Homework 4: Censoring and Panel Data

Xin Lin

```
# Load packages
library(tidyverse)
library(readxl)
library(magrittr)
library(janitor)
library(plyr)
library(formattable)
library(knitr)
library(kableExtra)
library(censReg)
library(panelr)
library(plm)

# abadon scientific notation in R
options(scipen = 999)

# set the starting number used to generate random sample
set.seed(999)

# import dataset
dat <- read_csv("data/dat_A4.csv")
dat_panel <- read_csv("data/dat_A4_panel.csv")
```

Exercise 1 Preparing the Data

1. Create new variables

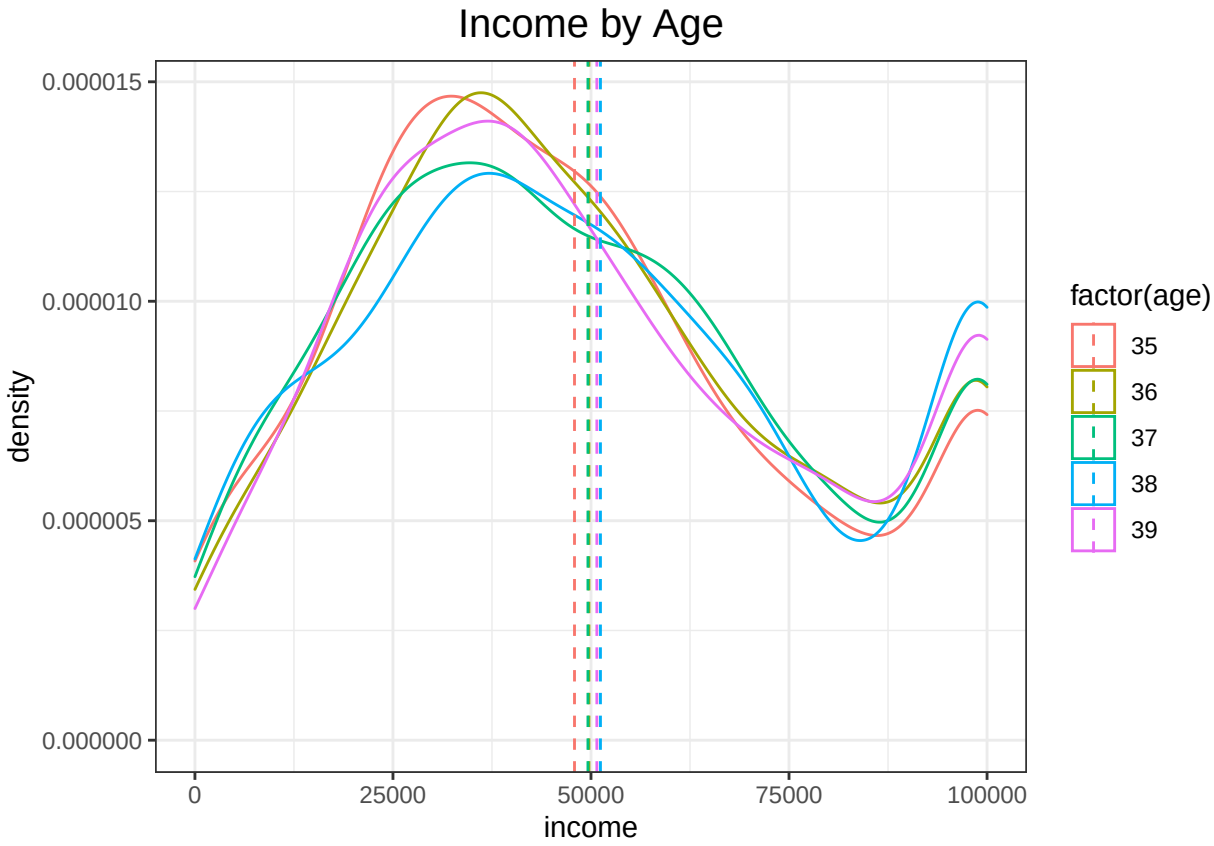
```
# create "age"
# age = 2019 - birth year
dat <- dat %>%
  mutate(age = 2019 - KEY_BDATE_Y_1997)

# create "work_exp"
# work_exp = sum of #weeks working at each job / 52
dat_work <- dat %>%
  select(contains("CV_WKSWK_JOB")) %>%
  mutate(work_exp = round(rowSums(., na.rm = TRUE)/52, digits = 2))
dat$work_exp <- dat_work$work_exp

# create edu-related variables
# the following only need to deal with the "ungraded" and NA: recode as 0
```

```r
# create "bio_father_edu"
dat <- dat %>%
  mutate(bio_father_edu = ifelse(CV_HGC_BIO_DAD_1997 == 95, 0, CV_HGC_BIO_DAD_1997))
# create "bio_mother_edu"
dat <- dat %>%
  mutate(bio_mother_edu = ifelse(CV_HGC_BIO_MOM_1997 == 95, 0, CV_HGC_BIO_MOM_1997))
# create "res_father_edu"
dat <- dat %>%
  mutate(res_father_edu = ifelse(CV_HGC_RES_DAD_1997 == 95, 0, CV_HGC_RES_DAD_1997))
# create "res_mother_edu"
dat <- dat %>%
  mutate(res_mother_edu = ifelse(CV_HGC_RES_MOM_1997 == 95, 0, CV_HGC_RES_MOM_1997))
# recode "YSCH.3113_2019" into numeric variable
dat <- dat %>%
  mutate(edu = ifelse(YSCH.3113_2019==1, 0,
                      ifelse(YSCH.3113_2019%in%c(2,3), 12,
                             ifelse(YSCH.3113_2019==4, 14,
                                    ifelse(YSCH.3113_2019==5, 16,
                                           ifelse(YSCH.3113_2019==6, 18,
                                                  ifelse(YSCH.3113_2019%in%c(7,8), 21, NA)))))))
```

