Short Guides to Microeconometrics
Fall 2025

Kurt Schmidheiny
University of Basel

Panel Data: Fixed and Random Effects

matrix-free

1 Introduction

In panel data, individuals (persons, firms, cities, ...) are observed at several points in time (days, years, before and after treatment, ...). This handout focuses on panels with relatively few time periods (small T) and many individuals (large N).

This handout introduces the two basic models for the analysis of panel data, the fixed effects model and the random effects model, and presents consistent estimators for these two models. The handout does not cover so-called dynamic panel data models.

Panel data are most useful when we suspect that the outcome variable depends on explanatory variables which are not observable but correlated with the observed explanatory variables. If such omitted variables are constant over time, panel data estimators allow to consistently estimate the effect of the observed explanatory variables.

2 The Econometric Model

Consider the multiple linear regression model for individual $i = 1, ..., N$ who is observed at several time periods $t = 1, ..., T$

$$y_{it} = \alpha + \beta_1 x_{it1} + ... + \beta_K x_{itK} + \gamma_1 z_{i1} + ... + \gamma_M z_{iM} + c_i + u_{it}$$

where y_{it} is the dependent variable, $x_{it1}, ..., x_{itK}$ are K time-varying explanatory variables, $z_{i1}, ..., z_{iM}$ are M time-invariant explanatory variables, α, β_k and γ_m are $K+M+1$ parameters, c_i is an *individual-specific effect* and u_{it} is an *idiosyncratic* error term.

We will assume throughout this handout that each individual i is observed in all time periods t. This is a so-called *balanced panel*. The treatment of unbalanced panels is straightforward but tedious.

The T observations for the N individuals are usually stored in the so-called *long format* in statistical software, i.e. there is one row for each individual i and each time period t:

i	t	y	x_1	$\dots$	x_K	z_1	$\dots$	z_M
1	1	y_{11}	x_{111}		x_{11K}	z_{11}		z_{1M}
1	2	y_{12}	x_{121}		x_{12K}	z_{11}		z_{1M}
	$\vdots$							
1	T	y_{1T}	x_{1T1}		x_{1TK}	z_{11}		z_{1M}
2	1	y_{21}	x_{211}		x_{21K}	z_{21}		z_{2M}
	$\vdots$							
2	T	y_{2T}	x_{2T1}		x_{2TK}	z_{21}		z_{2M}
	$\vdots$							
N	1	y_{N1}	x_{N11}		x_{N1K}	z_{N1}		z_{NM}
	$\vdots$							
N	T	y_{NT}	x_{NT1}		x_{NTK}	z_{N1}		z_{NM}

The data generation process (dgp) is described by:

PL1: Linearity

$$y_{it} = \alpha + \beta_1 x_{it1} + ... + \beta_K x_{itK} + \gamma_1 z_{i1} + ... + \gamma_M z_{iM} + c_i + u_{it}$$
where $E[u_{it}] = 0$ and $E[c_i] = 0$

The model is linear in parameters α, β_k, γ_m, effect c_i and error u_{it}.

PL2: Independence

$$\{x_{i11}...x_{iTK}, z_{i1}...z_{iM}, y_{i1}...y_{iT}\}_{i=1}^{N}$$
i.i.d. (independent and identically distributed)

The observations are independent across individuals but not necessarily across time. This is guaranteed by random sampling of individuals.

PL3: Strict Exogeneity

$E[u_{it}|x_{i11}...x_{iTK}, z_{i1}...z_{iM}, c_i] = 0$ (mean independent)

The idiosyncratic error term u_{it} is assumed uncorrelated with the explanatory variables of all past, current and future time periods of the same individual. This is a strong assumption which e.g. rules out lagged dependent variables. *PL3* also assumes that the idiosyncratic error is uncorrelated with the individual specific effect.

