# Computing Exercise: ECON 769

*Vladimir Smiljanic*

*2017-05-06*

### Question 1

The following code produces a 50% subsample of the source data, sampling on the individuals:

```
set.seed(42)

resamp<-function(data,replace,size){ #Resampling function
  unique_ind<-unique(data$id)
  samp_ind<-sample(unique_ind,size=size,replace=replace)
  do.call(rbind,lapply(samp_ind,function(x) data[data$id==x,]))
}

mydata<-resamp(source,replace=FALSE,size=266)
```

### Question 2

We can display the coefficients and their standard errors from the different models we will be estimating:

```
require(plm)
require(stargazer)
form<-as.formula(paste("lnhr~lnwg+kids+disab+ageh+agesq"))

pooled<- plm(form, data = mydata, model = "pooling",index=c("id"))
between <- plm(form, data = mydata, model = "between",index=c("id"))
fd<-plm(form, data = mydata, model = "fd",index=c("id"))
random<-plm(form, data = mydata, model = "random",index=c("id"),effect="individual")
within <- plm(form, data = mydata, model = "within",index=c("id"),effect="individual")
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mon, May 15, 2017 - 12:37:10

Table 1:

| | *Dependent variable:* | | | | |
|---|---|---|---|---|---|
| | lnhr | | | | |
| | Pooled | Within | Random | Between | First-Diff. |
| | (1) | (2) | (3) | (4) | (5) |
| Constant | 7.744*** | | 7.275*** | 8.041*** | |
| | (0.136) | | (0.170) | (0.335) | |
| lnwg | 0.097*** | 0.358*** | 0.204*** | 0.051* | 0.247*** |
| | (0.014) | (0.030) | (0.021) | (0.030) | (0.034) |
| kids | 0.009 | −0.004 | 0.006 | 0.007 | −0.012 |
| | (0.006) | (0.010) | (0.008) | (0.014) | (0.017) |
| disab | −0.106*** | −0.084*** | −0.085*** | −0.142* | −0.071** |
| | (0.027) | (0.031) | (0.029) | (0.075) | (0.031) |
| ageh | −0.017** | 0.001 | −0.007 | −0.026 | −0.059* |
| | (0.007) | (0.010) | (0.009) | (0.017) | (0.033) |
| agesq | 0.0002** | −0.00002 | 0.0001 | 0.0003 | 0.0004 |
| | (0.0001) | (0.0001) | (0.0001) | (0.0002) | (0.0004) |
| Constant | 7.744*** | | 7.275*** | 8.041*** | |
| | (0.136) | | (0.170) | (0.335) | |
| Observations | 2,660 | 2,660 | 2,660 | 266 | 2,394 |
| $R^2$ | 0.027 | 0.059 | 0.037 | 0.038 | 0.027 |
| Adjusted $R^2$ | 0.025 | −0.047 | 0.035 | 0.019 | 0.025 |
| *Note:* | | | | | *p<0.1; **p<0.05; ***p<0.01 |

To produce bootstrap standard errors, I built my own R function that would resample on the individuals. I spent a good deal of time attempting to find a package that would allow me to resample panel data but resulted in me building my own iterating technique. The code can be found the accompanying document (issue fitting into PDF):

Table 2: Linear Model: Bootstrap Standard Errors

| Coefficients | Pooled | Within | Random Effect | Between | First Difference |
|---|---|---|---|---|---|
| lnwg | 0.050 | 0.134 | 0.107 | 0.022 | 0.141 |
| kids | 0.009 | 0.017 | 0.013 | 0.009 | 0.020 |
| disab | 0.047 | 0.051 | 0.066 | 0.078 | 0.081 |
| ageh | 0.013 | 0.017 | 0.012 | 0.012 | 0.021 |
| agesq | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |

**Question 3**

### 3.1.1 Pooled OLS

Pooled OLS specifies constant coefficients and can be defined by the model:

$$y_{it} = \alpha + x'_{it}\beta + u_{it}, i = 1 \dots N$$
$$t = 1 \dots T$$

Since all of the data is pooled, the number of observations is equal to $N \times T$. If Cov$[u_{it}, x_{it}]$=0, then either $N \to \infty$ or $T \to \infty$ is sufficient for consistency. The pooled OLS estimator will be consistent if there is no correlation between $\alpha$ and $x_{it}$ and that the regressors are uncorrelated with the error term.

In addition, our model will be inconsistent if the true model is a fixed effects model where the constant term is heterogeneous for individual and is correlated with the regressors. Using pooled OLS estimation will result in an inconsistent $\beta$ as a result of the heterogeneity being placed into the error term, which would now be correlated with the regressors. If the underlying true model is random effects, then using pooled OLS will result in consistent estimator due to the constant term being iid.

