	0.00067	17.56295775	59.23411723
Time	0.700065608	0.278397378	2.51463	0.01681	0.134294068	1.265837148
Female?	1.921625308	4.669579663	0.41152	0.68327	-7.568102268	11.41135289
Yr	3.455208394	2.913244788	1.18603	0.24383	-2.465217298	9.375634085
H tot	0.085915112	0.033696017	2.54971	0.01546	0.017436567	0.154393657

i. Where *Time* is as before, *Female?* is a dummy variable (1 is female, 0 is male), *Yr* is the year of the student (1 to 4, with 1 being a freshman and 4 being a senior), and *H tot* is the total the student received on the relevant homework assignments (two assignments at 100 points each).

ii. Note that Time is lower than before—other variables are "doing the work" that only Time did—but it's still statistically significant.

iii. Now an additional minute results in an additional 0.7 points. An additional ten minutes yields 7 additional points.

iv. It might seem homework total, while statistically significant, doesn't matter that much (it's not practically significant). But keep in mind homework total ranges from 0 to 200 while Time maxed out at 55 (they had just under an hour to complete the exam).

v. Gender and Year don't matter; neither is statistically significant.

III. Multicollinearity

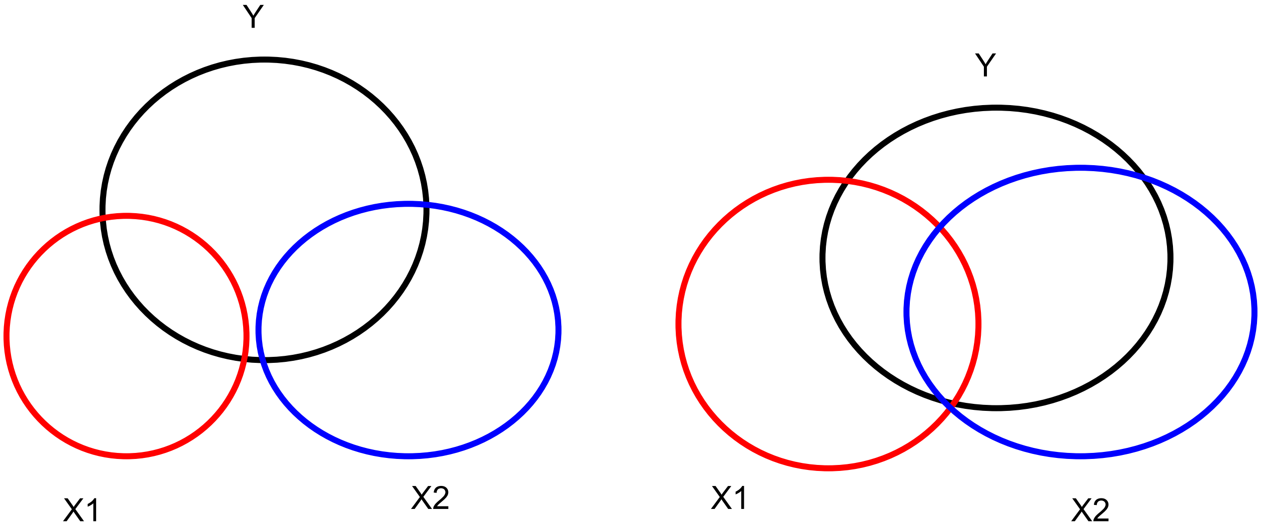
a. Suppose I run the previous regression but instead of just Homework Total, I use Homework Total and H4. But H4 isn't statistically significant and Homework Total's p increased (though it's still statistically significant). Why? It's because Homework Total and H4 are strongly correlated with each other.

b. If one or more pairs of our explanatory variables are highly correlated, we have multicollinearity.

i. It doesn't require perfect correlation (if there's perfect correlation, the program will drop one of the variables).

ii. <u>Example</u>: 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.

c. 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.



i. On the right, there is a portion of variation in Y that can be attributed to either X.

d. Imagine you're testing cupcake recipes with customers rating different types. Some of your recipes have lots of sugar and butter (type A), some have a moderate amount of each (type B), some have only a little of each (type C).

i. The type A cupcakes will certainly be most liked, followed by type B, and then by type C.

ii. But are the type A's liked because of the sugar or because of the butter? How important is each one? Can you lose some sugar and get the same level of enjoyment? You don't know because you have multicollinearity!

iii. You would need cupcakes with low amounts of sugar but lots of butter and vice versa. You need to reduce the correlation between the two explanatory variables.

e. You can notice multicollinearity if:

i. The F-test suggests your model as a whole is strong but none of your variables are significant.

ii. The regression coefficients change a lot when a variable is added or deleted.

iii. Check the correlation coefficient for each pair combination of your explanatory variables (e.g. five explanatory variables would mean ten pairs). A high correlation coefficient (such as ±0.8) suggests multicollinearity.[1] But there is no hard standard.

f. The easiest way to correct for multicollinearity is to remove variables. If it's functionally redundant with something else, why is it there?

[1] A more technical way to do this is Variance Inflation Factors (VIFs). The technique is beyond the scope of this course.