

Panel Analyses Report

David BYAMUNGU

2021-06-17

Contents

1	Prerequis	5
2	Introduction	7
3	Literature	9
4	The One-way Error Component Regression Model	11
4.1	INTRODUCTION	11
4.2	THE FIXED EFFECTS MODEL	13
4.3	THE RANDOM EFFECTS MODEL	16
5	Methods	17
6	Analyses	19
6.1	Netoyage de la base des données	19
6.2	Analyse descriptive des Varariales	25
7	Final Words	27

Chapter 1

Prerequis

Ceci est *une étude* des données de panel avec **Markdown**.

The **bookdown** package can be installed from CRAN or Github:

La structure de ce rapport est que chaque fichier RMD porte un chapitre et un thème bien spécifique de notre analyse

Nous avons utilisé XeLaTeX pour compiler ce document en PDF.

Chapter 2

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 le *modele à effet fixe* et le *modèle à effet aléatoire* et en fin produire un **modèle dynamique**

Nous expliquons Notre méthodologie dans la partie suivante.

Chapter 3

Literature

Chapter 4

The One-way Error Component Regression Model

4.1 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.

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

(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 X_{it} is the i 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 μ_i denotes the unobservable individual-specific effect and v_{it} 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

$$X_{it}$$

may contain a set of variables like experience, education, union membership, sex, race, etc. Note that α_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

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

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

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

$$u = Z_\mu \mu + v \quad (2.4)$$

$$y_{it} = \alpha + X'_{it} + U_{it}$$

, $i=1, \dots, N$; $t=1, \dots, T$ 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. α is a scalar, β is $K \times 1$ and X_{it} is the i th observation on K explanatory variables. disturbances, with it

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

where u_i denotes the unobservable individual-specific effect and v_{it} 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 X_{it} may contain a set of variables like experience, education, union membership, sex, race, etc. Note that α_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 α_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 $NT \times K$, $Z = [i_{NT}, X]$, $\delta = (\alpha', \beta')$ and i_{NT} is a vector of ones of dimension NT . Also, (2.2) can be written as

$$u = Z_\mu \mu + v$$

(2.4)

where $u = (u_{11}, \dots, u_{1T}, u_{21}, \dots, u_{2T}, \dots, u_{N1}, \dots, u_{NT})$ with the observations stacked such that the slower index is over individuals and the faster index is over time. $Z = IN \otimes T$ where IN is an identity matrix of dimension N , T is a vector of ones of dimension T and \otimes denotes Kronecker product. Z_μ 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 α_i if they are assumed to be fixed parameters. $\delta = (\alpha', \beta')$ and $\nu' = (\nu_{11}, \dots, \nu_{1T}, \dots, \nu_{N1}, \dots, \nu_{NT})$. Note that $Z_\mu Z_\mu' = I_N \otimes J_T$ where J_T is a matrix of ones of dimension T and $P = Z(Z'Z)^{-1}Z'$, the projection matrix on Z_μ , reduces to $IN \otimes J_T$ 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 $\text{rank}(P) = \text{tr}(P) = N$ and $\text{rank}(Q) = \text{tr}(Q) = N(T-1)$. 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 $P + Q = INT$. In fact, any two of these properties imply the third (see Graybill, 1961, theorem 1.68).

4.2 THE FIXED EFFECTS MODEL

In this case, the α_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

$$y = i_{NT} \alpha + X\beta + Z_\mu \mu + v = Z\delta + Z_\mu \mu + v \quad (2.5)$$

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

Note that Z is $NT \times (K+1)$ and Z , the matrix of individual dummies, is $NT \times N$. 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 α and β 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:

$$QY = QX + Qv \quad (2.6)$$

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 $\tilde{X} = QX$ with typical element

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

(2.7) with $\text{var}(\tilde{\beta}) = \sigma_v^2 (X'QX)^{-1} = \sigma_v^2 (\tilde{X}'\tilde{X})^{-1}$. $\tilde{\beta}$ 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

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

(2.8)

and averaging over time gives

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

(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) where we utilized the restriction that $\sum_{i=1}^n \mu_i = 0$. 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 $\alpha + \sum_{i=1}^n \mu_i$ are estimable from (2.8), and not α and μ_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 $\sigma^2 \mathbf{I}_{NT}$. Note that as $T \rightarrow \infty$ the FE estimator is consistent. However, if T is fixed and $N \rightarrow \infty$ as is typical in short labor panels, then only the FE estimator of δ 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. $H_0: \mu_1 = \mu_2 = \dots = \mu_{N-1} = 0$, 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

$$F_0 = \frac{\frac{RRSS - URSS}{N - 1}}{\frac{URSS}{NT - N - K}} \sim F_{N-1, N(T-1)-K} \quad (2.12)$$