$$y_{it} = \alpha + x'_{it}\beta + u_{it}, i = 1 \dots N$$
$$t = 1 \dots T$$

Where $u_{it} = \alpha_i - \alpha + \epsilon_{it}$. If there is a fixed effect for true model, then the $\alpha_i$ will be within the error term, be correlated with the regressors resulting in inconsistent estimator.

### 3.1.2 Between

Where the pooled OLS estimator uses variation over both time and cross-section to estimate $\beta$, the between estimator averages and individuals variables over time, thus only using variation between individuals. Using the individual-specific effects model ($y_{it} = \alpha_i + x'_{it}\beta + \epsilon_{it}$):

$$\bar{y}_i = \alpha + \bar{x}'_i\beta + (\alpha_i - \alpha_{it} + \bar{\epsilon}_i), i = 1 \dots N$$

where $\bar{y}_i = T^{-1}\sum_t y_{it}, \bar{\epsilon}_i = T^{-1}\sum_t \epsilon_{it}, \bar{x}_i = T^{-1}\sum_t x_{it}$.

The between estimator is consistent if the regressors $\bar{x}_i$ are independent of the composite error $(\alpha_i - \alpha_{it} + \bar{\epsilon}_i)$. This is fine for a constant coefficient or random effects model but will be inconsistent for a fixed effects model. The individual heterogeneity of the constant, which is correlated overtime, will be found in the error term. This correlation will be found within the $x_{it}$ and thus the $\bar{x}_i$ and will result in an inconsistent estimator.

### 3.1.3 Within

To ensure consistency, we require strict exogeneity conditional on the unobserved effect. That means we cannot have dependency on lagged variables, specifically feedback of y to future x's.

Where we have a model:

$$y_i = \alpha_i + x'_i\beta + \epsilon_{it}, i = 1 \dots N$$

We will require E[$\epsilon_{it}|x_{i1}, \ldots, x_{Nt}$]=0.

'Within' estimator takes advantage of the setup of panel data, where you can use the individual-specified deviations of regressors and dependent variable to find their association with their time-averaged values. The regression over time $\bar{y}_i = \alpha_+ \bar{x}_i'\beta + \bar{\epsilon}_i)$ can be subtracted from our individual-specified model to eliminate the individual varying constant.

$$y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i)'\beta + (\epsilon_{it} - \bar{\epsilon}_i), i = 1 \ldots N, t = 1 \ldots T$$

This will lead to a consistent estimator in fixed effects models, along with pooled OLS and random effects models. Though $\beta$s can be estimated consistently, the individual constant terms can also be estimated but will be inconsistent in short panels. OLS and GLS can be used for the 'within' estimators to achieve consistency.

Relating back to the exogeneity required, the condition that must hold for consistent estimation is:

$$E[\epsilon_{it} - \bar{\epsilon}_i | x_{i1} - \bar{x}_i, \ldots, x_{Nt}] = 0$$

Though the original strict exogeneity condition is sufficient.

A limitation is if any of the regressors are time-invariant, then they would be differenced out and you will not be able to find coefficients for them.

### 3.1.4 First Differences

First-differences measures the one-period change in regressors and dependent variables. By using the individual-specific effects model, and lagging it one period, you can difference out the $\alpha_i$ term and provide yourself with a consistent estimator using OLS.

$$y_{it-1} = \alpha_i + x_{it-1}'\beta + \epsilon_{it-1}, i = 1 \ldots N$$
$$t = 1 \ldots T$$

By differencing out the individual-specific constant term, we handle any correlations present within the data. As a result, a fixed effects model, would yield a consistent estimator (along with pooled OLS and random effects model).

$$(y_{it} - y_{it-1}) = (x_{it}' - x_{it-1}')\beta + (\epsilon_{it} - \epsilon_{it-1}), i = 1 \ldots N$$
$$t = 1 \ldots T$$

Issues do occur if the regressors are time-invariant and hence differencing two periods would result in zero ($x_{it}' - x_{it-1}' = 0$). In addition, we are also under the assumption that it is a fixed effects model and the constant term is time-invariant.

