### Education 265:  Problem Set 2

**Due:** Friday, January 15, 2020 at 1:00 p.m. in the class Canvas folder (Assignments tab 🡪 Problem Set 2). Turn in

* Your answer sheet as a Word document (titled “**LastName\_PS2**”)
* Type your answers in **bold directly** under the questions
* Your do-file (titled “**LastName\_PS2\_Code.do**”).

**Objectives:** In this problem set, we will continue to explore the class dataset, this time looking at associations between pairs of variables. We will also look at group differences in those associations. At the end of this problem set, you should have a sense of:

1. how to descriptively examine the association between two variables
2. how to visually exmaine the association between two variables
3. how to test and interpret correlation between two variables.

Amount of time it took you to finish this problem set: **5 hrs**

1. Set up your do-file. Start by setting the path directory (use globals or just the cd command) so that all files are saved in the right folder. Call the data into Stata (*use for\_p2.dta, clear*).
2. Last week we looked at the test scores at kindergarten, but we haven’t recode missing values on these variables in order to properly use them.
   1. Looking back at your answer from Problem Set 1, question 4 f), what values are currently used to indicate a missing case?

**math1-3: no missing case**

**math4: NOT ASCERTAINED, NOT APPLICABLE**

**math5-7: NOT ASCERTAINED**

**read1: no missing case**

**read2-4: NOT ASCERTAINED, NOT APPLICABLE**

**read5-7: NOT ASCERTAINED**

**genknow1-4: NOT ASCERTAINED, NOT APPLICABLE**

**science5-7: NOT ASCERTAINED**

* 1. There are 21 test scores in total (*read1, math1, genknow1, read2…science7*). Recode missing values to “.” for the first 6 of them (i.e., *read1, math 1, genknow1, read2, math2, genknow2)* *manually.* And then, use the loop command to recode the remaining 15 variables. Copy and paste the code that you used to recode these variables into your answer sheet.

**replace read1 = . if read1 < 0**

**replace read2 = . if read2 < 0**

**replace math1 = . if math1 < 0**

**replace math2 = . if math2 < 0**

**replace genknow1 = . if genknow1 < 0**

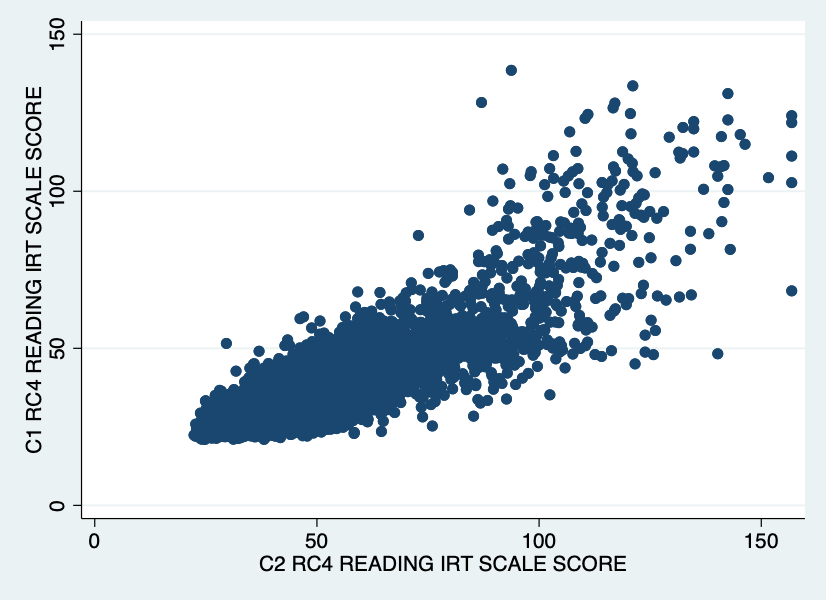
**replace genknow2 = . if genknow2 < 0**

