Computational and Empirical Methods

Assignment VII

Instructor: Prof. Gunnar Heins

Teaching Assistant: Pedro de Sousa Almeida

Student: C.H. (Chenhui Lu)

UFID: 76982846

Assignment 7

Instructor: Prof. Gunnar Heins

Teaching Assistant: Pedro de Sousa Almeida

Student: C.H.(Chenhui Lu)

UFID: 76982846

==========

```
# Empty Environment
rm(list = ls())
```

```
#library
library("data.table")
```

```
## Warning: package 'data.table' was built under R version 4.3.3
```

```
library("stargazer")
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics
Tables.
```

```
##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```
library("AER")
```

```
## Warning: package 'AER' was built under R version 4.3.2
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
## Loading required package: lmtest
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:data.table':
##
##     yearmon, yearqtr
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Loading required package: sandwich
```

```
## Loading required package: survival
```

```
## Warning: package 'survival' was built under R version 4.3.2
```

```
library("ggplot2")
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

```
library("ggrepel")
```

```
## Warning: package 'ggrepel' was built under R version 4.3.3
```

```
library("maps")
library("dplyr")
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:car':
##
##     recode
```

```
## The following objects are masked from 'package:data.table':
##
##     between, first, last
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library("mapproj")
```

1 OLS Regressions [35 points]

```
knitr::include_graphics("/Users/terrylu/Desktop/UF/fall/R and Matlab/R/After 2nd Midterm/Assignment/Problem_Set_7/HW7/1.1.png")
```

Suppose you want to estimate how price-elastic consumers of cars are, i.e. how much the price matters for a manufacturer's market share. For that purpose, suppose you run a regression

$$s_i = \alpha + \beta \cdot p_i + \epsilon_i \tag{1}$$

where s_i is firm i's market share and p_i the price that is charged by the firm.

```
knitr::include_graphics("/Users/terrylu/Desktop/UF/fall/R and Matlab/R/After 2nd Midterm/Assignment/Problem_Set_7/HW7/1.11.jpeg")
```

$$cov(S_i, P_i) = cov(\alpha + \beta P_i + \epsilon_i, P_i)$$

$$= cov(\alpha, P_i) + cov(\beta P_i, P_i) + cov(\epsilon_i, P_i)$$

$$= 0 + \beta\, var(P_i) + cov(\epsilon_i, P_i)$$

$$\beta = \frac{cov(S_i, P_i)}{var(P_i)} - \frac{cov(\epsilon_i, P_i)}{var(P_i)}$$

$$\Rightarrow \beta_{as} = \frac{\widehat{cov}(S_i, P_i)}{\widehat{var}(P_i)} \qquad \text{since. } E[cov(\epsilon_i, P_i)] = 0.$$

```
knitr::include_graphics("/Users/terrylu/Desktop/UF/fall/R and Matlab/R/After 2nd Midterm/Assignment/Problem_Set_7/HW7/1.2.jpeg")
```

2. The OLS estimator in this case estimates

$$\beta^{OLS} = \frac{\widehat{cov}(s_i, p_i)}{\widehat{var}(p_i)}$$

where $\widehat{cov}(s_i, p_i)$ and $\widehat{var}(p_i)$ are consistent estimators of the covariance and variance:

$$plim\ \widehat{cov}(s_i, p_i) = cov(s_i, p_i)$$
$$plim\ \widehat{var}(p_i) = var(p_i).$$

Show that

$$plim\ \beta^{OLS} = \beta$$

if $cov(\epsilon_i, p_i) = 0$.[1]

```
knitr::include_graphics("/Users/terrylu/Desktop/UF/fall/R and Matlab/R/After 2nd Midterm/Assignment/Problem_Set_7/HW7/1.22.jpeg")
```

As shown above: $cov(s_i, p_i) = \beta var(p_i) + cov(\varepsilon_i, p_i)$.

If $cov(\varepsilon_i, p_i) = 0$, $cov(\varepsilon_i, p_i) = \beta var(p_i)$.

$$plim\ \hat{\beta}^{OLS} = \frac{plim\ \widehat{cov}(s_i, p_i)}{plim\ \widehat{var}(p_i)} = \frac{cov(\varepsilon_i, p_i)}{var(p_i)}$$

since $\begin{cases} plim\ \widehat{cov}(s_i, p_i) = cov(\varepsilon_i, p_i) \\ plim\ \widehat{var}(p_i) = var(p_i) \end{cases}$

$$= \frac{\beta\, var(p_i)}{var(p_i)} = \beta.$$

Thus, $plim\ \hat{\beta}^{OLS} = \beta$.

```
knitr::include_graphics("/Users/terrylu/Desktop/UF/fall/R and Matlab/R/After 2nd Midterm/Assignment/Problem_Set_7/HW7/1.3.png")
```

3. Now suppose that the true relationship in the data is actually

$$s_i = \alpha + \beta \cdot p_i + \gamma \cdot q_i + \epsilon_i \tag{2}$$

where q_i is the quality of firm i's car. Would you expect q_i to be correlated with p_i? Why or why not?

Yes, I do believe that qi is correlated with pi. Since in the reality, higher priced car alwasy implies higher quality, otherwise the comsumers will not buy it. One the other hand, if a firm want to produce a new model which is high quality, it always has higher price due to the higher cost.

