Econometrics II - Assignment 1

Floris Holstege, Stanislav Avdeev

Exercise 1

A

We specify the model as follows:

$$Log(Earnings) = \beta_0 + Schooling\beta_1 + Age\beta_2 + Age^2\beta_3$$

We run an OLS regression with this model, the results of which can be found below. The only statistically significant variable (5% threshold) is the measure for schooling. On average, an increase of one year of schooling leads to a 22% increase in earnings.

Table 1: The effect of Schooling, Age and Age^2 on earnings

	Dependent variable:
	Log of Wage
Schooling	0.22^{***}
	(0.03)
Age	-0.34
	(0.52)
Age^2	-0.01
	(0.01)
Intercept	26.41^{***}
	(8.06)
Observations	416
R^2	0.82
F Statistic	604.261^{***} (df = 3; 412)
Note:	*p<0.1; **p<0.05; ***p<0.01

B

The problem is a selection on unobservables and observables. We have a selected sample, excluding the unemployed. This means we have no information on their potential earnings, if they were employed. We cannot just extrapolate our findings for the employed to the non-employed, since the selection influences both

unobservable variables (for example; the unemployed might have less motivation) and observable variables (for instance; the effect of years of schooling on earnings might be less strong for people who have not worked for some years). We model this selection bias in two steps. First, we indicate if the dependent variable is observed as follows:

$$I_i^* = Z_i'\gamma + V_i$$

The indicator variable I_i takes value 1 if $I_i^* > 0$, and 0 if $I_i^* \leq 0$. Second, we define our latent variable Y_i^*.

$$Y_i^* = X_i'\beta + U_i$$

We use these two definitions to define our observed dependent variable, Y_i.

$$Y_i = \begin{cases} Y_i^* & \text{for } I_i = 1 \\ Missing & \text{for } I_i = 0 \end{cases}$$

If we try to estimate Y_i with OLS, it will be consistent under one of two conditions. First, if U_i and V_i are independent. The intuition is that in this case, there is random sampling, which means the sample is not biased. But this is not the case here, since the observations are unobserved based on a criterion (unemployment). Second, if X_i and Z_i are uncorrelated. The intuition here is that when this holds, the variables that determine if one falls outside of the selected sample (Z_i) are unrelated to the the independent variables used in the selected sample (X_i), and thus the zero mean condition ($E(U_i|X_i) = 0$) still holds. But this is again, highly questionable here - for example, a low level of schooling makes it more likely someone becomes unemployed, and negatively affects their earnings. If neither of these two conditions hold, which appears to be the case here, the OLS is inconsistent.

C

The variable for the exclusion restriction should fulfill two conditions. First, the variable should have explanatory power when determining I_i. In this case, the variable should realistically influence the likelihood someone is unemployed. Second, the variable should be unrelated to the dependent variable Y_i, in this case earnings. From the variables available, our best candidate appears to be the dummy variable of married (1 if married, 0 if not). Regarding the first condition; When someone is married, this affect their willingness to continue working, since they share income with their partner. Regarding the second condition; if one assumes that wages are reflective of someones productivity, the fact that someone is married, should not meaningfully change one's productivity. However, there are some reasons to believe marriage still impacts wages - for one, married couples are more likely to have a baby, which likely affects ones earnings. But given the available variables, this appears the one that comes closest to fulfilling both criteria. Looking at the correlations, these confirm our intuitions; the correlation between marriage and unemployment is 0.17, but only 0.02 with logged wages.

