Panel Data Models

Mohammad Arshad Rahman
(webpage: https://www.arshadrahman.com)

Course: Econometrics
Chennai Mathematical Institute

Panel Data

A panel, or longitudinal, data set is one where there are repeated observations on the same N units: individuals, households, firms, countries, or any set of entities that remain stable through time.

Repeated observations create a potentially very large panel data sets. With N units and T time periods, we have NT observations.

- Advantage: Large sample, great for estimation.
- Disadvantage: Dependence! Observations are, likely, not independent.

Modeling the potential dependence creates different models.

Panel Data Sets

The National Longitudinal Survey of Youth (NLSY) is an example of panel data set. The same respondents were interviewed every year from 1979 to 1994. Since 1994 they have been interviewed every two years.

Panel data allows a researcher to study cross section effects (along N such as variation across firms) and time series effects (along T which is variation across time).

$$
\begin{array}{cc}
 & \text{Cross section} \\
\texttt{Time series} & \begin{pmatrix}
y_{11} & y_{21} & \dots & y_{i1} & \dots & y_{N1} \\
y_{12} & y_{22} & \dots & y_{i2} & \dots & y_{N2} \\
\dots & \dots & \dots & \vdots & \dots & \dots \\
y_{1t} & y_{2t} & \dots & y_{it} & \dots & y_{Nt} \\
\dots & \dots & \dots & \vdots & \dots & \dots \\
y_{1T} & y_{2T} & \dots & y_{iT} & \dots & y_{NT}
\end{pmatrix} = [y_1 \; y_2 \; \dots y_N]
\end{array}
$$

Panel Data Sets

Notation:

$$y_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT_i} \end{pmatrix}, \quad X_i = \begin{pmatrix} x_{i1,1} & x_{i2,1} & \cdots & x_{ik,1} \\ x_{i1,2} & x_{i2,2} & \cdots & x_{ik,2} \\ \vdots & \vdots & \ddots & \cdots \\ x_{i1,T_i} & x_{i2,T_i} & \cdots & x_{ik,T_i} \end{pmatrix}.$$

Stacking the model over the i's we have

$$y = X\beta + c + \varepsilon,$$

where X is $\sum_{i=1}^{N} T_i \times k$ matrix, β is a $k \times 1$ matrix, c is a $\sum_{i=1}^{N} T_i \times 1$ matrix associated with unobservable variables, and y and ε are $\sum_{i=1}^{N} T_i \times 1$ matrices.

Balanced and Unbalanced Panels

Notation: $y_{i,t},\ i = 1, \ldots, N;\ t = 1, \ldots, T_i$

Balanced Panel: number of observations is NT.

(This implies that every unit is surveyed in every time period. This is mathematically and notationally convenient.)

Unbalanced Panel: number of observations is $\sum_{i=1}^{N} T_i$.

Q: Is the fixed T_i assumption ever necessary? SUR models.

The NLSY data is unbalanced because some individuals have not been interviewed in some years. Some could not be located, some refused, and a few have died.

Panel Data Models

With panel data we can study different issues:

(1) *Cross sectional* variation (unobservable in time series data) vs *Time series* variation (unobservable in cross sectional data).

(2) Heterogeneity (observable and unobservable heterogeneity).

(3) Hierarchical structures (say zip codes, city, and state effects).

(4) Dynamics in economic behavior.

(5) Individual/group effects.

(6) time effects.

Panel Data Models: Example 1 - SUR

In Zellner's SUR formulation (no linear dependence on y_{it}) we have:

(A1) $y_{it} = x_{it}'\beta_i + \varepsilon_{it}$ – the DGP (x_{it}' includes an intercept)

(A2) $E[\varepsilon_i|X] = 0$,

(A3') $\text{Var}[\varepsilon_i|X] = \sigma_i^2 I_T = \sigma_{ii} I_T$ – groupwise heteroscedasticity

$\text{E}[\varepsilon_{it}\varepsilon_{jt}|X] = \sigma_{ij}$ – contemporaneous correlation

$\text{E}[\varepsilon_{it}\varepsilon_{js}|X] = 0$ – when $t \neq s$

(A4) $\text{Rank}(X) = \text{full rank}$

In (A1)-(A4), we have a generalized regression model with heteroscedasticity. OLS in each equation is fine, but not efficient. GLS is efficient.

If each equation is estimated separately, the we are not taking advantage of pooling i.e., using NT observations. Use LR or F tests to check if pooling (aggregation) can be done.

Panel Data Models: Example 2 - Pooling

Assumptions:

(A1) $y_{it} = x_{it}'\beta + z_i'\gamma + \varepsilon_{it}$ – the DGP (x_{it}' includes an intercept)

$i = 1, 2, \ldots, N$ – we have N individual, groups, or firms

$t = 1, 2, \ldots, T_i$ – usually $N >> T_i$

(A2) $E[\varepsilon_i|X, Z] = 0,$ – X and Z: exogenous

(A3) $\text{Var}[\varepsilon_i|X, Z] = \sigma_i^2 I$ – heteroscedasticity can be allowed

(A4) $\text{Rank}(X) = $ full rank

We think of X as a vector of observed characteristics. For example, firm size, market-to-book ratio, z-score, R & D expenditures, etc.

We think of Z as a vector of unobserved characteristics (`individual effects`). For example: quality of management, growth opportunities.

Panel Data Models: Basic Model

The DGP (A1) is linear:

$$y_{it} = \beta_1 + \sum_{j=2}^{k} x_{ij,t}\beta_j + \sum_{p=1}^{s} z_{ip}\gamma_p + \delta t + \varepsilon_{i,t}$$

Indices:

- i: individuals i.e., unit of observation,
- t: time period,
- k: number of observed explanatory variables,
- s: number of unobserved explanatory variables.

Time trend t allows for a shift of the intercepts over time, capturing time effects such as technological change, regulations, etc. But, if the implicit assumption of a constant rate of change is strong ($= \delta$), we use a set of dummy variables, one for each time period except reference period.

x_i: The variables of interest with associated parameter vector β.

z_i: The variables responsible for unobserved heterogeneity (& dependence on the y_i's). Usually, a nuisance component of the model.

The z_p variables are unobserved: Impossible to obtain information about each component in $\sum_{p=1}^{s} z_{ip}\gamma_p$. We define a term c_i, `the unobserved effect`, representing the joint impact of the z_p variables on y_i – like an index of unobservables for individual i:

$$c_i = \sum_{p=1}^{s} z_{ip}\gamma_p.$$

We can rewrite the regression model as:

$$y_{it} = \beta_1 + \sum_{j=2}^{k} x_{ij,t}\beta_j + c_i + \delta t + \varepsilon_{i,t}.$$

Panel Data Models: Basic Model

`Note`: If the X_j's are so comprehensive that they capture all relevant characteristics of individual i, c_i can be dropped and, then, pooled OLS may be used. But this situation is very unlikely.

In general, dropping c_i leads to missing variables problem: omitted variable bias!

We usually think of c_i as `contemporaneously exogenous` to the conditional error, That is, $E[\varepsilon_{it}|c_i] = 0$, $t = 1, \ldots, T$.

A stronger assumption: `Strict exogeneity` can also be imposed. Then, $E[\varepsilon_{it}|x_{i,1}, x_{i,2}, \ldots, x_{i,T}, c_i] = 0$ for $t = 1, \ldots, T$.

Panel Data Models: Basic Model

Strict exogeneity conditions on the whole history of x_i. Under this assumption:

$$E\left[y_{i,t}|x_{i,t}, c_i\right] = \beta_1 + \sum_{j=2}^{k} x_{ij,t}\beta_j + c_i + \delta t,$$

which implies that β_j's are partial effects holding c_i constant.

Violations of strict exogeneity are not rare. For example, if $x_{i,t}$ contains lagged dependent variables or if changes in ε_{it} affect $x_{i,t+1}$ (a "feedback" effect).

But to estimate β we still need to say something about the relation between $x_{i,t}$ and c_i. Different assumptions will give rise to different models.

Panel Data Models: Types

The basic DGP: $y_{i,t} = x'_{i,t}\beta + z'_i\gamma + \varepsilon_{i,t}$, & (A2)-(A4) apply.

Depending on how we model the heterogeneity in the panel, we have different models.

(1) Pooled (Constant Effect) Model

$z'_i\gamma$ is a constant term. $z'_i\gamma = \alpha$ (intercept, and uncorrelated with $x_{i,t}$!). Dependence on the $y_{i,t}$ may enter through the variance. That is, repeated observations on individual i are linearly independent. In this case,

$$y_{it} = x'_{i,t}\beta + \alpha + \varepsilon_{i,t} \quad (x'_{i,t} \text{ does not include intercept}).$$

OLS estimates α and β consistently. We estimate $k+1$ parameters.

(2) Fixed Effects Model (FEM)

The z_i's are correlated with X_i fixed effects:

$$E[z_i|X_i] = g(X_i) = \alpha^*.$$

The unobservable effects are correlated with included variables. So, pooled OLS will be inconsistent.

Assume $z_i'\gamma = \alpha_i$ (a constant, it does not vary with t). Then,

$$y_{i,t} = x_{i,t}'\beta + \alpha_i + \varepsilon_{i,t} \quad (x_{i,t}' \text{ does not include an intercept}).$$

The regression line is raised/lowered by a fixed amount for each individual i (the dependence created by the repeated observations!). In econometrics terms, this is the source of the `fixed effects`.

