Econometrics II - Assignment 2

Uncensored sloths

16 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. In order to demean the data, we have to subtract the mean from each regressors, the residuals as well as the dependent variable. Consider a static fixed-effect model

$$Y_{it} = \alpha + X_{it}\prime\beta + \eta_i + U_{it}$$

When we demean the data, the model is transformed to the following model

$$\tilde{Y_{it}} = \tilde{X_{it}}\prime\beta + \tilde{U_{it}} \equiv Y_{it} - \bar{Y_i} = (X_{it} - \bar{X_i})\prime\beta + (U_{it} - \bar{U_i})$$

As one can see, the fixed effect η_i drops out of the model as a result. Hence, we control for individual effects by transforming the model in a way that the fixed effects are not part of the model anymore. However, as a consequence, the fixed-effects model is not able to identify 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?

In order to consider time by adding a linear or polynomial term, the researcher would have to know the form of the relationship between time effect and the dependent variable. However, this can be quite challenging. This is why it simpler to consider time effects by adding a term for time. The fixed-effects model would then control for both, individual and time fixed effects. This is especially convenient when the researcher wants to control for time effects but at the same time time effects 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?

In comparison to the fixed-effects model, the random-effects model is able to estimate the coeffcients of time-invariant regressors. The reason is that instead of demeaning the data or taking the first difference, we estimate the model using (feasible) Generalized Least Squares (GLS). Therefore, it is also possible to do out-of sample predictions with the random-effects model as the time-invariant regressors are informative. If the assumptions hold, the random-effects model is more efficient than the fixed-effect model. However, the assumption on the stochactic strucutre of the individual effects are quite strict and if does not hold, the estimation is inconsistent.

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 - 17:12:43

Firstly, we can see that the index test score has a positive and significant effect at 1% on earnings. School has a positive and significant effect in both models, but the magnitude is lower in the model that includes the index test score (model 2) as a regressors. Age on the other side, has a slightly higher magnitude in the second model. Age squared is the same in both models, having a significant negative effect on earnings. Being black also a significant negative effect. However, the magnitude is considerably smaller in the second model. The living area and the regions all have a positive and significant effect in both models with the magnitude being higher for all regressors in the first model.

Especially, the estimations for school and being black differ substantially between both models. In case of school, a reason could be that individuals that attended school for a longer time, tend to achieve higher scores which is why including the test scores as a regressors lowers the estimated effect of school.

A reason why the (negative) effect of being black on the earnings is lower in the second model could be that individuals of black ethnicity tend to have lower scores in general. This could be an indicator for systemic discrimination within the school system. Due to the discriminatory structures, black individuals have a worse education on average which results in lower scores which and therefore lower earnings. Hence, including the test scores would lower the magnitude of the effect being black. However, this does not mean that being black has less impact on the earnings as we would have cross effects that are not considered at that point. If this is indeed the case, the current estimation is biased.

Note that one regional dummy (regs) was excluded. As all regional dummies are positive and significant, we can conclude that individuals living in the south earn relatively less to individuals living in other areas.

Moreover, R^2 is higher for the second model, indicating that the explanatory power is higher when we include the test scores as a regressors.

(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 + ethblack * school + urban + regne + regnc +
stargazer(pooled2, 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 - 17:12:44

The estimation shows that the interaction effect between schooling and being black is positive and significant at 1%. Attending school has therefore a higher impact on individuals of black ethnicity. The other estimates do not differ substantially from the previous estimated OLS model (that included test scores as a regressor. However, there is one exemption: being black has a substantially more negative effect on earnings than in the previous model. While the first model estimated that being black decreases the earnings by approximately 9.15%, the new model estimates, that being black decreases the earnings by approximately