D

We first estimate without the exclusion restriction, in which $X_i = Z_i$ for the Heckman estimator. Because these terms are equal, there is collinearity in the second step of the Heckman estimator, biasing the estimates. To explain this a bit more formally, consider the two steps in the Heckman estimator:

Step 1: Using a probit model, estimate I_i with: $\hat{I}_i = Z_i'\hat{\gamma}_i + V_i$

Step 2: Estimate β and $\rho\sigma$ with: $Y_i = X_i\beta + \rho\sigma\frac{\phi(Z_i'\hat{\gamma}_i)}{\Phi(Z_i'\hat{\gamma}_i)} + U_i^*$

If $X_i = Z_i$, then:

$$\frac{\phi(Z_i'\hat{\gamma}_i)}{\Phi(Z_i'\hat{\gamma}_i)} = \frac{\phi(X_i'\hat{\gamma})}{\Phi(X_i'\hat{\gamma})}$$
$$Y_i = X_i\beta + \rho\sigma\frac{\phi(X_i'\hat{\gamma})}{\Phi(X_i'\hat{\gamma})} + U_i^*$$

The latter regression has perfect collinearity. This means the standard errors for our estimators of β and $\rho\sigma$ will be inflated. To illustrate this difference, we first conduct the Heckman estimation where $X_i = Z_i$, with X_i consisting of the schooling, age, and age squared variables. We then conduct the Heckman estimation with the married variable added to Z_i. The table below reports the results of the second regression in the Heckman estimation:

Table 2: Sample Selection Model, with Heckman estimator

	Dependent variable:	
	logWage	
	Without exclusion restriction	With exlcusion restriction
schooling	0.303	0.215***
	(0.855)	(0.032)
age	1.423	−0.385
	(17.376)	(0.544)
age2	−0.039	−0.010
	(0.273)	(0.009)
invMillRatio	7.341	−0.174
	(72.233)	(0.617)
Constant	−6.436	27.209***
	(323.291)	(8.553)
Observations	416	416
Note:		*p<0.1; **p<0.05; ***p<0.01

E

When we use the ML estimation, we run into the same problem. When a regression equation has perfect collinearity, many iterative methods that maximize the log-likelihood function (such as Newton-Rapson) are unable to identify the variance of the estimator - leading to undefined standard errors in if we don't apply the exclusion restriction.

Table 3: Sample Selection Model, With MLE estimator

	Dependent variable:	
	logWage	
	Without exclusion restriction	With exclusion restriction
schooling	0.274	0.215***
	(Undefined.)	(0.032)
age	1.594	−0.379
	(Undefined.)	(0.538)
age2	−0.042	−0.011
	(Undefined.)	(0.009)
Constant	−6.423	27.091***
	(Undefined.)	(8.430)
Observations	666	666
Note:		*p<0.1; **p<0.05; ***p<0.01

F

We find the distribution of potential earnings for the non-employed in two steps. First, using the coefficients from the ML estimation of sample selection model (with the exclusion criterion), we predict the earnings of the unemployed. We use the ML estimation, since this is more efficient than the two-step Heckman estimator. Second, we use kernel density estimation (KDE) to estimate the probability density function of these earnings. The plot below shows the result of this method.

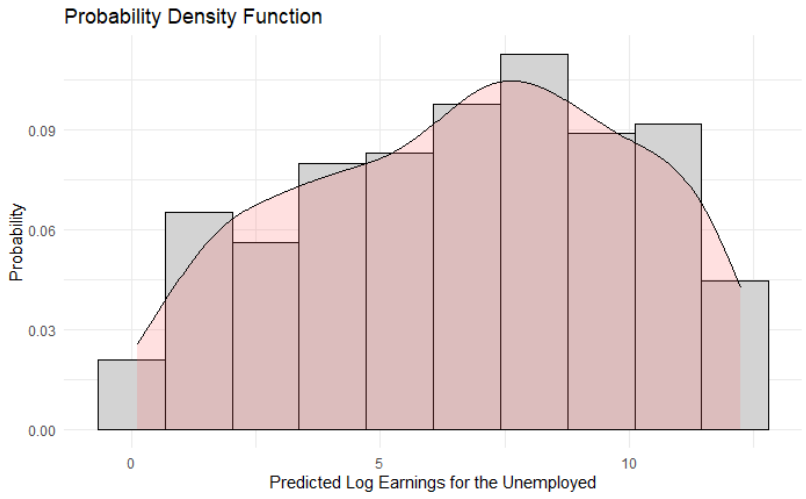

Figure 1:

The main downside to this method is that the model we specified for the employed might not be sufficient to capture certain dynamics that that are particularly important for the unemployed. For instance, the model does not capture how long someone has been out of a job (since those employed currently have a job), which might affect the earning potential of those currently unemployed if they were to have a job.