We have a lot of parameters: $k + N$. We have N individual effects! OLS can be used to estimate α and β consistently.

(3) Random Effects Model (REM)

The differences between individuals are random, drawn from a given distribution with constant parameters. We assume the z_i's are uncorrelated with the X_i. That is,

$$E[z_i|X_i] = \mu \quad \text{(if } X_i \text{ contains a constant term } \mu = 0, WLOG)$$

Add and subtract $E[z_i'\gamma] = \alpha$ to the basic DGP:

$$\begin{aligned} y_{i,t} &= x_{i,t}'\beta + E[z_i'\gamma] + (z_i'\gamma) - E[z_i'\gamma] + \varepsilon_{i,t} \\ &= x_{i,t}'\beta + \alpha + u_i + \varepsilon_{i,t}, \qquad u_i = (z_i'\gamma) - E[z_i'\gamma]. \end{aligned}$$

We have a compound (composed) error i.e., $u_i + \varepsilon_{i,t} = w_{i,t}$. This $w_{i,t}$ introduces contemporaneous correlation across the i group.

OLS estimates α and β consistently, but GLS will be efficient.

(4) Random Parameters/Coefficients Model

We introduce heterogeneity through β_i, but this may introduce additional N parameters. A solution is to model β_i. For example,

$$y_{i,t} = x'_{i,t}(\beta + h_i) + \alpha_i + \varepsilon_{i,t},$$

where h_i is a random vector that induces parameter variation, where $h_i \sim D(0, \sigma^2_{h_i})$. That is, we introduce heteroscedasticity.

Now the coefficients are different for each individual. It is possible to complicate the model by making them different though time:

$$\beta_{it} = (\beta + h_i) + \theta_t \qquad \text{where } \theta_t \sim D(0, \sigma^2_t).$$

Estimation technique: GLS, MLE.

Compact Notation

Model for individual i: $y_i = X_i\beta + c_i + \varepsilon_i$, where

X_i is a $T_i \times k$ matrix (does not include a column of 1's)

β is a $k \times 1$ matrix

$c_i = \sum_{p=1}^{s} z_{ip}\gamma_p$ is a $T_i \times 1$ matrix

y_i and ε_i are $T_i \times 1$ matrices

Stacking the model over i's: $y = X\beta + c + \varepsilon$

X is a $\sum_{i=1}^{N} T_i \times k$ matrix

β is a $k \times 1$ matrix

c, y, and ε are $\sum_{i=1}^{N} T_i \times k$ matrices

Or, $y = X^*\beta^* + \varepsilon$, with $X^* = [X\,\iota]$, $\sum_{i=1}^{N} T_i \times (k+1)$ matrix

$\beta^* = [\beta\, c]'$, $(k+1) \times 1$ matrix

Assumption for Asymptotics (Greene)

(1) Convergence of moments involving cross section X_i.

Ususally, we assume N increasing, T or T_i assumed fixed.

> `Fixed-T` asymptotics (see Greene)
>
> Time series characteristics are not relevant (may be nonstationary).
>
> If T is growing, need to treat as multivariate time series.

(2) Rank of matrices. X must have full column rank. However, X_i may not, if $T_i < k$.

(3) Strict exogeneity and dynamics. If $x_{i,t}$ contains $y_{i,t-1}$, then $x_{i,t}$ cannot be strictly exogenous. $x_{i,t}$ will be correlated with the unobservables in period $t-1$. Inconsistent OLS estimates!

Pooled Model

Pooled Model

General DGP: $y_{i,t} = x'_{i,t}\beta + c_i + \varepsilon_{i,t},$ & (A2)-(A4) apply.

The pooled model assumes that unobservable characteristics are uncorrelated with $x_{i,t}$ (does not include intercept). We can rewrite the DGP as:

$$y_{i,t} = x'_{i,t}\beta + \nu_i, \quad \text{where} \quad \nu_i = c_i + \varepsilon_{i,t} \text{ (compound error).}$$

To get a consistent estimator of β, we need $E[x'_{i,t}\nu_i] = 0$.

`Note`: $E[x'_{i,t}\varepsilon_{i,t}]$ is derived from (A2) $E[\varepsilon_{i,t}|x_{i,t}, c_i] = 0$. Then, to get consistency, we need $E[x'_{i,t}c_i = 0]$ for all t.

Given these assumptions, we can assume that $c_i = \alpha$ (a constant) independent of i. So, there is no heterogeneity. Then: $y_{i,t} = x'_{i,t}\beta + \alpha + \varepsilon_{i,t}$.

Pooled Model

We have the classical linear model with $k+1$ parameters:

$$y_{i,t} = x'_{i,t}\beta + \alpha + \varepsilon_{i,t}.$$

Estimation is done via OLS, BLUE and consistent.

Stacking the model over i's: $\quad y = X\beta + \alpha\iota + \varepsilon$

y, ι, and ε are $\sum_{i=1}^{N} T_i \times 1$ matrices

X is $\sum_{i=1}^{N} T_i \times k$ matrix

β is a $k \times 1$ matrix

We can re-write the pooled model as: $y = X^*\beta^* + \varepsilon$, with

$X^* = [X \;\; \iota], \quad \sum_{i=1}^{N} T_i \times (k+1)$ matrix

$\beta^* = [\beta \;\; c]', \quad (k+1) \times 1$ matrix

Pooled Model

In this context, OLS produces BLUE and consistent estimators. We refer to this as `pooled OLS estimation`.

Of course, if our assumption regarding the unobservable variables is wrong, we are in the presence of an omitted variable, c_i.

Then, we have potential bias and inconsistency of pooled OLS. The magnitude of these problems depends on how the true model behaves: "fixed" or "random".

Pooled Regression: Heterogeneity Bias

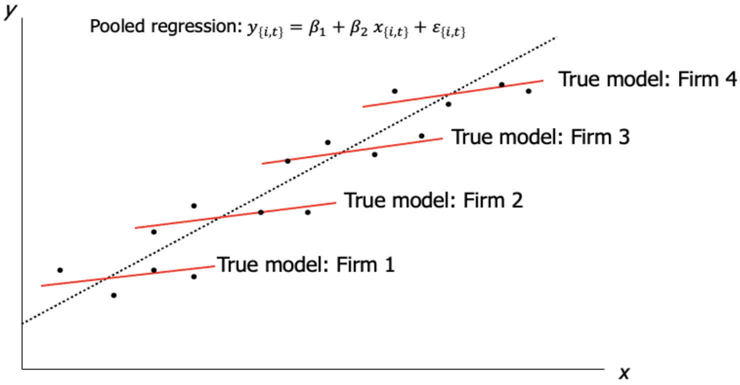

Figure 1: In the pooled regression model, there is no model for group/individual i heterogeneity. Thus, pooled regression may result in heterogeneity bias.

Pooled Regression: Within Transformation

We can estimate β ($\beta_1 = \alpha$) by centering the observations around their group/individual means. That is,

$$y_{it} = \beta_1 + \sum_{j=2}^{k} x_{ij,t}\beta_j + \varepsilon_{i,t}.$$

Subtracting the mean: $y_{it} - \bar{y}_i = \sum_{j=2}^{k}(x_{ij,t} - \bar{x}_{ij})\beta_j + (\varepsilon_{i,t} - \bar{\varepsilon}_i)$.

This method is called the **within-groups estimation** because the model explains the variations about the mean of dependent variable in terms of the variations about the means of explanatory variables for the set of observations relating to a given unit, i.

That is, the estimator reflects the time-series or within-individual i information reflected in the changes within individuals across time. β is estimated using the time-series information in the data.

Pooled Regression: Within Transformation

Disadvantage: There is a cost in the simplicity of the within-groups estimation. The intercept β_1 and any x_{ij} variable that remains constant for each individual (say, gender or College degree) will drop out of the model.

The elimination of the intercept may not matter, but the loss of the unchanging explanatory variables may be frustrating.

Obviously, if we are interested in the effect of gender in CEO compensation, within transformation will not work. But it will work well if we are interested on the effect of an independent Board of Directors, by looking at the compensation pre-/post-BOD.

Pooled Regression: Between Transformation

There is an additional alternative to estimate β, by expressing the model in terms of group/individual means. Model: $y_{i,t} = \beta_1 + \sum_{j=2}^{k} x_{ij,t}\beta_j + \varepsilon_{i,t}$.

Computing the mean:

$$\bar{y}_i = \beta_1 + \sum_{j=2}^{k} \bar{x}_{ij}\beta_j + \bar{\varepsilon}_i.$$

It is called **between estimator** because it relies on variations between individuals (say i & j). We are estimating β using the cross-sectional information in the data (the time-series individual i variation is gone!).

Pooled Regression: Between Transformation

Obviously, if we are interested on the effect of a new independent BOD during the tenure of a CEO on that CEO's compensation, between transformation will not work. But it will work well if we study the effect of gender on CEO's compensation.

Disadvantage: We loose observations (and power!), since we have only N data points.

`Remark`: Under the usual assumptions, pooled OLS using the between transformation is consistent and unbiased.