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	
	(1)	(2)
school	0.048***	0.046***
	(0.001)	(0.001)
age	0.078***	0.079***
	(0.004)	(0.004)
agesq	−0.001***	−0.001***
	(0.0001)	(0.0001)
ethblack	−0.096***	−0.295***
	(0.007)	(0.040)
urban	0.101***	0.102***
	(0.005)	(0.005)
regne	0.123***	0.124***
	(0.006)	(0.006)
regnc	0.017***	0.017***
	(0.005)	(0.005)
regw	0.072***	0.072***
	(0.007)	(0.007)
asvabc	0.011***	0.011***
	(0.0003)	(0.0003)
school:ethblack		0.016***
		(0.003)
Constant	−0.386***	−0.370***
	(0.051)	(0.051)
Observations	40,043	40,043
R^2	0.313	0.313
Adjusted R^2	0.313	0.313
F Statistic	2,023.536*** (df = 9; 40033)	1,824.813*** (df = 10; 40032)
Note:		*p<0.1; **p<0.05; ***p<0.01

-25.55 ($e^{-0.295} - 1 \approx -0.25547$). Therefore, the estimation provides evidence that there is discrimination on the labour market.

Note that R^2 does not increase, however, with the new model 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 + ethblack  * school + urban + regne + regnc
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 - 17:12:56

We estimate a random-fixed model with the same regressors as in the previous OLS model including the interaction effect between school and being black. Further, we assume that there are individual and time fixed effects. In the case of the individual effects, we do not have data on the social and economic backgrounds of the individuals, further educational aspects as well as character traits such as motivation. Aspects like these could have potentially an effect on earnings which why we should consider them by including individual fixed effects. The data set considers a relatively long time frame (1980-2000). During these 20 years, policy changes and macroeconomic dynamics also could have impacted the earnings of individuals which is why they should be considered by including a time fixed effect.

We also cluster the standard errors as for the panel data at hand one may relatively safely assume that errors are serially correlated across time for each individual, so that standard errors should be clustered by individual. As outlined by Abadie et al. (2017) a more rigorous justification is necessary for the use of clustering which is only shortly touched. First, one must assume heterogeneity in the treatment effect, i.e. of the schooling effect. Since the research question is concerned with this heterogeneity and a priori contemplations support the assertion of heterogeneity this assumption is made here. Second, one must assume either non-random sampling of the population or a non-random allocation of the treatment variable. While the former cannot safely be assessed here, it does not seem far-fetched to assume that the level of schooling differs non-randomly across sampled individuals, e.g. in correlation to variables such as ethnicity or some unobserved variables. Therefore, clustered standard errors (by individuals) are estimated and included in all estimation outputs presented if not stated otherwise.

While the estimates of schooling and test score only increase marginally, the magnitude of the estimator of age is substantially higher than in the OLS model. The estimates of being black and the interaction between being black and schooling are not significant in the random-fixed model. The esimate for living in the urban area is lower in magnitude in the random-fixed model. This is also the case, for the estimator which indicates whether someone lived in the north-east. The estimate for living in north-central region is not significant anymore and the estimate for living in the west region is higher in magnitude than before.

The most striking changes are the estimates for age, being black and its interaction effect with schooling. A reason why the estimate for age is higher in the random-fixed model is that this model type recognizes the panel structure of the data and therefore, realizes that the individuals earn more the older they become. However, note that random-effects model only is efficient if the stochastic assumption on the fixed effect holds. Therefore, we should conduct tests before making any conclusion on the heterogeneity of returns to schooling by ethnicity based on the random-effecs model.

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

The advantage of a random-effects model is that it is able estimate the effect of time-invariant regressors and that it can be used to make predictions outside the sample. Further, it is more efficient than the fixed-effects model but only if the stochastic structure is assumed correctly (which is a rather strict assumption).

A fixed-effects model, on the other hand, is not able to identify time-invariant regressors and therefore, we

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)
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)
school:ethblack	0.016***	−0.004
	(0.003)	(0.008)
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

cannot make any out of sample predictions. Moreover, a sufficient variation in the regressors is necessary. However, it is robust to the correlation between the omitted heterogeneity and the regressors.