PL4: Error Variance

a) $V[u_{it}|x_{i11}...x_{iTK}, z_{i1}...z_{iM}, c_i] = \sigma^2_u > 0$ and $< \infty$
$Corr[u_{it}, u_{is}|x_{i11}...x_{iTK}, z_{i1}...z_{iM}, c_i] = 0$ for all $s \neq t$
(homoscedastic and no serial correlation)

b) $V[u_{it}|x_{i11}...x_{iTK}, z_{i1}...z_{iM}, c_i] = \sigma^2_{u,it} > 0$ and $< \infty$
$Corr[u_{it}, u_{is}|x_{i11}...x_{iTK}, z_{i1}...z_{iM}, c_i] = 0$ for all $s \neq t$
(no serial correlation)

c) $V[u_{it}|x_{i11}...x_{iTK}, z_{i1}...z_{iM}, c_i] = \sigma^2_{u,it} > 0$ and $< \infty$
$Corr[u_{it}, u_{is}|x_{i11}...x_{iTK}, z_{i1}...z_{iM}, c_i] < 1$ and > -1 for all $s \neq t$

The remaining assumptions are divided into two sets of assumptions: the random effects model and the fixed effects model.

2.1 The Random Effects Model

In the random effects model, the individual-specific effect is a random variable that is uncorrelated with the explanatory variables.

RE1: Unrelated effects

$E[c_i|x_{i11}...x_{iTK}, z_{i1}...z_{iM}] = 0$

RE1 assumes that the individual-specific effect is a random variable that is uncorrelated with the explanatory variables of all past, current and future time periods of the same individual.

RE2: Effect Variance

a) $V[c_i|x_{i11}...x_{iTK}, z_{i1}...z_{iM}] = \sigma^2_c < \infty$ (homoscedastic)

b) $V[c_i|x_{i11}...x_{iTK}, z_{i1}...z_{iM}] = \sigma^2_{c,i}(x_{i11}...x_{iTK}, z_{i1}...z_{iM}) < \infty$
(heteroscedastic)

RE2a assumes constant variance of the individual specific effect.

RE3: Identifiability

$(1, x_{it1}, \cdots, x_{itK}, z_{i1}...z_{iM})$ are not linearly dependent and
$0 < V[x_{itk}] < \infty$ and $0 < \widehat{V}[x_{itk}]$

RE3 assumes that the regressors including a constant are not perfectly collinear, that all regressors (but the constant) have non-zero variance and not too many extreme values.

The random effects model can be written as

$$y_{it} = \alpha + \beta_1 x_{it1} + ... + \beta_K x_{itK} + \gamma_1 z_{i1} + ... + \gamma_M z_{iM} + v_{it}$$

where $v_{it} = c_i + u_{it}$. Assuming *PL2*, *PL4* and *RE1* in the special versions *PL4a* and *RE2a* leads to

$$V[v_{it}|x_{i11}...x_{iTK}, z_{i1}...z_{iM}] = \sigma^2_v = \sigma^2_c + \sigma^2_u \text{ for all } i, t$$
$$Cov[v_{it}, v_{is}|x_{i11}...x_{iTK}, z_{i1}...z_{iM}] = \sigma^2_c \text{ for all } i \text{ and } s \neq t$$
$$Cov[v_{it}, v_{js}|x_{i11}...x_{iTK}, x_{j11}...x_{jTK}, z_{i1}...z_{iM}, z_{j1}...z_{jM}] = 0 \text{ for all } s, t, i \neq j$$

This special case under *PL4a* and *RE2a* is therefore called the *equicorrelated random effects model.*

2.2 The Fixed Effects Model

In the fixed effects model, the individual-specific effect is a random variable that is allowed to be correlated with the explanatory variables.

FE1: Related effects

–

FE1 explicitly states the absence of the unrelatedness assumption in *RE1*.

FE2: Effect Variance

–

FE2 explicitly states the absence of the assumption in *RE2*.

FE3: Identifiability

$(\ddot{x}_{it1}, \cdots, \ddot{x}_{itK})$ are not linearly dependent and
$0 < V[\ddot{x}_{itk}] < \infty$ and $0 < \widehat{V}[\ddot{x}_{itk}]$ for all k
where $\ddot{x}_{itk} = x_{itk} - \bar{x}_{ik}$ and $\bar{x}_{ik} = 1/T \sum_t x_{itk}$

FE3 assumes that the time-varying explanatory variables are not perfectly collinear, that they have non-zero within-variance (i.e. variation over time for a given individual) and not too many extreme values. Hence, x_{it} cannot include a constant or any time-invariant variables. Note that only the parameters β but neither α nor γ are identifiable in the fixed effects model.

