Econ C142 Pset3

Johan Hallin

February 14 2019

Problem 1

a) Let

$$\mathrm{X}_i = \begin{bmatrix} 1 \\ x_2 \\ \vdots \\ x_i \end{bmatrix} \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_i \end{bmatrix}$$

$$\mathrm{y}_i = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \end{bmatrix} u_i = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_i \end{bmatrix}$$

(1)

We want to minimize

$$\hat{\beta} = (y_i - X`\beta) \tag{2}$$

FOC:

$$\frac{1}{N}\sum_i X_i(y_i - X`\beta) = 0 \tag{3}$$

Using the fact that the first covariate is 1 (row 1):

$$\frac{1}{N}\sum_i 1(y_i - X`\beta) = 0 \tag{4}$$

$$\frac{1}{N}\sum_i y_i = \frac{1}{N}\sum_i X`\beta \tag{5}$$

$$\bar{y} = \bar{x}\hat{\beta} \tag{6}$$

This is true because if we include a constant in our regression, the constant itself picks up any deviation and ensures that the predictions from the regression fit the mean perfectly.

b) Now changing the columnvector from (1).

$$\mathrm{X}_i = \begin{bmatrix} 1 \\ x_2 \\ \vdots \\ D_i \end{bmatrix} (7)$$

When x_i contains a dummy such that $D_i = 1$ when group G and $D_i = 0$ otherwise. Using the first order condition:

$$\frac{1}{N}\sum_{i} X_i(y_i - X`\beta) = 0 \tag{8}$$

Using only the row for dummy variable:

$$\frac{1}{N}\sum_{i} D_i(y_i - X`\beta) = 0 \tag{9}$$

For $D_i = 1$ we will only sum across group G thus:

$$\frac{1}{N}\sum_{i\in G}(y_i - X`\beta) = 0 \tag{10}$$

Distributing the sum and divide by N_g:

$$\frac{1}{N_g}\sum_{i\in G} y_i = \frac{1}{N_g}\sum_{i\in G} x_i\hat{\beta} \tag{11}$$

Using the definitions given to us we will have proven what we were set out to prove.

$$\bar{y} = \bar{x}\hat{\beta} \tag{12}$$

c)

Let

$$\tilde{\xi_i} = x_{ji} - x_{(\ j)i}\tilde{\pi}$$

(13)

From the first order condition

$$\hat{\beta} = [\frac{1}{N}\sum_{i} x_i`x_i]^{-}1[\frac{1}{N}\sum_{i} y_i u_i] \tag{14}$$

Lets expand the right hand side of the equation:

$$\frac{1}{N}\sum_{i}\hat{\xi}_i y_i = \frac{1}{N}\sum_{i}(\hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_j x_{ji} + ... + \hat{\beta}_k x_{ki} + \hat{u}_i) \tag{15}$$

From the first order condition from our regression we get that:

$$\frac{1}{N}\sum_{i}\hat{\xi}_i \hat{u}_i = 0 \tag{16}$$

This is true since ui is orthogonal to every covariate in the regression. From equation (13) we can see that xi is built on the covariates. Similarly, xi is orthogonal to every covariate except the jth. This comes from the first order

condition from our auxiliary regression.
Therefore we get the following true statement:

$$\frac{1}{N}\sum_i \hat{\xi}_i(\hat{\beta}_j x_{ji}) \tag{17}$$

Since beta is just a constant we may pull it out from the summation:

$$\hat{\beta}_j \frac{1}{N}\sum_i \hat{\xi}_i x_{ji}$$

(18)

Setting equal to the left hand side of equation (9):

$$\frac{1}{N}\sum_i \hat{\xi}_i y_i = \hat{\beta}_j \frac{1}{N}\sum_i \hat{\xi}_i x_{ji} \tag{19}$$

Solving for beta we get:

$$\hat{\beta}_j = [\frac{1}{N}\sum_i \hat{\xi}_i x_{ji}]^{-}1[\frac{1}{N}\sum_i \hat{\xi}_i y_i] \tag{20}$$