The main 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 could potentially have an impact on regressors such as schooling and test scores. 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 heterogeneous 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 - 17:13:08

We estimated a fixed-effects model with the same regressors as in the previous models. Following the same argumentation as with the random-effects model, we include individual and time fixed effects. Further, we cluster the standard errors. As the fixed-effects model is not able to identify time-invariant regressors, we do not have any estimation for being black and the test scores. Note that there also is no estimation for age! When considering time and individual fixed effects, we transform the model to the following form

$$Y_{it} - \bar{Y_i} - \bar{Y_t} + \bar{\bar{Y}} = (X_{it} - \bar{X_i} - \bar{X_t} + \bar{\bar{X}})\prime\beta + (U_{it} - \bar{U_i} - \bar{U_t} + \bar{\bar{U}})$$

where $\bar{Y_t} = \frac{1}{N}\sum_{i=1}^{N} Y_{it}$ and $\bar{\bar{Y}} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{i=1}^{T} Y_{it}$ (Cameron and Trivedi 2005, p. 738). Hence, after applying demeaning to the data age completely drops out. Nevertheless, we would proceed with both time and individuals fixed effects as there were important macroeconomic events between 1980 and 2000 that potentially could have had an impact on earnings. One example is the recession that took place between 1980 and 1982. Further, our main aim is to estimate the heterogeneity of returns to schooling of ethnicity and not the effect of age on earnings.

While school and age squared are estimated to have approximately the same effect as in previous models, the interaction between being black and school is now significant and negative. Therefore, being black and more years of schooling have not the same effect as for individuals of other ethnicities. The magnitude of the estimations for living in an urban area and living in the north-east region is lower in comparison to previous models. Living in the north-central region has no significant effect on earnings while living in the west region has a higher effect on earnings than before.

In comparison to the random effects model, the fixed-effects model provides us with evidence that there is discrimination on the labor market based on the ethnicity. Hence, we should test whether there is correlation between the individual fixed effects and the regressors to determine which model is more adequate.

(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.

To test whether there is correlation between regressors and the random effects (i.e. across individuals and time in the two-way specification adopted here), one can perform a Hausman test since under the null hypothesis of no correlation ($H_0 : \mathrm{E}[\eta_i|X_{i1}, \ldots, X_{iT}] = 0 \& \mathrm{E}[\lambda_t|X_{1t}, \ldots, X_{Nt}] = 0$) both estimators are consistent, while under the alternative hypothesis of correlation ($H_1 : \mathrm{E}[\eta_i|X_{i1}, \ldots, X_{iT}] \neq 0 \& \mathrm{E}[\lambda_t|X_{1t}, \ldots, X_{Nt}] \neq 0$) only the fixed-effect estimator is consistent. Under such asymptotic circumstances a Hausman test can be applied which is based on the following test statistic:

$$T = \left(\hat{\beta}_{FE} - \hat{\beta}_{RE}\right)' \left(\mathrm{Var}\left(\hat{\beta}_{FE}\right) - \mathrm{Var}\left(\hat{\beta}_{FE}\right)\right)^{-1} \left(\hat{\beta}_{FE} - \hat{\beta}_{RE}\right)$$

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)			−0.060***
			(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)	
school:ethblack	0.016***	−0.004	
	(0.003)	(0.008)	
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

Since this test statistic is based on the estimated variance of coefficients, standard errors are robustified against serial error correlation.

(Technical note: since the robustified variance-covariance matrix is passed to the respective model object no further robustification is necessary in the code below.)

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

```
##
##  Hausman Test
##
## data:  lnearnings ~ school + age + agesq + ethblack + I(ethblack * school) +  ...
## chisq = 2.1152, df = 6, p-value = 0.9088
## alternative hypothesis: one model is inconsistent
```

Given a p-value of below 1% we can reject the null hypothesis of the Hausman test. Therefore, the random effects model is inconsistent as the assumption of no correlation between the fixed effects and the regressors does not hold.

(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.

As an alternative to the Hausman test performed above, it is possible to perform a Wald test based on a Mundlak model which yields identical estimations to the within estimator. The model is specified as follows, where ω_i is a ransom effect uncorrelated with X_{it}.:

$$Y_{it} = X'_{it}\beta + \bar{X}_i\gamma + \omega_i + U_{it}$$

This model proposes a salient test for the null hypothesis of random effects, i.e. $H_0 : \gamma = 0$, against the alternative hypothesis of fixed effects, i.e. $H_1 : \gamma \neq 0$, by means of a Wald test. It can be shown that the Mundlak-test is asympototically equivalent to the Hausman test so that one would hope for consistency of their results.