2. Data visualizations

1) Plot the income data by age groups

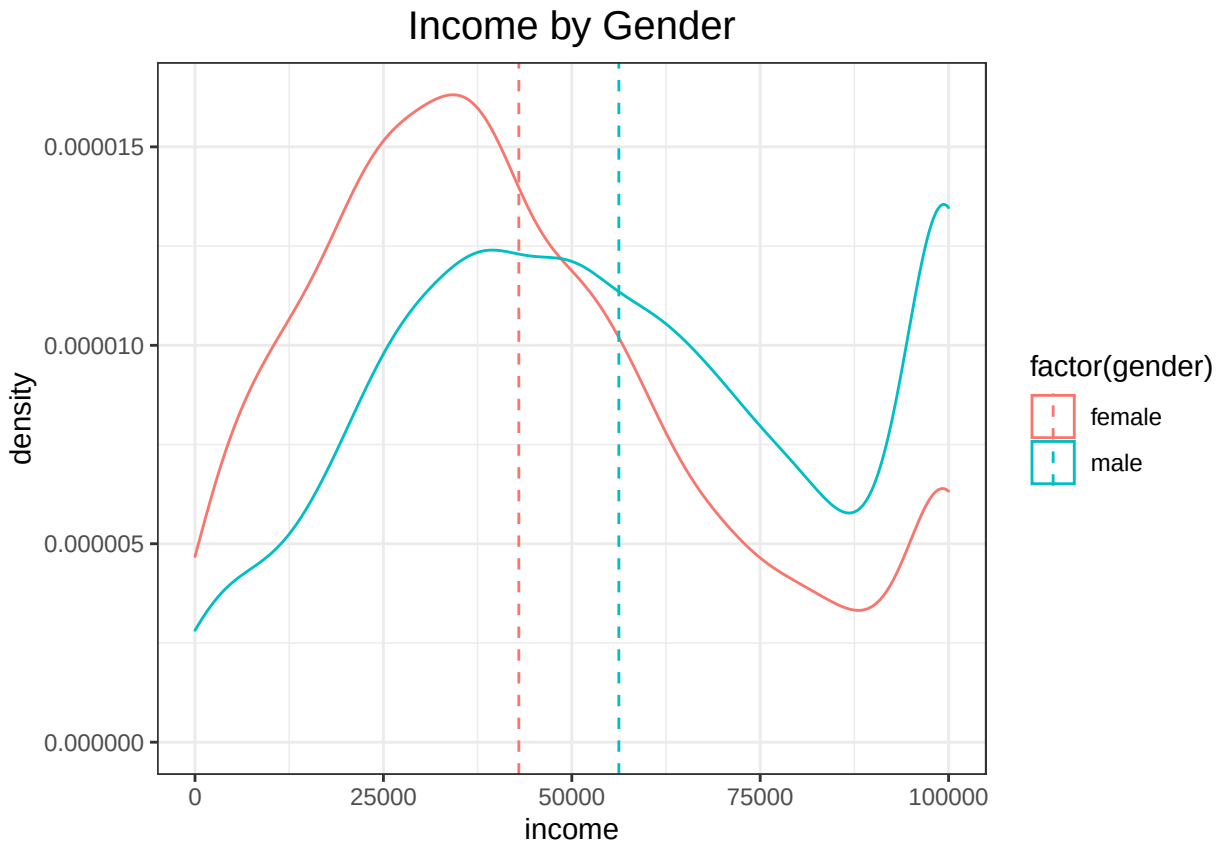
```r
# make the dat a data frame
dat <- dat %>%
  as.data.frame() %>%
  dplyr::rename(income = YINC_1700_2019)
# compute mean income by gener
mu1 <- ddply(dat, "age", summarise, grp.mean=mean(income, na.rm = T))
# plot
ggplot(dat, aes(income, color = factor(age))) +
  geom_density() +
  geom_vline(data=mu1, aes(xintercept=grp.mean, color=factor(age)), linetype="dashed") +
  xlab("income") +
  ylab("density") +
  labs(title = "Income by Age") +
  theme_bw() +
  theme(plot.title = element_text(face = "plain", size = 15, hjust = 0.5, color = "black"))
```



Interpretations: From the plot, we can observe that the distribution for each age group is almost the same, which means that the correlation between age and incoem may be insignificant. However, if looking at the mean income for each age group, we can observe that 38-year-old and 39-year-old participants have the highest income and 35-year-old participants have the least income, so there may be a positive correlation between age and income.

