**CEE-6/CH-36: Final graded assignment (take home)**

Answer the following questions using Stata

As part of this assignment you must construct, interpret both graphically and quantitatively, at least one categorical x categorical interaction, one categorical x continuous interaction, one continuous x continuous interaction, and one additional interaction of any kind. Note that you have to provide your reasoning as to why the two variables might interact but the interaction does not have to be statistically significant. Some of the data sets may have significant interactions and others might not. For all questions, you have to do whatever preliminary work is required to do the analysis you plan, then construct the model, then do enough diagnostic work to satisfy that your model is appropriate (i.e., CLINE for linear regression). Please note that you may have a satisfactory-looking model that, for example, doesn’t do perfectly on all the diagnostics. That’s OK and is part of the learning process. For example, suppose you do a linktest (as you should) and hat is significant, indicating your model contains important predictors but hat\_squared is also significant, meaning you are missing important variables. You should certainly address this but, in the end, you might not be sure what other variables to try and just explain that your model is good but still missing important information.

1. The data set school\_quality contains data that predicts school quality. The outcome variable for this question is whether a school is considered high quality or not and your job is to construct an appropriate model that best predicts whether or not a school is high quality.

2. The data set exercise.dta contains a subset of a bigger data set relating exercise to weight loss. The issue of physical and mental health has, if anything, come more to the forefront during the Covid-19 pandemic than at any other time. The data set is fairly self-explanatory. People were assigned to jogging or power-walking (their preference), swimming or no specific exercise program. All subjects were given teaching on the importance of exercise to health. Use the data to construct a model predicting pounds lost.

3. Conduct an appropriate analysis on the managers data set. This was a situation where a large Western US state was trying to determine the optimum number of managers to run large water supply plants. Keep in mind that the typical salary of one manager is in the six figures so budget constraints are, as usual, an issue. Be sure to do all necessary preliminary work and then write-up your findings in a brief report.

4. Use the crime.dta data to construct a linear regression with crime as the outcome variable.

5. Obtaining an adequate amount of essential minerals is always important, but even more so during pregnancy. Zinc is considered a very important nutrient during pregnancy and a hypothesis exists that vegetarian women might not get adequate amounts of zinc (zinc is most plentiful in animal products, especially red meat and certain fish). A study was done to examine this hypothesis. In the data set zinc.dta, four groups were compared: pregnant vegetarians (group = 1), non-pregnant vegetarians (group = 2), pregnant non-vegetarians (group = 3), and non-pregnant non-vegetarians (group = 4). Zinc levels (rounded) are given in µg/g of hair. Note that in any question that refers to “required preliminary testing,” I am referring to, as well as you can, checking the assumptions of whatever hypothesis test you are thinking of using.

a) Is there evidence that any group zinc level differs from any other? Be sure to give your null and alternative hypotheses and do any required preliminary analyses.

b) If there is evidence that at least one of these group zinc levels is significantly different than another, report all groups that are statistically different or not different and at what p-value. Briefly interpret the overall findings.

c) Suppose another research group reports that in a hypothesis test examining only non-pregnant vegetarians and non-pregnant non-vegetarians, a statistically significant result was obtained with a mean zinc difference of -30 μg/g (comparing the vegetarians to the non-vegetarians). Is this study concordant with the data you have? Should you use the results obtained for (b) to reply to these researchers or is there a better/simpler analysis you can do? If you decide to use the information from (b), state what mean zinc difference value you report to these other researchers. If you decide to do a different analysis, state Ho and Ha and do the analysis. Give your conclusions.

d) Is there evidence that pregnancy itself may stress zinc levels? Be sure to specify how you test this hypothesis. Give the null hypothesis and whether you reject it.