Econometrics II - Assignment 2

Uncensored sloths

13 Jan 2022

Question 1

(i) Why does the process of taking each observation relative to its individual-level mean have the effect of “controlling for individual effects”?\

The process of taking each observation relative to its individual-level mean is called demeaning. When apply demeaning, the individual fixed effects are dropping out from the model as a results. Hence, we control for the individual effects but we also are not able to estimate the impact of time-invariant regressors.

(ii) Two-way fixed effects with terms for both individual and time are often referred to as “controlling for individual and time effects”. Why might a researcher want to do this rather than just taking individual fixed effects and adding a linear/polynomial/etc. term for time?\

Firstly, it is easier to apply as you do not have to specify any linear or polynomial term (it is quite tricky as you need to have an idea what it the form of the impact). It especially is convenient to apply when the research only wants to control for time effects but time effects themselves are not part of the research question.

(iii) Why random effects is likely to do a better job of estimating the individual effects than fixed effects, if its assumptions hold?\

Because it is able to estimate timeinvariant regressors, we can also do out sample predictions, it is more efficient, etc. Look at the slides.

Question 2

```
# Load data
data <- read.csv("assignment2.csv")
lnearnings <- ln(data$earnings)
data <- cbind(data, lnearnings)
```

(i) First use pooled OLS to check the impact of including and excluding asvabc on the estimate of β_1 Present and explain the result.\

```
pooled1 <- plm(lnearnings ~ school + age + agesq + ethblack + urban + regne + regnc + regw, model = "poo
pooled2 <- plm(lnearnings ~ school + age + agesq + ethblack + urban + regne + regnc + regw + asvabc, mod
stargazer(pooled1, pooled2)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: So, Jan 16, 2022 - 13:17:57

The index test score has a positive significant effect at 1% level on ln earnings. Estimators for schooling is lower in the second model which makes sense if we assume that individuals that attended school for longer time are able to achieve higher results. Being black has also a negative significant effect. However, the magnitude is lower if the index test score is included.

One regional dummy was excluded. As all regional dummies are positive and significant, we can conclude that the dummy that was excluded, so south, indicates that individuals in the south earn less relative to individuals living in other areas.

The adjusted R^2 is higher for the model including the index test score.

What about age squared?

(ii) 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?\

```
pooled <- plm(lnearnings ~ school + age + agesq + ethblack + I(ethblack * school) + urban + regne + reg
stargazer(pooled)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: So, Jan 16, 2022 - 13:17:58

The interaction effect between schooling and being black is significant at 1% and positive. So attending school has a higher impact on black people. The other estimators do not change a lot in magnitude except for the estimator black ethnicity. This estimator decreases by X (34%), indicating that being black has a negative impact on earnings. Therefore, the data provide evidence that there is discrimination on the labour market. R^2 does not increase, however, we get a better picture of the dynamics on the labour market and heterogeneity of the effects.