```
knitr::include_graphics("/Users/terrylu/Desktop/UF/fall/R and Matlab/R/After 2nd Midterm/Assignment/Problem_Set_7/HW7/1.4.jpeg")
```

4. Since quality is to some extent a subjective measure, it is often difficult to measure or not available. Suppose because of that you still estimate regression (1), even though the true relationship in the data is given by (2). Show that this will result in ommited-variable bias if $cov(q_i, p_i) \neq 0$

Real: $S_i = \alpha + \beta \cdot P_i + \gamma \cdot q_i + \varepsilon_i$

Estimate: $S_i = \alpha + \beta \cdot P_i + u_i$ $\quad u_i = \gamma \cdot q_i + \varepsilon_i$

In estimate model situation.

$$cov(S_i, P_i) = cov(\alpha + \beta P_i + u_i, P_i)$$

$$= \beta var(P_i) + \underbrace{cov(u_i, P_i)}_{\text{mistakenly} = 0}$$

$$\beta^{ols} = \frac{cov(S_i, P_i)}{var(P_i)}$$

In real situation,

$$cov(S_i, P_i) = cov(\alpha + \beta \cdot P_i + u_i, P_i)$$

$$= cov(\alpha, P_i) + \beta var(P_i) + cov(\gamma q_i + \varepsilon_i, P_i)$$

$$= \beta var(P_i) + \gamma cov(q_i, P_i) + \underbrace{cov(\varepsilon_i, P_i)}_{=0}$$

$$\beta^{Real} = \frac{cov(S_i, P_i)}{var(P_i)} - \frac{\gamma cov(q_i, P_i)}{var(P_i)}$$

$\Rightarrow \beta' \neq \beta^2$, since $cov(q_i, P_i) \neq 0$ & $\gamma \neq 0$. $\Rightarrow$ ommited variable bias.

```
knitr::include_graphics("/Users/terrylu/Desktop/UF/fall/R and Matlab/R/After 2nd Midt
erm/Assignment/Problem_Set_7/HW7/1.5.jpeg")
```

5. Suppose you actually observe market shares and prices at different points in time t and $t+1$

$$s_{i,t} = \alpha + \beta \cdot p_{i,t} + \gamma \cdot q_i + \epsilon_{i,t} \tag{3}$$

$$s_{i,t+1} = \alpha + \beta \cdot p_{i,t+1} + \gamma \cdot q_i + \epsilon_{i,t+1} \tag{4}$$

but suppose that the quality of the car is actually the same for both periods. Because of that you now regress the change in the market share $\Delta s_i = s_{i,t+1} - s_{i,t}$ on the change in the price $\Delta p_i = p_{i,t+1} - p_{i,t}$. Show that this will allow you to estimate β consistently.

(3) − (4) :

$\Delta S_i = \beta(P_{it+1} - P_{it}) + \varepsilon_{it+1} - \varepsilon_{it}$. make $\varepsilon_{it+1} - \varepsilon_{it} = \Delta\varepsilon_i$

$= \beta \cdot \Delta P_i + \Delta\varepsilon_i$

$cov(\Delta S_i, \Delta P_i) = cov(\beta\Delta P_i + \Delta\varepsilon_i, \Delta P_i)$

$= \beta\, var(\Delta P_i) + cov(\Delta\varepsilon_i, \Delta P_i)$ assume = 0 in OLS

$\beta^{OLS} = \frac{cov(\Delta S_i, \Delta P_i)}{var(\Delta P_i)}$

$plim\ \beta^{OLS} = \frac{plim\ cov(\Delta S_i, \Delta P_i)}{plim\ var(\Delta P_i)} = \frac{\beta\, var(\Delta P_i)}{var(\Delta P_i)} = \beta^{Real}$.

```
knitr::include_graphics("/Users/terrylu/Desktop/UF/fall/R and Matlab/R/After 2nd Midterm/Assignment/Problem_Set_7/HW7/1.6.jpeg")
```

6. Would you be able to identify β in the regression above if prices do not change over time? Why or why not?

No. If $P_{t+1} = P_t$.

$\Delta S_i = \beta \cdot 0 + \Delta \varepsilon_i$

$\Delta S_i = \Delta \varepsilon_i$.

On the other hand.

$\text{plim}\, \beta^{OLS} = \frac{\beta \, \text{var}(\Delta P_i)}{\text{var}(\Delta P_i)}$ is meaningless,

since $\Delta P_i = 0$