3 Estimation with Pooled OLS

The *pooled OLS estimator* ignores the panel structure of the data and simply estimates α, β and γ by regressing y_{it} on a constant, on x_{it1}, ..., x_{itK} and on z_{i1}, ..., z_{iM}. In the special case with a constant and one regressor x_{it}, the resulting pooled OLS estimators of α and β are:

$$\widehat{\beta}^{POLS} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \bar{x})(y_{it} - \bar{y})}{\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \bar{x})^2}$$

$$\widehat{\alpha}^{POLS} = \bar{y} - \hat{\beta}_1 \bar{x}$$

where $\bar{y} = 1/NT \sum_i \sum_t y_{it}$ and $\bar{x} = 1/NT \sum_i \sum_t x_{it}$.

Random effects model: The pooled OLS estimator of α, β and γ is unbiased under *PL1*, *PL2*, *PL3*, *RE1*, and *RE3a* in small samples. Additionally assuming *PL4* and normally distributed idiosyncratic and individual-specific errors, it is normally distributed in small samples. It is consistent and approximately normally distributed under *PL1*, *PL2*, *PL3*, *PL4*, *RE1*, and *RE3* in samples with a large number of individuals ($N \to \infty$). However, the pooled OLS estimator is not efficient. More importantly, the usual standard errors of the pooled OLS estimator are incorrect and tests (t-, F-, z-, Wald-) based on them are not valid. Correct standard errors can be estimated with the so-called cluster-robust covariance estimator treating each individual as a cluster. Cluster-robust covariance matrix is consistent when the number of clusters $N \to \infty$. In practice we should have at least 50 clusters (see the handout on "Clustering in the Linear Model").

Fixed effects model: The pooled OLS estimators of α, β and γ are biased and inconsistent, because the variable c_i is omitted and potentially correlated with the other regressors.

4 Random Effects Estimation

The *random effects estimator* is the feasible generalized least squares (GLS) estimator. GLS transforms the data (dependent and explanatory variables) such that the error terms in the transformed model are uncorrelated across all N individuals *and* all time periods T. The GLS is similar to the weighted least squares (WLS) estimator but not easily expressed without matrix notation (see the handout on ”Heteroscedasticity in the linear model”).

The transformation of the model depends on the two unknown parameters σ_v^2 and σ_c^2 only. There are many different ways to estimate these two parameters. For example,

$$\widehat{\sigma}_v^2 = \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\widehat{v}_{it}^2 \quad , \quad \widehat{\sigma}_c^2 = \widehat{\sigma}_v^2 - \widehat{\sigma}_u^2$$

where

$$\widehat{\sigma}_u^2 = \frac{1}{NT - N}\sum_{t=1}^{T}\sum_{i=1}^{N}(\widehat{v}_{it} - \overline{\widehat{v}}_i)^2$$

and $\widehat{v}_{it} = y_{it} - \alpha^{POLS} - \widehat{\beta}_1^{POLS} x_{it1} - ... - \widehat{\beta}_K^{POLS} x_{itK} - \widehat{\gamma}_1^{POLS} z_{i1} - ... -$

$\widehat{\gamma}_M^{POLS} z_{iM}$ and $\overline{\widehat{v}}_i = 1/T \sum_{t=1}^T \widehat{v}_{it}$. The degree of freedom correction in $\widehat{\sigma}_u^2$ is also asymptotically important when $N \to \infty$.

Random effects model: We cannot establish small sample properties for the RE estimator. The RE estimator is consistent and asymptotically normally distributed under *PL1 - PL4*, *RE1*, *RE2* and *RE3* when the number of individuals $N \to \infty$ even if T is fixed. It can therefore be approximated in samples with many individual observations N as

$$\widehat{\alpha}^{RE} \overset{A}{\sim} N\left(\alpha, Avar\left[\widehat{\alpha}^{RE}\right]\right)$$

for all k

$$\widehat{\beta}_k^{RE} \overset{A}{\sim} N\left(\beta_k, Avar\left[\widehat{\beta}_k^{RE}\right]\right)$$

and for all m

$$\widehat{\gamma}_m^{RE} \overset{A}{\sim} N\left(\gamma_m, Avar\left[\widehat{\gamma}_m^{RE}\right]\right)$$