**foreach var of varlist read3-science7{**

**replace `var' = . if `var' < 0**

**}**

* 1. What is the mean and standard deviation of math4? **61.26345, 18.09301**
  2. What is the mean and standard deviation of read3 for Black females in the 4th SES quintile? **56.7144, 16.30564**
  3. How many standard deviations apart is the student in the the 99th percentile from the average student in the lowest SES quintile on the 8th grade science test *(Hint: use the “sum, detail” command.)* **(103.49 - 71.43378) / 17.654 = 1.8158049167**

1. Let’s now look at the association between two test scores.
   1. Generate a scatter plot for *read1* and *read2*. Paste your plot below and describe it in two sentences. **Although the relation between them is positive, there isn't a clear pattern or a consistent coefficient in it. Relation looks more diverged on higher values compared to lower values, however it's still unclear because there is a lot of overlap between the dots in lower values.**
   2. Calculate the correlation between *read1* and *read2*, interpret your result. (Hint: what percent of the variation in *read2* is explained by its relation with *read1*?)

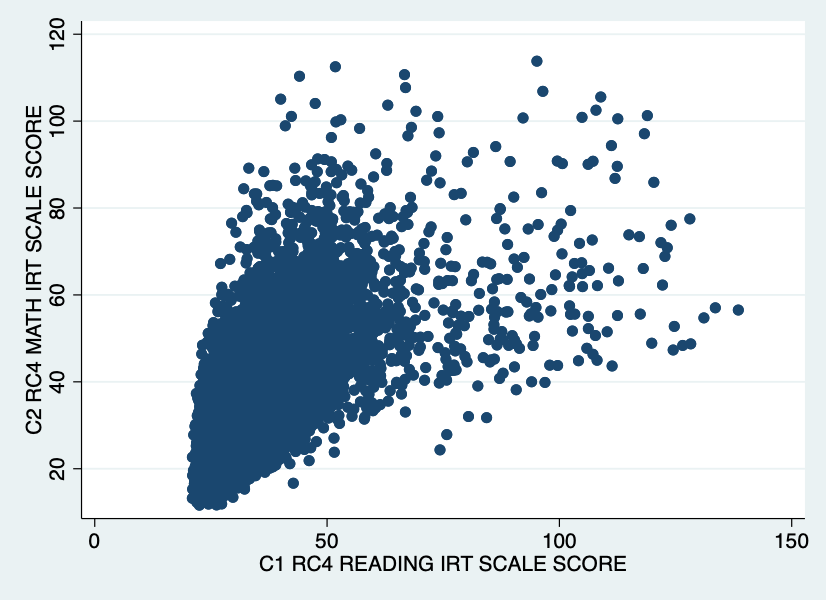
**0.8300**

**This shows the strength of the linear relation between read1 and read2 is high.**

* 1. Regress *read2* on *read1* (that is, predict *read2* using *read1*), explain the coefficient on *read1*, and then explain the R-squared in simple language.

**Coefficient: 1.148499, R-squared: 0.6889**

**The value of the R-squared shows that about 0.7 of the variance in read2 is explainable by the variance present in read1.**

* 1. Generate a scatter plot for *math2* and *read1*. Paste your plot below. Based on the plot did you expect the correlation between *math2* and *read1* to be greater or smaller than that between *read2* and *read1*? **Smaller, because the plot looks much more scattered comparing to the previous one.**

* 1. Calculate the correlation between *math2* and *read1*, does it align with your expectation in (d)? **0.6389, Yes**
  2. Calculate the correlation between *read2* and *math1* separately for boys or girls.

**Boys: 0.6703**

**Girls: 0.6823**

* 1. Calculate the correlation between *read1* and the *continuous measure of SES*, interpret your result. **0.3780, the statistical relationship between the reading scores and continuous measure of ses is not very strong.**
  2. (**BONUS**!) Predict *read1* using the *continuous measure of SES* ***seperately***for *each* of the SES quintiles using the loop command. Paste your code below and report the five slope coefficients.