```
knitr::include_graphics("/Users/terrylu/Desktop/UF/fall/R and Matlab/R/After 2nd Midterm/Assignment/Problem_Set_7/HW7/1.7.jpeg")
```

7. Suppose that quality is actually not constant over time and so you have to go back to running regression (1). But suppose you now consider using an instrumental variables (IV) approach to estimate β. IV requires you to find a variable that is correlated with your regressor of interest (i.e. here the price) but not with the error term (including quality).

   Suppose you have a measure of costs z_i and use it as an instrument for p_i and suppose it actually isn't correlated with neither q_i or ϵ_i. Show that the instrumental variables estimator, which is defined as

$$\beta^{IV} = \frac{\widehat{Cov}(z_i, s_i)}{\widehat{Cov}(z_i, p_i)},$$

   where $\widehat{Cov}(z_i, s_i)$ is a consistent estimator of $Cov(z_i, s_i)$ and $\widehat{Cov}(z_i, p_i)$ a consistent estimator of $Cov(z_i, p_i)$, will be consistent (i.e. $plim\ \beta^{IV} = \beta$) if

$$Cov(z_i, \epsilon_i) = 0 \text{ and } Cov(z_i, p_i) \neq 0.$$

$S_i = \alpha + \beta \cdot P_i + \varepsilon_i$

$P_i = \gamma + \delta \cdot Z_i + u_i$

$Cov(z_i, S_i) = cov(z_i, \alpha + \beta \cdot P_i + \varepsilon_i)$

$= cov(z_i, \alpha) + \beta\, cov(z_i, P_i) + cov(z_i, \varepsilon_i)$ ($cov(z_i, \alpha) = 0$)

$= \beta\, cov(z_i, P_i) + cov(z_i, \varepsilon_i)$

$\beta = \frac{cov(z_i, S_i)}{cov(z_i, P_i)} - \frac{cov(z_i, \varepsilon_i)}{cov(z_i, P_i)}$

$\Rightarrow \beta^{IV} = \frac{\widehat{cov}(z_i, S_i)}{\widehat{cov}(z_i, P_i)}$ since assume $cov(z_i, \varepsilon_i) = 0$.

since given, $cov(z_i, \varepsilon_i) = 0$, $cov(z_i, P_i) \neq 0$. $cov(z_i, S_i) = \beta cov(z_i, P_i)$,

$plim\, \beta^{IV} = plim \frac{\widehat{cov}(z_i, S_i)}{\widehat{cov}(z_i, P_i)} = \frac{plim\, \widehat{cov}(z_i, S_i)}{plim\, \widehat{cov}(z_i, P_i)} = \frac{cov(z_i, S_i)}{cov(z_i, P_i)} = \frac{\beta cov(z_i, P_i)}{cov(z_i, P_i)} = \beta^{Real}$

Thus. $plim\, \beta^{IV} = \beta^{Real}$. If $cov(z_i, \varepsilon_i) = 0$, $cov(z_i, P_i) \neq 0$.

2 Simulations, Omitted Variable Bias, and Controls [30 points]

```
knitr::include_graphics("/Users/terrylu/Desktop/UF/fall/R and Matlab/R/After 2nd Midterm/Assignment/Problem_Set_7/HW7/2.1.jpeg")
```

Suppose you want to again estimate the impact of schooling x_i on income y_i. Suppose income depends both on schooling and ability:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 a_i + \epsilon_i \tag{5}$$

1. To start, let's create an artificial dataset. Suppose the sample size is $N = 100$ and simulate income, schooling and ability for N hypothetical individuals.

   Suppose first ability is normally distributed with mean $\mu_a = 2$ and standard deviation $\sigma_a = 2.5$. In R for example you can draw N individuals from that distribution with the command

   $$a = rnorm(N, mean = 2, sd = 2.5).$$

   Suppose schooling has a random component x_{i1} that is normally distributed with mean 8 and sd = 2 but also depends on ability:

   $$x_i = x_{i1} + \gamma \cdot a_i.$$

   In R, you can therefore simulate x_i as

   $$x = rnorm(N, mean = 8, sd = 2) + c\_gamma * a.$$

   Finally, suppose the true coefficients are

   $$\beta_0 = 0.5$$
   $$\beta_1 = 1.5$$
   $$\beta_2 = 0.8$$

   Simulate this dataset when $\gamma = 0$ and compute the resulting income y_i for each individual, using the given values.

```
N <-100                                 # 100 samples
a <- rnorm(N, mean = 2, sd = 2.5)     # nomal random the ability

gamma <-  0                              # suppose the gamma = 1
x <- rnorm(N, mean = 8, sd = 2) + gamma * a   # get the education level based on gamm
a = 0

beta0 <- 0.5       # coefficents
beta1 <- 1.5
beta2 <- 0.8

epsilon <- rnorm(N, mean = 0, sd = 1)    # assume the error follows standard normal d
ist.

##### do we need to apply epsilon here?
y = beta0 + beta1 * x + beta2 * a

summary(x)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.373   6.698   7.890   7.931   9.427  12.948
```

```
knitr::include_graphics("/Users/terrylu/Desktop/UF/fall/R and Matlab/R/After 2nd Midt
erm/Assignment/Problem_Set_7/HW7/2.2.png")
```

2. Suppose γ is still equal to 0, i.e. ability does not matter for the decision to go to school. Also, suppose you observe income with measurement error, i.e. instead of observing income you observe

$$y_{i,obs} = y_i + \epsilon_i \tag{6}$$

with $\epsilon_i \sim N(0, \sigma_\epsilon = 3.5)$. In R for example you can construct $y_{i,obs}$ via

$$y\_obs = y + rnorm(N, mean = 0, sd = 3.5)$$

Run a regression of y_{obs} on schooling x. Is your estimated β_1 close to the true coefficient?

```
y_obs <- y + rnorm(N, mean = 0, sd = 3.5)
reg2.2 <- lm(y_obs ~ x)
summary(reg2.2)
```

```
##
## Call:
## lm(formula = y_obs ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.5999  -1.9571   0.1572   2.5153   9.2406
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.874      1.558   1.203    0.232
## x              1.534      0.190   8.071 1.79e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.938 on 98 degrees of freedom
## Multiple R-squared:  0.3993, Adjusted R-squared:  0.3932
## F-statistic: 65.14 on 1 and 98 DF,  p-value: 1.791e-12
```

```
cat("real β1 =", beta1, "\n")
```

```
## real β1 = 1.5
```

```
cat("estimated β1 =", coef(reg2.2)[2], "\n")
```

```
## estimated β1 = 1.533667
```

NO, it's not really close.