Useful Analysis of Variance Notation

The variance (total variation) quantifies the idea that each individual i–say, each firm–differs from the overall average. We can decompose the variance into two parts: a within-group/individual part and a between group/individual part:

$$\sum_{i=1}^{N}\sum_{i=1}^{T_i}(z_{it}-\overline{\overline{z}})^2=\sum_{i=1}^{N}\sum_{i=1}^{T_i}(z_{it}-\overline{z}_{i*})^2+\sum_{i=1}^{N}\sum_{i=1}^{T_i}T_i(z_{i*}-\overline{\overline{z}})^2.$$

`Total Variation = Within G. variation + Between G. variation`

`Within group variation`: Measures variation of individuals over time.

`Between group variation`: Measures variation of the means across individuals.

Pooled OLS with First Differences

For the DGP: : $y_{i,t} = x'_{i,t}\beta + c_i + \varepsilon_{i,t}$, & (A2)-(A4) apply.

Even with individual heterogeneity, we can use OLS to estimate β if we eliminate the cause of heterogeneity: c_i.

We can do this by taking first differences of the DGP. That is

$$\begin{aligned} \Delta y_{i,t} = y_{i,t} - y_{i,t-1} &= (x'_{i,t} - x'_{i,t-1})\beta + \Delta c_i + \Delta\varepsilon_{i,t} \\ &= \Delta x'_{i,t}\beta + u_{i,t}. \end{aligned}$$

`Note`: All time invariant variables, including c_i, disappear from the model (one "`diff`"). If the model has a time trend–economic fluctuations–, it also disappear, it becomes the constant term (the other "`diff`"). Thus, the method is usually called "`diffs in diffs`" (DD or DiD).

Pooled OLS with First Differences

With strict exogeneity of (X_i, c_i), the OLS regression of $\Delta y_{i,t}$ on $\Delta x_{i,t}$ is unbiased and consistent, but inefficient.

Why? The error is no longer $\varepsilon_{i,t}$, but $u_{i,t}$. The $\text{Var}(u)$ is given by:

$$\text{Var}\begin{pmatrix} \varepsilon_{i,2}-\varepsilon_{i,1} \\ \varepsilon_{i,3}-\varepsilon_{i,2} \\ \vdots \\ \varepsilon_{i,T_i-1}-\varepsilon_{i,T_i-2} \\ \varepsilon_{i,T_i}-\varepsilon_{i,T_i-1} \end{pmatrix} = \begin{pmatrix} 2\sigma_\varepsilon^2 & -\sigma_\varepsilon^2 & 0 & \dots & 0 \\ -\sigma_\varepsilon^2 & 2\sigma_\varepsilon^2 & -\sigma_\varepsilon^2 & \dots & 0 \\ 0 & -\sigma_\varepsilon^2 & 2\sigma_\varepsilon^2 & \dots & \vdots \\ \vdots & \vdots & \ddots & \dots & -\sigma_\varepsilon^2 \\ 0 & 0 & \dots & -\sigma_\varepsilon^2 & 2\sigma_\varepsilon^2 \end{pmatrix}.$$

That is, first differencing produces heteroscedasticity. Efficient estimation method: GLS, but GLS is complicated. Use OLS in first differences and use Newey-West SE/Panel Corrected SE (PCSE) with one lag.

OLS with First Diffs: Natural Experiment

In Finance & Economics, particularly in Corporate Finance, we apply the DD method when we use `natural experiments` (change in law, policy, or a regulation) to study the effect of x_t on y_t.

We have two periods: before and after the natural experiment (the `treatment`).

If we also have a well-defined control group, where the treatment was not administered–i.e., the natural experiment never occurred–, then, we can use DD estimation.

The number of groups, S, (treated or not treated) under consideration is usually small–typically 2. N is usually very large.

Diffs in Diffs: Natural Experiment 1

Example 1: We are interested in the effect of labor shocks on wages and employments. Natural experiment: The 1980 `Mariel boatlifs`, a temporary lifting of emigration restrictions in Cuba. Most of the `marielitis` (the 1980 Cuban immigrants) settled in Miami.

Two period: Before and after the 1980 Mariel boatlifs. Control group: Low skilled workers in Houston, LA, and Atlanta.

Calculate unemployment and wages of low skilled workers in both periods. Then, regress $\Delta y_{i,t}$ against a set of control variables (industry, education, age, etc.) and a treatment dummy:

$$\Delta y_{i,t} = y_{i,2} - y_{i,1} = \delta_0 + \delta_1 Tr_i + (x_{i,2} - x_{i,1})'\beta + u_{i,t}.$$

$H_0 : \delta_1 = 0$. Card (1990) found no effect of massive immigration.

Diffs in Diffs: Natural Experiment 2

Example 2: Suppose we are interested in the effect of a substantial increase in bank deposits on lending practices. We can use the shale revolution, which started around 2011, as a natural experiment.

Two period: Before and after the shale revolution (say, 2011). Control group: Banks in counties outside shale formation areas.

Measure lending practices (amount lent, FICO scores of loans, etc.), $y_{i,t}$, in both periods & regress $\Delta y_{i,t}$ on a set of control variables (size of county, size of bank, experience of bank employees, etc.) and a treatment dummy:

$$\Delta y_{i,t} = y_{i,2} - y_{i,1} = \delta_0 + \delta_1 Tr_i + (x_{i,2} - x_{i,1})'\beta + u_{i,t}.$$

$H_0 : \delta_1 = 0$. Glije (2011) rejects H_0, especially for counties dominated by small banks.

Application

Grunfeld Investment Study

Investment is modeled as a function of market value, capital, and firm. There are 200 observation on 10 firms from 1935-1954. Some versions of this data have 220 observations with data on the company American Steel.

Variable description is presented below.

invest: Gross investment in 1947 dollars.

value: Market value as of Dec. 31 in 1947 dollars

capital: Stock of plant and equipment in 1947 dollars

firm: General Motors (1), US Steel (2), General Electric (3), Chrysler (4), Atlantic Refining (5), IBM (6), Union Oil (7), Westinghouse (8), Goodyear (9), Diamond Match (10), (American Steel (11), only presented in AER package)

Grunfeld Investment Study

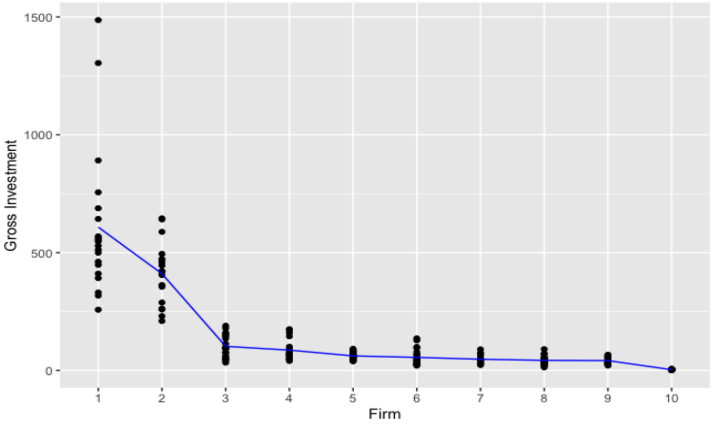

Figure 2: Heterogeneity across firms. The blue line connects the mean values of invest, using all available years across firms.

Grunfeld Investment Study

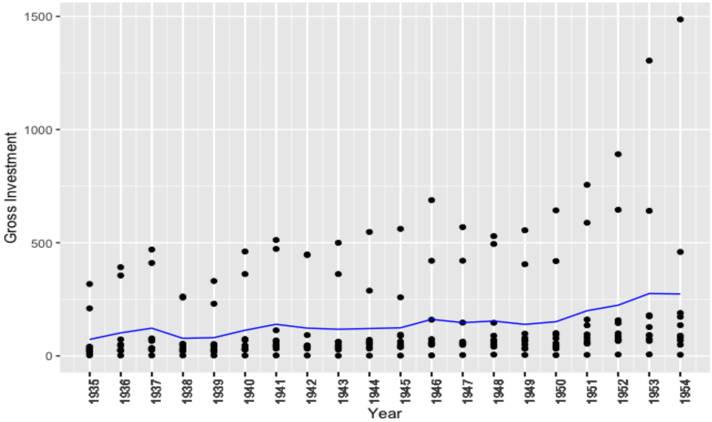

Figure 3: Time heterogeenity. The blue line connects the mean values of invest for all firms across years.

