

Assignment 1

David Gyarakı, Thao Le

Contents

| | | |
|----------|-------------------|----------|
| 1 | Question 1 | 2 |
| 1.1 | (i) | 2 |
| 1.2 | (ii) | 3 |
| 1.3 | (iii) | 3 |
| 1.4 | (iv) | 5 |
| 1.5 | (v) | 6 |
| 2 | Question 2 | 7 |

```
# load packages
if(!require(pacman)){install.packages("pacman")}

p_load(devtools,tidyverse,dplyr,ggplot2,latex2exp,cowplot,tseries,sampleSelection)

#load data
dfData = read.csv("assignment1_2023.csv")
attach(dfData)
```

1 Question 1

1.1 (i)

```
lm_model = lm(logwage ~ age + agesq + schooling, data = dfData)
summary(lm_model)
```

Call:

```
lm(formula = logwage ~ age + agesq + schooling, data = dfData)
```

Residuals:

| Min | 1Q | Median | 3Q | Max |
|---------|---------|--------|--------|--------|
| -3.3224 | -1.1782 | 0.0024 | 1.2208 | 3.1957 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|-----------|------------|---------|--------------|
| (Intercept) | 26.409280 | 8.057036 | 3.278 | 0.00113 ** |
| age | -0.341890 | 0.521078 | -0.656 | 0.51211 |
| agesq | -0.011142 | 0.008374 | -1.331 | 0.18408 |
| schooling | 0.215996 | 0.031534 | 6.850 | 2.71e-11 *** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.499 on 412 degrees of freedom

(250 observations deleted due to missingness)

Multiple R-squared: 0.8148, Adjusted R-squared: 0.8135

F-statistic: 604.3 on 3 and 412 DF, p-value: < 2.2e-16

1.2 (ii)

The sample selection problem here is to choose observations of the non-employed, which are those who have no income. The selection equation is then:

$$I_i = \begin{cases} 1 & \text{if } \log\text{wage} > 0 \\ 0 & \text{otherwise,} \end{cases}$$

and the second regression equation is:

$$Y_i^* = \mathbf{X}_i' \boldsymbol{\beta} + U_i.$$

We select a sample consisting of:

$$Y_i = \begin{cases} Y_i^* & \text{if } I_i = 1 \\ \text{missing} & \text{if } I_i = 0, \end{cases}$$

An OLS may fail in this context because the dependent variable (logwage) is missing for the non-employed sample, thus, it is not possible to derive an estimate of this variable for the non-employed

1.3 (iii)

The exclusion restriction variable is one that is included in \mathbf{Z}_1 but excluded from \mathbf{X}_1 , I would choose 'married' as a suitable candidate for the sample selection model. My motivation is that married people tends to have stable income, and thus, employed.

```
# Create I variable:
dfData = mutate(dfData, vI = if_else(logwage > 0, TRUE, FALSE))
dfData["vI"][is.na(dfData["vI"])] <- FALSE

# Heckman model with restriction
heckman_rest = heckit( vI ~ married+age + agesq + schooling, logwage ~ age + agesq + schoo
summary(heckman_rest)
```

```
-----
Tobit 2 model (sample selection model)
2-step Heckman / heckit estimation
666 observations (250 censored and 416 observed)
12 free parameters (df = 655)
Probit selection equation:
```

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|-----------|------------|---------|--------------|
| (Intercept) | -5.315285 | 5.293574 | -1.004 | 0.316 |
| married | 0.432572 | 0.100338 | 4.311 | 1.87e-05 *** |
| age | 0.332077 | 0.342618 | 0.969 | 0.333 |
| agesq | -0.005141 | 0.005512 | -0.933 | 0.351 |
| schooling | 0.018246 | 0.022309 | 0.818 | 0.414 |

Outcome equation:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|-----------|------------|---------|--------------|
| (Intercept) | 27.209400 | 8.517748 | 3.194 | 0.00147 ** |
| age | -0.385453 | 0.541932 | -0.711 | 0.47718 |
| agesq | -0.010459 | 0.008692 | -1.203 | 0.22932 |
| schooling | 0.214536 | 0.031874 | 6.731 | 3.69e-11 *** |

Multiple R-Squared:0.8148, Adjusted R-Squared:0.813

Error terms:

| | Estimate | Std. Error | t value | Pr(> t) |
|---------------|----------|------------|---------|----------|
| invMillsRatio | -0.1737 | 0.6148 | -0.283 | 0.778 |
| sigma | 1.4971 | NA | NA | NA |
| rho | -0.1160 | NA | NA | NA |

```
# Heckman model without restriction
```

```
heckman_unrest = heckit( vI ~ married+age + agesq + schooling, logwage ~ age + agesq + schooling)
```

```
summary(heckman_unrest)
```

Tobit 2 model (sample selection model)

2-step Heckman / heckit estimation

666 observations (250 censored and 416 observed)

13 free parameters (df = 654)

Probit selection equation:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|-----------|------------|---------|--------------|
| (Intercept) | -5.315285 | 5.293574 | -1.004 | 0.316 |
| married | 0.432572 | 0.100338 | 4.311 | 1.87e-05 *** |
| age | 0.332077 | 0.342618 | 0.969 | 0.333 |
| agesq | -0.005141 | 0.005512 | -0.933 | 0.351 |
| schooling | 0.018246 | 0.022309 | 0.818 | 0.414 |

