Assignment 1

Walter Verwer & Bas Machielsen

January 8, 2021

Question 1: The sample selection model.

A researcher aims to gain insight in the potential earnings of the non-employed. (In the data, the non-employed can be identified by a missing value for the earnings variable). She realizes that the sample of observed wages may be subject to sample selection.

(a) Run an OLS regression for log-earnings on schooling, age, and age squared. Present the results and comment on the estimates.

```
model_1 <- lm(data = data, formula = logWage ~ schooling + age + age2)

results_1 <- summary.lm(model_1)
coeffs_1 <- results_1$coefficients[,1]
pvals_1 <- results_1$coefficients[,4]

stargazer(model_1, style = "AER",
          font.size = "small",
          header = F, label = 'tab:q1_a_ols',
          title = 'OLS regression for log-earnings on schooling, age and age squared.')
```

The results show that one additional year of schooling has an effect of 0.216 on log(Wage), which means that one additional year of schooling has an estimated 21.600% effect on wages earned. This result is highly significant (p-value is 0.000) Another result, shown in figure 1 is that being one year older has an estimated effect of -0.342 on log(Wage). Thus, being one year older is estimated to have a -34.189% on wages. This result is not significant at a common level (p-value is 0.512). The third estimated coefficients is the one corresponding to the variable age squared. It has an estimated coefficient of -0.011, which represents the estimated effect of a one unit increase in age squared on log(Wage). This also means that a one unit increase in age squared is equal to an estimate effect of -1.114% on the linear representation of wage. This effect is not significant at common levels, because the p-value is 0.184. A note about both age and age squared is that even though both are not significant, their estimated sign is negative. This is against expectations, because it is normally the case that a more senior individual earns a higher wage. Finally, the estimated constant in the model is estimated at 26.409. This means that if all other variables are zero, then the log(Wage) will be equal to 26.409. Thus, if all other variables are zero, then wage is estimated to be equal to exp(26.409)=294716462317.194. This is an extremely high number given the characteristics of the wage variable. However, this result is highly significant, because the p-value is 0.001. Of course it is also not realistic, because for example it can not be the case that someone has earnings and has a zero age.

(b) Briefly discuss the sample selection problem that may arise in using these OLS estimates for the purpose of predicting the potential earnings of the non-employed. Formulate the sample selection model. In your answer, include an explanation why OLS may fail in this context.

An individual is only in this data set if they earn wages, i.e. if they are employed. Being employed itself is not randomly allocated, but rather, a function of e.g. age, age squared, and schooling. Also, employment

Table 1: OLS regression for log-earnings on schooling, age and age squared.

	logWage
schooling	0.216^{***}
	(0.032)
age	-0.342
	(0.521)
age2	-0.011
	(0.008)
Constant	26.409^{***}
	(8.057)
Observations	416
R^2	0.815
Adjusted R^2	0.813
Residual Std. Error	1.499 (df = 412)
F Statistic	604.261^{***} (df = 3; 412)
Notes:	***Significant at the 1 percent level.
	**Significant at the 5 percent level.
	*Significant at the 10 percent level.

status follows from labor supply and demand forces. Hence, the estimates of schooling on earnings are conditional on having earnings to begin with, whereas unbiased estimates must also include those individuals. Formally, $\mathbb{E}[\text{Earnings}] = \mathbb{E}[\text{Earnings}|\text{Having a job}] \cdot \mathbb{P}[\text{Having a job}] + \mathbb{E}[\text{Earnings}|\text{Not having a job}] \cdot (1 - \mathbb{P}[\text{Having a job}])$. The given estimation only concerns $\mathbb{E}[\text{Earnings}|\text{Having a job}]$.

The sample selection model is given by the following two equations.

$$\mathbb{P}[I_i = 1|Z] = \mathbf{\Phi}(Z_i'\gamma) \tag{1}$$

In equation 1, I_i denotes an indicator variable which is equal to 1 if we observe the wage of an individual, and Z_i denotes a vector of explanatory variables for the probability of an individual being employed or not.

Under our distributional assumptions, in the selection model $v_i \sim N(0,1)$, and therefore $P(I_i = 1) = P(I_i^* > 0) = \mathbf{\Phi}(Z_i'\gamma)$, where $\mathbf{\Phi}$ denotes a normal CDF. Note that the model used in this situation concerns a probit model, which has the goal to estimate the probability that an individual is employed.