Using the original auxiliary equation:

$$x_{ji} = x_{(\ j)i}\tilde{\pi} + \tilde{\xi}_i$$

(21)

$$\frac{1}{N}\sum_i \hat{\xi}_i(x_{(\ j)i}\tilde{\pi} + \tilde{\xi}_i) \tag{22}$$

From the first order condition of the auxiliary equation we get that Xi is orthogonal to xi:

$$\sum_i \xi_i \perp \hat{\ } x_{(\ j)i}\tilde{\pi} \tag{23}$$

So we have proven what we set out to prove which is:

$$\hat{\beta}_j = [\frac{1}{N}\sum_i \hat{\xi}_i{}^2]^{-}1[\frac{1}{N}\sum_i \hat{\xi}_i y_i] \tag{24}$$

Problem 2

a)
We start from our true model:

$$logwage = \beta_0 + \beta_1 imm + \beta_2 educ + u_i \tag{25}$$

The first step is to regress education on immigration and so creating an auxiliary model for education:

$$educ = \pi_0 + \pi_1 imm + \epsilon_i \tag{26}$$

Plugging this into equation (25):

$$logwage = \beta_0 + \beta_1 imm + \beta_2(\pi_0 + \pi_1 imm + \epsilon_i) + u_i \tag{27}$$

Simplifying:

$$logwage = (\beta_0 + \beta_2 \pi_0) + (\beta_1 + \pi_1 \beta_2) imm + (u_i + \epsilon_i \beta_2) \tag{28}$$

In this equation we have decomposed the true model to only including immigration as a covariate. As a result, there will be bias within our model that stem from the auxiliary model for education (26).

.

In our model for Women (See table 1) it is evident that there exist omitted variable bias. The constant from regression 1 is 2.886 whilst the true model predicts the constant to be 1.241. In addition β_1 for the first regression is estimated to be -1.80 whilst the true model is -0.010.

The same goes for men(See table 2). The constant from regression 1 is 3.156 whilst the true model predicts the constant to be 1.657. In addition β_1 for the first regression is estimated to be -0.245 whilst the true model is -0.075.

.

Regarding when we model by ethnicity we see that Hispanic coefficient is more negative, asians and other, the beta is overestimated. Note this is true for both cases, men and women.

Table 1: Relationships between Log Wages, Education, and Immigration Status for Working Women Age 35-44 in March 2012 Current Population Survery

	(Logwage) I	(Logwage) II	(Logwage) III	(Immigration Status) I	(Education) I
const	2.89***	1.23***	0.61***	14.45***	1.24***
	(0.01)	(0.03)	(0.02)	(0.03)	(0.03)
educ		0.11***	-0.03***		0.11***
		(0.00)	(0.00)		(0.00)
imm	-0.18***			-1.49***	-0.01
	(0.02)			(0.07)	(0.01)
N	10601	10601	10601	10601	10601
R2	0.01	0.22	0.04	0.04	0.22

Table 2: Relationships between Log Wages, Education, and Immigration Status for Working Men Age 35-44 in March 2012 Current Population Survery

	(Logwage) I	(Logwage) II	(Logwage) III	(Immigration Status) I	(Education) I
const	3.16***	1.61***	0.64***	14.19***	1.66***
	(0.01)	(0.03)	(0.02)	(0.03)	(0.03)
educ		0.11***	-0.03***		0.11***
		(0.00)	(0.00)		(0.00)
imm	-0.24***			-1.61***	-0.07***
	(0.02)			(0.07)	(0.01)
N	11306	11306	11306	11306	11306
R2	0.02	0.22	0.05	0.05	0.22

Table 3: Regressions for Immigrant Women: Asian, Hispanic and Other

	(Logwage) I	(Logwage) II	(Asian) I	(Hispanic) I	(Other) I	Education I
as	0.09***					0.51***
	(0.03)					(0.12)
const	2.89***	1.23***	-0.02**	0.63***	-0.00	14.45***
	(0.01)	(0.03)	(0.01)	(0.01)	(0.01)	(0.03)
educ		0.11***	0.01***	-0.04***	0.00***	
		(0.00)	(0.00)	(0.00)	(0.00)	
hisp	-0.43***					-3.42***
	(0.02)					(0.09)
oth	0.05					0.28**
	(0.03)					(0.12)
N	10601	10601	10601	10601	10601	10601
R2	0.04	0.22	0.00	0.13	0.00	0.13

