Assignment 1

David Gyaraki, Thao Le

Contents

```r
# 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)

```r
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
```

Looking at our OLS results, firstly, our F-statistic is significant, which means that there is an association between at least of one of the predictor variables and logwage. Thus, we can move on to interpret the coefficients. The adjusted R-square of 0.8135 means that 81.35% of the variance in the dependent variable can be explained by the model. Looking at our OLS estimate, only the OLS estimate of schooling is significant. Thus, we can only interpret the effect of the variable schooling. There is an association between the years of schooling and the salary of a person. Holding other variables constant, a year of schooling is associated with around 0.2160 units increase in log salary of an individual.

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:

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

and the second regression equation is:

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

We select a sample consisting of:

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

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 $\mathbf{Z_i}$ but excluded from $\mathbf{X_i}$, 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
```

```r
heckman_rest = heckit( vI ~ married+age + agesq +
                      schooling, logwage ~ age + agesq +
                      schooling, data = dfData)
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
--------------------------------------------
```

```r
# Heckman model without restriction
heckman_unrest = heckit( vI ~ married + age + agesq +
                           schooling, logwage ~ age + agesq
                         + schooling + married, data = dfData)
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
--------------------------------------------
```

Looking at the outcomes of the two model, we can see that the unrestricted model have a much higher standard error, this is because the unrestricted model run into the problem of multicollinearity (i.e., the Inverse Mill Ratio is almost perfectly colliniear to the rest of the explanatory variables). Because of multicolinearity, the variable schooling is no longer statistically significant in the model. Thus, we are more unsure of our estimates.

1.4 (iv)

```
# Maximum likelihood estimator, restricted
ML_rest = selection(vI ~ married+age + agesq + schooling,
                    logwage ~ age + agesq + schooling, data = dfData)
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
--------------------------------------------
```

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

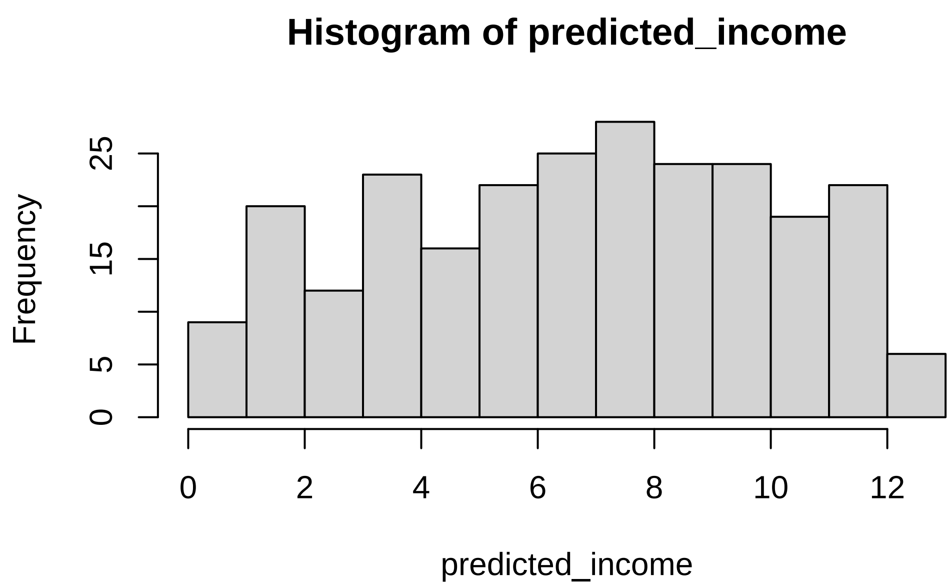
Similar to the situation in (iii), we can see that the unrestricted model have a much higher standard error, this is because the unrestricted model run into the problem of multicollinearity (i.e., the Inverse Mill Ratio is almost perfectly colliniear to the rest of the explanatory variables). Because of in increase in standard errors, the variable schooling is no longer statistically significant in the model. Thus, we are unsure of our estimates.