The second equation is concerned with explaining the wage of an individual. It is given by the following equation.

$$Y_i^* = X_i'\beta + U_i \tag{2}$$

In equation 2, Y_i^* denotes the observed log(Wage), X_i are the explanatory regressors for explaining log(Wage).

(c) Which variable in your data may be a suitable candidate as an exclusion restriction for the sample selection model?

For an exclusion restriction variable, we need a variable that is theoretically unrelated to earnings, but related to the probability of having a job. Empirically, we need a variable significantly different from zero in the selection equation and that it does not have an effect in the equation intended to explain the potential earnings of the non-employed.

A potential candidate for this is a variable that matters for being employed or not, but does not influence the height of the wage. We believe that marriage status could be a suitable candidate variable. The reason being that marriage status could matter for being employed or not, because if one is not married, the person is more likely to be the only person that has to provide the necessary funds of living. If someone is married however, than it is more likely that this person is not employed, because it might be the case that the partner provides. Marriage status is arguably a variable that does not provide a direct effect on the height of wages earned. This holds if we assume that wages only reflect an individual's marginal productivity of labor and that this is not influenced by marriage status. This is however doubtful if for example married individuals are happier and happiness influences one's productivity. Seeing as both criteria are met (given our assumption), we conclude that marriage status is a suitable candidate for an exclusion restriction.

(d) Estimate the sample selection model with the Heckman two-step estimator, both with and without the exclusion restriction and compare the outcomes.

For this question we are asked to estimate the Heckman two-step estimator, both with and without the exclusion variable. We have argued that married would be a suitable candidate for this variable. In the code below we have done the following. First we have estimated the sample selection model with the two-step approach, *including* the exclusion variable in the selection regression, and *excluding* it in the estimation regression. This is in a way as it should be done. The second thing we did was estimating the selection regression *without* the exclusion variable married and in the second stage using the exact same collection of independent variables in the estimation regression. One thing to note about our code is that we have used our own code to produce the two-step estimator as well as a package. Our own code produced the exact same results as the package. Seeing as the package provides us with more detailed results, we show it's results in our output table, displayed in table 2.

```
# Construct I (I=1 for y_i^* != na, else 0)
data$I <- ifelse(is.na(data$logWage) , 0, 1) # if na, then I is zero. Else 1.

## Two stage approach without package:
# Stage 1:
# probit <- glm(I ~ schooling + age + age2 + married,
#               family = binomial(link = "probit"),
#               data = data)
#
# # Construct variable for stage 2:
# Z_gamma <- cbind(1,data$schooling, data$age, data$age2, data$married) %*% coef(probit)
# inverse_mills_ratio <- dnorm(Z_gamma, mean=0, sd=1) / pnorm(Z_gamma, mean=0, sd=1)
#
# # Stage 2:
# sample_selection_2s <- lm_robust(logWage ~ schooling + age + age2 + inverse_mills_ratio,
#                                  se_type='HC1' ,data=data)

sample_selection_2s_with <- selection(I ~ schooling + age + age2 + married,
                                      logWage ~ schooling + age + age2, data=data,
                                      method='2step')
coeffs_sample_selection_2s_with <- sample_selection_2s_with$coefficients


## using package 'sampleSelection' to get maximum likelihood estimates:
sample_selection_2s_without <- selection(I ~ schooling + age + age2,
                                      logWage ~ schooling + age + age2, data=data,
                                      method='2step')
```

```
## Warning in heckit2fit(selection, outcome, data = data, weights = weights, :
## Inverse Mills Ratio is (virtually) collinear to the rest of the explanatory
```

```
## variables
```

```
coeffs_sample_selection_2s_without <- sample_selection_2s_without$coefficients


# Obtain output:
stargazer(sample_selection_2s_with, sample_selection_2s_without,
          style = "AER",
          font.size = "small",
          header = F, label = 'tab:q1_d',
          column.labels = c("Two-step with married", "Two-step without married"),
          title = 'Log earnings sample selection regression with two-step approach,
          with and without the exclusion variable.')
```

Table 2: Log earnings sample selection regression with two-step approach, with and without the exclusion variable.

	logWage	
	Two-step with married	Two-step without married
	(1)	(2)
schooling	0.215***	0.303
	(0.032)	(0.735)
age	−0.385	1.423
	(0.542)	(14.875)
age2	−0.010	−0.039
	(0.009)	(0.234)
Constant	27.209***	−6.436
	(8.518)	(276.123)
Observations	666	666
ρ	−0.116	1.290
Inverse Mills Ratio	−0.174 (0.615)	7.341 (61.129)
Notes:	***Significant at the 1 percent level.	
	**Significant at the 5 percent level.	
	*Significant at the 10 percent level.	

