UNIVERSITY OF NEVADA LAS VEGAS

LEE BUSINESS SCHOOL

ECO 772- ECONOMETRICS II

PROBLEM SET #04

STUDENT NAME: DANIEL EMAASIT

MAJOR: CIVIL ENGINEERING

INSTRUCTOR NAME: Prof. Ian McDonough

DATE SUBMITTED: 04/17/2015

SPRING 2015

Problem 17.6

Consider a family saving function for the population of all families in the United States: sav = β0 + β1inc + hhsize + β3educ + β4age + u,

where hhsize is household size, educ is years of education of the household head, and age is age of the household head. Assume that E(u|inc,hhsize,educ,age) = 0.

Part (i)

Suppose that the sample includes only families whose head is over 25 years old. If we use OLS on such a sample, do we get unbiased estimators of the βj? Explain.

The estimates from OLS will be unbiased because the sample chosen is based on an exogenous regressor. The population regression function for the family saving function is the same as the regression function in the sample with heads older than 25 years.

Part (ii)

Now, suppose our sample includes only married couples without children. Can we estimate all of the parameters in the saving equation? Which ones can we estimate?

No, we cannot estimate all the parameters. For example, since there is no variation in hhsize in the subpopulation, we would not be able to estimate β2. However we can estimate the intercept in the subpopulation becomes β0 + 2β2.

Assuming there is variation in inc, educ, and age among married people without children, we can still estimate β1, β3, and β4.

Part (iii)

Suppose we exclude from our sample families that save more than $25,000 per year. Does OLS produce consistent estimators of the βj?

OLS does not produce consistent estimators. This is because we would be selecting the sample on the basis of the dependent variable. A much better option would to use a truncated regression model.

Problem 17.7

Suppose you are hired by a university to study the factors that determine whether students admitted to the university actually come to the university. You are given a large random sample of students who were admitted the previous year. You have information on whether each student chose to attend, high school performance, family income, financial aid offered, race, and geographic variables. Someone says to you, "Any analysis of that data will lead to biased results because it is not a random sample of all college applicants, but only those who apply to this university." What do you think of this criticism?

Since the research is on this specific univeristy, I think it is appropriate to specify a model for this kind of data that consist of only applicants for this university. So I do not think that there is a sample selection problem.

However, if the pool of applicants changes in the near future, then there is a sample selection problem because the current students that apply may be different from students that may apply in the future.

Problem C17.7

Use the MROZ.RAW data for this exercise.

Part (i)

Using the 428 women who were in the workforce, estimate the return to education by OLS including exper, exper2, nwifeinc, age, kidslt6, and kidsge6 as explanatory variables. Report your estimate on educ and its standard error.

```
# Read the data into R
library(haven)

## Warning: package 'haven' was built under R version 3.1.3

w_data <- read_dta("MROZ.dta")

# Estimate the return to education by OLS
fit_lm <- lm(lwage ~ educ + exper + expersq + nwifeinc + age + kidslt6 +
kidsge6, data = w_data)

# Tabulate the model results
library(stargazer)

stargazer(fit_lm, type = "text", title = "Return to Education for Working
Women")

##
## Return to Education for Working Women
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                lwage
## ------------------------------------------------
## educ                          0.100***
##                                (0.015)
##
## exper                         0.041***
##                                (0.013)
##
## expersq                        -0.001*
##                               (0.0004)
```

```
##
## nwifeinc                      0.006*
##                               (0.003)
##
## age                           -0.004
##                               (0.005)
##
## kidslt6                       -0.056
##                               (0.089)
##
## kidsge6                       -0.018
##                               (0.028)
##
## Constant                      -0.358
##                               (0.318)
##
## ------------------------------------------------
## Observations                    428
## R2                             0.164
## Adjusted R2                    0.150
## Residual Std. Error      0.667 (df = 420)
## F Statistic           11.779*** (df = 7; 420)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

The coefficient and standard error on educ are 0.1 and 0.015 respectively.

Part (ii)

Now, estimate the return to education by Heckit, where all exogenous variables show up in the second-stage regression. In other words, the regression is log(wage) on educ, exper, exper2, nwifeinc, age, kidslt6, kidsge6, and estimated lambda. Compare the estimated return to education and its standard error to that from part (i).