Table 4: Regressions for Immigrant Men: Asian, Hispanic and Other

	(Logwage) I	(Logwage) II	(Asian) I	(Hispanic) I	(Other) I	Education I
as	0.07**					1.26***
	(0.03)					(0.12)
const	3.16***	1.61***	-0.07***	0.74***	-0.03***	14.19***
	(0.01)	(0.03)	(0.01)	(0.01)	(0.01)	(0.03)
educ		0.11***	0.01***	-0.04***	0.01***	
		(0.00)	(0.00)	(0.00)	(0.00)	
hisp	-0.47***					-3.63***
	(0.02)					(0.08)
oth	0.01					0.67***
	(0.03)					(0.12)
N	11306	11306	11306	11306	11306	11306
R2	0.05	0.22	0.01	0.17	0.01	0.18

In [83]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import statsmodels.api as sm
from stargazer.stargazer import Stargazer
import tabulate as tb
from statsmodels.iolib.summary2 import summary_col
```

In [2]:
```
df = pd.read_csv("ovb.csv")
df["const"] = 1
df.head()
```

Out[2]:

	state	age	wagesal	imm	hispanic	black	asian	educ	wage	logwage	female	f
0	11	44	18000	0	0	0	0	14	9.109312	2.209297	1	1
1	11	39	18000	0	0	0	0	14	18.000000	2.890372	0	0
2	11	39	35600	0	0	0	0	12	17.115385	2.839978	0	0
3	11	39	8000	0	0	0	0	14	5.128205	1.634756	1	0
4	11	39	100000	0	0	0	0	16	38.461538	3.649659	0	1

In [44]:
```
## Women
##Women: Asian = 1, hispanic = 1, imm = 1
wasian = dfwomen[(dfwomen.asian == 1)  & (dfwomen.hispanic == 0)]

#Women: Hispanic = 1, imm = 1
whisp = dfwomen[(dfwomen.hispanic == 1) & (dfwomen.imm== 1)]

#Women: Other
wother = dfwomen[(dfwomen.black == 1) & (dfwomen.imm == 1)]
wasian.count()
df['hisp'] = 0
df.loc[(df.imm == 1) & (df.hispanic == 1), 'hisp'] = 1
df['as'] = 0
df.loc[(df.imm == 1) & (df.asian == 1) & (df.hispanic == 0), 'as'] = 1
df['oth'] = 0
df.loc[(df.imm == 1) & (df.asian == 0) & (df.hispanic == 0), 'oth'] = 1
```

In [67]:
```
dfwomen = df[df.female == 1]
dfmen = df[df.female == 0]
```

In [111]:
```
reg1 = sm.OLS(endog=dfwomen['logwage'], exog=dfwomen[['const', 'imm']],)
.fit()


reg2 = sm.OLS(endog=dfwomen['logwage'], exog=dfwomen[['const', 'educ']])
.fit()


reg3 = sm.OLS(endog=dfwomen['imm'], exog=dfwomen[['const', 'educ']]).fit
()


reg4 = sm.OLS(endog=dfwomen['educ'], exog=dfwomen[['const', 'imm']]).fit
()


reg5 = sm.OLS(endog=dfwomen['logwage'], exog=dfwomen[['const', 'educ',
'imm']]).fit()
```

In [112]:
```
print(summary_col([reg1,reg2,reg3, reg4, reg5],stars=True,float_format='
%0.2f', regressor_order=['black'],
                  model_names = ['(Logwage)','(Logwage)','(Logwage)','(I
mmigration Status)','(Education)'], info_dict={'N':lambda x: "{0:d}".for
mat(int(x.nobs)),
                             'R2':lambda x: "{:.2f}".format(x.rsquared
)}).as_latex())
```