```
knitr::include_graphics("/Users/terrylu/Desktop/UF/fall/R and Matlab/R/After 2nd Midt
erm/Assignment/Problem_Set_7/HW7/2.3.png")
```

3. Rerun the regression from part (2) but use a larger sample of N = 100,000. Do you get closer to the true coefficient? How does a lower σ_ϵ affect how close you get to the true coefficient?

```
N <-100000                              # 100 samples
a <- rnorm(N, mean = 2, sd = 2.5)    # nomal random the ability

gamma <-  0                             # suppose the gamma = 1
x <- rnorm(N, mean = 8, sd = 2) + gamma * a   # get the education level based on gamm
a = 0

beta0 <- 0.5       # coefficents
beta1 <- 1.5
beta2 <- 0.8

y = beta0 + beta1 * x + beta2 * a

y_obs <- y + rnorm(N, mean = 0, sd = 3.5)
reg2.3 <- lm(y_obs ~ x)
summary(reg2.3)
```

```
##
## Call:
## lm(formula = y_obs ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.2750  -2.7243  -0.0005   2.7265  19.0127
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.191556   0.052604   41.66   <2e-16 ***
## x           1.488228   0.006372  233.56   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.031 on 99998 degrees of freedom
## Multiple R-squared:  0.353,  Adjusted R-squared:  0.353
## F-statistic: 5.455e+04 on 1 and 99998 DF,  p-value: < 2.2e-16
```

```
cat("real β1 =", beta1, "\n")
```

```
## real β1 = 1.5
```

```
cat("estimated β1 =", coef(reg2.3)[2], "\n")
```

```
## estimated β1 = 1.488228
```

YES, it's closer.

Now, try lower σϵ,for example 3.5, 0.35, 0.035, 0.0035, 0.00035, 0.000035:

```
N <-100000                              # 100 samples
a <- rnorm(N, mean = 2, sd = 2.5)     # nomal random the ability

gamma <-  0                             # suppose the gamma = 1
x <- rnorm(N, mean = 8, sd = 2) + gamma * a   # get the education level based on gamm
a = 0

beta0 <- 0.5       # coefficents
beta1 <- 1.5
beta2 <- 0.8

y = beta0 + beta1 * x + beta2 * a

σϵ_samples <- c(3.5, 0.35, 0.035, 0.0035, 0.00035, 0.000035)

for (i in σϵ_samples){
  y_obs <- y + rnorm(N, mean = 0, sd = i)
  reg2.3 <- lm(y_obs ~ x)
  summary(reg2.3)

  cat("when σϵ =", i, "\n")
  cat("real β1 =", beta1, "\n")
  cat("estimated β1 =", coef(reg2.3)[2], "\n")
  cat("  ", "\n")
  }
```

```
## when σϵ = 3.5
## real β1 = 1.5
## estimated β1 = 1.500859
##
## when σϵ = 0.35
## real β1 = 1.5
## estimated β1 = 1.502659
##
## when σϵ = 0.035
## real β1 = 1.5
## estimated β1 = 1.502266
##
## when σϵ = 0.0035
## real β1 = 1.5
## estimated β1 = 1.5022
##
## when σϵ = 0.00035
## real β1 = 1.5
## estimated β1 = 1.502202
##
## when σϵ = 3.5e-05
## real β1 = 1.5
## estimated β1 = 1.502201
##
```

It does not really matter, since the difference between real β1 and estimated β1 doesn't merge to 1.5

```
knitr::include_graphics("/Users/terrylu/Desktop/UF/fall/R and Matlab/R/After 2nd Midt
erm/Assignment/Problem_Set_7/HW7/2.4.png")
```

4. Now, suppose $\gamma = 1$, i.e. ability actually affects the schooling decision of each individual, x_i. But suppose you don't observe ability and therefore have to run the same regression as in part (2). With N = 100,000 and $\sigma_\epsilon = 3.5$, how does your estimated β compare to the true coefficient now? Is this what you would have expected based on what we saw in class?

```
N <-100000                              # 100 samples
a <- rnorm(N, mean = 2, sd = 2.5)    # nomal random the ability

gamma <-  1                             # suppose the gamma = 1
x <- rnorm(N, mean = 8, sd = 2) + gamma * a   # get the education level based on gamm
a = 0

beta0 <- 0.5       # coefficents
beta1 <- 1.5
beta2 <- 0.8

y = beta0 + beta1 * x + beta2 * a

y_obs <- y + rnorm(N, mean = 0, sd = 3.5)
reg2.3 <- lm(y_obs ~ x)
summary(reg2.3)
```

```
##
## Call:
## lm(formula = y_obs ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.5726  -2.5214  -0.0007   2.5110  15.0792
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.774916   0.038473  -72.13   <2e-16 ***
## x            1.987643   0.003663  542.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.718 on 99998 degrees of freedom
## Multiple R-squared:  0.7465, Adjusted R-squared:  0.7465
## F-statistic: 2.945e+05 on 1 and 99998 DF,  p-value: < 2.2e-16
```

```
cat("real β1 =", beta1, "\n")
```

```
## real β1 = 1.5
```

```
cat("estimated β1 =", coef(reg2.3)[2], "\n")
```

```
## estimated β1 = 1.987643
```

estimated β1 is higher than real β1. Yes, since the regression omitted the αi, the results has the omitted variable bias, which means the estimated β1 is biased. Since the cov(xi, αi) is positive, and gamma is positive, then estimated β1 should be always higher than real β1 (as the equation we shown above Problem 1.4)

```
knitr::include_graphics("/Users/terrylu/Desktop/UF/fall/R and Matlab/R/After 2nd Midt
erm/Assignment/Problem_Set_7/HW7/2.5.png")
```

5. Now, suppose you do observe ability. Include it as a control variable in your previous regression. Does this improve the estimate of β compared to the one you obtained in part (4)?

