Problem Set 3

Due on April 23

Please note:

- Submit your answers in a MS Word or PDF file together with the R code to canvas. Use well-complied tables and figures wherever necessary. Besides correctness, the clarity of your word file and R code will also affect grading.

1. Consider the data set cigarette.csv. It contains information of cigarette consumption in 2000 and 2006 of 45 states in U.S. The dummy variable *TAX* records a cigarette tax increase in 2003 happened in some states.

a. Fill the following table

	Average cigarette consumption		
	Before	After	After−Before
Control			
Treatment			
Treatment−Control			

b. Formulate a regression specification to find the causal effect of cigarette tax on consumption. Use regression table to report it.

2. Replicate the results in *The Colonial Origins of Comparative Development: An Empirical Investigation*. The data is taken from https://economics.mit.edu/faculty/acemoglu/data/ajr2001. We will focus on the following variables

Name of Variable	Note
Shortnam	Shortened name of the country
logpgp95	Log GDP per capita, PPP, 1995
avexpr	Average Expropriation Risk 1985-95
logem4	Log of settler mortality
baseco	Indicator for base sample
rich4	Indicator of neo-Europes

a. Open the dataset colony.dta. Construct two subsets of the data. One is the "base" sample with 64 observations. The other is the "base sample without neo-Europes".

b. Reproduce Figures 1 and 2 in the paper.

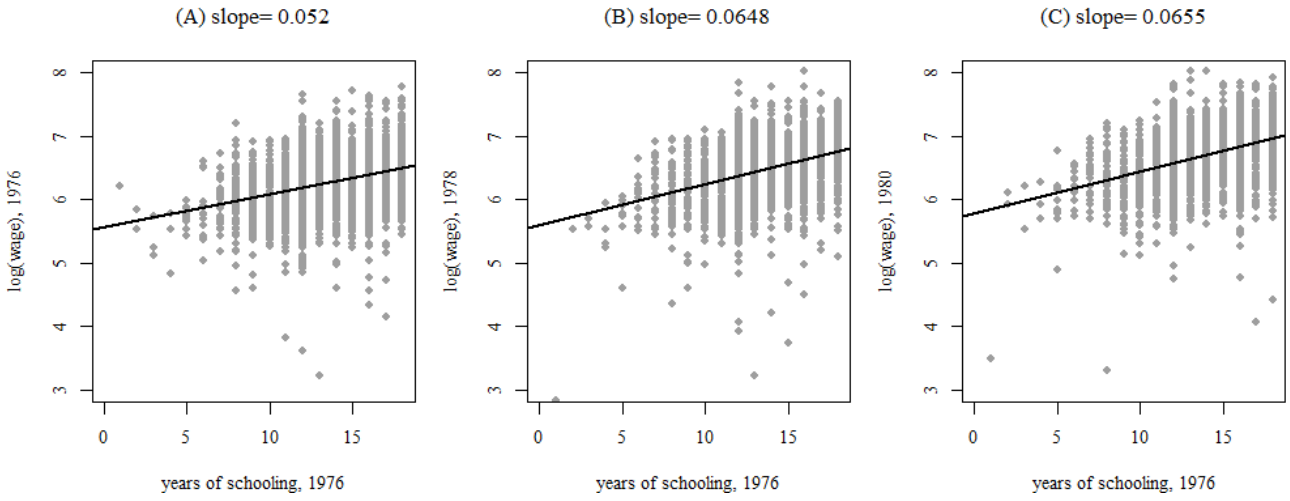
c. Reproduce Table 4, colume (1) to (4). Keep 3 digitis for decimal numbers.

3. The dataset card.csv is a partial dataset taken from David Card (1995) "Using Geographic Variation in College Proximity to Estimate the Return to Schooling".

It studies the classical question of returns to schooling using indicators of whether a person lives near a college as instrumental variables (IV).

id	sequential id runs from 1 to 5225
lwage76	log hourly wage in 1976
lwage78	log hourly wage in 1978
lwage80	log hourly wage in 1980
nearc2	grew up near 2-yr college
nearc4	grew up near 4-yr college
nearc4a	grew up near 4-yr public college
nearc4b	grew up near 4-yr private college
educ	education in 1976
age	age in 1976
kww	the kww score (psycnet.apa.org/record/1975-12557-001)
iq	a normed iq score
black	1 if black
smsa	in smsa (metropolitan city) in 1976
south	lived in south in 1966
famed	mom-dad education class 1-9

a. Reproduce this Figure. Note that the title of each diagram is the slope of the corresponding linear fitted line.



b. Run three OLS regressions using *lwage76, lwage78, lwage80* as the dependent variable, respectively. The regression use $educ, kww, iq, age, black, smsa, south$ and the categorical variable *famed* as regressors. Report the results in a regression table with the following format.

