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"?\

(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?\

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

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: Fr, Jan 14, 2022 - 18:31:53

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?\

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

```
# pooled3 <- plm(lnearnings ~ school + age + agesq + urban + regne + regnc + regw + regs + asvabc, mode
pooled4 <- plm(lnearnings ~ school + age + agesq + ethblack + I(ethblack * school) + urban + regne + reg
stargazer(pooled4)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Fr, Jan 14, 2022 - 18:31:53

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

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.\

```
# random1 <- plm(lnearnings ~ school + age + agesq  + urban + regne + regnc + regw + regs, model = "ran

# random2 <- plm(lnearnings ~ school + age + agesq + urban + regne + regnc + regw + regs + asvabc, mode

# random3 <- plm(lnearnings ~ school + age + agesq + ethblack + urban + regne + regnc + regw + regs + a

random4 <- plm(lnearnings ~ school + age + agesq + ethblack + I(ethblack  * school) + urban + regne + r
random4$vcov <- vcovHC(random4, cluster="group")
stargazer(random4)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Fr, Jan 14, 2022 - 18:32:04

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

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

```
# fixed1  <- plm(lnearnings ~ school + age + agesq  + urban + regne + regnc + regw, model = "within", i

# fixed2  <- plm(lnearnings ~ school + age + agesq + urban + regne + regnc + regw + asvabc, model = "wi

# fixed3  <- plm(lnearnings ~ school + age + agesq + ethblack + urban + regne + regnc + regw + asvabc, 

fixed4  <- plm(lnearnings ~ school + age + agesq + ethblack + I(ethblack  * school) + urban + regne + r
fixed4$vcov <- vcovHC(fixed4, cluster="group")
stargazer(fixed4)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Fr, Jan 14, 2022 - 18:32:15

(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.\

```
data$interaction <- data$ethblack*data$school

form <- lnearnings ~ school + age + agesq + ethblack +  interaction + urban + regne + regnc + regw + as

phtest(fixed4, random4)
```

```
##
##  Hausman Test
##
## data:  lnearnings ~ school + age + agesq + ethblack + I(ethblack * school) +  ...
## chisq = 20.366, df = 7, p-value = 0.004832
## alternative hypothesis: one model is inconsistent
```

Table 3:

	Dependent variable:
	lnearnings
school	0.050^{***}
	(0.003)
age	0.096^{***}
	(0.004)
agesq	-0.001^{***}
	(0.0001)
ethblack	-0.037
	(0.096)
I(ethblack ∗school)	-0.004
	(0.008)
urban	0.047^{***}
	(0.011)
regne	0.093^{***}
	(0.015)
regnc	-0.009
	(0.013)
regw	0.081^{***}
	(0.016)
asvabc	0.012^{***}
	(0.001)
Constant	-0.665^{***}
	(0.066)
Observations	40,043
R^2	0.309
Adjusted R^2	0.309
F Statistic	$4,337.823^{***}$
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 4:

	Dependent variable:
	lnearnings
school	0.051^{***}
	(0.006)
agesq	-0.001^{***}
	(0.0001)
I(ethblack $*$school)	-0.060^{***}
	(0.016)
urban	0.032^{***}
	(0.008)
regne	0.051^{**}
	(0.026)
regnc	-0.029
	(0.021)
regw	0.088^{***}
	(0.028)
Observations	40,043
R^2	0.020
Adjusted R^2	-0.113
F Statistic	34.587^{***} (df = 7; 35254)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
phtest(form, data = data, index = c("id", "time"), effect = "twoways")
```

```
##
##  Hausman Test
##
## data:  form
## chisq = 13.026, df = 7, p-value = 0.07148
## alternative hypothesis: one model is inconsistent
```

```
phtest(form, data = data, index = c("id", "time"), effect = "twoways", method = "aux", vcov = vcovHC)
```

```
##
##  Regression-based Hausman test, vcov: vcovHC
##
## data:  form
## chisq = 323.26, df = 7, p-value < 2.2e-16
## alternative hypothesis: one model is inconsistent
```

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

```
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 +
```

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

```
#mundlak2 <- feols(lnearnings ~ school + age + agesq + asvabc + mean_school + mean_age + mean_agesq + m

#mundlak3 <- plm(lnearnings ~ school + age + agesq + ethblack + asvabc + mean_school + mean_age + mean_

#mundlak4 <- plm(lnearnings ~ school + age + agesq + ethblack + ethblack  * school + asvabc + mean_scho
datafull <- na.omit(data)
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
```

```
summary(mundlak1)
```

