Endogeneity and Instrumental Variables Using the ivreg Package

*Henri Makika**

Julho 10, 2019

Contents

*State University of Campinas, São Paulo. E-mail : hd.makika@gmail.com

1 Introduction

Les variables instrumentales (IV) font référence à un ensemble de méthodes développées en économétrie à partir des années 1920 pour tirer des déductions causales dans des contextes où le traitement des intérêts ne peut être considéré de manière crédible comme attribué au hasard, même après avoir utilisé des covariables supplémentaires. Au cours des deux dernières décennies, ces méthodes ont attiré une attention considérable dans la littérature statistique. Bien que cette récente littérature statistique s'appuie sur la littérature économétrique antérieure, il existe néanmoins des différences importantes. Premièrement, la récente littérature statistique porte principalement sur le cas du traitement binaire. Deuxièmement, la littérature récente permet explicitement l'hétérogénéité des effets du traitement. Troisièmement, la littérature récente sur les variables instrumentales utilise explicitement le cadre de résultats potentiels utilisé par Neyman pour des expériences randomisées et généralisé en observation, voir par exemple Rubin (1974, 1978, 1990). Quatrièmement, dans les applications sur lesquelles se concentre cette littérature, y compris les expériences randomisées sur la non-conformité, les estimations en intention de traiter ou sous forme réduite présentent souvent un plus grand intérêt que dans les applications classiques d'équations simultanées en économétrie.

La littérature statistique récente a été motivée en partie par la littérature économétrique antérieure sur les variables instrumentales, à commencer par Wright (1928), [voir la discussion sur l'origine des variables instrumentales dans Stock et Trebbi, 2003]. Cependant, il existe également d'autres antécédents, en dehors de la littérature des variables instrumentales économétriques traditionnelles, notamment le travail de Zelen sur les modèles d'encouragement (Zelen, 1979, 1990). Les premiers articles de la littérature statistique récente incluent Angrist, Imbens et Rubin (1996), Robins (1989) et McClellan et Newhouse (1994). Les revues récentes incluent Rosenbaum (2010), Vansteelandt, Bowden, Babanezhad et Goetghebeur (2011) et Hernan et Robins (2006). Bien que ces revues incluent de nombreuses références à la littérature économique antérieure, il serait peut-être utile de discuter de la littérature économétrique de manière plus détaillée afin de fournir des informations de base et une perspective sur l'applicabilité des méthodes de variables instrumentales dans d'autres domaines.

Les méthodes de variables instrumentales constituent un élément central du canon de l'économétrie depuis la première moitié du XXe siècle et continuent de faire partie intégrante de la plupart des manuels de premier et de second cycle (par exemple, Angrist et Pischke, 2008; Bowden et Turkington, 1984; Greene, 2011; Hayashi, 2000; Manski, 1995; Stock et Watson, 2010; Wooldridge, 2002, 2008). Comme les statisticiens Fisher et Neyman (Fisher, 1925; Neyman, 1923), des économétriciens tels que Wright (1928), Working (1927), Tinbergen (1930) et Haavelmo (1943) étaient intéressés par des inférences causales, dans leur cas l'effet des politiques économiques sur le comportement économique. Cependant, contrairement à la littérature statistique sur l'inférence causale, le point de départ de ces économétriciens n'était pas l'expérience randomisée. Dès le départ, il a été reconnu que, dans les environnements étudiés, les causes ou les traitements n'étaient pas attribués à des unités passives (agents économiques dans leur environnement tels que des individus, des ménages, des entreprises ou des pays).

Au lieu de cela, les agents économiques influencent activement, ou même explicitement, le niveau de traitement qu'ils reçoivent. Le choix, plutôt que le hasard, a été le point de départ de la réflexion sur le mécanisme d'affectation dans la littérature en économétrie. Dans cette perspective, les unités recevant le traitement actif sont différentes de celles recevant le traitement de contrôle, pas seulement en raison de la réception du traitement: elles choisissent de recevoir le traitement actif car elles sont différentes pour commencer. Cela rend le traitement potentiellement endogène et crée ce que l'on appelle parfois dans la littérature économétrique le problème de sélection (Heckman, 1979).

Les premiers travaux d'économétrie sur les variables instrumentales n'avaient pas beaucoup d'impact sur la réflexion dans la communauté des statistiques. Bien que certains travaux techniques sur les propriétés d'échantillons volumineux de divers estimateurs aient été publiés dans des journaux de statistiques (par exemple, le très influent article d'Anderson et Rubin (1948)), les applications de non-économistes étaient rares. On ne sait pas exactement quelles en sont les raisons. L'une des possibilités est le fait que les premiers travaux sur les variables instrumentales étaient étroitement liés à des questions économiques de fond (par exemple, les interventions sur les marchés), en utilisant des concepts économiques théoriques qui peuvent sembler non pertinents ou difficiles à traduire dans d'autres domaines (par exemple, l'offre et la demande) .

Cela a peut-être suggéré aux non-économistes que les méthodes de variables instrumentales en général avaient une applicabilité limitée en dehors de l'économie. L'utilisation de concepts économiques n'était pas tout à fait inévitable, car les hypothèses critiques sur lesquelles reposent les méthodes de variables instrumentales sont substantielles et nécessitent une connaissance subtile de la matière(nous le verrons avec l'exemple ci-bas). Une autre raison peut être que, bien que les premiers travaux de Tinbergen et Haavelmo aient utilisé une notation très similaire à celle que Rubin (1974) a appelée par la suite la notation du résultat potentiel, la littérature s'est vite orientée vers une notation n'impliquant que des résultats réalisés ou observés. Voir pour une perspective historique Hendry et Morgan (1992) et Imbens (1997). Cette notation du résultat obtenu qui reste courante dans les manuels d'économétrie masque les liens entre les travaux de Fisher et Neyman sur les expériences randomisées et la littérature sur les variables instrumentales. Ce n'est que dans les années 1990 que les économétriciens sont revenus à la notation des résultats potentiels pour les questions de causalité (par exemple, Heckman, 1990, Manski, 1990; Imbens et Angrist, 1994), facilitant ainsi l'instauration d'un dialogue avec les statisticiens sur les méthodes de variable instrumentale.