Exercise 2

A

A key condition for OLS being the best linear unbiased estimator (BLUE) is that the errors are uncorrelated with the independent variables - the so-called exogeneity condition, or more formally, $plim(\frac{1}{n}X_i'\epsilon_i) = 0$. There are several reasons why this might not hold - one is that there is another variable that influences both X_i and y_i. In the case of schooling (X_i) and earnings (y_i), there are several variables that affect both; for instance, racist sentiments in society both affect people's ability to obtain schooling and find a job, and socio-economic privilege of one's family likely make both schooling and the job search easier. Given this, it seems implausible that this condition of OLS is satisfied.

B

A good set of instruments should fulfill two criteria: first, it should be relevant, or more formally: $Cov(Z_i, X_i) \neq 0$. Second, it should be valid: $Cov(Z_i, U_i) = 0$. Logically speaking, we suspect the distance variable does not fulfill the second criterion. How close one lives to a school can often be a function of parental wealth (since they can use their wealth to afford more expensive inner city houses closer to schools), which also tends to affect future earnings. On the other hand, regional subsidy policy appears more valid - since not all poor families necessarily get subsidies (they might not live in regions that adopt such policies).

To check validity of the instruments, we look at the correlation of errors terms from OLS regression and two instruments: distance and regional subsidies. For a joint instrument we use the Sargan test.

To check relevance of the instruments, we use F-test and as a rule of thumb 10. To check whether our regressors are endogenous we conduct the Hausman test. The results are shown below:

Table 4:

	Dependent variable: logWage			
	OLS	*instrumental variable*		
	OLS	Subsidy	Distance	Subsidy and Distance
schooling	0.216***	0.401***	0.470	0.408***
	(0.032)	(0.106)	(0.299)	(0.102)
age	−0.342	−0.233	−0.192	−0.229
	(0.521)	(0.546)	(0.587)	(0.547)
age2	−0.011	−0.013	−0.014	−0.013
	(0.008)	(0.009)	(0.010)	(0.009)
Constant	26.409***	23.694***	22.681**	23.589***
	(8.057)	(8.517)	(9.704)	(8.530)
Observations	416	416	416	416
Adjusted R^2	0.813	0.798	0.784	0.797
Weak instrument	-	43.320***	5.374**	23.894***
Hausman	-	3.634**	0.845	4.343***
Sargan test	-	-	-	0.052
Note:				*p<0.1; **p<0.05; ***p<0.01

The variables distance and regional subsidy are almost not correlated with OLS error terms (0.045 and 0.089, respectively), i.e. they are valid instruments. Sargan test for a joint relevance of distance and regional subsidy as instruments show that both instruments are valid and the model specification is correct. F-test is bigger than 10 when we use subsidy or subsidy and distance together. Thus, both instruments are relevant.

C

If all the independent variables are exogenous, then the OLS estimator is the most efficient consistent estimator - the IV estimator will still be consistent, but according to the Gaus-Markov theorem less efficient than OLS. We use the Hausman test to test for the exogeneity of the independent variable (schooling). In the Hausman test with distance and subsidy and only subsidy as the instrument variables, the Hausman test rejects the null hypothesis (exogenous independent variables, OLS is consistent) at a 5% threshold. This means we should prefer the IV estimator with distance and subsidy or only subsidy as the instrumental variables. However, as a joint instruments doesn't add a lot to the model with only subsidy as an instrument, we should prefer the IV estimator with only subsidy.