AR4

YING LIU

Econ 613

Ex. 1

1.1-1.2

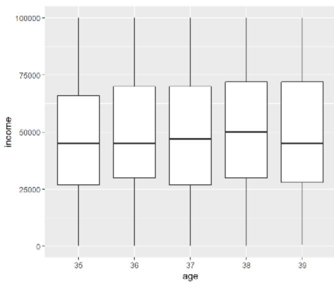
The new variables are created as follows.

For the education varible, I use YSCH-3113 to record edu variable, assume None is 0, GED(equivalent to high school level) takes 12y, associate degree takes 14y, BA takes 16y, MA takes 18y, phd takes 22y, Professional degree takes 22y. Then Recode the parents who are ungraded to 0.

	age	work_exp	edu
NA	38	0.0000000	NA
00	37	12.4230769	12
00	36	1.6923077	16
00	38	1.9230769	12
00	37	13.4615385	12
00	37	2.2500000	12
NA	36	2.3653846	0
00	38	4.1923077	16
00	37	3.2307692	18
00	35	5.0769231	18
00	37	11.9423077	16
00	38	14.9230769	12
00	35	0.0000000	12

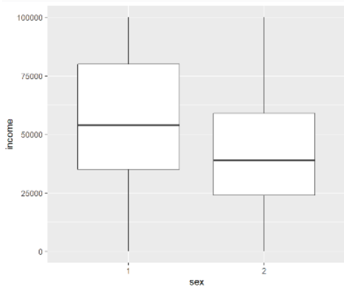
1.3

- **Boxplot**



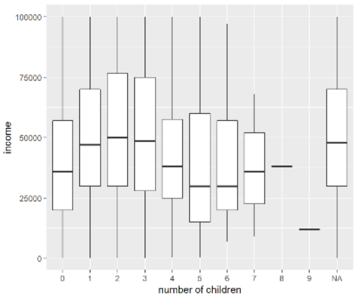
i)Boxplot of income by age groups shows that there is slightly positive relationship between age and income.

i)



ii)Boxplot of income by gender groups shows that male have greater wage than female.

ii)



iii)Boxplot of income grouped by number of children shows that families with fewer children(but not zero) have higher wage than others.

iii)

- **Table**

```
> table1

        35         36         37         38         39 
0.27777778 0.19444444 0.16666667 0.27777778 0.08333333 
```

i) Table of share of zero income grouped by age shows that younger people have a larger proportion in zero income group.

```
> table2

        1         2 
0.5833333 0.4166667 
```

ii) Table of share of zero income grouped by gender shows that male have higher proportion in zero income group.

```
> table3

           0          1          2          3
  0 0.00000000 0.13333333 0.00000000 0.06666667
  1 0.13333333 0.16666667 0.26666667 0.06666667
  2 0.10000000 0.00000000 0.00000000 0.03333333
  3 0.03333333 0.00000000 0.00000000 0.00000000
```

iii) Table of share of zero income grouped by number of children and martial status shows that people with fewer children and never married have a larger proportion in zero income group.

Ex. 2

2.1

i)

When using OLS model, the results are as follows:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2117.38    9513.01  -0.223    0.824    
age           365.63     255.72   1.430    0.153    
work_exp     1066.48      66.23  16.104   <2e-16 ***
edu          2310.87      85.92  26.894   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26030 on 5368 degrees of freedom
Multiple R-squared:  0.1655,	Adjusted R-squared:  0.1651 
F-statistic: 354.9 on 3 and 5368 DF,  p-value: < 2.2e-16
```

The model shows that for each one additional year of age, income will increase 365.63 but the correlation is insignificant; for one additional year of work experience, income will increase 1066.48; for one additional year of education, income will increase 2310.87.

ii) Selection problem: We wants to estimate the determinants of wage offers, but has access to wage observations for only those who work. Since people who work are selected non-randomly from the population, estimating the determinants of wages from the subpopulation who work may introduce bias.

2.2

Heckman model can deal with the selection problem: Heckman model suppose that we observe y only if the units of observation in that random sample make some decision. This allows us to characterize the sample selection bias that might emerge from attempting to estimate the regression with only the subsample for whom we observe y.

2.3

- **TWO STEP APPROACH**

The two-step approach first conducts a probit model regarding whether the individual is observed or not, in order to calculate the inverse mills ratio, or ‘nonselection hazard’. The second step is a standard linear model.

Step 1: Probit Model

```
Call:
glm(formula = observe_y ~ age + work_exp + edu + z, family = binomial(link = "probit"),
    data = NLSY97_full)

Deviance Residuals:
    Min       1Q    Median       3Q      Max
-3.3812   0.0963   0.1112   0.1264   0.2180