1.5 (v)

To specify the distribution of potential earnings for the non-employed, we first get a subsample of the unemployed individuals. Then, we use one of the restricted models in (iii) or (iv) to predict potential income of the non-employed and draw a histogram.

```r
# Get subsample of unemployed individuals
dfUnEmployed = dfData[dfData$vI == FALSE, ]

predicted_income = predict(ML_rest, newdata = dfUnEmployed)
hist(predicted_income)
```



The histogram does not give an apparent normal distribution. However, we can say that most predictions lies between 6 and 8, and the distribution is slightly left-skewed.

2 Question 2

2.1 (i)

```
# Get subsample of employed individuals
dfEmployed = dfData[dfData$vI == TRUE, ]

model0 = lm(logwage ~ schooling + age + agesq, data = dfEmployed)
summary(model0)
```

```
Call:
lm(formula = logwage ~ schooling + age + agesq, data = dfEmployed)

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 **
schooling    0.215996   0.031534   6.850 2.71e-11 ***
age         -0.341890   0.521078  -0.656  0.51211
agesq       -0.011142   0.008374  -1.331  0.18408
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.499 on 412 degrees of freedom
Multiple R-squared:  0.8148,    Adjusted R-squared:  0.8135
F-statistic: 604.3 on 3 and 412 DF,  p-value: < 2.2e-16
```

Looking at our OLS results, firstly, our F-statistic is significant, which means that there is an association between at least of one of the predictor variables and logwage. Thus, we can move on to interpret the coefficients. The OLS estimate of schooling is significant. Thus, we can only interpret the effect of the variable schooling. There is an *association* between the years of schooling and the salary of a person. Holding other variables constant, a year of schooling is associated with around 0.2160 increase in log salary of an individual.

However, we CANNOT discus the causal effect of schooling on income, because association is different from causation.

Regarding whether it is plausible that regularity conditions for applying OLS are satisfied. We believe it is plausible that some conditions such as homoskedasticity, X non-random, the error terms are normally distributed and has mean zero, no-auto correlation are satisfied, the condition that model is linear is also satisfied. However we are not sure if there is any multicollinearity between the explanatory variables yet. For example, the distance to school and the regional subsidy for school expenses might be correlated with the years of schooling of an individual. Specifically, people who have more subsidy and shorter distance to school could spend more years in school.

2.2 (ii)

```r
# Using distance as instrument variable
model1 = lm(schooling ~ distance, data = dfEmployed)
X.hat.1 = fitted.values(model1)

# Fit Linear regression model again using the fitted values of first step
model2 =lm(logwage ~ X.hat.1 + age + agesq, data = dfEmployed)
```

```r
summary(model2)
```

```
Call:
lm(formula = logwage ~ X.hat.1 + age + agesq, data = dfEmployed)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.4199 -1.2578 -0.0541  1.2115  3.5095 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept) 26.429485   8.687422   3.042   0.0025 **
X.hat.1      0.457857   0.284959   1.607   0.1089   
age         -0.460586   0.547910  -0.841   0.4010   
agesq       -0.009026   0.008804  -1.025   0.3059   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.577 on 412 degrees of freedom
Multiple R-squared:  0.795, Adjusted R-squared:  0.7935 
F-statistic: 532.6 on 3 and 412 DF,  p-value: < 2.2e-16
```

```r
# Using subsidy as instrument variable
model3 = lm(schooling ~ subsidy , data = dfEmployed)
X.hat.3 = fitted.values(model3)

# Fit Linear regression model again using the fitted values of first step

model4 =lm(logwage ~ X.hat.3 + age + agesq , data = dfEmployed)
summary(model4)
```

```
Call:
lm(formula = logwage ~ X.hat.3 + age + agesq, data = dfEmployed)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.4083 -1.2306 -0.0321  1.3012  3.6294 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) 25.889824   8.403013   3.081 0.002201 **
X.hat.3      0.410382   0.108613   3.778 0.000181 ***
age         -0.413512   0.540517  -0.765 0.444691
agesq       -0.009649   0.008684  -1.111 0.267186
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.555 on 412 degrees of freedom
Multiple R-squared:  0.8006,    Adjusted R-squared:  0.7992
F-statistic: 551.5 on 3 and 412 DF,  p-value: < 2.2e-16
```

```r
# Using subsidy and distance as instrument variable
model5 = lm(schooling ~ subsidy+distance, data = dfEmployed)
X.hat.5 = fitted.values(model3)

model6 =lm(logwage ~ X.hat.5 + age + agesq, data = dfEmployed)
summary(model6)
```

```
Call:
lm(formula = logwage ~ X.hat.5 + age + agesq, data = dfEmployed)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4083 -1.2306 -0.0321  1.3012  3.6294

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 25.889824   8.403013   3.081 0.002201 **
X.hat.5      0.410382   0.108613   3.778 0.000181 ***
age         -0.413512   0.540517  -0.765 0.444691
agesq       -0.009649   0.008684  -1.111 0.267186
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.555 on 412 degrees of freedom
Multiple R-squared:  0.8006,    Adjusted R-squared:  0.7992
F-statistic: 551.5 on 3 and 412 DF,  p-value: < 2.2e-16
```