Assuming the equicorrelated model (*PL4a* and *RE2a*), $\widehat{\sigma}_v^2$ and $\widehat{\sigma}_c^2$ are consistent estimators of σ_v^2 and σ_c^2, respectively. Then $\widehat{\alpha}_{RE}$, $\widehat{\beta}_{RE}$ and $\widehat{\gamma}_{RE}$ are asymptotically efficient and the asymptotic variance can be consistently estimated. Allowing for arbitrary conditional variances and for serial correlation of the combined error v_{it} (*PL4c* and *RE2b*), the asymptotic variance can be consistently estimated with the so-called cluster-robust covariance estimator treating each individual as a cluster (see the handout on "Clustering in the Linear Model"). In both cases, the usual tests (z-, Wald-) for large samples can be performed.

In practice, we can rarely be sure about equicorrelated errors and better always use cluster-robust standard errors for the RE estimator.

Fixed effects model: Under the assumptions of the fixed effects model (*FE1*, i.e. *RE1* violated), the random effects estimators of α, β and γ are biased and inconsistent, because the variable c_i is omitted and potentially correlated with the other regressors.

5 Fixed Effects Estimation

Subtracting time averages $\bar{y}_i = 1/T \sum_t y_{it}$ from the initial model

$$y_{it} = \alpha + \beta_1 x_{it1} + ... + \beta_K x_{itK} + \gamma_1 z_{i1} + ... + \gamma_M z_{iM} + c_i + u_{it}$$

yields the *within model*

$$\ddot{y}_{it} = \beta_1 \ddot{x}_{it1} + ... + \beta_K \ddot{x}_{itK} + \ddot{u}_{it}$$

where $\ddot{y}_{it} = y_{it} - \bar{y}_i$, $\ddot{x}_{itk} = x_{itk} - \bar{x}_{ik}$ and $\ddot{u}_{it} = u_{it} - \bar{u}_i$. Note that the individual-specific effect c_i, the intercept α and the time-invariant regressors $z_{i1}, ..., z_{iM}$ cancel.

The *fixed effects estimator* or *within estimator* of the slope coefficient β estimates the within model by OLS In the special case with one time-varying regressor, the resulting FE estimators of β is

$$\widehat{\beta}^{FE} = \frac{\sum_{i=1}^N \sum_{t=1}^T \ddot{x}_{it}\, \ddot{y}_{it}}{\sum_{i=1}^N \sum_{t=1}^T \ddot{x}_{it}^2}$$

The general case with K explanatory variables is derived likewise but not easily expressed without matrix notation. Note that the parameters α and γ are not estimated by the within estimator.

Random effects model and fixed effects model: The fixed effects estimator of β is unbiased under *PL1*, *PL2*, *PL3*, and *FE3* in small samples. Additionally assuming *PL4* and normally distributed idiosyncratic errors, it is normally distributed in small samples. Assuming homoscedastic errors with no serial correlation (*PL4a*), the variance $V[\widehat{\beta}_k^{FE}|x_{111}, ..., x_{NTK}]$ can be unbiasedly estimated with the usual OLS estimator in the transformed model. In the special case with one time-varying regressor, it is estimated as

$$\widehat{V}\left[\widehat{\beta}^{FE}|x_{111}, ..., x_{NTK}\right] = \frac{\widehat{\sigma}_u^2}{\sum_{i=1}^N \sum_{t=1}^T \ddot{x}_{it}^2}$$

where $\widehat{\sigma}_u^2 = \frac{1}{NT-N-K} \sum_i \sum_t \widehat{\ddot{u}}_{it}^2$ and $\widehat{\ddot{u}}_{it} = \ddot{y}_{it} - \widehat{\beta}^{FE} \ddot{x}_{it}$. Note the non-usual degrees of freedom correction. The usual z- and F-tests can be performed.