### 3.1.5 Random Effects

Random effects makes the assumption that the regressors and the idiosyncratic errors have no correlation. As a result, we can use both GLS and OLS to find consistent estimators. Once again, we need to have strict exogeneity between the x's and the $\alpha$. Even though the constant terms are individual-specific, they are assumed to be iid $[0, \sigma^2]$ and have no correlation with the error term which is also iid $[0, \sigma^2]$. As a result, using the random effects estimator on a true model that is pooled OLS or random effects will be consistent but inconsistent with fixed effects as the iid constant term assumption will not hold.

### 3.2.1 Pooled OLS

We cannot use the usual OLS standard errors because they are based on iid errors, something that is not true for pooled OLS. Errors are auto-correlated over time. In short panels, there is an assumption that there is independence over individuals.We must use panel-corrected standard errors to ensure that we are providing the correct inference.

We must correct the standard errors due to correlation that an individuals error term have over time. Each individual dependent variable $y_{it}$ will have a high correlation over time. Therefore, if the model over predicts the dependent variable in one year, it may also over predict for the same individual the next year. OLS treats the dependent variables over time as independent observations but as we see from our data this isn't true. Our standard errors will show a level of precision that is not correct and overstate how precise the estimator is.

The default standard errors presented by the regression output would be valid if we the errors were homoskedastic and not serially correlated (over time). Failure to control for homoskedasticity will give you a downward bias.

### 3.2.2 Between

The inclusion of $\alpha_i$ will help control for serial correlation but the large gap between the default standard errors and the bootstrap standard errors means there is still an issue. The default standard error still is under the assumption that the errors are homoskedastic. Since between estimators only uses cross-section variation, it only is under the homoskedastic assumption.

### 3.2.3 Within

The distribution of the 'within' estimator can be complicated because the $\epsilon_{it} - \bar{\epsilon}_i$ is correlated over time given individuals. The strong assumption is made that if $\epsilon_{it}$ is iid to provide for consistent standard errors. That is because we can relax the requirement for no serial correlation of error terms in the 'within' model and can apply the standard OLS results.

### 3.2.4 First Differences

First-difference estimators need to account for correlation over time in the error term $\epsilon_{it} - \epsilon_{i,t-1}$. You cannot use OLS standard errors for 'first-difference' model as they only apply if $\epsilon_{it}$ is a random walk and $\epsilon_{it} - \epsilon_{i,t-1}$ is iid. We need to use a robust standard error estimation by assuming $\epsilon_{it} - \epsilon_{i,t-1}$ is a MA(1) process.

### 3.2.5 Random Effects

The assumption is made that both $\alpha_i$ and $\epsilon_{it}$ are iid, important for when they are placed in the error term, $u_{it} = \alpha_i + \epsilon_{it}$. 'Random effects' doesn't require more then strict exogeneity and linear independence of transformed data for consistency. The assumptions of homoskedasticity are not important for consistency.

## Question 4

The Hausmen test assumes that the fixed effect and random effect estimators are consistent under the null. The alternative hypothesis is that only fixed effect is consistent. We are attempting to find if there is any correlation between the individual-specific effects and the regressors. We assume that the 'random effect' estimator is efficient under null, so no heteroskedasticity is present. If there is statistically significant difference between the estimators, then 'fixed effects' are present.