```
N <-100000                              # 100 samples
a <- rnorm(N, mean = 2, sd = 2.5)     # nomal random the ability

gamma <-  1                             # suppose the gamma = 1
x <- rnorm(N, mean = 8, sd = 2) + gamma * a   # get the education level based on gamm
a = 0

beta0 <- 0.5       # coefficents
beta1 <- 1.5
beta2 <- 0.8

y = beta0 + beta1 * x + beta2 * a

y_obs <- y + rnorm(N, mean = 0, sd = 3.5)
reg2.3 <- lm(y_obs ~ x + a)
summary(reg2.3)
```

```
##
## Call:
## lm(formula = y_obs ~ x + a)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.260  -2.371   0.013   2.355  15.300
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.484621   0.046323   10.46   <2e-16 ***
## x           1.502007   0.005506  272.81   <2e-16 ***
## a           0.798082   0.007055  113.13   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.494 on 99997 degrees of freedom
## Multiple R-squared:  0.7755, Adjusted R-squared:  0.7755
## F-statistic: 1.727e+05 on 2 and 99997 DF,  p-value: < 2.2e-16
```

```
cat("real β1 =", beta1, "\n")
```

```
## real β1 = 1.5
```

```
cat("estimated β1 =", coef(reg2.3)[2], "\n")
```

```
## estimated β1 = 1.502007
```

Yes, since we included the variable αi in, this elimated the β1's bias, and make the estimate unbiased.

3 Instrumental Variables [35 points]

```
knitr::include_graphics("/Users/terrylu/Desktop/UF/fall/R and Matlab/R/After 2nd Midt
erm/Assignment/Problem_Set_7/HW7/3.1.png")
```

For this problem, use the file "data_card.csv" which is posted on Canvas. This problem is taken from "Wooldridge - Econometric Analysis of Cross-Section and Panel Data". All variables are described in the file "card_variables.pdf". This problem is based on a paper by David Card who aimed at estimating the returns to college. To overcome the traditional identification problem, Card used the geographical distance of one's home to a college as instrument for schooling. The following questions ask you to replicate and understand Card's paper.

1. Read the first 11 pages of the paper (Card (1993))

```
# inport the data
dt <- fread("/Users/terrylu/Desktop/UF/fall/R and Matlab/R/After 2nd Midterm/Assignme
nt/Problem_Set_7/data_card.csv")
```

```
knitr::include_graphics("/Users/terrylu/Desktop/UF/fall/R and Matlab/R/After 2nd Midt
erm/Assignment/Problem_Set_7/HW7/3.2.png")
```

2. Estimate a log(wage) equation by OLS with *educ*, *exper*, $exper^2$, *black*, *south*, *smsa*, *reg*661 through *reg*668, and *smsa*66 as explanatory variables. Compare your results with Table 2, Column (2) in Card (1993).

```
knitr::include_graphics("/Users/terrylu/Desktop/UF/fall/R and Matlab/R/After 2nd Midt
erm/Assignment/Problem_Set_7/HW7/3.22.png")
```

Table 2: Estimated Regression Models for Log Hourly Earnings

	(1)	(2)	(3)	(4)	(5)
1. Education	0.074	0.075	0.073	0.074	0.073
	(0.004)	(0.003)	(0.004)	(0.004)	(0.004)
2. Experience	0.084	0.085	0.085	0.085	0.085
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
3. Experience-Squared /100	-0.224	-0.229	-0.230	-0.226	-0.229
	(0.032)	(0.032)	(0.032)	(0.032)	(0.032)
4. Black Indicator	-0.190	-0.199	-0.194	-0.194	-0.189
	(0.017)	(0.018)	(0.019)	(0.019)	(0.019)
5. Live in South	-0.125	-0.148	-0.146	-0.145	-0.146
	(0.015)	(0.026)	(0.026)	(0.026)	(0.026)
6. Live in SMSA	0.161	0.136	0.136	0.137	0.138
	(0.015)	(0.020)	(0.020)	(0.020)	(0.020)
7. Region in 1966 (8 indicators)	no	yes	yes	yes	yes
8. Live in SMSA in 1966	no	yes	yes	yes	yes
9. Parental Education[a] (main effects)	no	no	yes	yes	yes
10. Interacted Parental Education Classes[b]	no	no	no	yes	yes
11. Family Structure[c] (2 indicators)	no	no	no	no	yes
12. R-squared	0.291	0.300	0.301	0.303	0.304
13. P-value for family background effects	--	--	0.235	0.462	0.165

