Econ 142: HW 3

Sofia Guo

2/8/2019

1(a). Question

Show that if x_i contains a constant, then $\bar{y} = \bar{x}'\hat{\beta}$, where $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ and $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$

1(a) Proof:

Under implications of defining properties of the PRF, one FOC is $E[x_i\hat{u}_i] = 0$. Thus:

$\Rightarrow E[y_i] = E[x_i'\hat{\beta} + \hat{u}_i]$ by taking the expected value of the sample regression equation $y_i = x_i'\hat{\beta} + \hat{u}_i$
$\Rightarrow E[y_i] = E[x_i'\hat{\beta}] + E[\hat{u}_i]$ by distributing the expectation across sums
$\Rightarrow E[y_i] = E[x_i'\hat{\beta}] + 0$ from the sample FOC $E[x_i\hat{u}_i] = \frac{1}{N}\sum_{i=1}^{N} x_i\hat{u}_i = 0$
$\Rightarrow \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{1}{N}\sum_{i=1}^{N} x_i'\hat{\beta}$ from taking the expectation
$\therefore \bar{y} = \bar{x}'\hat{\beta}$ using the definitions $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ and $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ $\blacksquare$.

1(b). Question

Show that if x_i contains a dummy variable for membership in group g (which has N_g observations in the sample) then $\bar{y}_g = \bar{x}_g'\hat{\beta}$ where $\bar{y}_g = \frac{1}{N_g}\sum_{i\in g} y_i$, and $\bar{x}_g = \frac{1}{N_g}\sum_{i\in g} x_i$.

1(b) Proof:

The FOC require $\sum_{i=1}^{N} D_i(y_i - x_i'\hat{\beta}) = 0 = \sum_{i\in g}(y_i - x_i'\hat{\beta})$. Thus:
$\Rightarrow \frac{1}{N_g}\sum_{i\in g}(y_i - x_i'\hat{\beta}) = 0$ just by taking the expected value
$\Rightarrow \frac{1}{N_g}\sum_{i\in g} y_i = \frac{1}{N_g}\sum_{i\in g} x_i'\hat{\beta}$ distributing the sums and bringing second term over
$\therefore \bar{y}_g = \bar{x}_g'\hat{\beta}$ since $\bar{y}_g = \frac{1}{N_g}\sum_{i\in g} y_i$, and $\bar{x}_g = \frac{1}{N_g}\sum_{i\in g} x_i$ $\blacksquare$.

1(c). Question

Complete the proof of the Frisch-Waugh Theorem for the sample OLS regression coefficients by showing that the j^{th} row of $\hat{\beta}$ is:

$$\hat{\beta}_j = [\frac{1}{N}\sum_{i=1}^{N}\hat{\xi}_i^2]^{-1}[\frac{1}{N}\sum_{i=1}^{N}\hat{\xi}_i y_i] \tag{1}$$

1(c) Proof:

$\Rightarrow$ We know from the FOC for the sample regression that $\frac{1}{N}\sum_{i=1}^{N} x_i\hat{u}_i = 0$
$\Rightarrow$ Then define $\hat{u}_i = y_i - x_i'\hat{\beta}$ from the sample OLS regression $y_i = x_i'\hat{\beta} + \hat{u}_i$

$\Rightarrow$ We also know from the FOC that $\frac{1}{N}\sum_{i=1}^{N} x_{(\sim j)i}\hat{\xi}_i = 0$
$\Rightarrow$ Then define $\hat{\xi}_i = x_{ji} - x_{(\sim j)i}\hat{\pi}$ from the auxiliary regression $x_{ji} = x_{(\sim j)i}\hat{\pi} + \hat{\xi}_i$
$\Rightarrow$ We know that the sample regression equation is $y_i = \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_j x_{ji} + \cdots + \hat{\beta}_K x_{Ki} + \hat{u}_i$
$\Rightarrow$ Thus we can write

$$\frac{1}{N}\sum_{i=1}^{N}\hat{\xi}_i y_i = \frac{1}{N}\sum_{i=1}^{N}\hat{\xi}_i(\hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_j x_{ji} + \cdots + \hat{\beta}_K x_{Ki} + \hat{u}_i) \tag{2}$$

