Catherine Anne Prideaux & Connor Campbell
STA 371H--11:00

February 12, 2014 Scribing Notes

Review: T-Stats and Confidence Intervals
New: Dummy/Grouping Variables

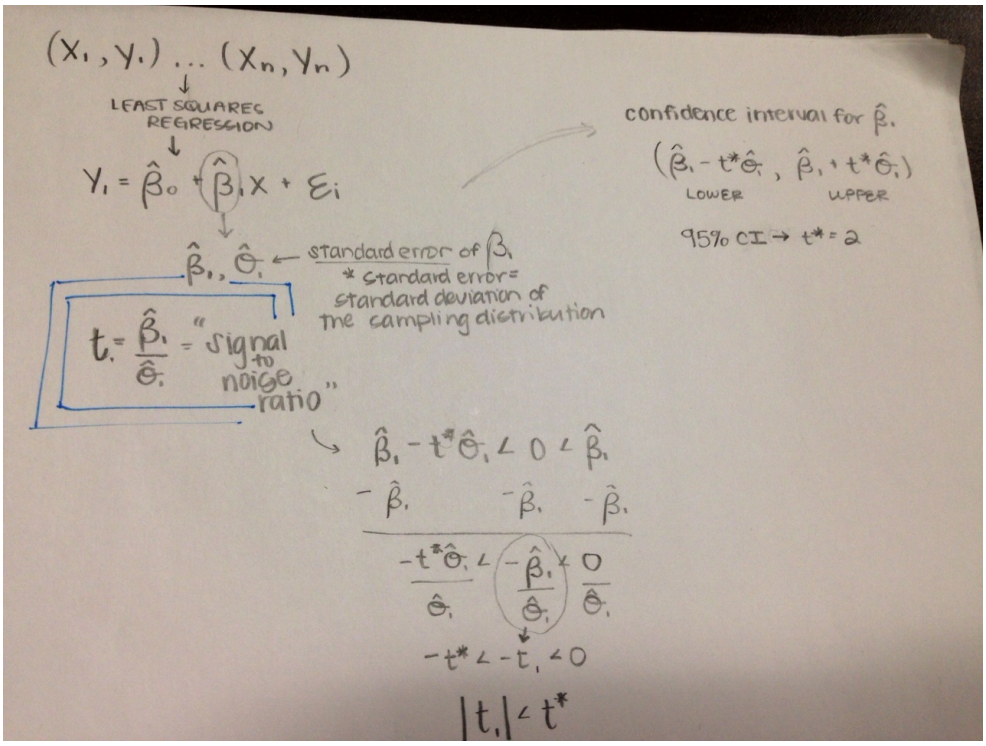
Today in class was a closed computer day - no R.

T-STATISTICS

A t-static, referred to as the “signal-to-noise” ratio, is a measure of the likelihood that the actual value of the parameter is not zero. It is computed by dividing the estimated value (either slope or intercept) by its standard error (the standard deviation of the sampling distribution). A t-stat is unitless (y/x over y/x), and the values come from the regression model.

We use t-statistics as a quick way to evaluate whether or not zero is in the confidence interval. The larger the absolute value of t, the less likely that the actual value of the parameter could be zero. (Dr. Scott does not use t-stats very often, but it is critical to understand for statistical literacy purposes.) Note that it is just a ratio and has no information about the nature of the data, in fact in using the t-statistic data is lost.

Below is the proof of how to get a t-statistic as well as the t-stat formula.