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.29010    1.61437   0.180    0.857
age          0.04884    0.04365   1.119    0.263
work_exp     0.01605    0.01255   1.280    0.201
edu          0.02173    0.01245   1.745    0.081 .
z            0.01225    0.05994   0.204    0.838
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Calculate the inverse mills ratio

```
> mills0 <- dnorm(probit_lp)/pnorm(probit_lp)
> summary(mills0)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.003149 0.013670 0.017584 0.018573 0.022081 0.056769
```

Step 2: Estimate via Linear Regression

```
Call:
lm(formula = YINC_1700_2019 ~ age + work_exp + edu + imr, data = NLSY97_full[observe_y,
    ])

Residuals:
   Min     1Q Median     3Q    Max
-78760 -19233  -3474  17913  85250

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -62370.2    27308.1  -2.284  0.02241 *
age           1464.3      532.2   2.751  0.00595 **
work_exp      1393.2      153.8   9.059  < 2e-16 ***
edu           2903.2      265.9  10.918  < 2e-16 ***
imr         501777.2   213185.2   2.354  0.01862 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26020 on 5367 degrees of freedom
Multiple R-squared:  0.1664,	Adjusted R-squared:  0.1658
F-statistic: 267.8 on 4 and 5367 DF,  p-value: < 2.2e-16
```

- **Maximum Likelihood**

```
select_ll <- function(par, X, Z, y, observe_y) {
    gamma        = par[1:5]
    lp_probit = Z %*% gamma

    beta    = par[6:9]
    lp_lm = X %*% beta
    pr=dnorm(lp_lm)
    pr[pr>0.999999] = 0.9999
    pr[pr<0.000001] = 0.0001
    sigma = par[10]
    rho      = par[11]
    rho = min(rho, 0.999999)
    rho = max(rho, -0.999999)
    pb = 1-pnorm(lp_probit[!observe_y])
    pb[pb <0.000001] = 0.000001
    ll = sum(log(pb)) +
       - log(sigma) +
       sum(dnorm(y, mean = lp_lm, sd = sigma, log = TRUE)) +
       sum(pnorm((lp_probit[observe_y] + rho/sigma * (y-lp_lm)) / sqrt(1-rho^2),
                      log.p = TRUE))

    -ll
}
```

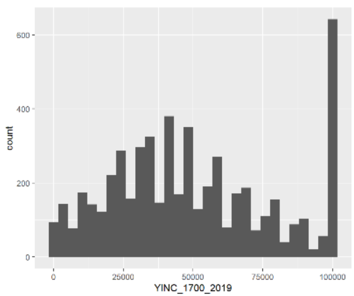
```
> fit_unbounded$par
 (Intercept)           age      work_exp           edu             z   (Intercept)           age
3.540156e+00  1.210650e+02  2.868821e+01  4.736516e+01  1.679832e-02 -5.803228e+04  1.849355e+03
    work_exp           edu
1.051907e+03  2.391092e+03  2.613859e+04 -3.959630e+01
```

- **Comparison**

The results show that all the coefficient increase when we use Heckman selection model and the coefficient of age turns to significance. Because people who are older and have less work experience and education background are less likely to obtain a job, the income in sample we observe are biased. When we use Heckman selection model, such problem has been solved and age\work experience\education are actually more important than the OLS model suggest.

Ex. 3

3.1



The income variable is upper censored at 100000.

3.2 I use the Tobit model to solve the censored problem. When data is censored such that while we observe the value, it is not the true value, which would extend beyond the range of the observed data. This is very commonly seen in cases where the dependent variable has been given some arbitrary cutoff at the lower or upper end of the range, often resulting in floor or ceiling effects respectively. The conceptual idea is that we are interested in modeling the underlying latent variable that would not have such restriction if it was actually observed.

3.3-3.4

```
tobit_ll <- function(par, X, y, ul = -Inf, ll = Inf) {

  # this function only takes a lower OR upper limit

  # parameters
  sigma = exp(par[length(par)])
  beta    = par[-length(par)]

  # create indicator depending on chosen limit(here we need upper limit 100000)
  if (!is.infinite(ll)) {
    limit = ll
    indicator = y > ll
  } else {
    limit = ul
    indicator = y < ul
  }

  # linear predictor
  beta = as.matrix(beta)
  lp = X %*% beta
  part1 = sum(indicator * log((1/sigma)*dnorm((y-lp)/sigma)))
  part2 = sum((1-indicator) * log(pnorm((lp-limit)/sigma)))
  pr= pnorm((lp-limit)/sigma)
  pr[pr>0.999999] = 0.999999
  pr[pr<0.000001] = 0.000001
  # log likelihood
  ll = part1 + part2
  -ll
}
```

```
> fit_tobit$par
(Intercept)         age    work_exp         edu   log_sigma
 -3898.0898    402.5147   1071.9333   2316.8192    208.2471
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2117.38    9513.01  -0.223    0.824
age           365.63     255.72   1.430    0.153
work_exp     1066.48      66.23  16.104   <2e-16 ***
edu          2310.87      85.92  26.894   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26030 on 5368 degrees of freedom
Multiple R-squared:  0.1655,    Adjusted R-squared:  0.1651 
F-statistic: 354.9 on 3 and 5368 DF,  p-value: < 2.2e-16
```