Grunfeld Investment Study

With function `lm()` it is straightforward to estimate the pooled OLS model. Regress `inv` on `capital`.

```
pooled_ols_lm <- lm(inv ~ capital, data = Grunfeld)
summary(pooled_ols_lm)

#Output

Residuals:
Min       1Q      Median      3Q      Max
-316.92  -96.45  -14.43       17.07   481.92
Coefficients:
Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)     14.23620    15.63927    0.91      0.364
capital         0.47722      0.03834   12.45     <2e-16
***---Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 162.9 on 198 degrees of freedom
Multiple R-squared:  0.439,Adjusted R-squared:  0.4362
F-statistic: 154.9 on 1 and 198 DF,  p-value: < 2.2e-16
```

Grunfeld Investment Study

We achieve the same coefficient estimates by using function `plm()` from package `plm`. First, an index has to be supplied, corresponding to the entity and/or time dimension of the panel. The argument `model=` is set to ``‘‘pooling’’``.

```
pooled_ols_plm <- plm(inv ~ capital, data = Grunfeld,
index = c("firm", "year"),
effect = "individual", model = "pooling")
summary(pooled_ols_plm)
```

Grunfeld Investment Study

```
#Output

Call:plm(formula = inv ~ capital, data = Grunfeld, effect = "individual",
model = "pooling", index = c("firm", "year"))

Balanced Panel: n = 10, T = 20, N = 200

Residuals:
Min.    1st Qu.  Median   3rd Qu.   Max.
-316.924   -96.450  -14.429   17.069    481.924

Coefficients:
  Estimate     Std. Error  t-value  Pr(>|t|)
(Intercept)    14.236205   15.639266      0.9103      0.3638
capital        0.477224     0.038339     12.4474     <2e-16
***---Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    9359900
Residual Sum of Squares: 5251000
R-Squared:      0.43899, Adj. R-Squared: 0.43616
F-statistic: 154.937 on 1 and 198 DF, p-value: < 2.22e-16
```

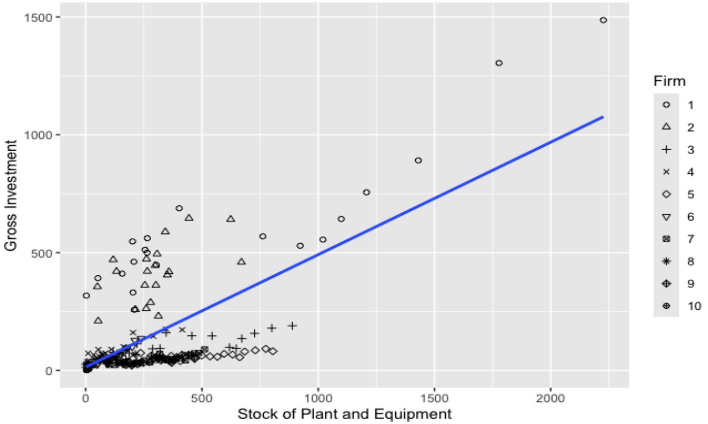

Figure 4: Scatter plot: Although firms could be distinguished by the variable firm, OLS estimation treats all observations as if they come from different entities and fits the regression line accordingly.

Main Models: FEM and REM

There are two main approaches to fitting models using panel data:

(1) Fixed effects regression.

(2) Random effects regressions.

The key difference between these two approaches is how the unobservable characteristics– the `individual effects`–are modeled.

Fixed effects (FE): The individuals are fixed. The differences between them are not interest, only β is interesting. No intent on generalizing the results.

Random effects (RE): The individuals come from a random sample drawn from a larger population, and the variance between them is interesting and can be informative about the larger population.

Fixed Effects Model

Fixed Effects Model (FEM)

The fixed effects model:

(A1) $y_{i,t} = x'_{i,t}\beta + c_i + \varepsilon_{i,t}$, ($x'_{i,t}$ does not include an intercept).

(A2) $E[\varepsilon_{i,t}|X_{i,s}, c_{i,s}] = 0$, for all t, s. (X_i and c_i are strictly exogenous).

The unobserved component, c_i, is arbitrarily correlated with $x_{i,t}$: $E[c_i|X_i] = g(X_i) = \mathrm{constant}_i$, which implies $Cov[x_{i,t}, c_i] \neq 0$.

Note 1: Under the FEM, pooled OLS omits c_i. So, the estimators are biased and inconsistent.

We summarize ("`control for`") these unobservable effects with α_i, a constant. All time invariant characteristics of individual i (location, gender, nationality, etc.) are swept away under this formulation.

Note 2: In a FEM, individuals serve as their own controls.

Estimation with Fixed Effects

Whatever effects the omitted variables have on the individual i at one time, they will also have the same effect at a later time, thus, their effects will be constant, or "`fixed`."

For this, we need the omitted variables to have time-invariant values with time-invariant effects. Typical example, a CEO's IQ/gender. We expect this variable to have the same effect at $t = 1$ or $t = 10$.

As we will see, FEM are estimated using the `within` transformation. Thus, if individuals do not change much (or at all) across time, a FEM may not work very well. We need within-individuals variability in the variables if we are to use individuals as their own controls.

FEM: Estimation via LSDV

Model for individual i is: $y_i = x_i'\beta + c_i + \varepsilon_i$, where all notations are as before and c_i is a $T_i \times 1$ vector. Recall, each individual has T_i observations.

Stacking over all individuals, the model can be expressed as,

$$y = X\beta + c + \varepsilon,$$

where all notations are as before and c, y, and ε are $\sum_i T_i \times 1$ vectors.

The FEM can be represented with a set of dummy/indicator variables:

$$y_{i,t} = x_{i,t}'\beta + \sum_{j=1}^{N} c_j d_{ij,t} + \varepsilon_{i,t}, \qquad \text{with } d_{ij,t} = 1, \text{ if } i = j.$$

FEM: Estimation via LSDV

The FE model assumes $c_i = \alpha_i$ (a constant, it does not vary with t):

$$y_i = X_i\beta + d_i\alpha_i + \varepsilon_i, \qquad \text{for each individual } i.$$

Staking the model, we have

$$\begin{aligned} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} &= \begin{pmatrix} X_1 & d_1 & 0 & \dots & 0 \\ X_2 & 0 & d_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ X_N & 0 & 0 & \dots & d_N \end{pmatrix} \begin{pmatrix} \beta \\ \alpha \end{pmatrix} + \varepsilon \\ &= [X \; D] \begin{pmatrix} \beta \\ \alpha \end{pmatrix} + \varepsilon = Z\delta + \varepsilon. \end{aligned}$$

The FEM is the CLM, but with many $(k + N)$ independent variables. OLS estimators are unbiased, consistent, efficient, but impractical if N is large.

FEM: LSDV Estimator

Avoid dummy variable trap: If a constant is present in the model, the number of dummy variable should be $N - 1$. The omitted individual or group becomes the reference category.

However, the choice of reference category is often arbitrary and thus, the interpretation of the α_i will not be particularly interesting.

Alternatively, we can drop the intercept β_1 and define dummy variables for all of the individuals. This is the more common approach, as done above. The α_i's now become the intercepts for each of the i's.

FEM: LSDV Estimator

If $E[\varepsilon_{i,t}|x_{i,s}, c_i] \neq 0$, then LSDV cannot be used. It is inconsistent. In this case, we need to use IV's. Or a good natural experiment.

Note: Least Squares (LS) is an estimator, not a model. Given the formulation with a lot of dummy variables, this particular LS estimator is called Least Squares Dummy Variable (LSDV) estimator.

FEM: Within Estimation

The OLS estimates of β and α are given by:

$$\begin{pmatrix} \hat{\beta} \\ \hat{\alpha} \end{pmatrix} = \begin{pmatrix} X'X & X'D \\ D'X & D'D \end{pmatrix}^{-1} \begin{pmatrix} X'y \\ D'y \end{pmatrix}$$

Using the Frisch-Waugh theorem: $\hat{\beta} = (X'M_D X)^{-1}(X'M_D y)$. where M_D is explained in the next slide.

In practice, we do not estimate $\{\alpha_i\}$–the c_i's–, they are not very interesting. Moreover, since we are in a `fixed-T` situation, $\hat{\alpha}_{N\times 1}$ is inconsistent. In addition, there is the potential `incidental parameter problem.`

FEM: Within Estimation

$$M_D = \begin{pmatrix} M_D^1 & 0 & \dots & 0 \\ 0 & M_D^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & M_D^N \end{pmatrix} \quad \text{(dummy variables are orthogonal)}$$

$$\begin{aligned} M_D^i &= I_{T_i} - d_i(d_i' d_i)^{-1} d_i = I_{T_i} - (1/T_i) d_i d_i' \\ X' M_D X &= \sum_{i=1}^{N} \sum_{t=1}^{T_i} (x_{i,t} - \bar{x}_i)(x_{i,t} - \bar{x}_i)' \\ X' M_D y &= \sum_{i=1}^{N} \sum_{t=1}^{T_i} (x_{i,t} - \bar{x}_i)(y_{i,t} - \bar{y}_i)' \end{aligned}$$

We subtract the group mean from each individual observation. Individual effects disappear. Now, OLS can easily be used to estimate the β's, using the demeaned data. This is the **within-groups** etimation.

FEM: Within Estimation

The within-groups method estimates the parameters using demeaned data. Since we do not include intercept β_1, we have,

$$y_{i,t} - \bar{y}_i = \sum_{j=2}^{k} (x_{ij,t} - \bar{x}_{ij})\beta_j + \delta^* + (\varepsilon_{i,t} - \bar{\varepsilon}).$$

Recall: It is called `within-groups/individuals` method because it relies on variations `within` individuals rather than `between` individuals.

For the usual asymptotic results, we need:

(A2) $E[\Delta\varepsilon_{i,t}|X_i] = 0$

(A3') $E[\Delta\varepsilon_i'\varepsilon_i|X_i, c_i] = \Sigma$ (different formulations are ok)

(A4) $E[\Delta X_i'\Delta X_i]$ has full rank.

FEM: Within Transformation Removes Effects

There are costs in the simplicity of the within-groups estimation:

(1) All `time invariant` variables (including constant) for each individual i drop out of the model. This eliminates all between-individuals variability (which may be contaminated by omitted variable bias) and leaves only within-subject variability to analyze.

(2) Dependent variables are likely to have smaller variances than in the original specification (measured as deviations from the i mean).

(3) The manipulation involves the loss of N degrees of freedom (we are estimating N means).

