Problem sets for BUEC 333
Part 3: Multiple linear regression and causality

I will indicate the relevant exercises for each week at the end of the Wednesday lecture. Numbered exercises are back-of-chapter exercises from Stock and Watson. Try to complete the exercises before going to the tutorials. In the tutorials, the TAs will help you if you have any difficulties.

Introduction to multiple linear regression

1. Exercises 6.1-6.4.
2. Exercise 6.5.
3. Exercise 6.6.
4. Exercise 6.7.
5. Exercise 6.8.
6. Exercise 6.9.
7. [Final, Summer 2014] Let W_i be an individual's wage, let $EDUC_i$ be their education, and let F_i be a dummy variable that equals 1 if and only if that person is female. Furthermore, let M_i be a dummy variable that equals 1 if and only if that person is male. Consider the following model:

$$W_i = \beta_0 + \beta_1 EDUC_i + \beta_2 F_i + \beta_3 M_i + \epsilon_i.$$

   a) What is the problem with using OLS to estimate the parameters in this regression equation?

   b) How do you solve this problem?

8. [Final, Summer 2014] Consider the following estimated regression equation that describes the relationship between a student's weight and height:

$$\widehat{WEIGHT} = 100 + 6.0\, HEIGHT$$

   a) A student has height 5. What is the regression's prediction for that student's weight?

   b) In the sample, the sample average of HEIGHT is 4. What can you say about the sample average for WEIGHT?

   Now, an additional variable is included, is ID, a student's SFU ID. Obviously, this is a nonsensical variable to include: it is not in any way related to a student's weight. The new estimated regression equation is

$$\widehat{WEIGHT} = 101.5 + 5.98\, HEIGHT + 0.02\, ID$$

   c) Someone's weight has nothing to do with their SFU ID. Still, the R^2 went up from 0.74 to 0.75. How is this possible?

   d) If the post office box number is not related to a student's weight, should the estimated coefficient not be equal to 0? How could it be that it is 0.02?

Topics in linear regression

1. Exercise 5.2, skip (b).
2. Exercise 5.5, (a) and (c).
3. Exercise 5.7
4. Exercise 5.10.
5. [Final, Summer 2014] Let D_i be a dummy variable. Consider the model that consists of the equation
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 D_i X_i + u_i$$
and the standard OLS assumptions
   a) Draw a graph to visualize this model.
   b) What is the interpretation of β_3?
6. Exercise 7.1, page 242.
7. Exercise 7.2, page 242.
8. Exercise 7.6, page 244.
9. During the lectures, we have discussed that, in the model
$$\log Y_i = \beta_0 + \beta_1 \log X_i + u_i$$
you can interpret β_1 as approximately measuring the expected percentage change in Y_i given a one-percent change in X_i, c.p. Similarly, we know that in the model
$$\log Y_i = \beta_0 + \beta_1 X_i + u_i,$$
we can interpret $\beta_1 \times 100\%$ as the expected percentage change in Y_i for a one-unit change in X_i. What is the interpretation of β_1 (or $\beta_1 \times 100\%$) in the model
$$Y_i = \beta_0 + \beta_1 \log X_i + u_i,$$
known as the linear-log model?
10. Exercise 8.7 (SW page 300).
11.

Instrumental variables

1. TBA

Panel data

1. (Final, Summer 2014). Consider the example used in the chapter on panel data. We have a panel data set on $n = 48$ U.S. states during $T = 7$ periods, from 1982 up to and including 1988. The total number of observations is 336.
   a) Is this a balanced panel? Explain.

b) For each state, in each time period, let Y_{it} denote the number of annual traffic deaths per 10000 in the population. Let X_{it} denote the beer tax in 1988 U.S. dollars. Temporarily ignore the data after 1982, so that we have a cross-section of 48 states. The estimated regression line gives

$$\hat{Y}_{i,1982} = 2.01 + 0.13X_{i,1982}.$$

If the Least Squares assumptions hold for this regression, how would you interpret the 0.13? Include in your answer: "tax on beer".

c) Alternatively, we can use fixed effects regression to estimate the effect fixed effects regression line is

$$\hat{Y}_{i,t} = \hat{\alpha}_i - 0.66X_{i,t}. \quad (0.29)$$

How would you interpret the -0.66?

d) Consider the results for the fixed effects regression. Do you think that the Least Square assumptions hold, i.e. do you believe that the 0.13 in the first result comes from an unbiased estimator? if YES: explain what causes the difference between 0.13 and -0.66. If NO: explain why the Least Square assumptions are unlikely to hold.

2. This question is about the code that we used during the lecture of Wednesday, July 17. You can find it on the course website or directly through [this link]. We are going to investigate the output of the commands on lines 236 and 250. Remember that the data set we are using is a random sample of 935 "young men" in the U.S. in 1980. A list of variables and descriptions is

```
1  [...]
2  # 1. wage                    monthly earnings
3  # 2. hours                   average weekly hours
4  # 3. IQ                      IQ score
5  # 4. KWW                     knowledge of world work score
6  # 5. educ                    years of education
7  # 6. exper                   years of work experience
8  # 7. tenure                  years with current employer
9  # 8. age                     age in years
10 [...]
11 # 12. urban                    =1 if live in SMSA
12 # 13. sibs                     number of siblings
13 # 14. brthord                  birth order
14 # 15. meduc                    mother's education
15 # 16. feduc                    father's education
16 [...]
```

Here is the first regression and its output:

```
1  > summary(lm(IQ~brthord,data=wageData.2))
2
3  [...]
4
5  Coefficients:
6              Estimate Std. Error t value Pr(>|t|)
7  (Intercept) 105.6634     0.8699 121.470  < 2e-16 ***
8  brthord      -1.6640     0.3129  -5.318 1.34e-07 ***
9  ---
10
11 [...]
```

a) What is the question we are trying to answer?

b) What results would you draw from this output, if you knew that all the OLS assumptions were satisfied?

c) Assumption (1) is unlikely to hold. Why?

d) To solve that problem, we are going to include additional regressors, namely the education of both parents ($feduc$ and $meduc$).

   i. Could you make an argument that would support $\text{cov}(feduc, brthord) \neq 0$ or $\text{cov}(meduc, brthord) \neq 0$?

   ii. Can you make an argument to support the statement "The regression coefficients on $feduc$ and $meduc$ are likely to be different from zero.

e) The output from that regression follows:

```
> summary(lm(IQ~brthord+feduc+meduc,data=wageData.2))

[...]

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  85.4160     2.6170  32.639  < 2e-16 ***
brthord      -1.0544     0.3712  -2.840 0.004647 **
feduc         0.9780     0.1970   4.963 8.83e-07 ***
meduc         0.8602     0.2333   3.687 0.000245 ***
---

[...]
```

If the OLS assumptions are valid, how do you interpret these results?

f) Do you believe that the conclusion under (e) is correct? Why / why not?