To estimate the model, one first needs to compute all values of $\bar{X}_{it}$:

```
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))
```

Since the Mundlak model is specified as an one-way random individual effect model, it needs to be estimated by means of (feasibly) GLS. In the output below 2 different estimations of the Mundlak model are obtained by means of different pre-programmed R-functions of the 'pml' package.

First, the model is estimated using the command plm which does, however, not allow for the inclusion of meaned regional dummies. This may be due to insufficient variance in the regional membership of individuals across time, i.e. insufficient mobility, which leads to high collinearity between the meaned dummy variables

in $\bar{X}_i$ and the non-meaned dummy variables in X_{it}. To verify the notion that the estimate of β should be identical to the within estimator the results of a one-way fixed individual effect model are depicted alongside. As one would expect, estimates are identical for all coefficients except for those of regional membership since the Mundlak model cannot correct for between effects because respective means are not included in $\bar{X}_i$ due to the discussed collinearity problem.

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

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

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

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

```
# 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) {
        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)
```

```
screenreg(list(mundlak_plm, fixed_oneway_plm), 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
```

(Technical note: to obtain sleek output for the pggls model we used code from Taroh Maru 2015)

To robustify our analysis, estimation of the Mundlak-model is also achieved by means of the plm-package function pggls which does allow for the inclusion of meaned regional dummies. Although discussion of the exact reasons for the possibility to include these means only for the estimation technique adopted by the pggls function are beyond the scope of this assignment (and the knowledge of the uncensored sloths), it is not unlikely to assume that it can allow for higher collinearity between regressors. As above, results of a one-way fixed individual effect model estimated by means of pggls are depicted alongside but yield rather puzzling divergences which are not further discussed for reasons described above.

(Technical note: Since we did not ascertain how to implement robustification of standard errors obtained through pggls, inferential results based on pggls-estimations obtained in this assignment should be read under the caveat of incorrect variance estimation.)

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

```
## 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
```

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

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

```
screenreg(list(mundlak_pggls, fixed_oneway_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(mundlak_plm, c(9:16))
```

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

```
## 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)
## 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
```

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

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

```
## Warning: 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
```

However, both estimation techniques yield the same conclusion once a Wald test is applied to test the joint significance of all coefficients in γ. In both cases there is very significant evidence against the null hypothesis $H_0 : \gamma = 0$ so that we can reject the null hypothesis of random effects. As one would expect, this is in line with the result of the Hausman test discussed above. Therefore, we conclude that the random effects mode is inconsistent for the panel data at hand, so that any inferences towards heterogeneity of schooling effects should be based on a fixed effect mode here.

To further robustify this conclusion, one should note that the Mundlak-model above is specified for the case of one-way fixed individual effects, while the preceding model specification and Hausman test results presume two-way fixed effects across individuals and time. Therefore, we also estimated the two-way Mundlak model as suggested by Wooldrige (2021), where $\bar{X}_t$ includes means over indivduals for all time-variant regressors:

$$Y_{it} = X'_{it}\beta + \left(\bar{X}_i \bar{X}_t\right)\gamma + \omega_i + U_{it}$$

To avoid collinearity probolems in estimation, (fesible) GLS results are obtained using the pggls function and yield the same conclusion towards the inconsistency of random effects estimation upon investigation of the Wald-test results.

(Technical note: for reasons unknown to the uncensored sloths exclusion of incomplete observations was necessary to perform the Wald test on the two-way Mundlak-model.)

```
data <- data %>%
  group_by(time) %>%
  mutate(mean_school_t = mean(school),
         mean_age_t = mean(age),
         mean_agesq_t = mean(agesq),
         mean_asvabc_t = mean(asvabc),
         mean_urban_t = mean(urban),
         mean_regne_t = mean(regne),
         mean_regnc_t = mean(regnc),
         mean_regw_t = mean(regw))

datafull <- na.omit(data)

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