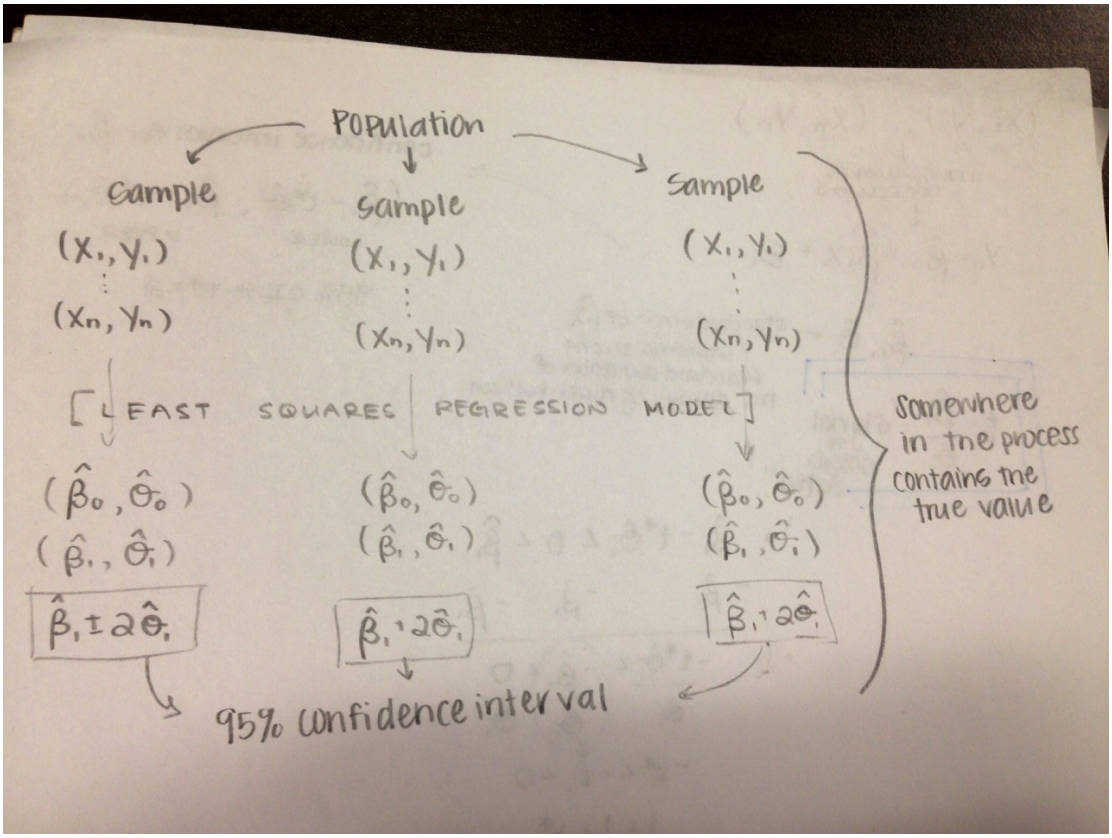


CONFIDENCE INTERVALS

What makes a confidence interval a confidence interval?

A confidence interval gives an estimated range of values, calculated from a given set of samples, which is likely to include an unknown population parameter. It is a range of values for the "truth" or the actual value. A confidence interval gives an upper value and a lower value, creating a range to minimize the effects of sampling error. A wider range width creates more certainty. It is frequently referred to as the "95% frequentist coverage property" which means that the interval covers the true value 95% of the time (% chosen). Confidence intervals assume that the distribution in the samples is the same as the distribution for the entire population.

Below is the process Dr. Scott outlined for getting a confidence interval. He compared creating a confidence interval to testing for default widgets in a factory. For our example, the worker sampled every 100th widget 1000 times and found 5% to be defective (50 defective, 950 okay). Therefore the "95% widget property" (ie--the 95% frequentist coverage property) states that 95% of widgets coming out of the factory should be okay (regardless of whether or not they were actually sampled).