FEM: Within Estimation

$\hat{\beta}$ is obtained by `within-groups` least squares (group mean deviations).

Then we use the normal equations to estimate $\hat{\alpha}_{N\times 1}$

$$
\begin{aligned}
& D'X\hat{\beta} + D'D\hat{\alpha} = D'y \\
\hat{\alpha}_{N\times 1} &= (D'D)^{-1}D'(y - X\hat{\beta}), \qquad \text{or,} \\
\hat{\alpha}_i &= \frac{1}{T_i}\sum_{i=1}^{T_i}(y_{i,t} - x'_{i,t}\hat{\beta}) = \bar{\hat{\varepsilon}}.
\end{aligned}
$$

`Note`: This is simple algebra–the estimator is just OLS. Note what $\hat{\alpha}_i$ is when $T_i = 1$. Follow this with $y_{i,t} - \hat{\alpha} - x'_{i,t}\hat{\beta} = 0$ if $T_i = 1$.

FEM: First-Difference (FD) Method

We can eliminate the individual FE using the `first-difference method` i.e., the unobserved effect is eliminated by subtracting observation in period $(t-1)$ from the observation in period t, for all time period:

$$
\begin{aligned}
y_{i,t} - y_{i,t-1} &= \sum_{j=2}^{k} (x_{ij,t} - x_{ij,t-1})\beta_j + (t-(t-1))\delta + \varepsilon_{i,t} - \varepsilon_{i,t-1} \\
\Delta y_{i,t} &= \sum_{j=2}^{k} \Delta x_{ij,t}\beta_j + \delta + \Delta\varepsilon_{i,t}.
\end{aligned}
$$

The error term is now $(\varepsilon_{i,t} - \varepsilon_{i,t-1})$. As before, differencing induces a moving average auto-correlation if $\varepsilon_{i,t}$ satisfies the CLM assumptions.

Note: If $\varepsilon_{i,t}$ is subject to AR(1) auto-correlation and ρ is close to 1, taking the first differences may approximately solve the problem.

FEM: Estimation – FE or FD?

Fixed-effects (or Within) Estimator

– Each variable is demeaned – i.e., subtracted by its average

– Dummy variable regression – i.e., put in a dummy variable for each cross-sectional unit, along with other explanatory variables. This may cause estimation difficulty when N is large.

FD Estimator

– Each variable is differenced once over time, so we are effectively estimating the relationship between change of variables.

FEM: Estimation – FE or FD?

`When` N `is large and` T `is small but greater than 2 (for` $T = 2$`, FE=FD)`

– FE is more efficient when $\varepsilon_{i,t}$ are serially uncorrelated while FD is more efficient when $\varepsilon_{i,t}$ follows a random walk ($\rho = 1$).

`When` T `is large and` N `is small`

– FD has advantage for processes with large positive auto-correlation. (If ρ is near 1, FD solves the nonstationary problem!)

– FE is more sensitive to nonnormality, heteroscedasticity, and serial correlation in $\varepsilon_{i,t}$.

– On the other hand, FE is less sensitive to violation of the strict exogeneity assumption. Then, FE is preferred when the processes are weakly dependent over time.

FEM: Calculation of $Var[\hat{\beta}|X]$

Since we have assumed strict exogeneity: $Cov[\varepsilon_{i,t},(x_{j,t},c_j)]=0$, we have OLS in the CLM. That is,

$$\begin{aligned}
\text{Asy.Var}[\hat{\beta}|X] &= (\sigma_\varepsilon^2/\sum_{i=1}^{N} T_i)\,\texttt{plim}\,[\sigma_\varepsilon^2/\sum_{i=1}^{N} T_i \sum_{i=1}^{N} X_i' M_D^i X_i]^{-1} \qquad \text{where,}\\
\hat{\sigma}_\varepsilon^2 &= \frac{\left(\sum_{i=1}^{N}\sum_{t=1}^{T_i}(y_{it}-\hat{\alpha}_i-x_{it}'\hat{\beta})\right)^2}{\left(\sum_{i=1}^{N} T_i - N - k\right)}.
\end{aligned}$$

PCSE Remark: All previous remarks apply to FEM.

We build the SE according to the type of data we have:

– If we do not suspect auto-correlated errors–not a strange situation–, we can rely on clustered White SE's.

– Otherwise, we use Driscoll and Kraay SE.

FEM: Testing for Fixed Effects

Under $H_0 : \alpha_i = \alpha$ for all i (i.e., no fixed-effects)

−Here, we test whether to pool or not to pool the data.

Different tests:

− F-test based on the LSDV dummy variable model: constant or zero coefficients for D. Test follows an $F_{N-1,NT-N-k}$ distribution.

− F-test based on FEM (the unrestricted model) vs pooled model (the restricted model). Test follows an $F_{N-1,NT-N-k}$ distribution.

An LR can also be done, usually, assuming normality. Test follows a χ^2_{N-1} distribution.

FEM: Hypothesis Testing

Based on estimated residuals of the fixed effects model.

(1) Estimate FEM: $y_{i,t} = x'_{i,t}\beta + \alpha_i + \varepsilon_{i,t}$, and store the residuals $\hat{\varepsilon}_{FE,i,t}$.

(2) Test as usual:

– For heteroscedastcity, use Breusch and Pagan (1980) test.

– For auto-correlation in AR(1) model, use Breusch and Godfrey (1981) test.

Application

Grunfeld Investment Study

(1) Least Squares Dummy Variable Estimation

(a) Fixed Effect Model using the `lm()` function.

With function `lm()` a FE model can be estimated by including dummy variables for all firms. This is the so called least squares dummy variable (LSDV) approach. Similarly to the pooled OLS model, I am regressing `inv` on `capital`. If there is a large number of individuals, the LSDV method is expensive from a computational point of view.

```
fe_model_lm <- lm(inv ~ capital + factor(firm), data = Grunfeld)
summary(fe_model_lm)
```

NOTE: One firm dummy variable is dropped to avoid the dummy variable trap.

```
Call:lm(formula = inv ~ capital + factor(firm), data = Grunfeld)

Residuals:
Min         1Q        Median     3Q         Max
-190.715  -20.835    -0.459     21.383     293.687

Coefficients:
                 Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)     367.61297   18.96710    19.382    < 2e-16***
capital           0.37075    0.01937    19.143    < 2e-16***
factor(firm)2   -66.45535   21.23578    -3.129    0.00203**
factor(firm)3  -413.68214   20.66845   -20.015    < 2e-16***
factor(firm)4  -326.44100   22.54586   -14.479    < 2e-16***
factor(firm)5  -486.27841   20.34373   -23.903    < 2e-16***
factor(firm)6  -350.86559   22.69651   -15.459    < 2e-16***
factor(firm)7  -436.78321   21.11352   -20.687    < 2e-16***
factor(firm)8  -356.47246   22.86642   -15.589    < 2e-16***
factor(firm)9  -436.17028   21.21684   -20.558    < 2e-16***
factor(firm)10 -366.73127   23.64118   -15.512    < 2e-16***
---Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 63.57 on 189 degrees of freedom
Multiple R-squared:  0.9184,Adjusted R-squared:  0.9141
F-statistic: 212.7 on 10 and 189 DF,  p-value: < 2.2e-16
```

Grunfeld Investment Study

(b) Fixed Effect Model using the `lm()` function, excluding the intercept.

Next up, I calculate the same model but drop the constant (intercept) by adding -1 to the formula, so that no coefficient (level) of firm is excluded. Note that this does not alter the coefficient estimate of capital!

```
fe_model_lm_nocons <- lm(inv ~ capital + factor(firm) -1, data = Grunfeld)
summary(fe_model_lm_nocons)
```

```
Call:lm(formula = inv ~ capital + factor(firm) - 1, data = Grunfeld)

Residuals:
Min        1Q       Median   3Q      Max
-190.715  -20.835   -0.459   21.383  293.687

Coefficients:
                 Estimate  Std. Error  t value    Pr(>|t|)
capital           0.37075    0.01937   19.143    < 2e-16***
factor(firm)1   367.61297   18.96710   19.382    < 2e-16***
factor(firm)2   301.15762   15.31806   19.660    < 2e-16***
factor(firm)3   -46.06917   16.18939   -2.846    0.00492**
factor(firm)4    41.17196   14.40645    2.858    0.00474**
factor(firm)5  -118.66544   17.05605   -6.957    5.52e-11***
factor(firm)6    16.74738   14.35657    1.167    0.24487
factor(firm)7   -69.17024   15.46733   -4.472    1.33e-05***
factor(firm)8    11.14050   14.31023    0.778    0.43725
factor(firm)9   -68.55731   15.34015   -4.469    1.35e-05***
factor(firm)10    0.88169   14.21425    0.062    0.95061
---Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 63.57 on 189 degrees of
freedomMultiple R-squared:  0.9439,Adjusted R-squared:  0.9407
F-statistic: 289.3 on 11 and 189 DF,  p-value: < 2.2e-16
```

Grunfeld Investment Study

Scatter Plot

Due to the introduction of firm dummy variables each firm has its own intercept with the y axis! For comparison, I plot the fitted values from the pooled OLS model (blue dashed line). Its slope is more steep compared to the LSDV approach as influential observations of firm 1 lead to an upward bias.

```
ggplot(data = broom::augment(fe_model_lm),
aes(x = capital, y = .fitted)) +
geom_point(aes(color = `factor(firm)`)) +
geom_line(aes(color = `factor(firm)`)) +
geom_line(data=broom::augment(pooled_ols_lm),
aes(x = capital, y =.fitted),
color = "blue", lty="dashed", linewidth = 1) +
labs(x = "Stock of Plant and Equipment", y = "Fitted Values (inv ~ capital)",
color = "Firm")
```

