**SUMMARIZING SOME KEY CONCEPTUAL PIECES IN REGRESSION ANALYSIS**

**CATEGORY 1:**

THINGS THAT EXIST BUT THAT WE DO NOT KNOW, AND CAN NEVER KNOW FOR SURE

The main, big, important thing: β1

The exact functional form of the relationship between X and Y

The true amount of unexplained variation in Y in the population (i.e. the true “ui” values)

* Even in the population, the Y values are not perfectly determined by X (or even many X’s) – there is still likely to be some random variation around even the true regression line.
* In other words, even if we know β0 and β1, and thus have the truest regression line, it does not mean that every person’s value would be perfectly on that line (i.e. as long as there are two people with the same X but different Y, they can’t both be on the line)

**CATEGORY 2:**

THINGS THAT WE CAN ESTIMATE TO TRY TO LEARN ABOUT WHAT WE DON’T KNOW

and are estimated via OLS (ESTIMATOR + DATA ESTIMATE)

**CATEGORY 3:**

THINGS THAT WE CAN USE TO TRY TO DECIDE IF OUR ESTIMATES ARE USEFUL

 can be calculated after you have gotten the OLS estimates, so that you can see what your model predicts for any given Xi – you can see if that seems useful by taking the next step:

 can be calculated for a given observation using their actual Yi minus (predicted value)

*  can be used to help calculate R2 to assess “goodness of fit”
* can be used to help calculate the standard error of 

These are usually done automatically within the software, but I want it to be clear to you that calculation of R2 and the standard error of  both depend upon 

Once we have the standard error of , then we can do hypothesis tests (t-statistics, confidence intervals) to establish our level of confidence in , and in particular assess whether we have convincingly established whether the real β1 is positive or negative, or if we are still not sure whether there is a nonzero relationship between X and Y.