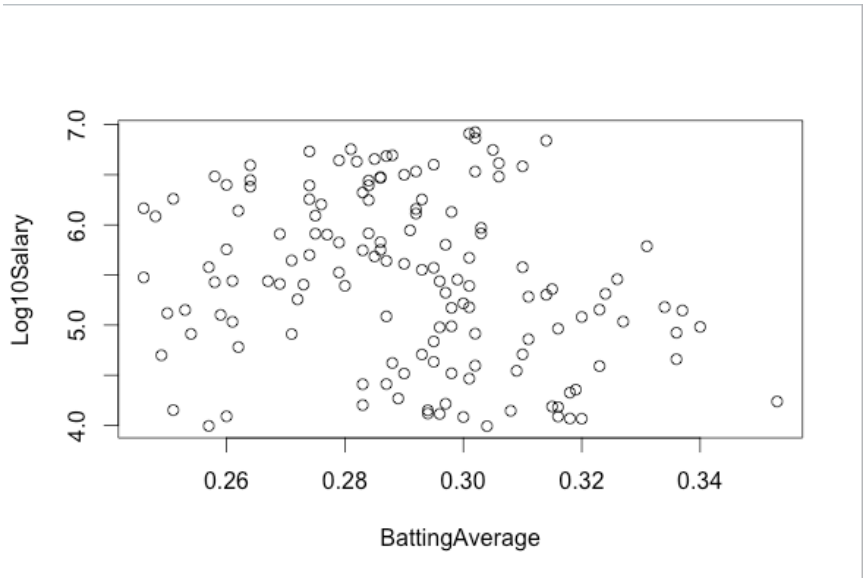


The methods we have learned so far in class to get confidence intervals (such as bootstrapping and two standard deviations) all sufficiently meet the 95% frequentist coverage property. Additionally, to get a confidence interval in R, you can use the code confint(x, 0.95).

GROUPING VARIABLES IN REGRESSION

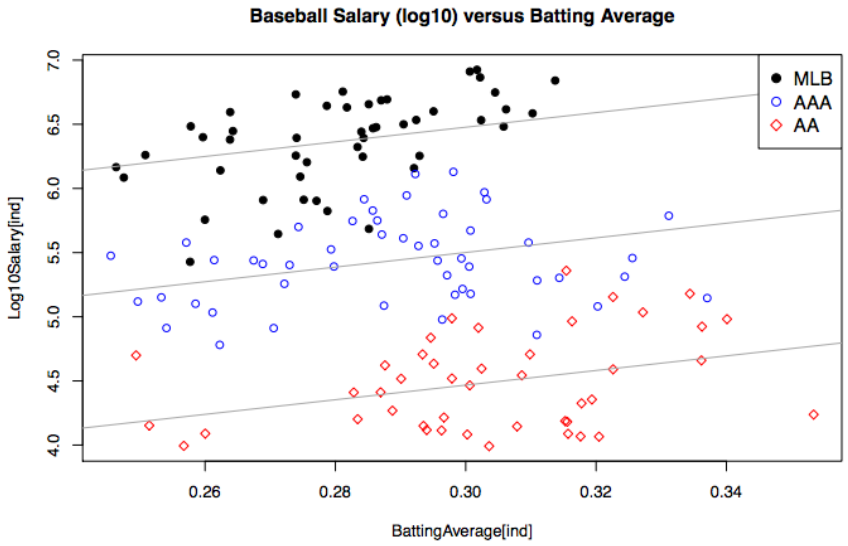
also see chapter 5 of Dr. Scott's course packet (http://jgscott.github.io/STA371/files/05-GroupingVariables.pdf)

We first looked at a graph of baseball players' batting average and their $\log_{10}$ salary.



