Econometrics II - Assignment 2

Floris Holstege, Stanislav Avdeev

1

As a base model we use the following model specification:

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

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 preferred, since its the only consistent one of the two.

However, a crucial assumption we make when applying the Hausman test, in this way is that at least the fixed effects model is consistent. If both models are inconsistent, the Hausman test will indicate that there is no difference between the two models, because the variance of β is similar for both - but we cannot then say that we prefer the random effects over the fixed effects model - both are inconsistent.

In this specific case, there are good reasons to believe that the fixed effects model is also inconsistent. This is because the assumption of strict exogeneity, $E[U_{it}|X_{i1},...,X_{iT},\eta_i] = 0$, is likely to be violated. For instance, it is likely that there are certain unobserved variables, such as general socio-economic privilege and networks (social capital) affect both an individuals earnings, and their ability to obtain schooling. Thus, we cannot trust the Hausman test in this case.

7

In order to test whether or not to use random or fixed effects, we use the Mundlak estimation. We define this as follows:

$$\eta_i = \bar{X}_i\gamma + \omega_i$$
$$Y_{it} = X_{it}'\beta + \bar{X}_i\gamma + \omega_i + U_i$$

In this case, $\bar{X}_i$ is the average of each regressor, per worker, over time, and ω_i is a random element, assumed to be uncorrelated with X_{it}. Using the Mundlak estimation, we can subsequently test if $\gamma = 0$ (H_0), or if $\gamma \neq 0$. If we fail to reject H_0, then random effects is appropriate, since the average of the regressors is unrelated to the dependent variable. The table below shows the results of this last regression.

In this table, we do not show the “average test score”, since this is already a constant. We test the joint probability of $\gamma = 0$ with a wald test. It is statistically significant, and thus we reject H_0, indicating that fixed effects is more appropriate for this case.

	Model 1
(Intercept)	-1.23^{***}
	(0.22)
Schooling	0.05^{***}
	(0.00)
AGE	0.08^{***}
	(0.00)
AGESQ	-0.00^{***}
	(0.00)
Ethnicity	-0.09^{***}
	(0.02)
URBAN	0.03^{***}
	(0.01)
REGNE	0.05^{***}
	(0.01)
REGNC	-0.03^{*}
	(0.01)
REGW	0.08^{***}
	(0.01)
Test_score	0.01^{***}
	(0.00)
Schooling_AVG	0.00
	(0.00)
AGE_AVG	0.07^{***}
	(0.02)
AGESQ_AVG	-0.00^{***}
	(0.00)
URBAN_AVG	0.10^{***}
	(0.02)
REGNE_AVG	0.06^{**}
	(0.02)
REGNC_AVG	0.04^{*}
	(0.02)
REGW_AVG	-0.03
	(0.02)
R^2	0.31
Num. obs.	40043

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 3: Statistical models

8

9

The Verbeek and Nijman test is simply a Hausman test between a model with an unbalanced dataset, and a model with a balanced dataset. This exercise asks us to apply it to a model with a dataset where each individual is in at least 5 panels. It seems that using this last model is inappropriate to test if there is attrition bias in the full, unbalanced dataset, since this latter model could still portray attrition bias: some individuals might have dropped out (e.g. less than 5 panels) for a variety of reasons (unmotivated, etc.) that might influence the results. We therefore also conduct a Verbeek and Nijman test where we comare the unbalanced dataset to a balanced dataset, with only workers that appear in all models. In both cases, the H_0 of the Verbeek and Nijman test (that there is no attrition bias) is rejected.