```
dt[ , exper2 := (exper^2)]
reg3.2 <- lm(log(wage) ~ educ + exper + exper2 + black + south + smsa + reg661 + reg6
62 + reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + smsa66, data = dt)

summary(reg3.2)
```

```
##
## Call:
## lm(formula = log(wage) ~ educ + exper + exper2 + black + south +
##     smsa + reg661 + reg662 + reg663 + reg664 + reg665 + reg666 +
##     reg667 + reg668 + smsa66, data = dt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.62326 -0.22141  0.02001  0.23932  1.33340
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.7393765  0.0715282  66.259  < 2e-16 ***
## educ         0.0746933  0.0034983  21.351  < 2e-16 ***
## exper        0.0848320  0.0066242  12.806  < 2e-16 ***
## exper2      -0.0022870  0.0003166  -7.223 6.41e-13 ***
## black       -0.1990123  0.0182483 -10.906  < 2e-16 ***
## south       -0.1479550  0.0259799  -5.695 1.35e-08 ***
## smsa         0.1363845  0.0201005   6.785 1.39e-11 ***
## reg661      -0.1185698  0.0388301  -3.054 0.002281 **
## reg662      -0.0222026  0.0282575  -0.786 0.432092
## reg663       0.0259703  0.0273644   0.949 0.342670
## reg664      -0.0634942  0.0356803  -1.780 0.075254 .
## reg665       0.0094550  0.0361174   0.262 0.793504
## reg666       0.0219476  0.0400984   0.547 0.584183
## reg667      -0.0005888  0.0393793  -0.015 0.988072
## reg668      -0.1750058  0.0463394  -3.777 0.000162 ***
## smsa66       0.0262417  0.0194477   1.349 0.177327
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3723 on 2994 degrees of freedom
## Multiple R-squared:  0.2998, Adjusted R-squared:  0.2963
## F-statistic: 85.48 on 15 and 2994 DF,  p-value: < 2.2e-16
```

Two results are the same.

```
knitr::include_graphics("/Users/terrylu/Desktop/UF/fall/R and Matlab/R/After 2nd Midt
erm/Assignment/Problem_Set_7/HW7/3.3.png")
```

3. Estimate a reduced form equation for educ containing all explanatory variables from part (2) and the dummy variable nearc4. Does nearc4 have an economically and statistically significant correlation with educ?

```
reg3.3 <- lm(log(wage) ~ educ + exper + exper2 + black + south + smsa + nearc4 + reg6
61 + reg662 + reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + smsa66, data = dt
)

summary(reg3.3)
```

```
##
## Call:
## lm(formula = log(wage) ~ educ + exper + exper2 + black + south +
##     smsa + nearc4 + reg661 + reg662 + reg663 + reg664 + reg665 +
##     reg666 + reg667 + reg668 + smsa66, data = dt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.62661 -0.22235  0.01862  0.23944  1.33226
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.7353773  0.0716219  66.116  < 2e-16 ***
## educ         0.0744417  0.0035060  21.233  < 2e-16 ***
## exper        0.0847311  0.0066247  12.790  < 2e-16 ***
## exper2      -0.0022854  0.0003166  -7.218 6.66e-13 ***
## black       -0.2001590  0.0182786 -10.950  < 2e-16 ***
## south       -0.1476166  0.0259810  -5.682 1.46e-08 ***
## smsa         0.1347577  0.0201562   6.686 2.73e-11 ***
## nearc4       0.0182541  0.0168929   1.081 0.279973
## reg661      -0.1198127  0.0388460  -3.084 0.002059 **
## reg662      -0.0235321  0.0282835  -0.832 0.405471
## reg663       0.0268518  0.0273758   0.981 0.326741
## reg664      -0.0632290  0.0356801  -1.772 0.076479 .
## reg665       0.0109031  0.0361412   0.302 0.762917
## reg666       0.0258109  0.0402563   0.641 0.521464
## reg667       0.0023874  0.0394744   0.060 0.951778
## reg668      -0.1729954  0.0463754  -3.730 0.000195 ***
## smsa66       0.0199851  0.0202908   0.985 0.324737
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3723 on 2993 degrees of freedom
## Multiple R-squared:  0.3001, Adjusted R-squared:  0.2964
## F-statistic: 80.21 on 16 and 2993 DF,  p-value: < 2.2e-16
```

No, the variable near4 is not significant as shown above. Which means near4 does not aaffect the log(wage) directly, satisfying one of the requirement of IV.

```
knitr::include_graphics("/Users/terrylu/Desktop/UF/fall/R and Matlab/R/After 2nd Midt
erm/Assignment/Problem_Set_7/HW7/3.4.png")
```

4. Estimate the log(wage) equation by IV, using nearc4 as an instrument for educ. Compare the coefficient on education with that obtained from part (2)

