SESSION 2010

Economics Department

Economic Data Analysis

Notes on Panel Data Analysis

"*I see that you are an enthusiast of this new science. Would you care to try another word? Trash.*"

"*Why not? It doesn't matter that you're a skeptic. Not in the least. What was it again, trash? Very well ... trash, trashcan, ashcan, trashman. Trashmass, trashmic, catatrashmic. Trashmass, trashmosh. In a large enough scale, trashmos. And-of course - macrotrashm! Tichy, you come up with the best words! Really, just think of it, macrotrashm!*"

"*I'm afraid I don't follow. It's nonsense to me.*" ...

"*Secondly, macrotrashm is nonsense so far, yet we can already guess its sense-to-be, its future significance. The word observe, implies nothing less than a new psychozoic theory! Implies that the stars are of artificial origin!*"

"*Now where do you get that?*"

"*From the word itself. Macrotrasm indicates, or rather suggests, this image: in the course of many eons the Universe filled up with trash, the wastes of various civilizaitons. The wastes got in the way, of course, hampering astronomers and cosmonauts, and so enormous incinerators were built, all at extremely high temperatures, observe, to burn the trash, and with sufficient mass to pull it in from space themselves. Gradually space clears up and behold, there are your stars, those selfsame furnaces, and the dark nebulae-this is the trash that remains to be removed.*"

"*You can't be serious! The Universe nothing but one big trash disposal? You don't really think that's possible? Professor!*" (The Futurological Congress, Lem, 1974)

Contents

1 Independently Pooled Cross Sections

- If a random sample from a population is obtained at different points in time we obtain an *independently pooled cross section*.
- To account for the possibility of differently distributed populations in different years, time dummies are usually included.

1.1 Policy Analysis with Pooled Cross Sections

Independently pooled cross sections can be used to analyze so-called *natural experiments*. A natural experiment occurs when there is an exogenous source of variation in the explanatory variables, i.e., when, say because of a policy change, the environment in which individuals, households, firms, countries, etc., operate, is changed. A natural experiment will always determine a *control group*, not affected by the policy change, and a *treatment group*, which is considered affected by the change. This occurrence is particularly useful wherever estimates seem particularly susceptible to omitted variable bias. For instance, consider the impact on house prices of the construction of a new incinerator. It is very likely that the incinerator will be build in areas where the house prices are already low, and are likely to remain low after the construction of the incinerator. A standard OLS regression would incorrectly attribute this difference in prices to the effect of the incinerator, and hence bias the estimated coefficient.

1.2 Difference–in–Difference (DD) Estimator

DD estimation consists of identifying a policy intervention (treatment) and then compare the difference in outcomes after and before the intervention for the groups affected by it to this difference for unaffected groups. In order to control for preexisting heterogeneity between control and treatment groups, at least two years of data are required, one before the policy intervention ad one after.

1.3 The Case of an Incinerator

1.3.1 Data

Interactive R example

```
line 1
line 2
line 3
line 4
line 5
line 6
line 7
```

- Kiel and McClain (1995): "House Prices During Siting Decisions Stages: The Case of an Incinerator from Rumor Through Operation," *Journal of Environmental Economics and Management*, 28, 241–255.

```
year      age       agesq     nbh       cbd       intst     lintst    price
rooms     area      land      baths     dist      ldist     wind      lprice
y81       larea     lland     y81ldist  lintstsq  nearinc   y81nrinc  rprice
lrprice

  Obs:   321

  1. year                     1978 or 1981
```

```
 2. age                         age of house
 3. agesq                       age^2
 4. nbh                         neighborhood #, 1 to 6
 5. cbd                         dist. to central bus. dstrct, feet
 6. intst                       dist. to interstate, feet
 7. lintst                      log(intst)
 8. price                       selling price
 9. rooms                       # rooms in house
10. area                        square footage of house
11. land                        square footage lot
12. baths                       # bathrooms
13. dist                        dist. from house to incinerator, feet
14. ldist                       log(dist)
15. wind                        perc. time wind incin. to house
16. lprice                      log(price)
17. y81                         =1 if year == 1981
18. larea                       log(area)
19. lland                       log(land)
20. y81ldist                    y81*ldist
21. lintstsq                    lintst^2
22. nearinc                     =1 if dist <= 15840
23. y81nrinc                    y81*nearinc
24. rprice                      price, 1978 dollars
```

```
25. lrprice                    log(rprice)
```

1.3.2 Estimations

- Consider the model for 1981

$$rprice_{i,1981} = \alpha_1 + \alpha_2\ nearinc_{i,1981} + \epsilon_{i,1981}$$

where $nearinc$ is a dummy defined as

$$nearinc = \begin{cases} 1, & \text{if } distance \leq 15840 \text{ feet;} \\ 0, & \text{otherwise.} \end{cases}$$

- What is the interpretation of the OLS estimators $\widehat{\alpha}_1$ and $\widehat{\alpha}_2$?

- $\widehat{\alpha}_1$ estimates the 1981 average price of all houses distant from the incinerator, $\widehat{\alpha}_1 + \widehat{\alpha}_2$ estimates the 1981 average price of all houses near to the incinerator,

$$\widehat{\alpha}_1 = \overline{rprice}_{1981,nearinc=0},$$
$$\widehat{\alpha}_1 + \widehat{\alpha}_2 = \overline{rprice}_{1981,nearinc=1},$$

and hence α_2 estimates the 1981 difference in average price between houses that are distant and those that are close to the incinerator,[1]

$$\widehat{\alpha}_2 = \overline{rprice}_{1981,nearinc=1} - \overline{rprice}_{1981,nearinc=0}.$$

[1] When regressing rprice on a constant, OLS yields the sample average of $rprice$ as the fitted coefficient. OLS will return the value of α_1 that

$$\min_{\alpha_1} S(\alpha_1) = \sum_{i=1}^{n} (rprice_i - \alpha_1)^2,$$

i.e, the solution to

$$\frac{\partial S(\alpha_1)}{\partial \alpha_1} = \sum_{i=1}^{n} -2(rprice_i - \alpha_1) = 0,$$

Interactive R example

```
dd <- read.table("e:/statmeth/KIELMC.dat", header=T)
> names(dd)
 [1] "year"     "age"      "agesq"    "nbh"      "cbd"      "intst"    "lintst"   "price"
 [9] "rooms"    "area"     "land"     "baths"    "dist"     "ldist"    "wind"     "lprice"
[17] "y81"      "larea"    "lland"    "y81ldist" "lintstsq" "nearinc"  "y81nrinc" "rprice"
[25] "lrprice"
> attach(dd)