2) Plot the income data by gender groups

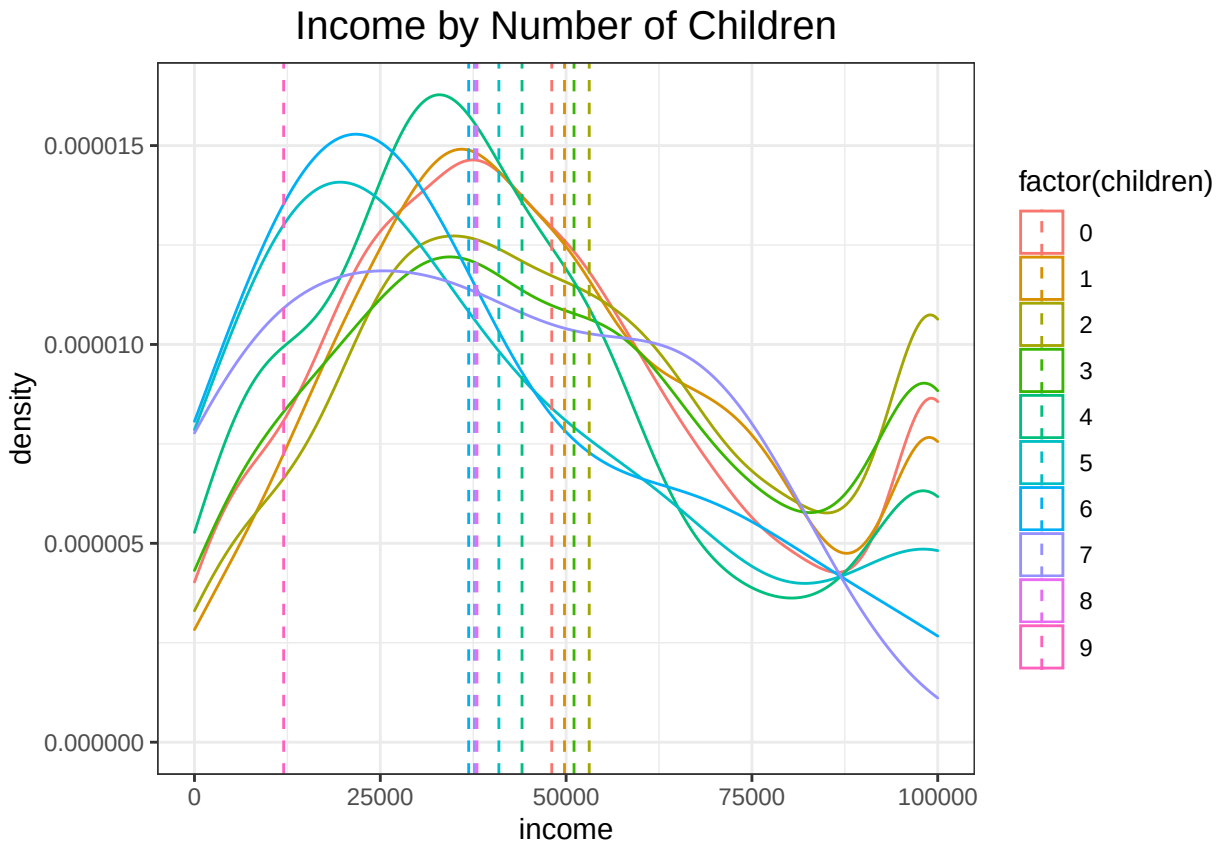
```
# create text gender variable
dat <- dat %>%
  mutate(gender = ifelse(KEY_SEX_1997 == 1, "male", "female"))
# compute mean income by gener
mu2 <- ddply(dat, "gender", summarise, grp.mean=mean(income, na.rm = T))
# plot
ggplot(data=dat, aes(income, color = factor(gender))) +
  geom_density() +
  geom_vline(data=mu2, aes(xintercept=grp.mean, color=gender), linetype="dashed") +
  xlab("income") +
  ylab("density") +
  labs(title = "Income by Gender") +
  theme_bw() +
  theme(plot.title = element_text(face = "plain", size = 15, hjust = 0.5, color = "black"))
```



Interpretations: From the plot, we can observe from both the distribution and mean that male participants earn significantly higher income than females do.

3) Plot the income data by number of children

```
# treate NA in CV_BIO_CHILD_HH_U18_2019 as 0
dat <- dat %>%
  mutate(children = ifelse(is.na(CV_BIO_CHILD_HH_U18_2019), 0, CV_BIO_CHILD_HH_U18_2019))
# compute mean income by gener
mu3 <- ddply(dat, "children", summarise, grp.mean=mean(income, na.rm = T))
# plot
ggplot(data=dat, aes(income, color = factor(children))) +
  geom_density() +
  geom_vline(data=mu3, aes(xintercept=grp.mean, color=factor(children)), linetype="dashed") +
  xlab("income") +
  ylab("density") +
  labs(title = "Income by Number of Children") +
  theme_bw() +
  theme(plot.title = element_text(face = "plain", size = 15, hjust = 0.5, color = "black"))
```



Interpretations: From the plot, we can observe from both the distribution and mean that participants with 4 or more children earns significantly lower than the other participants do. Participants with 1, 2, or 3 children earns the highest income among all participants. This suggests that the number of children might be negatively correlated with income.

4) Table the share of "0" in the income data by age groups

```
# find the share of "0" in income by age
share1 <- dat %>%
  tabyl(age, income) %>%
  adorn_totals(where = c("row", "col")) %>%
  adorn_percentages() %>%
  select(age, "0") %>%
  dplyr::rename("share of zero income" = "0")
# make the table
ans <- as.data.frame(share1)
kable(ans) %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"), full_width=FALSE)
```

age	share of zero income
35	0.0056465
36	0.0038738
37	0.0032591
38	0.0053362
39	0.0017741
Total	0.0040071

Interpretations: The share of zero-income is very low accross all age groups, but 35-year-old and 38-year-old participants have the highest zero-income share.

5) Table the share of "0" in the income data by gender groups

```
# find the share of "0" in income by gender
share2 <- dat %>%
  tabyl(gender, income) %>%
  adorn_totals(where = c("row", "col")) %>%
  adorn_percentages() %>%
  select(gender, "0") %>%
  dplyr::rename("share of zero income" = "0")
# make the table
ans <- as.data.frame(share2)
kable(ans) %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"), full_width=FALSE)
```

gender	share of zero income
female	0.0034208
male	0.0045662
Total	0.0040071

Interpretations: The share of zero-income is very low accross gender groups, but male participants have higher zero-income share than females do.