For our results we first note an important difference between the two choices of independent variable selections. The inverse Mills ratio of the model that does not include marriage in the selection stage is perfectly collinear with the rest of the explanatory variables. The result of this is that our standard errors are much larger for the model that is estimated without marriage than for the model that is estimated with marriage. From this we can conclude that excluding marriage leads to a large loss of estimation efficiency. This is also to be expected, because multicollinearity causes the variance of the estimator to be inflated. Another observation that can be made is the change in the point estimates of the coefficients. For example, the constant changes from 27.209 with marriage, to -6.436 without marriage. Interestingly, we do find a positive coefficient for the age variable and a negative coefficient for age squared (albeit both are not significant). These findings hint towards there being a diminishing marginal effect of age on log(Wage). This finding is in-line with what one would expect to find in reality. Something else that can be observed is that $\rho > 1$ for the model without marriage included in the selection equation. This is not possible given the fact that ρ represents a correlation coefficient and thus should be between 0 and 1 in absolute value. A final observation that can be made is the large difference in the inverse Mills ratio coefficients and standard error. That is

the coefficient for the model with marriage is smaller and negative, in comparison to the model without marriage. For the standard errors, the model with marriage appears to be more efficient than the model without marriage.

(e) Estimate the sample selection model with Maximum Likelihood, both with and without the exclusion restriction and compare the outcomes.

```
# Construct I (I=1 for y_i^* != na, else 0)
sample_selection_ml_with <- selection(I ~ schooling + age + age2 + married,
                                      logWage ~ schooling + age + age2, data=data,
                                      method='ml')
coeffs_sample_selection_ml_with <- sample_selection_ml_with$estimate[6:9]


## using package 'sampleSelection' to get maximum likelihood estimates:
sample_selection_ml_without <- selection(I ~ schooling + age + age2,
                                      logWage ~ schooling + age + age2, data=data,
                                      method='ml')
```

```
## Warning in heckit2fit(selection, outcome, data = data, printLevel =
## printLevel, : Inverse Mills Ratio is (virtually) collinear to the rest of the
## explanatory variables
```

```
coeffs_sample_selection_ml_without <- sample_selection_ml_without$coefficients

# Obtain output:
stargazer(sample_selection_ml_with, sample_selection_ml_without,
          style = "AER",
          font.size = "small",
          header = F, label = 'tab:q1_e', digits=3,
          column.labels = c("ML with married", "ML without married"),
          title = 'Log earnings sample selection regression with maximum likelihood,
          with and without the exclusion variable.')
```

In table 3 we have displayed our estimation results using maximum likelihood, for the model with and without the exclusion restrictions. Again we observe a change in the point estimates as a result of the change in variable selection. The estimates are rather comparable to the estimates obtained for the two-step Heckman estimator without using marriage as an exclusion variable. An interesting observation to be made is that we are unable to retrieve standard errors for the model where we have excluded marriage. This is likely due to the fact that the perfect multicollinearity causes the optimizer used for maximum likelihood to converge to zero for the standard errors, or there are problems with invertibility of the inner product of the dependent variables. For the estimate of the correlation coefficient ρ we notice a similar result for the models where we used the two-stage approach. We observe that the correlation coefficient moves from being negative under the model with marriage to 1.000 under the model without marriage.

(f) On the basis of your results, how would you specify the distribution of potential earnings for the non-employed?

For this question we can characterize the distribution of potential earning for the non-employed by using our model based on maximum likelihood and with marriage. The reason for the choice of the maximum likelihood model is that it is more efficient than the two-step approach, because we have hetereoskedastic errors for the two-step approach. However, we could have possibly chosen to estimate the model with the simple OLS model. There are two reasons for this. First, the estimates of the models do not differ much. Second, and perhaps more informative, in the two-step Heckman estimator, the inverse Mills ratio appears the have an insignificant coefficient, which hints towards the absence of a sample selection bias. In the end

Table 3: Log earnings sample selection regression with maximum likelihood, with and without the exclusion variable.

	logWage	
	ML with married	ML without married
	(1)	(2)
schooling	0.215***	0.274
	(0.032)	(Inf.000)
age	−0.379	1.594
	(0.538)	(Inf.000)
age2	−0.011	−0.042
	(0.009)	(Inf.000)
Constant	27.091***	−6.423
	(8.430)	(Inf.000)
Observations	666	666
Log Likelihood	−1,186.617	−1,281.546
ρ	−0.099 (0.374)	1.000 (Inf.000)
Notes:	***Significant at the 1 percent level.	
	**Significant at the 5 percent level.	
	*Significant at the 10 percent level.	

we do choose for the sample selection model based on maximum likelihood, we believe that the theoretical argument for a selection bias is strong in this case.