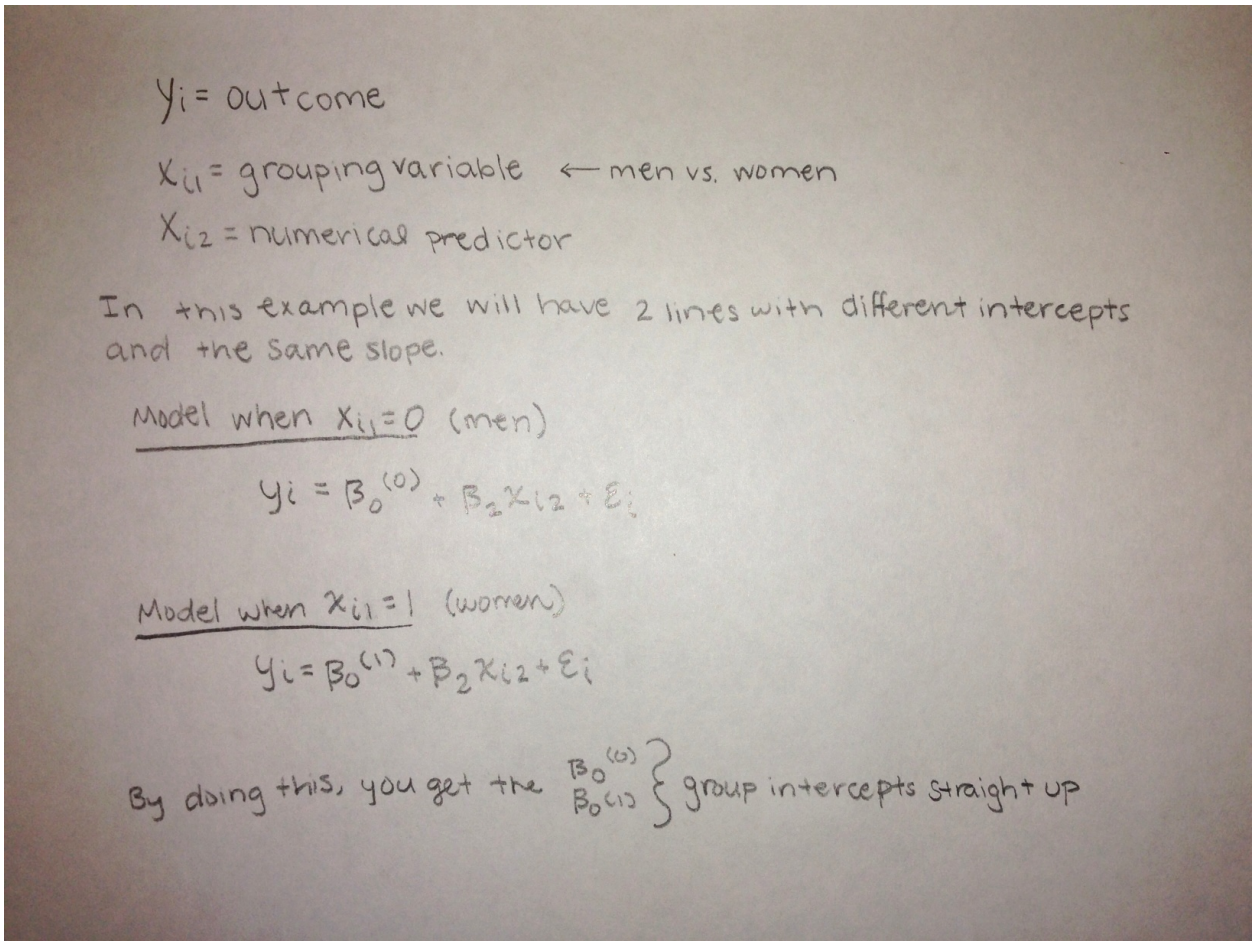
However, when you look at the graph, why is the trend downward? This is counter intuitive because it seems that the better one bats, the more they would be paid.

This can be explained through the aggregation paradox. This is where the trend that holds for individuals does not hold for groupings of individuals. In the case of the batting average, we can see the the data is aggregated across all leagues of baseball. When we disaggregate the data, the expected trend becomes much more clear (see below). There are now positive slopes showing that salaries increase with a higher batting average.