Il sied de noter que, les premiers travaux en économétrie sont utiles pour comprendre la littérature relative aux variables instrumentales modernes et, en outre, potentiellement utiles pour améliorer les applications de ces méthodes et identifier les instruments potentiels. Ces méthodes peuvent en fait être utiles dans de nombreux contextes que les statisticiens étudient. L'exposition au traitement est rarement une simple question de hasard ou de choix. Les deux aspects sont importants et aident à comprendre quand les inférences causales sont crédibles et quand elles ne le sont pas. Ce faisant, notre objectif dans ce papier est donc de démontrer techniquement comment identifier les instruments potentiels dans l'analyse économique (partie 1) et résoudre ce problème à partir d'utilisation des données en panel(partie 2).

2 Variables instrumentales

À la base, les variables instrumentales changent les incitations pour les agents de choisir un niveau de traitement particulier, sans affecter les résultats potentiels associés à ces niveaux de traitement. Prenons un exemple de programme de formation professionnelle dans lequel le chercheur s'intéresse à l'effet moyen du programme de formation sur les gains. Chaque individu est caractérisé par deux résultats de gains potentiels, les gains en fonction de la formation et les gains en l'absence de formation. Chaque personne choisit de participer ou non en fonction de ses avantages nets perçus. Comme indiqué dans Athey et Stern (1998), il est important que ces avantages nets diffèrent des revenus. Vous le ferez par les coûts associés à la participation à ce régime. Supposons qu'il existe une variation dans les coûts encourus par les individus pour participer au programme de formation. Les coûts sont définis au sens large et peuvent inclure le temps de trajet pour se rendre dans les installations du programme ou l'effort requis pour s'informer sur le programme. De plus, supposons que ces coûts soient indépendants des résultats potentiels. C'est une hypothèse forte, souvent rendue plus plausible par le conditionnement sur des covariables. Les mesures du coût de la participation peuvent alors servir de variables instrumentales et aider à identifier les effets causals du programme. En fin de compte, nous comparons les gains des personnes à faible coût de participation au programme avec ceux des personnes à coût de participation élevé et attribuons la différence de gains moyens à la hausse du taux de participation au programme entre les deux groupes.

Dans presque tous les cas, l'hypothèse selon laquelle il n'y a pas d'effet direct de la modification des incitations sur les résultats potentiels est controversée et doit être évaluée au cas par cas. La deuxième partie de l'hypothèse, selon laquelle les coûts sont indépendants des résultats potentiels, peut-être après l'ajustement des covariables, est qualitativement très différente. Dans certains cas, il est satisfait par la conception, par exemple si les incitations sont randomisées. Dans les études d'observation, il s'agit d'une hypothèse de fond, non fondée, qui peut être plus plausible ou du moins à peu près durable après le conditionnement en covariables. Par exemple, dans un certain nombre d'études, les chercheurs ont utilisé la distance physique par rapport aux installations comme instrument d'exposition aux traitements disponibles dans ces installations. De telles études incluent McClellan et Newhouse (1994) et Baiocchi, Small, Lorch et Rosenbaum (2010) qui utilisent la distance à des hôpitaux dotés de capacités particulières comme instrument de traitement associé à ces capacités, après conditionnement à une distance du centre médical le plus proche, et Card. (1995), qui utilise la distance aux collèges comme un instrument pour fréquenter un collège.

3 Packages & définitions

```
library(readxl)      # read data excel
library(stargazer)   # Various programing for tables
library(ggplot2)     # Various programing tools for plotting data
library(AER)         # Applied Econometrics tests
library(tidyverse)   # Modern data science library
library(plm)         # Panel data analysis library
library(car)         # Companion to applied regression
library(gplots)      # Various programing tools for plotting data
library(tseries)     # For timeseries analysis
library(lmtest)      # For hetoroskedasticity analysis
```

Nous prénons l'exemple de salaire et vérifions s'il est sujet d'endogénéité de la variable expérience. Il s'agit donc d'un modèle avec une supposée variable instrumental. *Une variable explicative est endogène*, si toute variable explicative dans un modèle de régression linéaire qui est corrélée au terme d'erreur.

Problème d'endogénéité: deux ou plusieures variables sont déterminées conjointement dans un modèle, par exemple: variables prix et quantité dans un système d'offre et de demande. *Causes d'endogénéité*: (i) omission de variables pertinentes en corrélation avec $x_1, ..., x_k$; (ii) erreurs de mesure en $x_1, ..., x_k$, par exemple, un proxy mal spécifié; (iii) la simultanéité entre y et une ou plusieures variables explicatives. *Conséquences*: les estimateurs du moindre carré ordinaire (MCO) deviennent biaisés, incohérents et inefficaces. *Solution*: employer des variables instrumentales dans le but de faciliter la recherche d'estimateurs cohérents.

Une fois qu'on a un modèle où la covariance entre les variables explicatives et les erreurs sont non nulles, une variable z est considérée, (i) Pas de corrélation avec μ ne peut être observée et (ii) en corrélation qu'avec x. On appelle alors z de variable instrumentale ou tout simplement z est l'instrument de x.

Lorsque la régression réalisée pause un problème d'endogénéite, la conséquence ce que les estimateurs des moindres carrés ordinaires deviennent biaisés. Par conséquant, il est souhaitable d'utiliser les estimateurs des moindres carrés en deux étapes (MQ2E). Lorsque plusieures variables instrumentales sont requises pour une variable endogène (c'est-à-dire que l'équation structurelle est suridentifiée). Supposons le modèle structurel suivant :

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \mu_1,$$

(i) on suppose que l'une des variables explicatives est endogène y_2 de l'équation structurelle ci-dussus. En supposant alors l'existence de z_2 et z_3 corrélés à y_2 et non corrélés à μ_1; On arrive à la forme réduite en régressant y_2 sur z_1, z_2, z_3 (notez que l'estimateur z_1 sous-jacent au régresseur exogène β_2 est utilisé comme son propre instrument dans la matrice d'instrument Z). La forme réduite nécessite que les coefficients de z_2, z_3 soient non nuls pour que les variables instrumentales soient valides. Un test de restriction de Wald est alors appliqué pour la détection. Baum (2006) considère que si la statistique F dépasse 10, l'instrument est considéré comme fort.

(ii) l'estimation de l'équation sous forme réduite, on vérifie si les coefficients de z_2, z_3 sont non nuls; on utilise (ou non) l'estimateur y_2 en tant que variable instrumentale de y_1 sur z_1 et l'estimateur de y_2 (c'est-à-dire une VI).

Il est important de noter que (i) l'estimateur MQ2E est moins efficace que la méthode MCO lorsque les variables explicatives sont, en fait, exogènes; (ii) les MCO et MQ2E fournissent des estimateurs cohérents si la condition d'exogénéité est satisfaite; (iii) un test d'endogénéité d'une variable explicative doit être effectué pour vérifier s'il est nécessaire d'utiliser MQ2E.

Le test d'endogénéité Hausman est basé sur la comparaison des estimations OLS et MQ2E, afin de déterminer si les différences sont significativement différentes de zéro. La forme structurelle du modèle à estimer est la suivante :

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \mu,$$

```
MROZ <- read_excel("~/Videos/Inverno 2019/Dados/MROZ.xlsx")

head(MROZ)
```

```
## # A tibble: 6 x 6
##    WAGE  EDUC EXPER FATHEDUC MOTHEDUC HUSEDUC
##   <dbl> <dbl> <dbl>    <dbl>    <dbl>   <dbl>
## 1  3.35    12    14        7       12      12
## 2  1.39    12     5        7        7       9
## 3  4.55    12    15        7       12      12
## 4  1.10    12     6        7        7      10
## 5  4.59    14     7       14       12      12
## 6  4.74    12    33        7       14      11
```

```
dados = subset(MROZ, !is.na(MROZ$WAGE))

n = nrow(dados)
lwage = log(dados$WAGE)
educ = dados$EDUC
exper = dados$EXPER
exper2 = (dados$EXPER)^2
motheduc = dados$MOTHEDUC
fatheduc = dados$FATHEDUC
huseduc = dados$HUSEDUC

reg1 = lm (lwage ~ educ + exper + exper2)
```

Nous prenons la variable éducation de la mère (Motheduc) comme instrument. Et nous vérifions s'il existe une corrélation entre la variable instrumentale (Motheduc) et éducation.

```
reg.aux = lm (educ ~ motheduc)

stargazer(reg1, reg.aux, type = "text", digits = 4, column.labels = c("", ""),
          keep.stat = c('n','rsq','adj.rsq','f'), out = "mrd.txt")
```

```
##
## ===============================================================
##                               Dependent variable:
##              ---------------------------------------------------
##                        lwage                       educ
##
##                         (1)                        (2)
## ---------------------------------------------------------------
## educ                0.1075***
##                      (0.0141)
##
## exper               0.0416***
##                      (0.0132)
##
## exper2              -0.0008**
##                      (0.0004)
##
## motheduc                                        0.2674***
```

```
##                                                (0.0309)
##
## Constant                -0.5220***              10.1145***
##                          (0.1986)                (0.3109)
##
## ---------------------------------------------------------------
## Observations               428                     428
## R2                        0.1568                  0.1498
## Adjusted R2               0.1509                  0.1478
## F Statistic  26.2862*** (df = 3; 424) 75.0493*** (df = 1; 426)
## ===============================================================
## Note:                                *p<0.1; **p<0.05; ***p<0.01
```

Notez que la corrélation entre les variables est statistiquement différente (égale) à zéro. Ceci suggère que educ est une variable endogène. Ainsi, le modèle structurel est estimé en prenant en compte la variable instrumentale:

```
reg.iv = ivreg(lwage ~ educ + exper + exper2 | motheduc + exper + exper2)

stargazer(reg1, reg.aux, reg.iv, type = "text", digits = 4, column.labels = c("", "", ""),
          keep.stat = c('n','rsq','adj.rsq','f'), out = "mrd.txt")
```

```
##
## ============================================================================
##                                  Dependent variable:
##              ---------------------------------------------------------------
##                        lwage                    educ              lwage
##                         OLS                     OLS            instrumental
##                                                                  variable
##
##                         (1)                     (2)                (3)
## ----------------------------------------------------------------------------
## educ                0.1075***                                      0.0493
##                      (0.0141)                                     (0.0374)
##
## exper               0.0416***                                     0.0449***
##                      (0.0132)                                     (0.0136)
##
## exper2              -0.0008**                                     -0.0009**
##                      (0.0004)                                     (0.0004)
##
## motheduc                                    0.2674***
##                                              (0.0309)
##
## Constant            -0.5220***              10.1145***            0.1982
##                      (0.1986)                (0.3109)            (0.4729)
##
## ----------------------------------------------------------------------------
## Observations           428                     428                 428
## R2                    0.1568                  0.1498              0.1231
## Adjusted R2           0.1509                  0.1478              0.1169
## F Statistic  26.2862*** (df = 3; 424) 75.0493*** (df = 1; 426)
## ============================================================================
## Note:                                          *p<0.1; **p<0.05; ***p<0.01
```

On remarque que la valeur de l'estimateur educ diminue (augmente) lorsque l'effet corrélé avec le motheduc

est isolé. Nous considérons à présent un modèle avec plus d'une variable instrumentale (motheduc, fatheduc e huseduc).

```
reg2.aux = lm (educ ~ motheduc + fatheduc + huseduc + exper + exper2)
```

Pour évaluer si les instruments sont puissants, le test de restriction de Wald est utilisé:

```
linearHypothesis(reg2.aux, c("motheduc = 0", "fatheduc = 0", "huseduc = 0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## motheduc = 0
## fatheduc = 0
## huseduc = 0
##
## Model 1: restricted model
## Model 2: educ ~ motheduc + fatheduc + huseduc + exper + exper2
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    425 2219.2
## 2    422 1274.4  3    944.85 104.29 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Par conséquent, les instruments peuvent être considérés comme forts (faibles), lorsque le résultat du test F soit supérieur (inférieur) à 10. Pour notre cas nous remarquons que les instruments sont forts.

```
reg2.iv = ivreg(lwage ~ educ + exper + exper2 | motheduc + fatheduc + huseduc + exper + exper2)

stargazer(reg2.aux, reg.iv, reg2.iv, type = "text", digits = 4,
          column.labels = c("", "", "", ""),
          keep.stat = c('n','rsq','adj.rsq','f'), out = "mrd.txt")
```

```
##
## ============================================================
##                           Dependent variable:
##               ---------------------------------------------
##                          educ                    lwage
##                          OLS                instrumental
##                                               variable
##
##                          (1)                (2)       (3)
## ------------------------------------------------------------
## motheduc             0.1142***
##                       (0.0308)
##
## fatheduc             0.1061***
##                       (0.0295)
##
## huseduc              0.3753***
##                       (0.0296)
##
## educ                                      0.0493   0.0804***
##                                          (0.0374)  (0.0218)
##
## exper                  0.0375           0.0449*** 0.0431***
```

```
##                       (0.0343)         (0.0136)  (0.0133)
##
## exper2                -0.0006          -0.0009** -0.0009**
##                       (0.0010)         (0.0004)  (0.0004)
##
## Constant             5.5383***          0.1982    -0.1869
##                       (0.4598)         (0.4729)  (0.2854)
##
## ------------------------------------------------------------
## Observations            428               428       428
## R2                     0.4286           0.1231    0.1495
## Adjusted R2            0.4218           0.1169    0.1435
## F Statistic  63.3037*** (df = 5; 422)
## ============================================================
## Note:                           *p<0.1; **p<0.05; ***p<0.01
```

Notez que l'inclusion de plus de variables instrumentales pour expliquer la variable educ a rendu cette dernière statistiquement significative (non significative) par rapport au modèle avec une variable instrumentale. À présent nous vérifions l'homoscédasticité des erreurs.

```
bptest(reg2.iv)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  reg2.iv
## BP = 11.709, df = 3, p-value = 0.008449
```

Si la variance de l'erreur n'est pas constante (hétéroscédasticité), la statistique robuste est utilisée pour corriger le modèle:

```
summary(reg2.iv, vcov. = sandwich)
```

```
##
## Call:
## ivreg(formula = lwage ~ educ + exper + exper2 | motheduc + fatheduc +
##     huseduc + exper + exper2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.08378 -0.32135  0.03538  0.36934  2.35829
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.1868572  0.2998514  -0.623 0.533510
## educ         0.0803918  0.0216016   3.722 0.000225 ***
## exper        0.0430973  0.0152347   2.829 0.004893 **
## exper2      -0.0008628  0.0004197  -2.056 0.040413 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6693 on 424 degrees of freedom
## Multiple R-Squared: 0.1495,  Adjusted R-squared: 0.1435
## Wald test: 9.278 on 3 and 424 DF,  p-value: 5.913e-06
```

Nous appliquons à présent le test de Hausman et Sargan pour vérifier s'il y a nécessité d'utiliser les instruments dans la régression.

```r
summary(reg2.iv, vcov. = sandwich, diagnostics = TRUE)
```

```
##
## Call:
## ivreg(formula = lwage ~ educ + exper + exper2 | motheduc + fatheduc +
##     huseduc + exper + exper2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.08378 -0.32135  0.03538  0.36934  2.35829
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.1868572  0.2998514  -0.623 0.533510
## educ         0.0803918  0.0216016   3.722 0.000225 ***
## exper        0.0430973  0.0152347   2.829 0.004893 **
## exper2      -0.0008628  0.0004197  -2.056 0.040413 *
##
## Diagnostic tests:
##                  df1 df2 statistic p-value
## Weak instruments   3 422   108.139  <2e-16 ***
## Wu-Hausman         1 423     3.256  0.0719 .
## Sargan             2  NA     1.115  0.5726
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6693 on 424 degrees of freedom
## Multiple R-Squared: 0.1495,  Adjusted R-squared: 0.1435
## Wald test: 9.278 on 3 and 424 DF,  p-value: 5.913e-06
```

Pour le test de *Wu-Hausman* l'hypothèse nulle H_0: MCO = MQ2E, la variable n'est pas endogène, et H_1: la variable est endogène en supposant que les VIs sont exogènes. Pour vérifier l'exogénéité des VIs, le test de *Sargan* est utilisé, où H_0: tous les VIs ne sont pas corrélés avec l'erreur (les instruments sont valides) et H_1 : le contraire.

L'endogénéité peut se définir, de façon assez large, comme étant la corrélation existante entre une ou plusieurs variables indépendantes et le terme d'erreur de la régression. La complexité des décisions économiques ajoutée à l'information limitée dont dispose le chercheur fait que la quasi-totalité des études peut être sujette à ce biais.

Après une courte description dans la première partie des sources potentielles d'endogénéité (les variables omises, la simultanéité, les erreurs de mesure) et de leur impact sur l'estimation, des techniques de traitement des biais introduits à l'aide de chocs exogènes ou de variables instrumentales, nous abordons dans la deuxième partie l'économétrie de données en panel, c'est-à-dire à des techniques qui reposent plus nettement sur des hypothèses de modélisation, comme la méthode des effets fixes et des moments généralisés (voir Arellano et Bond, 1991).

Baltagi (2008) pense que la prise en compte de l'hétérogénéité des individus ou des firmes est fondamentale dans les études réalisées à l'aide de données de panel, ce que ne font pas la plupart des études en coupe instantanée. Par ailleurs, il constate que les études en séries temporelles sont souvent affectées par des corrélations sérielles, ce qui est sans doute moins vrai pour les données de panel qui ajoutent de la variabilité dans l'échantillon traité. Enfin, il assure que cette économétrie de panel est précieuse pour construire des modèles explicatifs dynamiques.

Par définition, on appelle *données de panel* des données qui comprennent plusieurs observations au cours du temps pour un même individu statistique (ou plusieurs individus). Nous démontrons, à l'aide d'un exemple,

comment utiliser les *packages* de RStudio pour traiter ce problème.

4 Importation et analyse exploratoire des données