(iii) Perform the analysis for heterogeneous schooling effects using the random effects model. Present the results and compare the outcomes with the pooled OLS results obtained before. Interpret the outcomes.\

```
random <- plm(lnearnings ~ school + age + agesq + ethblack + I(ethblack  * school) + urban + regne + reg
random$vcov <- vcovHC(random, cluster="group")
stargazer(pooled, random)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: So, Jan 16, 2022 - 13:18:11

We assume two-way. Time: f.e. exogeneous macroeconomic impacts that lower earnings. Individual: character traits, unobserved educational aspects, economic background, social background, etc.

Table 1:

	Dependent variable:	
	lnearnings	
	(1)	(2)
school	0.070^{***}	0.048^{***}
	(0.001)	(0.001)
age	0.074^{***}	0.078^{***}
	(0.004)	(0.004)
agesq	-0.001^{***}	-0.001^{***}
	(0.0001)	(0.0001)
ethblack	-0.192^{***}	-0.096^{***}
	(0.007)	(0.007)
urban	0.106^{***}	0.101^{***}
	(0.005)	(0.005)
regne	0.143^{***}	0.123^{***}
	(0.006)	(0.006)
regnc	0.031^{***}	0.017^{***}
	(0.005)	(0.005)
regw	0.085^{***}	0.072^{***}
	(0.007)	(0.007)
asvabc		0.011^{***}
		(0.0003)
Constant	−0.079	-0.386^{***}
	(0.051)	(0.051)
Observations	40,043	40,043
R^2	0.292	0.313
Adjusted R^2	0.292	0.313
F Statistic	$2,063.759^{***}$ (df = 8; 40034)	$2,023.536^{***}$ (df = 9; 40033)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 2:

	Dependent variable:
	lnearnings
school	0.046^{***}
	(0.001)
age	0.079^{***}
	(0.004)
agesq	-0.001^{***}
	(0.0001)
ethblack	-0.295^{***}
	(0.040)
I(ethblack $*$school)	0.016^{***}
	(0.003)
urban	0.102^{***}
	(0.005)
regne	0.124^{***}
	(0.006)
regnc	0.017^{***}
	(0.005)
regw	0.072^{***}
	(0.007)
asvabc	0.011^{***}
	(0.0003)
Constant	-0.370^{***}
	(0.051)
Observations	40,043
R^2	0.313
Adjusted R^2	0.313
F Statistic	$1{,}824.813^{***}$ (df = 10; 40032)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 3:

	Dependent variable:	
	lnearnings	
	(1)	(2)
school	0.046***	0.050***
	(0.001)	(0.003)
age	0.079***	0.096***
	(0.004)	(0.004)
agesq	−0.001***	−0.001***
	(0.0001)	(0.0001)
ethblack	−0.295***	−0.037
	(0.040)	(0.096)
I(ethblack ∗school)	0.016***	−0.004
	(0.003)	(0.008)
urban	0.102***	0.047***
	(0.005)	(0.011)
regne	0.124***	0.093***
	(0.006)	(0.015)
regnc	0.017***	−0.009
	(0.005)	(0.013)
regw	0.072***	0.081***
	(0.007)	(0.016)
asvabc	0.011***	0.012***
	(0.0003)	(0.001)
Constant	−0.370***	−0.665***
	(0.051)	(0.066)
Observations	40,043	40,043
R^2	0.313	0.309
Adjusted R^2	0.313	0.309
F Statistic	1,824.813*** (df = 10; 40032)	4,337.823***
Note:	*p<0.1; **p<0.05; ***p<0.01	

The interaction between ethnicity and schooling is not significant anymore, as well as one of the regions. The magnitude of age increased. Magnitude of urban decreased. So all in all, we can see that the magnitudes of the estimators are different from OLS. R^2 does not have any interpretative power as the Gauss-Markov assumptions do not hold (nochmal nachschauen)

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

Note that disadvanteges of fixed effext estimation are the following:\ ⇒ time-invariant regressors are not identified ⇒ out of sample prediction is impossible consequently (time-invariant regressors do not add any information) ⇒ need sufficient variation in X

On ther hand, it is robust to correlation between the omitted heterogeneity and the regressors.

Note that the random-effects model can be used to make predictions outside the sample and time-invariant regressors are informative on this. However, it assumes strongers assumptions than fixed effects. Also, it is more efficient than fixed effects if the stochastic structure is correct.

The question is whether η_i and λ_t are uncorrelated to the regressors. Considering that η_i could include dimensions as character traits or social and economic backgrounds (of parents), we have to assume that this condition does not hold as these dimension have an impact on schooling f.e. Therefore, we would plead for using a fixed effects model despite the disadvantage that it does not identify time-invariant regressors. \

v) Apply the fixed effects estimator to analyze the heterogenous schooling effects. Interpret the outcomes.\