```
reg3.4 <- ivreg(log(wage) ~ educ + exper + exper2 + black + south + smsa + reg661 + r
eg662 + reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + smsa66 | nearc4 + exper
+ exper2 + black + south + smsa + reg661 + reg662 + reg663 + reg664 + reg665 + reg666
+ reg667 + reg668 + smsa66, data = dt)

summary(reg3.4)
```

```
##
## Call:
## ivreg(formula = log(wage) ~ educ + exper + exper2 + black + south +
##     smsa + reg661 + reg662 + reg663 + reg664 + reg665 + reg666 +
##     reg667 + reg668 + smsa66 | nearc4 + exper + exper2 + black +
##     south + smsa + reg661 + reg662 + reg663 + reg664 + reg665 +
##     reg666 + reg667 + reg668 + smsa66, data = dt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83164 -0.24075  0.02429  0.25208  1.42760
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.7739653  0.9349470   4.037 5.56e-05 ***
## educ         0.1315038  0.0549637   2.393  0.01679 *
## exper        0.1082711  0.0236586   4.576 4.92e-06 ***
## exper2      -0.0023349  0.0003335  -7.001 3.12e-12 ***
## black       -0.1467758  0.0538999  -2.723  0.00650 **
## south       -0.1446715  0.0272846  -5.302 1.23e-07 ***
## smsa         0.1118083  0.0316620   3.531  0.00042 ***
## reg661      -0.1078142  0.0418137  -2.578  0.00997 **
## reg662      -0.0070465  0.0329073  -0.214  0.83046
## reg663       0.0404445  0.0317806   1.273  0.20325
## reg664      -0.0579171  0.0376059  -1.540  0.12364
## reg665       0.0384576  0.0469387   0.819  0.41267
## reg666       0.0550887  0.0526597   1.046  0.29559
## reg667       0.0267580  0.0488287   0.548  0.58374
## reg668      -0.1908912  0.0507113  -3.764  0.00017 ***
## smsa66       0.0185311  0.0216086   0.858  0.39119
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3883 on 2994 degrees of freedom
## Multiple R-Squared: 0.2382,  Adjusted R-squared: 0.2343
## Wald test: 51.01 on 15 and 2994 DF,  p-value: < 2.2e-16
```

```
cat("β1(OLS) =", coef(reg3.2)[2], "\n")
```

```
## β1(OLS) = 0.07469326
```

```
cat("β1(IV) =", coef(reg3.4)[2], "\n")
```

```
## β1(IV) = 0.1315038
```

β1(IV) is significant too, moreover the coefficent of β1(IV) is higher than earlier estimates β1(OLS). Which means increasing every 1 unit of education imput, the wage will increase 13.15%.

```
knitr::include_graphics("/Users/terrylu/Desktop/UF/fall/R and Matlab/R/After 2nd Midt
erm/Assignment/Problem_Set_7/HW7/3.5.png")
```

5. Now use nearc2 along with nearc4 as instruments for educ. First estimate the reduced form for educ and comment on whether nearc2 or nearc4 is more strongly related to educ. How do the 2SLS estimates compare with the earlier estimates?

We regression with model: educ = β0 + β1 * nearc2 + β2 * nearc4 + βi * X + ϵ; X contains all control variables.

```
reg3.51 <- lm(educ ~ nearc2 + nearc4 + exper + exper2 + black + south + smsa + reg661
+ reg662 + reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + smsa66, data = dt)
summary(reg3.51)
```

```
##
## Call:
## lm(formula = educ ~ nearc2 + nearc4 + exper + exper2 + black +
##     south + smsa + reg661 + reg662 + reg663 + reg664 + reg665 +
##     reg666 + reg667 + reg668 + smsa66, data = dt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5851 -1.3845 -0.0823  1.2765  6.2930
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.677e+01  2.163e-01  77.528  < 2e-16 ***
## nearc2       1.230e-01  7.743e-02   1.589 0.112256
## nearc4       3.206e-01  8.784e-02   3.650 0.000267 ***
## exper       -4.123e-01  3.369e-02 -12.237  < 2e-16 ***
## exper2       8.479e-04  1.650e-03   0.514 0.607379
## black       -9.452e-01  9.391e-02 -10.065  < 2e-16 ***
## south       -4.191e-02  1.355e-01  -0.309 0.757162
## smsa         4.014e-01  1.048e-01   3.830 0.000131 ***
## reg661      -1.688e-01  2.041e-01  -0.827 0.408286
## reg662      -2.690e-01  1.478e-01  -1.820 0.068884 .
## reg663      -1.902e-01  1.458e-01  -1.305 0.192022
## reg664      -3.772e-02  1.892e-01  -0.199 0.841990
## reg665      -4.371e-01  1.903e-01  -2.297 0.021703 *
## reg666      -5.022e-01  2.097e-01  -2.395 0.016679 *
## reg667      -3.775e-01  2.079e-01  -1.816 0.069511 .
## reg668       3.820e-01  2.454e-01   1.557 0.119683
## smsa66       7.825e-05  1.069e-01   0.001 0.999416
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.94 on 2993 degrees of freedom
## Multiple R-squared:  0.4776, Adjusted R-squared:  0.4748
## F-statistic:   171 on 16 and 2993 DF,  p-value: < 2.2e-16
```

As shown above, nearc4 is significant while nearc2 is not. Then we will take nearc4 as IV.| 2SLS: log(wage) = β0 + β1 * educ + βi * X + ϵ; educ = α0 + α1 * nearc4 + u

β1(IV) = est_cov(nearc4, log(wage))/est_cov(nearc4,educ)