6) Table the share of "0" in the income data by number of children

```
# find the share of "0" in income by gender
share3 <- dat %>%
  tabyl(children, income) %>%
  adorn_totals(where = c("row", "col")) %>%
  adorn_percentages() %>%
  select(children, "0") %>%
  dplyr::rename("share of zero income" = "0")
# make the table
ans <- as.data.frame(share3)
kable(ans) %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"), full_width=FALSE)
```

children	share of zero income
0	0.0029889
1	0.0064103
2	0.0046296
3	0.0062422
4	0.0000000
5	0.0000000
6	0.0000000
7	0.0000000
8	0.0000000
9	0.0000000
Total	0.0040071

Interpretations: There are no observations with zero income if they have four or more children, and this might be caused by that there are fewer observations with four or more children in our dataset. From the table, we can see that participants with only one child have the highest zero-income share.

7) Table the share of "0" in the income data by marital status

```
# recode the marital status variable
dat <- dat %>%
  mutate(marital_status = ifelse(CV_MARSTAT_COLLAPSED_2019 == 1, "married",
                                 ifelse(CV_MARSTAT_COLLAPSED_2019 == 2, "separated",
                                        ifelse(CV_MARSTAT_COLLAPSED_2019 == 3, "divorced",
                                               ifelse(CV_MARSTAT_COLLAPSED_2019 == 4, "widowed", "single
  mutate(marital_status = ifelse(is.na(marital_status), "unanswered", marital_status))
# find the share of "0" in income by gender
share4 <- dat %>%
  tabyl(marital_status, income) %>%
  adorn_totals(where = c("row", "col")) %>%
  adorn_percentages() %>%
  select(marital_status, "0") %>%
  dplyr::rename("share of zero income" = "0")
# make the table
ans <- as.data.frame(share4)
kable(ans) %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"), full_width=FALSE)
```

marital_status	share of zero income
divorced	0.0012048
married	0.0062247
separated	0.0298507
single	0.0040726
unanswered	0.0000000
widowed	0.0000000
Total	0.0040071

Interpretations: From the table, we can observe that separated participants have significantly higher zero-income share than the others.

Exercise 2 Heckman Selection Model

1. Specify and estimate an OLS model to explain the income variable

Answer: The OLS model I am going to estimate is as follows:

$$ln(income)_i = \beta_0 + \beta_1 age_i + \beta_2 male_i + \beta_3 ubran_i + \beta_4 minority_i + \beta_5 work\_exp_i + \beta_6 edu_i + \epsilon_i$$

```
# recode the gender variable: 1 if male and 0 if female
# recode the urban variable: 1 if urban and 0 if rural
# recode the marital_status variable: 1 if married and 0 if single (never married, separated, widowed,
# recode the race_ethnicity variable: 1 if minority (black or hispanic) and 0 if not minority
dat <- dat %>%
  mutate(male = ifelse(gender == "male", 1, 0),
         urban = ifelse(CV_URBAN.RURAL_2019 == 1, 1, 0),
         married = ifelse(marital_status == "married", 1, 0),
         minority = ifelse(KEY_RACE_ETHNICITY_1997 %in% c(1,2), 1, 0)) %>%
  dplyr::rename("id" = PUBID_1997)
```

```
# keep only useful variables and keep observations whose incoem > 0
dat_ols <- dat %>%
  select(income, age, male, urban, married, minority, work_exp, children, edu, id) %>%
  filter(income>0) %>%
  na.omit() %>%
  mutate(ln_income = log(income))

# use lm() to estimate the model specified above
reg_ols <- lm(ln_income ~ age + male + urban + minority + work_exp + edu, data = dat_ols)
summary(reg_ols)
```

```
##
## Call:
## lm(formula = ln_income ~ age + male + urban + minority + work_exp +
##     edu, data = dat_ols)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7232 -0.2809  0.1318  0.4587  2.1715
##
## Coefficients:
##              Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  8.936452   0.282145  31.673 < 0.0000000000000002 ***
## age          0.006904   0.007542   0.915                 0.36
## male         0.370923   0.021131  17.554 < 0.0000000000000002 ***
## urban        0.141752   0.027119   5.227          0.000000179 ***
## minority    -0.098499   0.021853  -4.507          0.000006705 ***
## work_exp     0.035436   0.001960  18.082 < 0.0000000000000002 ***
## edu          0.064942   0.002637  24.625 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7636 on 5305 degrees of freedom
## Multiple R-squared:  0.2086, Adjusted R-squared:  0.2077
## F-statistic: 233.1 on 6 and 5305 DF,  p-value: < 0.00000000000000022
```

Interpretations: From the above results of OLS, we can observe that the coefficient on age is not significant since their P-value are less than 0.05. The coefficients on male, urban, minority, work_exp, and edu are significant. I will interpret a coefficent on a dummy variable and a coefficient on a continuous/discrete variable here: The coefficient on male β_2 means that the average income is 37.09% higher for males compared to females', holding other variables constant. The coefficient on edu β_6 means that with an additional year of education, average income increases by 6.49%, holding other variables constant.

Selection Problem: In the OLS above, the observations whose income are less than or equal to 0 or missing (not reported) are excluded from regression. Also, the missingness in income data may be non-random, and it is conditional on some exogenous factors, for example, housewives / husands who are unemployed don't have any occupational income, so they may choose not to report their income. Therefore, since their income are excluded / not observable, there might be some bias when estimating the model using the (sub)sample.