The FE estimator is consistent and asymptotically normally distributed under *PL1 - PL4* and *FE3* when the number of individuals $N \to \infty$ even if T is fixed. It can therefore be approximated in samples with many individual observations N as

$$\widehat{\beta}_k^{FE} \overset{A}{\sim} N\left(\beta_k, Avar\left[\widehat{\beta}_k^{FE}\right]\right)$$

Assuming homoscedastic errors with no serial correlation (*PL4a*), the asymptotic variance can be consistently estimated as the usual OLS estimator in the transformed model. In the special case with one time-varying regressor, it is estimated as

$$\widehat{Avar}\left[\widehat{\beta}^{FE}\right] = \frac{\widehat{\sigma}_u^2}{\sum_{i=1}^N \sum_{t=1}^T \ddot{x}_{it}^2}$$

where $\widehat{\sigma}_u^2 = \frac{1}{NT-N}\sum_i \sum_t \widehat{\ddot{u}}_{it}^2$.

Allowing for heteroscedasticity and serial correlation of unknown form (*PL4c*), the asymptotic variance $Avar[\widehat{\beta}_k]$ can be consistently estimated with the so-called cluster-robust covariance estimator treating each individual as a cluster (see the handout on “Clustering in the Linear Model”). In both cases, the usual tests (z-, Wald-) for large samples can be performed.

In practice, the idiosyncratic errors are often serially correlated (violating *PL4a*) when $T > 2$. Bertrand, Duflo and Mullainathan (2004) show that the usual standard errors of the fixed effects estimator are drastically understated in the presence of serial correlation. It is therefore advisable to always use cluster-robust standard errors for the fixed effects estimator.

6 Random Effects vs. Fixed Effects Estimation

The random effects model can be consistently estimated by both the RE estimator or the FE estimator. We would prefer the RE estimator if we can be sure that the individual-specific effect really is an unrelated effect (*RE1*). This is usually tested by a (Durbin-Wu-)Hausman test. However, the Hausman test is only valid under homoscedasticity and cannot include time fixed effects.

The unrelatedness assumption (*RE1*) is better tested by running an auxiliary regression (Wooldridge 2010, p. 332, eq. 10.88, Mundlak, 1978):

$$y_{it} = \alpha + \beta_1 x_{it1} + ... + \beta_K x_{itK} + \gamma_1 z_{i1} + ... + \gamma_M z_{iM} + \lambda_1 \overline{x}_{i1} + ... + \lambda_K \overline{x}_{iK} + \delta_t + u_{it}$$

where $\overline{x}_{ik} = 1/T \sum_t x_{itk}$ for $k = 1, ..., K$ are the time averages of all time-varying regressors. Include time fixed δ_t if they are included in the RE and FE estimation. A joint Wald-test (or F-test) on H_0: $\lambda_k = 0$ for all $k = 1, .., K$ tests *RE1*. Use cluster-robust standard errors to allow for heteroscedasticity and serial correlation.

Note: Assumption *RE1* is an extremely strong assumption and the FE estimator is almost always much more convincing than the RE estimator. Not rejecting *RE1* does not mean accepting it. Interest in the effect of a time-invariant variable is no sufficient reason to use the RE estimator.

7 Least Squares Dummy Variables Estimator (LSDV)

The least squares dummy variables (LSDV) estimator is pooled OLS including a set of $N - 1$ dummy variables which identify the individuals and hence an additional $N - 1$ parameters. Note that one of the individual dummies is dropped because we include a constant. Time-invariant explanatory variables, $z_{i1}, ..., z_{iM}$, are dropped because they are perfectly collinear with the individual dummy variables.

The LSDV estimators of $\beta_1, ..., \beta_K$ are numerically identical with the FE estimators and therefore consistent under the same assumptions. The LSDV estimators of the additional parameters for the individual-specific dummy variables, however, are inconsistent as the number of parameters goes to infinity as $N \to \infty$. This so-called *incidental parameters* problem generally biases all parameters in *non-linear* fixed effects models like the probit model.

