David Youngberg
BSAD 210—Montgomery College

LECTURE 30: BASICS OF MULTIVARIABLE REGRESSIONS

I. Control

a. A multiple regression has more than one independent variable.

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

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

i. Example: It makes sense that life expectancy depends on both diet and exercise.

c. The nice thing about this is that this technique lets you control for other factors.

i. 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.

d. 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).

ii. You can use multiple dummies for a variable with a few categories (White? Black? Asian? Hispanic?)

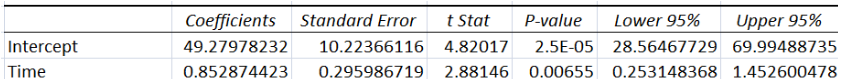
iii. You always want to have a number of dummies equal to one minus the number of categories. If the dummy is "Female?" then you know 1=F and 0=M. Adding "Male?" is redundant.

e. 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.

II. Time and Total, Basic

a. A friend of mine once proposed that a graph of the total on the exam and the time it takes to complete the exam should look like an inverted-U shape. Students who take little time don't know anything and gave up; those who take a lot of time don’t know anything and are just desperate.

b. I’m suspicious of the claim so a while back I recorded the time, in minutes, each student took when he or she handed in the 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

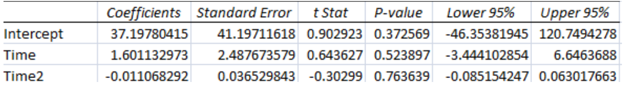
   i. I’m claiming spending more time on your exam should increase your total, all other things equal. The logic is that you’ll complete work you didn’t get to, you’ll check your answers and fix mistakes, you’ll have more time to remember things, etc.

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

c. Would an inverted-U fit better? One way to check is to add a variable, Time^2, to the equation.

   i. To do this, I squared each Time observation and use both variables to predict Total.

   ii. Now we have two variables—Time and Time^2—to predict our equation. Here are the results:



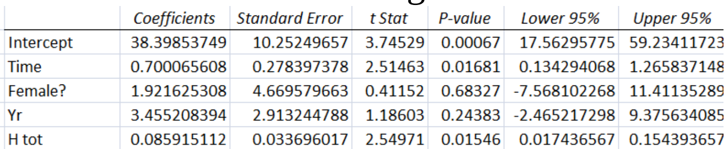
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	37.19780415	41.19711618	0.902923	0.372569	-46.35381945	120.7494278
Time	1.601132973	2.487673579	0.643627	0.523897	-3.444102854	6.6463688
Time2	-0.011068292	0.036529843	-0.30299	0.763639	-0.085154247	0.063017663

   iii. Neither variable is statistically significant anymore. Time^2 is just redundant. As we will discuss later, redundancy in variable selection can ruin a model.

III. 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, and *H tot* is the total the student received on the relevant homeworks.

   ii. Note that Time is lower than before—other variables are “doing the work” that only Time did—but it’s still significant.