2. Explain why the Heckman model can deal with the selection problem

Answer: Heckman model corrects for selection bias using a two-stage estimator. In the first step, run a probit regression to predict the probability of earning an (occupational) income and then compute IMR=pdf(income>0)/cdf(income>0). In the second step, run OLS aagin, including IMR as a new explanatory variable, to rule out the effect of selection.

3. Estimate a Heckman selection model

1) In the first stage of predicting the probability of earning some income, whether earning some income might be affected by the number of children in family and marital status since housewives/husbands are more likely to be not employed and not earn any income; therefore, I predict the probability of earning using the probit specified as

$$Pr(income\_reported = 1) = \Phi(\beta_0 + \beta_1 age_i + \beta_2 male_i + \beta_3 ubran_i + \beta_4 minority_i$$
$$+\beta_5 work\_exp_i + \beta_6 edu_i + \beta_7 married_i + \beta_8 children_i + \epsilon_i)$$

Then, I compute $IMR = \frac{\phi(X\hat{\beta})}{\Phi(X\hat{\beta})}$, where X are variables specified above.

2) In the second stage, I will estimate the following model:

$$ln(income)_i = \beta_0 + \beta_1 age_i + \beta_2 male_i + \beta_3 ubran_i + \beta_4 minority_i + \beta_5 work\_exp_i + \beta_6 edu_i + \beta_7 imr_i + \epsilon_i$$

```
# keep only useful variables
# create a dummy to indicate whether an observation earns an income
# 1 if income > 0; 0 otherwise
dat_heckman <- dat %>%
  mutate(income_non_missing = ifelse(income==0 | is.na(income), 0, 1)) %>%
  select(income, age, male, urban, married, minority, work_exp, children, edu, id) %>%
  na.omit() %>%
  mutate(income_reported = ifelse(is.na(income) | income == 0, 0, 1))

# first stage: probit model for selection mechanism
reg_probit <- glm(income_reported ~ age + male + urban + minority + work_exp + edu
                  + married + children, data = dat_heckman)

# compute IMR
pred <- reg_probit$linear.predictors
imr <- dnorm(pred) / pnorm(pred)
dat_heckman$imr <- imr
```

```r
# keep observations whose income is greater than 0
dat_heckman <- dat_heckman %>%
  filter(income>0) %>%
  mutate(ln_income = log(income))

# second stage: regression for selected sample
reg_heckman <- lm(ln_income ~ age + male + urban + minority + work_exp + edu + imr,
                  data = dat_heckman)
summary(reg_heckman)
```

```
##
## Call:
## lm(formula = ln_income ~ age + male + urban + minority + work_exp +
##     edu + imr, data = dat_heckman)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6945 -0.2707  0.1339  0.4588  2.1682
##
## Coefficients:
##               Estimate Std. Error t value             Pr(>|t|)
## (Intercept) -15.653069   4.736905  -3.304             0.000958 ***
## age           0.032743   0.009017   3.631             0.000284 ***
## male          0.341157   0.021842  15.619 < 0.0000000000000002 ***
## urban         0.097632   0.028352   3.444             0.000579 ***
## minority     -0.019896   0.026527  -0.750             0.453268
## work_exp      0.044291   0.002593  17.084 < 0.0000000000000002 ***
## edu           0.080053   0.003920  20.423 < 0.0000000000000002 ***
## imr          80.622188  15.503550   5.200          0.000000207 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7617 on 5304 degrees of freedom
## Multiple R-squared:  0.2127, Adjusted R-squared:  0.2116
## F-statistic: 204.6 on 7 and 5304 DF,  p-value: < 0.00000000000000022
```