```
dataPanel101 = read_csv("https://github.com/ds777/sample-datasets/blob/master/dataPanel101.csv?raw=true
```

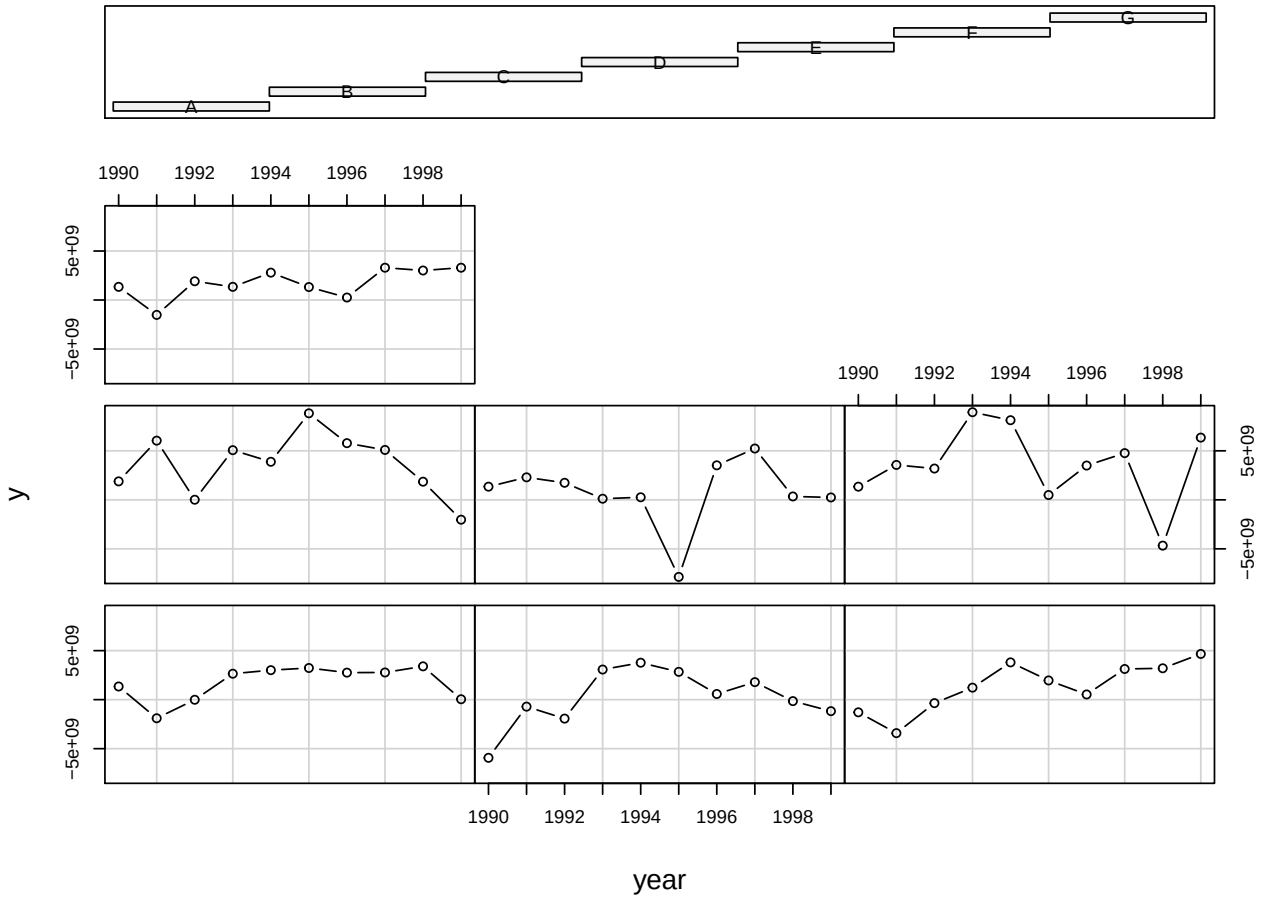
```
dataPanel101 = plm.data(dataPanel101, index = c("country","year"))
```

```
## Warning: use of 'plm.data' is discouraged, better use 'pdata.frame' instead
```

