Econ 21410 - Problem Set V

Selection Part II*

John Eric Humphries†

May 19, 2014

This homework should be done in LaTeX The homework will be graded on correctness, but will also heavily weight clarity and professionalism. Its better to not do some of the harder parts than to turn in an incomprehensible document. Your R script as well as a log-file should be submitted. Alternatively, use knitr to print your code inline in your latex document.

SUBMISSION: The homework must be emailed to Oliver and myself by 2p.m. Monday, May 5th. The email must include a pdf with the filename lastname_pset5.pdf and R code called lastname_pset5_code.R where "lastname" should be replaced with your last name. The subject of your email should be [ECON 21410: pset5 submission]

If you are struggling, please use the github page to ask for help!* Remember that asking and answering questions on our github page, coming to office hours to ask questions, and contributing to the class wiki are all worth participation credit, which is 10% of your grade in this class.

In this homework we will be estimating selection-corrected regressions on real data. We will use a pre-packaged function, build our own two-step estimator. and our own semi-parametric estimator.

Building a data set

1. Go to usa.ipums.org and click on "Select Data".

2. First, click on "Change Sample" and select only the 2012 ACS (not the 3 or 5 year) and click "submit sample selection".

3. Now navigate the variable selection menu and select AGE, NCHILD, NCHLT5, SEX, MARST, RACE, EDUC, EMPSTAT, WKSWORK2, UHRSWORK, and INCWAGE.

4. Once selected click "View Cart", then "Create data extract". The data set will be really large, but we can make it a bit smaller right away by next clicking on "Select cases" and check "AGE" and "SEX" and click "submit". Select ages 25 - 55 and select females and click "Submit".

5. Next, select "change" in the "Data format" row, in the new screen select "Comma Delimited" and click and click "Submit.

*Please email johneric@uchicago.edu and obrowne@uchicago.edu if you have questions.
†Please email johneric@uchicago.edu and obrowne@uchicago.edu if you have questions.

6. Finally, click "Submit extract". At this point you will need to wait until ipums emails you that your data is ready.

7. At some point it will ask you to log-in / sign up. Go ahead and make an account, it only takes a couple of minutes.

8. When you get your email download your data as well as the codebooks/documentation!

This is going to be a pretty large data set (over 100MB), so avoid adding additional variables or making multiple copies.

Getting started

1. Read the data in to R (hint, use the "Import Dataset" button under the "Environment" tab. Make sure to include headers)

2. WKSWORK2 is an indicator variable for ranges of weeks worked. Construct a new variable called "wks" that has the same number of values, but has values equal to the middle of each bin (for example, equal to 48.5 if WKSWORK2==5).

3. Make a histogram of INCWAGE. What is funny about it (hint, read the documentation available with your download)? For now, set the problematic cases equal to NA (UPDATE: I have been told that once you select down to the stated sample that you no longer have topcoded observations. If this is the case, just make note of this.)

4. Using "RACE", construct a variable called "white" which is equal to 1 if the individual is white and 0 otherwise (we call such binary variables dummy variables). Make another dummy variable for if the individual is black.

5. If you did not select ages 25-55 when downloading your data, drop all ages less than 25 and greater than 55. Similarly drop all males if they are included.

6. Using MARST construct a dummy variable for if the individual is currently married called "married" and a dummy variable for if the individual was ever previously married (so divorced, widowed, separated, etc) called "wasmarried".

7. Using WAGEINC, "wks" and UHRSWORK to construct our best guess and individual's hourly wages. Also construct log hourly wages. These two variables will have many "NaN", "Inf", and "-Inf" values. Set these all equal to "NA"(hint: $is.na(acs\$hrw) = is.nan(acs\$hrw)$).

8. Make a histogram of your hourly wage variable. What are the smallest (non-zero) and largest hourly wages you find?

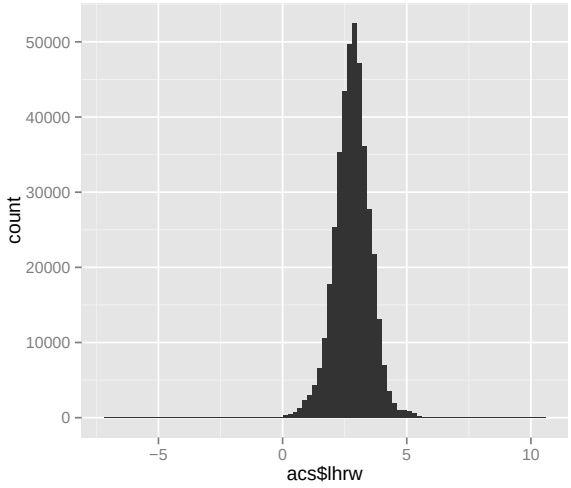
9. Make a dummy variable called "emp" which is equal to 1 if EMPSTAT==1, and equal to 0 if EMPSTAT==2 or EMPSTAT==3. Set this variable equal to "NA" if EMPSTAT==0. Read the documentation and explain how we should interpret the "emp" variable (1 sentence).

