

Utilisation du paquet plm: application en relations industrielles

Milène R. E. Lokrou
Étudiante au doctorat en relations industrielles

Université Laval

14 mai 2019



Faculté des sciences sociales
Département des
relations industrielles



Structure de l'atelier

1 Introduction

- Rappel : données de panel
- Estimation sur des données de panel avec R
- Objectifs de l'atelier

2 Les modèles linéaires

3 Présentation du paquet plm

- La structure des données
- Les modèles d'estimation et l'interface
- L'approche logicielle pour estimer

4 Cas pratique : syndicalisation et chômage aux États-Unis

- Description de la base de données utilisée
- Application du paquet plm

5 Conclusion

6 Bibliographie sommaire

Rappel : données de panel

Les données dites de panel renferment deux dimensions :

- ① Des données transversales
- ② Des séries temporelles

Rappel : données de panel

Avantages liés à l'utilisation de données de panel :

- ① Meilleur contrôle de l'hétérogénéité
- ② Données plus riches d'informations
- ③ Possibilité de capter certaines dynamiques
- ④ Capacité de produire des analyses plus complexes

Rappel : données de panel

Dans un panel typique, le nombre d'individus n est grand et celui des périodes T est petit.

Cependant, un panel peut présenter une ou plusieurs caractéristiques (Park, 2011) :

- court
- long
- cylindré
- non cylindré
- fixe
- rotatif

Introduction

Rappel : données de panel

	airline	year	load	cost	output	fuel
1.	1	1	.534487	13.9471	-.0483954	11.57731
2.	1	2	.532328	14.01082	-.0133315	11.61102
3.	1	3	.547736	14.08521	.0879925	11.61344
4.	1	4	.540846	14.22863	.1619318	11.71156
5.	1	5	.591167	14.33236	.1485665	12.18896
6.	1	6	.575417	14.4164	.1602123	12.48978
7.	1	7	.594495	14.52004	.2550375	12.48162
8.	1	8	.597409	14.65482	.3297856	12.6648
9.	1	9	.638522	14.78597	.4779284	12.85868
10.	1	10	.676287	14.99343	.6018211	13.25208
11.	1	11	.605735	15.14728	.4356969	13.67813
12.	1	12	.61436	15.16818	.4238942	13.81275
13.	1	13	.633366	15.20081	.5069381	13.75151
14.	1	14	.650117	15.27014	.6001049	13.66419
15.	1	15	.625603	15.3733	.6608616	13.62121
16.	2	1	.490851	13.25215	-.652706	11.55017
17.	2	2	.473449	13.37018	-.626186	11.62157
18.	2	3	.503013	13.56404	-.4228269	11.68405
19.	2	4	.512501	13.8148	-.2337306	11.65092
20.	2	5	.566782	14.00113	-.1708536	12.27989

Figure – exemple de données de panel (Source : Park, 2011, p. 4)

Estimation sur des données de panel avec R

Dans R, plusieurs paquets permettent de produire une estimation sur des données de panel. On peut citer, entre autres, les paquets :

- phtt (pour l'effet de temps) ;
- splm (pour des données spatiales) ;
- lme4 (pour les modèles linéaires mixtes) ;
- nlme (pour les modèles linéaires et non linéaires mixtes) ;
- plm (pour les modèles linéaires).

Objectifs de l'atelier

Dans le cadre de cet atelier, nous allons utiliser le paquet plm. Quatre objectifs sont par ailleurs visés :

- ➊ Présenter le paquet plm et l'intérêt de celui-ci
- ➋ Offrir une analyse de données de panel sur la base d'un cas pratique
- ➌ Spécifier les modèles adéquats (tests de spécification)
- ➍ Estimer des modèles à effets fixes et à effets aléatoires

Les modèles linéaires

Tel que mentionné par Brigitte Dormont (1989, p.20), «le choix le plus fréquent en économétrie des données de panel consiste à adopter une spécification en terme de modèle à erreurs composées».

