Executive Summary for Lecture Set #1

There are three main ideas contained in the first set of lecture slides, and they both relate to regressions. In order to contextualize the discussion, let's specify a simple regression:

$$Y = b_0 + b_1 X$$

Using this regression, we can review the three main lessons imparted within these slides.

Lesson #1: Regressions allow us to examine how a **change** in X will **change** Y (on average). As you'll see, a one-unit change in X will change Y by b_1 (on average). This is handy to know because (for example) if we allow X to represent advertising expenses, and allow Y to represent our profits, then the regression will tell us that every additional dollar spent on advertising will change our profits by b_1, on average. Note that this idea focuses on just **one** of the coefficients in the regression: b_1. So, whenever I ask about the way in which one factor **changes** another factor, I'm really asking about a **single coefficient** from our regression.

Lesson #2: Regressions also allow us to perform predictions about Y. Specifically, if we know the **value** of X, then we can predict the average value of Y in this case. As an example, if we happen to know that X has a value of 25, then we can use the regression to make a prediction about the value of Y in this case by plugging this value directly into our regression:

$$Our\ Best\ Guess\ about\ Y = Estimate\ of\ b_0 + (Estimate\ of\ b_1)(25)$$

Note that unlike "Lesson #1" (that focuses on only one coefficient), our prediction incorporates the **entire right-hand-side** of the regression: it's necessary to calculate a sum that includes **both** b_0 **and** b_1, ***along with a specific numerical value for X***. So, whenever I ask for your best guess about a particular Y variable, while providing you with specific numerical values for a right-hand-side variable, I'm really asking you to make a prediction by calculating a sum that involves the entire right-hand-side of the regression.

Lesson #3: If we omit an important right-hand-side variable from our regression, then coefficient estimate will probably be wrong. In particular, suppose we estimate the original regression at the top of this page (with one small change in the name of the "X" variable):

$$Y = b_0 + b_1 X_1$$

This regression dictates that values of Y depend **ONLY** on values of X_1, and are completely unrelated to **ANY OTHER FACTORS**. This is a very strong assumption, and can lead to big problems. In particular, suppose that values of Y depend on two variables (not just one): X_1 and X_2. In this case, we'd need to estimate something called a "multiple regression" that looks like this (and we'll discuss the details for interpreting this regression in class):

$$Y = c_0 + c_1 X_1 + c_2 X_2$$

The key point to understand here is that our estimates of b_1 and c_1 are different: $c_1 \neq b_1$, and this is a more general problem called "Omitted Variable Bias" (or OVB, for short). This will constitute the major problem to be solved in future lectures.