10. Create $acs\$fEmp = factor(EMPSTAT)$. This is a variable of class "factor" which we will discuss in class.

```r
# Section 1: data cleaning ========================

# Loading Data
acs <- read.csv("~/Dropbox/teaching/computationalEconomics/modules/n4_selection/usa_00011.cs
names(acs)

# Filtering by age
acs <- acs[(acs$AGE > 25 & acs$AGE < 55), ]

# Handling top codes
is.na(acs$INCWAGE) <- acs$INCWAGE[acs$INCWAGE == 999999]

# Building log wage income and log hourly wage.
vals <- 0:6
wkcat <- c(0, 6.5, 20, 35, 43, 48.5, 51)
for (i in 1:7) acs$wks[acs$WKSWORK2 == vals[i]] <- wkcat[i]

# Hrly and log wages
acs$lninc <- log(acs$INCWAGE)
acs$hrw <- (acs$INCWAGE/(acs$UHRSWORK * acs$wks))
is.na(acs$hrw) <- is.nan(acs$hrw)
acs$hrw[acs$hrw == Inf] <- NA
acs$lhrw <- log(acs$hrw)
is.na(acs$lhrw) <- is.infinite(acs$lhrw)

# Race
acs$white <- acs$RACE == 1
acs$black <- acs$RACE == 2

# Married
acs$married <- acs$MARST == 1
acs$wasmarried <- (acs$MARST > 1 & acs$MARST < 6)

# Empstat
acs$emp <- acs$EMPSTAT == 1
is.na(acs$emp) <- acs$EMPSTAT == 0

# Redefining Education levels to make them more
# interpretable
acs$fEDUC <- NA
acs$fEDUC[acs$EDUC >= 1 & acs$EDUC <= 5] <- "0:<HS"
acs$fEDUC[acs$EDUC == 6] <- "1:HS diploma"
acs$fEDUC[acs$EDUC >= 7 & acs$EDUC <= 9] <- "2:Some College"
acs$fEDUC[acs$EDUC == 10] <- "3:Bachelors"
acs$fEDUC[acs$EDUC == 11] <- "4:Graduate"
acs$fEDUC <- factor(acs$fEDUC)
acs <- acs[complete.cases(acs), ]

wageDistn <- qplot(acs$lhrw, binwidth = 0.2)
ggsave(file = "wageDistn.pdf", path = "~/n4_selection")
```

```
# ======================
```

Figure 1: Wage Distribution



Selection Correction

The original Heckman 1976 and 1979 papers considered women's labor supply. The Z, or "exclusion restrictions" were variables related to marriage and the household such as marital status, number of children, and number of young children. We will use number of children and number of children less than 5 as our Z.

Using a pre-written package

1. Run a regression of Log wages on the education factor, age, age^2, white, black, married, and was married. Make a table and discuss your results (2-4 sentences). Why do I ask you to include both age and age squared (1-2 sentences)?

2. Run a probit regression of emp on the education factor, age, age^2, white, black, married, was married, number of children, and number of children under 5. Make a table and discuss your results (2-4 sentences.)

3. Using the package "sampleSelection", use the command "heckit()" to run the 2-step heckman selection procedure discussed in class. Your selection equation should be the same equation as used in your probit above. Your outcome equation should be the same as the one used in the regression above. Make a table of the results. How do your coefficients in your outcome equation differ from those in your linear regression above? Are the coefficients you get from the "selection" stage the same as the ones you get in the probit above?

4. Do you think the number of children and the number of children under age 5 are valid exclusion restrictions? Why or why not?

```r
# Section 2: package regressions
# ========================

# OLS
OLS.lhrw <- lm(lhrw ~ fEDUC + AGE + I(AGE^2) + white +
    black + married + wasmarried, data = acs)

# Probit
probit.emp <- glm(emp ~ fEDUC + AGE + I(AGE^2) + white +
    black + married + wasmarried + NCHILD + NCHLT5,
    family = binomial(link = "probit"), data = acs)

# Heckit
heckit.lhrw <- heckit(emp ~ fEDUC + AGE + I(AGE^2) +
    white + black + married + wasmarried + NCHILD +
    NCHLT5, lhrw ~ fEDUC + AGE + I(AGE^2) + white +
    black + married + wasmarried, data = acs)

# ========================
```

Table 1:

	Dependent variable:
	emp
fEDUC1:HS diploma	0.257*** (0.012)
fEDUC2:Some College	0.366*** (0.012)
fEDUC3:Bachelors	0.493*** (0.013)
fEDUC4:Graduate	0.638*** (0.014)
AGE	0.050*** (0.004)
I(AGE^2)	−0.0005*** (0.00005)
white	0.028*** (0.009)
black	−0.038*** (0.012)
married	0.042*** (0.008)
wasmarried	−0.063*** (0.009)
NCHILD	−0.029*** (0.003)
NCHLT5	−0.081*** (0.007)
Constant	−0.097 (0.076)
Observations	416,574
Log Likelihood	−109,637.400
Akaike Inf. Crit.	219,300.900
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 2:

	Dependent variable:		
	lhrw		
	OLS	*Heckman selection*	*OLS*
	(1)	(2)	(3)
fEDUC1:HS diploma	0.277*** (0.005)	0.326*** (0.008)	0.326*** (0.008)
fEDUC2:Some College	0.483*** (0.005)	0.549*** (0.009)	0.549*** (0.009)
fEDUC3:Bachelors	0.814*** (0.005)	0.890*** (0.011)	0.890*** (0.011)
fEDUC4:Graduate	1.066*** (0.005)	1.151*** (0.013)	1.151*** (0.012)
AGE	0.054*** (0.001)	0.061*** (0.002)	0.061*** (0.002)
I(AGE^2)	−0.001*** (0.00002)	−0.001*** (0.00002)	−0.001*** (0.00002)
white	−0.006** (0.003)	−0.004 (0.003)	−0.004 (0.003)
black	−0.057*** (0.004)	−0.065*** (0.004)	−0.065*** (0.004)
married	0.059*** (0.003)	0.058*** (0.003)	0.058*** (0.003)
wasmarried	−0.008** (0.003)	−0.013*** (0.004)	−0.013*** (0.004)
invMillsRatio			0.449*** (0.065)
Constant	0.948*** (0.026)	0.675*** (0.047)	0.675*** (0.045)
Observations	416,574	416,574	384,860
R^2	0.194	0.207	0.207
Adjusted R^2	0.194	0.207	0.207
ρ		0.679	
Inverse Mills Ratio		0.449*** (0.067)	
Residual Std. Error	0.655 (df = 416563)		0.623 (df = 384848)
F Statistic	10,009.500*** (df = 10; 416563)		9,147.615*** (df = 11; 384848)
Note:			*p<0.1; **p<0.05; ***p<0.01

The two-step procedure by hand

Now we will build our own two-step selection correction estimates.

1. In words explain the steps to estimating the heckman selection correction when we assume normality.

2. Write down the equation for the inverse-mills ratio for working women.

3. Using the saved output of your probit model above to extract $X\beta + Z\alpha$ for each individual.

4. Estimate the inverse-mills ratio for each individual.

5. Run your regression from above, but include the inverse-mills ratio as a regressor. Do you get the same point estimates on your other coefficents as the heckit command above? What about standard errors? Why are the standard errors you get here incorrect. Make a table of your regression results.

```
# Section 3: Two step regression by Hand
# ========================

# Extract Inverse Mills Ratio from Probit Estimates
acs$invMillsRatio <- invMillsRatio(probit.emp)$IMR1
# Run Regression on existing model with inverse
# mills ratio as regressor
heckit.lhrw.hand <- lm(lhrw ~ fEDUC + AGE + I(AGE^2) +
    white + black + married + wasmarried + invMillsRatio,
    data = acs, subset = (emp == TRUE))

# ========================
```

Relaxing the normality assumption.

1. In words explain the steps to estimating the heckman selection proceedure when we use a polynomial of the probability of being employed rather than the inverse-mills ratio.

2. Estimate a probit with the same regressors as the probit above, but fully interacted (warning, this may take several minutes to run on your computer). Explain the risk of over-fitting and how we may go about choosing the number of parameters to estimate in a smarter way than we just did.

3. Extract the probabilities from your probit (the fitted values). Make a histogram.

4. Run the original regression above, but include the probabilities directly. First include the probability, then add a 2nd and 3rd order term (i.e., squared and cubed), then add the 4th-10th ordered term. Make a table of the coefficients for these three regressions.

5. How do the coefficients on your education categories, race dummies, marriage dummies, age, and age squared change as you increase the length oft he polynomial of probabilities above?

6. How do the coefficients on your education categories, race dummies, marriage dummies, age, and age squared compare to the coefficients from the "heckit()" command above which assumed normality?