Regression linéaire simple

Le cas simple d'un modèle d'analyse de données de panel à une seule variable explicative (x) peut donc être spécifié comme suit :

$$y_{it} = x_{it}b + u_{it}$$

avec $i = 1, \dots, N$ et $t = 1, \dots, T$ et le terme d'erreur $u_{it} = \alpha_i + \varepsilon_{it}$

Les modèles linéaires

Regression linéaire simple

Le cas simple d'un modèle d'analyse de données de panel à une seule variable explicative (x) peut donc être spécifié comme suit :

$$y_{it} = x_{it}b + u_{it}$$

Où :

i est une unité statistique communément appelée individu

t est la période

x_{it} est la variable indépendante et b la pente

α_i et ε_{it} sont des perturbations aléatoires non corrélées, d'espérances nulles.

La structure des données

Dans R, l'identification d'une base de données va reposer sur un **data frame**.

«Un **data frame** est une liste de classe **data.frame** dont tous les éléments sont de la même longueur (ou comptent le même nombre de lignes si les éléments sont des matrices).» (Goulet, 2016, p. 30).

La structure des données

Puisque les données de panel renferment deux dimensions (i.e. individuelle et temporelle), un argument **index** doit être ajouté afin d'indiquer la structure des données. Cet argument peut prendre quatre formes (Croissant et Millo, 2008) :

- ➊ NULL
- ➋ Une chaîne de caractères
- ➌ Un vecteur de deux chaînes de caractères
- ➍ Un entier

Les modèles d'estimation et l'interface

4 modèles d'estimation sont fournis par le paquet plm (Croissant et Millo, 2008, p.5) :

- ❶ plm : l'estimation classique des données de panel. Dans ce modèle, la fonction lm est utilisée pour transformer les données.
- ❷ pvcm : l'estimation des modèles avec des coefficients variables.
- ❸ pgmm : l'estimation avec la méthode des moments généralisée.
- ❹ pggls : l'estimation avec des moindres carrés généralisés faisables.

Les modèles d'estimation et l'interface

3 arguments sont communs aux modèles d'estimation précédents :

- ❶ l'index : i et t pour chaque observation
- ❷ l'effet : les effets individuels, les effets temporels ou les deux
- ❸ le modèle : à effets fixes ou à effets aléatoires

Présentation du paquet plm

L'approche logicielle pour estimer

Comment obtenir un paquet plm qui fonctionne ?

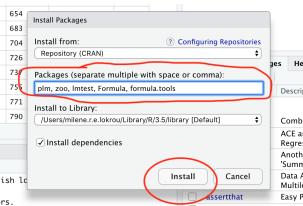
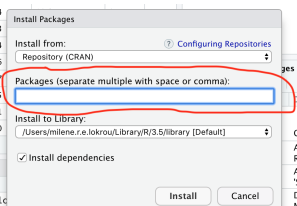
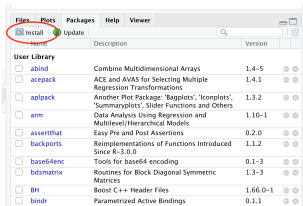
- ➊ Étape 1 : téléchargez *RStudio Desktop*
(<https://www.rstudio.com/products/rstudio/download/>).
- ➋ Étape 2 : Installez les paquets plm, lmtest, zoo, formula, formula.tools.

Présentation du paquet plm

L'approche logicielle pour estimer

Si le paquet formula est préalablement installé, vous pourrez avoir accès au paquet plm en utilisant la commande `library("plm")` dans Rstudio.

Si vous n'avez aucun paquet installé, voici quelques captures d'écran qui vous seront utiles pour les trouver, puis les installer.



L'approche logicielle pour estimer

En fonction des arguments utilisés, le paquet plm permet d'estimer :

- Des effets fixes (within)
- Des données fusionnées (pooling)
- La première différence (fd)
- Les variations inter-individuelles (between)
- Des modèles à erreurs composées (random)

Présentation du paquet plm

L'approche logicielle pour estimer

L'utilisation générale du paquet plm consiste à indiquer la structure du modèle, les données et l'approche d'estimation choisie.