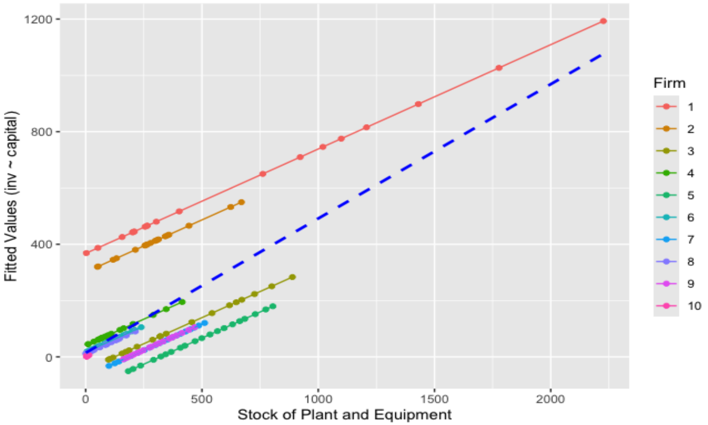

Figure 5: Fitted values from fixed-effects and pooled OLS models.

Grunfeld Investment Study

(2) Within Groups Estimator

(a) Fixed Effects Model using `plm()` function.

The same coefficient estimates as with the LSDV approach can be computed with function `plm()`. The argument `model=` is now set to "`within`". This is the within estimator with n entity-specific intercepts.

```
fe_model_plm <- plm(inv ~ capital, data = Grunfeld,
    index = c("firm", "year"),
    effect = "individual", model = "within")
summary(fe_model_plm)
```

The coefficient of capital indicates how much `inv` changes over time, on average per firm, when `capital` increases by one unit.

Grunfeld Investment Study

```
Call:plm(formula = inv ~ capital, data = Grunfeld, effect = "individual",
model = "within", index = c("firm", "year"))

Balanced Panel: n = 10, T = 20, N = 200

Residuals:
Min.          1st Qu.     Median     3rd Qu.     Max.
-190.71466   -20.83474   -0.45862   21.38262    293.68714

Coefficients:

        Estimate    Std. Error  t-value    Pr(>|t|)
capital 0.370750    0.019368    19.143     < 2.2e-16***

Total Sum of Squares: 2244400, Residual Sum of Squares: 763680
R-Squared: 0.65973, Adj. R-Squared: 0.64173
F-statistic: 366.446 on 1 and 189 DF, p-value: < 2.22e-16
```

Grunfeld Investment Study

```
fixef(fe_model_plm)
```

With function `fixef()` the fixed effects, i.e. the constants for each firm can be extracted. Compare them with the coefficients of the LSDV approach (w/o the constant)- they must be identical.

```
1          2          3          4          5          6          7
367.61297  301.15762  -46.06917  41.17196 -118.66544  16.74738  -69.17024

8          9          10
11.14050  -68.55731   0.88169
```

Testing for Fixed-Effects

With the function `pFtest()` one can test for fixed effects with the null hypothesis that pooled OLS is better than FE.

```
pFtest(fe_model_plm, pooled_ols_plm)

# Output
F test for individual effects
data:  inv ~ capital
F = 123.39, df1 = 9, df2 = 189, p-value < 2.2e-16
alternative hypothesis: significant effects
```

Alternatively, this test can be carried out by jointly assessing the significance of dummy variables in the LSDV approach. The results are identical.

Testing for Fixed-Effects

```
Joint significane test with LSDV approach
car::linearHypothesis(fe_model_lm, hypothesis.matrix =
matchCoefs(fe_model_lm, "firm"))

# Output
Hypothesis:
factor(firm)2 = 0, factor(firm)3 = 0, factor(firm)4 = 0,
factor(firm)5 = 0, factor(firm)6 = 0, factor(firm)7 = 0
factor(firm)8 = 0, factor(firm)9 = 0, factor(firm)10 = 0

Model 1: restricted model
Model 2: inv ~ capital + factor(firm)

Res.Df     RSS    Df     Sum of Sq      F       Pr(>F)
1     198 5250996
2     189 763680   9      4487316      123.39   < 2.2e-16***
---Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In both cases the null hypothesis is rejected in favor of the alternative that there are significant fixed effects.

Grunfeld Investment Study

(3) First Difference Estimator

There is another way of estimating a FE model by specifying `model = "fd"` in the function `plm()`.

```
fe_model_fd<- plm(inv ~ capital -1, data = Grunfeld,
index = c("firm", "year"),
effect = "individual", model = "fd")
summary(fe_model_fd)
```

The coefficient of capital is now different compared to the LSDV approach and within-groups estimator. This is because the coefficients and standard errors of the first-difference model are only identical to the previously obtained results when there are two time periods. For longer time series, both the coefficients and standard errors will be different.

```
Oneway (individual) effect First-Difference Model

Call:plm(formula = inv ~ capital - 1, data = Grunfeld, effect =
"individual",     model = "fd", index = c("firm", "year"))

Balanced Panel: n = 10, T = 20, N = 200
Observations used in estimation: 190

Residuals:
 Min.     1st Qu.   Median     Mean     3rd Qu.    Max.
-240.4  -11.7     0.1         3.5      12.6       333.2

Coefficients:
             Estimate   Std. Error   t-value   Pr(>|t|)
capital      0.230780   0.059639     3.8696    0.00015***
---Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares: 584410
Residual Sum of Squares: 561210
R-Squared: 0.04476, Adj. R-Squared: 0.04476
F-statistic: 14.9739 on 1 and 189 DF, p-value: 0.00014998
```

With Two Periods

Let's verify that the results from LSDV and within approaches are identical to the FD approach when there are two time periods only. Let's drop all years except 1935 and 1936 from the Grunfeld dataset and estimate the model again.

```
# Within estimation (two periods)
fe_model_plm_check <- plm(inv ~ capital, data =Grunfeld,
subset = year %in% c(1935, 1936),
index = c("firm", "year"), effect = "individual", model = "within")
lmtest::coeftest(fe_model_plm_check)

# Output
t test of coefficients:
        Estimate  Std. Error t value Pr(>|t|)
capital  0.91353    0.85333   1.0705   0.3122
```

With Two Periods

Let's verify that the results from LSDV and within approaches are identical to the FD approach when there are two time periods only. Let's drop all years except 1935 and 1936 from the Grunfeld dataset and estimate the model again.

```
# FD estimation (two periods)
fe_model_fd_check<- plm(inv ~ capital -1, data = Grunfeld,
subset = year %in% c(1935, 1936),
index = c("firm", "year"), effect = "individual", model ="fd")
lmtest::coeftest(fe_model_fd_check)

