Problem Set 2

Due on April 9th 17:00

Please note:

- All question shall be relevant from class materials but not every part is taught in class. In other word, some questions are a little "innovative" so to train your programming and problem solving skill.
- It is normal to get stuck. Keep working. Go back to class code and exercise in the textbook for practice. Discuss, try, and retry.
- Submit your answers (in Microsoft Word or PDF format) and your code. Your answer shall be well written. Graph and Table shall be well-formatted. Your code shall be easy for TA to run and check. Your grade will be affected if your code does not provide proper output, or it is confusing so that TA cannot run it.
- **Use proper regression model to analyze the empirical questions. Report the result both by words and regression results. Use Table and Figures to report the result wherever necessary.**

1. In this exercise, we use the data set stock_price.csv to study prediction of company's stock price. The variables are

company	name of the company
peratio	price-equity ratio (dependent variable)
risk	a measure of a company's riskiness, a high measure implies a high risk
earn	the median percentage earnings growth rate in the past five years
div	the median percentage dividend growth rate in the past five years

a. Generate the log of price-equity ratio as a new dependent variable *logpe*. Compile a table of summary statistics.

b. Run six linear regressions by OLS.

$$
\begin{aligned}
\text{Model I:} \quad & peratio = \beta_1 + \beta_2 div + e \\
\text{Model II:} \quad & peratio = \beta_1 + \beta_2 div + \beta_3 earn + e \\
\text{Model III:} \quad & peratio = \beta_1 + \beta_2 div + \beta_3 earn + \beta_4 risk + e \\
\text{Model IV:} \quad & logpe = \beta_1 + \beta_2 div + e \\
\text{Model V:} \quad & logpe = \beta_1 + \beta_2 div + \beta_3 earn + e \\
\text{Model VI:} \quad & logpe = \beta_1 + \beta_2 div + \beta_3 earn + \beta_4 risk + e
\end{aligned}
$$

Report all six regression results in one table. (coefficient, standard error, n, adjusted R^2).

c. For all six regression specifications above, plot six scatter-plots which shows both the data and fitted value. Use *div* as x-asix and adjusted R^2 as diagram title.

d. Conduct two F-tests. One compares model I with III, the other compares model IV and VI. Report the null hypotheses, test statistics, p-values, and conclusion.

2. Consider the data set cigarette.csv. It contains information of cigarette consumption in 2000 and 2006 of 45 states in U.S. Dummy variable *TAX* records a cigarette tax increase in 2003 happened in some states as the . Use an appropriate estiamtion technique to determine the impact of the cigarette tax increase on the consumption of cigarettes.

3. Consider the data set hospital_choice.csv with the following variables:

Ducla	dummy for whether patient *i* goes to ULCA medical center
distance	the distance from patient *i*'s home to UCLA medical center
income	the income of patient *i* (thousand USD)
old	dummy for whether patient *i* is older than 75

This is a survey conducted by UCLA medical center to study what kind of patient goes to UCLA medical center for treatment. Use a proper model to estimate the marginal effect of income to whether one goes to UCLA and comment the result

4. Consider the data set drug_price.csv. This is a data set about the price information of a medicine produced by a pharmaceutical company. This drug is sold in 32 countries and the Competition Commission is investigating whether the company practice international price discrimination. Here are the variables:

p.r	price of the drug in country *i* relative to U.S. price
cv	consumption volume of the drug in country *i*
cv.r	overall consumption volume of drugs in country *i* relative to U.S
GDP.r	per capita GDP of country *i* relative to U.S.
p.control	dummy for price control in country *i*
p.comp	dummy for price competition is encouraged in country *i*
patent	dummy for whether the drug is protected by patent in country *i*

Investigate whether the pharmaceutical company try to sell the drug for higher price in countries with higher per capita GDP. Use several regression results and determine the best linear regression model.

5. Consider the data set woman_educ.csv. with the following variables:

inlf	=1 if in lab force, 1975
hours	hours worked, 1975
kidslt6	# kids < 6 years
kidsge6	# kids 6-18
age	woman's age in yrs
educ	years of schooling
wage	estimated wage from earn, hrs
repwage	representative wage at interview in 1976
hushrs	hours worked by husband, 1975
husage	husband's age
huseduc	husband's years of schooling
huswage	husband's hourly wage, 1975
faminc	family income, 1975
mtr	fed. marginal tax rte facing woman
motheduc	mother's years of schooling
fatheduc	father's years of schooling
unem	unemployment rate in county of resident
city	=1 if live in a metropolitan city
exper	actual labor narket experience
nwifeinc	(faminc - wage*hours)/1000
lwage	log(wage)
expersq	exper^2

The goal of this empirical exercise is to study return to education of married working woman. The goal is to estimate the coefficient of *educ* in the following model

$$\log(wage) = \beta_1 + \beta_2 educ + \beta_3 x_3 + \cdots + \beta_K x_K + \varepsilon.$$

As we mentioned in class, education is a person's endogenous choice. So *educ* is correlated with a person's intrinsic ability like IQ. Because *wage* is also related to IQ, directly regress *wage* on *educ* will lead to endogeneity.

a. Do some basic OLS regressions and record the results.

b. Explain which variable(s) are potentially good instrument for *educ* and explain why they are valid.

c. Use the instrument you selected above, conduct a first-stage regression to show the instrument is strong enough.

d. Show the two-stage-least-square estimation of return to education. Be careful to use the right standard errors.

6. Consider the data set jtrain.csv with the following variables:

hrsemp	total training hours / total trained employee
grant	= 1 if received grant
employ	# employees at plant
d88	= 1 if year = 1988
d89	= 1 if year = 1989

The data covers three years: 1987, 1988, 1989. Run a "pooling" regression, a fixed-effect regression, and a random-effect regression. Interpret the results.

7. Use Monte Carlo simulation to show the validity of t-test.

Consider the following data generating process: $x_i \sim N(2, 2^2)$, $e_i \sim N(0,1)$, $\beta_1 = 1$, $\beta_2 = 0.1$, $y_i = \beta_1 + \beta_2 x_i + e_i$. You can shall the validity of t-test by the following procedure:

a. Set $n = 100$, generate $S = 1000$ samples of x_i and y_i.

b. For each simulated sample, use lm() function to find the OLS estimated coefficeint $\hat{\beta}_2$, its standard error $SE(\hat{\beta}_2)$, and t-test statistic. Plot the density of t-test statistic.

c. Use R to find the 0.025 and 0.975 ts quantile of student-t distribution with $n-1$ degree of freedom. What is the rejection range of the $t$test of the signifiance of $\hat{\beta}_2$

d. Calculate among the S t-statistics, how many of them fall into the rejection range? What is the power of the test.

e. Increase sample size from 100 to 1000, do the same exercise above. Does the power of the test increase?