```
head(dataPanel101)
```

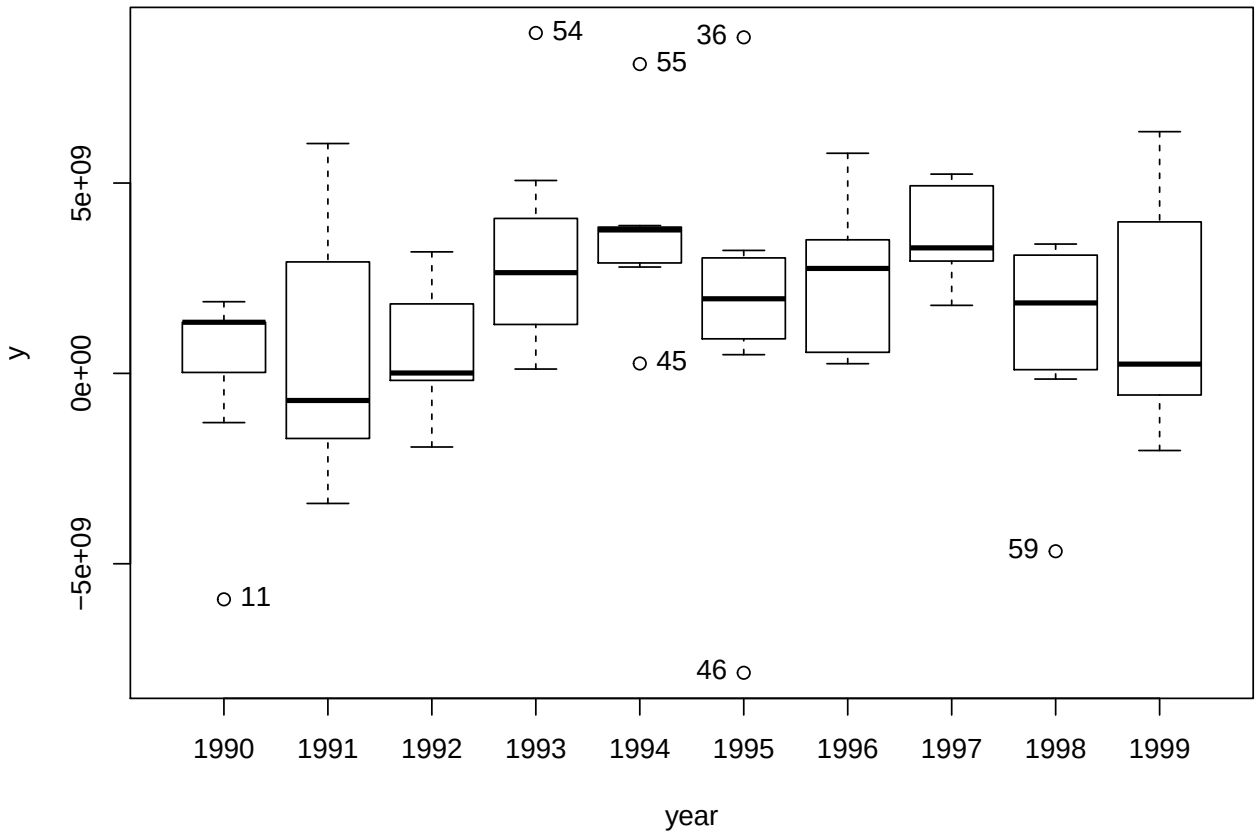
```
##   country year           y y_bin        x1         x2          x3
## 1       A 1990  1342787840     1 0.2779037 -1.1079559  0.28255358
## 2       A 1991 -1899660544     0 0.3206847 -0.9487200  0.49253848
## 3       A 1992   -11234363     0 0.3634657 -0.7894840  0.70252335
## 4       A 1993  2645775360     1 0.2461440 -0.8855330 -0.09439092
## 5       A 1994  3008334848     1 0.4246230 -0.7297683  0.94613063
## 6       A 1995  3229574144     1 0.4772141 -0.7232460  1.02968040
##     opinion
## 1 Str agree
## 2     Disag
## 3     Disag
## 4     Disag
## 5     Disag
## 6 Str agree
```

```
coplot(y ~ year|country, type = "b", data = dataPanel101)
```



Les barres en haut indiquent la position des pays de gauche à droite, en commençant par la ligne du bas.

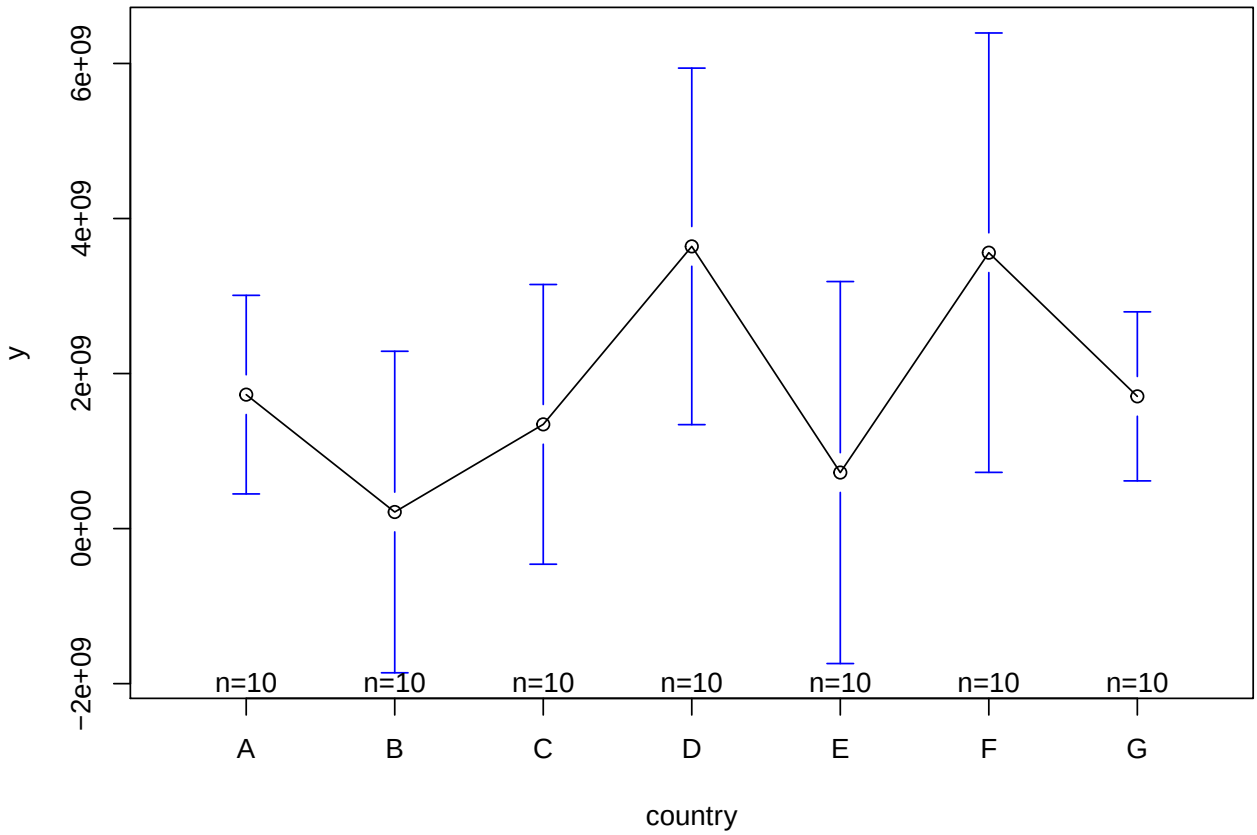
```
scatterplot(y ~ year|country, data = dataPanel101)
```