8 First Difference Estimator

Subtracting the lagged value $y_{i,t-1}$ from the initial model

$$y_{it} = \alpha + \beta_1 x_{it1} + ... + \beta_K x_{itK} + \gamma_1 z_{i1} + ... + \gamma_M z_{iM} + c_i + u_{it}$$

yields the *first-difference model*

$$\dot{y}_{it} = \beta_1 \dot{x}_{it1} + ... + \beta_K \dot{x}_{itK} + \dot{u}_{it}$$

where $\dot{y}_{it} = y_{it} - y_{i,t-1}$, $\dot{x}_{it} = x_{it} - x_{i,t-1}$ and $\dot{u}_{it} = u_{it} - u_{i,t-1}$. Note that the individual-specific effect c_i, the intercept α and the time-invariant regressors $z_{i1}, ..., z_{iM}$ cancel. The *first-difference estimator* (FD) of the slope coefficients $\beta_1, ..., \beta_K$ estimates the first-difference model by OLS. In the special case with one time-varying regressor, the resulting FE estimators of β is

$$\widehat{\beta}^{FD} = \frac{\sum_{i=1}^{N} \sum_{t=2}^{T} \dot{x}_{it}\, \dot{y}_{it}}{\sum_{i=1}^{N} \sum_{t=2}^{T} \dot{x}_{it}^2}$$

The general case with K explanatory variables is derived likewise but not easily expressed without matrix notation. Note that the parameters α and γ are not estimated by the FD estimator. In the special case $T = 2$, the FD estimator is numerically identical to the FE estimator.

Random effects model and fixed effects model: The FD estimator is a consistent estimator of β under the same assumptions as the FE estimator. It is less efficient than the FE estimator if u_{it} is not serially correlated (*PL4a*).

9 Fixed Effects vs. First Difference Estimation

Given the fixed effects model (*PL1*, *PL2*, *PL3*, *FE3*), both the fixed effects and the first difference estimator of β are consistent. Hence, the two estimators should be similar in large samples. In practice, however, the two estimator often differ substantially. The reason for this is typically a misspecification of the timing in the linear model. *PL1* assumes that changes in x_{it} have only an instantaneous effect on y_{it} at time t. In practice, effects often need several periods to materialize. Such patterns are called *dynamic treatment effects*. In this situation, the first difference estimator will only pick up the instantaneous effect at time t while the fixed effects estimator picks up an average of the dynamic treatment effects.

10 Time Fixed Effects

We often also suspect that there are time-specific effects δ_t which affect all individuals in the same way

$$y_{it} = \alpha + \beta_1 x_{it1} + ... + \beta_K x_{itK} + \gamma_1 z_{i1} + ... + \gamma_M z_{iM} + \delta_t + c_i + u_{it}\,.$$

We can estimate this extended model by including a dummy variable for $T-1$ time periods with one period serving as the reference period. Assuming a fixed number of time periods T and the number of individuals $N \to \infty$, both the RE estimator and the FE estimator are consistent using time dummy variables under above conditions. Estimation with both individual fixed effects and time fixed effects is called *two-way fixed effects* estimation.

11 Heterogeneous Effects

PL1 assumes that the parameters β_k are constant across individuals i and time t. However, in reality, effects likely differ across i and t, i.e. the effects are heterogeneous and researchers seek to estimate an average treatment effect $ATE_k = E[\beta_{itk}]$. Unfortunately, the linear panel estimators discussed in this handout $\widehat{\beta}_k$ are in general not unbiased estimators for ATE_k (see e.g. de Chaisemartin and D'Haultfœuille, 2020).

An exception is the two-way fixed effects estimation in a panel with two time periods $t = 1, 2$ with a dependent variable y_{it} and a single explanatory variable d_{it} which takes the value $d_{i2} = 1$ if an individual i is treated in period 2 and $d_{it} = 0$ otherwise: $y_{it} = \beta_0 + \beta_1 d_{it} + \delta_t + c_i + u_{it}$ with $\delta_1 = 0$. In this case, the two-way fixed effects estimator is equivalent to the average first difference $\Delta y_{i2} = y_{i2} - y_{i1}$ in the treated group minus the average difference in the control group (differences-in-differences estimator). $\widehat{\beta}_1$ can be interpreted as the average treatment effect on the treated ($ATET$) even if the individual effects β_{it1} are heterogeneous provided that the expected change from period 1 to 2 in the treated group would have been identical to the expected change in the control group (common trends assumption).

12 Implementation in Stata 17

Stata provides a series of commands that are especially designed for panel data. See `help xt` for an overview.

