Binary Dependent Variables Assignment

Jose Fernandez

MSBA Data Analytics III

Question 1

We will examine a paper by Anastasia Semykina entitled, “Self-employment among women: Do children matter more than we previously thought?”. You are provided the following data.

March CPS white women (NLSY_white_women_JAE.csv): 12,624 women for several year. You observe the following variables

Variable	definitions:
id	unique individual ID
year	year
working	=1 if working, otherwise
self_empl	=1 if self-employed, otherwise
age	age in years
agesq	age squared
educ	years of schooling, truncated at 20 years
edu_0_11	=1 if has 0-11 years of schooling, 0 otherwise
edu_12	=1 if has 12 years of schooling, 0 otherwise
edu_13_15	=1 if has 13-15 years of schooling, 0 otherwise
edu_16plus	=1 if has 16 or more years of schooling, 0 otherwise
married	=1 if married, 0 if not
d_ch_1_5	=1 if has children ages 0 to 5, otherwise
d_ch_0	=1 if has a newborn (<1 years old), otherwise
d_ch_1_5_alt	=1 if has children ages 1 to 5, otherwise
d_ch_6_17	=1 if has children ages 6 to 17, otherwise
rotter_score	locus of control
sesteem_score1	self esteem score
urban	=1 if urban location, 0 otherwise
afqt_1	AFQT score
south	=1 if South region, otherwise
northeast	=1 if Northeast region, otherwise

Variable	definitions:
northcen	=1 if North Central region, otherwise
west	=1 if West region, otherwise
sp_inc1000	spouse’s income in thousands of dollars
samesex	=1 if the first two children have the same gender, otherwise
policever	=1 if ever stopped by police for other than minor traffic offense in 1980, 0 otherwise
unemp_rate	unemployment rate in percentage points
m_sp_inc1000	individual time mean of sp_inc1000
m_married	individual time mean of married

```
options(digits = 5)
library(texreg)
library(stargazer)
library(sampleSelection)
library(mfx)
library(tidyverse)
library(readr)
NLSY <- read_csv("NLSY_white_women_JAE (2).csv")
```

a. Estimate a linear probability model of self employment and a separate linear probability model of working.

```
ols.work<-lm(working~age+agesq+educ+married+d_ch_1_5+d_ch_0+d_ch_6_17+rotter_score+sesteem_score
1+urban+afqt_1+south+northeast+northcen+sp_inc1000+policever+unemp_rate+factor(year),data = NLS
Y)
ols.work2 <-lm(working~age+agesq+edu_12+edu_13_15+edu_16plus+married+d_ch_1_5+d_ch_0+d_ch_6_17+r
otter_score+sesteem_score1+urban+afqt_1+south+northeast+northcen+sp_inc1000+policever+unemp_rate
+factor(year),data = NLSY)
ols.self<-lm(self_empl~age+agesq+educ+married+d_ch_1_5+d_ch_0+d_ch_6_17+rotter_score+sesteem_sco
re1+urban+afqt_1+south+northeast+northcen+sp_inc1000+policever+unemp_rate+factor(year),data = NL
SY)
ols.self2 <-lm(self_empl~age+agesq+edu_12+edu_13_15+edu_16plus+married+d_ch_1_5+d_ch_0+d_ch_6_17
+rotter_score+sesteem_score1+urban+afqt_1+south+northeast+northcen+sp_inc1000+policever+unemp_ra
te+factor(year),data = NLSY)
stargazer(ols.work,ols.self, type="html",keep=c("age","agesq","educ","married","d_ch_1_5","d_ch_
0","d_ch_6_17","rotter_score","sesteem_score1","afqt_1","sp_inc1000","policever","unemp_rate"),c
ovariate.labels = c("Age","Age Squared","Education (Years)","Married","Children between 1 to 5 y
rs","New Borns","School Aged Children","Rotter Score","Self-Esteem Score","AFQT Score","Current
 Spousal Income","Ever Stopped by Police","Unemployment Rate"),dep.var.labels = c("Working","Sel
f-Employed"),header = FALSE)
```