```
# Estimate a model by Heckit method
library(sampleSelection)

fit_hkt <- heckit(selection = inlf ~ educ + exper + expersq + nwifeinc + age
+ kidslt6 + kidsge6,
                  outcome = lwage ~ educ + exper + expersq + nwifeinc + age +
kidslt6 + kidsge6,
                  method = "2step", data = w_data)

# Display the results
summary(fit_hkt)
```

```
## --------------------------------------------
## Tobit 2 model (sample selection model)
## 2-step Heckman / heckit estimation
## 753 observations (325 censored and 428 observed)
## 19 free parameters (df = 735)
```

```
## Probit selection equation:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.270077   0.508593   0.531  0.59556
## educ         0.130905   0.025254   5.183 2.82e-07 ***
## exper        0.123348   0.018716   6.590 8.37e-11 ***
## expersq     -0.001887   0.000600  -3.145  0.00173 **
## nwifeinc    -0.012024   0.004840  -2.484  0.01320 *
## age         -0.052853   0.008477  -6.235 7.63e-10 ***
## kidslt6     -0.868328   0.118522  -7.326 6.24e-13 ***
## kidsge6      0.036005   0.043477   0.828  0.40786
## Outcome equation:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.5602853  0.4587672  -1.221 0.222370
## educ         0.1187172  0.0340507   3.486 0.000518 ***
## exper        0.0598358  0.0336730   1.777 0.075987 .
## expersq     -0.0010523  0.0006381  -1.649 0.099566 .
## nwifeinc     0.0038434  0.0044919   0.856 0.392492
## age         -0.0111580  0.0134792  -0.828 0.408053
## kidslt6     -0.1880451  0.2308275  -0.815 0.415533
## kidsge6     -0.0122255  0.0296063  -0.413 0.679775
## Multiple R-Squared:0.1649,   Adjusted R-Squared:0.1489
## Error terms:
##                Estimate Std. Error t value Pr(>|t|)
## invMillsRatio   0.2885     0.4636   0.622    0.534
## sigma           0.6896         NA      NA       NA
## rho             0.4183         NA      NA       NA
## --------------------------------------------
```

The estimated return to education is 11.9%, (Coefficient = 0.1187 and Standard Error = 0.034). The estimated return to education is larger than without the Heckit corrections, but the Heckit standard error is more than two times larger.

Part (iii)

Using only the 428 observations for working women, regress estimated lambda on educ, exper, exper2, nwifeinc, age, kidslt6, and kidsge6. How big is the R-squared? How does this help explain your findings from part (ii)? (Hint: Think multicollinearity.)

```
# Obtain the inverse mills ration as lambda
myProbit <- glm(inlf ~ educ + exper + expersq + nwifeinc + age + kidslt6 +
kidsge6,
                family = binomial(link = "probit" ), data = w_data)
w_data$IMR <- invMillsRatio(myProbit)$IMR1

# Regress estimated lambda
fit_lm2 <- lm(IMR ~ educ + exper + expersq + nwifeinc + age + kidslt6 +
kidsge6, data = w_data)

# Show the model results
stargazer(fit_lm2, type = "text", title = "Model Results")
```

```
##
## Model Results
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                  IMR
## ------------------------------------------------
## educ                            -0.077***
##                                  (0.002)
##
## exper                           -0.077***
##                                  (0.001)
##
## expersq                         0.001***
##                                 (0.00004)
##
## nwifeinc                        0.007***
##                                 (0.0003)
##
## age                             0.031***
##                                  (0.001)
##
## kidslt6                         0.536***
##                                  (0.007)
##
## kidsge6                         -0.024***
##                                  (0.003)
##
## Constant                        0.735***
##                                  (0.033)
##
## ------------------------------------------------
## Observations                      753
## R2                               0.968
## Adjusted R2                      0.967
## Residual Std. Error      0.091 (df = 745)
## F Statistic         3,169.943*** (df = 7; 745)
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

The model produces a high R squared value of 0.968. This shows that there is multicollinearity among the regressors in the second stage regression, which led to the large standard errors in part (ii).