**S-040: Introduction to Applied Data Analysis**

Harvard Graduate School of Education

Problem Set C: “Statistical inference and confidence intervals”

**DUE: Monday, October 21, 2019 (by 11:59 PM)**

**This is an *individual* assignment; although you are welcome to discuss the assignment with your peers, all written work must be your own.**

**Purpose of the problem set**

In this problem set, you will demonstrate your ability to read and interpret the results of a simple regression. This will include interpreting the estimated regression coefficients (Unit 3), as well as getting some practice using the 95% Confidence Intervals associated with those coefficients (Unit 5).

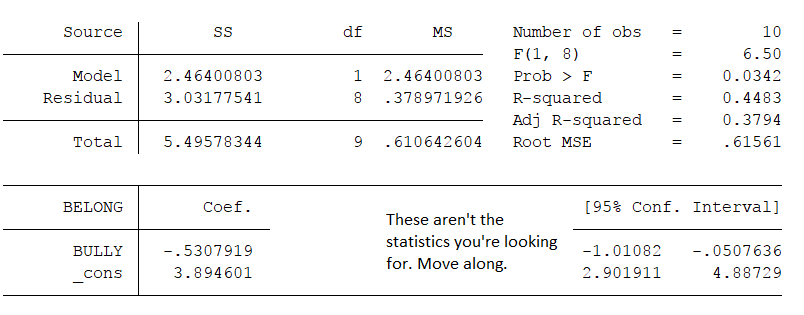
The data used in this assignment were collected by the Making Caring Common project. These data are *not* publicly available; given the nature of this assignment, we don’t want you to be able to access the data themselves. Sorry (not sorry)! They were obtained by surveying students in participating schools around the country. We will use them to explore the association between the sense of belonging students experience in school () and students’ experiences of bullying (). Both of these variables are *scale scores*. They are computed by averaging students’ responses to a number of individual items. The items measure how much students feel a sense of belonging and community in their schools (higher values indicate a stronger sense of belonging). The items measure how frequently students experience different sorts of bullying (higher values indicate more frequent bullying; a 0 corresponds to a student who reports never having experienced any of the kinds of bullying, a 4, which is not observed in this dataset, to a student who reported experiencing every kind of bullying multiple times per week).

**Formatting guidelines for the assignment**

Your write-up should be submitted as a Word document with one-inch margins all around, using Times New Roman 12-point font.

**The Scenario**

To answer the questions in this problem set, you will be asked to consider the following regression output from Stata (R users can ask for help making connections with R output). It depicts the findings of a simple linear regression with which we examined the question: *What is the relationship sense of belonging at school and frequency of bullying experiences among students in schools sampled by Making Caring Common?*



We have redacted the t-statistic and p-values for pedagogical purposes. We want you to focus your attention on the coefficients and 95% confidence intervals.  
  
**Note: These data were modified for the purposes of this problem set! We deleted a lot of observations!**

**1. Stating the relationship**

Assuming a linear relationship between sense of belonging at school () and frequency of experience of bullying (); write the **population equation** and the **fitted equation**. **Interpret the coefficients in the fitted equation**. When the outcome is a scale score, as in this example, the unit of measurement is a *point* or *scale point*.

**Population Equation is BELONGi = + BULLYi + i**

– Intercept which controls the height of the line in the population. Sense of belonging for a student whose observed frequency of bullying is zero in the population

- Slope between predictor and outcome which controls the steepness of the line in the population. How much we think two students who differ by 1 scale point on frequency of bullying will differ on sense of belonging in the population

- Residual value in the population. Extent to which a student’s observed sense of belonging is different from the predicted sense of belonging in population

BELONGi – Sense of belonging for a student i in the population

BULLY*i –* Frequency of bullying for a student i in the population

**Fitted Equation is = + BULLYi**

= 3.89 -0.53BULLYi

= Intercept in the sample. We predict that for a student whose observed frequency of bullying is zero will have a sense of belonging score of 3.89 scale point

= Slope between predictor and outcome in the sample. We estimate that 1 scale point difference in frequency of bullying is associated with -0.53 scale point difference in sense of belonging for students

**<Submit your equations and sentences>.**

**2. Plausible values for**

Give a **range of plausible population values** for the slope of the association between and .  
The range of plausible population values for the slope of association between sense of belonging at school (BELONG) and frequency of bullying (BULLY) is -1.01 to -0.05 scale points

**<Submit range of plausible values>.**

**3. Plausible values for students who have never experience bullying**

Give a **range of plausible mean values** of sense of belonging at school () for students who have never experienced bullying at school ().

The range of plausible mean values of sense of belonging at school (BELONG) and students who have never experienced bullying at school () is 2.90 to 4.89 scale points

**<Submit range of plausible values>.**

**4. Generalizing to the population**

Is the association between and statistically significant at the 0.05 level? How do you know?

We can identify the statistical significance of the association between BELONG and BULLY at the 0.05 level through the Confidence Interval (CI)

Confidence Interval (CI) of (range of plausible values for the true slope) in the association between BELONG and BULLYis from -1.01 to -0.05 scale points. Since, 0 is not a plausible value of (according to the 95% CI), we will reject H0 (null hypothesis) and conclude that we have evidence of an association between student’s sense of belonging at school (BELONG) and students’ experiences of bullying (BULLY) which is statistically significant at the 0.05 level

**<Submit your answer (Yes/No) and explanation>.**

**Challenge Question! (Just for fun…this won’t be graded!)**

If we regressed in the opposite direction ( on ) would the association be statistically significant at the 0.05 level? How do you know and why?

**YES,** p value will remain the same

**<Submit your answer (Yes/No) and explanation>.**

**Another Challenge Question! (So much ungraded fun!)**

Give a ballpark estimate of the standard error for the association.

Standard error of the association would be between -0.24

**<Submit your estimated standard error>.**

**Yet Another Challenge Question! (Why ask for a grade, it’s so much fun!)**

You may have noticed that this dataset contains only 10 observations. It was created by taking a (very small) random subset of the full dataset, which had over 9,000 observations. Someone might object that, because we have taken a subset of the complete dataset, our results are not trustworthy. What is wrong with this perspective?

We have chosen the 95% confidence interval which will hint us the true value in the population

**<Submit your estimated standard error>.**

**A final word**

Remember that this is an individual assignment, where you will be working on your own.

Please abide by the course and school policies on plagiarism. If someone else makes a contribution to your work, either advertently or inadvertently, you must recognize her or his contribution explicitly in your manuscript. If you do this consistently, you cannot be accused of plagiarizing (intentionally or otherwise) their contribution.

Please remember that this is not a competition amongst class members. Everyone’s work will be graded on its quality, not on a curve, and we hope that the class can function integrally as a team whose members help each other when needed or requested.