# Output

t test of coefficients:
         Estimate   Std. Error  t value  Pr(>|t|)
capital  0.91353    0.85333     1.0705   0.3122
```

Random Effects Model

The Random Effects Model (REM)

DGP: $y_{i,t} = x_{i,t}'\beta + z_i'\gamma + \varepsilon_{i,t}, \qquad$ – obs. for individual i at time t.

When the observed characteristics are constant for each i (e.g., gender), a FEM is not an effective tool because such variables cannot be included.

An alternative approach, known as random effects model (REM) that, subject to two conditions, provides a solution to this problem.

(1) It is possible to treat each of the unobserved z_p variables as being drawn randomly from a given distribution.

The z_p variables are distributed independently of all the X_i variables, i.e., $E[z_i' X_j] = 0$.

(1) Randomly drawn unobserved Z_p variables

The $c_i = z_i'\gamma$ may be treated as a RV (thus, the name of this approach) drawn from a given distribution. Let, $u_i = c_i - \alpha$, where $E[c_i|X] = \alpha$. Then,

$$
\begin{aligned}
y_{i,t} &= \beta_1(=\alpha) + \sum_{j=2}^{k} x_{ij,t}\beta_j + u_i + \delta t + \varepsilon_{i,t} \\
&= \beta_1(=\alpha) + \sum_{j=2}^{k} x_{ij,t}\beta_j + \delta t + w_{i,t}, \qquad \text{where } \; w_{i,t} = u_i + \varepsilon_{i,t}.
\end{aligned}
$$

We deal with the unobserved effect by subsuming it into a compound disturbance term $w_{i,t}$. We assume that $u_i \sim D(0, \sigma_u^2)$. Then,

$$E[w_{i,t}] = E[u_i] + E[\varepsilon_{i,t}] = 0.$$

The zero mean assumption ($E[u_i] = 0$) is not crucial, since any nonzero component will be absorbed by the intercept β_1.

The Random Effects Model (REM)

(2) z_p is independent of all x_j variables

Otherwise, u_i (& $w_{i,t}$) will not be uncorrelated with x_j. The RE estimation will be biased and inconsistent.

Note: We would have to use the FEM, even if the first condition seems to be satisfied.

If conditions (1) and (2) are satisfied, we can use the REM, and OLS will work, but there is a complication: $w_{i,t}$ is heteroscedastic.

REM: *aka* Error Components Model

Model: $y_{i,t} = x'_{i,t}\beta + u_i + \varepsilon_{i,t} = x'_{i,t}\beta + w_{i,t} \quad (x'_{i,t} \text{ incl. a col. of 1's}).$

REM Assumptions:

$$
\begin{aligned}
E[\varepsilon_{i,t}|X_i] &= 0, \qquad E[\varepsilon_{i,t}^2|X_i] = \sigma_\varepsilon^2 \\
E[u_i|X_i] &= 0, \qquad E[u_i^2|X_i] = \sigma_u^2 \\
E[u_i\varepsilon_{j,t}|X_i] &= 0, \qquad (u \ \& \ \varepsilon \text{ are independent}) \\
E[u_i u_j|X_i] &= 0, \qquad (i \neq j, \text{no cross-correlation of RE}) \\
E[\varepsilon_{i,t}\varepsilon_{j,t}|X_i] &= 0, \qquad (i \neq j, \text{no cross-correlation of errors, } \varepsilon_{i,t}) \\
E[\varepsilon_{i,t}\varepsilon_{j,s}|X_i] &= 0, \qquad (t \neq s, \text{there is no auto-correlation for } \varepsilon_{i,t})
\end{aligned}
$$

$$
\begin{aligned}
\sigma_{w_{it}}^2 &= \sigma_{u_i+\varepsilon_{it}}^2 = \sigma_{u_i}^2 + \sigma_{\varepsilon_{it}}^2 + 2\sigma_{u_i,\varepsilon_{it}} = \sigma_u^2 + \sigma_\varepsilon^2 \\
\sigma_{w_{it1},w_{it2}} &= \sigma_{(u_i+\varepsilon_{it1}),(u_i+\varepsilon_{it2})} = \sigma_u^2.
\end{aligned}
$$

REM: Notation (Greene)

$$
\begin{aligned}
\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} &= \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{pmatrix} \beta + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{pmatrix} + \begin{pmatrix} u_1 \iota_1 \\ u_2 \iota_2 \\ \vdots \\ u_N \iota_N \end{pmatrix} \qquad \begin{matrix} T_1 \text{ obs} \\ T_2 \text{ obs} \\ \vdots \\ T_N \text{ obs} \end{matrix} \\
&= X\beta + \varepsilon + u, \qquad \sum_{i=1}^{N} T_i \text{ observations} \\
&= X\beta + w.
\end{aligned}
$$

In all that follows, except where explicitly noted, X, X_i, and x'_{it} contain a constant term as the first element. To avoid notational clutter, in those cases, x'_{it} etc. will simply denote the counterpart without the constant term. Use of the symbol k for the number of variables will thus be context specific but will usually include the constant term.

REM: Notation (Greene)

$$
\begin{aligned}
Var[\varepsilon_i + u_i\iota] &= \begin{pmatrix} \sigma_\varepsilon^2 + \sigma_u^2 & \sigma_u^2 & \dots & \sigma_u^2 \\ \sigma_u^2 & \sigma_\varepsilon^2 + \sigma_u^2 & \dots & \sigma_u^2 \\ \vdots & \vdots & \ddots & \dots \\ \sigma_u^2 & \sigma_u^2 & \dots & \sigma_\varepsilon^2 + \sigma_u^2 \end{pmatrix} \\
&= \sigma_\varepsilon^2 I_{T_i} + \sigma_u^2 \iota\iota' (T_i \times T_i) \\
&= \sigma_\varepsilon^2 I_{T_i} + \sigma_u^2 \iota\iota' = \Omega_i .
\end{aligned}
$$

$$
Var[w|X] = \begin{pmatrix} \Omega_1 & 0 & \dots & 0 \\ 0 & \Omega_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \dots \\ 0 & 0 & \dots & \Omega_N \end{pmatrix} \quad \text{(differ in the dimension } T_i\text{)}
$$

Note: If $E[\varepsilon_{i,t}, \varepsilon_{j,t}|X_i] = 0$ $(i \neq j)$ or $E[\varepsilon_{i,t}, \varepsilon_{i,s}|X_i] = 0$ $(t \neq s)$, we no longer have this nice diagonal-type structure for $Var[w|X]$.

REM: Estimation via GLS

The standard approach to estimate RE model is via GLS. If σ_u^2 and σ_ε^2 are known, then GLS is most efficient. Standard results for GLS: (1) Consistent and asymptotically normal, (2) Unbiased, and (3) Inefficient.

$$\begin{aligned} \hat{\beta}_{RE} &= \left[X'\Omega^{-1}X\right]^{-1}\left[X'\Omega^{-1}y\right] = \left[\sum_{i=1}^{N} X_i'\Omega_i^{-1}X_i\right]^{-1}\left[\sum_{i=1}^{N} X_i'\Omega_i^{-1}y_i\right] \\ \Omega_i^{-1} &= \frac{1}{\sigma_\varepsilon^2}\left[I_{T_i} - \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_u^2}\iota\iota'\right] \end{aligned}$$

Note that Ω_i^{-1} depends on i only through T_i.

As usual, the matrix $\Omega^{-1/2} = P$ will be used to transform the data.

The matrix $\Omega^{-1/2} = P$ is used to transform the data. That is,

$$\begin{aligned} y_{it} - \theta \bar{y}_i &= (x_{it} - \theta_i \bar{x}_i)'\beta + \nu_{it}, \qquad \text{where} \\ \theta_i &= 1 - \sqrt{\sigma_\varepsilon^2/(\sigma_\varepsilon^2 + T_i \sigma_u^2)} \\ Asy.Var[\hat{\beta}_{GLS}] &= (X'\Omega^{-1}X)^{-1} = \sigma_\varepsilon^2 (X'^{*}X^{*})^{-1}. \end{aligned}$$

We call the transformed data: quasi time-demeaned data. As expected, GLS is just pooled OLS with the transformed data.

Note: The RE can be seen as mixture of two estimators

– when $\theta_i = 0$ ($\sigma_u = 0$) $\rightarrow$ Pooled OLS estimator

– when $\theta_i = 1$ ($\sigma_\varepsilon = 0$ or $\sigma_u \rightarrow \infty$) (LSDV estr., u_i's becomes the FE)

Then, the bigger (smaller) the variance of the unobserved effect–i.e., individual heterogeneity is bigger–, the closer it is to FE (pooled OLS). Also, when T is large, it becomes more like FE.

REM: FGLS - Estimators for the Variances

Naturally, σ_ε^2 and σ_u^2 are unknown, which calls for FGLS to estimate the RE model. We first need to estimate the variance components. Assuming a balanced panel, the usual steps are:

(1) Start with a consistent estimator of β. For example, pooled OLS, $\hat{\beta}$.

(2) Compute $\sum_{i=1}^N \sum_{t=1}^T (y_{i,t} - x_{i,t}'\hat{\beta})^2$ – estimates $\sum_{i=1}^N \sum_{t=1}^T (\sigma_\varepsilon^2 + \sigma_u^2)$.

(3) Divide by a function of NT. For example: $NT - k - 1$

- We estimate σ^2 as, $\hat{\sigma}^2_{pooled} = (\hat{\varepsilon'}_{pooled}\hat{\varepsilon}_{pooled})/(NT - k - 1)$.
- We will use $\hat{\sigma}^2_{pooled}$ to estimate the sum: $\sigma_\varepsilon^2 + \sigma_u^2$

(4) Use LSDV estimation to get $\hat{\alpha}_i$ and $\hat{\beta}_{LSDV}$. Keep residuals $\hat{\epsilon}_{FE,i,t}$.

(5) Compute $\Sigma_i \Sigma_t (y_{i,t} - \hat{\alpha}_i - x_{i,t}'\hat{\beta}_{LSDV})^2$, – estimates $\sum_{i=1}^N \sum_{t=1}^T \sigma_\varepsilon^2$.

(6) To estimate σ_ε^2, divide by $NT - k - N$: $\hat{\sigma}_\varepsilon^2 = \sum_{i=1}^N \sum_{t=1}^T (\hat{\epsilon}_{FE,i,t})^2/(NT - k - N)$.

(7) Estimate σ_u^2 as $\hat{\sigma}_u^2 = \hat{\sigma}^2_{pooled} - \hat{\sigma}_\varepsilon^2$.

REM: FGLS - Estimators for the Variances

Feasible GLS requires (only) consistent estimators of σ_ε^2 and σ_u^2.

Candidates:

(1) From the robust LSDV estimator: $\hat{\sigma}_\varepsilon^2 = \dfrac{\sum_{i=1}^N \sum_{t=1}^{T_i} (y_{it} - \hat{\alpha}_i - x_{it}'\hat{\beta}_{LSDV})^2}{\sum_{i=1}^N T_i - k - N}$.

(2) From the pooled OLS estimator: $\hat{\sigma}_\varepsilon^2 + \hat{\sigma}_u^2 = \dfrac{\sum_{i=1}^N \sum_{t=1}^{T_i} (y_{it} - \hat{\alpha}_{OLS} - x_{it}'\hat{\beta}_{OLS})^2}{\sum_{i=1}^N T_i - k - 1}$.

(3) From the group means regression: $\hat{\sigma}_\varepsilon^2 / \bar{T} + \hat{\sigma}_u^2 = \dfrac{\sum_{i=1}^N (\bar{y}_{it} - \tilde{\alpha} - \bar{x}_i'\tilde{\beta}_{Means})^2}{\sum_{i=1}^N T_i - k - 1}$.

(4) (Wooldridge) $E[w_{it} w_{is} | X_i] = \sigma_u^2$ if $t \neq s$, $\hat{\sigma}_u^2 = \dfrac{\sum_{i=1}^N \sum_{t=1}^{T_i - 1} \sum_{s=t+1}^{T_i} \hat{w}_{it} \hat{w}_{is}}{\sum_{i=1}^N T_i - k - N}$

Note: In the above, $x_{i,t}$ does not contain the intercept.

REM: Practical Problems with FGLS

All of the preceding estimators regularly produce negative estimates of σ_u^2. Estimation is even more complicated in unbalanced panels.

A bulletproof solution (originally used in TSP, now LIMDEP and others):

(1) From robust LSDV estimator: $\hat{\sigma}_\varepsilon^2 = \dfrac{\sum_{i=1}^N \sum_{t=1}^{T_i} (y_{it} - \hat{\alpha}_i - x_{it}'\hat{\beta}_{LSDV})^2}{\sum_{i=1}^N T_i}$.

(2) From pooled OLS estimator: $\hat{\sigma}_\varepsilon^2 + \hat{\sigma}_u^2 = \dfrac{\sum_{i=1}^N \sum_{t=1}^{T_i} (y_{it} - \hat{\alpha}_{OLS} - x_{it}'\hat{\beta}_{OLS})^2}{\sum_{i=1}^N T_i} \geq \hat{\sigma}_\varepsilon^2$.

$$\hat{\sigma}_u^2 = \frac{\sum_{i=1}^N \sum_{t=1}^{T_i} (y_{it} - \hat{\alpha}_{OLS} - x_{it}'\hat{\beta}_{OLS})^2 - \sum_{i=1}^N \sum_{t=1}^{T_i} (y_{it} - \hat{\alpha}_i - x_{it}'\hat{\beta}_{LSDV})^2}{\sum_{i=1}^N T_i}.$$

Bullet proof solution: Do not correct by degrees of freedom. Then, given that the unrestricted RSS (LSDV) will be lower than the restricted (pooled OLS) RSS, $\hat{\sigma}_u^2$ will be positive.

Testing for Random Effects: LM Test

The most common test for the RE model vs a pooled OLS model is an LM test based on Breusch & Pagan (1980). $H_0 : \sigma_u^2 = 0$ (i.e., no common effects). As with all LM test, it is based on the constrained model (here, pooled OLS model). Let $\hat{\varepsilon}$ be pooled OLS residuals.

Breusch and Pagan Lagrange Multiplier statistic: Assuming normality (and for convenience now, a balanced panel)

$$LM = \frac{NT}{2(T-1)}\left[\frac{\sum_{i=1}^{N} T\bar{\hat{\varepsilon}}^2}{\sum_{i=1}^{n} \hat{\varepsilon}_{it}^2} - 1\right]^2 = \frac{NT}{2(T-1)}\left[\frac{\sum_{i=1}^{N} T\bar{\hat{\varepsilon}}^2 - \sum_{i=1}^{n} \hat{\varepsilon}_{it}^2}{\sum_{i=1}^{n} \hat{\varepsilon}_{it}^2}\right]^2 \overset{H_0}{\sim} \chi_1^2.$$

For unbalanced panels, the scale in front becomes $\frac{(\sum_{i=1}^{N} T_i)^2}{2\sum_{i=1}^{N} T_i(T_i-1)}$.

FE vs RE: Understanding Differences

Suppose, we want to study the effect of an MBA on stock trading, controlling for other factors such as income and experience. We have a panel, with individuals measured annually over 10 years. We expect some year-to-year correlation with a given individual i (with unobservable individual-level effects accounting for part of the correlation between yearly trading within the same i).

To understand the difference between FE & RE, we ask:

- Does a regression coefficient for an MBA represent a comparison of two i's, one with an MBA and one without one? $\rightarrow$ Between Effect (RE).

- Or does it compare two yearly trading records from the same i who happened to receive an MBA in the interim? $\rightarrow$ Within Effect (FEM).

FE vs RE

Q: RE estimation or FE estimation?

Case for RE:

– Under no omitted variables –or if the omitted variables are uncorrelated with $x_{i,t}$ in the model– then a REM is probably best: It produces unbiased and efficient estimates, & uses all the data available.

– RE can deal with observed characteristics that remain constant for each individual. In FE, they have to be dropped from model.

– In contrast with FE, RE estimates a small number of parameters – We do not lose N degrees of freedom.

– Philosophically speaking, a REM is more attractive: Why should we assume one set of unobservables fixed and the other random?

FE vs RE

Case against RE:

- If either of the conditions for using RE is violated, we should use FE.

`Condition (1):` Randomly drawn unobserved z_p variables. This is a reasonable assumption in many cases: Many of the panels are designed to be a random sample (for example, NLSY).

But, it would not be a reasonable assumption if the units of observation in the panel data set were data from the S&P 500 firms.

`Condition (2):` z_p is independent of all of the x_j variables.

A violation of condition (2) causes inconsistency in the RE estimation.

FE vs RE

FE estimation is always consistent. On the other hand, a violation of condition (2) causes inconsistency in the RE estimation.

In summary, the FE model always produce consistent estimates (under the null and alternative hypotheses), while the RE model is more efficient under the null hypothesis, but inconsistent under the alternative hypothesis.

To decide between FE and RE, we utilize the Hausman test.

FE vs RE: Hausman Test

The null hypothesis for Hausman test is $H_0 : E(c_i|X_i) = 0$. A Wald statistic is computed as follows:

$$W = [\hat{\beta}_{FE} - \hat{\beta}_{RE}]' \Big[V(\hat{\beta}_{FE}) - V(\hat{\beta}_{RE}) \Big]^{-1} [\hat{\beta}_{FE} - \hat{\beta}_{RE}] \sim \chi^2_{k_{FE}},$$

with $V(\hat{\beta}_{FE}) = \hat{\sigma}^2_{\varepsilon}[X' M_D X]^{-1}$, with $\hat{\sigma}^2_{\varepsilon} = \frac{\left(\sum_{i=1}^{N} \sum_{t=1}^{T_i} (y_{it} - \hat{\alpha}_i - x'_{it}\hat{\beta}_{FE}) \right)^2}{\left(\sum_{i=1}^{N} T_i - N - k \right)}$, and $V(\hat{\beta}_{RE}) = \left[\sum_{i=1}^{N} X' \hat{\Omega}_i X \right]^{-1}$, where $\hat{\beta}_{RE}$ only includes coefficients associated with regressors also included in the FE model (thus no time-invariant elements), and $V(\hat{\beta}_{RE})$ is adjusted accordingly.

Application

Grunfeld Investment Study

The RE model (also called Partial Pooling Model) assumes, in contrast to the FE model, that any variation between entities is random and not correlated with the regressors used in the model. If there are reasons to believe that differences between entities influence the dependent variable, a RE model should be preferred. This also means that time-invariant variables (like a person's gender) can be taken into account as regressors. The entity's error term (unobserved heterogeneity) is hence not correlated with the regressors.

To break down the difference between FE and RE:

(1) the FE model assumes that an individual (entity) specific effect is correlated with the independent variables.

(2) while the RE model assumes that an individual (entity) specific effect is not correlated with the independent variables.

With function `plm()` the RE model can be estimated. The argument `model=` is set to value `"random"`.