Stata requires panel data in the so-called *long form*: there is one line for every individual and every time observation. The very powerful Stata command `reshape` helps transforming data into this format. Before working with panel data commands, we have to tell Stata the variables that identify the individual and the time period. For example, load data and define individuals (variable *idcode*) and time periods (variable *year*)

```
webuse nlswork.dta
xtset idcode year
```

Stata provides descriptive statistics for panel data with the commands

```
xtdescribe
xtsum
```

The pooled OLS estimator with corrected standard errors is calculated with the standard ols command `regress`:

```
generate ttl_exp2 = ttl_exp^2
reg ln_wage grade ttl_exp ttl_exp2, vce(cluster idcode)
```

where the `vce` option was used to report correct cluster-robust standard errors. This command multiplies $\widehat{Avar}$ with $(NT-1)/(NT-M-K-1) \cdot N/(N-1)$ as a small sample correction and uses $N-1$ degrees of freedom for t- and F-tests.

The random effects estimator is calculated by the Stata command `xtreg` with the option `re`:

```
xtreg ln_wage grade ttl_exp ttl_exp2, re
```

Stata reports asymptotic z- and Wald-tests with random effects estimation. Cluster-robust standard errors are reported with:

```
xtreg ln_wage grade ttl_exp ttl_exp2, re vce(cluster idcode)
```

Since version 10, Stata always assumes clustering with robust standard errors in random and fixed effects estimations. So we could also just use

```
xtreg ln_wage grade ttl_exp ttl_exp2, re vce(robust)
```

The fixed effects estimator is calculated by the Stata command `xtreg` with the option `fe`:

```
xtreg ln_wage ttl_exp ttl_exp2, fe
```

Note that the effect of time-constant variables like *grade* is not identified by the fixed effects estimator. The parameter reported as *_cons* in the Stata output is the average fixed effect $1/N\sum_i c_i$. This command uses $NT-N-K$ degrees of freedom for t- and F-tests. Cluster-robust standard errors are reported with the `vce` option:

```
xtreg ln_wage ttl_exp ttl_exp2, fe vce(cluster idcode)
```

This command multiplies $\widehat{Avar}$ with $(NT-1)/(NT-N-K)\cdot N/(N-1)$ as a small correction and reports reports cluster-robust t- and F-tests with $N-1$ degrees of freedom. The latter is particularly useful with large T (see Stock and Watson, 2008).

The Hausman test is calculated by

```
xtreg ln_wage grade ttl_exp ttl_exp2, re
estimates store b_re
xtreg ln_wage ttl_exp ttl_exp2, fe
estimates store b_fe
hausman b_fe b_re, sigmamore
```

and the auxiliary regression version by

```
regress ln_wage grade ttl_exp ttl_exp2
tegen ttl_exp_mean = mean(ttl_exp) if e(sample), by(idcode)
egen ttl_exp2_mean = mean(ttl_exp2) if e(sample), by(idcode)
regress ln_wage grade ttl_exp ttl_exp2 ///
    ttl_exp_mean ttl_exp2_mean, vce(cluster idcode)
test ttl_exp_mean ttl_exp2_mean
```

Note that the time averages are generate with the sample used in both the random effects and the pooled OLS estimation.

13 Implementation in R 4.4.3

The R package `plm` provides a series of functions and data structures that are especially designed for panel data.

The `plm` package works with data stored in a dataframe in the so-called *long form.* Long form data means that there is one line for every individual and every time observation. For example, load data

```
library(haven)
nlswork <- read_dta("https://www.stata-press.com/data/r17/nlswork.dta")
```

where individuals are defined by *idcode* and time periods by *year*.

Pooled OLS with cluster-robust standard errors can be estimated with a standard regression and the packages `lmtest` and `sandwich`

```
pols1 <- lm(ln_wage~grade+ttl_exp+I(ttl_exp^2), data = nlswork)
library(lmtest)
library(sandwich)
coeftest(pols1, vcov = vcovCL, cluster = ~idcode)
```

This command multiplies $\widehat{Avar}$ with $(NT-1)/(NT-M-K-1)\cdot N/(N-1)$ as a small sample correction.