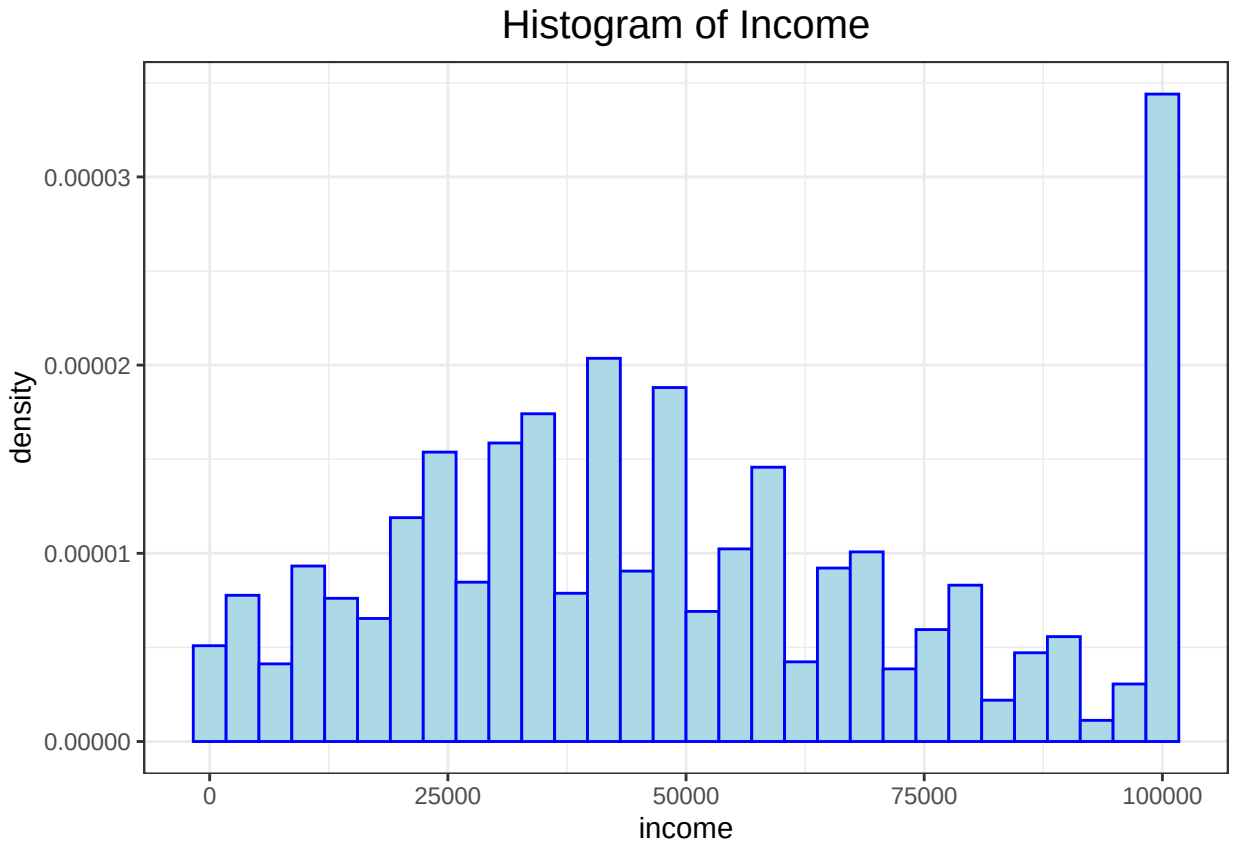
Interpretations: From the table below, we can observe that both the signs and values of the coefficients remain similar to OLS in the Heckam model. In OLS, the age coefficient is not significant; however, in the Heckman selection model, the age coefficient is significant, but the minority coefficient is not significant. Also, from the Heckman model where the selection bias is ruled out, we can see that the coefficient on imr is positive. It suggests that the observed income are higher on average.

	OLS coefficient	OLS t-value	Heckman Model coefficient	Heckman Model t-value
age	0.0069042	0.9153998	0.0327434	3.6314810
male	0.3709230	17.5537798	0.3411569	15.6190654
urban	0.1417524	5.2270349	0.0976322	3.4435781
minority	-0.0984995	-4.5073530	-0.0198963	-0.7500325
work_exp	0.0354364	18.0818386	0.0442913	17.0838923
edu	0.0649420	24.6252373	0.0800527	20.4229475
imr	0.0000000	0.0000000	80.6221878	5.2002405

Exercise 3 Censoring

1. Plot a a histogram to check the distribution of the income variable. What might be the censored value here?

```
ggplot(data=dat, aes(income)) +
  geom_histogram(aes(y=..density..), color="blue", fill="lightblue") +
  xlab("income") +
  ylab("density") +
  labs(title = "Histogram of Income") +
  theme_bw() +
  theme(plot.title = element_text(face = "plain",size = 15, hjust = 0.5, color = "black"))
```



Answer: From the histogram above, we can observe that the data is top coded at $100,000.

2. Propose a model to deal with the censoring problem

Answer: Tobit model.

3. Estimate the model above

```
# use the same dataset as in the ols
dat_tobit <- dat_ols
```

```r
# estimate the tobit model by optimizing the likelihood function
# y is censored for values above 100,000
tobit_like <- function(par, x1, x2, x3, x4, x5, x6, y){
  d <- log(100000)  # censored value
  y_hat <- 1*par[1] + x1*par[2] + x2*par[3] + x3*par[4] + x4*par[5] + x5*par[6] + x6*par[7]
  sigma <- par[8]
  cdf <- pnorm((y_hat-d)/sigma)
  pdf <- dnorm((y-y_hat)/sigma)
  cdf[cdf>0.999999] = 0.999999
  cdf[cdf<0.000001] = 0.000001
  log_like <- sum(ifelse(y == d, log(pdf), log(cdf))-log(sigma))
  return(-log_like)
}

# specify X and y
x1 <- as.matrix(dat_tobit$age)
x2 <- as.matrix(dat_tobit$male)
x3 <- as.matrix(dat_tobit$urban)
x4 <- as.matrix(dat_tobit$minority)
x5 <- as.matrix(dat_tobit$work_exp)
x6 <- as.matrix(dat_tobit$edu)
y <- as.matrix(dat_tobit$ln_income)

# optimize the likelihood function
tobit_res <- optim(runif(7,-20,20), fn = tobit_like, method="BFGS",
                   control=list(trace=6,REPORT=1,maxit=1000),
                   x1=x1, x2=x2, x3=x3, x4=x4, x5=x5, x6=x6, y = y, hessian=TRUE)
tobit_res$par
```

```r
# estimate the tobit model using the package
reg_tobit <- censReg(ln_income ~ age + male + urban + minority + work_exp + edu,
                     right = log(100000), data = dat_tobit)
summary(reg_tobit)
```

```
##
## Call:
## censReg(formula = ln_income ~ age + male + urban + minority +
##     work_exp + edu, right = log(100000), data = dat_tobit)
##
## Observations:
##          Total  Left-censored     Uncensored Right-censored
##           5312              0           4681            631
##
## Coefficients:
##              Estimate Std. error t value             Pr(> t)
## (Intercept)  8.615897   0.314580  27.389 < 0.0000000000000002 ***
## age          0.012271   0.008405   1.460                0.144
## male         0.433350   0.023560  18.394 < 0.0000000000000002 ***
## urban        0.165208   0.030151   5.479          0.0000000427 ***
## minority    -0.131099   0.024307  -5.393          0.0000000691 ***
## work_exp     0.038331   0.002185  17.544 < 0.0000000000000002 ***
## edu          0.074391   0.002941  25.293 < 0.0000000000000002 ***
## logSigma    -0.177082   0.010545 -16.792 < 0.0000000000000002 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Newton-Raphson maximisation, 5 iterations
## Return code 1: gradient close to zero
## Log-likelihood: -6404.656 on 8 Df
```

Interpretations: From the table below, we can observe that both the signs and values of the coefficients remain similar to OLS in the Tobit model. In both OLS and Tobit model, the age coefficient is not significant. Also, from the Tobit model, we can see that the coefficient on logSigma is the estimated standrd error of the regression. The value is comparable to the root mean squared error that would be obtained in OLS regression.