Oneway (individual) effect General FGLS model

Call: pggls(formula = 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, data = datafull, effect = “individual”, model = “random”, index = c(“id”))

Unbalanced Panel: n = 4740, T = 1-18, N = 39800

Residuals: Min. 1st Qu. Median Mean 3rd Qu. Max. -1.72655 -0.21870 0.04607 0.05032 0.31548 3.04595

Coefficients: Estimate Std. Error z-value Pr(>|z|)
(Intercept) -1.0095e+00 2.1125e-01 -4.7784 1.767e-06 ***school 4.9100e-02 4.2397e-03 11.5810 < 2.2e-16***
age 6.7116e-02 3.3005e-03 20.3353 < 2.2e-16 ***agesq -6.7648e-04 5.3317e-05 -12.6879 < 2.2e-16***
ethblack -3.1254e-01 8.5602e-02 -3.6511 0.0002611 ***I(ethblack*** **school) -4.5467e-02 1.3714e-02 -3.3154 0.0009152** *urban 1.2519e-02 5.7480e-03 2.1780 0.0294052*
regne 7.4015e-02 1.6398e-02 4.5138 6.369e-06 ***regnc -1.6577e-02 1.4092e-02 -1.1764 0.2394452***
regw 8.4370e-02 1.6040e-02 5.2600 1.440e-07 asvabc 1.0830e-02 6.9762e-04 15.5239 < 2.2e-16
mean_school -3.6043e-03 4.9620e-03 -0.7264 0.4676057
mean_age 6.5444e-02 1.5844e-02 4.1305 3.620e-05 mean_agesq -1.3083e-03 2.7463e-04 -4.7639 1.899e-06 ***I(ethblack*** **mean_school) 6.2688e-02 1.5011e-02 4.1761 2.966e-05** *mean_urban 1.0273e-01 1.4624e-02 7.0249 2.143e-12* ** mean_regne 1.6098e-02 2.1454e-02 0.7504 0.4530256
mean_regnc 1.6924e-02 1.8440e-02 0.9178 0.3587391
mean_regw -2.8874e-02 2.1871e-02 -1.3202 0.1867571
— Signif. codes: 0 ‘’ ***0.001*** ’’ *0.01* ” 0.05 ‘.’ 0.1 ’ ’ 1 Total Sum of Squares: 10060 Residual Sum of Squares: 6965.2 Multiple R-squared: 0.30766

```
stargazer(mundlak2)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Fr, Jan 14, 2022 - 18:32:53

```
stargazer(fixed4)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Fr, Jan 14, 2022 - 18:32:57

```
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
```

Table 5:

	Dependent variable:
	lnearnings
school	0.053^{***}
	(0.004)
age	0.078^{***}
	(0.003)
agesq	-0.001^{***}
	(0.00004)
ethblack	-0.302^{***}
	(0.089)
I(ethblack ∗school)	-0.062^{***}
	(0.012)
urban	0.044^{***}
	(0.005)
regne	0.091^{***}
	(0.010)
regnc	-0.008
	(0.009)
regw	0.080^{***}
	(0.011)
asvabc	0.012^{***}
	(0.001)
mean_school	-0.002
	(0.005)
mean_age	0.071^{***}
	(0.016)
mean_agesq	-0.001^{***}
	(0.0003)
I(ethblack ∗mean_school)	0.079^{***}
	(0.013)
Constant	-1.236^{***}
	(0.218)
Observations	40,043
R^2	0.369
Adjusted R^2	0.368
F Statistic	$14{,}641.510^{***}$
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 6:

	Dependent variable:
	lnearnings
school	0.051^{***}
	(0.006)
agesq	-0.001^{***}
	(0.0001)
I(ethblack $*$school)	-0.060^{***}
	(0.016)
urban	0.032^{***}
	(0.008)
regne	0.051^{**}
	(0.026)
regnc	-0.029
	(0.021)
regw	0.088^{***}
	(0.028)
Observations	40,043
R^2	0.020
Adjusted R^2	-0.113
F Statistic	34.587^{***} (df = 7; 35254)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
## 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 -6 497.86  < 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?\

(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(fixed4, 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
```