Executive Summary for Lecture Set #2

This set of slides illustrates how we can use regressions to compare average differences in some Y variable between two (or more) groups – groups of people or time periods. We accomplish this by using something called a "dummy variable": a variable equal to either 1 if it represents data collected from a particular group, and 0 if it is not collected from this group. Let's allow X to represent such a dummy variable in our simple, bivariate regression:

$$Y = b_0 + b_1 X$$

Lesson #1: Let's suppose that X is equal to one if a person in our sample has a first name that starts with letter L, and zero if the person's first name does not start with the letter L; and suppose the Y represents weekly earnings. In this case, what's our prediction about the average value of Y (weekly earnings) for people whose first name does not start with L? Perform the prediction by setting X = 0:

$$Our\ Best\ Guess\ about\ Y = Estimate\ of\ b_0 + (Estimate\ of\ b_1)(0) = b_0$$

Note that b_0 represents the average value of Y for the group NOT represented by dummy variable X (we call this group the "***Omitted Reference Category***").

Lesson #2: What's our prediction about the average value of Y (weekly earnings) for people whose first name *does* start with L? Again, perform the prediction by setting X = 1:

$$Our\ Best\ Guess\ about\ Y = Estimate\ of\ b_0 + (Estimate\ of\ b_1)(1) = b_0 + b_1$$

Note that $(b_0 + b_1)$ represents the average value of Y for the group represented by dummy variable.

Lesson #3: What's the difference in the average value of Y between the two groups here? Subtract our result in Lesson #1 from our result in Lesson #2:

$$(Average\ Y\ for\ the\ group\ with\ X = 1) - (Average\ Y\ for\ the\ group\ with\ X = 0)$$
$$= (b_0 + b_1) - (b_0) = b_1$$

The coefficient b_1 shows us the average difference in weekly earnings between the two groups! And this result generalizes: we can compare the average difference in any Y between two groups of people, or two time periods (or any two groupings we like). Two other points: (i) we can also make comparison across more than just two groups (please refer to extra posted materials for that) and (ii) we'll need to be careful to ***NOT*** interpret this result in a *causal* way.

Lesson #4: We can use multiple dummy variables (let's call them X_1, X_2 and X_3) in a single regression, too:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3$$

The same general principles will apply to this type of regression as the bivariate case:

(i) b_0 will still represent the average value of Y for the omitted reference category – the group NOT represented by ***ANY*** dummy variables in our regression.

(ii) b_1, b_2 and b_3 each represent the average difference in Y between the omitted reference group and each group represented by the dummy variables X_1, X_2 and X_3, respectively.