Econometrics II: Assignment 2

Walter Verwer & Bas Machielsen

January 14, 2021

Question 1

First use pooled OLS to check the impact of including and excluding ASVABC on the estimate of α_1. Present and explain the result.

	Dependent variable:	
	(1)	(2)
ASVABC	0.011***	
	(0.000)	
AGE	0.078***	0.074***
	(0.004)	(0.004)
AGESQ	-0.001***	-0.001***
	(0.000)	(0.000)
S	0.048***	0.070***
	(0.001)	(0.001)
ETHBLACK	-0.096***	-0.192***
	(0.007)	(0.007)
URBAN	0.101***	0.106***
	(0.005)	(0.005)
REGNE	0.004	0.075***
	(0.011)	(0.011)
REGNC	-0.103***	-0.036***
	(0.011)	(0.011)
REGW	-0.048***	0.018
	(0.011)	(0.011)
REGS	-0.120***	-0.068***
	(0.011)	(0.011)
Observations	40,043	40,043
R^2	0.313	0.292
Adjusted R^2	0.313	0.292
Residual Std. Error	0.416(df = 40033)	0.423(df = 40034)
F Statistic	2023.536*** (df = 9.0; 40033.0)	2063.759*** (df = 8.0; 40034.0)
Note:		*p<0.1; **p<0.05; ***p<0.01

The inclusion of the proxy for ability decreases the estimate for the coefficient of schooling. Hence, given all other standard assumptions, ability and schooling are positively correlated, and the omission of a proxy for ability overestimates the impact of schooling. Something else that we are able to observe is that because of the inclusion of the test scores, the adjusted-R^2 increases by approximately 0.2 points. This implies that including test scores add about 2% to the model's ability to explain the observed variation. Finally, we

could do a very rough t-test to compare both estimates of schooling. We observe a standard-error of about 0.001, and we observe a change in the coefficients of roughly 0.022. This implies that the t-value of the corresponding t-test is roughly 22. What this implies is that the observed difference in the estimate of the coefficient is significantly different.

Question 2

Perform a pooled OLS analysis to obtain insight in the heterogeneity of returns to schooling by ethnicity. Present the results and comment on the outcomes: what are the conclusions based on this?

	Dependent variable:		
	Interaction	Not Black	Black
	(1)	(2)	(3)
BLACKxS	0.016***		
	(0.003)		
ASVABC	0.011***	0.010***	0.014***
	(0.000)	(0.000)	(0.001)
AGE	0.079***	0.084***	0.035***
	(0.004)	(0.004)	(0.010)
AGESQ	-0.001***	-0.001***	-0.000*
	(0.000)	(0.000)	(0.000)
S	0.046***	0.046***	0.061***
	(0.001)	(0.001)	(0.003)
ETHBLACK	-0.295***	0.000***	0.096
	(0.040)	(0.000)	(0.112)
URBAN	0.102***	0.111***	0.006
	(0.005)	(0.005)	(0.017)
REGNE	0.007	-0.007	0.090***
	(0.011)	(0.012)	(0.031)
REGNC	-0.100***	-0.114***	-0.009
	(0.011)	(0.011)	(0.031)
REGW	-0.045***	-0.061***	0.073**
	(0.011)	(0.012)	(0.034)
REGS	-0.116***	-0.127***	-0.058**
	(0.011)	(0.011)	(0.029)
Observations	40,043	35,223	4,820
R^2	0.313	0.300	0.315
Adjusted R^2	0.313	0.299	0.314
Note:			*p<0.1; **p<0.05; ***p<0.01

For the pooled models, we cluster the standard errors on the individual level, allowing for correlation in the error-term between observations belonging to the same individual. We can see that the interaction effect is significant: that is to say, there is a significant difference between blacks and non-black in the influence of schooling on earnings. When we split up the sample into blacks and non-black, we get a similar view: the point estimate for the effect of schooling seems to be slightly lower for black people than for non-black people. As seen in the pooled regression with interaction effect, the differential impact is statistically significant. This follows from the fact that the change is roughly 0.015 and if we take the highest standard error for the two (0.003), we would obtain a t-statistic of about 5. Which means that the observed difference is very likely to be because of ethnicity. Interestingly, the interaction approach results a very similar estimate and standard error as the difference between the two models. This implies that both methods give similar results. To conclude, we observe that there is a racial difference in the influence of schooling on earnings for blacks and non-blacks. It appears that blacks benefit more from schooling than whites, in terms of earnings.

Question 3

Perform the analysis for heterogenous schooling effects using the random effects model. Present the results and compare the outcomes with the pooled OLS results obtained before. Interpret the outcomes.

