Assignment 1

Students

January 6, 2023

1 Question 1

1.1 i)

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

	Dependent variable:
	logwage
schooling	0.216^{***}
	(0.032)
age	-0.342
	(0.521)
I(age^2)	-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)
Note:	*p<0.1; **p<0.05; ***p<0.01

From Table 1 can be observed that only the intercept and `schooling` are significant. Both are significant at the 1%-significance level. For a given worker, an additional year of schooling is associated with a $(e^{0.216} - 1) \cdot 100 \approx 24.11\%$ increase in wage. The intercept and explanatory variables explain 81.5% of the variation in `logwage`.

1.2 ii)

`logwage` is only observed when a worker earns a wage.

Let Z_i', γ, and V_i denote some exclusion restriction, the corresponding coefficient, and error term for individual i,

respectively. Then, we have the selection equation

$$I_i^* = Z_i^{'}\gamma + V_i,$$

where I_{i} takes value 1 if the wage for individual i $I_i^* > 0$, and value 0 otherwise. In addition, let Y_i^* denote the latent variable `logwage`, $X_i^{'}$ the regressor(s), β the coefficient (vector) and U_i the error term. The second regression equation is

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

However, instead of the true `logwage` Y_i^* we observe

1.3 iii)

1.4 iv)

1.5 v)

2 Question 2

2.1 i)

An OLS model of the form

$$earnings = \beta_0 + \beta_1\ schooling + \epsilon$$

likely does not yield consistent estimates as the exogeneity assumption $E[\mathbf{X}\epsilon]$ is going to be violated. In this case, the residual contains individual's characteristics such as intelligence, which should both influence the amount of schooling an individual seeks, as well as directly impact the income of an individual.

2.2 ii)

Table 2 displays the results for the different IV models and the OLS estimation.

Based on the F-statistics of the IV regressions, we would choose the model that only uses the *subsidy* as an instrumental variable, as its F-statistic is the highest out of all 3 IV models, even higher than when using both instrumental variables. As the estimates of these two models seem to be the same, we choose the simpler model which only contains the *subsidy* instrumental variable. Note that *distance* appears to be a weak instrument, as its model's F-statistic is low and its estimates are close to the, likely inconsistent, OLS estimates.

2.3 iii)

The OLS results are similar to the results with the IV regression where we use *distance* as an instrumental variable. The estimates for the other two IV regressions differ strongly from the OLS estimates, but are similar to each other.

If there is no endogeneity, than we would use the OLS model, as it is efficient in this context. Assuming that we have valid instruments, we can test for endogeneity by the Hausman test.

We present the results for the Hausman test in Table 3, clearly we cannot reject the Hausman test for the *distance* instrument variable. However, we can reject the Hausman test on $\alpha = 0.05$ for the *subsidy* and for both instrumental variables.

These results and the theoretical background suggest, that there is an endogeneity problem and that *distance* is a weak instrument. Hence, we would choose the simplest model which seems to address this endogeneity issue, which is the model with the *subsidy* instrument.

Table 2: Regression for log-earnings on schooling.

	Dependent variable:			
	logwage			
	OLS	*instrumental variable*		
	(1)	(2)	(3)	(4)
schooling	0.101	0.100	0.730***	0.658***
	(0.073)	(0.626)	(0.262)	(0.245)
Constant	5.778***	5.789	1.599	2.078
	(0.512)	(4.163)	(1.753)	(1.639)
IV variable	-	distance	subsidy	both
F Statistic IV	-	5.6453	41.1994	23.0049
Observations	416	416	416	416
R^2	0.005	0.005	−0.176	−0.137
Adjusted R^2	0.002	0.002	−0.179	−0.139
Residual Std. Error (df = 414)	3.467	3.467	3.768	3.705
F Statistic	1.950 (df = 1; 414)			
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 3: P-values of Hausman test for the IV models

IV	pvalue
IV Distance	0.9979
IV Subsidy	0.0061
IV Both	0.0103