```

$$\sum_{i=1}^{n} rprice_i = \sum_{i=1}^{n} \alpha_1,$$

$$\sum_{i=1}^{n} rprice_i = n\alpha_1,$$

$$\alpha_1 = \frac{1}{n}\sum_{i=1}^{n} rprice_i = \overline{rprice}.$$

If we add the incinerator indicator, *nearinc*, the minimization becomes

$$\min_{\alpha_1,\alpha_2} S(\alpha_1,\alpha_2) = \left\{\sum_{i=1}^{n}(rprice_i - \alpha_1 - \alpha_2 nearinc)^2\right\},$$

which can be rewritten as

$$\min_{\{\alpha_1,\alpha_2\}} S(\alpha_1,\alpha_2) = \left\{\sum_{\{i|nearinc=0\}}(rprice_i - \alpha_1)^2 + \sum_{\{i|nearinc=1\}}(rprice_i - \alpha_1 - \alpha_2 nearinc)^2\right\} =$$

$$= \left\{\min_{\{\alpha_1,\alpha_2\}} \sum_{\{i|nearinc=0\}}(rprice_i - \alpha_1)^2\right\} + \left\{\min_{\{\alpha_1,\alpha_2\}} \sum_{\{i|nearinc=1\}}(rprice_i - \alpha_1 - \alpha_2)^2\right\} =$$

$$= \left\{\min_{\{\alpha_1,\alpha_2\}} \sum_{\{i|nearinc=0\}}(rprice_i - \alpha_1)^2\right\} + \left\{\min_{\{\alpha_1,\alpha_2\}} \sum_{\{i|nearinc=1\}}(rprice_i - \gamma)^2\right\}$$

so that the fitted value of α_1 is the average of prices for homes close to the incinerator only, the fitted value of $\alpha_1 + \alpha_2 = \gamma$ is the average of home prices far from the incinerator, and, therefore the fitted value for α_2 is the difference in home's average prices for distant and the average for close to the incinerator.

```
> lm( rprice ~ nearinc, data = subset(dd, year==1981) )

Call:
lm(formula = rprice ~ nearinc, data = subset(dd, year == 1981))

Coefficients:
(Intercept)      nearinc
     101308       -30688
```

Estimation Results

(i) The fitted regression line is

$$\widehat{rprice} = \underset{(32.754)}{101307.5145} - \underset{(-5.266)}{30688.27376}\, nearinc$$

(ii) The negative coefficient for *nearinc* suggests, as expected from "theory," that being close to the incinerator lowers house values.

(iii) We estimate that being close to the incinerator determines a fall of \$30688.27376 in the average selling price of homes. Figure ?? illustrates the regression results.

Model for 1978

$$rprice_{i,1978} = \beta_1 + \beta_2 \, nearinc_{i,1978} + \epsilon_{i,1978}$$

- What is the interpretation of the OLS estimators $\widehat{\beta}_1$ and $\widehat{\beta}_2$?
- Analogously to the 1981 case,

$$\widehat{\beta}_1 = \overline{rprice}_{1978,nearinc=0},$$

$$\widehat{\beta}_1 + \widehat{\beta}_2 = \overline{rprice}_{1978,nearinc=1},$$

  and hence

$$\widehat{\beta}_2 = \overline{rprice}_{1978,nearinc=1} - \overline{rprice}_{1978,nearinc=0}.$$

Interactive R example

```
> lm( rprice ~ nearinc, data = subset(dd, year==1978) )

Call:
lm(formula = rprice ~ nearinc, data = subset(dd, year == 1978))

Coefficients:
(Intercept)      nearinc
      82517       -18824
```

Estimation Results

(i) The fitted regression line is

$$\widehat{rprice} = \underset{(31.094)}{82517.22764} - \underset{(-3.968)}{18824.37050}\, nearinc$$

(ii) The negative coefficient for *nearinc* suggests that being close to the "to be build" incinerator has a negative impact on home values. How can we interpret this findings?

(iii) We estimate that being close to the "to be constructed" incinerator determines a fall of $18824.37050 in the average selling price of homes.

1.3.3 Economic Significance

The impact of the construction of the incinerator on values of homes is given by

$$\widehat{\delta} = \widehat{\alpha}_2 - \widehat{\beta}_2 = -30688.27376 - (-18824.37050) = -11863.90 dollars$$

which is also known as the *difference-in-differences estimator* as it can be expressed as a difference of differences

$$\widehat{\delta} = \left(\overline{rprice}_{1981,nearinc=1} - \overline{rprice}_{1981,nearinc=0}\right) - \left(\overline{rprice}_{1978,nearinc=1} - \overline{rprice}_{1978,nearinc=0}\right)$$

1.3.4 Statistical Significance

To test whether the impact is significantly different from zero, H_0: $\delta = 0$, we need the standard error (square root of the variance) of $\widehat{\delta}$. We can estimate the following pooled model

$$rprice = \gamma_1 + \gamma_2\, y81 + \gamma_3\, nearinc + \delta\, (y81 \cdot nearinc) + \nu$$

$$\frac{\partial rprice}{\partial nearinc} = \begin{cases} \gamma_3 + \delta, & \text{if } year = 81; \\ \gamma_3, & \text{if } year = 78. \end{cases}$$

$$intercept = \begin{cases} \gamma_1 + \delta, & \text{if } year = 81; \\ \gamma_3, & \text{if } year = 78. \end{cases}$$

Table 1: Summary of impacts

	$nearinc = 1$	$nearinc = 0$
year=78	$\gamma_1 + \gamma_3$	γ_1
year=81	$\gamma_1 + \gamma_2 + \gamma_3 + \delta$	$\gamma_1 + \gamma_2$

$$rprice = \gamma_1 + \gamma_2\, y81 + \gamma_3\, nearinc + \delta\,(y81 \cdot nearinc) + \nu$$

$$\frac{\partial rprice}{\partial nearinc} = \begin{cases} \gamma_3 + \delta\, y81, & \text{if } nearinc = 1; \\ \gamma_1 + \gamma_2 y81, & \text{if } nearinc = 0. \end{cases}$$

Table 2: Summary of impacts

	$nearinc = 1$	$nearinc = 0$
year=78	$\gamma_1 + \gamma_3$	γ_1
year=81	$\gamma_1 + \gamma_2 + \gamma_3 + \delta$	$\gamma_1 + \gamma_2$

Note that (from Table 2):

- γ_1 captures the house prices for houses far from the incinerator in 1978.
- γ_2 captures the change in house prices for all houses from 1978 to 1981.
- γ_3 captures the effect of the location of the house not due to the presence of the incinerator.

Omitting any of the two dummies can bias δ.

Interactive R example

```
> lm( rprice ~ y81 + nearinc + y81nrinc, data = dd )

Call:
lm(formula = rprice ~ y81 + nearinc + y81nrinc, data = dd)

Coefficients:
(Intercept)          y81      nearinc     y81nrinc
      82517        18790       -18824       -11864

```

2 Panel Data

We will reproduce the results from the following paper:

- Stern D. I. and Mick S. Common (2001), Is there an environmental Kuznets curve for sulfur?, *Journal of Environmental Economics and Management*, 40(2).