Table 1: Random effects model

	Dependent variable:
	EARNINGS
ASVABC	0.011^{***}
	(0.001)
AGE	0.078^{***}
	(0.003)
AGESQ	-0.001^{***}
	(0.00004)
S	0.051^{***}
	(0.002)
ETHBLACK	-0.029
	(0.077)
URBAN	0.044^{***}
	(0.005)
REGNE	-0.358^{***}
	(0.049)
REGNC	-0.459^{***}
	(0.048)
REGW	-0.370^{***}
	(0.048)
REGS	-0.452^{***}
	(0.048)
BLACKxS	-0.004
	(0.006)
Observations	40,043
R^2	0.366
Adjusted R^2	0.366
F Statistic	$252{,}174.700^{***}$
Note:	*p<0.1; **p<0.05; ***p<0.01

The random effects model assumes that $\mathbb{E}[\eta_i|X_1, \ldots, X_n] = 0$, in words, that the individual-specific effects are uncorrelated to the predictor variables. In this model, the point estimate for schooling is now close to the point estimate for schooling in the pooled OLS regression including the proxy for ability. Hence, the random effects estimator looks a lot like the pooled estimator, indicating that the contribution from the

within group estimator is marginal. This can also be observed when looking at the decomposition of the explained variance: the between R-squared is larger than the within R-squared, indicating the model does a better job explaining the changes between individuals rather than individuals over time. There seems to be no differences in returns to education between individuals of different ethnicity: the interaction coefficient is insignificant.

Question 4

A priori, would you plead for using fixed effects estimation or random effects estimation? Explain your answer.

A priori, it would make more sense to use fixed-effects rather than random effects, because it is very likely that the unobservable individual components η_i are correlated to the predictor variables X rather than being random. For example, η_i can be interpreted as being some measure of ability or innate willingness to exert effort, and that is likely related to age, schooling and test score. A possible correlation would violate the randomness of η_i required by random effects, and hence, fixed effects would be preferred.

Question 5

Apply the fixed effects estimator to analyze the heterogenous schooling effect. Interpret the outcomes.

Table 2:

	Dependent variable:
	EARNINGS
AGE	0.078^{***}
	(0.003)
AGESQ	-0.001^{***}
	(0.00004)
S	0.053^{***}
	(0.004)
URBAN	0.028^{***}
	(0.006)
REGNE	0.051^{***}
	(0.015)
REGNC	-0.026^{*}
	(0.013)
REGW	0.089^{***}
	(0.015)
BLACKxS	-0.062^{***}
	(0.012)
Observations	40,043
R^2	0.253
Adjusted R^2	0.152
F Statistic	$1{,}494.723^{***}$ (df = 8; 35270)
Note:	*p<0.1; **p<0.05; ***p<0.01

Because of multicollinearity, one coefficient from the dummies is dropped.

Question 6

Fixed effects estimation may not be as effcient as random effcient estimation, but is robust to correlation between regressors and the random effcient. Can we perform a Hausman test in this context? Perform the test you propose. Fixed effects estimation may not be as efficient as random effects estimation, but is robust to correlation between regressors and the random effect. Can we perform a Hausman test in this context? Perform the test you propose.

The test tests the null hypothesis that the unique errors are not correlated with the regressors.

```
phtest(fixed_effects, random_effects)
```

```
##
##  Hausman Test
##
## data:  formula
## chisq = 192.51, df = 8, p-value < 2.2e-16
## alternative hypothesis: one model is inconsistent
```

The null hypothesis is rejected, implying that the unique parts are correlated with the regressors, and hence, random effects is an inconsistent estimator.

Question 7

Perform Mundlak estimation of the model. Present the results of estimation and test for the joint sigificance of the within-group means.

For this question we first need to estimate the time mean of the regressor variables, for every individual. Then we include those estimates in a random effects model and apply a wald test on the coefficients of the time meaned regressors.