- Interpretation: The tobit model shows that when the income would not have such upper restriction, then all the coefficient of age\work experience\education on income will increase.

Ex. 4

4.1

Ability bias: People with traits the labor market values (intelligence, work ethic, conformity, etc.) tend to get more education. Since employers have some ability to detect these valued traits, people with more education would have earned above-average incomes even if their education were only average.

Punchline: Standard estimates overstate the effect of education on worker productivity and income.

4.2-4.3

- **Within Estimator**

```
Call:
lm(formula = income_dif ~ edu_dif + martial_dif + workexp_dif, 
    data = Q4_demeaned)

Residuals:
    Min      1Q  Median      3Q     Max 
-141025   -9256    -625    7881  275910 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2.281e-11  7.022e+01    0.00        1    
edu_dif     1.505e+03  2.378e+01   63.30   <2e-16 ***
martial_dif 1.621e+04  2.210e+02   73.35   <2e-16 ***
workexp_dif 2.912e+03  2.626e+01  110.86   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20110 on 82004 degrees of freedom
Multiple R-squared:  0.309,     Adjusted R-squared:  0.3089 
F-statistic: 1.222e+04 on 3 and 82004 DF,  p-value: < 2.2e-16
```

- **Between Estimator**

```
> summary(between_model)

Call:
lm(formula = income ~ edu + martial + work_exp, data = ave)

Residuals:
   Min     1Q Median     3Q    Max 
-53193  -9314  -2416   5844 288229 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1774.16     501.08   3.541 0.000401 ***
edu          1202.36      43.99  27.330  < 2e-16 ***
martial      9030.94     569.53  15.857  < 2e-16 ***
work_exp     2173.43      75.80  28.672  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15600 on 8596 degrees of freedom
Multiple R-squared:  0.2257,	Adjusted R-squared:  0.2255 
F-statistic: 835.4 on 3 and 8596 DF,  p-value: < 2.2e-16
```

- **Difference(any) Estimator**

```
> summary(fd_model)

Call:
lm(formula = income_fd ~ edu_fd + martial_fd + workexp_fd, data = FD)

Residuals:
    Min      1Q  Median      3Q     Max 
-211035   -5889   -2172    4258  321617 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4087.71      67.99   60.12  < 2e-16 ***
edu_fd         68.77      22.40    3.07  0.00214 ** 
martial_fd   2359.35     225.02   10.48  < 2e-16 ***
workexp_fd    955.65      29.63   32.25  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17100 on 73404 degrees of freedom
  (因为不存在，8600个观察量被删除了)
Multiple R-squared:  0.01578,	Adjusted R-squared:  0.01574 
F-statistic: 392.3 on 3 and 73404 DF,  p-value: < 2.2e-16
```

- **Interpretation:**

Within estimator: within estimator removes the unobserved differences between groups. This is because they are time invariant

Between estimator: The between estimator is in general biased in the same way as pooled OLS. To see this, write out the general panel data model

$$y_{it}=\beta' x_{it}+\gamma' z_i+\eta_i+u_{it} (i=1,...,N;t=1,...,T)$$

where the x variables are time-varying, the z variables are time invariant and η_i is the time-invariant individual effect. The between model is the cross-sectional equation

$$\bar{y}_i=\beta'\bar{x}_i+\gamma' z_i+\eta_i+\bar{u}_i$$

where
$\bar{y}_i=\frac{1}{T}\sum_{t=1}^{T} y_{it}, \bar{x}_i=\frac{1}{T}\sum_{t=1}^{T} x_{it}, \bar{u}_i=\frac{1}{T}\sum_{t=1}^{T} u_{it}$.

After averaging and deriving the between estimator, the individual effect η_i does not drop out of the equation. For all intents and purposes, the between estimator is useful in considering the random effects model rather than an estimator in its own right.

First difference estimator: The First-Difference (FD) estimator is obtained by running a pooled OLS from delta yit on deltaxit. The FD estimator wipes out time invariant omitted variables ci using the repeated observations over time.