For the dataset used by Common and Stern the first column is the time index (years from 1960 to 1990) whereas the second column contains the individual country index (there are 74 countries in the sample). The correspondence between codes and countries is provided in Table ??. The third column contains the populations, the fourth SO_2 emissions. The fifth is the GDP in real 1990 international dollars. The sixth column contains the SO_2 concentration per capita, and the last column a OECD/non-OECD dummy. The emission data comes from ASL and Associates; GDP and population is taken from the Penn World Table.

```
1960    54  17910   1099.72 7258    0.0614  1
1961    54  18270   1076.06 7261    0.0589  1
1962    54  18614   1073.68 7605    0.05768 1
1963    54  18963   1087.53 7876    0.05735 1
1964    54  19326   1142.22 8244    0.0591  1
1965    54  19678   1206.56 8664    0.06132 1
1966    54  20049   1174.17 9093    0.05857 1
1967    54  20411   1304.04 9231    0.06389 1
1968    54  20744   1328.19 9582    0.06403 1
...
```

height1	ALGERIA	95	JAPAN
14	EGYPT	97	KOREA,
18	GHANA	98	KUWAIT
22	KENYA	100	MALAYSIA
25	MADAGASCAR	102	MYANMAR
30	MOROCCO	106	PHILIPPINES
31	MOZAMBIQUE	108	SAUDI ARABIA
32	NAMIBIA	109	SINGAPORE
34	NIGERIA	110	SRI LANKA
41	SAFRICA	111	SYRIA
44	TANZANIA	112	TAIWAN
46	TUNISIA	113	THAILAND
48	ZAIRE	116	AUSTRIA
49	ZAMBIA	117	BELGIUM
50	ZIMBABWE	119	CYPRUS
52	BARBADOS	120	CZECHOSLOVAKIA
54	CANADA	121	DENMARK
60	GUATEMALA	122	FINLAND
62	HONDURAS	123	FRANCE
64	MEXICO	125	WGERMANY
65	NICARAGUA	126	GREECE
71	TRINIDAD&TOBAGO	129	IRELAND
72	U.S.A.	130	ITALY
73	ARGENTINA	131	LUXEMBOURG
74	BOLIVIA	133	NETHERLANDS
75	BRAZIL	134	NORWAY
76	CHILE	136	PORTUGAL
77	COLOMBIA	137	ROMANIA
81	PERU	138	SPAIN
83	URUGUAY	139	SWEDEN

2.1 Advantages of Panel Data

- more informative data that can provide more reliable estimates
- allow to estimate and test more complex models, models with less restrictive assumptions and therefore more realistic
- can control for individual heterogeneity and unobservable or missing variables

2.2 Disadvantages of Panel Data

- problems with design and data collection
  - coverage
  - non-response
  - recall
- measurement errors
  - unclear survey questions
  - memory errors
  - deliberate distorsions
- sample selection problems

3 Accounting for Unobserved Effects with Panel Methods

Consider looking at the relationship between emissions and growth.

- Individual countries have many unique characteristics that are difficult to quantify, yet we might wish to include them in the set of variables that determine pollution to avoid the curse of omitted variables bias (which in this instance is known as heterogeneity bias).
- These characteristics include aspects of geography, history, preferences, and natural resource endowments that are constant over the years in which we observe a cross section of countries, i.e., time-invariant.
- In a cross-section of countries we cannot condition on such characteristics without quantifying them.

In general, we can classify the unobserved factors that affect the dependent variable into three types.

1. Time-invariant, those that are constant over time and different for each individual (preferences)
2. Individual-invariant, those that vary over time but are constant for each individual (OIL shocks)
3. Individual and time variant, those that very both with time and individuals (trade)

   Country and time effects (fixed effects) Apart from the growth, trade and structural change variables, there are a number of country specific factors that influence energy requirements. Examples of such factors are resource endowments, climate, geographical location and culture. These aspects of a country either do not change or change very slowly over time. Following earlier studies we control for these factors by including country specific dummy variables. In addition to these we also include a dummy variable for each year. This allows us to control for factors that evolve over time and impact all countries, for example, world energy prices and technological developments. (Suri and Chapman, 1998)

4 General Form of Panel Model

- The general form of the panel data model used in this study is given by the equation

$$Y_{it} = \mu + \sum_{j=1}^{k} \beta_j \, X_{j_{it}} + u_{it}, \tag{1}$$

  with $i = 1, \ldots, N$ and $t = 1, \ldots, T_i$ representing groups and types respectively. k the total number of regressors.
- The error component, u_{it}, in Equation 2, can take different structures.
- The specification of error components can depend solely on the contry/individual, one-way error component, to which the observation belongs or both on the country and year, two-way error component.
- If the specification depends on country/individual, then we have $u_{it} = \alpha_i + \epsilon_{it}$. The term α_i is intended to capture the heterogeneity across countries and ϵ_{it} is the classical error term with zero mean and a constant variance.
- Moreover, the individual effects, α_i, can be assumed to be either *fixed* or *random*.
  - If assumed fixed, the α_is can be estimated by including a dummy variable for each country i,The N α_is can then be estimated by ordinary least squares
  - If random the α_is are assumed to be IID with mean zero and homoskedastic covariance matrix, σ_α, and independent of the ϵ_{it}.
- The nature of the error structures leads to different estimation procedures depending on the specification. The latter assumptions will be tested in order to select the appropriate model. For further details on panel methods see Verbeek (2008) or Greene (2008).

4.1 Unobserved Individual–Specific Effects

Panel methods allow to account for the effects of any combination of omitted variables that remain constant over time. The general model we are interested in estimating is a single equation with individual effects of the form

$$Y_{it} = \mu + \sum_{j=1}^{K} \beta_j \, X_{j\,it} + \underbrace{\alpha_i + \epsilon_{it}}_{\text{one-wayerror component}} \,, \tag{2}$$

with $i = 1, \ldots, N,\ t = 1, \ldots, T_i$.

1. $X_{1it}, \ldots, X_{Kit}$ are K regressors (independent variables).
2. The α_i capture the effects of those variables that are specific to the i*th* individual and that are constant over time (time invariant).
3. $\epsilon_{it} \sim IID(0, \sigma_\epsilon^2)$.
4. For *balanced panels* $T_i = T$. To keep notation simple we will assume that the panels we consider are balanced.

There are two main approaches to estimate models like (??), the *fixed effects model* (FEM) and the *random effects model* (REM).

4.1.1 Fixed Effects Model

In the fixed effects model the individual-specific dummies are treated as fixed constants. A practical implementation of the fixed effects model is by augmenting the standard regression model with dummy variables

$$y_{it} = \alpha_1 D_{1it} + \alpha_2 D_{2it} + \cdots + \beta_1\, x_{it1} + \cdots + \beta_K\, x_{itK} + \epsilon_{it} \tag{3}$$

where D_{jit} is an individual-specific dummy variable defined as

$$D_{jit} = \begin{cases} 1, & \text{if } i = j \text{ ;} \\ 0, & \text{otherwise.} \end{cases}$$

- Note that there is no general intercept α in the model to avoid the dummy variable trap.
- This estimator for the β's goes under the name of LSDV, **Least Squares with Dummy Variables**, estimator.
- This model can be estimated by OLS after creating N, D_i dummies.
- A computationally simpler approach consists of regressing $(y_{it} - y_{i.})$ on $(x_{it} - x_{i.})$, using OLS with no constant, where $y_{i.}$ is the individual-specific mean of y and $x_{i.}$ are the N individual-specific means of x_{it}.
- The transformation that produces observations in **deviations from individual means** is called the *within* transformation.
- The individual-specific intercepts are recovered by calculating $\widehat{\alpha}_i = y_{i.} - x_{i.1}\widehat{\beta}_1^{FE} - \cdots - x_{i.K}\widehat{\beta}_K^{FE}$.