Our results are shown below. In the wald test output, one can see that the time meaned regressors are highly jointly significant away from zero. This indicates that it is better to use the fixed effects model instead of the random effects model.

```
# 1. Estimate time means per individual and variable:
    # We moeten mergen, de mundlak part moet constant erbij toegevoegd worden, for elke t.
    # Dus voor elke individu neem de mean over de tijd en maak de dimensies gelijk aan Y_it

# 2. Use pggls() to estimate the feasible GLS of the model, use method = random:
    # Idem

# 3. Apply Wald test

# wald.test(vcov(ppgls(model)), b=coeffs(ppgls(models)), Terms = 10:17, df = earnings)

library(aod)

# get average over time per worker
earnings <- py$earnings

X_hat <- earnings %>%
  group_by(ID) %>%
  summarise(across(everything(), mean))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
colnames(X_hat)[c(-1)] <- paste(colnames(X_hat)[c(-1)], "MEAN", sep = "_")

# add to individual variables
earnings_with_mean <- merge(py$earnings, X_hat, by = "ID")

mundlak <- pggls(EARNINGS ~  S + AGE + AGESQ + ETHBLACK + URBAN +
                   REGNE + REGNC + REGW  + ASVABC + BLACKxS +
                   S_MEAN + AGE_MEAN + AGESQ_MEAN + ETHBLACK_MEAN +
                   URBAN_MEAN + REGNE_MEAN + REGNC_MEAN +
                   REGW_MEAN  + ASVABC_MEAN + BLACKxS_MEAN,
                 data=earnings_with_mean,
                 model="random",
                 index = c("ID")
                 )
```

```
## Warning: for argument 'model' to pggls(), the value 'random' has been renamed as 'pooling'
```

```
summary(mundlak)
```

```
## Oneway (individual) effect General FGLS model
##
## Call:
## pggls(formula = EARNINGS ~ S + AGE + AGESQ + ETHBLACK + URBAN +
##     REGNE + REGNC + REGW + ASVABC + BLACKxS + S_MEAN + AGE_MEAN +
##     AGESQ_MEAN + ETHBLACK_MEAN + URBAN_MEAN + REGNE_MEAN + REGNC_MEAN +
##     REGW_MEAN + ASVABC_MEAN + BLACKxS_MEAN, data = earnings_with_mean,
##     model = "random", index = c("ID"))
##
## Unbalanced Panel: n = 4765, T = 1-18, N = 40043
##
## Residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -1.74792 -0.23017  0.03351  0.03576  0.29999  3.05220
##
## Coefficients:
##                 Estimate  Std. Error  z-value  Pr(>|z|)
## (Intercept)  -1.2510e+00  2.1951e-01  -5.6988 1.206e-08 ***
## S             5.3364e-02  3.5788e-03  14.9113 < 2.2e-16 ***
## AGE           8.0426e-02  2.5753e-03  31.2294 < 2.2e-16 ***
## AGESQ        -8.7286e-04  4.3586e-05 -20.0259 < 2.2e-16 ***
## ETHBLACK     -3.4345e-01  8.8672e-02  -3.8732 0.0001074 ***
## URBAN         2.5952e-02  5.8146e-03   4.4632 8.075e-06 ***
## REGNE         5.0441e-02  1.5064e-02   3.3486 0.0008124 ***
## REGNC        -3.3550e-02  1.3140e-02  -2.5533 0.0106697 *
## REGW          8.5575e-02  1.5112e-02   5.6627 1.490e-08 ***
## ASVABC        1.1163e-02  7.2404e-04  15.4176 < 2.2e-16 ***
## BLACKxS      -6.3085e-02  1.1558e-02  -5.4581 4.812e-08 ***
## S_MEAN       -5.0132e-03  4.5625e-03  -1.0988 0.2718655
## AGE_MEAN      6.6373e-02  1.6363e-02   4.0564 4.984e-05 ***
## AGESQ_MEAN   -1.3773e-03  2.8409e-04  -4.8482 1.246e-06 ***
## URBAN_MEAN    1.0664e-01  1.5239e-02   6.9978 2.599e-12 ***
```

```
## REGNE_MEAN    6.4214e-02  2.0909e-02   3.0712 0.0021324 **
## REGNC_MEAN    3.8575e-02  1.8123e-02   2.1285 0.0332961 *
## REGW_MEAN    -3.1238e-02  2.1732e-02  -1.4374 0.1506011
## BLACKxS_MEAN  8.2673e-02  1.3422e-02   6.1597 7.287e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Total Sum of Squares: 10099
## Residual Sum of Squares: 6930.9
## Multiple R-squared: 0.3137
```

```
stargazer(mundlak, header=FALSE)
```

```
##
## % Error: Unrecognized object type.
```

```
wald.test(vcov(mundlak), b=coef(mundlak), Terms = 10:17)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 480.8, df = 8, P(> X2) = 0.0
```

Question 8

What are your overall conclusions from the analysis of heterogeneity in returns to schooling by ethnicity?

Question 9

To gain insight in the impact of nonresponse and attrition, the researcher applies a variant of the Verbeek and Nijman-test (see lecture slides). He defines the dummy variable d_i which is 1 if the individual is in the panel for more than 5 waves, and is zero otherwise. Apply the Verbeek and Nijman test with this definition of d_i (otherwise equal to the definition at the lecture slides). Draw conclusions and address practical problems you possibly met in implementing the test.