7. Would you say the differences between "lm()", "heckit()", and your semi-parametric approach are economically meaningful?

```
# Section 4: Semi-parametric Selection Correction
# ======================

# Run fully interacted probit on all varaibles
probit.interacted.emp <- glm(emp ~ (factor(EDUC) +
    AGE + I(AGE^2) + white + black + married + wasmarried +
    NCHILD + NCHLT5)^2, family = binomial(link = "probit"),
    data = acs)
# Predict the probabilities
predictions <- predict(probit.interacted.emp, type = "response")
# Generate Histogram of fitted values
qplot(predictions, binwidth = 0.01)

# Initialize a table to store all coefficents
coef.table <- matrix(rep(NA, 21 * 9), ncol = 9)
# Loop over polynomials of degree 2 to 10
for (i in 2:10) {
    # Run regression on existing model but adding in
    # regressors for the probabilities raised to the
    # ith degree
    semiparam.heckit <- lm(lhrw ~ fEDUC + AGE + I(AGE^2) +
        white + black + married + wasmarried + I(poly(predictions,
        i, raw = TRUE)), data = acs, subset = (acs$emp ==
        TRUE))

    # Store coefficents in the table
    l <- length(semiparam.heckit$coefficients)
    coef.table[1:l, i - 1] <- semiparam.heckit$coefficients
}

# Name table rose and columns
rownames(coef.table) <- names(semiparam.heckit$coefficients)
rownames(coef.table)[12:21] <- paste("predictions^",
    1:10)
colnames(coef.table) <- paste("poly", 2:10)

# ========================
```

Table 3:

	poly 2	poly 3	poly 4	poly 5	poly 6	poly 7	poly 8	poly 9	poly 10
fEDUC1:HS diploma	0.280	0.281	0.281	0.280	0.280	0.280	0.280	0.280	0.281
fEDUC2:Some College	0.481	0.483	0.482	0.482	0.482	0.482	0.482	0.482	0.482
fEDUC3:Bachelors	0.795	0.795	0.795	0.794	0.793	0.793	0.792	0.792	0.793
fEDUC4:Graduate	1.028	1.026	1.028	1.032	1.034	1.034	1.035	1.035	1.035
AGE	0.054	0.054	0.054	0.054	0.055	0.055	0.055	0.055	0.055
I(AGEˆ2)	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001
whiteTRUE	-0.013	-0.014	-0.013	-0.012	-0.012	-0.012	-0.012	-0.012	-0.012
blackTRUE	-0.062	-0.062	-0.062	-0.061	-0.060	-0.060	-0.060	-0.060	-0.061
marriedTRUE	0.060	0.060	0.060	0.059	0.058	0.058	0.057	0.057	0.057
wasmarriedTRUE	0.005	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006

Table 4:

	poly 2	poly 3	poly 4	poly 5	poly 6	poly 7	poly 8	poly 9	poly 10
predictionsˆ 1	-10	19	308	-3,042	21,516	21,516	-112,096	-112,096	677,955
predictionsˆ 2	6	-28	-548	7,798	-70,468	-70,468	432,017	432,017	-2,935,848
predictionsˆ 3		13	427	-9,902	122,178	122,178	-907,748	-907,748	7,046,769
predictionsˆ 4			-123	6,231	-118,322	-118,322	1,111,053	1,111,053	-10,064,805
predictionsˆ 5				-1,555	60,703	60,703	-769,311	-769,311	8,370,128
predictionsˆ 6					-12,892	-12,892	248,361	248,361	-3,382,747
predictionsˆ 7									
predictionsˆ 8							-14,567	-14,567	398,727
predictionsˆ 9									
predictionsˆ 10									-43,135

Side Projects

- Program some or all of this assignment in C++ using Rcpp or in Julia (up to 4 points).
- Program some or all of assignment 2 or 3 in C++ using Rcpp or in Julia (up to 5 points, raising the reward since we had no takers!).
- If you scored lower than 70 on one of your first 3 assignments, write a 2-3 page report documenting was wrong with your code, using the solutions fix your code so that is works properly. (up to 3 points each).
- Write a guide on the class wiki about how to work with NA, NaN, Inf values in R. Provide examples and discussion (up to 1.5 points for up to 3 people).
- Write down 3 research ideas. For each explain the idea in 4-10 sentences. Write 1-4 sentences on why it is interesting, write 1-4 sentences on what data you would use, write 1-4 sentences on what methods it would use. These ideas can be very rough. If you would like feedback on one or all of them, let us know (up to 1 point each).
- Find a paper that applies a selection correction method to data and write a 1-3 page report on it (up to 2.5 points).
- Find 2-3 news articles where the author is not taking selection into account. Discuss how this may bias results (up to 2.5 points)
- Write a guide for using stargazer or texreg to make latex tables on the wiki. Include examples (1.5 points for up to 3 people for both stargazer and texreg).