	Dependent variable:	
	Working	Self-Employed
	(1)	(2)
Age	-0.003	0.010**
	(0.005)	(0.004)

Age Squared	-0.00003	-0.0001
	(0.0001)	(0.0001)
Education (Years)	0.014^{***}	-0.001
	(0.001)	(0.001)
Married	0.053^{***}	0.001
	(0.005)	(0.004)
Children between 1 to 5 yrs	-0.170^{***}	0.044^{***}
	(0.004)	(0.004)
New Borns	-0.014^{*}	-0.006
	(0.007)	(0.006)
School Aged Children	-0.052^{***}	0.022^{***}
	(0.004)	(0.004)
Rotter Score	0.001	-0.002^{***}
	(0.001)	(0.001)
Self-Esteem Score	0.002^{***}	0.001^{**}
	(0.0005)	(0.0004)
AFQT Score	0.001^{***}	0.0001
	(0.0001)	(0.0001)
Current Spousal Income	-0.002^{***}	0.001^{***}
	(0.0001)	(0.0001)
Ever Stopped by Police	-0.013^{*}	0.027^{***}
	(0.007)	(0.005)
Unemployment Rate	-0.006^{*}	0.001
	(0.003)	(0.003)
Observations	33,365	28,228
R^2	0.120	0.027
Adjusted R^2	0.119	0.025
Residual Std. Error	0.339 (df = 33329)	0.253 (df = 28192)
F Statistic	129.720^{***} (df = 35; 33329)	21.935^{***} (df = 35; 28192)
Note:		*p<0.1;* **p<0.05;** p<0.01

```
stargazer(ols.work2,ols.self2, type="html",keep=c("age","agesq","edu_12","edu_13_15","edu_16plus","married","d_ch_1_5","d_ch_0","d_ch_6_17","rotter_score","sesteem_score1","afqt_1","sp_inc1000","policever","unemp_rate"),covariate.labels = c("Age","Age Squared","High School Graduate","Some College","College Graduate","Married","Children between 1 to 5 yrs","New Borns","School Aged Children","Rotter Score","Self-Esteem Score","AFQT Score","Current Spousal Income","Ever Stopped by Police","Unemployment Rate"),dep.var.labels = c("Working","Self-Employed"),header = FALSE)
```

	Dependent variable:	
	Working	Self-Employed
	(1)	(2)
Age	-0.003	0.010^{**}
	(0.005)	(0.004)
Age Squared	-0.00004	-0.0001
	(0.0001)	(0.0001)
High School Graduate	0.141^{***}	0.0002
	(0.007)	(0.006)
Some College	0.164^{***}	-0.003
	(0.008)	(0.007)

College Graduate	0.167***	-0.006
	(0.009)	(0.008)
Married	0.045***	0.001
	(0.005)	(0.004)
Children between 1 to 5 yrs	-0.167***	0.044***
	(0.004)	(0.004)
New Borns	-0.013*	-0.006
	(0.007)	(0.006)
School Aged Children	-0.047***	0.022***
	(0.004)	(0.004)
Rotter Score	0.001	-0.002***
	(0.001)	(0.001)
Self-Esteem Score	0.002***	0.001**
	(0.0005)	(0.0004)
AFQT Score	0.001***	0.0001
	(0.0001)	(0.0001)
Current Spousal Income	-0.002***	0.001***
	(0.0001)	(0.0001)
Ever Stopped by Police	-0.011*	0.027***
	(0.007)	(0.005)
Unemployment Rate	-0.006*	0.001
	(0.003)	(0.003)
Observations	33,365	28,228
R^2	0.128	0.027
Adjusted R^2	0.127	0.025
Residual Std. Error	0.337 (df = 33327)	0.253 (df = 28190)
F Statistic	132.210*** (df = 37; 33327)	20.746*** (df = 37; 28190)
Note:		*p<0.1;* ***p<0.05;*** p<0.01

Note, I included year dummies in my regression. It is ok if you did not, but it may change some of your responses to the questions below. I also use the location characteristics like urban and region, but do not report them in the table. It is fine if you reported them.