Si l'on utilise des données déjà contenues dans `r` (e.g. `data("Grunfeld", package = "Ecdat")`), un usage basique du paquet plm équivalra à :

- `R> grun.fe <- plm(inv ~ value + capital, data = Grunfeld, model = "within")`
- `R> grun.re <- plm(inv ~ value + capital, data = Grunfeld, model = "random")`

Présentation du paquet plm

L'approche logicielle pour estimer

Il existe deux modèles à erreurs composées :

Le modèle à erreurs composées de type I

$$u_{it} = \alpha_i + \varepsilon_{it}$$

Le modèle à erreurs composées de type II

$$u_{it} = \alpha_i + \lambda_t + \varepsilon_{it}$$

Où :

λ_t est l'effet inobservé du temps

Présentation du paquet plm

L'approche logicielle pour estimer

Ces modèles peuvent être spécifiés dans le paquet plm. Par exemple, pour spécifier le modèle de type II :

```
R> grun.twfe <- plm(inv ~ value + capital, data = Grunfeld, model =  
"within", + effect = "twoways")
```

```
R> fixef(grun.twfe, effect = "time")
```

L'approche logicielle pour estimer

Enfin, pour éviter des risques de mauvaise spécification des modèles, plusieurs auteurs (e.g. Greene, 2011 ; Park, 2011) suggèrent l'utilisation de tests tels que celui de Hausman ou encore ceux du multiplicateur de Lagrange.

Présentation du paquet plm

L'approche logicielle pour estimer

Pour le test de Hausman, **phtest** permet de l'exécuter. Les arguments principaux sont deux objets **panelmodel** ou une **formule**. Une application classique du test de Hausman consiste à comparer des modèles à effets fixes et aléatoires.

On aurait donc, en utilisant les données Grunfeld contenues dans le paquet plm :

```
R> gw <- plm(inv ~ value + capital, data = Grunfeld, model = "within")
```

```
R> gr <- plm(inv ~ value + capital, data = Grunfeld, model = "random")
```

```
R> phtest(gw, gr)
```

Présentation du paquet plm

L'approche logicielle pour estimer

plmtest permet d'effectuer des tests du multiplicateur de Lagrange. L'argument principal est un objet **plm** auquel doivent impérativement être ajoutés deux arguments en lien respectivement avec le type de test et les effets testés. On aurait donc, en utilisant les données Grunfeld contenues dans le paquet plm :

```
R> g <- plm(inv ~ value + capital, data = Grunfeld, model = "pooling")  
R> plmtest(g, effect = "twoways", type = "bp")
```

"bp" fait ici référence au test de Breusch et Pagan.

Cas pratique : syndicalisation et chômage aux États-Unis

Description de la base de données utilisée

- Les données s'étendent de 2001 à 2017 et concernent 44 États américains.
- Notre base de données contient donc $n = 44$ États and $t = 18$ années.
- On assume que dans des États où il y a un fort taux de syndicalisation, mesuré par le nombre de salariés membres d'un syndicat en pourcentage, il y a un faible taux de chômage.
- On utilise comme variables de contrôle les salaires, les régions (i.e. Nord-est, Sud, Mid-ouest, Ouest) et la crise de 2008-2009.

Cas pratique : syndicalisation et chômage aux États-Unis

Les régions sont divisées comme suit (Wilson, 2002, p. 9) :

Nord-est	Sud	Mid-ouest	Ouest
Connecticut	Alabama	Illinois	Alaska
Delaware	Arkansas	Indiana	Arizona
Maine	Florida	Iowa	California
Maryland	Georgia	Kansas	Colorado
Massachusetts	Kentucky	Michigan	Hawaii
New Hampshire	Louisiana	Minnesota	Idaho
New Jersey	Mississippi	Missouri	Montana
New York	North Carolina	Nebraska	Nevada
Pennsylvania	Oklahoma	North Dakota	New Mexico
Rhode Island	South Carolina	Ohio	Oregon
Vermont	Tennessee	South Dakota	Utah
	Texas	Wisconsin	Washington
	Virginia		Wyoming
	West Virginia		