The way we characterize the distribution is by predicting the values of the log(Wage) of the unemployed individuals. This is done by simply filling in the observed data for the unemployed in the model and then predict via the estimated model parameters. The code that makes the prediction is shown below, as well as a kernel density plot of the predicted log(Wage) of the unemployed.

```
# Predict log(Wage):
data$est_log_wage<-NaN
for (i in c(1:nrow(data))){
  data$est_log_wage[i] <- coeffs_sample_selection_ml_with %*% cbind(1,
                                                                    data$schooling,
                                                                    data$age,
                                                                    data$age2)[i,1:4]
}

# Dummy that is 1 for being unemployed:
data$d_unem <- ifelse(is.na(data$logWage) , 1, NaN) # if na, then d_unem is 1. Else NaN.

# Construct vector of unemployed log(Wage) predictions:
log_wage_unemployed <- na.omit(data$d_unem * data$est_log_wage)

# Histogram + kernel density plot of log(Wage) predictions of the unemployed:
hist(log_wage_unemployed, # histogram
 col="gray", # column color
 border="black",
 prob = TRUE, # show densities instead of frequencies
 xlab = "log(Wage)",
```

```
 xlim = c(0,14),
 main = "")
lines(density(log_wage_unemployed), # density plot
 lwd = 2, # thickness of line
 col = "black",
 xlim = c(0,14))
```

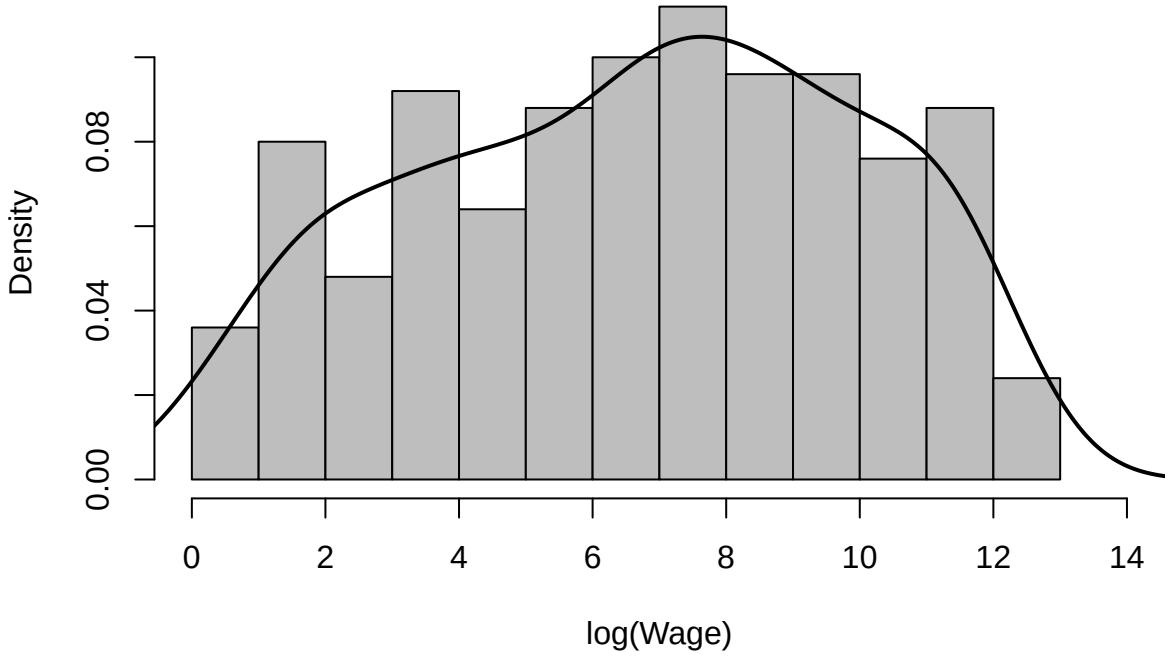

Figure 1: Historgram and kernel density plot of predicted $log(Wage)$ of the unemployed.