I also reported a second set of tables that include education as a set of dummies.

b. Estimate the same two models from a) using either probit or logit and report the marginal effects.

```
probit.work<-probitmfx(working~age+agesq+educ+married+d_ch_1_5+d_ch_0+d_ch_6_17+rotter_score+sesteem_score1+urban+afqt_1+south+northeast+northcen+sp_inc1000+policever+unemp_rate+factor(year),data = NLSY)
probit.self<-probitmfx(self_empl~age+agesq+educ+married+d_ch_1_5+d_ch_0+d_ch_6_17+rotter_score+sesteem_score1+urban+afqt_1+south+northeast+northcen+sp_inc1000+policever+unemp_rate+factor(year),data = NLSY)
htmlreg(list(probit.work,probit.self),omit.coef = "(year)|(urban)|(south)|(north)",custom.coef.names= c("Age","Age Squared","Education (Years)","Married","Children between 1 to 5 yrs","New Borns","School Aged Children","Rotter Score","Self-Esteem Score","AFQT Score","Current Spousal Income","Ever Stopped by Police","Unemployment Rate"), digits = 4)
```

	Model 1	**Model 2**
Age	-0.0052	0.0126***
p < 0.001; *p < 0.01;* p < 0.05		

	Model 1	**Model 2**
	(0.0045)	(0.0037)
Age Squared	-0.0000	-0.0001*
	(0.0001)	(0.0001)
Education (Years)	0.0157***	-0.0016*
	(0.0011)	(0.0008)
Married	0.0243***	0.0102**
	(0.0047)	(0.0035)
Children between 1 to 5 yrs	-0.1720***	0.0426***
	(0.0050)	(0.0040)
New Borns	-0.0164*	-0.0021
	(0.0064)	(0.0053)
School Aged Children	-0.0471***	0.0184***
	(0.0043)	(0.0035)
Rotter Score	0.0007	-0.0018**
	(0.0008)	(0.0007)
Self-Esteem Score	0.0022***	0.0010**
	(0.0005)	(0.0004)
AFQT Score	0.0012***	0.0001
	(0.0001)	(0.0001)
Current Spousal Income	-0.0012***	0.0003***
	(0.0001)	(0.0000)
Ever Stopped by Police	-0.0102	0.0296***
	(0.0064)	(0.0061)
Unemployment Rate	-0.0065*	0.0014
	(0.0033)	(0.0027)
Num. obs.	33365	28228
Log Likelihood	-12245.4057	-6863.8169
Deviance	24490.8114	13727.6338
AIC	24562.8114	13799.6338
BIC	24865.7608	14096.5643

p < 0.001; *p < 0.01;* p < 0.05

Statistical models

```
probit.work2<-probitmfx(working~age+agesq+edu_12+edu_13_15+edu_16plus+married+d_ch_1_5+d_ch_0+d_ch_6_17+rotter_score+sesteem_score1+urban+afqt_1+south+northeast+northcen+sp_inc1000+policever+unemp_rate+factor(year),data = NLSY)
probit.self2<-probitmfx(self_empl~age+agesq+edu_12+edu_13_15+edu_16plus+married+d_ch_1_5+d_ch_0+d_ch_6_17+rotter_score+sesteem_score1+urban+afqt_1+south+northeast+northcen+sp_inc1000+policever+unemp_rate+factor(year),data = NLSY)
htmlreg(list(probit.work2,probit.self2),omit.coef = "(year)|(urban)|(south)|(north)",custom.coef.names= c("Age","Age Squared","High School Graduate","Some College","College Graduate","Married","Children between 1 to 5 yrs","New Borns","School Aged Children","Rotter Score","Self-Esteem Score","AFQT Score","Current Spousal Income","Ever Stopped by Police","Unemployment Rate"), digits = 4)
```