We have simply applied a fixed effects panel data regression on both the full data set, as well as on the data set where we only take the individuals into account that are 5 or more times observed in the data set. We have done this by applying a filter by counting how many times a specific individual is in the data set, if the individual is counted 5 or more times, then it is included in the fixed effects regression. Our conclusion based on a Haussman test between the two estimated coefficient vectors for both data sets is that there is a significant difference. This tells us that there is attrition bias present in the data. However, we do have to note that this specification of the Verbeek and Nijman test does not compare a fully balanced model with our original unbalanced model. It actually compares two unbalanced models with each other. Even though this is the case we still obtain evidence of attrition bias due to an unobserved variable that is causing the attrition. The reason is, as stated before, that the two estimated coefficient vectors are different from each other. A practical problem that can arise (we did not have this problem) is that the sample size of the fully balanced model becomes simply too small to infere meaningful results. By taking out individuals that left before being in there for the 5^{th}, this prevents small sample size problems. A practical problem we did have was that we had to programme this filter ourselves, because it is a deviation from the standard

Verbeek and Nijman methodology. But still this was a minor challenge to overcome, given the ease that this is programmed in R.

```
earnings <- py$earnings %>%
  mutate(BLACKxS = ETHBLACK * S)

frequencies <- earnings %>%
  count(ID)

earnings <- dplyr::left_join(earnings, frequencies, by = "ID")

# Unbalanced model:
unbalanced_formula <- paste0(py$y, " ~ ", paste(py$ivs2, collapse = " + "))

unbalanced_fixed_effects <- plm(formula = unbalanced_formula,
                      data = earnings,
                      index = c("ID","TIME"),
                      model = "within")

# Balanced model:
partial_unbalanced_formula <- paste0(py$y, " ~ ", paste(py$ivs2, collapse = " + "))

partial_unbalanced_fixed_effects <- plm(formula = partial_unbalanced_formula,
                      data = earnings %>%
                        filter(n >= 5),
                      index = c("ID","TIME"),
                      model = "within")

summary(unbalanced_fixed_effects)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = unbalanced_formula, data = earnings, model = "within",
##     index = c("ID", "TIME"))
##
## Unbalanced Panel: n = 4765, T = 1-18, N = 40043
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -2.3043412 -0.1281976  0.0065233  0.1341122  2.7466990
##
## Coefficients: (1 dropped because of singularities)
##            Estimate  Std. Error  t-value  Pr(>|t|)
## AGE      7.8070e-02  2.6052e-03  29.9668 < 2.2e-16 ***
## AGESQ   -8.2876e-04  4.4144e-05 -18.7739 < 2.2e-16 ***
## S        5.2799e-02  3.5268e-03  14.9708 < 2.2e-16 ***
## URBAN    2.8161e-02  5.8654e-03   4.8013 1.583e-06 ***
## REGNE    5.1121e-02  1.4972e-02   3.4145 0.0006396 ***
## REGNC   -2.5614e-02  1.3072e-02  -1.9594 0.0500715 .
## REGW     8.8853e-02  1.4953e-02   5.9422 2.839e-09 ***
## BLACKxS -6.1817e-02  1.1525e-02  -5.3639 8.195e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Total Sum of Squares:    3662.8
## Residual Sum of Squares: 2735.4
## R-Squared:      0.25319
## Adj. R-Squared: 0.15215
## F-statistic: 1494.72 on 8 and 35270 DF, p-value: < 2.22e-16
```

```
summary(partial_unbalanced_fixed_effects)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = partial_unbalanced_formula, data = earnings %>%
##     filter(n >= 5), model = "within", index = c("ID", "TIME"))
##
## Unbalanced Panel: n = 3686, T = 5-18, N = 37219
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -2.3063166 -0.1295646  0.0084185  0.1356819  2.7463007
##
## Coefficients: (1 dropped because of singularities)
##            Estimate  Std. Error  t-value  Pr(>|t|)
## AGE      8.0323e-02  2.6330e-03  30.5065 < 2.2e-16 ***
## AGESQ   -8.6305e-04  4.4599e-05 -19.3513 < 2.2e-16 ***
## S        5.0630e-02  3.6873e-03  13.7308 < 2.2e-16 ***
## URBAN    2.8787e-02  5.9600e-03   4.8300 1.371e-06 ***
## REGNE    5.4910e-02  1.5387e-02   3.5686 0.0003594 ***
## REGNC   -3.9541e-02  1.3497e-02  -2.9296 0.0033959 **
## REGW     7.9656e-02  1.5460e-02   5.1523 2.588e-07 ***
## BLACKxS -6.2404e-02  1.2048e-02  -5.1797 2.235e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    3449
## Residual Sum of Squares: 2553.3
## R-Squared:      0.2597
## Adj. R-Squared: 0.17815
## F-statistic: 1470.1 on 8 and 33525 DF, p-value: < 2.22e-16
```

```
phtest(unbalanced_fixed_effects, partial_unbalanced_fixed_effects)
```

```
##
##  Hausman Test
##
## data:  unbalanced_formula
## chisq = 79.03, df = 8, p-value = 7.663e-14
## alternative hypothesis: one model is inconsistent
```