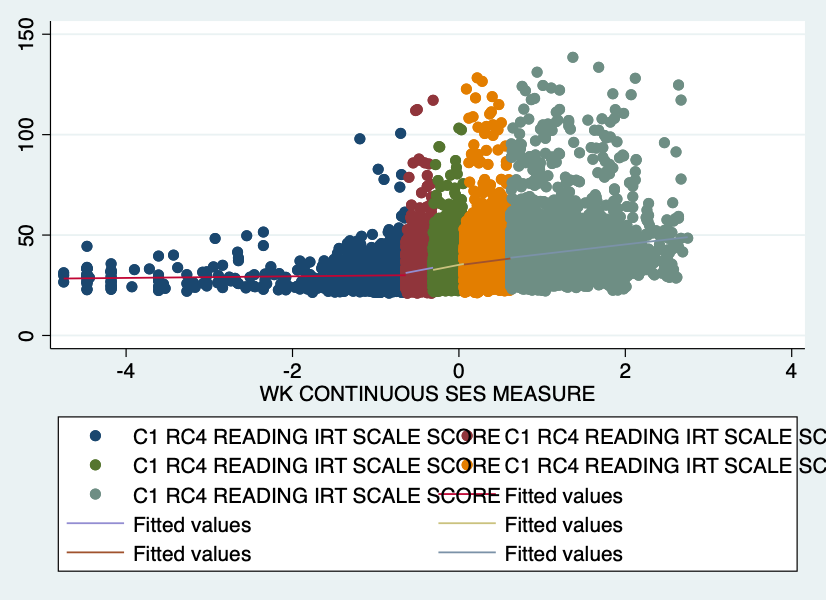
**forvalues i=1(1)5 {**

**display "SES `i'"**

**reg read1 wksesl if wksesq5 == `i'**

**}**

**Slopes: 0.3971564, 7.933094, 8.143944, 5.391485, 4.819744**

Then, produce a scatter plot of *read1* and the *continuous measure of SES* for each of the five groups *in one graph,* using different colors for each quintile and including a line of the relationship within each quintile (hint -- use the command “lfit”*).* Paste your scatter plot below.

1. Is the percent of students on free lunch in a school a significant predictor of reading gains between fall and spring of kindergarten? Interpret the coefficient on predictor variable.

**Coefficient of read1 without s2kflnch control: 1.148499**

**Coefficient of read1 with s2kflnch control: 1.150878**

**Coefficient of s2kflnch control: .0003339**

**As it is clear from the coefficients with and without free lunch variable control, the value has not changed noticeably and also the coefficient of the the free lunch variable itself is close to zero, so we can interpret that this is not a appropriate control variable.**

* 1. What happens to this relationship when you control for a student’s (continuous) SES? Use words and numbers to describe your answer. Why?

**Coefficient of read1: 1.127997**

**Coefficient of wksesl: .6909102**

**Although the coefficient of wksesl is significant, it seems that the relation between read1 and read2 is not independent from wksesl. The value of R-squared and std error has also not change noticeably but it also shows that ses and reading scores are somewhat realated.**

1. Skim through all the variables, pick a non-test-score variable that you think would be hihgly correlated with *read1*. Explain your reasoning in one sentence. **c1r4rpb3**

**assessments on reading abilities of kindergarten can be highly correlated with the reading scores.**

* 1. Calculate the correlation between the variable you picked above (make sure to recode missing value if needed) and *read1*, interpret your result in one sentence. **0.8765, There is a strong statistical relationship between how well kindergarteners do in reading and their assessment measures in Ending Sounds.**
  2. Pick a non-test-score variable that you think would be very weakly correlated with *read1*. Explain your reasoning in one sentence. **c1r4mpb8 reading score should not be correlated with the cognitive assessments of fractions.**
  3. Run correlation between the variable you picked above (make sure to recode missing value if needed) and *read1*, interpret your result in one sentence.

**0.0443, There isn't a strong statistical correlation between the reading score and the cognitive assessment of fractions.**

1. Save the dataset (*save week2hw\_clean.dta, replace*).