```
fixed  <- plm(lnearnings ~ school + age + agesq + ethblack + I(ethblack  * school) + urban + regne + reg
fixed$vcov <- vcovHC(fixed, cluster="group")
stargazer(pooled, random, fixed)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: So, Jan 16, 2022 - 13:18:25

Interaction is significant and negative which differs from our OLS estimation. According to this estimation being black and in school has a lower impact on earnings than for individuals who are not black. Age is not identified in plm which could be cause by assuming time and individual fixed effects as this is not when we only assume individual effect. When we do it with feols, we cannot include both variables! This might be due to the time-invariance of age square - however, this would not explain why this is not an issue with the dummies where we also have a low time-invariance.

the way twoways is calculated, age becomes zero so it drops out. However, we still would proceed with twoways as macroeconomic dynamics can have significant effects on earnings. COnsidering that the time frame includes f.e. the oil crisis, we should not ignore it. Further, age square partially incorporates the impact of age on earnings. \

(vi) 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.\

We can perform a hausman test between our fixed and random model with robust standard errors. Note that we included the robust standard errors in the list. It is important to consider the robust standard errors as the test is based on the estimated coeffcient variances. P value is below 1%. Hence, we can reject our null hypothesis. Therefore, the random effects model is inconsistent as the assumption of no correlation between the fixed effects and the regressors does not hold.

```
phtest(fixed, random)
```

```
##
##  Hausman Test
##
## data:  lnearnings ~ school + age + agesq + ethblack + I(ethblack * school) +  ...
```

Table 4:

	Dependent variable:		
	lnearnings		
	(1)	(2)	(3)
school	0.046***	0.050***	0.051***
	(0.001)	(0.003)	(0.006)
age	0.079***	0.096***	
	(0.004)	(0.004)	
agesq	−0.001***	−0.001***	−0.001***
	(0.0001)	(0.0001)	(0.0001)
ethblack	−0.295***	−0.037	
	(0.040)	(0.096)	
I(ethblack ∗school)	0.016***	−0.004	−0.060***
	(0.003)	(0.008)	(0.016)
urban	0.102***	0.047***	0.032***
	(0.005)	(0.011)	(0.008)
regne	0.124***	0.093***	0.051**
	(0.006)	(0.015)	(0.026)
regnc	0.017***	−0.009	−0.029
	(0.005)	(0.013)	(0.021)
regw	0.072***	0.081***	0.088***
	(0.007)	(0.016)	(0.028)
asvabc	0.011***	0.012***	
	(0.0003)	(0.001)	
Constant	−0.370***	−0.665***	
	(0.051)	(0.066)	
Observations	40,043	40,043	40,043
R^2	0.313	0.309	0.020
Adjusted R^2	0.313	0.309	−0.113
F Statistic	1,824.813*** (df = 10; 40032)	4,337.823***	34.587*** (df = 7; 35254)
Note:			*p<0.1; **p<0.05; ***p<0.01

```
## chisq = 20.366, df = 7, p-value = 0.004832
## alternative hypothesis: one model is inconsistent
```

(vii) Perform Mundlak estimation of the model. Present the results of estimation and test for the joint significance of the within-group means.\

Wald Test for both significant - jointly significant, covariance between fixed effects and regressors. In line Hausman test. Address issues with plm (does not includes means) but regressors are in line with fixed effects model (partially). Pggls is shit - nothing works. Less observations, not consistent with the fixed effect model.But including mean dummies works. We get the error singluraties in plm, so a reasong might be that the rows are too similar and plm needs more variance. Maybe pggls is better with that.

```
data <- data %>%
  group_by(id) %>%
  mutate(mean_school = mean(school),
         mean_age = mean(age),
         mean_agesq = mean(agesq),
         mean_asvabc = mean(asvabc),
         mean_urban = mean(urban),
         mean_regne = mean(regne),
         mean_regnc = mean(regnc),
         mean_regw = mean(regw))
```

```
mundlak2 <- plm(lnearnings ~ school + age + agesq + ethblack + I(ethblack  * school) + urban + regne + r
```

```
## Warning in pdata.frame(data, index): column 'time' overwritten by time index
```

```
mundlak2$vcov <- vcovHC(mundlak2, cluster="group")

mundlak1 <- pggls(lnearnings ~ school + age + agesq + ethblack + I(ethblack  * school) + urban + regne +
```