from the previous step
$\Rightarrow$ Since $\frac{1}{N}\sum_{i=1}^{N} x_{(\sim j)i}\hat{\xi}_i = 0$ we know that $\hat{\xi}_i \perp x_{(\sim j)i}$
$\Rightarrow$ Since $\hat{\xi}_i = x_{ji} - x_{(\sim j)i}\hat{\pi}$ we know that $\hat{\xi}_i \perp \hat{u}_i$ because $\frac{1}{N}\sum_{i=1}^{N} x_i\hat{u}_i = 0$ $(x_i \perp \hat{u}_i)$
$\Rightarrow$ Thus we know that all the terms in 2 except $\hat{\beta}_j x_{ji}$ equal 0; Thus:
$\Rightarrow$ $\frac{1}{N}\sum_{i=1}^{N}\hat{\xi}_i y_i = \hat{\beta}_j \frac{1}{N}\sum_{i=1}^{N} x_{ji}$ by simplification
$\Rightarrow$ Using $x_{ji} = x_{(\sim j)i}\hat{\pi} + \hat{\xi}_i$ we substitute into $\frac{1}{N}\sum_{i=1}^{N}\hat{\xi}_i x_{ji} = \frac{1}{N}\sum_{i=1}^{N}\hat{\xi}_i(x_{(\sim j)i}\hat{\pi} + \hat{\xi}_i) = \frac{1}{N}\sum_{i=1}^{N}\hat{\xi}_i^2$ because $\hat{\xi}_i \perp x_{(\sim j)i}$ using the FOC for $\hat{\pi}$
$\Rightarrow$ $\frac{1}{N}\sum_{i=1}^{N}\hat{\xi}_i y_i = \hat{\beta}_j \frac{1}{N}\sum_{i=1}^{N}\hat{\xi}_i^2$ by substitution
$\therefore \hat{\beta}_j = [\frac{1}{N}\sum_{i=1}^{N}\hat{\xi}_i^2]^{-1}[\frac{1}{N}\sum_{i=1}^{N}\hat{\xi}_i y_i]$ $\blacksquare$.

2. OVB Dataset

```r
#load libraries
library(dplyr)
library(ggplot2)
library(magrittr)
library(xlsx)
library(reshape2)
library(stargazer)
library(lubridate)
library(lmtest)
library(ivpack)
library(kableExtra)

#import dataset
ovb_raw <- read.csv("/Users/sofia/Box/Cal (sofiaguo@berkeley.edu)/2018-19/Spring 2019/Econ 142/PSETS/PSE
```

2(a). Question:

Write an expression for the OLS estimate of the coefficient on immigrant statues from logwage = constant, immigrant statues, if the true model is logwage = constant, education, immigrant status.

2(a) Solution:

$$\hat{\beta}^0_{imm} = \hat{\beta}_{imm} + \hat{\pi}_2\hat{\beta}_{educ} \tag{3}$$

where $\hat{\pi}_2$ is the coefficient on immigration from regressing immigration on education (the omitted variable):

$$education_i = \hat{\pi}_1 + \hat{\pi}_2 immigration_i + \hat{\xi}_i \tag{4}$$

2(b). Question:

Estimate the 5 models and show values for (a), first female then male.

1. logwage = constant, immigrant statues
2. logwage = constant, education
3. immigrant status = constant, education
4. education = constant, immigrant status
5. logwage = constant, education, immigrant status

2(b) Solution:

Values for the female regressions:

$\hat{\pi}_{2_{female}} = -1.49214$

$\hat{\beta}_{imm_{female}} = -0.179986$

$\hat{\beta}_{educ_{female}} = 0.113853$

$\hat{\beta}^{0}_{imm_{female}} = -0.179986 + (-1.49214) * (0.113853) = -0.3498706$

Values for the male regressions:

$\hat{\pi}_{2_{male}} = -1.61176$

$\hat{\beta}_{imm_{male}} = -0.24477$

$\hat{\beta}_{educ_{male}} = 0.105620$

$\hat{\beta}^{0}_{imm_{male}} = -1.61176 + (-0.24477) * (0.105620) = -1.637613$

Code:

```
#estimate model 1 for both genders

#filter females only
fem <- ovb_raw %>%
  dplyr::filter(female == 1)

#filter males only
male <- ovb_raw %>%
  dplyr::filter(female == 0)

#run regression 1 for females
reg_fem_1 <- summary(lm(logwage ~ imm, data = fem))
reg_fem_1
```

```
##
## Call:
## lm(formula = logwage ~ imm, data = fem)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -1.5001 -0.4407 -0.0206  0.4066  3.2851
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.886378   0.007154  403.48   <2e-16 ***
## imm         -0.179986   0.016532  -10.89   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.664 on 10599 degrees of freedom
## Multiple R-squared:  0.01106,    Adjusted R-squared:  0.01097
## F-statistic: 118.5 on 1 and 10599 DF,  p-value: < 2.2e-16
```

```r
#run regression 1 for males

reg_male_1 <- summary(lm(logwage ~ imm, data = male))
reg_male_1
```

```
##
## Call:
## lm(formula = logwage ~ imm, data = male)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.76962 -0.42464 -0.00445  0.41578  3.08032
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.15592    0.00728  433.51   <2e-16 ***
## imm         -0.24477    0.01558  -15.71   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6844 on 11304 degrees of freedom
## Multiple R-squared:  0.02136,    Adjusted R-squared:  0.02127
## F-statistic: 246.7 on 1 and 11304 DF,  p-value: < 2.2e-16
```

```r
#estimate model 2 for both genders

#run regression 2 for females
reg_fem_2 <- summary(lm(logwage ~ educ, data = fem))
reg_fem_2
```

```
##
## Call:
## lm(formula = logwage ~ educ, data = fem)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1316 -0.3413  0.0034  0.3555  3.3868
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.234852   0.029811   41.42   <2e-16 ***
## educ        0.114153   0.002064   55.30   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5882 on 10599 degrees of freedom
## Multiple R-squared:  0.2239, Adjusted R-squared:  0.2238
## F-statistic:  3058 on 1 and 10599 DF,  p-value: < 2.2e-16
```

```r
#run regression 2 for males
reg_male_2 <- summary(lm(logwage ~ educ, data = male))
reg_male_2
```

```
##
## Call:
## lm(formula = logwage ~ educ, data = male)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.38111 -0.35217  0.01881  0.35307  3.08724
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.609458   0.027146   59.29   <2e-16 ***
## educ        0.107897   0.001917   56.28   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6115 on 11304 degrees of freedom
## Multiple R-squared:  0.2189, Adjusted R-squared:  0.2188
## F-statistic:  3167 on 1 and 11304 DF,  p-value: < 2.2e-16
```

```r
#estimate model 3 for both genders

#run regression 3 for females
reg_fem_3 <- summary(lm(imm ~ educ, data = fem))
reg_fem_3
```

```
##
## Call:
## lm(formula = imm ~ educ, data = fem)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51850 -0.22201 -0.13306 -0.07376  0.98554
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.607445   0.019329   31.43   <2e-16 ***
## educ        -0.029649   0.001339  -22.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3814 on 10599 degrees of freedom
## Multiple R-squared:  0.04424,    Adjusted R-squared:  0.04415
## F-statistic: 490.6 on 1 and 10599 DF,  p-value: < 2.2e-16
```

```r
#run regression 3 for males
reg_male_3 <- summary(lm(imm ~ educ, data = male))
reg_male_3
```

```
##
## Call:
## lm(formula = imm ~ educ, data = male)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.54939 -0.27436 -0.15213 -0.09101  0.97011
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.641066   0.017880   35.85   <2e-16 ***
## educ        -0.030559   0.001263  -24.20   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4028 on 11304 degrees of freedom
## Multiple R-squared:  0.04925,    Adjusted R-squared:  0.04917
## F-statistic: 585.6 on 1 and 11304 DF,  p-value: < 2.2e-16
```

```r
#estimate model 4 for both genders

#run regression 4 for females
reg_fem_4 <- summary(lm(educ ~ imm, data = fem))
reg_fem_4
```

```
##
## Call:
## lm(formula = educ ~ imm, data = fem)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.9597  -1.4518  -0.4518   1.5482   7.0403
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.45183    0.02915  495.76   <2e-16 ***
## imm         -1.49214    0.06737  -22.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.706 on 10599 degrees of freedom
## Multiple R-squared:  0.04424,    Adjusted R-squared:  0.04415
## F-statistic: 490.6 on 1 and 10599 DF,  p-value: < 2.2e-16
```

```r
#run regression 2 for males
reg_male_4 <- summary(lm(educ ~ imm, data = male))
reg_male_4
```

```
##
## Call:
## lm(formula = educ ~ imm, data = male)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.5776  -2.1894  -0.5776   1.8106   7.4224
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.18939    0.03111   456.1   <2e-16 ***
## imm         -1.61176    0.06660   -24.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.925 on 11304 degrees of freedom
## Multiple R-squared:  0.04925,    Adjusted R-squared:  0.04917
## F-statistic: 585.6 on 1 and 11304 DF,  p-value: < 2.2e-16
```

```
#estimate model 5 for both genders

#run regression 5 for females
reg_fem_5 <- summary(lm(logwage ~ educ + imm, data = fem))
reg_fem_5
```

```
##
## Call:
## lm(formula = logwage ~ educ + imm, data = fem)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1318 -0.3439  0.0038  0.3552  3.3943
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.240988   0.031170  39.814   <2e-16 ***
## educ         0.113853   0.002112  53.915   <2e-16 ***
## imm         -0.010101   0.014981  -0.674      0.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5883 on 10598 degrees of freedom
## Multiple R-squared:  0.2239, Adjusted R-squared:  0.2238
## F-statistic:  1529 on 2 and 10598 DF,  p-value: < 2.2e-16
```

```
#run regression 2 for males
reg_male_5 <- summary(lm(logwage ~ educ + imm, data = male))
reg_male_5
```

```
##
## Call:
## lm(formula = logwage ~ educ + imm, data = male)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.38334 -0.36121  0.01482  0.35657  3.14132
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  1.657239   0.028614  57.917  < 2e-16 ***
## educ         0.105620   0.001964  53.780  < 2e-16 ***
## imm         -0.074534   0.014263  -5.226 1.77e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6108 on 11303 degrees of freedom
## Multiple R-squared:  0.2208, Adjusted R-squared:  0.2206
## F-statistic:  1601 on 2 and 11303 DF,  p-value: < 2.2e-16
```