```
reg3.52 <- ivreg(log(wage) ~ educ + exper + exper2 + black + south + smsa + reg661 +
reg662 + reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + smsa66 | nearc2 + near
c4 + exper + exper2 + black + south + smsa + reg661 + reg662 + reg663 + reg664 + reg6
65 + reg666 + reg667 + reg668 + smsa66, data = dt)

summary(reg3.52)
```

```
##
## Call:
## ivreg(formula = log(wage) ~ educ + exper + exper2 + black + south +
##     smsa + reg661 + reg662 + reg663 + reg664 + reg665 + reg666 +
##     reg667 + reg668 + smsa66 | nearc2 + nearc4 + exper + exper2 +
##     black + south + smsa + reg661 + reg662 + reg663 + reg664 +
##     reg665 + reg666 + reg667 + reg668 + smsa66, data = dt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.93841 -0.25068  0.01932  0.26519  1.46998
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.3396868  0.8945378   3.733 0.000192 ***
## educ         0.1570594  0.0525782   2.987 0.002839 **
## exper        0.1188149  0.0228061   5.210 2.02e-07 ***
## exper2      -0.0023565  0.0003475  -6.781 1.43e-11 ***
## black       -0.1232778  0.0521500  -2.364 0.018147 *
## south       -0.1431945  0.0284448  -5.034 5.08e-07 ***
## smsa         0.1007530  0.0315193   3.197 0.001405 **
## reg661      -0.1029760  0.0434224  -2.371 0.017779 *
## reg662      -0.0002287  0.0337943  -0.007 0.994602
## reg663       0.0469556  0.0326490   1.438 0.150484
## reg664      -0.0554084  0.0391828  -1.414 0.157437
## reg665       0.0515041  0.0475678   1.083 0.279006
## reg666       0.0699968  0.0533049   1.313 0.189237
## reg667       0.0390596  0.0497499   0.785 0.432446
## reg668      -0.1980371  0.0525350  -3.770 0.000167 ***
## smsa66       0.0150626  0.0223360   0.674 0.500132
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4053 on 2994 degrees of freedom
## Multiple R-Squared: 0.1702,  Adjusted R-squared: 0.166
## Wald test: 47.07 on 15 and 2994 DF,  p-value: < 2.2e-16
```

```
cat("β1(IV1) =", coef(reg3.4)[2], "\n")
```

```
## β1(IV1) = 0.1315038
```

```
cat("β1(IV2) =", coef(reg3.52)[2], "\n")
```

```
## β1(IV2) = 0.1570594
```

β1(IV2) is significant too, moreover the coefficent of β1(IV2) is higher than earlier estimates β1(IV1). Which means increasing every 1 unit of education imput, the wage will increase 15.7%. However, the IV nearc2 is not a strong instrumental variable, so we need to be more cautious and discuss more about using nearc2 and nearc4 as IVs.

```
knitr::include_graphics("/Users/terrylu/Desktop/UF/fall/R and Matlab/R/After 2nd Midt
erm/Assignment/Problem_Set_7/HW7/3.6.png")
```

6. For a subset of the men in the sample, IQ score is available. Regress iq on nearc4. Is IQ score uncorrelated with nearc4?

IQ = β0 + β1 * nearc4 + ϵ

```
reg3.6 <- lm(IQ ~ nearc4, data = dt)
summary(reg3.6)
```

```
##
## Call:
## lm(formula = IQ ~ nearc4, data = dt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.611  -9.207   0.793  10.793  45.793
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 100.6106     0.6275 160.347  < 2e-16 ***
## nearc4        2.5962     0.7455   3.483 0.000507 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 2059 degrees of freedom
##   (949 observations deleted due to missingness)
## Multiple R-squared:  0.005856,   Adjusted R-squared:  0.005373
## F-statistic: 12.13 on 1 and 2059 DF,  p-value: 0.0005071
```

No, since the regression shows nearc4 significantly related with IQ with coefficent of 2.5962. Therefore, nearc4 is not a completely exogenous instrumental variable, cov(nearc4, ϵ_log(wage)) ≠ 0

```
knitr::include_graphics("/Users/terrylu/Desktop/UF/fall/R and Matlab/R/After 2nd Midt
erm/Assignment/Problem_Set_7/HW7/3.7.png")
```

7. Now regress iq on nearc4 along with smsa66, reg661, reg662, and reg669. Are iq and nearc4 correlated? What do you conclude about the importance of controlling for the 1966 location and regional dummies in the log(wage) equation when using nearc4 as an IV for educ?

```
reg3.7 <- lm(IQ ~ nearc4 + smsa66 + reg661 + reg662 + reg669, data = dt)
summary(reg3.7)
```

```
##
## Call:
## lm(formula = IQ ~ nearc4 + smsa66 + reg661 + reg662 + reg669,
##     data = dt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.415  -9.607   0.615  10.548  45.548
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  99.3847     0.7017 141.642  < 2e-16 ***
## nearc4        0.8681     0.8217   1.056  0.29088
## smsa66        1.3545     0.8028   1.687  0.09170 .
## reg661        4.7681     1.5468   3.083  0.00208 **
## reg662        5.8081     0.9018   6.441 1.47e-10 ***
## reg669        1.8447     1.1517   1.602  0.10938
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.21 on 2055 degrees of freedom
##   (949 observations deleted due to missingness)
## Multiple R-squared:  0.03019,    Adjusted R-squared:  0.02783
## F-statistic: 12.79 on 5 and 2055 DF,  p-value: 2.867e-12
```

IQ is not correlated with nearc4 any more. When we using nearc4 as IV, it's important to include these control variables in the IV regression model to ensure the exogeneity of the instrumental variable, on the other hand, to ensure cov(nearc4, ϵ_log(wage)) = 0