Econometrics II - Assignment 2

Floris Holstege, Stanislav Avdeev

1

As a base model we use the following model specification:

$$Log(Earnings) = \alpha_0 + \alpha_1 Schooling_{it} + \alpha_2 AGE_{it} + \alpha_3 AGE^2_{it} + \alpha_4 ETHNICITY_i + \alpha_5 URBAN_{it} + \alpha_6 REGNE_{it} + \alpha_7 REGNCit + \alpha_8 REGW_{it}$$

In order to check the impact of ability, we include a variable $ASVABC$ in the base model. Including an ability allows to account for an omitted variavble bias that most likely occurs in the base model without it. To account for hetero , we use robust standard errors.

The results of the base model with and without ability variable are as follows. When we do not account for ability in the model specification, returns to one year of education are higher, i.e. returns to a year of education are 7% and statistically significant at the 1% level. When we inlude the ability variable, returns to one year of education become 4.8% and remain significant at the 1% level. The underlying reason for such a drop is that higher ability students tend to get more education, thus, they tend to get higher earnings.

Table 1: OLS pooled model with and without ability variable

	Dependent variable:	
	EARNINGS	
	(1)	(2)
Schooling	0.070***	0.048***
	(−0.00004)	(−0.00004)
Test_score		0.011***
		(−0.00004)
Ethnicity	−0.192***	−0.096***
	(−0.00004)	(−0.0002)
Constant	−0.079***	−0.386***
	(0.004)	(0.004)
Observations	40,043	40,043
Adjusted R^2	0.292	0.313
Note:	*p<0.1; **p<0.05; ***p<0.01	

From now on we are going to include the ability variable in all model specifications due to two reasons. First, we deal with an omitted variable bias problem (theoretical problem). Second, including this variable improves our model: adjusted R squared is slightlty higher. Moreover, for the sake of saving the space we don't report all coefficients in tables unless they are important for a specific question. Besides, in our model robust s.e. are smaller than conventional s.e. The most probable cause for that is that residual variance goes up with the value of x (for example, earnings are more variable for those with more schooling) (Angrist and Pischke, 2009).

2

To measure an amount of discrimination on the labour market, we estimate three OLS pooled models: including a cross effect of schooling and ethnicity and two models separated by ethnicity. As we can see, there is a statistically significant difference between returns to education by ethnicity. Interaction term between years of schooling and ethnicity yields a statistically significant effect of 1.6% at the 1% level. Having estimated a model separately for black and other give returns to a year of education of 6.1% and 4.6% respectively, which are significant at the 1% level. Based on these results we can conclude that there is not a discrimination against black people.

To estimate heterogenous returns to schooling in ipcoming models, we would use a model with an interaction term due to two reasons. First, it contains all observations in one dataset. Second, as we are not intrested in heterogenous effects between regressors other than schooling, a model with an interaction term is preferred to separated models.

Table 2: OLS pooled model with heterogeneous effects by ethnicity

	Dependent variable:		
	EARNINGS		
	(1)	(2)	(3)
Schooling	0.046***	0.046***	0.061***
	(−0.00001)	(−0.00000)	(−0.0004)
Test_score	0.011***	0.010***	0.014***
	(−0.00002)	(−0.00002)	(−0.0001)
Ethnicity	−0.295***		
	(−0.001)		
Schooling:Ethnicity	0.016***		
	(0.00004)		
Constant	−0.370***	−0.437***	0.038
	(0.004)	(0.005)	(0.036)
Observations	40,043	35,223	4,820
Adjusted R^2	0.313	0.299	0.314
Note:	*p<0.1; **p<0.05; ***p<0.01		

3

To exploit the panel data structure of our data, we use a panel model with random effects. Random effects model assumes that $E[\eta_i|X_{i1},...,X_{iT}] = 0$, i.e. individual effects are not correlated. As it can be seen in the overall table, there seems to be no difference between returns to education by ethnicity, i.e. an interaction term between scooling and ethnicity is not statistically significant. The results of the OLS pooled model differ from random fixed effects model. In the pooled OLS model indivudually specific effects are not being taken into account, thus, orthogonality assumption of the error term is violated. Therefore, one has to rely on the reults of panel model estimators.

4

A priori, a Fixed-effects model seems to make more sense, as it is highly likely that unobserved individual effects would be correleted with our regressors, i.e. $E[\eta_i | X_{i1}, ..., X_{iT}] \neq 0$. For example, one unobserved individual effect could be the motivation of a worker. The schooling and age of worker are likely to have some impact on their motivation. In this case, there is endogeneity between the regressors (schooling, age) and individual effects,thus, fixed-effects model is the preferred choice and gives us consistent estimates.

5

6

To decide between fixed or random effects we run a Hausman test. Under the H_0, $E[\eta_i | X_{i1}, ..., X_{iT}] = 0$, and the random effects and fixed effects model are both consistent. In this case, the random effects model is preferred, since its more efficient than the fixed effects model. However, if H_A is true, and $E[\eta_i | X_{i1}, ..., X_{iT}] \neq 0$, then the fixed effects model is preffered, since its the only consistent one of the two. Performing the Hausman test on our random effect and fixed effect model, we reject the H_0 that both models are consistent, and thus the fixed effects model is preferred.

7

8

9