Dep Var.	*lwage76*		*lwage78*		*lwage80*	
	(1)	(2)	(3)	(4)	(5)	(6)
	OLS	IV	OLS	IV	OLV	IV
educ						
⋮						

c. Based on the above OLS model specification, run several first-stage regressions for potential IVs *nearc2, nearc4, nearc4a, nearc4b.* Use *t test* results to determine which two IVs are the most correlated with *educ.*

d. Based on the specification in part (b), run three IV two-stage-least-square regressions using the two strongest IVs you found in part (c). Report the result side by side with corresponding OLS results (refer to the table format above). Make some comments about the results.

e. In wage data, there are many missing values (NA). If a person has NA in his wage, it means he is not participating the labor force at that time (unemployed or not working).

Create three new indicator variables, *work76*, *work78*, *work80*, that indicate whether a person participates the labor force (has wage record). Use logit and probit model to see how education affect the probability of labor participation. Include the same control variables as in part (b). Report your results in a table.

4. In this exercise, we are going to study the property of two-stage-least-square (TSLS) estimator using instrumental variables (IV). Use the following data generating process to generate a sample for $i = 1, 2, ..., n$.

True coefficient values: $\beta_1 = 1$, $\beta_2 = 0.5$, $\beta_3 = 2$.

Error terms:

$$\begin{pmatrix} e_i \\ u_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}\right).$$

Right-hand side variables: $z_{1i} \sim$ i.i.d.$N(1, 1^2)$, $z_{2i} \sim$ i.i.d.Gamma(2), $z_{3i} \sim$ i.i.d.$N(2, 0.5^2)$, $x_{1i} =$ i.i.d.$N(1, 2^2)$.

$$x_{2i} = z_{1i} + 0.5z_{2i} + 0.01z_{3i} + u_i$$

$$y_i = \beta_1 + \beta_2 x_{1i} + \beta_3 x_{2i} + e_i$$

Store the data generated from each simulation as a $(n \times 8)$ matrix or data frame $\{y_i, x_{1i}, x_{2i}, z_{1i}, z_{2i}, z_{3i}, e_i, u_i\}_{i=1}^n$.

a. Let sample size n takes value $n \in \{100, 200, 300, ..., 5000\}$. Compute $\hat{\beta}_{OLS} = (\hat{\beta}_{1,OLS}, \hat{\beta}_{2,OLS}, \hat{\beta}_{3,OLS})$ for each case. Graphically show that OLS estimator is inconsistent.

b. By the setting of DGP, x_2 suffers endogeneity problem. Each of z_1, z_2, z_3 can serve as a valid instrument. Set $n = 100$, use each one of them to perform TSLS estimation. You shall report the coefficient estimates and standard errors. Instead of hand-programing the estimator, use "ivreg" function in "AER" package so that the standard errors are correct.

[Hint: compute three sets of estimates. Use z_1, compute $\hat{\beta}^{z_1}_{TSLS} = (\hat{\beta}^{z_1}_{1,TSLS}, \hat{\beta}^{z_1}_{2,TSLS}, \hat{\beta}^{z_1}_{3,TSLS})$ and their standard errors. Then do the same thing for z_2 and get $\hat{\beta}^{z_1}_{TSLS}$; for z_3, get $\hat{\beta}^{z_3}_{TSLS}$.]

c. Set $n = 100$, compute another set of TSLS estimation by using all three IVs (z_1, z_2, z_3). Record coefficient estimate $\hat{\beta}^{z_1,z_2,z_3}_{TSLS}$ and standard errors. Report your results in a table like the following. Which regression results you think is the best?

	OLS	TSLS z_1	TSLS z_2	TSLS z_3	TSLS z_1, z_2, z_3
$\hat{\beta}_1$	?	?	?	?	?
$SE(\hat{\beta}_1)$	?	?	?	?	?
$\hat{\beta}_2$	?	?	?	?	?
$SE(\hat{\beta}_2)$	?	?	?	?	?
$\hat{\beta}_3$	?	?	?	?	?
$SE(\hat{\beta}_2)$	?	?	?	?	?

d. Comment on the estimation result of using only z_3 as IV.