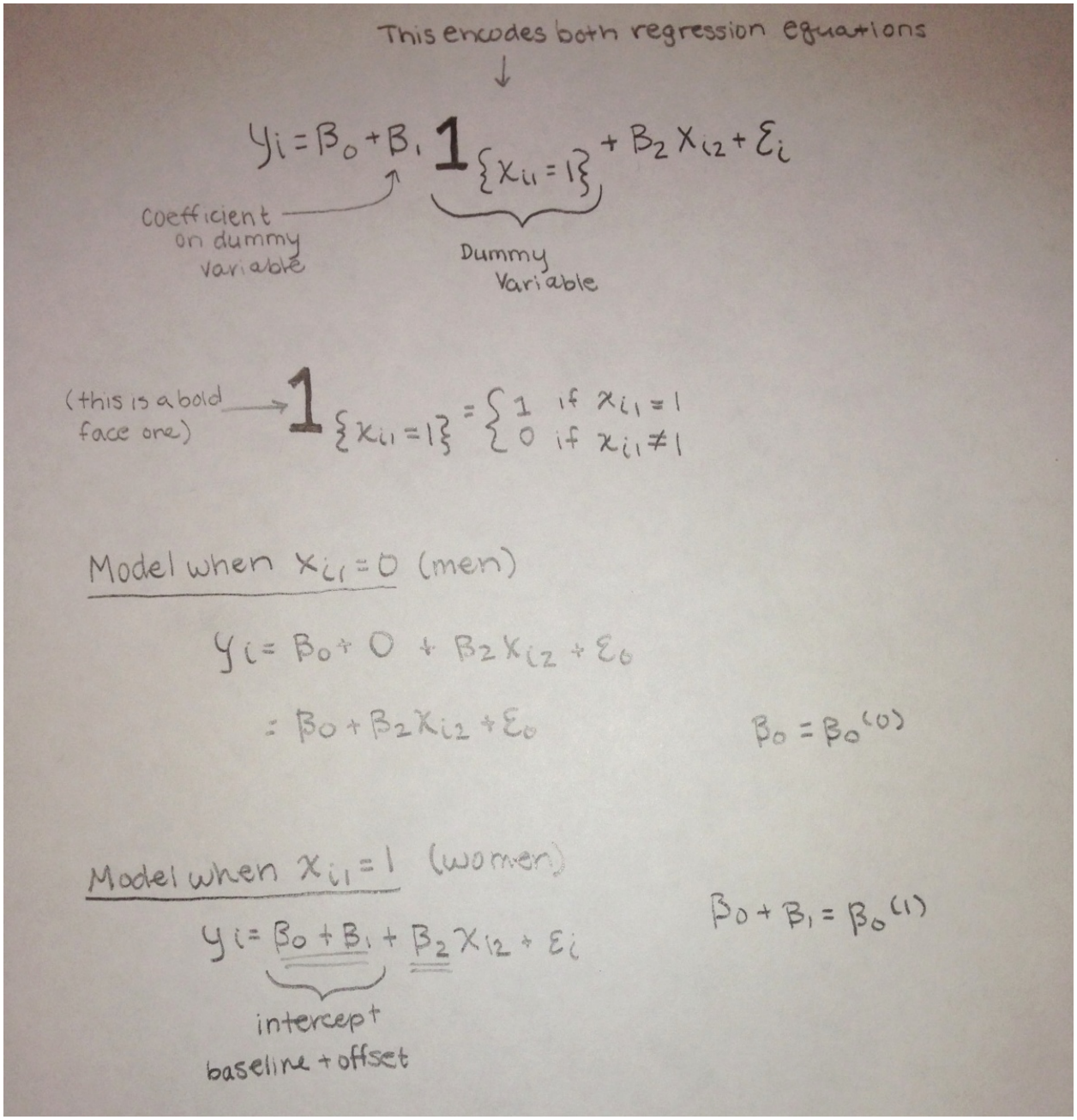


Dr. Scott showed us other examples of the aggregation paradox on an in-class powerpoint. One example was the Berkeley women's case. Just like the batting average and salary example. We looked into aggregating versus disaggregating the data.

We then learned how to do this using the Ten Mile Race (Cherry Blossom Run) example. This is where the data contained aggregated men and women, and we chose to separate it assigning men the value 0 and women the value 1 (see below). This example is for two variables, but you can see it worked with more in the course packet.



The second way to do this is though Baseline/Offset form (also known as Dummy Variable form) (see below).



Dr. Scott ended class by stating a video on how to do this in R would be posted. Follow the link to the video here. http://www.youtube.com/watch?v=wMKmQxBxUBE