Applied Regression and Time Series Analysis
Homework 1: OLS Estimation

Jeffrey Yau

January 1, 2016

Instructions

The weekly assignment serves two purposes: (1) Extend the materials taught in the asynchronized materials; some new concepts or techniques are introduced in the weekly assignment. (2) Ensure that you have learned the concepts, techniques, theories, statistical models covered in a specific week. Below are some guidelines:

- **Submission: Submit 2 files. Missing one of the two files will result in a 50% reduction in grade.**
    1. A report (in pdf format) detailing your answers and all the steps to arrive at your answers
    2. A well-documented R-script, jupyter notebook, or Rmd file detailing all of the codes used to arrive at your answers.
- Late submission will not receive any credit.
- Answers need to be typed. Latex, which you will likely have to use in R markdown, is great for typesetting documents with mathematical symbols.
- All the steps used to arrive at your final answers need to be shown clearly. These steps are as important as the final answer.
- The final answer of each question needs to be very easy identified; the use of bold fonts, highlights, or circling will help.
- This is a group project. Form a group with 3 or 4 people.
- Although this is a group project, we encourage you to attempt all of the exercises before discussing with your teammates. Do not use the "division-of-labor" approach. Each of the students in a group is expected to make sufficient contribution to the lab. If any of your teammate does not make sufficient contribution, please contact your instructor.
- DO NOT copy and paste or even leverage on the solutions we gave to the students in previous semesters. Violation will be reported to the Director of the MIDS program and the Office that oversees UC Berkeley Academic Integrity. In any case, the lab has various subtle changes that make those answers not directly applicable.

Data

The file **birthweight_w271.RData** contains data from the 1988 National Health Interview Survey, which may have been modified by the instructors to test your proficiency. This survey is conducted by the U.S. Census Bureau and has collected data on individual health metrics since 1957. Like all surveys, a full analysis would require advanced techniques such as those provided by the R survey package. For this exercise, however, you are to treat the data as a true random sample. You will use this dataset to practice interpreting OLS coefficients.

Exercises

Question 1:
Load the birthweight dataset. Note that the actual data is provided in a data table named "data".

Use the following procedures to load the data:

Step 1: put the provided R Workspace birthweight_w271.RData in the directory of your choice.

Step 2: Load the dataset using this command: *load*("*birthweight.Rdata*")

Question 2:
Examine the basic structure of the data set using desc, str, and summary to examine all of the variables in the data set. How many variables and observations in the data?

These commands will be useful:

1. desc
2. str(data)
3. summary(data)

Question 3:
As we mentioned in the live session, it is important to start with a question (or a hypothesis) when conducting regression modeling. In this execrise, we are in the question: **"Do mothers who smoke have babies with lower birth weight?"**

The dependent variable of interested is *bwght*, representing birthweight in ounces. Examine this variable using both tabulated summary and graphs. Specifically,

1. Summarize the variable *bwght*: *summary*(*data*$*bwght*)
2. You may also use the quantile function: *quantile*(*data*$*bwght*). List the following quantiles: 1%, 5%, 10%, 25%, 50%, 75%, 90%, 95%, 99%
3. Plot the histogram of *bwght* and comment on the shape of its distribution. Try different bin sizes and comment how it affects the shape of the histogram. Remember to label the graph clearly. You will also need a title for the graph.

4. This is a more open-ended question: Have you noticed anything "strange" with the *bwght* variable and the shape of histogram this variable? If so, please elaborate on your observations and investigate any issues you have identified.

Question 4:
Examine the variable *cigs*, which represents number of cigarettes smoked each day by the mother while pregnant. Conduct the same analysis as in question 3.

Question 5:
Generate a scatterplot of *bwght* against *cigs*. Based on the appearance of this plot, how much of the variation in *bwght* do you think can be explained by *cigs*?

Question 6:
Estimate the simple linear regression of *bwght* on *cigs*. What coefficient estimates and the standard errors associated with the coefficient estimates do you get? Interpret the results. Note that you may have to "take care of" any potential data issues before building a regression model.

Question 7:
Now, introduce a new independent variable, *faminc*, representing family income in thousands of dollars. Examine this variable using the same analysis as in question 3. In addition, produce a scatterplot matrix of *bwght*, *cigs*, and *faminc*. Use the following command (as a starting point):

$library(car)$
$scatterplot.matrix(\ bwght + cigs + faminc, data = data2)$
Note that the car package is needed in order to use the scatterplot.matrix function.

Question 8:
Regress *bwgth* on both *cigs* and *faminc*. What coefficient estimates and the standard errors associated with the coefficient estimates do you get? Interpret the results.

Question 9:
Explain, in your own words, what the coefficient on *cigs* in the multiple regression means, and how it is different than the coefficient on *cigs* in the simple regression? Please provide the intuition to explain the difference, if any.

Question 10:
Which coefficient for *cigs* is more negative than the other? Suggest an explanation for why this is so.