- (2) *Computational warning.* One computational caution for those using the Within regression given by (2.10). The s^2 of 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:

$$y_i = Z_i \delta + \mu_i i_T + v_i \quad (2.13)$$

where y_i is $T \times 1$, $Z_i = [1_T, X_i]$, 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 $T \times 1$. In general, $E(v_i, v_i') = \Omega_i$ for $i = 1, 2, \dots, N$, 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 $N \rightarrow \infty$ and T fixed. Performing the Within transformation on this set of equations (2.13) one gets

$$\tilde{y}_i = \tilde{X}_i\beta + \tilde{v}_i \quad (2.14)$$

where

$$\tilde{y} = Qy$$

,

$$\tilde{X} = QX$$

and

$$\tilde{v} = Qv$$

, with

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

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:

$$N^{\frac{1}{2}}(\tilde{\beta} - \beta) \sim N(0, M^{-1}VM^{-1}) \quad (2.15)$$

where

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

Note that

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

and

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

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

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

where

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

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

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

4.3 THE RANDOM EFFECTS MODEL

Chapter 5

Methods

We describe our methods in this chapter.

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

$$i : 1 \rightarrow N$$

et

$$t : 1 \rightarrow T$$

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).

Chapter 6

Analyses

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

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

6.1 Netoyage de la base des données

Apperçue globale des données

Voici la structure de la base des données

```
## Rows: 3,310
## Columns: 6
## $ `N°` <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 1~
## $ `Country/destination` <chr> "AFRIQUE DU SUD", "AFRIQUE DU SUD", "AFRIQUE DU ~
## $ Year <dbl> 2011, 2011, 2011, 2013, 2013, 2013, 2013, 2013, ~
## $ Goods <fct> "GRUES SUR PNEUMATIQUE", "CAMION FAMIL", "CAMION~
## $ Weight <dbl> 13500, 12000, 24000, 183, 19520, 19520, 19520, 1~
## $ Taxe <dbl> 0, 0, 0, 0, 264771, 272817, 283220, 264142, 0, 0~
```

Table 6.1: 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

Voici les modalités de la variable **Goods** qui signifie **Marchandises**

La variable **Goods** a 740 modalités

Faisons la caractérisation des niveaux des marchandises dont l'encodage fait défaut

```
class(taxe_df$Goods)

## [1] "character"

Usage de tm et Stringr

## Warning: Unknown levels in `f`: equipements protection

## [1] "Autres Marchandises"
## [2] "Bois"
## [3] "Machines et appareils domestique"
## [4] "Médicaments et plantes médicinales"
## [5] "Poissons, viande et oeufs"
## [6] "Matériels de construction"
## [7] "Matériel Informatique et Electroniques"
## [8] "Véhicules, camions, Motos et acc"
## [9] "Vêtements, tissus et acc et chaussure"
## [10] "boissons, bières et limonades"
## [11] "Machine us Industriel"
## [12] "Article Ménage et Campement"
## [13] "sacs, sachets et emballages"
## [14] "Papiers et fournitures de bureaux"
## [15] "Produits alimentaires, prep et huiles"
## [16] "café arabica"
## [17] "Minerais et dérivés"
## [18] "engins et tracteurs"
## [19] "Cigarette et papier cigarettes et tabac"
## [20] "construction préfabriquées"
## [21] "cadres et conteneurs"
## [22] "Pièces de Réchange appareils"
## [23] "Générateurs, batterie et piles"
## [24] "etuis en plastique ou textile"
## [25] "Pétrole et dérivés et huile de graissage"
## [26] "boissons, bières, liqueurs et limonades"
## [27] "produits beauté"
## [28] "peaux des bêtes"

##           n      % val%
## Autres Marchandises      160  4.8  4.8
## Bois                     140  4.2  4.2
## Machines et appareils domestique      30  0.9  0.9
## Médicaments et plantes médicinales      34  1.0  1.0
## Poissons, viande et oeufs              12  0.4  0.4
## Matériels de construction              27  0.8  0.8
## Matériel Informatique et Electroniques      32  1.0  1.0
## Véhicules, camions, Motos et acc      113  3.4  3.4
## Vêtements, tissus et acc et chaussure      100  3.0  3.0
## boissons, bières et limonades           15  0.5  0.5
## Machine us Industriel                 54  1.6  1.6
```