```
##
##  Hausman Test
##
## data:  lnhr ~ lnwg + kids + disab + ageh + agesq
## chisq = 58.309, df = 5, p-value = 2.716e-11
## alternative hypothesis: one model is inconsistent
```

```
##
##  Hausman Test
##
## data:  lnhr ~ lnwg
## chisq = 54.227, df = 1, p-value = 1.786e-13
## alternative hypothesis: one model is inconsistent
```

We can conclude that with both specifications, we produce a large statistic which would reject our null hypothesis. This means we would reject that our individual-specific effects are uncorrelated with regressors and reject that the 'random effects' estimator is consistent.

## Question 5

We will take the exponetial of the 'lnhr' variable and divide it by 365. This will produce the number of hours worked per day to the nearest hour:

```
mydata$hr<-round(exp(mydata$lnhr)/365)
```

## Question 6

Using our new count variable, we will estimate the pooled, within and random effects models:

```r
require(pglm)
form<-as.formula(paste("hr~lnwg+kids+disab+ageh+agesq"))

summary(glm(form,family="poisson",data=mydata))
```

```
##
## Call:
## glm(formula = form, family = "poisson", data = mydata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4372  -0.3746   0.0095   0.3257   2.8604
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.9876454  0.1728180  11.501   <2e-16 ***
## lnwg         0.0197940  0.0182261   1.086   0.2775
## kids         0.0102421  0.0077490   1.322   0.1863
## disab       -0.0686939  0.0352954  -1.946   0.0516 .
## ageh        -0.0131688  0.0089430  -1.473   0.1409
## agesq        0.0001483  0.0001100   1.348   0.1777
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 954.68  on 2659  degrees of freedom
## Residual deviance: 943.98  on 2654  degrees of freedom
## AIC: 10543
##
## Number of Fisher Scoring iterations: 4
```

```r
summary(pglm(form,model="within",family=poisson,data=mydata,index="id"))
```

```
## ------------------------------------------
## Maximum Likelihood estimation
## Newton-Raphson maximisation, 2 iterations
## Return code 2: successive function values within tolerance limit
## Log-Likelihood: -4261.452
## 5  free parameters
## Estimates:
##         Estimate Std. error t value Pr(> t)
## lnwg   6.391e-02  5.025e-02   1.272   0.203
## kids   8.870e-04  1.534e-02   0.058   0.954
## disab -3.589e-02  4.962e-02  -0.723   0.469
## ageh   2.923e-03  1.644e-02   0.178   0.859
## agesq -2.861e-05  2.091e-04  -0.137   0.891
## ------------------------------------------
```

```r
summary(pglm(form,model="random",family=poisson,data=mydata,index="id",effect="individual"))
```

```
## ------------------------------------------
## Maximum Likelihood estimation
## Newton-Raphson maximisation, 13 iterations
## Return code 2: successive function values within tolerance limit
```

```
## Log-Likelihood: -5245.773
## 7  free parameters
## Estimates:
##               Estimate Std. error t value Pr(> t)
## (Intercept)  1.9259228        Inf       0       1
## lnwg         0.0231170        Inf       0       1
## kids         0.0096498        Inf       0       1
## disab       -0.0606032        Inf       0       1
## ageh        -0.0105335        Inf       0       1
## agesq        0.0001174        Inf       0       1
## sigma       92.0771738        Inf       0       1
## --------------------------------------------
```

Unfortunately, bootstrapping with "random effect" and "fixed effect" proved to be difficult to solve using R at this point in time. I was not able to produce Standard Errors that were correct and a result I will not be presenting them at this time.

**Question 7**