	OLS coefficient	OLS t-value	Tobit Model coefficient	Tobit Model coefficient
age	0.0069042	0.9153998	0.0122708	1.460010
male	0.3709230	17.5537798	0.4333503	18.393774
urban	0.1417524	5.2270349	0.1652079	5.479331
minority	-0.0984995	-4.5073530	-0.1310991	-5.393455
work_exp	0.0354364	18.0818386	0.0383310	17.543696
edu	0.0649420	24.6252373	0.0743908	25.293102
logSigma	0.0000000	0.0000000	-0.1770818	-16.792226

Exercise 4 Panel Data

1. Explain the potential ability bias when trying to explain the determinants of wages

Answer: There is a causal relationship between education and income. Education is a source of an accumulation of competences, and most productive individuals have an interest in studying for the longest period, entailing the probability of ability bias.

2. Exploit the panel dimension of the data to propose a model to correct for the ability bias

```
# clean the panal data set
# rename inconsistent variables
dat_panel <- dat_panel %>%
  dplyr::rename(CV_HIGHEST_DEGREE_EVER_EDT_1998 = CV_HIGHEST_DEGREE_9899_1998,
                CV_HIGHEST_DEGREE_EVER_EDT_1999 = CV_HIGHEST_DEGREE_9900_1999,
                CV_HIGHEST_DEGREE_EVER_EDT_2000 = CV_HIGHEST_DEGREE_0001_2000,
                CV_HIGHEST_DEGREE_EVER_EDT_2001 = CV_HIGHEST_DEGREE_0102_2001,
                CV_HIGHEST_DEGREE_EVER_EDT_2002 = CV_HIGHEST_DEGREE_0203_2002,
                CV_HIGHEST_DEGREE_EVER_EDT_2003 = CV_HIGHEST_DEGREE_0304_2003,
                CV_HIGHEST_DEGREE_EVER_EDT_2004 = CV_HIGHEST_DEGREE_0405_2004,
                CV_HIGHEST_DEGREE_EVER_EDT_2005 = CV_HIGHEST_DEGREE_0506_2005,
                CV_HIGHEST_DEGREE_EVER_EDT_2006 = CV_HIGHEST_DEGREE_0607_2006,
                CV_HIGHEST_DEGREE_EVER_EDT_2007 = CV_HIGHEST_DEGREE_0708_2007,
                CV_HIGHEST_DEGREE_EVER_EDT_2008 = CV_HIGHEST_DEGREE_0809_2008,
                CV_HIGHEST_DEGREE_EVER_EDT_2009 = CV_HIGHEST_DEGREE_0910_2009)

# convert wide to long
dat_panel_long <- dat_panel %>%
```

```
  long_panel(prefix='_', begin  = 1997, end = 2019, label_location = "end")

# create total work experience variable
dat_panel_work <- dat_panel_long %>%
  rowwise() %>%
  select(contains("CV_WKSWK_JOB")) %>%
  mutate(work_exp = round(rowSums(.-0, na.rm = TRUE)/52, digits = 2))
dat_panel_long$work_exp <- dat_panel_work$work_exp

# rename year, income
# create age, male, married variables
# recode edu into numeric variables
# keep only useful variables
# delete observations with NA's
dat_panel_long <- dat_panel_long %>%
  dplyr::rename(year = wave,
                income = "YINC-1700",
                edu = CV_HIGHEST_DEGREE_EVER_EDT) %>%
  rowwise() %>%
  dplyr::mutate(age = year - KEY_BDATE_Y,
                male = ifelse(KEY_SEX == 1, 1, 0),
                married = ifelse(CV_MARSTAT_COLLAPSED == 1, 1, 0),
                edu = ifelse(edu==0, 0,
                             ifelse(edu%in%c(1,2), 12,
                                    ifelse(edu==3, 14,
                                           ifelse(edu==4, 16,
                                                  ifelse(edu==5, 18,
                                                         ifelse(edu%in%c(6,7), 21, NA)))))) %>%
  select(id, year, income, age, male, married, work_exp, edu) %>%
  na.omit()
```