Cas pratique : syndicalisation et chômage aux États-Unis

Provenance des données utilisées

- Local Area Unemployment Statistics (LAUS) (pour le taux de chômage)
- Quarterly Census of Employment and Wages (QCEW) (pour les salaires)
- Current Population Survey (CPS) (pour la densité syndicale)

Cas pratique : syndicalisation et chômage aux États-Unis

Application du paquet plm

L'interface de RStudio, une fois les paquets utiles au bon fonctionnement du paquet plm installés, est le suivant :

```
> install.packages(c("plm", "zoo", "lmtest", "Formula", "formula.tools"))
Installing packages into '/Users/milene.r.e.lokrou/Library/R/3.5/library'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/macosx/el-capitan/contrib/3.5/plm_1.7-0.tgz'
Content type 'application/x-gzip' length 2230696 bytes (2.1 MB)
=====
downloaded 2.1 MB

trying URL 'https://cran.rstudio.com/bin/macosx/el-capitan/contrib/3.5/zoo_1.8-5.tgz'
Content type 'application/x-gzip' length 1086374 bytes (1.0 MB)
=====
downloaded 1.0 MB

trying URL 'https://cran.rstudio.com/bin/macosx/el-capitan/contrib/3.5/lmtest_0.9-37.tgz'
Content type 'application/x-gzip' length 348281 bytes (340 KB)
=====
downloaded 340 KB

trying URL 'https://cran.rstudio.com/bin/macosx/el-capitan/contrib/3.5/Formula_1.2-3.tgz'
Content type 'application/x-gzip' length 177499 bytes (173 KB)
=====
downloaded 173 KB

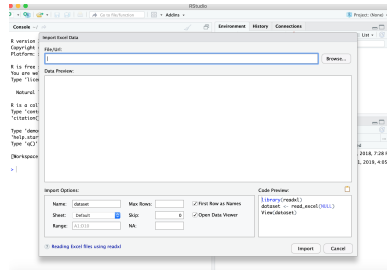
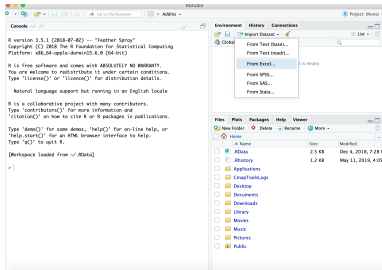
trying URL 'https://cran.rstudio.com/bin/macosx/el-capitan/contrib/3.5/formula.tools_1.7.1.tgz'
Content type 'application/x-gzip' length 86499 bytes (84 KB)
=====
```

Cas pratique : syndicalisation et chômage aux États-Unis

Application du paquet plm

Pour commencer l'analyse des données de panel, il nous faut importer la base de données **RAQC** dans R.

On peut importer une base de données dans R de plusieurs manières. J'ai pour habitude de choisir **Excel**.



Cas pratique : syndicalisation et chômage aux États-Unis

Application du paquet plm

Une fois la base de données importée, notre interface est la suivante :

The screenshot shows the RStudio interface with the RAQC dataset loaded. The Environment pane on the right shows 'RAQC' with 748 observations and 6 variables. The console on the left shows the R prompt and the output of the `View(RAQC)` command, which displays a table of data for Alabama from 2001 to 2012.