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

## Warning: column 'time' overwritten by time index
```

```
fixed_oneway  <- plm(lnearnings ~ school + age + agesq + ethblack + I(ethblack  * school) + urban + regn
fixed_oneway$vcov <- vcovHC(fixed_oneway, cluster="group")

fixed_pggls <- pggls(lnearnings ~ school + age + agesq + ethblack + I(ethblack  * school) + urban + regn
```

```
## Warning in pdata.frame(data, index): column 'time' overwritten by time index
```

```
# pggls (http://tarohmaru.web.fc2.com/R/ExerciseDiagnostics.html)
extract.pggls <- function (model, include.rsquared = TRUE, include.adjrs = TRUE,
    include.nobs = TRUE, ...)
{
    s <- summary(model, ...)
    coefficient.names <- rownames(s$CoefTable)
    coefficients <- s$CoefTable[, 1]
    standard.errors <- s$CoefTable[, 2]
    significance <- s$CoefTable[, 4]
    rs <- s$rsqr
    n <- length(s$resid)
    gof <- numeric()
    gof.names <- character()
    gof.decimal <- logical()
    if (include.rsquared == TRUE) {
```

```r
        gof <- c(gof, rs)
        gof.names <- c(gof.names, "R$^2$")
        gof.decimal <- c(gof.decimal, TRUE)
    }
    if (include.nobs == TRUE) {
        gof <- c(gof, n)
        gof.names <- c(gof.names, "Num. obs.")
        gof.decimal <- c(gof.decimal, FALSE)
    }
    tr <- createTexreg(coef.names = coefficient.names, coef = coefficients,
        se = standard.errors, pvalues = significance, gof.names = gof.names,
        gof = gof, gof.decimal = gof.decimal)
    return(tr)
}

setMethod("extract", signature = className("pggls", "plm"),
          definition = extract.pggls)
```

```r
screenreg(list(mundlak2, fixed_oneway), digits=3, single.row=TRUE)
```

```
##
## ====================================================================
##                         Model 1                Model 2
## --------------------------------------------------------------------
## (Intercept)                -1.236 (0.215) ***
## school                      0.053 (0.006) ***      0.053 (0.006) ***
## age                         0.078 (0.004) ***      0.078 (0.004) ***
## agesq                      -0.001 (0.000) ***     -0.001 (0.000) ***
## ethblack                   -0.302 (0.086) ***
## ethblack * school          -0.062 (0.016) ***     -0.062 (0.016) ***
## urban                       0.044 (0.007) ***      0.028 (0.008) ***
## regne                       0.091 (0.015) ***      0.051 (0.026) *
## regnc                      -0.008 (0.012)         -0.026 (0.022)
## regw                        0.080 (0.016) ***      0.089 (0.028) **
## asvabc                      0.012 (0.001) ***
## mean_school                -0.002 (0.006)
## mean_age                    0.071 (0.016) ***
## mean_agesq                 -0.001 (0.000) ***
## ethblack * mean_school      0.079 (0.017) ***
## --------------------------------------------------------------------
## s_idios                     0.278
## s_id                        0.307
## R^2                         0.369                  0.253
## Adj. R^2                    0.368                  0.152
## Num. obs.               40043                  40043
## ====================================================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

```r
screenreg(list(mundlak1, fixed_pggls), digits=3, single.row=TRUE)
```

```
##
## ====================================================================
##                         Model 1                Model 2
## --------------------------------------------------------------------
## (Intercept)                -1.009 (0.211) ***
## school                      0.049 (0.004) ***      0.065 (0.003) ***
## age                         0.067 (0.003) ***      0.048 (0.002) ***
## agesq                      -0.001 (0.000) ***     -0.000 (0.000) ***
## ethblack                   -0.313 (0.086) ***
## ethblack * school          -0.045 (0.014) ***     -0.048 (0.009) ***
## urban                       0.013 (0.006) *       -0.005 (0.004)
## regne                       0.074 (0.016) ***      0.034 (0.012) **
## regnc                      -0.017 (0.014)         -0.078 (0.010) ***
## regw                        0.084 (0.016) ***      0.080 (0.012) ***
## asvabc                      0.011 (0.001) ***
## mean_school                -0.004 (0.005)
## mean_age                    0.065 (0.016) ***
## mean_agesq                 -0.001 (0.000) ***
## ethblack * mean_school      0.063 (0.015) ***
## mean_urban                  0.103 (0.015) ***
## mean_regne                  0.016 (0.021)
## mean_regnc                  0.017 (0.018)
## mean_regw                  -0.029 (0.022)
## --------------------------------------------------------------------
## R^2                         0.308                  0.727
## Num. obs.               39800                  40043
## ====================================================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