```
## [1] "11" "54" "45" "55" "46" "36" "59"
```

4.1 Hétérogénéité entre pays

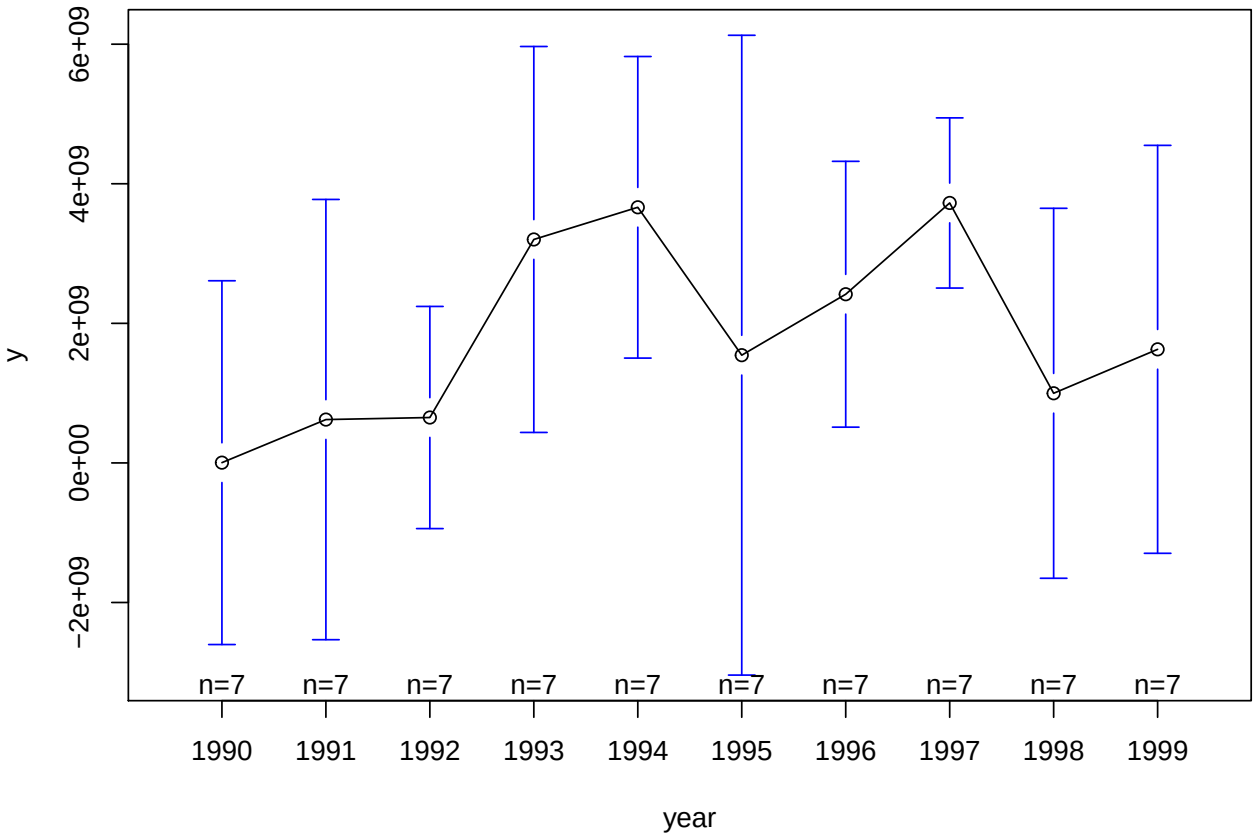
```
plotmeans(y ~ country, data = dataPanel101)
```



La fonction *plotmeans* dessine un intervalle de confiance de 95% autour des moyennes.

4.2 Hétérogénéité entre les années

```
plotmeans(y ~ year, data = dataPanel101)
```



5 Modélisation de données en panel

Les avantages de l'analyse des données de panel :

i. *Différences entre individus et périodes*: les modèles de données de panel nous permettent d'utiliser des variables binaires pour contrôler les différences entre les unités transversales (individus) et les périodes. Les données transversales ne fournissent pas suffisamment de degrés de liberté pour une telle analyse;

ii. *Degrés de liberté*: la taille d'échantillon d'une donnée de panel est le nombre d'unités transversales multiplié par le nombre de périodes. Dans les données transversales (séries chronologiques), nous n'avons que le nombre d'unités transversales (périodes);

iii. *Contrôler le biais de variable omis*: nous pouvons contrôler les éléments non observables liés à la fois aux régresseurs et à la régression (biais de variable omis) à l'aide de variables binaires ou de la transformation *Within*.

5.1 Modèle OLS de base

Le modèle de régression en utilisant l'estimateur OLS ne prend pas en compte l'hétérogénéité entre pays ni entre années.