	States	Code	Year	Unemployment rate	wages	union membership
1	Alabama	AL	2001	5.100000	568	10.0
2	Alabama	AL	2002	5.900000	587	9.1
3	Alabama	AL	2003	6.025000	607	8.1
4	Alabama	AL	2004	5.700000	631	9.7
5	Alabama	AL	2005	4.500000	654	10.2
6	Alabama	AL	2006	4.066667	683	8.8
7	Alabama	AL	2007	3.975000	704	9.5
8	Alabama	AL	2008	5.716667	726	9.8
9	Alabama	AL	2009	10.991667	737	10.9
10	Alabama	AL	2010	10.541667	755	10.1
11	Alabama	AL	2011	9.616667	771	10.0
12	Alabama	AL	2012	7.983333	790	9.2

Showing 1 to 12 of 748 entries

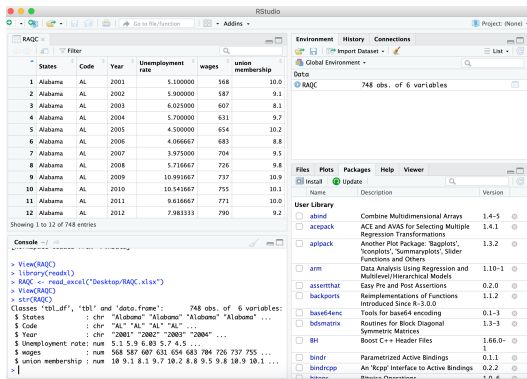
Console

```
Type "demo()" for some demos, "help()" for on-line help, or  
"help.start()" for an HTML browser interface to help.  
Type "q()" to quit R.  
  
[Workspace loaded from ~/.RData]  
  
> View(RAQC)  
> library(readxl)  
> RAQC <- read_excel("Desktop/RAQC.xlsx")  
> View(RAQC)  
> |
```

Cas pratique : syndicalisation et chômage aux États-Unis

Application du paquet plm

On veut voir la structure de notre base de données. On utilise la commande : `str`



The screenshot shows the RStudio interface. The top-left pane displays a data frame with 12 rows and 6 columns: States, Code, Year, Unemployment rate, wages, and union membership. The top-right pane shows the Environment tab with the RAQC dataset listed as 748 observations of 6 variables. The bottom-left pane shows the console output of the `str` command, which displays the structure of the RAQC dataset, including the number of observations (748) and the data types of the variables.

RAQC

	States	Code	Year	Unemployment rate	wages	union membership
1	Alabama	AL	2001	5.100000	568	10.0
2	Alabama	AL	2002	5.900000	587	9.1
3	Alabama	AL	2003	6.025000	607	8.1
4	Alabama	AL	2004	5.700000	631	9.7
5	Alabama	AL	2005	4.500000	654	10.2
6	Alabama	AL	2006	4.066667	683	8.8
7	Alabama	AL	2007	3.975000	704	9.5
8	Alabama	AL	2008	5.716667	726	9.8
9	Alabama	AL	2009	10.991667	737	10.9
10	Alabama	AL	2010	10.541667	755	10.1
11	Alabama	AL	2011	9.616667	771	10.0
12	Alabama	AL	2012	7.983333	790	9.2

Showing 1 to 12 of 748 entries

```
> View(RAQC)
> library(readxl)
> RAQC <- read_excel("Desktop/RAQC.xlsx")
> View(RAQC)
> str(RAQC)
Classes 'tbl_df', 'tbl' and 'data.frame':    748 obs. of  6 variables:
 $ States      : chr  "Alabama" "Alabama" "Alabama" "Alabama" ...
 $ Code       : chr  "AL" "AL" "AL" "AL" ...
 $ Year       : chr  "2001" "2002" "2003" "2004" ...
 $ Unemployment rate: num  5.1 5.9 6.03 5.7 4.5 ...
 $ wages      : num  568 587 607 631 654 683 704 726 737 755 ...
 $ union membership: num  10 9.1 8.1 9.7 10.2 8.8 9.5 9.8 10.9 10.1 ...
```