```
re_model_plm <- plm(inv ~ capital, data = Grunfeld,
index = c("firm", "year"), effect = "individual", model = "random")
summary(re_model_plm)

# Output
Oneway (individual) effect Random Effect Model

Call:plm(formula = inv ~ capital, data = Grunfeld, effect = "individual",
model = "random", index = c("firm", "year"))

Balanced Panel: n = 10, T = 20, N = 200

Effects:
                var       std.dev   share
idiosyncratic  4040.63    63.57     0.135
individual    25949.52   161.09     0.865
theta: 0.9121
```

Grunfeld Investment Study

```
Residuals:
Min.       1st Qu.    Median   3rd Qu.      Max.
-164.0821  -22.2955   -3.7463   16.9121  319.9564

Coefficients:
               Estimate   Std. Error  z-value   Pr(>|z|)
(Intercept)    43.246697  51.411319   0.8412    0.4002
capital        0.372120   0.019316    19.2652   <2e-16***

---Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Total Sum of Squares: 2299300, Residual Sum of Squares: 799910
R-Squared: 0.65211, Adj. R-Squared: 0.65036
Chisq: 371.149 on 1 DF, p-value: < 2.22e-16
```

The coefficients in the RE model include both the within-entity and between-entity effects. When having data with multiple entities and time periods the coefficient of capital represents the average effect on `investment` when `capital` changes across years and between firms by one unit.

Grunfeld Investment Study

A decision between a fixed and random effects model can be made with the Hausman test, which checks whether the individual error terms are correlated with the regressors. The null hypothesis states that there is no such correlation (RE). The alternative hypothesis is that a correlation exists (FE). The test is implemented in function `phtest()`

```
phtest(fe_model_plm, re_model_plm)

# Output
Hausman Test
data:  inv ~ capitalchisq = 0.93423, df = 1, p-value = 0.3338
alternative hypothesis: one model is inconsistent
```

The null hypothesis cannot be rejected here, hence we should use a RE model.

Thank you!