Alternatively, pooled OLS with corrected standard errors is estimated by the package `plm` with the function `plm` and its model option `pooling`:

```
library(plm)
pols2 <- plm(ln_wage~grade+ttl_exp+I(ttl_exp^2), model="pooling",
             data = nlswork, index=c("idcode", "year"))
summary(pols2)
coeftest(pols2, vcov=vcovHC(pols2, cluster="group", type="HC1"))
```

where `coeftest` reports cluster-robust standard errors. `cluster="group"` defines the clusters by the individual identifier set by the option `index` in plm, i.e. the variable `idcode` in the example. This command multiplies $\widehat{Avar}$ with $(NT-1)/(NT-M-K-1)$ but not with $N/(N-1)$ and uses $NT-M-K-1$ degrees of freedom for t- and F-tests.

The random effects estimator is calculated by `plm` option `random`:

```
re <- plm(ln_wage~grade+ttl_exp+I(ttl_exp^2), model="random",
          data = nlswork, index=c("idcode", "year"))
summary(re)
```

Cluster-robust standard errors are reported with

```
coeftest(re, vcov=vcovHC(re, cluster="group", type="HC1"))
```

The fixed effects estimator is calculated by `plm` option `within`

```
fe <- plm(ln_wage ~ grade + ttl_exp + I(ttl_exp^2), model="within",
      data=nlswork, index=c("idcode", "year"))
summary(fe)
```

Note that effects of time-constant variables like `grade` are not identified by the fixed effects estimator. This command uses $NT-N-K$ degrees of freedom for t- and F-tests. Cluster-robust standard errors are given by:

```
coeftest(fe, vcov=vcovHC(fe, cluster="group", type="HC1"))
```

This command multiplies $\widehat{Avar}$ with $(NT-1)/(NT-K-1)$ as a small sample correction and uses $NT-K-1$ degrees of freedom for t- and F-tests.

The Hausman test is calculated by estimating RE and FE and then comparing the estimates:

```
phtest(fe, re)
```

and the auxiliary regression version by

```
vars <- c("idcode", "year", "ln_wage", "grade", "ttl_exp")
sample <- nlswork[complete.cases(nlswork[,vars]),vars]
sample$ttl_exp_mean <- ave(sample$ttl_exp, sample$idcode, FUN = mean)
sample$ttl_exp2_mean <- ave(sample$ttl_exp^2, sample$idcode, FUN = mean)
aux <- plm(ln_wage~grade+ttl_exp+I(ttl_exp^2)+ttl_exp_mean+ttl_exp2_mean,
          model="pooling", data = sample, index=c("idcode", "year"))
summary(aux)
waldtest(aux, .~.-ttl_exp_mean-ttl_exp2_mean,
         vcov=vcovHC(aux, cluster="group", type="HC1"))
```

Note that the dataset was reduced to the sample used in the random effects and the pooled OLS estimation before generating the time averages.

References

Introductory textbooks

Stock, James H. and Mark W. Watson (2020), Introduction to Econometrics, 4th Global ed., Pearson. Chapter 10.

Wooldridge, Jeffrey M. (2009), Introductory Econometrics: A Modern Approach, 4th ed., South-Western Cengage Learning. Ch. 13 and 14.

Angrist, Joshua D. and Jörn-Steffen Pischke (2009), Mostly Harmless Econometrics: An Empiricist's Companion, Princeton University Press. Chapter 5.

Advanced textbooks

Cameron, A. Colin and Pravin K. Trivedi (2005), Microeconometrics: Methods and Applications, Cambridge University Press. Chapter 21.

Wooldridge, Jeffrey M. (2010), Econometric Analysis of Cross Section and Panel Data, MIT Press. Chapter 10.

Articles

Manuel Arellano (1987), Computing Robust Standard Errors for Within-Group Estimators, Oxford Bulletin of Economics and Statistics, 49, 431–434.

Bertrand, M., E. Duflo and S. Mullainathan (2004), How Much Should We Trust Differences-in-Differences Estimates?, Quarterly Journal of Economics, 119(1), 249–275.

de Chaisemartin, Clément and Xavier D'Haultfœuille (2020), Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects, American Economic Review 2020, 110(9), 2964–2996.

Mundlak, Y. (1978), On the pooling of time series and cross section data, Econometrica, 46, 69–85.

Stock, James H. and Mark W. Watson (2008), Heteroskedasticity-Robust Standard Errors for Fixed Effects Panel Data Regression, Econometrica, 76(1), 155–174. [advanced]