Environment History Connections

Global Environment -

Data

RAQC 748 obs. of 6 variables

Files Plots Packages Help Viewer

Install Update

User Library

Name	Description	Version
<input type="checkbox"/> abind	Combine Multidimensional Arrays	1.4-5
<input type="checkbox"/> acepack	ACE and AVAS for Selecting Multiple Regression Transformations	1.4.1
<input type="checkbox"/> aplpack	Another Plot Package: 'Bagplots', 'complots', 'Summaryplots', 'Slider Functions and Others	1.3.2
<input type="checkbox"/> arm	Data Analysis Using Regression and Multilevel/Hierarchical Models	1.10-1
<input type="checkbox"/> assertthat	Easy Pre and Post Assertions	0.2.0
<input type="checkbox"/> backports	Reimplementations of Functions Introduced Since R-3.0.0	1.1.2
<input type="checkbox"/> base64enc	Tools for base64 encoding	0.1-3
<input type="checkbox"/> bdsmatrix	Routines for Block Diagonal Symmetric Matrices	1.3-3
<input type="checkbox"/> BH	Boost C++ Header Files	1.66.0-1
<input type="checkbox"/> bindr	Parameterized Active Bindings	0.1.1
<input type="checkbox"/> bindrcpp	An 'Rcpp' Interface to Active Bindings	0.2.2
<input type="checkbox"/> bit64	Bit64 Operations	1.0-6

Cas pratique : syndicalisation et chômage aux États-Unis

Application du paquet plm

Nous allons créer de nouvelles variables dans notre base de données :

- ➊ Étape 1 : Créez une variable GFC, variable de contrôle pour la crise de 2008.
- ➋ Étape 2 : Créez des variables binaires de régions (i.e. Nord-est, Sud, Mid-ouest, Ouest), variables de contrôle pour les régions.

Cas pratique : syndicalisation et chômage aux États-Unis

Application du paquet plm

Notre modèle sera spécifié comme suit :

$$\ln Chom_{it} = \beta_1 \ln Synd_{it} + z_i \alpha + \beta_2 GFC + \beta_3 \ln Sal_{it} + u_{it}$$

Avec $i = 1, \dots, N$ et $t = 1, \dots, T$.

Où z_i contient un terme constant et les variables binaires de régions (i.e. Mid-ouest, Sud, Ouest et Nord-est).

Rappel : pour éviter une multicolinéarité entre les variables binaires de régions, n-1 régions seront incluses dans le modèle.

Cas pratique : syndicalisation et chômage aux États-Unis

Application du paquet plm

Où :

InChom représente l'évolution du taux de chômage

InSynd représente l'évolution du taux de syndicalisation

GFC représente la variable *crise de 2008*

InSal représente la croissance des salaires en pourcentage

Cas pratique : syndicalisation et chômage aux États-Unis

Application du paquet plm

La démarche de programmation avec R se fera durant l'atelier.

Les codes utilisés pour ce cas pratique se trouvent dans l'annexe ***Codes pour cas pratique _ paquet plm.***

Conclusion

Ce qu'il faut retenir !

- Le paquet plm offre des outils intéressants pour faire une analyse de données de panel avec R.
- Il existe plusieurs paquets qui permettent d'effectuer une analyse de données de panel avec R (e.g. lme4 et nlme).
- Cependant, l'intérêt du paquet plm réside dans le fait, à mon sens, qu'il s'adresse à des personnes familières avec le jargon de l'économétrie ou qui s'inscrivent dans une approche économétrique.
- Le paquet plm renferme des tests de spécification et d'autres tests utiles pour manipuler et analyser des données de panel et pour en interpréter les résultats.

Bibliographie sommaire

-  Joshua D Angrist et Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.
-  Badi Baltagi. *Econometric analysis of panel data*. John Wiley & Sons, 2008.
-  Yves Croissant, Giovanni Millo et al. "Panel data econometrics in R: The plm package". In : *Journal of statistical software* 27.2 (2008), p. 1-43.
-  Brigitte Dormont. "Petite apologie des données de panel". In : *Economie & prévision* 87.1 (1989), p. 19-32.
-  Cheng Hsiao. *Analysis of panel data*. 2^e éd. 54. Cambridge university press, 2003.
-  Jeffrey M Wooldridge. *Econometric Analysis of Cross-Section and Panel Data*. MIT press, 2002.