After using 3 options of instrument variables, we can see that ‘distance’ is not a good instrument variable as the part of ‘schooling’ not explained by ‘distance’ (stored in ‘X.hat.1’) is not statistically significant in the linear model. On the other hand, after adding ‘subsidy’ as an instrument variable, the part of ‘schooling’ not correlated with ‘subsidy’ is statistically significant in explaining the changes in ‘logwage’. We will only use ‘subsidy’ as the instrument variable, not a combination of both ‘subsidy’ and ‘schooling’, because using both variables can lead to the issue of over-identification.

2.3 (iii)

```
# OLS outcomes
model0$coefficients
```

```
(Intercept)   schooling         age       agesq
 26.4092796   0.2159957  -0.3418898  -0.0111417
```

```
# IV outcomes
model4$coefficients
```

```
 (Intercept)      X.hat.3          age        agesq
25.889824395  0.410381734 -0.413511745 -0.009648645
```

Above are the OLS and IV estimates, in which, X.hat.3 in the IV estimates is the part of ‘schooling’ that is not correlated with subsidy.

Looking at the outcomes, we can see that the IV estimate of schooling coefficient is higher than the OLS estimate. This might be because the effect of subsidy on schooling is eliminated, making the number of years of schooling smaller, and thus it needs higher weights to predict the income. Both OLS and IV estimates yielded significant effect from the schooling variable.

We would prefer OLS in the case where there is no correlation between explanatory variables and the error terms.

To decide between OLS and IV, we first perform a t-test to check the relevance of the instrument (i.e., checking whether ‘subsidy’ and ‘schooling’ are correlated).

```
summary(model3)
```

```
Call:
lm(formula = schooling ~ subsidy, data = dfEmployed)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.5382 -1.5382  0.0031  1.7324  4.8151 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.83150    0.30298  15.947  < 2e-16 ***
subsidy      0.27067    0.04217   6.419 3.77e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.238 on 414 degrees of freedom
Multiple R-squared:  0.09051,	Adjusted R-squared:  0.08831 
F-statistic:  41.2 on 1 and 414 DF,  p-value: 3.769e-10
```

We can see that there is statistically sigfificant evidence that there is an association between 'subsidy' and 'schooling'. Thus, 'subsidy' is a relevant instrument. Moreover, we also need to check the validity of the instrument using the Sargan test:

```r
sargan_test = lm(model3$residuals ~ subsidy + age + agesq, data = dfEmployed)

test_statistics <- summary(sargan_test)$r.squared*nrow(dfEmployed)

print(1-pchisq(test_statistics,1))  # prints p-value
```

```
[1] 0.01679377
```

We can see that the p-value is 0.0167793, which is significant at $\alpha = 0.05$. This means that our instrument variable is valid.

Using the result of both the t-test and the Sargan test, we choose to use the IV estimate instead of the OLS one as the IV estimate utilises a relevant and valid instrument variable.