	Model 1	**Model 2**
Age	-0.0063	0.0125***

p < 0.001; *p < 0.01;* p < 0.05

	Model 1	Model 2
	(0.0045)	(0.0037)
Age Squared	-0.0000	-0.0001*
	(0.0001)	(0.0001)
High School Graduate	0.0901***	-0.0030
	(0.0054)	(0.0058)
Some College	0.0959***	-0.0063
	(0.0046)	(0.0063)
College Graduate	0.1139***	-0.0109
	(0.0053)	(0.0066)
Married	0.0192***	0.0103**
	(0.0047)	(0.0035)
Children between 1 to 5 yrs	-0.1705***	0.0426***
	(0.0050)	(0.0040)
New Borns	-0.0158*	-0.0021
	(0.0064)	(0.0053)
School Aged Children	-0.0437***	0.0184***
	(0.0043)	(0.0035)
Rotter Score	0.0010	-0.0018**
	(0.0008)	(0.0007)
Self-Esteem Score	0.0020***	0.0010**
	(0.0005)	(0.0004)
AFQT Score	0.0012***	0.0001
	(0.0001)	(0.0001)
Current Spousal Income	-0.0012***	0.0003***
	(0.0001)	(0.0000)
Ever Stopped by Police	-0.0102	0.0298***
	(0.0065)	(0.0061)
Unemployment Rate	-0.0065*	0.0014
	(0.0033)	(0.0027)
Num. obs.	33365	28228
Log Likelihood	-12178.6213	-6863.7516
Deviance	24357.2426	13727.5033
AIC	24433.2426	13803.5033
BIC	24753.0226	14116.9299

p < 0.001; *p < 0.01;* p < 0.05

Statistical models

c. Are your estimates between parts a and b similar? Please interpret your results.

Our Results between the linear probability model and the probit model are similar. There are a few exceptions. The effects of police and marriage are different. Marital status is statistically significant in the probit specification of self-employment, but not in the linear specification model. Education is found to be statistically significant for self-employment in the probity specification, but not in the linear specification. Finally, “ever being stopped by police” reduces the likelihood of working in the linear specification, but not in the probit specification.

Across both models we find that the likelihood of self-employment increases and decreasing rate. If you did not use year fixed effects, then you would find this result for working too. Marriage is found to increase the likelihood of work, but having children decreases the likelihood of work. Self-esteem increases both the likelihood of work and self-employment. The amount your spouse earns decreases your likelihood of work, but increases your likelihood of self-employment if you do work. The spousal income can be viewed as additional capital for starting a business. Similarly, if you are ever stopped by police, we find this decreases your likelihood of work, but increases your likelihood of self-employment. These people may be pushed into entrepreneurial activities.

d. Consider what variables would affect your likelihood of working, but not necessarily your likelihood of becoming self-employed?

We will focus our attention on children. The presents of children decreases the amount of time available for work. Similarly, this decrease amount of time for work would make it more difficult to have your own business. On the other hand, the presence of children could make it easier to have a family owned business.

For this reason it would seem that children is a useful instrument for the likelihood of work. Although, it could fail the exclusion restriction when thinking about family businesses.

e. Now consider a sample selection model where the woman first decided whether or not to work and then decides if she should be self-employed. What variable do you choose as your instrument? That is, what variable affects your decision to work, but not your decision to be self-employed? Provide some reasoning for your answer.