```
\begin{table}
\caption{}
\begin{center}
\begin{tabular}{lccccc}
\hline
      & (Logwage) I & (Logwage) II & (Logwage) III & (Immigration Statu
s) I & (Education) I  \\
\midrule
\midrule
const & 2.89***     & 1.23***      & 0.61***       & 14.45***
     & 1.24***        \\
      & (0.01)      & (0.03)       & (0.02)        & (0.03)
     & (0.03)         \\
educ  &             & 0.11***      & -0.03***      &
     & 0.11***        \\
      &             & (0.00)       & (0.00)        &
     & (0.00)         \\
imm   & -0.18***    &              &               & -1.49***
     & -0.01          \\
      & (0.02)      &              &               & (0.07)
     & (0.01)         \\
N     & 10601       & 10601        & 10601         & 10601
     & 10601          \\
R2    & 0.01        & 0.22         & 0.04          & 0.04
     & 0.22           \\
\hline
\end{tabular}
\end{center}
\end{table}
```

In [113]:
```
reg1 = sm.OLS(endog=dfmen['logwage'], exog=dfmen[['const', 'imm']],).fit
()


reg2 = sm.OLS(endog=dfmen['logwage'], exog=dfmen[['const', 'educ']]).fit
()


reg3 = sm.OLS(endog=dfmen['imm'], exog=dfmen[['const', 'educ']]).fit()


reg4 = sm.OLS(endog=dfmen['educ'], exog=dfmen[['const', 'imm']]).fit()


reg5 = sm.OLS(endog=dfmen['logwage'], exog=dfmen[['const', 'educ', 'imm'
]]).fit()
```

In [114]:
```
print(summary_col([reg1,reg2,reg3, reg4, reg5],stars=True,float_format='
%0.2f', regressor_order=['black'],
                  model_names = ['(Logwage)','(Logwage)','(Logwage)','(I
mmigration Status)','(Education)'], info_dict={'N':lambda x: "{0:d}".for
mat(int(x.nobs)),
                              'R2':lambda x: "{:.2f}".format(x.rsquared
)}).as_latex())
```

```
\begin{table}
\caption{}
\begin{center}
\begin{tabular}{lccccc}
\hline
      & (Logwage) I & (Logwage) II & (Logwage) III & (Immigration Statu
s) I & (Education) I  \\
\midrule
\midrule
const & 3.16***     & 1.61***      & 0.64***       & 14.19***
     & 1.66***        \\
      & (0.01)      & (0.03)       & (0.02)        & (0.03)
     & (0.03)         \\
educ  &             & 0.11***      & -0.03***      &
     & 0.11***        \\
      &             & (0.00)       & (0.00)        &
     & (0.00)         \\
imm   & -0.24***    &              &               & -1.61***
     & -0.07***       \\
      & (0.02)      &              &               & (0.07)
     & (0.01)         \\
N     & 11306       & 11306        & 11306         & 11306
     & 11306          \\
R2    & 0.02        & 0.22         & 0.05          & 0.05
     & 0.22           \\
\hline
\end{tabular}
\end{center}
\end{table}
```

In [108]:
```
#Women by et

reg = sm.OLS(dfwomen.logwage, dfwomen[['const','as','hisp','oth']]).fit
()
reg1 = sm.OLS(dfwomen.logwage, dfwomen[['const','educ']]).fit()

reg2 = sm.OLS(dfwomen['as'], dfwomen[['const','educ']]).fit()

reg3 = sm.OLS(dfwomen['hisp'], dfwomen[['const','educ']]).fit()

reg4 = sm.OLS(dfwomen['oth'], dfwomen[['const','educ']]).fit()

reg5 = sm.OLS(dfwomen.educ, dfwomen[['const','as','hisp','oth']]).fit()
```

In [109]:
```
print(summary_col([reg,reg1,reg2, reg3, reg4,reg5],stars=True,float_form
at='%0.2f', regressor_order=['black'],
                  model_names = ['(Logwage)','(Logwage)','(Asian)','(His
panic)','(Other)','Education'], info_dict={'N':lambda x: "{0:d}".format(
int(x.nobs)),
                                'R2':lambda x: "{:.2f}".format(x.rsquared
)}).as_latex())
```