2(c). Question:

Redo 5 models for Females and Males, distinguishing 3 groups of immigrants. Put and include results in 2 new tables similar to the table in lecture 5.

```r
#add dummies for asian, hispanic and other immigrants

#make female data frame
ovb_race_imm_fem <- ovb_raw %>%
  mutate(asian_imm = as.numeric(asian == 1 & imm ==1 & hispanic ==0),
         hispanic_imm = as.numeric(hispanic == 1 & imm ==1 & asian ==0),
         other_imm = as.numeric(imm ==1 & asian ==0 & hispanic==0)) %>%
  dplyr::filter(female ==1)

#make male data frame
ovb_race_imm_male <- ovb_raw %>%
  mutate(asian_imm = as.numeric(asian == 1 & imm ==1 & hispanic ==0),
         hispanic_imm = as.numeric(hispanic == 1 & imm ==1 & asian ==0),
         other_imm = as.numeric(imm ==1 & asian ==0 & hispanic==0)) %>%
  dplyr::filter(female ==0)
```

```r
#run regression 1 for females
reg_fem_1_imm <- lm(logwage ~ imm + asian_imm + hispanic_imm + other_imm, data = ovb_race_imm_fem)

#run regression 1 for males
reg_male_1_imm <- lm(logwage ~ imm + asian_imm + hispanic_imm + other_imm, data = ovb_race_imm_male)
```

```r
#estimate model 2 for both genders

#run regression 2 for females
reg_fem_2_imm <- lm(logwage ~ educ + asian_imm + hispanic_imm + other_imm, data = ovb_race_imm_fem)

#run regression 2 for males
reg_male_2_imm <- lm(logwage ~ educ + asian_imm + hispanic_imm + other_imm, data = ovb_race_imm_male)
```

```r
#estimate model 3 for both genders

#run regression 3 for females
reg_fem_3_imm <- lm(imm ~ educ + asian_imm + hispanic_imm + other_imm, data = ovb_race_imm_fem)

#run regression 3 for males
reg_male_3_imm <- lm(imm ~ educ + asian_imm + hispanic_imm + other_imm, data = ovb_race_imm_male)
```

```r
#estimate model 4 for both genders
```

```r
#run regression 4 for females
reg_fem_4_imm <- lm(educ ~ imm + asian_imm + hispanic_imm + other_imm, data = ovb_race_imm_fem)

#run regression 4 for males
reg_male_4_imm <- lm(educ ~ imm + asian_imm + hispanic_imm + other_imm, data = ovb_race_imm_male)
```

```r
#estimate model 5 for both genders

#run regression 5 for females
reg_fem_5_imm <- lm(logwage ~ educ + imm + asian_imm + hispanic_imm + other_imm, data = ovb_race_imm_fem

#run regression 5 for males
reg_male_5_imm <- lm(logwage ~ educ + imm + asian_imm + hispanic_imm + other_imm, data = ovb_race_imm_ma
```