A similar answer to those above would suffice. I am not too picky here about the response.

f. Estimate a Heckman two-step equation to correct for sample selection in your self-employment equation.

```
heck.1<-heckit(working~age+agesq+educ+married+d_ch_1_5+d_ch_0+d_ch_6_17+rotter_score+sesteem_score1+urban+afqt_1+south+northeast+northcen+sp_inc1000+policever+unemp_rate + factor(year), self_empl~age+agesq+educ+married+rotter_score+sesteem_score1+urban+afqt_1+south+northeast+northcen+sp_inc1000+policever+unemp_rate+factor(year), data = NLSY, method = "2step")
htmlreg(list(heck.1),omit.coef = "(year)|(urban)|(south)|(north)",custom.coef.names= c("Constant","Age","Age Squared","Education (Years)","Married","Rotter Score","Self-Esteem Score","AFQT Score","Current Spousal Income","Ever Stopped by Police","Unemployment Rate","Inverse Mills Ratio","STD","Rho"), digits = 4, include.selection = FALSE)
```

	Model 1
Constant	-0.2580***
	(0.0679)
Age	0.0118**
	(0.0039)
Age Squared	-0.0001*
	(0.0001)
Education (Years)	0.0008
	(0.0009)
Married	0.0142***
	(0.0038)
Rotter Score	-0.0017*
	(0.0007)

p < 0.001; *p < 0.01;* p < 0.05

	Model 1
Self-Esteem Score	0.0014***
	(0.0004)
AFQT Score	0.0004***
	(0.0001)
Current Spousal Income	0.0002***
	(0.0001)
Ever Stopped by Police	0.0249***
	(0.0056)
Unemployment Rate	-0.0000
	(0.0029)
Inverse Mills Ratio	0.1360***
	(0.0120)
STD	0.2642
Rho	0.5147
R^2	0.0247
Adj. R^2	0.0235
Num. obs.	33365
Censored	5137
Observed	28228

$p < 0.001$; *$p < 0.01$;* $p < 0.05$

Statistical models

```
heck.2<-heckit(working~age+agesq+edu_12+edu_13_15+edu_16plus+married+d_ch_1_5+d_ch_0+d_ch_6_17+rotter_score+sesteem_score1+urban+afqt_1+south+northeast+northcen+sp_inc1000+policever+unemp_rate + factor(year), self_empl~age+agesq+edu_12+edu_13_15+edu_16plus+married+rotter_score+sesteem_score1+urban+afqt_1+south+northeast+northcen+sp_inc1000+policever+unemp_rate+factor(year), data = NLSY, method = "2step")
htmlreg(list(heck.2),omit.coef = "(year)|(urban)|(south)|(north)",custom.coef.names= c("Constant","Age","Age Squared","High School Graduate","Some College","College Graduate","Married","Rotter Score","Self-Esteem Score","AFQT Score","Current Spousal Income","Ever Stopped by Police","Unemployment Rate","Inverse Mills Ratio","STD","Rho"), digits = 4, include.selection = FALSE)
```

	Model 1
Constant	-0.2682***
	(0.0678)
Age	0.0117**
	(0.0039)
Age Squared	-0.0001*
	(0.0001)
High School Graduate	0.0258***
	(0.0069)
Some College	0.0263***
	(0.0080)
College Graduate	0.0227**
	(0.0085)

$p < 0.001$; *$p < 0.01$;* $p < 0.05$

	Model 1
Married	0.0133^{***}
	(0.0038)
Rotter Score	-0.0016^{*}
	(0.0007)
Self-Esteem Score	0.0014^{**}
	(0.0004)
AFQT Score	0.0003^{***}
	(0.0001)
Current Spousal Income	0.0002^{***}
	(0.0001)
Ever Stopped by Police	0.0252^{***}
	(0.0056)
Unemployment Rate	-0.0001
	(0.0029)
Inverse Mills Ratio	0.1390^{***}
	(0.0124)
STD	0.2646
Rho	0.5255
R^2	0.0246
Adj. R^2	0.0234
Num. obs.	33365
Censored	5137
Observed	28228

$p < 0.001$; *$p < 0.01$;* $p < 0.05$

Statistical models

g. Are there any differences between your results from c) and f)

We see from the regressions that the inverse mills ratio is statistically significant, which indicates sample selection is present.

The positive coefficient indicates the OLS coefficients are too large due to the sample selection bias