Question 2: Earnings and Schooling

The same researcher is interested in estimating the causal effect of schooling on earnings for employed individuals only. As a consequence, she performs the subsequent analysis on the (sub)sample of employed individuals.

(a) Discuss the estimation of the causal effect of schooling on earnings by OLS. In particular, address whether or not it is plausible that regularity conditions for applying OLS are satisfied.

It is not plausible that the regularity conditions are satisfied. In particular, an observable such as an individual's **ability** might be correlated with the wage, but also with the decision to live close to a school. Hence, the estimates suffer from endogeneity.

(b) The researcher has collected data on two potential instrumental variables subsidy and distance for years of schooling.

- **distance measures the distance between the school location and the residence of the individual while at school-going age.**
- **subsidy is an indicator depending on regional subsidies of families for covering school expenses.**

The researcher has the option to use only distance as an instrumental variable, or to use only the instrumental variable subsidy, or to use both distance and subsidy as instrumental variables. Perform instrumental variables estimation for these three options. Which option do you prefer? Include in your answer the necessary analyses and numbers on which you base your choice.

```
firstoption <- ivreg(data = data, formula =
                          logWage ~ age + age2 + schooling | distance + age + age2)

secondoption <- ivreg(data = data, formula =
                           logWage ~ age + age2 + schooling | subsidy + age + age2)

thirdoption <- ivreg(data = data, formula =
                           logWage ~ age + age2 + schooling | subsidy + distance + age
                     + age2)
```

```
stargazer(firstoption, secondoption, thirdoption, font.size = "small",
          style = "AER",
          header = F)
```

We consider that the second option, to include only subsidy as an instrument, is the best option. The reason is that distance is unlikely to satisfy the exclusion restriction: distance is (to a certain extent) an endogenous variable: wealthier (or more able) parents may choose to live closer to school, and invest more in the education of their children (or genetically transmit ability). Since a potentially endogenous instrument must not be used as such, we prefer the estimates in equation 2. However, we see that the results show that distance has no predictive power in schooling, thus showing that the endogeneity is very small. Conditional on subsidy being a good instrument, then, the potential endogeneity does not substantially changes the estimates of schooling on earnings.

(c) Compare the IV estimates with the OLS outcomes. Under which conditions would you prefer OLS over IV? Perform a test and use the outcome of the test to support your choice between OLS and IV. Motivate your choice.

We first observe that the OLS estimate $\beta = 0.216$ is about half the magnitude of the IV-estimate. This means that the bias generated by OLS likely *downplays* the actual effect (if the IV estimates satisfy the exclusion

Table 4:

	logWage		
	(1)	(2)	(3)
age	−0.192	−0.233	−0.229
	(0.587)	(0.546)	(0.547)
age2	−0.014	−0.013	−0.013
	(0.010)	(0.009)	(0.009)
schooling	0.470	0.401***	0.408***
	(0.299)	(0.106)	(0.102)
Constant	22.681**	23.694***	23.589***
	(9.704)	(8.517)	(8.530)
Observations	416	416	416
R^2	0.786	0.799	0.798
Adjusted R^2	0.784	0.798	0.797
Residual Std. Error (df = 412)	1.613	1.560	1.565
Notes:	***Significant at the 1 percent level.		
	**Significant at the 5 percent level.		
	*Significant at the 10 percent level.		

restriction). In case we would not trust the IV assumptions, we would prefer to trust the (conservative) estimate that downplays the effect, i.e. the OLS estimates. We can test whether the OLS estimates are substantially different from the IV estimates by conducting a Hausman test:

```
hoi <- summary(
    secondoption,
        diagnostics=TRUE)

hoi$diagnostics
```

```
##                  df1 df2 statistic      p-value
## Weak instruments   1 412 43.319777 1.416463e-10
## Wu-Hausman         1 411  3.634148 5.730274e-02
## Sargan             0  NA        NA           NA
```

The null hypothesis in the Hausman test is exogeneity of the *schooling* variable. As becomes clear, the null hypothesis is marginally rejected, implying the *schooling* is endogenous, but only marginally so. Hence, we would prefer to trust the IV estimates in this case.