2(c) Tables:

```r
stargazer(reg_fem_1_imm,
          reg_fem_2_imm,
          reg_fem_3_imm,
          reg_fem_4_imm,
          reg_fem_5_imm,
          type = "latex", title = "Female Immigrant Regression
          Results",
          header = F,
          multicolumn = F,
          column.sep.width = '0.1pt',
          single.row = T,
          omit.stat = c("f", "ser"))
```

Table 1: Female Immigrant Regression Results

	Dependent variable:				
	logwage	logwage	imm	educ	logwage
	(1)	(2)	(3)	(4)	(5)
imm	0.106 (0.218)			−2.341*** (0.859)	0.368* (0.196)
educ		0.112*** (0.002)	−0.0003*** (0.0001)		0.112*** (0.002)
asian_imm	−0.020 (0.220)	0.029 (0.027)	0.999*** (0.001)	2.851*** (0.866)	−0.339* (0.198)
hispanic_imm	−0.543** (0.219)	−0.054** (0.021)	0.998*** (0.001)	−1.092 (0.863)	−0.421** (0.197)
other_imm	−0.057 (0.220)	0.016 (0.028)	0.999*** (0.001)	2.624*** (0.867)	−0.351* (0.198)
Constant	2.886*** (0.007)	1.269*** (0.033)	0.005*** (0.002)	14.452*** (0.028)	1.267*** (0.033)
Observations	10,601	10,601	10,601	10,601	10,601
R^2	0.038	0.224	0.994	0.134	0.225
Adjusted R^2	0.038	0.224	0.994	0.134	0.224
Note:					*p<0.1; **p<0.05; ***p<0.01

```r
stargazer(reg_male_1_imm,
          reg_male_2_imm,
          reg_male_3_imm,
          reg_male_4_imm,
          reg_male_5_imm,
          type = "latex", title = "Male Immigrant Regression
          Results",
          header = F,
          multicolumn = F,
          column.sep.width = '0.1pt',
          single.row = T,
          omit.stat = c("f", "ser"))
```

Table 2: Male Immigrant Regression Results

	Dependent variable:				
	logwage	logwage	imm	educ	logwage
	(1)	(2)	(3)	(4)	(5)
imm	−0.141 (0.195)			−2.273*** (0.787)	0.097 (0.177)
educ		0.105*** (0.002)	−0.0003*** (0.0001)		0.105*** (0.002)
asian_imm	0.213 (0.197)	−0.060** (0.027)	0.999*** (0.001)	3.529*** (0.795)	−0.156 (0.178)
hispanic_imm	−0.331* (0.195)	−0.092*** (0.019)	0.997*** (0.001)	−1.366* (0.789)	−0.188 (0.177)
other_imm	0.155 (0.197)	−0.057** (0.027)	0.999*** (0.001)	2.947*** (0.795)	−0.154 (0.178)
Constant	3.156*** (0.007)	1.671*** (0.031)	0.006*** (0.002)	14.189*** (0.029)	1.671*** (0.031)
Observations	11,306	11,306	11,306	11,306	11,306
R^2	0.051	0.221	0.994	0.176	0.221
Adjusted R^2	0.051	0.221	0.994	0.176	0.221
Note:					*p<0.1; **p<0.05; ***p<0.01