Table 6.2: Modalités de la variable Goods à l'importation des donnees

x
0
3Café vert arabica, en feve K3
Abats comestibles,congeles,de chevaux,anes,mulets,ovins ou caprins
ABATS COMESTIBLES;CONGELES;DE CHEVAUX;ANES;MULETS;OVINS
Accessoires de radio diffusion
Accessoires de vehicules
Accumulateurs electriques
Acide acetique
ages de 5 ans ou moins
ages de plus de 5 ans
Agés de plus de 5 ans ou moins
Alcaloides du quinquina et leurs derives;
ALCOOL ETYLIQUE NON DENATURE
ambulance d'une cylindree excedant 2500 cm3
Antennes
Antennes et reflecteurs d'antennes
antennes et reflateurs
Appareils d'eclairage electriques
Appareils d'eclairage non electriques
Appareils d'eclairages electriques
Appareils du n°84.14
Appareils electrothermiques pour la
appareils pour la reception,la conversion et la transmission
Art et materiel d'athletisme
Articles confectionnes en textiles
Articles d'economie domestique,en
Articles de bureau
ARTICLES DE BUREAU
Articles de bureau ou de la papeterie
Articles de friperie
ARTICLES DE FRIPERIE
Articles et materiel d'athletisme
Ashok Layland
ASPIRATEUR ET ACCESSOIRES
Autes bois sciés
AUTRE MACHINE ET APPAREIL A IMPRIMER
AUTRE MINERAIS DE TITANE (Coltant)
AUTRE PARTIE DE PLANTE
AUTRE PEAUX
AUTRE PREP ALIMENTAIRE
Autre vehicules automobiles a usages speciaux
Autres
AUTRES
Autres bois scies
Autres abats comestibles frais ou refrigerés de chevaux,anes,mulets,ovins,caprins
Autres abats comestibles,congeles,de chevaux,anes,mulets,ovins ou caprins
Autres accessoires de tuyauterie en fonte
Autres accumulateurs electriques
Autres appareils elevateurs, a action continue pour marchandises
Autres armes
Autres art de bureau ou de papeterie en papier
Autres Art de menage
Autres articles d'economie domestique, en
autres articles de bureau

## Article Menange et Campement	37	1.1	1.1
## sacs, sachetsn emballages	6	0.2	0.2
## Papiers et fournitures de bureaux	24	0.7	0.7
## Produits alimentaires,prep et huiles	68	2.1	2.1
## caféarabica	1303	39.4	39.5
## Minéraux et dérivés	1053	31.8	31.9
## engins et tracteurs	18	0.5	0.5
## Cigarette et papier cigarettes et tabac	14	0.4	0.4
## constructionprefabriquees	7	0.2	0.2
## cadreset conteneurs	1	0.0	0.0
## Pièces de Réchange appareils	6	0.2	0.2
## Générateurs,batterie et piles	15	0.5	0.5
## etuis en plastique ou textile	1	0.0	0.0
## Pétrole et dérivées et huile de graissage	4	0.1	0.1
## boissons, bières,liqueurs et limonades	2	0.1	0.1
## produits beaute	10	0.3	0.3
## peauxdes betes	14	0.4	0.4
## NA	10	0.3	NA

netoyage de la variable country_desti qui est un facteur dans le quel nous retrouvons les niveaux rédon-
dants (sur l'identifiant des pays)

```
## [1] "AFRIQUE DU SUD"
## [2] "ALGERIE"
## [3] "ALLEMAGNE"
## [4] "AMERIQUE LATINE"
## [5] "ANGLETERRE"
## [6] "ANGOLA"
## [7] "ARABIE"
## [8] "ASIE"
## [9] "AUSTRALIE"
## [10] "BELGIQUE"
## [11] "BURUNDI"
## [12] "CANADA"
## [13] "CHINE"
## [14] "CHYPRE"
## [15] "CONGO BRAZA"
## [16] "CZECH REP"
## [17] "DOMBASI SIMBA"
## [18] "EMIRATES ARABES UNIES"
## [19] "ESPAGNE"
## [20] "FRANCE"
## [21] "GABON"
## [22] "GRANDE BRATAGNE"
## [23] "GRECE"
## [24] "HONG KONG"
## [25] "ILE MAURICE"
## [26] "INDE"
## [27] "ITALIE"
## [28] "J WOLFF"
## [29] "JAPON"
## [30] "KENYA"
## [31] "KP - Corée, République Populaire démocra"
```

```
## [32] "LIBAN"
## [33] "LUXEMBOURG"
## [34] "MADRID"
## [35] "MALAISIE"
## [36] "MAROC"
## [37] "NERLAND"
## [38] "NERETHERLAND"
## [39] "NIGERIA"
## [40] "NOUVELLE ZELANDE"
## [41] "OUGANDA"
## [42] "PANAMA"
## [43] "PAYS BAS"
## [44] "PHILLIPINE"
## [45] "POLOGNE"
## [46] "PORTUGAL"
## [47] "R-U"
## [48] "RDC"
## [49] "RDC/BELGIQUE"
## [50] "RDC/BUNIA"
## [51] "RDC/CHINE"
## [52] "RDC/ETATS UNIS"
## [53] "RDC/FRANCE"
## [54] "RDC/MALAISIE"
## [55] "RDC/OUGANDA"
## [56] "RDC/R-U"
## [57] "RDC/RWANDA"
## [58] "RDC/SINGAPOUR"
## [59] "RDC/SUISSE"
## [60] "REP TCHEQUE"
## [61] "ROYAUME UNI"
## [62] "RWANDA"
## [63] "SENEGAL"
## [64] "SINGAPOUR"
## [65] "SKN"
## [66] "SOMALIE"
## [67] "SOUDAN"
## [68] "SUCAFINA"
## [69] "SUD SOUDAN"
## [70] "SUEDE"
## [71] "SUISSE"
## [72] "SUITZERLAND"
## [73] "Swaziland"
## [74] "SWEDEN"
## [75] "SWITZERLAND"
## [76] "TANZANIE"
## [77] "TCHAD"
## [78] "THAILANDE"
## [79] "TWIN TRADING"
## [80] "TZ"
## [81] "UAE"
## [82] "UNION EUROPEENNE"
## [83] "USA"
```