Outcome equation:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|----------|------------|---------|-----------|
| (Intercept) | 94.14571 | 30.04111 | 3.134 | 0.0018 ** |
| age | -3.68541 | 2.91861 | -1.263 | 0.2071 |
| agesq | 0.04058 | 0.04759 | 0.853 | 0.3941 |
| schooling | 0.03527 | 0.19887 | 0.177 | 0.8593 |

```

married      -4.30249      NaN      NaN      NaN
Multiple R-Squared:0.8153, Adjusted R-Squared:0.8131
Error terms:
              Estimate Std. Error t value Pr(>|t|)
invMillsRatio -17.976      NaN      NaN      NaN
sigma          13.399      NA       NA       NA
rho            -1.342      NA       NA       NA
-----

```

- STILL NEED TO COMPARE OUTCOMES

1.4 (iv)

```

# Maximum likelihood estimator, restricted
ML_rest = selection(vI ~ married+age + agesq + schooling, logwage ~ age + agesq + schoolin
summary(ML_rest)

```

```

-----
Tobit 2 model (sample selection model)
Maximum Likelihood estimation
Newton-Raphson maximisation, 2 iterations
Return code 8: successive function values within relative tolerance limit (reltol)
Log-Likelihood: -1186.617
666 observations (250 censored and 416 observed)
11 free parameters (df = 655)
Probit selection equation:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.347695   5.290476  -1.011    0.312
married      0.432671   0.100314   4.313 1.86e-05 ***
age          0.334151   0.342394   0.976    0.329
agesq       -0.005174   0.005508  -0.939    0.348
schooling    0.018294   0.022308   0.820    0.412
Outcome equation:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.091220   8.430218   3.214 0.00138 **
age         -0.378997   0.537729  -0.705 0.48118
agesq       -0.010560   0.008627  -1.224 0.22139
schooling    0.214749   0.031784   6.757 3.13e-11 ***
Error terms:
              Estimate Std. Error t value Pr(>|t|)
sigma      1.49568     0.06006  24.902 <2e-16 ***
rho       -0.09931     0.37382  -0.266    0.791

```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
-----
```

```
# Maximum likelihood estimator, unrestricted
ML_unrest = selection(vI ~ married + age + agesq + schooling, logwage ~ age + agesq + schooling)
summary(ML_unrest)
```

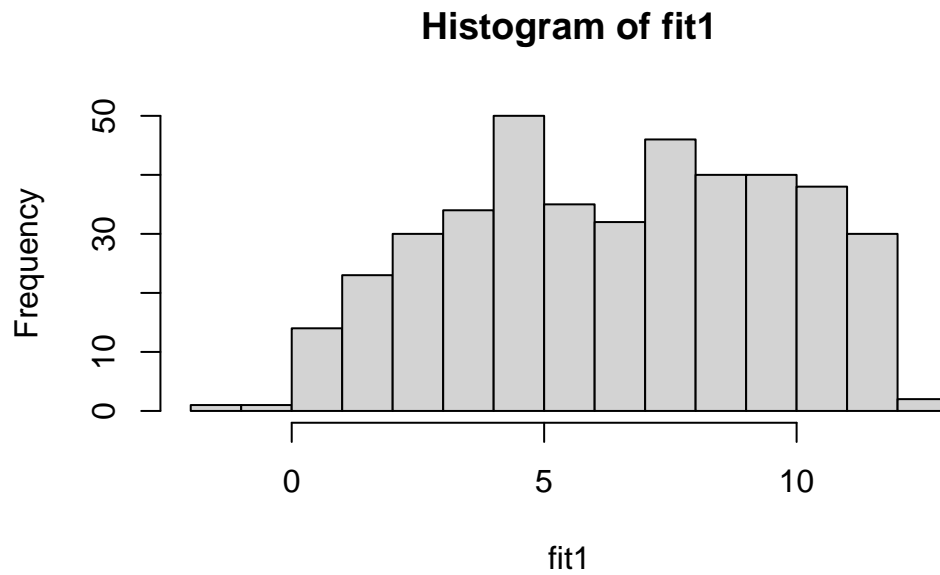
```
-----
Tobit 2 model (sample selection model)
Maximum Likelihood estimation
Newton-Raphson maximisation, 2 iterations
Return code 3: Last step could not find a value above the current.
Boundary of parameter space?
Consider switching to a more robust optimisation method temporarily.
Log-Likelihood: -1501.802
666 observations (250 censored and 416 observed)
12 free parameters (df = 654)
Probit selection equation:
      Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.315285    5.597136  -0.950    0.343
married      0.432572    0.099200   4.361 1.51e-05 ***
age          0.332077    0.362958   0.915    0.361
agesq       -0.005141    0.005849  -0.879    0.380
schooling    0.018246    0.021917   0.833    0.405
Outcome equation:
      Estimate Std. Error t value Pr(>|t|)
(Intercept) 94.10255    35.31853   2.664   0.0079 **
age         -4.29420     2.29499  -1.871   0.0618 .
agesq        0.05014     0.03695   1.357   0.1752
schooling    0.08000     0.14232   0.562   0.5742
married     -3.79447     0.54690  -6.938 9.57e-12 ***
Error terms:
      Estimate Std. Error t value Pr(>|t|)
sigma      7.648         NaN     NaN     NaN
rho       -0.990         NaN     NaN     NaN
-----
```

- STILL NEED TO COMPARE OUTCOMES

1.5 (v)

get fitted values => plot histogram

```
fit1 <- fitted(heckman_rest)
hist(fit1)
```



The distribution is relatively normal?

2 Question 2