4.1.2 Random Effects Model

If random the α_is are assumed to be IID with mean zero and homoskedastic covariance matrix, σ_α, and independent of the ϵ_{it}. The REM can be represented by the equation

$$y_{it} = \mu + \beta_1\, x_{it1} + \cdots + \beta_K\, x_{itK} + \alpha_i + \epsilon_{it} \tag{4}$$

with

1. $E[\alpha_i] = 0$; $\mathrm{var}[\alpha_i] = \sigma^2_\alpha$.
2. $\mathrm{cov}[\epsilon_{it}, \alpha_i] = 0$.
3. $\mathrm{var}[\epsilon_{it} + \alpha_i] = \sigma^2 = \sigma^2_\epsilon + \sigma^2_\alpha$.
4. $\mathrm{corr}[\epsilon_{it} + \alpha_i, \epsilon_{is} + \alpha_i] = \rho = \frac{\sigma^2_u}{\sigma^2}$, for $s \neq t$.

- Equation 4 includes a general intercept α. The dummy variable trap is avoided by assuming that the the expectation individual-specific errors, α_i, is zero.
- The individual-specific effect is now represented as a stochastic component α_i, of the same type as the error term ϵ_{it}.
- The REM requires is estimated by GLS (Generalized Least Squares). Basically, the estimation involves using OLS to regress $(y_{it} - \theta y_{i.})$ on $(1 - \theta)$ and $(x_{it} - \theta x_{i.})$ where $(1 - \theta)$ corresponds to the constant term. The GLS estimate of the REM model is equivalent to applying the OLS method to the original data transformed by removing a fraction θ of the individual-specific means, $y_{i.}$ and $x_{i.}$ (instead of the whole means as with the within transformation).
- θ is calculated as $1 - \vartheta$ where $\vartheta = \frac{\sigma^2_\epsilon}{\sigma^2_\epsilon + T\sigma^2_\alpha}$.

4.2 Unobserved Time–Specific Effects

In many cases we would like to include time-specific effects in our models

$$Y_{it} = \mu + \sum_{j=1}^{K} \beta_j \, X_{j\,it} + \underbrace{\alpha_i + \lambda_t + \epsilon_{it}}_{\text{two-way error component}}$$

with $i = 1, \ldots, N, \; t = 1, \ldots, T_i$ and where α_i and λ_t are respectively individual and time specific effects.

4.3 Unobserved Time and Individual–Varying Effects

If no candidate variable such as openness to trade is available, we might have to resort to lagged dependent variables.

$$y_{it} = \gamma_1 y_{i(t-1)} + \beta_1 \, x_{it1} + \cdots + \beta_k \, x_{itk} + \lambda_t + \alpha_i + \epsilon_{it}$$

with $i = 1, \ldots, N, \; t = q + 1, \ldots, T_i$, where α_i and λ_t are respectively individual and time specific effects.

4.4 First difference (FD) Estimator

One approach to account for individual heterogeneity is simply to get rid of it by first differencing.

$$y_{it} = \beta\ x_{it} + \alpha_i + \epsilon_{it} \tag{5}$$

$$y_{i(t-1)} = \beta\ x_{i(t-1)} + \alpha_i + \epsilon_{i(t-1)} \tag{6}$$

Subtracting (6) from (5), we get

$$y_{it} - y_{i(t-1)} = \beta(\ x_{it} - x_{i(t-1)}) + (\epsilon_{it} - \epsilon_{i(t-1)})$$

$$\Delta y_{it} = \beta \Delta x_{it} + \Delta \epsilon_{it} \tag{7}$$

Assuming that $\Delta\epsilon_{it}$ are uncorrelated with Δx_{it1}, we can estimate (7) by OLS. This is called that first-differenced estimator.

5 Model Specification

5.1 OLS Vs. Individual Effects

The (Breusch and Pagan) LM-test is used to test the hypothesis that individual effects are significant. If these tests do not provide any evidence for individual effects, then the model can simply be estimated by ordinary least squares (OLS). A large value of the LM statistic argues in favor of the of a panel data model

5.2 Fixed Vs. Random Effects

- The estimated parameters can differ quite substantially if T is small and N large.
- So if the number of individual, say countries, is small and we are interested in the α_i, it makes sense to prefer the FE estimator.
- With large population, if interested in making inference about the population, the RE estimator might be more appropriate.
- Even in the case of a large population of individual countries, we might still opt for the FE estimator.
  - If the individual effects, α_i, are correlated with the regressors. In this case the RE estimator is inconsistent.
  - The Hausman test is used to decide whether the regressors are correlated with the individual effect. A small Hausman statistic argues in favor of the random effect model, a large in favor of a fixed effects one.

  Sometimes, the RE might be preferable, even when the individual effects are correlated with the regressors.

  - The FE estimator allows α_i and x_{it} to be arbitrarily correlated by eliminating, as in the FD estimator, the individual effects and all time-invariant effects. We cannot therefore include in our model variables that are constant over time, e.g., in the EKC example, OECD/non OECD dummy.

6 Reading Data

Interactive R example

```
stern.dat <- read.table("e:/jan/stern2.dat",header=T)
attach(stern.dat)

> names(stern.dat)
[1] "year"    "country" "pop"     "so"      "gdppc"   "sopc"    "oe"

> head(stern.dat)
  year country   pop      so gdppc    sopc oe
1 1960      54 17910 1099.72  7258 0.06140  1
2 1961      54 18270 1076.06  7261 0.05890  1
3 1962      54 18614 1073.68  7605 0.05768  1
4 1963      54 18963 1087.53  7876 0.05735  1
5 1964      54 19326 1142.22  8244 0.05910  1
6 1965      54 19678 1206.56  8664 0.06132  1

```

7 Descriptive Statistics

We can now summarise the data in many ways.

Interactive R example

```
> summary(stern.dat)
      year         country             pop                 so                gdppc     
 Min.   :1960   Min.   :  1.00   Min.   :    231   Min.   :    0.01   Min.   :  303  
 1st Qu.:1967   1st Qu.: 62.00   1st Qu.:   5063   1st Qu.:   14.67   1st Qu.: 1548  
 Median :1975   Median : 94.50   Median :  12764   Median :  101.22   Median : 3566  
 Mean   :1975   Mean   : 90.69   Mean   :  47466   Mean   :  703.03   Mean   : 5360  
 3rd Qu.:1983   3rd Qu.:123.00   3rd Qu.:  32979   3rd Qu.:  448.50   3rd Qu.: 7728  
 Max.   :1990   Max.   :147.00   Max.   :1133683   Max.   :14213.89   Max.   :80831  
      sopc                  oe        
 Min.   :8.900e-07   Min.   :0.0000  
 1st Qu.:1.960e-03   1st Qu.:0.0000  
 Median :9.673e-03   Median :0.0000  
 Mean   :2.150e-02   Mean   :0.3108  
 3rd Qu.:2.780e-02   3rd Qu.:1.0000  
 Max.   :4.656e-01   Max.   :1.0000  
```

Interactive R example

```
  aggregate(stern.dat, by = list( oe ),  FUN= "mean" )
  Group.1 year   country       pop         so     gdppc       sopc oe
1       0 1975  75.58824 54261.85  545.0277 3614.351 0.02061154  0
2       1 1975 124.17391 32397.30 1053.3722 9230.491 0.02347759  1

```

8 Creating and Transforming Variables

For the whole sample create squared and log terms.

Interactive R example

```
lso <- log(sopc)
lgdp <- log(gdppc)
lgdp2 <- lgdp^2
```

9 Plotting Data

Interactive R example

```
plot(lgdp,lso)
```

The resultiong LimDep plot is shown in figure 1.

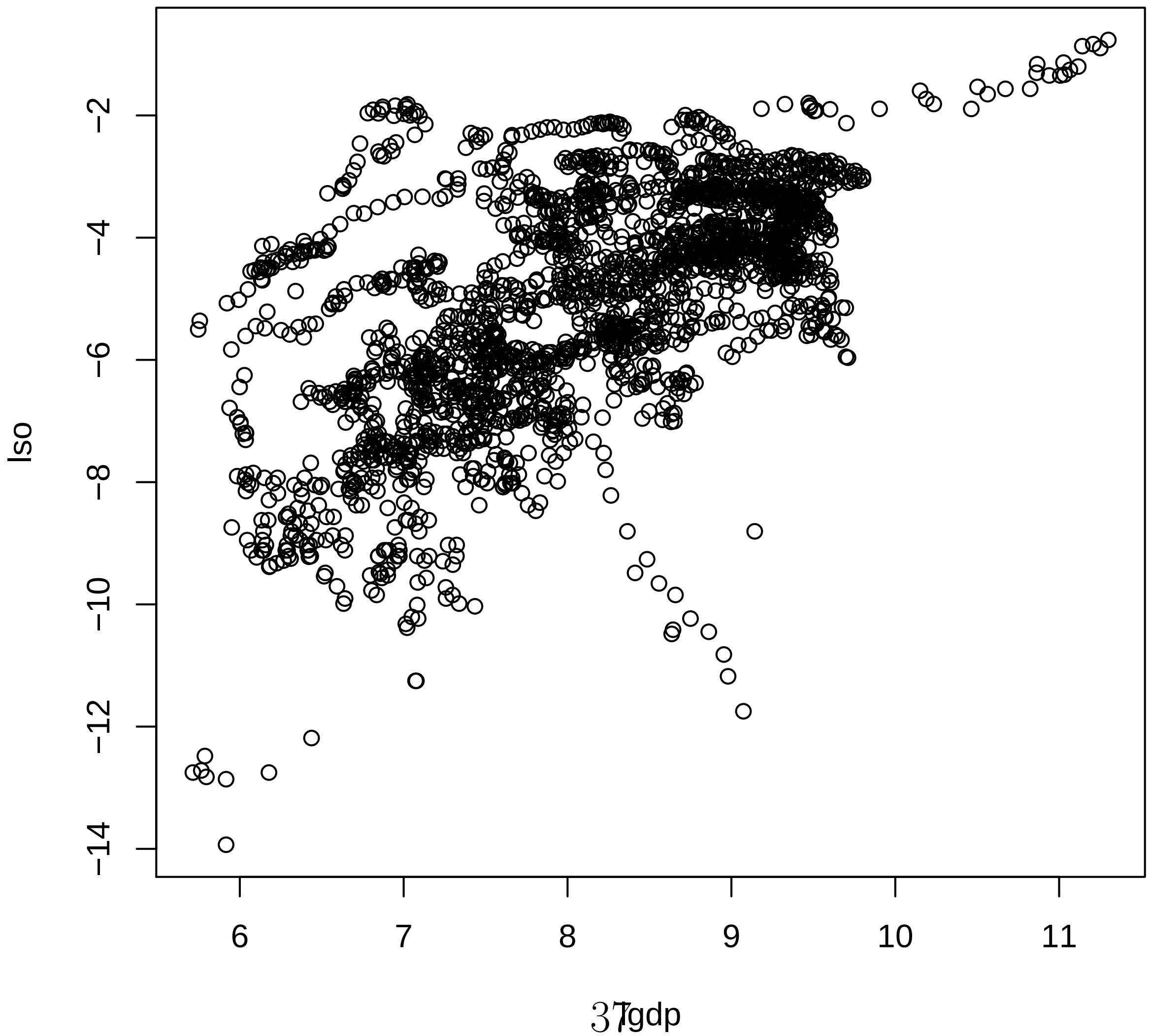


Figure 1: R plot

10 Running Panel Regressions in R

Interactive R example

```
library(plm)

stern.fra <- data.frame(
    cbind(
    country = country,
    year=year,
    oe=oe,
    lso = log(sopc),
    lgdp = log(gdppc),
    lgdp2 = lgdp^2
    )
)

stern.plm <- plm.data(stern.fra, index = c("country","year"))

```

10.1 OLS

Interactive R example

```
> stern.ols <- plm(lso ~ lgdp  + lgdp2, data=stern.plm, model = "pooling")
> summary(stern.ols)
Oneway (individual) effect Pooling Model

Call:
plm(formula = lso ~ lgdp + lgdp2, data = stern.plm, model = "pooling")

Balanced Panel: n=74, T=31, N=2294

Residuals :
    Min.  1st Qu.   Median  3rd Qu.     Max.
-7.80000 -0.85100 -0.00892  0.83200  4.65000

Coefficients :
            Estimate Std. Error t-value Pr(>|t|)
(Intercept) -17.0282     1.7525   -9.72  < 2e-16 ***
lgdp          1.8209     0.4380    4.16  3.3e-05 ***
lgdp2        -0.0418     0.0271   -1.54     0.12
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Total Sum of Squares:    8110
```

```
Residual Sum of Squares: 5090
F-statistic: 679.914 on 2 and 2291 DF, p-value: <2e-16

```

10.2 Fixed (within) Individual Country Effects

Interactive R example

```
> stern.fe <- plm( lso ~ lgdp  + lgdp2, data=stern.plm,
   model = "within", effect = "individual")
> summary(stern.fe)
Oneway (individual) effect Within Model

Call:
plm(formula = lso ~ lgdp + lgdp2, data = stern.plm, effect = "individual",
    model = "within")

Balanced Panel: n=74, T=31, N=2294

Residuals :
   Min. 1st Qu.  Median 3rd Qu.    Max.
-4.3300 -0.1860  0.0291  0.2360  3.2900

Coefficients :
      Estimate Std. Error t-value Pr(>|t|)
lgdp    3.2528     0.3533    9.21  < 2e-16 ***
lgdp2  -0.1525     0.0213   -7.16  1.1e-12 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

```

```
Total Sum of Squares:    888
Residual Sum of Squares: 756
F-statistic: 192.867 on 2 and 2218 DF, p-value: <2e-16
```

10.3 Accessing Fixed Country Individual Effects

The estimated individual effects can be retrieved using the command *fixef(stern.fe)*. The option *dmean* produces differences from the mean. This is useful to contrast the fixed effects later with the random effects.

Interactive R example

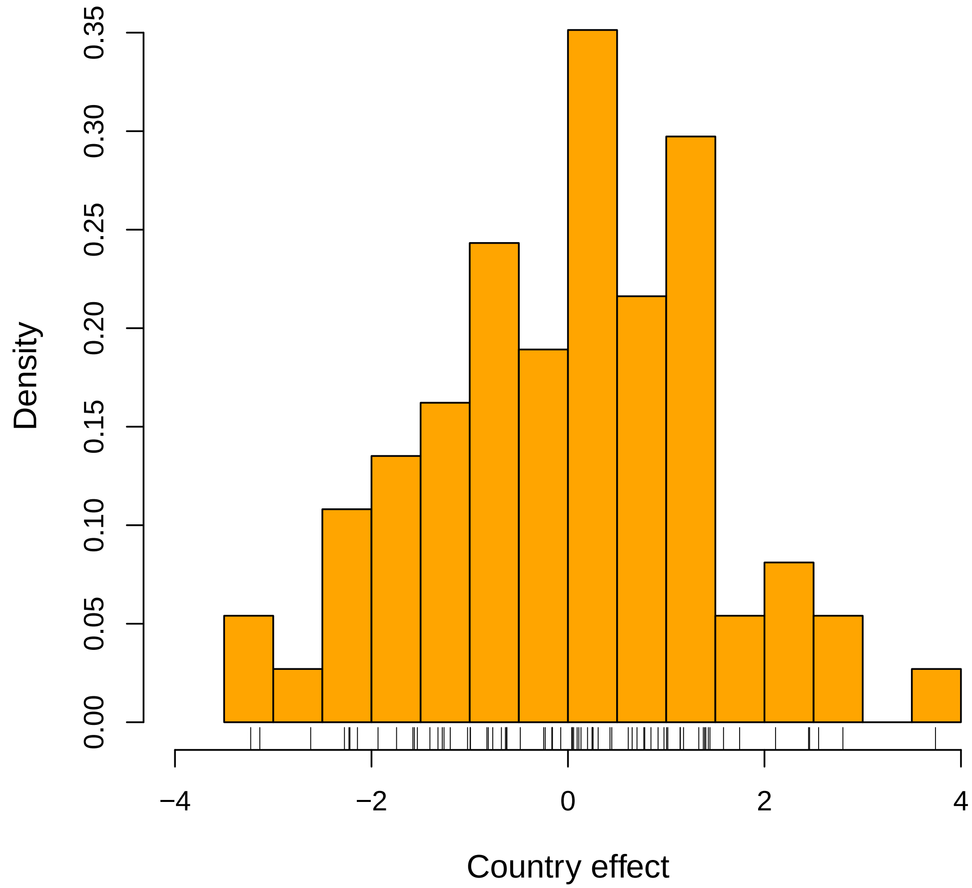
```

> summary(fixef(stern.fe,effect ="individual",type = 'dmean'))
     Estimate Std. Error t-value Pr(>|t|)
1   -2.229050   1.475348 -1.5109  0.13082
14  -0.231877   1.454547 -0.1594  0.87334
18  -1.578418   1.432819 -1.1016  0.27063
22  -0.765428   1.421201 -0.5386  0.59018
25  -2.618924   1.436167 -1.8236  0.06822 .
```

To produce a histogram of the fixed effects:

Interactive R example

```
hist(fixef(stern.fe,effect ="individual",type = 'dmean'))
```


Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
−4
−2
0
2
4
Country effect

10.4 Random Country Individual Effects

Interactive R example

```
> stern.re <- plm( lso ~ lgdp  + lgdp2,
           data=stern.plm,  model="random", effect = "individual")
> summary(stern.re)
Oneway (individual) effect Random Effect Model
   (Swamy-Arora's transformation)

Call:
plm(formula = lso ~ lgdp + lgdp2, data = stern.plm, effect = "individual",
    model = "random")

Balanced Panel: n=74, T=31, N=2294

Effects:
                var std.dev share
idiosyncratic 0.341   0.584  0.15
individual    1.934   1.391   0.85
theta:  0.925

Residuals :
   Min. 1st Qu.  Median 3rd Qu.    Max.
-4.5700 -0.1680  0.0551  0.2520  3.0400

```

```
Coefficients :
            Estimate Std. Error t-value Pr(>|t|)
(Intercept) -21.3789     1.4538  -14.71  < 2e-16 ***
lgdp          3.2663     0.3498    9.34  < 2e-16 ***
lgdp2        -0.1521     0.0211   -7.20  7.9e-13 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Total Sum of Squares:    928
Residual Sum of Squares: 782
F-statistic: 213.906 on 2 and 2291 DF, p-value: <2e-16
```

10.5 Fixed vs Random Country Effects

We can plot:

1. a histogram of the fixed effects:

Interactive R example

```
fief <- fixef(stern.fe,effect ="individual",type = 'dmean')

hist( fief, freq=F, br=20, xlab="Country effect",main="",col="orange")
```

2. A reference normal

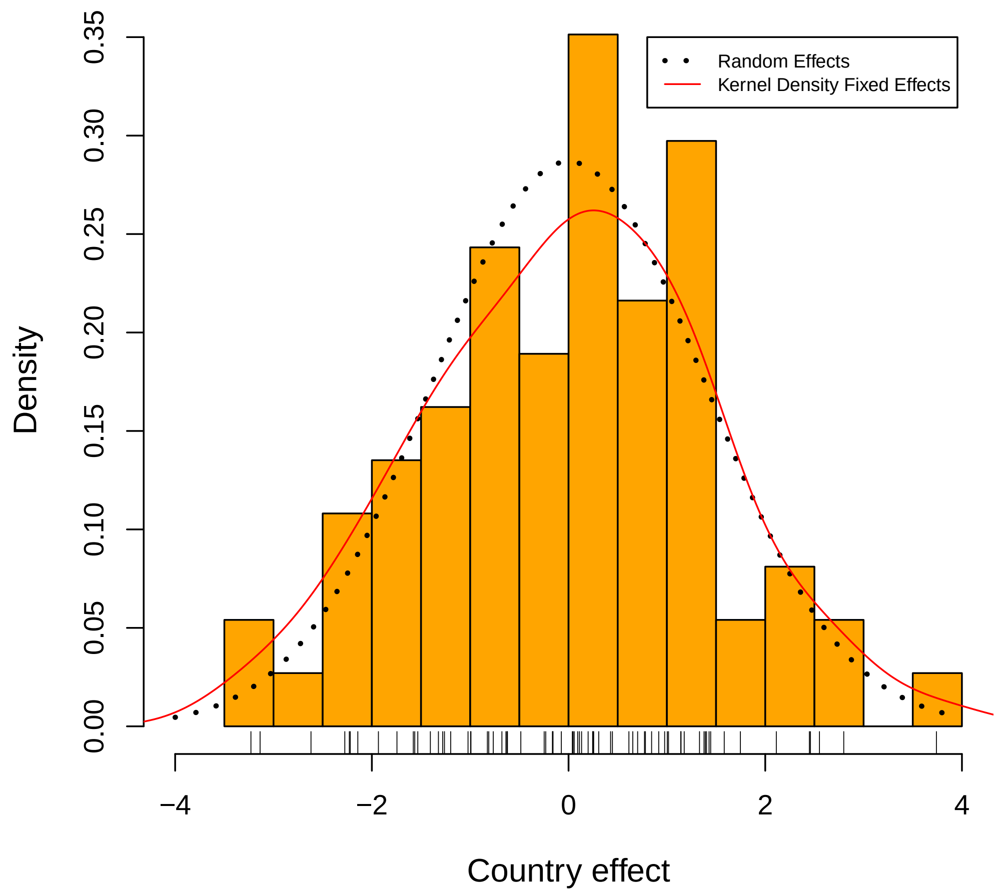
Interactive R example

```
 curve(dnorm(x,sd=1.3906),add=T,lty=3,lwd=2)
```

3. An a nonparametric Kernel estimator:

Interactive R example

```
lines(density( fief ),col="red")
```


Random Effects
Kernel Density Fixed Effects
Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
−4
−2
0
2
4
Country effect

10.6 Fixed Country and Time Effects

Interactive R example

```
> stern2.fe <- plm( lso ~ lgdp  + lgdp2, data=stern.plm,
       model = "within", effect = "twoways")
> summary(stern2.fe)
Twoways effects Within Model

Call:
plm(formula = lso ~ lgdp + lgdp2, data = stern.plm, effect = "twoways",
    model = "within")

Balanced Panel: n=74, T=31, N=2294

Residuals :
   Min. 1st Qu.  Median 3rd Qu.    Max.
-4.4200 -0.1800  0.0209  0.2220  3.2700

Coefficients :
      Estimate Std. Error t-value Pr(>|t|)
lgdp    3.8465     0.3653   10.53  < 2e-16 ***
lgdp2  -0.1706     0.0215   -7.94  3.1e-15 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

```

```
Total Sum of Squares:    850
Residual Sum of Squares: 728
F-statistic: 182.654 on 2 and 2188 DF, p-value: <2e-16
```

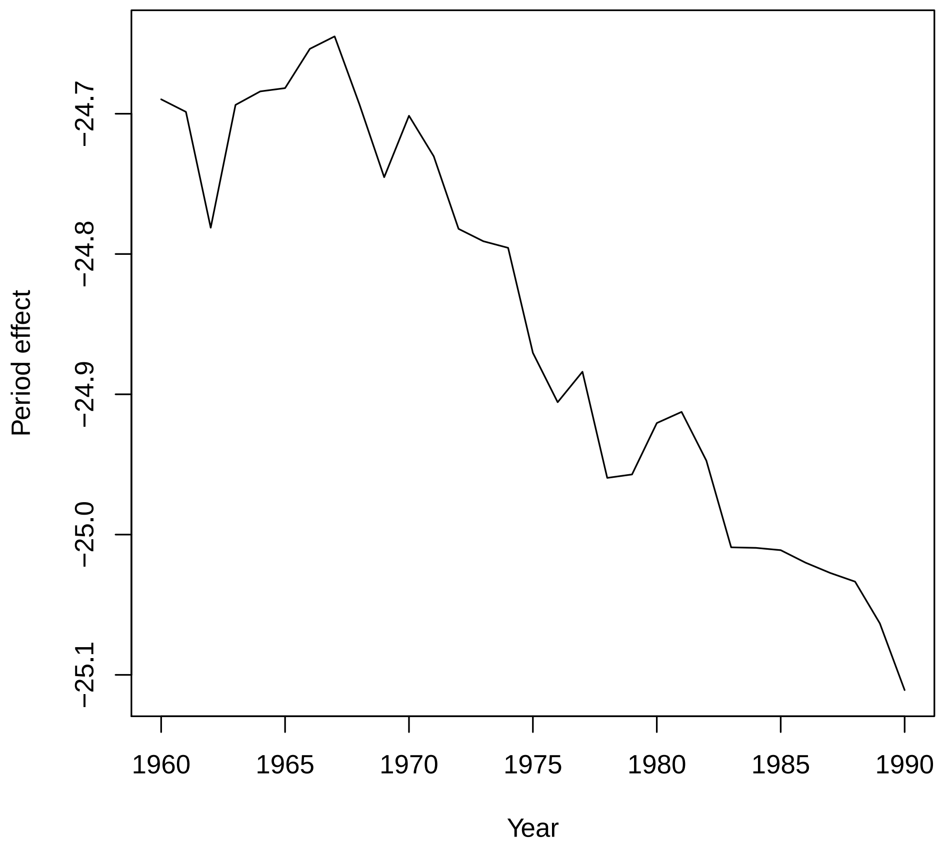
Interactive R example

```
> summary(fixef(stern2.fe,effect ="time"))
     Estimate Std. Error t-value  Pr(>|t|)
1960 -24.6897     1.5494 -15.935 < 2.2e-16 ***
1961 -24.6987     1.5500 -15.935 < 2.2e-16 ***
1962 -24.7813     1.5517 -15.971 < 2.2e-16 ***
1963 -24.6937     1.5530 -15.900 < 2.2e-16 ***
1964 -24.6841     1.5543 -15.881 < 2.2e-16 ***
1965 -24.6817     1.5556 -15.866 < 2.2e-16 ***
1966 -24.6537     1.5564 -15.840 < 2.2e-16 ***
1967 -24.6449     1.5571 -15.827 < 2.2e-16 ***
1968 -24.6933     1.5583 -15.846 < 2.2e-16 ***
1969 -24.7453     1.5602 -15.861 < 2.2e-16 ***
1970 -24.7014     1.5621 -15.813 < 2.2e-16 ***
1971 -24.7304     1.5637 -15.815 < 2.2e-16 ***
1972 -24.7821     1.5643 -15.842 < 2.2e-16 ***
...
```

Interactive R example

```
fe.1 <- summary(fixef(stern2.fe, effect ="time"))[,1]
plot(1960:1990,fe.1,type="l",xlab="Year",ylab="Period effect",ylim=rang)
```

The following Figure shows the time effects.



10.7 First-differences Estimator

Interactive R example

```
> stern.fd <- plm( lso ~ lgdp  + lgdp2,
      data=stern.plm,  model = "fd", effect = "individual")
> summary(stern.fd)
Oneway (individual) effect First-Difference Model

Call:
plm(formula = lso ~ lgdp + lgdp2, data = stern.plm, effect = "individual",
    model = "fd")

Balanced Panel: n=74, T=31, N=2294

Residuals :
    Min.  1st Qu.   Median  3rd Qu.     Max.
-3.37000 -0.07210 -0.00618  0.05760  4.09000

Coefficients :
            Estimate Std. Error t-value Pr(>|t|)
(intercept) -0.00378    0.00623   -0.61   0.5442
lgdp         2.07434    0.73995    2.80   0.0051 **
lgdp2       -0.08990    0.04623   -1.94   0.0519 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```

Total Sum of Squares:    168
Residual Sum of Squares: 165
F-statistic: 22.3866 on 2 and 2217 DF, p-value: 2.37e-10
```

11 Model specification: OLS Vs. Individual Effects, Fixed Vs. Random Effects

- First, the (Breusch and Pagan) LM-test is used to test **the hypothesis that individual effects are significant**. If these tests do not provide any evidence for individual effects, then the model can simply be estimated by ordinary least squares (OLS).
- Second, the Hausman test is used to decide whether the **regressors are correlated with the individual effect**.
- The large value of the LM statistic argues in favor of the of a panel data model, the small Hausman statistic argues in favor of the random effect model. A small Hausman statistic argues in favor of the random effect model, a large in favor of a fixed effects one.

- The LM test is based on the idea that if the variance of the individual effects is zero then the random effects model model reduces to the pooled OLS model.
- The null hypothesis tested is

$$\mathbf{H}_0\colon \sigma_\alpha^2 = 0$$

Interactive R example

```
> plmtest(  lso ~ lgdp  + lgdp2, data=stern.plm, type="bp", effect = c("twoways"))

        Lagrange Multiplier Test - two-ways effects (Breusch-Pagan)

data:  lso ~ lgdp + lgdp2
chisq = 24163, df = 2, p-value < 2.2e-16
alternative hypothesis: significant effects
```

- The LM statistic is 24163 and assessed against the χ^2 distribution with 2 degrees of freedom is significant at any conventional level level (p-value ≈ 0.00).
- We can conclude that country effects are considerably important and that panel methods are necessary to determine the impact of income factors on the emissions.

- To decide between fixed and random effects approach to panel data estimation, we can use the Hausman test (1978),
- The null hypothesis tested is

$$\mathbf{H}_0\text{: } \mathrm{corr}(\alpha_i, X_{i,t}) = 0$$

  that the regressors are correlated with the individual effect.
- If the individual effects, α_i, are correlated with the regressors the RE estimator is inconsistent.
- A small Hausman statistic argues in favor of the random effect model, a large in favor of a fixed effects one.

Interactive R example

```
> phtest( stern.fe, stern.re )

        Hausman Test

data:  lso ~ lgdp + lgdp2
chisq = 6.2707, df = 2, p-value = 0.04348
alternative hypothesis: one model is inconsistent
```

- The Hausman statistic is 6.2707 which assessed against the χ^2 distribution with 2 degrees of freedom is significant at the 5 per cent level (p-value 0.04348).
- The fixed effect model is the preferred model.

12 Selecting a Subsample

To select a subsample of observation (for example to compute the regression only for OECD or for non-OECD countries) for example, to include only the OECD sample, we can issue the following commands.

Interactive R example

```
> summary(stern.fe.oe)
Oneway (individual) effect Within Model

Call:
plm(formula = lso ~ lgdp + lgdp2, data = stern.plm,
   subset = oe == 1, effect = "individual", model = "within")

Balanced Panel: n=23, T=31, N=713

Residuals :
    Min.  1st Qu.   Median  3rd Qu.     Max.
-1.11000 -0.16000  0.00648  0.15000  0.96400

Coefficients :
      Estimate Std. Error t-value  Pr(>|t|)
lgdp  12.24450    0.73052  16.761 < 2.2e-16 ***
lgdp2 -0.67121    0.04124 -16.276 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
Total Sum of Squares:    70.142
Residual Sum of Squares: 45.805
F-statistic: 182.773 on 2 and 688 DF, p-value: < 2.22e-16
```

13 Other Tests

Autocorrelation:

- Breusch-Godfrey test for Panel models

Interactive R example

```
library(lmtest)


pbgtest(stern.fe, order = 4)

> pbgtest(stern.fe, order = 4)

        Breusch-Godfrey/Wooldridge test for serial correlation in panel models

data:  lso ~ lgdp + lgdp2
chisq = 1567.839, df = 4, p-value < 2.2e-16
alternative hypothesis: serial correlation in idiosyncratic errors
```

- Durbin-Watson test for Panel models

Interactive R example

```
pdwtest(stern.fe)
> pdwtest(stern.fe)

        Durbin-Watson test for serial correlation in panel models

data:  lso ~ lgdp + lgdp2
DW = 0.3516, p-value < 2.2e-16
alternative hypothesis: serial correlation in idiosyncratic errors
```

14 Robust Standard Errors/t-statistics

Interactive R example

```
### We discovered serial autocorrelation we must decide what to do
### 1) evidence of dynamic misspecification could go dynamic

#### 2) or use robust standard errors

 > coeftest(stern.fe,vcovHC)

 t test of coefficients:

        Estimate Std. Error t value Pr(>|t|)
 lgdp   3.252832   1.297629  2.5068  0.01226 *
 lgdp2 -0.152532   0.077606 -1.9655  0.04949 *
 ---
 Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

15 References

15.1 Environmental Economics References

– Andreoni, J. and Levinson, A (2001) , "The Simple Analytics of the Environmental Kuznets Curve," *Journal of Public Economics*, 80, 269–286.

– Grossman G. M., Krueger A. B. (1995), "Economic growth and the Environment," *Quarterly Journal of Economics*, **110** (2), pp. 353–377.

– Kiel and McClain (1995): "House Prices During Siting Decisions Stages: The Case of an Incinerator from Rumor Through Operation," *Journal of Environmental Economics and Management*, 28, 241–255.

– Perman, R., Ma, Y., McGilvray, J. and Common, M. S. (1999), *Natural Resources and Environmental Economics*, 2nd Edition, Longmans.

– Panayotou, T. (2000), "Economic Growth, Environment, Kuznets Curve," *CID Working Paper*, No. 56.

– Stern, D. I. and Common, M. S. (2001), "Is There an Environmental Kuznetz Curve for Sulfur?" *Journal of Environmental Economics and Management*, 41, pp. 162–178.

15.2 Panel Data References

The most useful books on panel data are

– Baltagi, B.H. (2001), *Econometric Analysis of Panel Data*, 2nd edition, Wiley, Chichester.

– Hsiao, C. (1986), *Analysis of Panel Data*, Cambridge University Press, Cambridge.

– Wooldridge, J.M. (2002), *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge, Massachusetts.

Good general introductions of panel data methods are available as chapters in the following texts:

- Greene, W. (1997), *Econometric Analysis*, Prentice Hall.
- Jhonston, J. and J. DiNardo (1997), *Econometric Methods*, 4th ed., McGraw Hill.
- Wooldridge, J.M. (2001), *Introductory Econometrics: A Modern Approach*, South-Western College Publishing.

Good updated introductions to panel data methods are available as chapters in the following texts:

- Baltagi, B.H. (1998), Panel Data Methods, in *Handbook of Applied Economic Statistics*, Edited by Ullah and Giles, Marcel Dekkar.
- Verbeek, M. (2000), *A Guide to Modern Econometrics*, Wiley.

A few relevant papers are:

- Arellano, M. and Bond, S.R. (1991), "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations, *Review of Economic Studies*, 58, 277–297.
- Breitung, J. and Meyer, W. (1994), "Testing for Unit Roots in Panel Data: Are wages on different bargaining levels cointegrated?", *Applied Economics*, 26, 353–361.
- Hausman, J.A. and Taylor, W.E. (1979), "Panel Data and Unobservable Individual Effects", *Econometrica*, 49, 1377–1399.

If you are interested in dynamic panels and some of the estimators and techniques presented in the above-mentioned papers you might want to connsult

- Arellano, Bond and Doornik, Dynamic panel data estimation using DPD for Ox,
  available at the following URL:
  http://www.nuff.ox.ac.uk/Users/Doornik/software/dpdox.zip

– Doornik and Draisma, Introduction to Ox,
available at the following URL:
File-URL: http://www.nuff.ox.ac.uk/Users/Doornik/doc/oxtutor.zip

15.3 Other Econometrics References

– Fieller, E.C. (1940), "The biological standardization of Insulin" *Suppl to J.R.Statist.Soc*, 7, pp. 1–64.