1) Estimate the model using Within Estimator

```
# create new data set to estimate within estimator
dat_within <- dat_panel_long

# calculate the mean income / edu / work_exp / married for each observation
dat_within$mean_income <- ave(dat_within$income, dat_within$id, FUN=function(x)mean(x, na.rm=T))
dat_within$mean_edu <- ave(dat_within$edu, dat_within$id, FUN=function(x)mean(x, na.rm=T))
dat_within$mean_work_exp <- ave(dat_within$work_exp, dat_within$id, FUN=function(x)mean(x, na.rm=T))
dat_within$mean_married <- ave(dat_within$married, dat_within$id, FUN=function(x)mean(x, na.rm=T))

# calculate the difference for each observation
dat_within <- dat_within %>%
  rowwise() %>%
  dplyr::mutate(income_diff = income - mean_income,
                edu_diff = edu - mean_edu,
                work_exp_diff = work_exp - mean_work_exp,
                married_diff = married - mean_married)

# estimate the model
```

```r
reg_within <- lm(income_diff ~ edu_diff + married_diff + work_exp_diff, data = dat_within)
summary(reg_within)
```

```
##
## Call:
## lm(formula = income_diff ~ edu_diff + married_diff + work_exp_diff,
##     data = dat_within)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -142657   -9109    -655    7775  276129
##
## Coefficients:
##                            Estimate           Std. Error t value
## (Intercept)       0.00000000001245    69.71834268464114    0.00
## edu_diff       1614.43833662861266    22.28931571108003   72.43
## married_diff  15382.39779299426664   220.76091151679455   69.68
## work_exp_diff  2819.24812026030168    26.22458977192016  107.50
##                           Pr(>|t|)
## (Intercept)                      1
## edu_diff      <0.0000000000000002 ***
## married_diff  <0.0000000000000002 ***
## work_exp_diff <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19970 on 82004 degrees of freedom
## Multiple R-squared:  0.3189, Adjusted R-squared:  0.3188
## F-statistic: 1.28e+04 on 3 and 82004 DF,  p-value: < 0.00000000000000022
```

2) Estimate the model using Between Estimator

```r
# create new data set to estimate between estimator
dat_between <- dat_within

# estimate the model
reg_between <- lm(mean_income ~ mean_edu + mean_married + mean_work_exp, data = dat_between)
summary(reg_between)
```

```
##
## Call:
## lm(formula = mean_income ~ mean_edu + mean_married + mean_work_exp,
##     data = dat_between)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -46496  -9540  -2566   5966 288941
##
## Coefficients:
##               Estimate Std. Error t value            Pr(>|t|)
## (Intercept)    3679.37     189.48   19.42 <0.0000000000000002 ***
```

```
## mean_edu       1282.69      14.85   86.38 <0.0000000000000002 ***
## mean_married   8446.63     181.11   46.64 <0.0000000000000002 ***
## mean_work_exp  1611.30      24.97   64.54 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15080 on 82004 degrees of freedom
## Multiple R-squared:  0.1728, Adjusted R-squared:  0.1728
## F-statistic:  5711 on 3 and 82004 DF,  p-value: < 0.00000000000000022
```

3) Estimate the model using Difference (any) Estimator

```r
# create new data set to estimate difference estimator
dat_difference <- dat_panel_long

# calculate
dat_difference$fir_income <- ave(dat_difference$income, dat_difference$id, FUN=function(x)x[1])
dat_difference$fir_edu <- ave(dat_difference$edu, dat_difference$id, FUN=function(x)x[1])
dat_difference$fir_married <- ave(dat_difference$married, dat_difference$id, FUN=function(x)x[1])
dat_difference$fir_work_exp <- ave(dat_difference$work_exp, dat_difference$id, FUN=function(x)x[1])
dat_difference$fd_income <- dat_difference$income - dat_difference$fir_income
dat_difference$fd_edu <- dat_difference$edu - dat_difference$fir_edu
dat_difference$fd_married <- dat_difference$married - dat_difference$fir_married
dat_difference$fd_work_exp <- dat_difference$work_exp - dat_difference$fir_work_exp

# estimate the model
reg_difference <- lm(fd_income ~ fd_edu + fd_married + fd_work_exp, data = dat_difference)
summary(reg_difference)
```

```
##
## Call:
## lm(formula = fd_income ~ fd_edu + fd_married + fd_work_exp, data = dat_difference)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -106518  -12497   -6245    7122  311896
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)  7378.01     138.84   53.14 <0.0000000000000002 ***
## fd_edu        813.38      14.14   57.51 <0.0000000000000002 ***
## fd_married  14728.12     207.25   71.06 <0.0000000000000002 ***
## fd_work_exp  2542.88      26.07   97.53 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25590 on 82004 degrees of freedom
## Multiple R-squared:  0.236,  Adjusted R-squared:  0.236
## F-statistic:  8444 on 3 and 82004 DF,  p-value: < 0.00000000000000022
```

3. Interpret the results from each model and explain why different models yield different parameter estimates

Here are the results of the three estimators:

	within	between	first-difference
education	1614.438	1282.693	813.3776
married	15382.398	8446.627	14728.1180
work experience	2819.248	1611.299	2542.8812

Within Estimator: **education**: Income will increase by 1614.44 on average for each additional year increase in education, controlling for all time-invariant heterogeneity. **married**: it is a dummy variable, which cannot be interpreted. **work experience**: Income will increase by 2819.25 on average for each additional year increase in work experience, controlling for all time-invariant heterogeneity.

Between Estimator: **education**: Income will increase by 1282.69 on average for each additional year increase in education, controlling for individual heterogeneity. **married**: it is a dummy variable, which cannot be interpreted. **work experience**: Income will increase by 1611.30 on average for each additional year increase in work experience, controlling for individual heterogeneity.

First-Difference Estimator: **education**: Income in year t will increase by 813.38 on average for each additional year increase in education in year t-1. **married**: it is a dummy variable, which cannot be interpreted. **work experience**: Income in year t will increase by 2542.88 on average for each additional year increase in work experience in year t-1.

We can observe from the above table that the three estimators produce very different results, but all of the coefficients are significant, and their values are positive. The three estimators generate different results since within estimator takes the time variation and discard the individual effects; between estimator takes the individual effects and discard the time variation; and first Difference estimator takes both effects.