Table 6.3: 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

```
## [84] "WALTER MATTER"
## [85] "ZAMBIE"

## [1] "AFRIQUE DU SUD"      "ALGERIE"      "ALLEMAGNE"
## [4] "AMERIQUE LATINE"    "GRANDE BRATAGNE" "ANGOLA"
## [7] "ARABIE"             "ASIE"         "AUSTRALIE"
## [10] "BELGIQUE"           "BURUNDI"      "CANADA"
## [13] "CHINE"              "CHYPRE"       "CONGO BRAZA"
## [16] "REP TCHEQUE"        "NA"           "EMIRATES ARABES UNIES"
## [19] "ESPAGNE"            "FRANCE"       "GABON"
## [22] "GRECE"              "HONG KONG"    "ILE MAURICE"
## [25] "INDE"               "ITALIE"       "JAPON"
## [28] "KENYA"              "KP - Corée"   "LIBAN"
## [31] "LUXEMBOURG"         "MALAISIE"     "MAROC"
## [34] "NERLAND"            "PAYS BAS"     "NIGERIA"
## [37] "NOUVELLE ZELANDE"   "OUGANDA"      "PANAMA"
## [40] "PHILLIPINE"         "POLOGNE"      "PORTUGAL"
## [43] "ROYAUME UNI"        "RDC"          "USA"
## [46] "RWANDA"             "SINGAPOUR"    "SUISSE"
## [49] "SENEGAL"            "SOMALIE"      "SOUDAN"
## [52] "SUD SOUDAN"         "SUEDE"        "Swaziland"
## [55] "TANZANIE"           "TCHAD"        "THAILANDE"
## [58] "UNION EUROPEENNE"  "ZAMBIE"
```

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 élargués de la base des données, car nous avons jugé qu' aucune méthode d'imputation n'est applicable pour ce genre de situation. Nous avons fait la même chose pour les variables tels que **Les marchandises**.

6.1.1 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.

```
DBase <- taxe_df %>%
  select(Year, Country_dest, Goods, Weight, Taxe) %>%
  group_by(Year, Country_dest, Goods) %>%
  summarise(Weight=sum(Weight), Taxe=sum(Taxe), .groups = "drop") %>% drop_na()

correlate(DBase) %>% kable(caption = "Table de corrélation entre les variables quantitatives")
```



```
#plot_correlate(DBase)
```

```
df <- pdata.frame(DBase, index = c("Year", "Country_dest"))
```

```
## Warning in pdata.frame(DBase, index = c("Year", "Country_dest")): duplicate couples (id-time) in r  
## to find out which, use, e.g., table(index(your_pdataframe), useNA = "ifany")
```

```
DF <- df %>% pivot_wider(names_from = Goods, values_from = c(Taxe, Weight))
```

```
DB <- pdata.frame(DBase, index=c("Year", "Goods"))
```

```
## Warning in pdata.frame(DBase, index = c("Year", "Goods")): duplicate couples (id-time) in resultin  
## to find out which, use, e.g., table(index(your_pdataframe), useNA = "ifany")
```

6.2 Analyse descriptive des Varariales

Conversion des données en modèle des panels des données

Chapter 7

Final Words

We have finished a nice book.