BUEC-333, Summer 2014
Hand-in assignment #1

Deadline: June 26, 12:30

Rules

1. Use the data that is provided on the course website, and read the accompanying data description.
2. Use **R** to answer the questions. Other software is not accepted.
3. Using an answer from a fellow student or a tutor will result in a score of 0 for the hand-in part of your grade. That means that you will be assigned 0 points for **both** hand-in assignments.
4. Acknowledge your sources. For example, if you use code from a website, then refer to that website in your answer.
5. The **open labs** are there for you to get help and ask questions: use them!
6. For each question:
    - Give a short written answer
    - Write (or type) the code that you used to get that outcome: **without** code, you get **0 points.** Exceptions: purely theoretical questions, such as questions 3 and 4.
    - Attach a hard copy of the R output supporting the answer

Setting

You are going to investigate the effect of class size on student performance. The background of the available data set can be found in the data description file on the course website. **Read the data description file!** For an introduction to the topic, see the slides and your notes from week 1.

- read_scr: Average reading score
- math_scr: Average math score
- testscr: Average of reading and math score
- enrl_tot: Total number of students enrolled

- teachers: Total number of teachers (FTE)
- str: Number of students per teacher
- calw_pct: Percentage of students in CalWorks public assistance program
- meal_pct: Percentage of students qualifying for a reduced price lunch
- avginc: Average income, 1000 $
- el_pct: Percentage of English learners
- computer: Number of computers
- comp_stu: Number of computers per student
- expn_stu: Expenditure per student

Questions

First, download the data set "testscores_california_1999.csv", from the course website. There, you also find a description of the data.

1. Install, then start RStudio.[1] Set your working directory to the folder where you saved your data. Now, load and inspect the data. You could, for example, cut and paste the following commands

```
#Read the data file and save the results
csdata <- read.csv("testscores_california_1999.csv")
#Have a quick look at the data
#Just to check that the data import went well
#First 6 lines
head(csdata)
#Last 6 lines
tail(csdata)
#Display a summary of the data
summary(csdata)
```

   From the output from that last command, you can read the maximum value for "comp_stu". What is that value? What does it mean? Are you surprised?

2. Give the sample mean, minimum, maximum, and sample standard deviation of math_scr.[2]

3. Is the sample standard deviation you computed under (2) an estimand, an estimator, or an estimate?

[1] More information about R and RStudio at this section of the course website: http://www.sfu.ca/~cmuris/2014-Summer-333/index.html#R

[2] Remember: Always provide a short, written, answer; write/type the R code you used to obtain the answer; and attach a hard copy of the R output that supports your written answer. Exception: purely theoretical questions.

4. Is the number you got a random variable?

5. Construct a 99% confidence interval for the mean of "math_scr". Also, construct a 59% confidence interval for the mean of "avg_inc".

6. For the previous question, what is the population you chose?

7. From question (5): What is the interpretation of the 99% confidence interval that you computed?

8. Estimate the coefficients in the linear regression of "math_scr" on "str". Include a constant. Report the coefficient estimates and standard errors, and interpret the coefficient estimates.

9. Construct a 95% confidence interval for the regression coefficient of "str".

10. You expect the coefficient of str to be negative. Formulate an appropriate null and alternative hypothesis; formulate a decision rule (use significance level of 5%); use R to compute the necessary values; draw the conclusion.

11. What is the interpretation of the conclusion of the test in (10)? What do you conclude about the effect of class size on student test performance?

12. There other variables that could have an effect on student test scores. List two or three such variables in this data set, and explain why they could have an effect on test scores, Do you expect the effects to be positive or negative?

13. Estimate the coefficients in the linear regression of "math_scr" on "str" and the variables you came up with under (12). Include a constant. Report the coefficient estimate and standard error for "str".

14. For bonus points, explain (or speculate about) the difference between your answer under (13) versus (8).

15. If you could gather additional data to answer this question in a better way, what data would you gather?