The dataset provided is an example file from Stock and Watson (2020). It is a dataset for the California Standardized Testing and Reporting (STAR)), containing information about test performance, school characteristics and student demographic backgrounds for 420 K-6 and K-8 districts in California during 1998 and 1999. Please answer the following questions by analyzing the provided dataset. You may use any kind of statistic software you are comfortable with. This assignment is due on **May 2nd**. Please send me a file of your answers along with a code file at my school email address.

Please download the dataset here:
https://www.dropbox.com/s/74b4ve2zsg9ybt7/Assignment_1%28STAR%29.dta?dl=0

In case you are not using STATA, below please find the labels for relevant variables:

Variable Names	**Labels**
dist_cod	District code
county	County
district	District
enrl_tot	Total enrollment
teachers	Number of teachers
computer	Number of computers
testscr	Average test score (=(read_scr+math_scr)/2)
comp_stu	Computers per student (=computer/enrl_tot)
expn_stu	Expenditures per student
str	Student teacher ratio (=enrl_tot/teachers)
avginc	District average income (in $1,000)
el_pct	Percent of English learners
read_scr	Average Reading Score
math_scr	Average Math Score

1. Please state the CLRM (classic linear regression model) assumptions 1-6. Briefly discuss the required assumptions for OLS estimators to be 1) unbiased; 2) BLUE; and 3) BUE.

2. Education economist have long believed that class size is a major determinant to student performance. Please perform a univariate regression of testscr (average test scores) on str (student teacher ratio) with homoskedasticity assumption. Please report the OLS estimates and standard errors for β_0 and β_1.

3. Please report the *t*-statistic and the *p*-value for β_1 estimates (coefficient for str). Please report the 95% confidence interval for it. Suppose we are testing against a null hypothesis $H_0: \beta_1 = 0$. Please state your statistical inference.
(e.g. under what confidence level, we can/cannot reject blablabla)

4. Please report the OLS estimates and standard errors for β_0 and β_1 under the heteroskedasticity assumption. How would your answer for question 3. change?

5. What is the R-square for the above regression? How do you interpret this number?

6. Please provide the summary statistics for str (mean, std., median, p25, p50, p75, min, max). Suppose we have an additional education district with a student teacher ratio of merely 5. What is your best guess of the average test score of this district according to the above univariate model?

7. Now, please include avginc (district average income) and expn_stu (expenditures per student) into the set of regressors. What happens to the estimated coefficient of str? How does R-square change accordingly? Please provide your intuitive explanation to the changes.

8. Lastly, please perform a joint test of the hypotheses of avginc=0 and expn_stu=0.