```
## 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
```

```
screenreg(list(mundlak_twoway, fixed), digits=3, single.row=TRUE)
```

```
##
## ========================================================================
##                            Model 1                Model 2
## ------------------------------------------------------------------------
## (Intercept)                  -2.272 (0.331) ***
## school                        0.046 (0.005) ***      0.051 (0.006) ***
## age                           0.093 (0.006) ***
## agesq                        -0.001 (0.000) ***     -0.001 (0.000) ***
## ethblack                      1.041 (0.220) ***
## ethblack * school            -0.021 (0.015)         -0.060 (0.016) ***
```

```
## urban                          0.028 (0.007) ***      0.032 (0.008) ***
## regne                          0.067 (0.017) ***      0.051 (0.026) *
## regnc                         -0.011 (0.015)         -0.029 (0.021)
## regw                           0.081 (0.017) ***      0.088 (0.028) **
## asvabc                         0.011 (0.001) ***
## mean_school                   -0.002 (0.005)
## mean_age                       0.059 (0.016) ***
## mean_agesq                    -0.001 (0.000) ***
## ethblack * mean_school         0.042 (0.016) *
## mean_urban                     0.096 (0.015) ***
## mean_regne                     0.026 (0.022)
## mean_regnc                     0.014 (0.019)
## mean_regw                     -0.030 (0.023)
## mean_school_t                  0.195 (0.024) ***
## mean_age_t                    -0.147 (0.020) ***
## mean_agesq_t                   0.002 (0.000) ***
## ethblack * mean_school_t      -0.108 (0.016) ***
## mean_urban_t                   0.192 (0.246)
## mean_regne_t                   0.231 (0.420)
## mean_regnc_t                   0.733 (0.328) *
## mean_regw_t                    2.106 (0.534) ***
## ------------------------------------------------------------------------
## R^2                            0.300                  0.020
## Num. obs.                  35292                  40043
## Adj. R^2                                             -0.113
## ========================================================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

```
waldtest(mundlak_twoway, c(9:24))
```

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

## Warning: 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 + mean_school_t + mean_age_t + mean_agesq_t +
##     I(ethblack * mean_school_t) + mean_urban_t + mean_regne_t +
##     mean_regnc_t + mean_regw_t
## Model 2: lnearnings ~ school + age + agesq + ethblack + I(ethblack * school) +
##     urban + regne + regnc + mean_regnc_t + mean_regw_t
##   Res.Df  Df  Chisq Pr(>Chisq)
## 1  35265
## 2  35281 -16 617.57  < 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?

As discussed above, the Hausmant test and Wald test(s) indicate clearly provide evidence for incosistency of random effects estimation so that the analysis of heterogeneity in returns to schooling should be based on the specified fixed effects model. Although this precludes identification of the effect of black ethnicity on wages,

this does not pose a problem for the research question which is concerned with the identified cross effect of schooling and ethnicity. Fixed effects estimation suggests that this interaction effect between schooling and black ethnicity is significant and negative, so that the returns to schooling are lower for people of colour classified as black here. Based on this fiding one can conlcude that there is evidence for discrimination by ethnicity on the labour market, which awards schooling less so for people of colour. Furthermore, one could conclude that there may be some discriminatory notion in schooling itself so that people of colour on average obtain education which is of lower quality and thus less rewarded on the labour market.

(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
```

To test attrition bias in panel data, we estimate the model once with an unbalanced panel and once with a balanced panel where we only include data on individuals who participated in more than five waves. The test itself is a Hausman type test based on $\hat{\beta}_{balanced} - \hat{\beta}_{unbalanced}$.

Generally, we already had practical problems all along with the Hausman test as we estimated robust standard errors but they are not automatically included in the estimated model. Hence, we substituted the covariance matrix of the model with the covariance matrix containing the robust standard errors.

As the p-value is around 0.6872, we cannot reject the null hypothesis, indicating that there is no attrition bias. As the estimation with the unbalanced data is more efficient, one should proceed with the unbalanced panel data.

References

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. (2017). When should you adjust standard errors for clustering? (No. w24003). National Bureau of Economic Research.

Cameron, A. C., & Trivedi, P. K. (2005). Microeconometrics: methods and applications. 10th edition, Cambridge university press.

Wooldridge, J. (2021). Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators.

Maru, Taroh (2015). Model Diagnostic Exercises. Online: http://tarohmaru.web.fc2.com/R/ExerciseDiagnostics.html [16.01.2022].