```
\begin{table}
\caption{}
\begin{center}
\begin{tabular}{lcccccc}
\hline
      & (Logwage) I & (Logwage) II & (Asian) I & (Hispanic) I & (Other)
I & Education I  \\
\midrule
\midrule
as    & 0.09***     &              &           &              &
   & 0.51***      \\
      & (0.03)      &              &           &              &
   & (0.12)       \\
const & 2.89***     & 1.23***      & -0.02**   & 0.63***      & -0.00
   & 14.45***     \\
      & (0.01)      & (0.03)       & (0.01)    & (0.01)       & (0.01)
   & (0.03)       \\
educ  &             & 0.11***      & 0.01***   & -0.04***     & 0.00***
   &              \\
      &             & (0.00)       & (0.00)    & (0.00)       & (0.00)
   &              \\
hisp  & -0.43***    &              &           &              &
   & -3.42***     \\
      & (0.02)      &              &           &              &
   & (0.09)       \\
oth   & 0.05        &              &           &              &
   & 0.28**       \\
      & (0.03)      &              &           &              &
   & (0.12)       \\
N     & 10601       & 10601        & 10601     & 10601        & 10601
   & 10601        \\
R2    & 0.04        & 0.22         & 0.00      & 0.13         & 0.00
   & 0.13         \\
\hline
\end{tabular}
\end{center}
\end{table}
```

In [115]:
```
##Men

reg = sm.OLS(dfmen.logwage, dfmen[['hisp','as','oth','const']]).fit()
reg1 = sm.OLS(dfmen.logwage, dfmen[['educ','const']]).fit()

reg2 = sm.OLS(dfmen['as'], dfmen[['educ','const']]).fit()

reg3 = sm.OLS(dfmen['hisp'], dfmen[['educ','const']]).fit()

reg4 = sm.OLS(dfmen['oth'], dfmen[['educ','const']]).fit()

reg5 = sm.OLS(dfmen.educ, dfmen[['const','as','hisp','oth']]).fit()
```

In [117]:
```
print(summary_col([reg,reg1,reg2, reg3, reg4,reg5],stars=True,float_form
at='%0.2f', regressor_order=['black'],
                   model_names = ['(Logwage)','(Logwage)','(Asian)','(His
panic)','(Other)','Education'], info_dict={'N':lambda x: "{0:d}".format(
int(x.nobs)),
                                 'R2':lambda x: "{:.2f}".format(x.rsquared
)}).as_latex())
```

```
\begin{table}
\caption{}
\begin{center}
\begin{tabular}{lcccccc}
\hline
      & (Logwage) I & (Logwage) II & (Asian) I & (Hispanic) I & (Other)
I & Education I  \\
\midrule
\midrule
as    & 0.07**      &              &           &              &
   & 1.26***      \\
      & (0.03)      &              &           &              &
   & (0.12)       \\
const & 3.16***     & 1.61***      & -0.07***  & 0.74***      & -0.03**
*  & 14.19***     \\
      & (0.01)      & (0.03)       & (0.01)    & (0.01)       & (0.01)
   & (0.03)       \\
educ  &             & 0.11***      & 0.01***   & -0.04***     & 0.01***
   &              \\
      &             & (0.00)       & (0.00)    & (0.00)       & (0.00)
   &              \\
hisp  & -0.47***    &              &           &              &
   & -3.63***     \\
      & (0.02)      &              &           &              &
   & (0.08)       \\
oth   & 0.01        &              &           &              &
   & 0.67***      \\
      & (0.03)      &              &           &              &
   & (0.12)       \\
N     & 11306       & 11306        & 11306     & 11306        & 11306
   & 11306        \\
R2    & 0.05        & 0.22         & 0.01      & 0.17         & 0.01
   & 0.18         \\
\hline
\end{tabular}
\end{center}
\end{table}
```