Assignment 1

Walter Verwer & Bas Machielsen

January 4, 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.

```r
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',
          caption = 'OLS regression for log-earnings on schooling, age and age squared.')
```

The results show that 1 additional year of schooling has an effect of 0.216 on log(Wage), which means that 1 additional year of schooling has an estimated 21.6% effect on wages earned. This result is highly significant (p-value is 2.706×10^{-11}) Another result shown in figure 2 is that being 1 year older has an estimated effect of -0.342 on log(wage). Thus, being 1 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).

(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 dataset 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^2, and schooling. 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 $\mathbb{E}[\text{Earnings}|\text{Having a job}]$.

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

Table 1:

	logWage
schooling	0.216^{***}
	(0.032)
age	-0.342
	(0.521)
age2	-0.011
	(0.008)
Constant	26.400^{***}
	(8.060)
Observations	416
R^2	0.815
Adjusted R^2	0.813
Residual Std. Error	1.500 (df = 412)
F Statistic	604.000^{***} (df = 3; 412)
Notes:	***Significant at the 1 percent level.
	**Significant at the 5 percent level.
	*Significant at the 10 percent level.

Table 2:

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

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.

```r
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)
```

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

Table 3:

	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	22.700**	23.700***	23.600***
	(9.700)	(8.520)	(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)	1.610	1.560	1.560
Notes:	***Significant at the 1 percent level. **Significant at the 5 percent level. *Significant at the 10 percent level.		

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 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: