Panel Analyses Report

David BYAMUNGU

2021-10-18

Contents

1 Introduction **5**

2 The One-way Error Component Regression Model **7**

2.1 INTRODUCTION 7

2.2 THE FIXED EFFECTS MODEL 9

2.3 THE RANDOM EFFECTS MODEL 13

3 The Two-way Error Component Regression Model **19**

3.1 NTRODUCTION 19

3.2 THE FIXED EFFECTS MODEL 20

3.3 THE RANDOM EFFECTS MODEL 21

3.4 REFERENCES 26

4 Test of Hypotheses with Panel Data **27**

4.1 TESTS FOR POOLABILITY OF THE DATA 27

4.2 TESTS FOR INDIVIDUAL AND TIME EFFECTS 31

4.3 HAUSMAN'S SPECIFICATION TEST 31

5 Analyses **33**

5.1 Netoyage de la base des données 33

5.2 Modele avec les pays comme individus 35

5.3 Modèle avec les marchandises comme individus 42

6 Conclusion **45**

Chapter 1

Introduction

Dans ce document nous cherchons à modéliser les taxes perçues dans différents pays , formant un panel dont la période est 10 ans. Ainsi, nous expliquerons la variable taxe par:

(1) Le poids des Marchandises
(2) La qualité des Marchandises

Le but est d'arbitrer entre:

(1) le modèle *pooling*
(2) le *modele à éffet fixe* et
(3) le *modèle à éffet aléatoire*

et en fin produire un `modèle dynamique` permettant d'expliquer la variation du taxe au cours du temps, avec comme variable dépendante additionnelle le taxe décalé

Nous expliquons Notre méthodologie dans la partie suivante.

Chapter 2

The One-way Error Component Regression Model

2.1 INTRODUCTION

A panel data regression differs from a regular time-series or cross-section regression in that it has a double subscript on its variables, i.e.

$$y_{it} = \alpha + X_{it}^{'}\beta + u_{it}$$
$$i = 1, ..., N; t = 1, ..., T \tag{2.1}$$

(2.1)

with $ i $ denoting households, individuals, firms, countries, etc. and t denoting time. The i subscript, therefore, denotes the cross-section dimension whereas t denotes the time-series dimension.

$$\alpha$$

is a scalar,

$$\beta$$

is

$$K \times 1$$

and Xi t is the $ it_{th} $ observation on K explanatory variables. Most of the panel data applications utilize a one-way error component model for the disturbances, with

$$u_{it} = \mu_i + v_{it}$$

(2.2)

where i denotes the unobservable individual-specific effect and i t denotes the remainder disturbance. For example, in an earnings equation in labor economics, yi t will measure earnings of the head of the household, whereas

$$X_{it}$$

may contain a set of variables like experience, education, union membership, sex, race, etc. Note that i is time-invariant and it accounts for any individual-specific effect that is not included in the regression. In this case we could think of it as the individual's unobserved ability. The remainder disturbance

$$v_{it}$$

varies with individuals and time and can be thought of as the usual disturbance in the regression. Alternatively, for a production function utilizing data on firms across time,

$$y_{it}$$

will measure output and

$$X_{it}$$

will measure inputs. The unobservable firm-specific effects will be captured by the

$$\mu_i$$

and we can think of these as the unobservable entrepreneurial or managerial skills of the firm's executives. Early applications of error components in economics include Kuh (1959) on investment, Mundlak (1961) and Hoch (1962) on production functions and Balestra and Nerlove (1966) on demand for natural gas. In vector form (2.1) can be written as

$$y = \alpha i_{NT} + X\beta + u = Z\delta + u \tag{2.2}$$

(2.3)

$$y = \alpha i_{NT} + X\beta + u = Z\delta + u \tag{2.3}$$

(2.3)

where $ y $ is $ NT \times 1 $, $ X $ is $ NT \times K $, $ Z = [_{NT} , X] $, $ ^{'} = (^{ {'}, }) $ and $\iota_N T$ is a vector of ones of dimension NT. Also, (2.2) can be written as

$$u = Z_\mu \mu + v \tag{2.4}$$

(2.4)

$$y_{it} = \alpha + X_{it}^{'} + U_{it}$$

, $ i=1,...,N ; t=1,...,T $ with i denoting households, individuals, firms, countries, etc. and t denoting time. The i subscript, therefore, denotes the cross-section dimension whereas t denotes the time-series dimension. $ $ is a scalar, $ $ is $ K \times 1 $ and $X_{it} $is the $ it^{th} $ observation on K explanatory variables. disturbances, with it

$$u_{it} = u_i + v_{it}$$

where i denotes the unobservable individual-specific effect and νit denotes the remainder disturbance. For example, in an earnings equation in labor economics, y_{it} will measure earnings of the head of the household, whereas Xi t may contain a set of variables like experience, education, union membership, sex, race, etc. Note that i is time-invariant and it accounts for any individual-specific effect that is not included in the regression. In this case we could think of it as the individual's unobserved ability. The remainder disturbance $ {it} $ varies with individuals and time and can be thought of as the usual disturbance in the regression. Alternatively, for a production function utilizing data on firms across time, $ y_{it} $ will measure output and $ X_{it} $ will measure inputs. The unobservable firm-specific effects will

be captured by the $ _i $ and we can think of these as the unobservable entrepreneurial or managerial skills of the firm's executives. Early applications of error components in economics include Kuh (1959) on investment, Mundlak (1961) and Hoch (1962) on production functions and Balestra and Nerlove (1966) on demand for natural gas. In vector form (2.1) can be written as

$$y = \alpha i_{NT} + X\beta + u = Z\delta + u$$

where y is $ NT \times 1$, X is $NT \times K$, $ Z = [i_{NT} , X] $, $ ^{‘}=(^ ’, ^{’})$ and i_{NT} is a vector of ones of dimension NT . Also, (2.2) can be written as

$$u = Z_{\mu}\mu + v \tag{2.5}$$

(2.4)

where $ u^{‘} = (u_{11}, . . . , u_{1T} , u_{21}, . . . , u_{2T}, . . . , u_{N1}, . . . , u_{NT}) $ with the observations stacked such that the slower index is over individuals and the faster index is over time.$ Z_ = IN T $ where IN is an identity matrix of dimension N, T is a vector of ones of dimension T and $ $ denotes Kronecker product. $ Z_ $ is a selector matrix of ones and zeros, or simply the matrix of individual dummies that one may include in the regression to estimate the i if they are assumed to be fixed parameters. $ ^{’} = (_1, . . . , _N)$ and $\nu^{'} = (\nu 11, ..., \nu_{1T}, ..., \nu_{N1}, ..., \nu_{NT})$. Note that $Z_ Z^{‘} = I_N J_T $ where J_T is a matrix of ones of dimension T and $ P = Z (Z^{’} Z){−1} Z^{’} $, the projectionmatrix on $ Z_ $, reduces to $ IN J_T $ where $J _T= JT /T $. P is a matrix which averages the observation across time for each individual, and Q = INT − P is a matrix which obtains the deviations from individual means. For example, regressing y on the matrix of dummy variables Z_{μ} gets the predicted values $ P_y $ which has a typical element

$$\bar{y}_i = \sum_{t=1}^{T} \frac{y_{it}}{T}$$

repeated T times for each individual. The residuals of this regression are given by Qy which has a typical element

$$(y_{it} - \bar{y}_{i.})$$

P and Q are (i) symmetric idempotent matrices, i.e.

$P^{'} = P$ and $ P^2 = P $. This means that $rank(P) = tr(P) = N$ and $rank(Q) = tr(Q) = N(T−1)$. This uses the result that the rank of an idempotent matrix is equal to its trace (see Graybill, 1961,theorem 1.63). Also, (ii) P and Q are orthogonal, i.e. $PQ=0 $ and (iii) they sum to the identity matrix $P + Q = I_{NT}$. In fact, any two of these properties imply the third (see Graybill, 1961, theorem 1.68).

2.2 THE FIXED EFFECTS MODEL

In this case, the i are assumed to be fixed parameters to be estimated and the remainder disturbances stochastic with v_{it} independent and identically distributed $IID(0, \sigma_v^2)$. The X_{it} are assumed independent of the v_{it} for all i and t. The fixed effects model is an appropriate specification if we are focusing on a specific set of N firms, say, IBM, GE,Westinghouse, etc. and our inference is restricted to the behavior of these sets of firms. Alternatively, it could be a set of N OECD countries, or N American states. Inference in this case is conditional on the particular N firms, countries or states that are observed. One can substitute the disturbances given by (2.4) into (2.3) to get

$$y = \alpha i_{IT} + X\beta + Z_{\mu}\mu + v = Z\delta + Z_{\mu}\mu + v \tag{2.6}$$

(2.5)

and then perform ordinary least squares (OLS) on (2.5) to get estimates of $, and µ $

Note that Z is $ NT ×(K+1)$ and Z , the matrix of individual dummies, is$NT ×N $ NT × N. If N is large, (2.5) will include too many individual dummies, and the matrix to be inverted by OLS is large and of dimension $(N + K)$. In fact, since and are the parameters of interest, one can obtain the LSDV (least squares dummy variables) estimator from (2.5), by premultiplying the model by Q and performing OLS on the resulting transformed model:

$$Qy = QX\beta + Qv \tag{2.7}$$

(2.6)

This uses the fact that $QZ_\mu = Qi_{NT} = 0$, since $PZ_\mu = Z_\mu$ the Q matrix wipes out the individual effects. This is a regression of $\tilde{y} = QY$ with element $(y_{it} - \bar{y}_{i.})$ on $\breve{X} = QX$ with typical element

$$\tilde{\beta} = (X'QX)^{-1} X'Qy \tag{2.8}$$

(2.7) (2.7) with $\text{var}(\tilde{\beta}) = \sigma_v^2 (X'QX)^{-1} = \sigma_v^2 \left(\widetilde{X}'\tilde{X}\right)^{-1}$. $\tilde{\beta}$ could have been obtained from (2.5) using results on partitioned inverse or the Frisch–Waugh–Lovell theorem discussed in Davidson and MacKinnon (1993, p. 19). This uses the fact that P is the projection matrix on Z_μ and $Q = I_{NT} - P$ (see problem 2.1). In addition, generalized least squares (GLS) on (2.6), using the generalized inverse, will also yield $\tilde{\beta}$ (see problem 2.2).

Note that for the simple regression

$$y_{it} = \beta x_{it} + \mu_i + v_i \tag{2.9}$$

(2.8)

and averaging over time gives

$$\bar{y}_{i.} = \beta \bar{x}_{i.} + \mu_i + \bar{v}_i \tag{2.10}$$

(2.9)

Therefore, subtracting (2.9) from (2.8) gives

$$y_{it} - \bar{y}_{i.} = \beta(x_{it} - \bar{x}_{i.}) + (v_{it} - \bar{v}_{i.})$$

(2.10)

Also, averaging across all observations in (2.8) gives

$$\bar{y}_{..} = \alpha + \beta \bar{x}_{..} + \bar{v}_{..}$$

(2.11)i =0 where we utilized the restriction that $ *{i=1}^{n} i=0 $ This is an arbitrary restriction on the dummy variable coefficients to avoid the dummy variable trap, or perfect multicollinearity; see Suits (1984) for alternative formulations of this restriction. In fact only β and $ (+* {i})$ are estimable from (2.8), and not and i separately, unless a restriction like*

$$\sum_{i=1}^{n} \mu_i = 0$$

is imposed. In this case, $\tilde{\beta}$ is obtained from regression (2.10),

$$\tilde{\alpha} = \bar{y}_{..} - \tilde{\beta} \bar{x}_{..}$$

can be recovered from (2.11) and

$$\tilde{\mu}_i = \bar{y}_{i.} - \tilde{\alpha} - \tilde{\beta}\bar{X}_{i.}$$

from (2.9). For large labor or consumer panels, where N is very large,regressions like (2.5) may not be feasible, since one is including (N − 1) dummies in the regression. This fixed effects (FE) least squares, also known as least squares dummy variables (LSDV), suffers from a large loss of degrees of freedom. We are estimating (N − 1) extra parameters, and too many dummies may aggravate the problem of multicollinearity among the regressors. In addition, this FE estimator cannot estimate the effect of any time-invariant variable like sex, race, religion, schooling or union participation. These time-invariant variables are wiped out by the Q transformation, the deviations from means transformation (see (2.10)). Alternatively, one can see that these time-invariant variables are spanned by the individual dummies in (2.5) and therefore any regression package attempting (2.5) will fail, signaling perfect multicollinearity. If (2.5) is the true model, LSDV is the best linear unbiased estimator (BLUE) as long as v_{it} is the standard classical disturbance with mean 0 and variance–covariance matrix $ v^{2} I_{NT} $. Note that as $T \to \infty$ the FE estimator is consistent. However, if T is fixed and $N \to \infty$ as is typical in short labor panels, then only the FE estimator of is consistent; the FE estimators of the individual effects $\alpha + \mu_i$ are not consistent since the number of these parameters increases as N increases. This is the incidental parameter problem discussed by Neyman and Scott (1948) and reviewed more recently by Lancaster (2000). Note that when the true model is fixed effects as in (2.5), OLS on (2.1) yields biased and inconsistent estimates of the regression parameters. This is an omission variables bias due to the fact that OLS deletes the individual dummies when in fact they are relevant.

(1) *Testing for fixed effects.* One could test the joint significance of these dummies, i.e. H0; $\mu_1 = \mu_2 = ... = \mu_{N-1} = 0$, by performing an F-test. (Testing for individual effects will be treated extensively in Chapter 4.) This is a simple Chow test with the restricted residual sums of squares (RRSS) being that of OLS on the pooled model and the unrestricted residual sums of squares (URSS) being that of the LSDV regression. If N is large, one can perform the Within transformation and use that residual sum of squares as the URSS. In this case

$$F_0 = \frac{\dfrac{RRSS - URSS}{N-1}}{\dfrac{URSS}{NT-N-K}} \sim F_{N-1,N(T-1)-K}$$

(2.12)

(2) *Computational warning.* One computational caution for those using theWithin regression given by (2.10). The s^2 f this regression as obtained from a typical regression package divides the residual sums of squares by NT − K since the intercept and the dummies are not included. The proper s^2 , say s^{*2} from the LSDV regression in (2.5), would divide the same residual sums of squares by $N(T-1) - K$. Therefore, one has to adjust the variances obtained from the Within regression (2.10) by multiplying the variance–covariance matrix by

$$\frac{s^2}{s^{*2}}$$

or simply by multiplying by $[NT - K]/[N(T-1) - K]$

(3) *Robust estimates of the standard errors.* For the Within estimator, Arellano (1987) suggests a simple method for obtaining robust estimates of the standard errors that allow for a general variance–covariance matrix on the v_{it} as in White (1980). One would stack the panel as an equation for each individual:

$$y_i = Z_i\delta + \mu i_{it} + v_i$$

(2.13)

where y_i is T $\times$ \$Z_i =[l_T,X_i] \$, X_i is T $\times$ K, μ_i is a scalar, $\delta' = (\alpha, \beta')$, i_T is a vector of ones of dimension T and v_i is T x 1 . In general, $E(v_i, v_i') = \Omega_i$ for i = 1, 2, . . . , N, where Ω_i is a positive definite matrix of dimension T . We still assume $E(v_i, v_j') = 0$ for $i \neq j$ T is assumed small and N large as in household or company panels, and the asymptotic results are performed for $N \to \infty$ and T fixed. Performing the Within transformation on this set of equations (2.13) one gets

$$\tilde{y}_i = \widetilde{X}_i \beta + \tilde{v}_i$$

(2.14)

where

$$\tilde{y} = Qy$$

,

$$\widetilde{X} = QX$$

and

$$\tilde{v} = Qv$$

, with

$$\tilde{y} = (\tilde{y}_1', ..., \tilde{y}_N')'$$

and

$$\tilde{y}_i = (I_T - \bar{J}_T) y_i$$

Computing robust least squares on this system, as described by White (1980), under the restriction that each equation has the same β one gets the Within estimator of which has the following asymptotic distribution:

$$N^{\frac{1}{2}}(\tilde{\beta} - \beta) \sim N(0, M^{-1} V M^{-1})$$

(2.15)

where

$$M = \frac{p \lim(\widetilde{X}' \widetilde{X})}{N}$$

Note that

$$\widetilde{X}_i = (I_T - \bar{J}_T) X_i$$

and

$$\widetilde{X}' diag[\Omega_i] Q \widetilde{X}$$

(see problem 2.3). In this case, V is estimated by

$$\frac{\widetilde{V} = \sum_{i=1}^{N} \widetilde{X}_i' \tilde{u}_i \tilde{u}_i' \widetilde{X}_i'}{N}$$

where

$$\tilde{u}_i = \tilde{y}_i - \widetilde{X}_i \tilde{\beta}_i$$

. Therefore, the robust asymptotic variance– covariance matrix of β is estimated by

$$\text{var}(\tilde{\beta}) = \left(\widetilde{X}' \tilde{X}\right)^{-1} \left[\sum_{i=1}^{N} \widetilde{X}_i' \tilde{u}_i \tilde{u}_i' \widetilde{X}_i\right] \left(\widetilde{X}' \tilde{X}\right)^{-1}$$

2.3 THE RANDOM EFFECTS MODEL

There are too many parameters in the fixed effects model and the loss of degrees of freedom can be avoided if the i can be assumed random. In this case $ *i* IID(0, ^2)$, $ _{it} IID(0, _v^2) $ and the μ_i are independent of the v_{it} . In addition, the X_{it} are independent of the μ_i and v_{it},for all i and t. The random effects model is an appropriate specification if we are drawing N individuals randomly from a large population. This is usually the case for household panel studies. Care is taken in the design of the panel to make it "representative" of the population we are trying to make inferences about. In this case, N is usually large and a fixed effects model would lead to an enormous loss of degrees of freedom. The individual effect is characterized as random and inference pertains to the population from which this sample was randomly drawn.

But what is the population in this case? Nerlove and Balestra (1996) emphasize Haavelmo's (1944) view that the population "consists not of an infinity of individuals, in general, but of an infinity of decisions" that each individual might make. This view is consistent with a random effects specification. From (2.4), one can compute the variance–covariance matrix

$$\begin{aligned}\Omega = E(uu') &= Z_\mu E(\mu\mu') Z'_\mu + E(vv') \\ &= \sigma^2_\mu (I_N \otimes J_T) + \sigma^2_v (I_N \otimes I_T)\end{aligned} \tag{2.11}$$

This implies a homoskedastic variance $\text{var}(u_{it}) = \sigma^2_\mu + \sigma^2_v$f for all i and t, and an equicorrelated block-diagonal covariance matrix which exhibits serial correlation over time only between the disturbances of the same individual. In fact,

$$\begin{aligned}\text{cov}(u_{it}, u_{js}) &= \sigma^2_\mu + \sigma^2_v \quad \text{for} \quad i = j, t = s \\ &= \sigma^2_\mu \qquad\qquad \text{for } i = j, t \neq s\end{aligned}$$

and zero otherwise. This also means that the correlation coefficient between μ_{it} and μ_{js} is

$$\begin{aligned}\rho = \text{correl}(u_{it}, u_{js}) &= 1 \qquad\qquad\qquad\qquad\qquad\qquad \text{for } i = j, t = s \\ &= \sigma^2_\mu / (\sigma^2_\mu + \sigma^2_v) \quad \text{for } i = j, t \neq s\end{aligned}$$

and zero otherwise. In order to obtain the GLS estimator of the regression coefficients, we need $ ^{-1} $. This is a huge matrix for typical panels and is of dimension $NT \times NT$. No brute force inversion should be attempted even if the researcher's application has a small N and T .1 We will follow a simple trick devised by Wansbeek and Kapteyn (1982b, 1983) that allows the derivation of $ ^{-1} $ ans $\Omega^{-1/2}$ Essentially, one replaces J_T by $T\bar{J}_T$ and I_T by $(E_T + \bar{J}_T)$ where E_T is by definition $(I_T - \bar{J}_T)$. In this case

$$\Omega = T\sigma^2_\mu (I_N \otimes \bar{J}_T) + \sigma^2_v (I_N \otimes E_T) + \sigma^2_v (I_N \otimes \bar{J}_T)$$

Collecting terms with the same matrices, we get

$$\Omega = (T\sigma^2_\mu + \sigma^2_v)(I_N \otimes \bar{J}_T) + \sigma^2_v (I_N \otimes E_T) = \sigma^2_1 P + \sigma^2_v Q \tag{2.12}$$

where $\sigma^2_1 = T\sigma^2_\mu + \sigma^2_v$. (2.18) is the spectral decomposition representation of Ω, with σ^2_1 being the first unique characteristic root ofΩ of multiplicity $N(T-1)$. It is easy to verify, using the properties of P and Q, that

$$\Omega^{-1} = \frac{1}{\sigma^2_1} P + \frac{1}{\sigma^2_v} Q \tag{2.13}$$

and

$$\Omega^{-1/2} = \frac{1}{\sigma_1} P + \frac{1}{\sigma_v} Q \tag{2.14}$$

In fact, $\Omega^r = (\sigma_1^2)^r P + (\sigma_v^2)^r Q$ where r is an arbitrary scalar. Now we can obtain GLS as a weighted least squares. Fuller and Battese (1973, 1974) suggested premultiplying the regression equation given in (2.3) by $\sigma_v \Omega^{-1/2} = Q + (\sigma_v/\sigma_1) P$ and performing OLS on the resulting transformed regression. In this case, $y^* = \sigma_v \Omega^{-1/2} y$ has a typical element $y_{it} - \theta \bar{y}_{i.}$ where $\theta = 1 - (\sigma_v/\sigma_1)$ (see problem 2.4). This transformed regression inverts a matrix of dimension $(K+1)$ and can easily be implemented using any regression package.

The best quadratic unbiased (BQU) estimators of the variance components arise naturally from the spectral decomposition of Ω . In fact, $Pu \sim (0, \sigma_1^2 P)$ and $Qu \sim (0, \sigma_v^2 Q)$ and

$$\hat{\sigma}_1^2 = \frac{u' P u}{\text{tr}(P)} = T \sum_{i=1}^{N} \bar{u}_{i.}^2 / N \tag{2.15}$$

and

$$\hat{\sigma}_v^2 = \frac{u' Q u}{\text{tr}(Q)} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} (u_{it} - \bar{u}_{i.})^2}{N(T-1)} \tag{2.16}$$

provide the BQU estimators of σ_1^2 andσ_v^2 , respectively (see problem 2.5).

These are analyses of variance-type estimators of the variance components and are minimum variance-unbiased under normality of the disturbances (see Graybill, 1961). The true disturbances are not known and therefore (2.21) and (2.22) are not feasible. Wallace and Hussain (1969) suggest substituting OLS residual $\hat{u}_{\text{OLS}}$ instead of the true u. After all, under the random effects model, the OLS estimates are still unbiased and consistent, but no longer efficient. Amemiya (1971) shows that these estimators of the variance components have a different asymptotic distribution from that knowing the true disturbances. He suggests using the LSDV residuals instead of the OLS residuals. In this case $\tilde{u} = y - \tilde{\alpha} \iota_{NT} - X\tilde{\beta}$ where and $\bar{X}'_{..}$ is a $1 \times K$

vector of averages of all regressors. Substituting these $\hat{u}$ for u in (2.21) and (2.22) we get the Amemiya-type estimators of the variance components. The resulting estimates of the variance components have the same asymptotic distribution as that knowing the true disturbances:

$$\begin{pmatrix} \sqrt{NT}\left(\hat{\sigma}_v^2 - \sigma_v^2\right) \\ \sqrt{N}\left(\hat{\sigma}_\mu^2 - \sigma_\mu^2\right) \end{pmatrix} \sim N\left(0, \begin{pmatrix} 2\sigma_v^4 & 0 \\ 0 & 2\sigma_\mu^4 \end{pmatrix}\right) \tag{2.17}$$

where $\hat{\sigma}_\mu^2 = \left(\hat{\sigma}_1^2 - \hat{\sigma}_v^2\right)/T$.[3]

Swamy and Arora (1972) suggest running two regressions to get estimates of the variance components from the corresponding mean square errors of these regressions. The first regression is the Within regression, given in (2.10), which yields the following s^2:

$$\hat{\sigma}_v^2 = \left[y' Q y - y' Q X \left(X' Q X\right)^{-1} X' Q y\right] / [N(T-1) - K] \tag{2.18}$$

The second regression is the Between regression which runs the regression of averages across time, i.e.

$$\bar{y}_{i.} = \alpha + \bar{X}'_{i.} \beta + \bar{u}_{i.} \quad i = 1, \ldots, N \tag{2.19}$$

(2.25)

This is equivalent to premultiplying the model in (2.5) by P and running OLS. The only caution is that the latter regression has NT observations because it repeats the averages T times for each individual,

while the cross-section regression in (2.25) is based on N observations. To remedy this, one can run the cross-section regression

$$\sqrt{T}\bar{y}_{i.} = \alpha\sqrt{T} + \sqrt{T}\bar{X}'_{i.}\beta + \sqrt{T}\bar{u}_{i.} \tag{2.20}$$

(2.26)

where one can easily verify that var $\left(\sqrt{T}\bar{u}_{i.}\right) = \sigma_1^2$. This regression will yield an s^2 given by

$$\hat{\sigma}_1^2 = \left(y'Py - y'PZ\left(Z'PZ\right)^{-1}Z'Py\right)/(N-K-1) \tag{2.21}$$

(2.27)

Note that stacking the following two transformed regressions we just performed yields

$$\begin{pmatrix} Qy \\ Py \end{pmatrix} = \begin{pmatrix} QZ \\ PZ \end{pmatrix}\delta + \begin{pmatrix} Qu \\ Pu \end{pmatrix} \tag{2.22}$$

(2.28)

and the transformed error has mean 0 and variance–covariance matrix given by

$$\begin{pmatrix} \sigma_v^2 Q & 0 \\ 0 & \sigma_1^2 P \end{pmatrix}$$

Problem 2.7 asks the reader to verify that OLS on this system of 2NT observations yields OLS on the pooled model (2.3). Also, GLS on this system yields GLS on (2.3). Alternatively, one could get rid of the constant by running the following stacked regressions:

$$\begin{pmatrix} Qy \\ \left(P-\bar{J}_{NT}\right)y \end{pmatrix} = \begin{pmatrix} QX \\ \left(P-\bar{J}_{NT}\right)X \end{pmatrix}\beta + \begin{pmatrix} Qu \\ \left(P-\bar{J}_{NT}\right)u \end{pmatrix} \tag{2.23}$$

(2.29)

This follows from the fact that Q NT = 0 and (P − J⁻NT) NT = 0. The transformed error has zero mean and variance–covariance matrix

$$\begin{pmatrix} \sigma_v^2 Q & 0 \\ 0 & \sigma_1^2\left(P-\bar{J}_{NT}\right) \end{pmatrix}$$

OLS on this system yields OLS on (2.3) and GLS on (2.29) yields GLS on (2.3). In fact,

$$\begin{aligned} \hat{\beta}_{\text{GLS}} &= \left[\left(X'QX/\sigma_v^2\right) + X'\left(P-\bar{J}_{NT}\right)X/\sigma_1^2\right]^{-1}\left[\left(X'Qy/\sigma_v^2\right)\right. \\ &\quad \left.+X'\left(P-\bar{J}_{NT}\right)y/\sigma_1^2\right] \\ &= \left[W_{XX}+\phi^2 B_{XX}\right]^{-1}\left[W_{Xy}+\phi^2 B_{Xy}\right] \end{aligned} \tag{2.24}$$

(2.30)

with var $\left(\hat{\beta}_{\text{GLS}}\right) = \sigma_v^2\left[W_{XX}+\phi^2 B_{XX}\right]^{-1}$. Note that $W_{XX} = X'QX, B_{XX} = X'\left(P-\bar{J}_{NT}\right)X$ and $\phi^2 = \sigma_v^2/\sigma_1^2$. Also, the Within estimator of β is $\underline{\beta}_{\text{Within}} = W_{XX}^{-1}W_{Xy}$ and the Between estimator of β is $\hat{\beta}_{\text{Between}} = B_{XX}^{-1}B_{Xy}$. This shows that $\hat{\beta}_{\text{GLS}}$ is a matrix weighted average of $\tilde{\beta}_{\text{Within and}}\ \hat{\beta}_{\text{Between}}$ weighing each estimate by the inverse of its corresponding variance. In fact

$$\hat{\beta}_{\text{GLS}} = W_1\tilde{\beta}_{\text{Within}} + W_2\hat{\beta}_{\text{Between}} \tag{2.25}$$

(2.31) where

$$W_1 = [W_{XX} + \phi^2 B_{XX}]^{-1} W_{XX}$$

and

$$W_2 = [W_{XX} + \phi^2 B_{XX}]^{-1} (\phi^2 B_{XX}) = I - W_1$$

This was demonstrated by Maddala (1971). Note that (i) if $\sigma_\mu^2 = 0$ then $\phi^2 = 1$ and $\hat{\beta}_{\text{GLS}}$ reduces to $\hat{\beta}_{\text{OLs.}}$ (ii) If $T \to \infty$, then $\phi^2 \to 0$ and $\hat{\beta}_{\text{GLS}}$ tends to $\tilde{\beta}_{\text{Within.}}$ Also, if W_{XX} is huge compared to B_{XX} then $\hat{\beta}_{\text{GLS}}$ will be close to $\tilde{\beta}_{\text{Within}}$. However, if B_{XX} dominates W_{XX} then $\hat{\beta}_{\text{GLS}}$ tends to $\hat{\beta}_{\text{Between.}}$ In other words, the Within estimator ignores the Between variation, and the Between estimator ignores the Within variation. The OLS estimator gives equal weight to the Between and Within variations. From (2.30), it is clear that $\text{var}\left(\tilde{\beta}_{\text{Within}}\right) - \text{var}\left(\hat{\beta}_{\text{GLS}}\right)$ is a positive semidefinite matrix, since ϕ^2 is positive. However, as $T \to \infty$ for any fixed $N, \phi^2 \to 0$ and both $\hat{\beta}_{\text{GLS}}$ and $\tilde{\beta}_{\text{Within}}$ have the same asymptotic variance.

Another estimator of the variance components was suggested by Nerlove (1971a). His zuggestion is to estimate σ^2 as $\sum^N \cdot(\hat{u} \cdot -\hat{\pi})^2/(N-1)$ where $\hat{u}.$ are the dummv coefficients estimates from the LSDV regression. σ_v^2 is estimated from the Within residual sums of squares divided by NT without correction for degrees of freedom. [4]

Note that, except for Nerlove's (1971a) method, one has to retrieve $\hat{\sigma}_\mu^2$ as $\left(\hat{\sigma}_1^2 - \hat{\sigma}_v^2\right)/T$. In this case, there is no guarantee that the estimate of $\hat{\sigma}_\mu^2$ would be nonnegative. Searle (1971) has an extensive discussion of the problem of negative estimates of the variance components in the biometrics literature. One solution is to replace these negative estimates by zero. This in fact is the suggestion of the Monte Carlo study by Maddala and Mount (1973). This study finds that negative estimates occurred only when the true σ_μ^2 was small and close to zero. In these cases OLS is still a viable estimator. Therefore, replacing negative $\hat{\sigma}_\mu^2$ by zero is not a bad sin after all, and the problem is dismissed as not being serious. [5]

How about the properties of the various feasible GLS estimators of β Under the random effects model, GLS based on the true variance components is BLUE, and all the feasible GLS estimators considered are asymptotically efficient as either N or $N \to \infty$. Maddala and Mount (1973) compared OLS, Within, Between, feasible GLS methods, MINQUE, Henderson's method III, true GLS and maximum likelihood estimation using their Monte Carlo study. They found little to choose among the various feasible GLS estimators in small samples and argued in favor of methods that were easier to compute. MINQUE was dismissed as more difficult to compute and the applied researcher given one shot at the data was warned to compute at least two methods of estimation, like an ANOVA feasible GLS and maximum likelihood to ensure that they do not yield drastically different results. If they do give different results, the authors diagnose misspecification.

Taylor (1980) derived exact finite sample results for the one-way error component model. He compared the Within estimator with the Swamy–Arora feasible GLS estimator. He found the following important results:

(1) Feasible GLS is more efficient than LSDV for all but the fewest degrees of freedom.
(2) The variance of feasible GLS is never more than 17% above the Cramer–Rao lower bound.
(3) More efficient estimators of the variance components do not necessarily yield more efficient feasible GLS estimators.

These finite sample results are confirmed by the Monte Carlo experiments carried out by Maddala and Mount (1973) and Baltagi (1981a).

Bellmann, Breitung and Wagner (1989) consider the bias in estimating the variance components using the Wallace and Hussain (1969) method due to the replacement of the true disturbances by OLS residuals, also the bias in the regression coefficients due to the use of estimated variance components rather than

the true variance components. The magnitude of this bias is estimated using bootstrap methods for two economic applications. The first application relates product innovations, import pressure and factor inputs using a panel at the industry level. The second application estimates the earnings of 936 full-time working German males based on the first and second wave of the German Socio-Economic Panel. Only the first application revealed considerable bias in estimating $ _ ^2 $. However, this did not affect the bias much in the corresponding regression coefficients

2.3.1 Fixed vs Random

Having discussed the fixed effects and the random effects models and the assumptions underlying them, the reader is left with the daunting question, which one to choose? This is not as easy a choice as it might seem. In fact, the fixed versus random effects issue has generated a hot debate in the biometrics and statistics literature which has spilled over into the panel data econometrics literature. Mundlak (1961) and Wallace and Hussain (1969) were early proponents of the fixed effects model and Balestra and Nerlove (1966) were advocates of the random error component model. In Chapter 4, we will study a specification test proposed by Hausman (1978) which is based on the difference between the fixed and random effects estimators. Unfortunately, applied researchers have interpreted a rejection as an adoption of the fixed effects model and nonrejection as an adoption of the random effects model.6 Chamberlain (1984) showed that the fixed effects model imposes testable restrictions on the parameters of the reduced form model and one should check the validity of these restrictions before adopting the fixed effects model (see Chapter 4). Mundlak (1978) argued that the random effects model assumes exogeneity of all the regressors with the random individual effects. In contrast, the fixed effects model allows for endogeneity of all the regressors with these individual effects. So, it is an "all" or "nothing" choice of exogeneity of the regressors and the individual effects, see Chapter 7 for a more formal discussion of this subject.

Hausman and Taylor (1981) allowed for some of the regressors to be correlated with the individual effects, as opposed to the all or nothing choice. These over-identification restrictions are testable using a Hausman-type test (see Chapter 7). For the applied researcher, performing fixed effects and random effects and the associated Hausman test reported in standard packages like Stata, LIMDEP, TSP, etc., the message is clear: Do not stop here. Test the restrictions implied by the fixed effects model derived by Chamberlain (1984) (see Chapter 4) and check whether a Hausman and Taylor (1981) specification might be a viable alternative (see Chapter 7).

Chapter 3

The Two-way Error Component Regression Model

3.1 NTRODUCTION

Wallace and Hussain (1969), Nerlove (1971b) and Amemiya (1971), among others, the regression model given by (2.1), but with two-way error components disturbances:

$$u_{it} = \mu_i + \lambda_t + v_{it} \quad i = 1, ..., N; t = 1, ..., T \tag{3.1}$$

(3.1)

where μ_i denotes the unobservable individual effect discussed in Chapter 2, λ_t denotes the unobservable time effect and v_{it} is the remainder stochastic disturbance term. Note that λ_t is individual-invariant and it accounts for any time-specific effect that is not included in the regression. For example, it could account for strike year effects that disrupt production; oil embargo effects that disrupt the supply of oil and affect its price; Surgeon General reports on the ill-effects of smoking, or government laws restricting smoking in public places, all of which could affect consumption behavior. In vector form, (3.1) can be written as

$$u = Z_\mu \mu + Z_\lambda \lambda + v \tag{3.2}$$

(3.2)

where Z_μ, μ and v were defined earlier. $Z_\lambda = i_N \otimes I_T$ is the matrix of time dummies that one may include in the regression to estimate the λ_t if they are fixed parameters, and $\lambda' = (\lambda_1, ..., \lambda_T)$ Note that

$$Z_\lambda Z_\lambda' = J_N \otimes I_T$$

and the projection on Z_λ is

$$Z_\lambda \left(Z_\lambda' Z_\lambda\right)^{-1} Z_\lambda' = \bar{J}_N \otimes I_T$$

This last matrix averages the data over individuals, i.e., if we regress y on Z_λ, the predicted values are given by

$$(\bar{J_N} \otimes I_T) y$$

which has typical element $\bar{y}_{.t} = \sum_{i=1}^{N} y_{it}/N$.

3.2 THE FIXED EFFECTS MODEL

If the μ_i and λ_t are assumed to be fixed parameters to be estimated and the remainder disturbances stochastic with $v_{it} \sim \text{IID}(0, \sigma_v^2)$, then (3.1) represents a two-way fixed effects error component model. The X_{it} are assumed independent of the v_{it} for all i and t. Inference in this case is conditional on the particular N individuals and over the specific time periods observed. Recall that Z_λ, the matrix of time dummies, is $NT \times T$. If N or T is large, there will be too many dummy variables in the regression $\{(N-1)+(T-1)\}$ of them, and this causes an enormous loss in degrees of freedom. In addition, this attenuates the problem of multicollinearity among the regressors. Rather than invert a large $(N+T+K-1)$ matrix, one can obtain the fixed effects estimates of β by performing the following Within transformation given by Wallace and Hussain (1969):

$$Q = E_N \otimes E_T = I_N \otimes I_T - I_N \otimes \bar{J}_T - \bar{J}_N \otimes I_T + \bar{J}_N \otimes \bar{J}_T \tag{3.3}$$

(3.3)

where $E_N = I_N - \bar{J}_N$ and $E_T = I_T - \bar{J}_T$. This transformation "sweeps" the μ_i and λ_t effects. In fact, $\tilde{y} = Qy$ has a typical element $\tilde{y}_{it} = (y_{it} - \bar{y}_{i.} - \bar{y}_{it} + \bar{y}..)$ where $\bar{y}_{..} = \sum_i \sum_t y_{it}/\ NT$, and one would perform the regression of $\tilde{y} = Qy$ on $\widetilde{X} = QX$ to get the Within estimator $\tilde{\beta} = (X'QX)^{-1} X'Qy$

Note that by averaging the simple regression given in (2.8) over individuals, we get

$$\bar{y}_{.t} = \alpha + \beta \bar{x}_{.t} + \lambda_t + \bar{v}_{.t} \tag{3.4}$$

(3.4)

where we have utilized the restriction that $\sum_i \mu_i = 0$ to avoid the dummy variable trap. Similarly the averages defined in (2.9) and (2.11) still hold using $\sum_t \lambda_t = 0$ and one can deduce that

$$(y_{it} - \bar{y}_{i.} - \bar{y}_{.t} + \bar{y}_{..}) = (x_{it} - \bar{x}_{i.} - \bar{x}_{.t} + \bar{x}_{..})\,\beta + (v_{it} - \bar{v}_{i.} - \bar{v}_{.t} + \bar{v}_{..}) \tag{3.5}$$

(3.5)

OLS on this model gives $\tilde{\beta}$ the Within estimator for the two-way model. Once again, the Within estimate of the intercept can be deduced from

$$\tilde{\alpha} = \bar{y}_{.} - \tilde{\beta}\bar{x}$$

and those of μ_i and λ_i are given by

$$\tilde{\mu}_i = (\bar{y}_{i.} - \bar{y}_{..}) - \tilde{\beta}\,(\bar{x}_{i.} - \bar{x}_{..}) \tag{3.6}$$

(3.6)

$$\tilde{\lambda}_t = (\bar{y}_{.t} - \bar{y}_{..}) - \tilde{\beta}\,(\bar{x}_{.t} - \bar{x}_{..}) \tag{3.7}$$

(3.7)

Note that the Within estimator cannot estimate the effect of time-invariant and individualinvariant variables because the Q transformation wipes out these variables. If the true model is a two-way fixed effects model as in (3.2), then OLS on (2.1) yields biased and inconsistent estimates of the regression coefficients. OLS ignores both sets of dummy variables, whereas the one-way fixed effects estimator considered in Chapter 2 ignores only the time dummies. If these time dummies are statistically significant, the one-way fixed effects estimator will also suffer from omission bias.

3.2.1 Testing for Fixed Effects

As in the one-way error component model case, one can test for joint significance of the dummy variables:

$$H_0 : \mu_1 = ... = \mu_{N-1} = 0 \quad \text{and} \quad \lambda_1 = ... = \lambda_{T-1} = 0$$

The restricted residual sums of squares (RRSS) is that of pooled OLS and the unrestricted residual sums of squares (URSS) is that from the Within regression in (3.5). In this case,

$$F_1 = \frac{(\text{RRSS} - \text{URSS})/(N+T-2)}{\text{URSS}/(N-1)(T-1)-K} \overset{H_0}{\sim} F_{(N+T-2),(N-1)(T-1)-K} \tag{3.8}$$

(3.8)

Next, one can test for the existence of individual effects allowing for time effects, i.e. $H_2 : \mu_1 == u_N = 0$ allowing $\lambda \neq 0$ for $t = 1 ... \quad T-1$

The URSS is still the Within residual sum of squares. However, the RRSS is the regression with time-series dummies only, or the regression based upon

$$(y_{it} - \bar{y}_{.t}) = (x_{it} - \bar{x}_{.t})\beta + (u_{it} - \bar{u}_{.t}) \tag{3.9}$$

(3.9)

In this case the resulting F -statistic is $F_2 \overset{H_0}{\sim} F_{(N-1),(N-1)(T-1)-K}$. Note that F_2 differs from F_0 in (2.12) in testing for $\mu_i = 0$. The latter tests $H_0 : \mu_i = 0$ assuming that $\lambda_t = 0$, whereas the former tests $H_2 : \mu_i = 0$ allowing $\lambda_t \neq 0$ for $t = 1, ..., T-1$. Similarly, one can test for the existence of time effects allowing for individual effects, i.e.

$$H_3 : \lambda_1 = ... = \lambda_{T-1} = 0 \quad \text{allowing} \quad \mu_i \neq 0; i = 1, ..., (N-1) \tag{3.10}$$

The RRSS is given by the regression in (2.10), while the URSS is obtained from the regression (3.5). In this case, the resulting F-statistic is $F_3 \sim^{H_0} F_{(T-1),(N-1)(T-1)-K}$.

Computational Warning

As in the one-way model, s^2 from the regression in (3.5) as obtained from any standard regression package has to be adjusted for loss of degrees of freedom. In this case, one divides by$ (N − 1)(T − 1) − K $ and multiplies by $(NT − K) $to get the proper variance–covariance matrix of the Within estimator.

3.3 THE RANDOM EFFECTS MODEL

If $\mu_i \sim \text{IID}\left(0, \sigma_\mu^2\right), \lambda_t \sim \text{IID}\left(0, \sigma_\lambda^2\right)$ and $v_{it} \sim \text{IID}\left(0, \sigma_v^2\right)$ independent of each other, then this is the two-way random effects model. In addition, X_{it} is independent of μ_i, λ_t and v_{it} for all i and t. Inference in this case pertains to the large population from which this sample was randomly drawn. From (3.2), one can compute the variance-covariance matrix

$$\begin{aligned} \Omega = E(uu') &= Z_\mu E(\mu\mu') Z_\mu' + Z_\lambda E(\lambda\lambda') Z_\lambda' + \sigma_v^2 I_{NT} \\ &= \sigma_\mu^2 (I_N \otimes J_T) + \sigma_\lambda^2 (J_N \otimes I_T) + \sigma_v^2 (I_N \otimes I_T) \end{aligned} \tag{3.11}$$

(3.10)

The disturbances are homoskedastic with $\text{var}(u_{it}) = \sigma_\mu^2 + \sigma_\lambda^2 + \sigma_v^2$ for all i and t,

$$\begin{aligned} \operatorname{cov}\left(u_{it}, u_{js}\right) &= \sigma_\mu^2 \quad i = j, t \neq s \\ &= \sigma_\lambda^2 \quad i \neq j, t = s \end{aligned} \tag{3.12}$$

(3.11)

and zero otherwise. This means that the correlation coefficient

$$\begin{aligned} \operatorname{correl}\left(u_{it}, u_{js}\right) &= \sigma_\mu^2 / \left(\sigma_\mu^2 + \sigma_\lambda^2 + \sigma_v^2\right) \quad i = j, t \neq s \\ &= \sigma_\lambda^2 / \left(\sigma_\mu^2 + \sigma_\lambda^2 + \sigma_v^2\right) \quad i \neq j, t = s \\ &= 1 \quad i = j, t = s \\ &= 0 \quad i \neq j, t \neq s \end{aligned} \tag{3.13}$$

(3.12)

In order to get Ω^{-1}, we replace J_N by $N\bar{J}_N, I_N$ by $E_N + \bar{J}_N, J_T$ by $T\bar{J}_T$ and I_T by $E_T + \bar{J}_T$ and collect terms with the same matrices. This gives

$$\Omega = \sum_{i=1}^{4} \lambda_i Q_i \tag{3.14}$$

(3.13)

where $\lambda_1 = \sigma_v^2, \lambda_2 = T\sigma_\mu^2 + \sigma_v^2, \lambda_3 = N\sigma_\lambda^2 + \sigma_v^2$ and $\lambda_4 = T\sigma_\mu^2 + N\sigma_\lambda^2 + \sigma_v^2$. Correspond- ingly, $Q_1 = E_N \otimes E_T, Q_2 = E_N \otimes \bar{J}_T, Q_3 = \bar{J}_N \otimes E_T$ and $Q_4 = \bar{J}_N \otimes \bar{J}_T$, respectively. The λ_i are the distinct characteristic roots of Ω and the Q_i are the corresponding matrices of eigenprojectors. λ_1 is of multiplicity $(N-1)(T-1), \lambda_2$ is of multiplicity $(N-1), \lambda_3$ is of multiplicity $(T-1)$ and λ_4 is of multiplicity 1.[1] Each Q_i is symmetric and idempotent with its rank equal to its trace. Moreover, the Q_i are pairwise orthogonal and sum to the identity matrix. The advantages of this spectral decomposition are that

$$\Omega^r = \sum_{i=1}^{4} \lambda_i^r Q_i \tag{3.15}$$

(3.14)

where r is an arbitrary scalar so that

$$\sigma_v \Omega^{-1/2} = \sum_{i=1}^{4} \left(\sigma_v / \lambda_i^{1/2}\right) Q_i \tag{3.16}$$

(3.15)

and the typical element of $y^* = \sigma_\nu \Omega^{-1/2} y$ is given by

$$y_{it}^* = y_{it} - \theta_1 \bar{y}_{i.} - \theta_2 \bar{y}_{.t} + \theta_3 \bar{y}_{..} \tag{3.17}$$

(3.16)

where $\theta_1 = 1 - \left(\sigma_v / \lambda_2^{1/2}\right), \theta_2 = 1 - \left(\sigma_v / \lambda_3^{1/2}\right)$ and $\theta_3 = \theta_1 + \theta_2 + \left(\sigma_v / \lambda_4^{1/2}\right) - 1$. As a result, GLS can be obtained as OLS of y^* on Z^*, where $Z^* = \sigma_\nu \Omega^{-1/2} Z$. This transformation was first derived by Fuller and Battese (1974), see also Baltagi (1993).

The best quadratic unbiased (BQU) estimators of the variance components arise naturally from the fact that $Q_i u \sim (0, \lambda_i Q_i)$. Hence,

$$\hat{\lambda}_i = u' Q_i u / \operatorname{tr}(Q_i) \tag{3.18}$$

(3.17)

is the BQU estimator of λ_i for $i = 1, 2, 3$. These ANOVA estimators are minimum variance unbiased (MVU) under normality of the disturbances (see Graybill, 1961). As in the one-way error component model, one can obtain feasible estimates of the variance components by replacing the true disturbances by OLS residuals (see Wallace and Hussain, 1969). OLS is still an unbiased and consistent estimator under the random effects model, but it is inefficient and results in biased standard errors and t -statistics. Alternatively, one could substitute the Within residuals with $\tilde{u} = y - \tilde{\alpha}\iota_{NT} - X\tilde{\beta}$, where $\tilde{\alpha} = \bar{y}_{..} - \bar{X}'_{..}\tilde{\beta}$ and $\tilde{\beta}$ is obtained by the regression in (3.5). This is the method proposed by Amemiya (1971). In fact, Amemiya (1971) shows that the Wallace and Hussain (1969) estimates of the variance components have a different asymptotic distribution from that knowing the true disturbances, while the Amemiya (1971) estimates of the variance components have the same asymptotic distribution as that knowing the true disturbances:

$$\begin{pmatrix} \sqrt{NT}\left(\hat{\sigma}_v^2 - \sigma_v^2\right) \\ \sqrt{N}\left(\hat{\sigma}_\mu^2 - \sigma_\mu^2\right) \\ \sqrt{T}\left(\hat{\sigma}^2 - \sigma_\mu^2\right) \end{pmatrix} \sim N\left(0, \begin{pmatrix} 2\sigma_v^4 & 0 & 0 \\ 0 & 2\sigma_\mu^4 & 0 \\ 0 & 0 & 2\sigma^4 \end{pmatrix}\right) \tag{3.19}$$

(3.18)

SubstitutingOLSorWithin residuals instead of the true disturbances in (3.17) introduces bias in the corresponding estimates of the variance components. The degrees of freedom corrections that make these estimates unbiased depend upon traces of matrices that involve the matrix of regressors X. These corrections are given in Wallace and Hussain (1969) and Amemiya (1971), respectively. Alternatively, one can infer these correction terms from the more general unbalanced error component model considered in Chapter 9. Swamy and Arora (1972) suggest running three least squares regressions and estimating the variance components from the corresponding mean square errors of these regressions. The first regression corresponds to the Within regression which transforms the original model by $Q_1 = E_N \otimes E_T$. This is equivalent to the regression in (3.5), and yields the following estimate of σ_v^2 :

$$\hat{\lambda}_1 = \hat{\sigma}_v^2 = \left[y' Q_1 y - y' Q_1 X \left(X' Q_1 X\right)^{-1} X' Q_1 y\right] / [(N-1)(T-1) - K] \tag{3.20}$$

(3.19)
(à arranger)

The second regression is the Between individuals regression which transforms the original model by

$$Q_2 = E_N \otimes \bar{J}_T$$

This is equivalent to the regression of

$$(\bar{y}_{i.} - \bar{y}_{i..})$$

on

$$(\bar{X}_{i.} - \bar{X}_{i..})$$

and yields the following estimate of $ _2=T _v^2 + _v^2 $:

$$\hat{\lambda}_2 = \left[y' Q_2 y - y' Q_2 X \left(X' Q_2 X\right)^{-1} X' Q_2 y\right] / [(N-1) - K] \tag{3.21}$$

(3.20)

from which one obtains $\hat{\sigma}_\mu^2 = \left(\hat{\lambda}_2 - \hat{\sigma}_v^2\right)/T$. The third regression is the Between time-periods regression which transforms the original model by $Q_3 = \bar{J}_N \otimes E_T$. This is equivalent to the regression of $(\bar{y}_{.t} - \bar{y}_{..})$ on $\left(\bar{X}_{.t} - \bar{X}_{..}\right)$ and yields the following estimate of $\lambda_3 = N\sigma_\lambda^2 + \sigma_v^2$

$$\hat{\lambda}_3 = \left[y'Q_3y - y'Q_3X\left(X'Q_3X\right)^{-1}X'Q_3y\right]/[(T-1)-K] \tag{3.22}$$

(3.21)

from which one obtains

$$(\widehat{\widehat{\sigma_\lambda^2}} = \hat{\hat{\lambda}}_3 - \hat{\hat{\lambda}}_v)/N$$

Stacking the three transformed regressions just performed yields

$$\begin{pmatrix} Q_1y \\ Q_2y \\ Q_3y \end{pmatrix} = \begin{pmatrix} Q_1X \\ Q_2X \\ Q_3X \end{pmatrix}\beta + \begin{pmatrix} Q_1u \\ Q_2u \\ Q_3u \end{pmatrix} \tag{3.23}$$

(3.22)

since $Q_i\iota_{NT} = 0$ for $i = 1, 2, 3$, and the transformed error has mean 0 and variance-covariance matrix given by diag $[\lambda_i Q_i]$ with $i = 1, 2, 3$. Problem 3.4 asks the reader to show that OLS on this system of $3NT$ observations yields the same estimator of β as OLS on the pooled model (2.3). Also, GLS on this system of equations (3.22) yields the same estimator of β as GLS on (2.3). In fact,

$$\begin{aligned}\hat{\beta}_{\text{GLS}} &= \left[\left(X'Q_1X\right)/\sigma_v^2 + \left(X'Q_2X\right)/\lambda_2 + \left(X'Q_3X\right)/\lambda_3\right]^{-1} \\ &\quad \times \left[\left(X'Q_1y\right)/\sigma_v^2 + \left(X'Q_2y\right)/\lambda_2 + \left(X'Q_3y\right)/\lambda_3\right] \\ &= \left[W_{XX} + \phi_2^2B_{XX} + \phi_3^2C_{XX}\right]^{-1}\left[W_{Xy} + \phi_2^2B_{Xy} + \phi_3^2C_{Xy}\right]\end{aligned} \tag{3.24}$$

(3.23)

with var $\left(\hat{\beta}_{\text{GLS}}\right) = \sigma_v^2\left[W_{XX} + \phi_2^2B_{XX} + \phi_3^2C_{XX}\right]^{-1}$. Note that $W_{XX} = X'Q_1X, B_{XX} = X'Q_2X$ and $C_{XX} = X'Q_3X$ with $\phi_2^2 = \sigma_v^2/\lambda_2, \phi_3^2 = \sigma_v^2/\lambda_3$. Also, the Within estimator of β is $\tilde{\beta}_W = W_{XX}^{-1}W_{Xy}$, the Between individuals estimator of β is $\hat{\beta}_B = B_{XX}^{-1}B_{Xy}$ and the Between timeperiods estimator of β is $\hat{\beta}_C = C_{XX}^{-1}C_{Xy}$. This shows that $\hat{\beta}_{\text{GLS}}$ is a matrix-weighted average of $\tilde{\beta}_W, \hat{\beta}_B$ and $\hat{\beta}_C$. In fact,

$$\hat{\beta}_{\text{GLS}} = W_1\tilde{\beta}_W + W_2\hat{\beta}_B + W_3\hat{\beta}_C \tag{3.25}$$

(3.24)

where

$$\begin{aligned}W_1 &= \left[W_{XX} + \phi_2^2B_{XX} + \phi_3^2C_{XX}\right]^{-1}W_{XX} \\ W_2 &= \left[W_{XX} + \phi_2^2B_{XX} + \phi_3^2C_{XX}\right]^{-1}\left(\phi_2^2B_{XX}\right) \\ W_3 &= \left[W_{XX} + \phi_2^2B_{XX} + \phi_3^2C_{XX}\right]^{-1}\left(\phi_3^2C_{XX}\right)\end{aligned} \tag{3.26}$$

This was demonstrated by Maddala (1971). Note that (i) if $\sigma_\mu^2 = \sigma_\lambda^2 = 0, \phi_2^2 = \phi_3^2 = 1$ and $\hat{\beta}_{\text{GLS}}$ reduces to $\hat{\beta}_{\text{OLs}}$; (ii) as T and $N \to \infty, \phi_2^2$ and $\phi_3^2 \to 0$ and $\hat{\beta}_{\text{GLS}}$ tends to $\tilde{\beta}_W$; (iii) if $\phi_2^2 \to \infty$ with ϕ_3^2 finite, then $\hat{\beta}_{\text{GLS}}$ tends to $\hat{\beta}_B$; (iv) if $\phi_3^2 \to \infty$ with ϕ_2^2 finite, then $\hat{\beta}_{\text{GLS}}$ tends to $\hat{\beta}_C$. Wallace and Hussain (1969) compare $\hat{\beta}_{\text{GLS}}$ and $\tilde{\beta}_{\text{Within}}$ in the case of nonstochastic (repetitive) X and find that both are (i) asymptotically normal, (ii) consistent and unbiased and that

(iii) $\hat{\beta}_{\text{GLS}}$ has a smaller generalized variance (i.e. more efficient) in finite samples. In the case of non-stochastic (nonrepetitive) X they find that both $\hat{\beta}_{\text{GLS}}$ and $\tilde{\beta}_{\text{Within}}$ are consistent, asymptotically unbiased and have equivalent asymptotic variance-covariance matrices, as both N and $T \to \infty$. The last statement can be proved as follows: the limiting variance of the GLS estimator is

$$\frac{1}{NT} \lim_{\substack{N\to\infty \\ T\to\infty}} \left(X'\Omega^{-1}X/NT\right)^{-1} = \frac{1}{NT} \lim_{\substack{N\to\infty \\ T\to\infty}} \left[\sum_{i=1}^{3} \frac{1}{\lambda_i} \left(X'Q_iX/NT\right)\right]^{-1} \tag{3.27}$$

(3.25)

but the limit of the inverse is the inverse of the limit, and

$$\lim_{\substack{N\to\infty \\ T\to\infty}} \frac{X'Q_iX}{NT} \quad \text{for } i = 1, 2, 3 \tag{3.28}$$

(3.26)

all exist and are positive semidefinite, since

$$\lim_{\substack{N\to\infty \\ T\to\infty}} \left(X'X/NT\right)$$

is assumed finite and positive definite. Hence

$$\lim_{\substack{N\to\infty \\ T\to\infty}} \frac{1}{\left(N\sigma_\lambda^2 + \sigma_v^2\right)} \left(\frac{X'Q_3X}{NT}\right) = 0$$

and

$$\lim_{\substack{N\to\infty \\ T\to\infty}} \frac{1}{\left(T\sigma_\mu^2 + \sigma_v^2\right)} \left(\frac{X'Q_2X}{NT}\right) = 0$$

Therefore the limiting variance of the GLS estimator becomes

$$\frac{1}{NT} \lim_{\substack{N\to\infty \\ T\to\infty}} \sigma_v^2 \left(\frac{X'Q_1X}{NT}\right)^{-1}$$

which is the limiting variance of the Within estimator. One can extend Nerlove's (1971a) method for the one-way model, by estimating σ_μ^2 as $\sum_{i=1}^{N} \left(\hat{\mu}_i - \hat{\mu}\right)^2/(N-1)$ and σ_λ^2 as $\sum_{t=1}^{T} \left(\hat{\lambda}_t - \bar{\hat{\lambda}}\right)^2/(T-1)$ where the $\hat{\mu}_i$ and $\hat{\lambda}_t$ are obtained as coefficients from the least squares dummy variables regression (LSDV). σ_v^2 is estimated from the Within residual sums of squares divided by NT. Baltagi (1995, appendix 3) develops two other methods of estimating the variance components. The first is Rao's (1970) minimum norm quadratic unbiased estimation (MINQUE) and the second is Henderson's method III as described by Fuller and Battese (1973). These methods require more notation and development and may be skipped in a brief course on this subject. Chapter 9 studies these estimation methods in the context of an unbalanced error component model.

Baltagi (1981a) performed a Monte Carlo study on a simple regression equation with twoway error component disturbances and studied the properties of the following estimators: OLS, the Within estimator and six feasible GLS estimators denoted by WALHUS, AMEMIYA, SWAR, MINQUE, FUBA and NERLOVE corresponding to the methods developed by Wallace and Hussain (1969), Amemiya (1971), Swamy and Arora (1972), Rao (1972), Fuller and Battese (1974) and Nerlove (1971a), respectively. The mean square error of these estimators was computed relative to that of true GLS, i.e. GLS knowing the true variance

components. To review some of the properties of these estimators: OLS is unbiased, but asymptotically inefficient, and its standard errors are biased; see Moulton (1986) for the extent of this bias in empirical applications. In contrast, the Within estimator is unbiased whether or not prior

information about the variance components is available. It is also asymptotically equivalent to the GLS estimator in case of weakly nonstochastic exogenous variables. Early in the literature, Wallace and Hussain (1969) recommended the Within estimator for the practical researcher, based on theoretical considerations but more importantly for its ease of computation. In Wallace and Hussain's (1969, p. 66) words the "covariance estimators come off with a surprisingly clear bill of health". True GLS is BLUE, but the variance components are usually not known and have to be estimated. All of the feasible GLS estimators considered are asymptotically efficient. In fact, Prucha (1984) showed that as long as the estimate of σ_v^2 is consistent, and the probability limits of the estimates σ_μ^2 and σ_λ^2 are finite, the corresponding feasible GLS estimator is asymptotically efficient. Also, Swamy and Arora (1972) proved the existence of a family of asymptotically efficient two-stage feasible GLS estimators of the regression coefficients. Therefore, based on asymptotics only, one cannot differentiate among these twostage GLS estimators. This leaves undecided the question of which estimator is the best to use. Some analytical results were obtained by Swamy (1971) and Swamy and Arora (1972). These studies derived the relative efficiencies of (i) SWAR with respect to OLS, (ii) SWAR with respect to Within and (iii) Within with respect to OLS. Then, for various values of N, T, the variance components, the Between groups, Between time-periods and Within groups sums of squares of the independent variable, they tabulated these relative efficiency values (see Swamy, 1971 , chapters II and III; Swamy and Arora, 1972, p. 272). Among their basic findings is the fact that, for small samples, SWAR is less efficient than OLS if σ_μ^2 and σ_λ^2 are small. Also, SWAR is less efficient than Within if σ_μ^2 and σ_λ^2 are large. The latter result is disconcerting, since Within, which uses only a part of the available data, is more efficient than SWAR, a feasible GLS estimator, which uses all of the available data.

3.4 REFERENCES

Chapter 4

Test of Hypotheses with Panel Data

4.1 TESTS FOR POOLABILITY OF THE DATA

The question of whether to pool the data or not naturally arises with panel data. The restricted model is the pooled model given by (2.3) representing a behavioral equation with the same parameters over time and across regions. The unrestricted model, however, is the same behavioral equation but with different parameters across time or across regions. For example, Balestra and Nerlove (1966) considered a dynamic demand equation for natural gas across 36 states over six years. In this case, the question of whether to pool or not to pool boils down to the question of whether the parameters of this demand equation vary from one year to the other over the six years of available data. One can have a behavioral equation whose parameters may vary across regions. For example, Baltagi and Griffin (1983) considered panel data on motor gasoline demand for 18 OECD countries. In this case, one is interested in testing whether the behavioral relationship predicting demand is the same across the 18 OECD countries, i.e. the parameters of the prediction equation do not vary from one country to the other.

These are but two examples of many economic applications where time-series and crosssection data may be pooled. Generally, most economic applications tend to be of the first type, i.e. with a large number of observations on individuals, firms, economic sectors, regions, industries and countries but only over a few time periods. In what follows, we study the tests for the poolability of the data for the case of pooling across regions keeping in mind that the other case of pooling over time can be obtained in a similar fashion. For the unrestricted model, we have a regression equation for each region given by

$$y_i = Z_i\delta_i + \mu_i \text{i=1,2,...,N} \tag{4.1}$$

(4.1)

where

$$y' = (y_i 1, ..., y_{iT})$$

, $Z_i = [i_T, X_i]$ ans X_i is \$T $\times$K \$.

The important thing to notice is that δ_1 is different for every regional equation. We want to test the hypothesis $H_0 : \delta_l = \delta$ for all i, so that undcr H_0 we can write the restricted model given in (4.1) as

$$y = Z\delta + \mu \tag{4.2}$$

where

$$Z' = (Z_1', Z_2', \dots, Z_N')$$

and

$$u' = (u'_1, u'_2, ..., u'_N)$$

. The unrestricted model can also be written as

$$y = \begin{pmatrix} Z_1 & 0 & ... & 0 \\ 0 & Z_2 & ... & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & ... & Z_N \end{pmatrix} \begin{pmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_N \end{pmatrix} + u = Z^* \delta^* + u \tag{4.3}$$

where

$$\delta^{*\prime} = (\delta'_1, \delta'_2, ..., \delta'_N)$$

and

$$Z = Z^* I^*$$

with

$$I^* = (\iota_N \otimes I_{K'})$$

, an

$$NK' \times K'$$

matrix, with

$$K' = K + 1$$

. Hence the variables in Z are all linear combinations of the variables in

$$Z^*$$

.

4.1.1 Test for Poolability under

$$u \sim N(0, \sigma^2 I_{NT})$$

Assumption 4. 1

$$\mu \sim N(0, \sigma^2 I_{NT})$$

Under assumption 4.1, the minimum variance unbiased estimator for δ in equation (4.2) is

$$\hat{\delta}_{\text{OLS}} = \hat{\delta}_{mle} = (Z'Z)^{-1} Z'y \tag{4.4}$$

and therefore

$$y = Z\hat{\delta}_{\text{OLS}} + e \tag{4.5}$$

implying that

$$e = \left(I_{NT} - Z(ZZ')^{-1} Z'\right) y = My = M(Z\delta + \mu) = M\mu$$

since $MZ = 0$. Similarly, under assumption 4.1, the MVU for δ_i is given by

$$\hat{\delta}_{i,\text{OLS}} = \hat{\delta}_{i,mle} = (Z'_i Z_i)^{-1} Z'_i y_i \tag{4.6}$$

therefore

$$y_i = Z_i \hat{\delta}_{i,\text{OLS}} + e_i \tag{4.7}$$

implying that

$$e_i = \left(I_T - Z_i \left(Z_i' Z_i\right)^{-1} Z_i'\right) y_i = M_i y_i = M_i \left(Z_i \delta_i + u_i\right) = M_i u_i$$

since $M_{i} Z_{i}=0 $ and this is true for $i = 1, 2, ..., N$. Also, let

$$M^* = I_{NT} - Z^* \left(Z^{*\prime} Z^*\right)^{-1} Z^{*\prime} = \begin{pmatrix} M_1 & 0 & ... & 0 \\ 0 & M_2 & ... & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & ... & M_N \end{pmatrix}$$

One can easily deduce that $y = Z^* \hat{\delta}^* + e^*$ with $e^* = M^* y = M^* u$ and $\hat{\delta}^* = \left(Z^{*\prime} Z^*\right)^{-1} Z^{*\prime} y$. Note that both M and M^* are symmetric and idempotent with $MM^* = M^*$. This easily follows since

$$\begin{aligned} Z \left(Z' Z\right)^{-1} Z' Z^* \left(Z^{*\prime} Z^*\right)^{-1} Z^{*\prime} &= Z \left(Z' Z\right)^{-1} I^{*\prime} Z^{*\prime} Z^* \left(Z^{*\prime} Z^*\right)^{-1} Z^{*\prime} \\ &= Z \left(Z' Z\right)^{-1} Z' \end{aligned} \tag{4.8}$$

(Non Numéroté)

This uses the fact that $Z = Z^* I^*$. Under assumption 4.1, $e'e - e^{*\prime} e^* = u' \left(M - M^*\right) u$ and $e^{*\prime} e^* = u' M^* u$ are independent since $\left(M - M^*\right) M^* = 0$. Also, both quadratic forms when divided by σ^2 are distributed as χ^2 since $\left(M - M^*\right)$ and M^* are idempotent. Dividing these quadratic forms by their respective degrees of freedom and taking their ratio leads to the following test statistic: [1]

$$F_{obs} = \frac{\left(e'e - e^{*\prime}\right) / \left(\text{tr}\left(M\right) - \text{tr}\left(M^*\right)\right)}{e^{*\prime} e^* / \text{tr}\left(M\right)}$$

$$F_{obs} = \frac{\left(e_1' e - e_2' e_1 - e' e_2 - ... - e_N' e\right) / \left(N - 1\right) K'}{\left(e_1' e - e_2' e_1 - e' e_2 - ... - e_N' e\right) / N \left(T - K'\right)} \tag{4.9}$$

Under H_0, F_{obs} is distributed as an

$$F\left(\left(N - 1\right) K', N \left(T - K'\right)\right)$$

Hence the critical region for this test is defined as

$$\left\{ F_{obs} > \left(\left(N - 1\right) K', NT - NK', \alpha_0\right) \right\}$$

where α_0 denotes the level of significance of the test. This is exactly the Chow test presented by Chow (1960) extended to the case of N linear regressions. Therefore if an economist has reason to believe that assumption 4.1 is true, and wants to pool his data across regions, then it is recommended that he or she test for the poolability of the data using the Chow test given in (4.8). However, for the variance component model $\mu \sim \left(0, \Omega\right)$ and not $\left(0, \sigma^2 I_{NT}\right)$ Therefore, even if we assume normality on the disturbances two questions remain: (1) is the Chow test still the right test to perform when $\mu \sim N\left(0, \Omega\right)$? and (2) does the Chow statistic still have an F-distribution when $\mu \sim N\left(0, \Omega\right)$? The answer to the first question is no, the Chow test given in (4.8) is not the right test to perform. However, as will be shown later, a generalized Chow test will be the right test to perform. As for the second question, it is still relevant to ask because it highlights the problem of economists using the Chow test assuming erroneously that μ is $N\left(0, \sigma^2 I_{NT}\right)$ when in fact it is not. For example, Toyoda (1974), in treating the case where the μ_i are heteroskedastic,

found that the Chow statistic given by (4.8) has an approximate $F-$ Distribution where the degree of freedom of the denominator depends upon the true variances. Hence for specific values of these variances, Toyoda demonstrates how wrong it is to apply the Chow test in case of heteroskedastic variances.

Having posed the two questions above, we can proceed along two lines: the first is to find the approximate distribution of the Chow statistic (4.8) in case $\mu \sim N(0, \Omega)$ and therefore show how erroneous it is to use the Chow test in this case (this is not pursued in this book). The second route, and the more fruitful, is to derive the right test to perform for pooling the data in case $\mu \sim N(0, \Omega)$ This is done in the next subsection.

4.1.2 Test for Poolability under the General Assumption

$$u \sim N(0, \Omega)$$

Assumption 4.2

$$u \sim N(0, \Omega)$$

In case Ω is known up to a scalar factor, the test statistic employed for the poolability of the data would be simple to derive. All we need to do is transform our model (under both the null and alternative hypotheses) such that the transformed disturbances have a variance of $\sigma^2 I_{NT}$, then apply the Chow test on the transformed model. The later step is legitimate because the transformed disturbances have homoskedastic variances and the analysis of the previous subsection applies in full. Given \$ =^2 \$, we premultiply the restricted model given in (4.2) by $\sum^{-1/2}$ and we call $\sum^{-1/2} y = \dot{y}$, $\sum^{-1/2} Z = \dot{Z}$ and $\sum^{-1/2} \mu = \dot{\mu}$ Hence

$$\dot{y} = \dot{Z}\delta + \dot{u} \tag{4.10}$$

with

$$E\left(\dot{u}\dot{u}'\right) = \sum^{-1/2} E(u, u') \sum^{-1/2\prime} = \sigma^2 I_{NT}$$

. Similarly, we premultiply the unrestricted model given in (4.3) by $\sum^{-1/2}$ and we call $\sum^{-1/2} Z^* = Z^*$. Therefore

$$\dot{y} = \dot{Z}^*\delta^* + \dot{u} \tag{4.11}$$

with $E(\dot{u}\dot{u}') = \sigma^2 I_{NT}$.

At this stage, we can test $H_0 : \delta_i = \delta$ for every $i = 1, 2, ..., N$, simply by using the Chow statistic, only nowon the transformed models (4.9) and (4.10) since they satisfy assumption 4.1 of homoskedasticity of the normal disturbances. Note that

$$\dot{Z} = \dot{Z}^* I^*$$

, which is simply obtained from \$Z=Z^* I* \$ by premultiplying by $\sum^{-1/2}$.

Defining

$$\dot{M} = I_{NT} - \dot{Z}\left(\dot{Z}'\dot{Z}\right)^{-1}\dot{Z}'$$

and

$$\dot{M^*} = I_{NT} - \dot{Z^*}\left(\dot{Z^{*\prime}}\dot{Z^*}\right)^{-1}\dot{Z^{*\prime}}$$

it is easy to show that $\dot{M}$ and $\dot{M^*}$ are both symmetric, idempotent and such that

$$\dot{M}\dot{M^*} = \dot{M^*}$$

Once again the conditions for lemma 2.2 of Fisher (1970) are satisfied, and the test statistic

$$\dot{F}_{obs} = \frac{\left(\dot{e}'\dot{e} - \dot{e}^{*\prime}\dot{e}^{*}\right) / \left(\operatorname{tr}\left(\dot{M}\right) - \operatorname{tr}\left(\dot{M}^{*}\right)\right)}{\dot{e}^{*\prime}\dot{e}^{*} / \operatorname{tr}\left(\dot{M}^{*}\right)} \sim F\left(\left(N-1\right)K', N\left(T-K'\right)\right) \tag{4.12}$$

where $\dot{e} = \dot{y} - \dot{Z}\hat{\dot{\delta}}_{OLS}$ and $\hat{\dot{\delta}}_{OLS} = \left(\dot{Z'}\dot{Z}\right)^{-1}\dot{Z}'\dot{y}$ implying that $\dot{e} = \dot{M}\dot{y} = \dot{M}\dot{u}$. Similarly,$\dot{e^*} = \dot{y} - \widehat{\dot{Z}^*\dot{\delta}^*_{OLS}}$ and $\widehat{\dot{\delta}^*_{OLS}} = \left(\dot{Z^{*\prime}}\dot{Z}\right)^{-1}\dot{Z^{*\prime}}\dot{y}$ implying that $\dot{e^*} = \dot{M}^*\dot{y} = \dot{M}^*\dot{u}$. Using the fact that $\dot{M}$ and $\dot{M}^*$ are symmetric and idempotent, we can rewrite (4.11) as

$$\dot{F}_{obs} = \frac{\left(\dot{y}'\dot{M}\dot{y} - \dot{y}'\dot{M}^*\dot{y}\right) / \left(N-1\right)K'}{\dot{y}'\dot{M}\dot{y} / N\left(T-K'\right)}$$

$$\dot{F}_{obs} = \frac{\left(y'\sum^{-1/2}\dot{M}\sum^{-1/2}y - \sum^{-1/2}\dot{M}^*\sum^{-1/2}y\right) / \left(N-1\right)K'}{\sum^{-1/2}\dot{M}^*\sum^{-1/2}y / N\left(T-K'\right)} \tag{4.13}$$

But

$$\dot{M} = I_{NT} - \sum^{-1/2} Z\left(Z'\sum^{-1} Z\right)^{-1} Z'\sum^{-1/2'}$$

And

$$\dot{M}^* = I_{NT} - \sum^{-1/2} Z^*\left(Z^*\sum^{-1} Z\right)^{-1} Z^{*\prime}\sum^{-1/2'}$$

so that

$$\sum^{-1/2}\dot{M}\sum_{-1/2} = \sum^{-1} - \sum^{-1} Z\left(Z'\sum^{-1} Z\right)^{-1} Z'$$

and

$$\sum^{-1/2}\dot{M}^*\sum_{-1/2} = \sum^{-1} - \sum^{-1} Z^*\left(Z^{*\prime}\sum^{-1} Z^*\right)^{-1} Z^{*\prime}\sum^{-1}$$

Hence we can write (4.12) in the form

4.2 TESTS FOR INDIVIDUAL AND TIME EFFECTS

4.3 HAUSMAN'S SPECIFICATION TEST

Chapter 5

Analyses

Nous faisons *application* des méthodes présentées dans le chapitre précédant pour l'analyse des données de pannel

Avant de passefr à la modélisation, nous ferons une description de nos variables d'interet d'une manière statique : nos prédicteurs et la variables réponses

5.1 Netoyage de la base des données

Aperçue globale des données

Voici la structure de la base des données

La variables **Goods** a 740 modalités. En observant de près les modalités, on trouve que ces modalités sont rédondantes.Pour cela, il faudra une récodification des cettes variales.

Ainsi, donc les modalités representant la même marchandise seront groupées ensemble , pour des raisons de simplification.

Usage de tm et Stringr pour manipuler les chaines des caractères

```
##  [1] "Autres Marchandises"
```

Table 5.1: Echantillon de la base des données

N°	Country/destination	Year	Goods	Weight	Taxe
1	AFRIQUE DU SUD	2011	GRUES SUR PNEUMATIQUE	13500	0
2	AFRIQUE DU SUD	2011	CAMION FAMIL	12000	0
3	AFRIQUE DU SUD	2011	CAMION SOMUL	24000	0
4	AFRIQUE DU SUD	2013	Café vert arabica k4	183	0
5	AFRIQUE DU SUD	2013	Café vert arabica k4	19520	264771
6	AFRIQUE DU SUD	2013	Café vert arabica k4	19520	272817
7	AFRIQUE DU SUD	2013	Café vert arabica k4	19520	283220
8	AFRIQUE DU SUD	2013	Café vert arabica k4	19520	264142
9	AFRIQUE DU SUD	2013	CAMION	24000	0
10	AFRIQUE DU SUD	2017	Instruments et appareils du n°90.15	654	0

Table 5.2: Table de corrélation entre les variables quantitatives

var1	var2	coef_corr
Weight	Year	-0.1727414
Taxe	Year	-0.1965648
Year	Weight	-0.1727414
Taxe	Weight	0.6699457
Year	Taxe	-0.1965648
Weight	Taxe	0.6699457

```
##  [2] "Bois"
##  [3] "Machines et appareils domestique"
##  [4] "Médicaments et plantes médécinales"
##  [5] "Poissaons, viande et oeufs"
##  [6] "Matériels de construction"
##  [7] "Materiel Informatique et Electroniques"
##  [8] "Véhicules,camions,Motos et acc"
##  [9] "Vetements,tissus et acc et chaussure"
## [10] "boissons, bières et limonades"
## [11] "Machine us Ingsutriel"
## [12] "Article Menange et Campement"
## [13] "sacs, sachetsn emballages"
## [14] "Papiers et fournitures de bureaux"
## [15] "Produits alimentaires,prep et huiles"
## [16] "caféarabica"
## [17] "Minérais et dérivés"
## [18] "engins et tracteurs"
## [19] "Cigarette et papier cigarettes et tabac"
## [20] "constructionprefabriquees"
## [21] "cadreset conteneurs"
## [22] "Pièces de Réchange appareils"
## [23] "Générateurs,baterie et piles"
## [24] "etuis en plastique ou textile"
## [25] "Pétrole et dérivées et huile de graissage"
## [26] "boissons, bières,liqueurs et limonades"
## [27] "produits beaute"
## [28] "peauxdes betes"
```

netoyyage de la variable country_desti qui est un facteur dans le quel nous retrouvons les niveaux rédondants (sur l'identifiant des pays)

Dans la base des données il y a des entreprises que l'on a enregistré à la place des pays. ces genre des cas ont été traité par remplacement avec le *NA* pour **Not Available** et ces dernier on été élagués de la base des données, car nous avions jugé qu' aucune méthode d'imputation n'est applicable pour ce genre de situation. Nous avions fait la même chose pour les variables telles que **Les marchandises**.

Nouvelle base de données pour les analyses

Regroupement des variables pour la synthèse pour rendre la base des données simple à exploiter, éliminer les NA dans les observations telsque les pays et les valeurs pour les marchandises et les taxes.

Les données de panel, ou données longitudinales possèdent les deux dimensions précédentes (individuelle

et temporelle). En effet, il est souvent intéressant d'identifier l'effet associé à chaque individu (un effet qui ne varie pas dans le temps, mais qui varie d'un individu à un autre). Cet effet peut être fixe ou aléatoire.

Par conséquent, le modèle en données de panel s'écrit comme un modèle à double indice qui prend la forme suivante :

$$Y_{it} = \alpha_i \sum_k \beta_{ki} x_{ki} + \epsilon_{it}$$

avec

$$i : 1 \to N$$

et

$$t : 1 \to T$$

La double dimension qu'offrent les données de panel est un atout majeur. En effet, si les données en séries temporelles permettent d'étudier l'évolution des relations dans le temps, elles ne permettent pas de contrôler l'hétérogénéité entre les individus. A l'inverse, les données en coupes transversales permettent d'analyser l'hétérogénéité entre les individus mais elles ne peuvent pas tenir compte des comportements dynamiques, puisque la dimension temporelle est exclue du champ d'analyse.

Ainsi, en utilisant des données de panel, on pourra exploiter les deux sources de variation de l'information statistique : - Temporelle où variabilité intra-individuelle (within) - et individuelle ou variabilité inter-individuelle (Between).

5.2 Modele avec les pays comme individus

5.2.1 Agregation des données avec la méthode reduce

5.2.2 Analyse descriptive des Vararíales

Conversion des données en modèle des panels des données

```
##  [1] "AFRIQUE DU SUD"
##  [2] "ALGERIE"
##  [3] "ALLEMAGNE"
##  [4] "AMERIQUE LATINE"
##  [5] "GRANDE BRATAGNE"
##  [6] "ANGOLA"
##  [7] "ARABIE"
##  [8] "ASIE"
##  [9] "AUSTRALIE"
## [10] "BELGIQUE"
## [11] "BURUNDI"
## [12] "CANADA"
## [13] "CHINE"
## [14] "CHYPRE"
## [15] "CONGO BRAZA"
## [16] "REP TCHEQUE"
## [17] "NA"
```

```
## [18] "EMIRATES ARABES UNIES"
## [19] "ESPAGNE"
## [20] "FRANCE"
## [21] "GABON"
## [22] "GRECE"
## [23] "HONG KONG"
## [24] "ILE MAURICE"
## [25] "INDE"
## [26] "ITALIE"
## [27] "JAPON"
## [28] "KENYA"
## [29] "KP - Corée, République Populaire démocra"
## [30] "LUXEMBOURG"
## [31] "MALAISIE"
## [32] "MAROC"
## [33] "NERLAND"
## [34] "PAYS BAS"
## [35] "NIGERIA"
## [36] "NOUVELLE ZELANDE"
## [37] "OUGANDA"
## [38] "PANAMA"
## [39] "PHILLIPINE"
## [40] "POLOGNE"
## [41] "PORTUGAL"
## [42] "ROYAUME UNI"
## [43] "RDC"
## [44] "USA"
## [45] "RWANDA"
## [46] "SINGAPOUR"
## [47] "SUISSE"
## [48] "SENEGAL"
## [49] "SOMALIE"
## [50] "SOUDAN"
## [51] "SUD SOUDAN"
## [52] "SUEDE"
## [53] "Swaziland"
## [54] "TANZANIE"
## [55] "TCHAD"
## [56] "THAILANDE"
## [57] "UNION EUROPEENNE"
## [58] "ZAMBIE"
```

5.2.3 Overall Variations

```
##
## ================================================================================================
## Statistic  N    St. Dev.          Min         Pctl(25)      Pctl(75)          Max
## ------------------------------------------------------------------------------------------------
## Taxes     228 24,567,143.00 -8,268,474.00 -8,268,474.00 -5,225,015.00 153,754,050.00
## Weight    228  407,973.40    -218,369.60   -196,559.50    -7,893.40    2,883,127.00
## ------------------------------------------------------------------------------------------------
```

L'écart-type des taxes pour les individus à travars le temps est 24,567,143.00 et celle des poids des marchandises est de 407,973.40 Kg .

5.2.4 Between Variations

```
##
## ================================================================================
## Statistic N      Mean       St. Dev.   Min Pctl(25)  Pctl(75)      Max
## --------------------------------------------------------------------------------
## Taxes     58 4,215,140.00 9,499,457.00  0     0     3,345,864.0 47,724,880
## Weight    58  132,331.80   196,253.90  100  19,665   120,394.6   940,804
## --------------------------------------------------------------------------------
```

De ce tableau, on peut lire que:

- Les taxes varient de varient de 4,215,140.00 en moyenne d'un individu à un autre
- Les poids des marchandises varient de 132,331.80 dollars d'un pays à un autre.

5.2.5 Within variations

```
##
## ================================================================================
## Statistic  N    St. Dev.        Min       Pctl(25)   Pctl(75)      Max
## --------------------------------------------------------------------------------
## Taxes     228 20,714,497.00 -47,724,880 -3,291,776.0 175,658.1 133,691,390
## Weight    228  322,405.60    -870,261    -63,771.4   29,042.9   2,624,477
## --------------------------------------------------------------------------------
```

De ce tableau, on peut lire que :

- L'écart moyen des taxes dans un pays, d'une année à une autre est de 20,714,497.00 dollars.
- L'écart moyen des poids des marchandises est de 322,405.60 kg d'une année à une autre, pour un pays.

5.2.6 Modélisation

Modèle à polaire, considérant que le modèle est le même pour tous les pays. ce la suppose l'absence des éffets aléatoire et des éffets aléatoires.

```
## Pooling Model
##
## Call:
## plm(formula = Taxes ~ Weight, data = p_DF_aggreg, model = "pooling")
##
## Unbalanced Panel: n = 58, T = 1-10, N = 228
##
## Residuals:
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
```

```
## -76127299  -3525554    -38253         0    822609 135485918
##
## Coefficients:
##                Estimate  Std. Error t-value Pr(>|t|)
## (Intercept) -8.7251e+05  1.3302e+06 -0.6559   0.5125
## Weight       4.1859e+01  2.8796e+00 14.5367   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    1.37e+17
## Residual Sum of Squares: 7.0803e+16
## R-Squared:      0.48321
## Adj. R-Squared: 0.48092
## F-statistic: 211.315 on 1 and 226 DF, p-value: < 2.22e-16
```

Le modèle s'écrit:

$$Y = 4.1859e^{01}X - 8.7251e^{05}$$

Avec:

- Y : Les taxes recoltées pour les différentes marchandises dans tout les pays
- X : Le poids globales des toutes les marchandises ayant traversé la douane pour chaque période concernant notre étude, et pour chaque individus

Notons que, pour ce type de panel, nos individus sont les pays et le temps est mésuré en année , à partir de l'année 2010 jusqu'en 2020.

La variable poids, dans les données des bases a été agrées, en calculant la somme de tous les poids pour pour les différentes marchandises n pour une année Ainsi, nous avions:

$$P_{i.t} = \sum_{j=1}^{n} Pi..$$

i indiquant les marchandises , de la prémière à la n-ième marchandise. Ainsi donc, nous avions somme les poids de chaque marchandise pour trouver le poids annuel des marchanides ayant traversé la douane d'un pays. Il en est de même pour les taxes.

Le coefficient associé au poids des marchandises est significatif car le p-value associé est inférieur à 0,05. Tansdis que le terme indépendant n'est pas significativement différent de 0 vu que le p-value associé est supérieur au seuil.

Ainsi donc, quand le poids augmente d'une unité, les taxes augmentent de 4,185 unités, pour tous les pays, et à travers le temps. Cette augmentation est significativement différent de 0.

En testant la significativité globale du modèle, nous utilisant le Langranhe Multiplier test.

```
##
##  Lagrange Multiplier Test - (Honda) for unbalanced panels
##
## data:  Taxes ~ Weight
## normal = 0.89501, p-value = 0.1854
## alternative hypothesis: significant effects
```

Le test pour le modèle `pooling` est non significatif vu que le p-value associé au test est supérieur au seuil de 5% (0,05). Ainsi donc, nous adoptons l'hypothèse nulle du test selon laquelle le modèle polaire n'est pas significatif.

Modele à éffet fixes. Nous produisons ce modèle en utilisant la methode `Within` qui traduit les variations de la variables dependante en fonction des variables dépendantes, mais pour un individus à travers le temps.

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = Taxes ~ Weight, data = p_DF_aggreg, effect = "individual",
##     model = "within")
##
## Unbalanced Panel: n = 58, T = 1-10, N = 228
##
## Residuals:
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -69551626  -1715042         0         0   1537006 114103679
##
## Coefficients:
##        Estimate Std. Error t-value  Pr(>|t|)
## Weight  39.6087     3.8914  10.178 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    9.7404e+16
## Residual Sum of Squares: 6.0386e+16
## R-Squared:      0.38005
## Adj. R-Squared: 0.16728
## F-statistic: 103.601 on 1 and 169 DF, p-value: < 2.22e-16
```

ici, nous modélisons les éffets individuels. Le coefficient des taxes signifie, que les pays augmentaient leurs importations d'une unitées, chacune, nous aurions une augmentation de 39,6087 dollars pour ce pays.

EXtraction des effets fixes Ce sont les effects individuels spécifiques pour chaque pays.

On remarque que:

- La chine a 21758301 dollars de taxe de plus que la moyenne des taxes de tous les pays, et ce, d'une manière significative.
- HONG KONG 21758301 dollars de plus que la moyenne des pays, et ce, significativement.
- Les effets individuels de tous les autres pays ne sont pas significativement différents de 0 vu que les p-values associées, respectivement pour chaque pays sont de loin supérieur au seuil chacune.

Le modèle à éffets aléatoire

```
## Oneway (individual) effect Between Model
##
## Call:
## plm(formula = Taxes ~ Weight, data = p_DF_aggreg, model = "between")
##
```

Table 5.3: Effets individuels

	Estimate	Std. Error	t-value	Pr(>|t|)
AFRIQUE DU SUD	-2984326.03	8460861	-0.3527213	0.7247369
ALGERIE	-1165271.62	18903294	-0.0616438	0.9509193
ALLEMAGNE	-1632580.61	7723181	-0.2113871	0.8328399
AMERIQUE LATINE	2202439.71	13368089	0.1647535	0.8693349
GRANDE BRATAGNE	-3470934.54	6056452	-0.5730970	0.5673408
ANGOLA	-15589167.27	18967690	-0.8218801	0.4123031
ARABIE	-142028.25	18902837	-0.0075136	0.9940139
ASIE	38936.72	18902846	0.0020598	0.9983589
AUSTRALIE	285990.29	18902686	0.0151296	0.9879466
BELGIQUE	3926237.34	7009603	0.5601226	0.5761375
BURUNDI	-315018.56	6301266	-0.0499929	0.9601871
CANADA	6630399.23	10924084	0.6069524	0.5446970
CHINE	21758300.68	6513971	3.3402513	0.0010301
CHYPRE	2015623.85	18902888	0.1066305	0.9152086
CONGO BRAZA	355028.22	18902684	0.0187819	0.9850373
REP TCHEQUE	1577815.74	18903000	0.0834691	0.9335774
NA	-1030865.82	13367444	-0.0771176	0.9386212
EMIRATES ARABES UNIES	1635013.00	7146910	0.2287720	0.8193227
ESPAGNE	358452.36	9451857	0.0379240	0.9697930
FRANCE	-3760302.34	6321394	-0.5948533	0.5527371
GABON	270146.82	18902687	0.0142915	0.9886143
GRECE	-167122.77	13366412	-0.0125032	0.9900389
HONG KONG	14181338.15	6715252	2.1118103	0.0361728
ILE MAURICE	-18631113.63	18994748	-0.9808561	0.3280658
INDE	2932397.32	9457208	0.3100701	0.7568897
ITALIE	-3579816.54	13371879	-0.2677123	0.7892472
JAPON	-11074.78	10913538	-0.0010148	0.9991915
KENYA	-367019.02	6683619	-0.0549132	0.9562725
KP - Corée, République Populaire démocra	218970.88	18902915	0.0115840	0.9907712
LUXEMBOURG	3908067.60	9453578	0.4133956	0.6798409
MALAISIE	9323874.62	7990237	1.1669084	0.2448913
MAROC	-3041138.54	13370444	-0.2274523	0.8203470
NERLAND	-1458943.80	18905086	-0.0771720	0.9385780
PAYS BAS	-2135869.03	6707725	-0.3184193	0.7505601
NIGERIA	69469.46	13366251	0.0051974	0.9958592
NOUVELLE ZELANDE	-964653.67	9452267	-0.1020553	0.9188338
OUGANDA	-9660163.23	6259154	-1.5433657	0.1246125
PANAMA	1902251.68	13367426	0.1423050	0.8870086
PHILLIPINE	377090.25	18902684	0.0199490	0.9841076
POLOGNE	-189313.83	13366333	-0.0141635	0.9887163
PORTUGAL	-1520165.36	9453188	-0.1608098	0.8724353
ROYAUME UNI	-1520165.36	13367521	-0.1137208	0.9095941
RDC	-10887866.60	8575838	-1.2695980	0.2059733
USA	-3702024.25	7157453	-0.5172265	0.6056744
RWANDA	-2745202.66	6429503	-0.4269697	0.6699449
SINGAPOUR	-1271537.64	7722149	-0.1646611	0.8694075
SUISSE	-6428958.73	6084800	-1.0565604	0.2922202
SENEGAL	-164360.36	18902760	-0.0086950	0.9930727
SOMALIE	182031.20	18902817	0.0096298	0.9923280
SOUDAN	370990.51	18902684	0.0196263	0.9843646
SUD SOUDAN	-3003470.69	18905608	-0.1588667	0.8739636
SUEDE	-1129587.86	13367521	-0.0845024	0.9327570
Swaziland	-975538.34	7726665	-0.1262561	0.8996793
TANZANIE	-373840.11	6683928	-0.0559312	0.9554627

```
## Unbalanced Panel: n = 58, T = 1-10, N = 228
## Observations used in estimation: 58
##
## Residuals:
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -17710356  -1366186    300948         0    907729  22825793
##
## Coefficients:
##                Estimate  Std. Error t-value  Pr(>|t|)
## (Intercept) -9.2150e+05  9.0924e+05 -1.0135    0.3152
## Weight       3.8816e+01  3.8643e+00 10.0449 3.886e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    5.1437e+15
## Residual Sum of Squares: 1.8358e+15
## R-Squared:      0.64309
## Adj. R-Squared: 0.63671
## F-statistic: 100.9 on 1 and 56 DF, p-value: 3.886e-14
```

Ainsi donc, le modèle within s'écrit:

$$Y = +3.8816e^{01}X - 9.2150e^{05}$$

Pour fragmentation d'une unité pour le poids des marchandises, on observe une augmentation des taxes de 38,81 unitées d'un pays à un autre. cela signifie que, si un pays à importé un 1Kg plus qu'un autre, nous observerons une différence de 38,82 dollars pour les taxes.

```
##
##  F test for individual effects
##
## data:  Taxes ~ Weight
## F = 0.51147, df1 = 57, df2 = 169, p-value = 0.998
## alternative hypothesis: significant effects
```

Le choix du modèle en utlisant le test d'Hausman

Ce test consiste à comparer les coefficients de ce s deux modèles, pour en fint chhoisir le modèle qui est consistant entre les modèle à éffet individuel et le modèle à éffet aléatoire.

```
##
##  Hausman Test
##
## data:  Taxes ~ Weight
## chisq = 0.020872, df = 1, p-value = 0.8851
## alternative hypothesis: one model is inconsistent
```

Ce test étant non qignificatif, c'est-à-dire , nous adoptons l'hypothèse nulle : les coefficients du modèle à éffet fixe et ceux du modèle à éffet aléatoire ne sont pas significativement différents. Ainsi donc, nous utiliserons les modèles à éffets aléatoire car celui-ci est efficien (éficace).

Effet aleatoire

```
## Oneway (individual) effect Random Effect Model
##    (Wallace-Hussain's transformation)
##
## Call:
## plm(formula = Taxes ~ Weight, data = p_DF_aggreg, model = "random",
##     random.method = "walhus")
##
## Unbalanced Panel: n = 58, T = 1-10, N = 228
##
## Effects:
##                      var   std.dev share
## idiosyncratic 3.578e+14 1.891e+07     1
## individual    0.000e+00 0.000e+00     0
## theta:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0       0       0       0       0       0
##
## Residuals:
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -76127299  -3525554    -38253         0    822609 135485918
##
## Coefficients:
##                Estimate  Std. Error z-value Pr(>|z|)
## (Intercept) -8.7251e+05  1.3302e+06 -0.6559   0.5119
## Weight       4.1859e+01  2.8796e+00 14.5367   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    1.37e+17
## Residual Sum of Squares: 7.0803e+16
## R-Squared:      0.48321
## Adj. R-Squared: 0.48092
## Chisq: 211.315 on 1 DF, p-value: < 2.22e-16
```

5.3 Modèle avec les marchandises comme individus

Modèle à polaire, considérant les deux effets

```
## Pooling Model
##
## Call:
## plm(formula = Taxes ~ Weight, data = p_DF_aggreg, model = "pooling")
##
## Unbalanced Panel: n = 58, T = 1-10, N = 228
##
## Residuals:
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -76127299  -3525554    -38253         0    822609 135485918
##
## Coefficients:
```

```
##                Estimate  Std. Error t-value Pr(>|t|)
## (Intercept) -8.7251e+05  1.3302e+06 -0.6559   0.5125
## Weight       4.1859e+01  2.8796e+00 14.5367   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    1.37e+17
## Residual Sum of Squares: 7.0803e+16
## R-Squared:      0.48321
## Adj. R-Squared: 0.48092
## F-statistic: 211.315 on 1 and 226 DF, p-value: < 2.22e-16
```

Modele à éffet aléatoire

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = Taxes ~ Weight, data = p_DF_aggreg, effect = "individual",
##     model = "within")
##
## Unbalanced Panel: n = 58, T = 1-10, N = 228
##
## Residuals:
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -69551626  -1715042         0         0   1537006 114103679
##
## Coefficients:
##        Estimate Std. Error t-value  Pr(>|t|)
## Weight  39.6087     3.8914  10.178 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    9.7404e+16
## Residual Sum of Squares: 6.0386e+16
## R-Squared:      0.38005
## Adj. R-Squared: 0.16728
## F-statistic: 103.601 on 1 and 169 DF, p-value: < 2.22e-16
```

l'expression du modèle est : $ Y=39,6087 X$

Ainsi donc, ce coefficient étant significatif, traduit que l'augmentation du poids des marchnise de une unité entrainte une augmentation des taxes perçues de 39,6087 dollars, pour la même marchandise

Le modèle à éffets aléatoire

```
## Oneway (individual) effect Between Model
##
## Call:
## plm(formula = Taxes ~ Weight, data = p_DF_aggreg, model = "between")
##
## Unbalanced Panel: n = 58, T = 1-10, N = 228
## Observations used in estimation: 58
```

```
##
## Residuals:
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -17710356  -1366186    300948         0    907729  22825793
##
## Coefficients:
##                Estimate  Std. Error t-value  Pr(>|t|)
## (Intercept) -9.2150e+05  9.0924e+05 -1.0135    0.3152
## Weight       3.8816e+01  3.8643e+00 10.0449 3.886e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    5.1437e+15
## Residual Sum of Squares: 1.8358e+15
## R-Squared:      0.64309
## Adj. R-Squared: 0.63671
## F-statistic: 100.9 on 1 and 56 DF, p-value: 3.886e-14
```

Le modèle s'écrit de la manière suivante:

$$Y = 3.8816e^{01}X - 9.2150e^{05}$$

Le coefficient associé au poids étant significatif, on trouve que: si une marchandise dépassait en poids une autre de 1Kg, les taxes perçues pour cette marchandises dépasseraient cette autre de 38,8 dollars.

Condidérant le R-Carré ajutsté, Ce modèle explique à 63,7% les variations des taxes en fonction des poids des marchandises.

Pour arbitrer le choix entre le modèle à effet fixe et le modèle à éffet aléatoire, nous allons utiliser le test d'Hausman qui consite à comprer les coefficients de ces deux modèles

```
## # A tibble: 1 x 5
##   statistic p.value parameter method       alternative
##       <dbl>   <dbl>     <int> <chr>        <chr>
## 1      9.03 0.00265         1 Hausman Test one model is inconsistent
```

Le test d'Hausman étant significatif pour ces données, nous optons donc le modèle à éffet à éffet fixe pour estimer les taxes sachant que les individus sont les marchandises car celui-ci est consistant.

On suppose donc une variation significative entre les marchandises.

Chapter 6

Conclusion

Dans ce document nous venons de modeliser les taxes perçues dans les différents pays sur les marchandises à leurs destinations. Le modèle a pour variable exogène les **taxes** et comme variable endogène **les poids des marchandises** .

Nous avions manipulé ces données de manière à en extraire deux bases pour les données de panel:

- une base des données dans laquelle les individus sont les pays. *dans cette base, nous avions aggrégé les poids des tous les marchandises en une seul variable* `weight` *en faisant leur somme.il en est de même pour les variables taxes*
- une base des données dans laquelle les individus sont les marchandises. *dans cette base, nous avions utlisé un format large de notre dataframe. nous avions agrégé la variable poids et la variable taille en faisant la sommes, pour chaque ligne, pour tous les pays.

La modélisation a donc consister à estimer :

- Le modèle **pooling**
- Le modèle **within**
- le modèle **between**

Pour chaque estimation, nous avions fait les test de spécification du modèle pour identifier , le quel de ces trois modèles sus indiqués represente au mieux nos données. Nous avions possédé au test de **Langrade Multiplier**, et au test d'**Hausman**.