```
waldtest(mundlak1, c(9:16))
```

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

## Warning in pdata.frame(data, index): column 'time' overwritten by time index

## Wald test
##
## Model 1: lnearnings ~ school + age + agesq + ethblack + I(ethblack * school) +
##     urban + regne + regnc + regw + asvabc + mean_school + mean_age +
##     mean_agesq + I(ethblack * mean_school) + mean_urban + mean_regne +
##     mean_regnc + mean_regw
## Model 2: lnearnings ~ school + age + agesq + ethblack + I(ethblack * school) +
##     urban + regne + regnc + mean_regnc + mean_regw
##   Res.Df Df  Chisq Pr(>Chisq)
## 1  39781
## 2  39789 -8 446.93  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
waldtest(mundlak2, c(9:16))
```

```
## Warning in modelUpdate(objects[[i - 1]], objects[[i]]): more terms specified
## than existent in the model: 15, 16

## Warning in modelUpdate(objects[[i - 1]], objects[[i]]): column 'time'
## overwritten by time index

## Wald test
##
## Model 1: lnearnings ~ school + age + agesq + ethblack + I(ethblack * school) +
##     urban + regne + regnc + regw + asvabc + mean_school + mean_age +
##     mean_agesq + I(ethblack * mean_school)
## Model 2: lnearnings ~ school + age + agesq + ethblack + I(ethblack * school) +
##     urban + regne + regnc
##   Res.Df Df  Chisq Pr(>Chisq)
## 1  40028
## 2  40034 -6 422.44  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

1. Hausmant test and wald test indicate correlation between fixed effects and regressors. So we should use the fixed effects model. Our estimations have shown that the interaction effect between schooling and being black is significant and negative. So there evidence for discrimination on the labour market. etc.

(ix) To gain insights on the impact of nonresponse and attrition, the researcher applies a variant of the Verbeek and Nijman-test. 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 in the lecture slides). Draw conclusions and address practical problems you possibly met in implementing the test.\

```
data <- data %>% group_by(id) %>%
  mutate(dummy = ifelse(length(id) > 5, 1, 0))
```

```
fixedbalanced  <- plm(lnearnings ~ school + age + agesq + ethblack + ethblack  * school + urban + regne

fixedbalanced$vcov <- vcovHC(fixedbalanced, cluster="group")

phtest(fixed, fixedbalanced)
```

```
##
##  Hausman Test
##
## data:  lnearnings ~ school + age + agesq + ethblack + I(ethblack * school) +  ...
## chisq = 3.9223, df = 6, p-value = 0.6872
## alternative hypothesis: one model is inconsistent
```

We already had practical problems all along. The question is how to include robust standard errors in the Hausmantest. P value is very high so we can reject H_0. Hence, there is no evidence for attrition bias. As the estimation with the unbalanced data is more efficient, we should use that one.

References