```
ols <- lm(y ~ x1, data = dataPanel101)

stargazer(ols, type = "text", digits = 3, column.labels = c(""),
          keep.stat = NULL, out = "panel.txt")
```

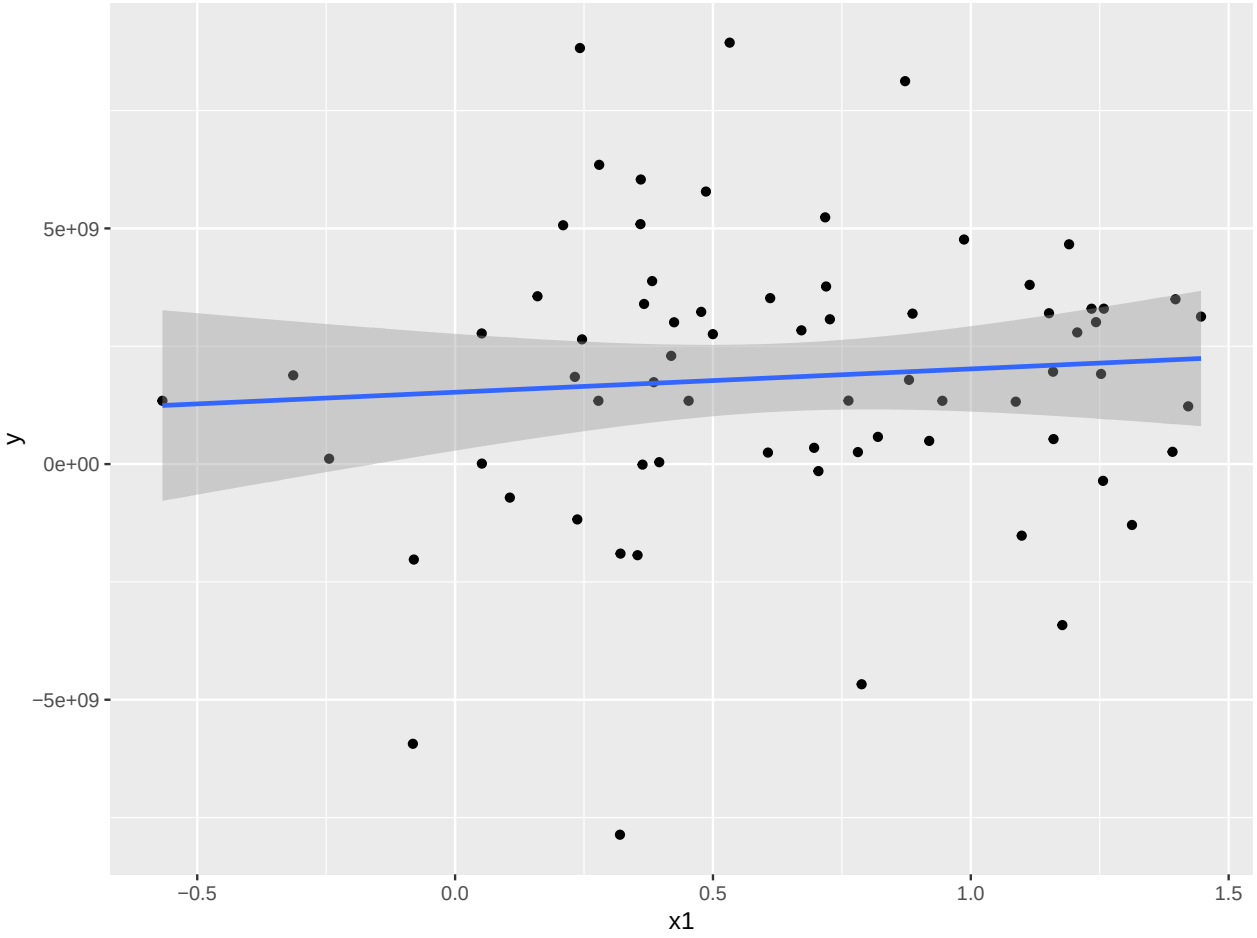
```
##
```

```
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                   y
##
## ------------------------------------------------
## x1                         494,988,865.000
##                           (778,861,258.000)
##
## Constant                 1,524,319,101.000**
##                           (621,072,623.000)
##
## ------------------------------------------------
## Observations                     70
## R2                              0.006
## Adjusted R2                     -0.009
## Residual Std. Error 3,028,276,250.000 (df = 68)
## F Statistic              0.404 (df = 1; 68)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

5.2 Graphique de Dispersion entre y et x_1

```
yhat = ols$fitted

ggplot(dataPanel101, aes(x = x1, y = y))+ geom_point() +
  geom_smooth(method = lm)
```



5.3 Estimateur à effets fixes

On fait maintenant l'hypothèse que les effets individuels β_i sont représentés par des constantes (d'où l'appellation estimateur à effets fixes).

i. Effets fixes spécifiques à un pays utilisant des variables binaires

```
fixed.dum = lm(y ~ x1 + factor(country) - 1, data = dataPanel101)

stargazer(fixed.dum, type = "text", digits = 3, column.labels = c(""),
          keep.stat = NULL, out = "panel.txt")
```

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                   y
##
## ------------------------------------------------
## x1                       2,475,617,742.000**
##                          (1,106,675,596.000)
##
## factor(country)A           880,542,434.000
##                           (961,807,055.000)
```

```
##
## factor(country)B        -1,057,858,320.000
##                         (1,051,067,687.000)
##
## factor(country)C        -1,722,810,680.000
##                         (1,631,513,767.000)
##
## factor(country)D       3,162,826,916.000***
##                          (909,459,152.000)
##
## factor(country)E         -602,621,958.000
##                         (1,064,291,688.000)
##
## factor(country)F        2,010,731,852.000*
##                         (1,122,809,099.000)
##
## factor(country)G         -984,717,393.000
##                         (1,492,723,120.000)
##
## ------------------------------------------------
## Observations                    70
## R2                             0.440
## Adjusted R2                    0.368
## Residual Std. Error 2,795,552,578.000 (df = 62)
## F Statistic            6.095*** (df = 8; 62)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

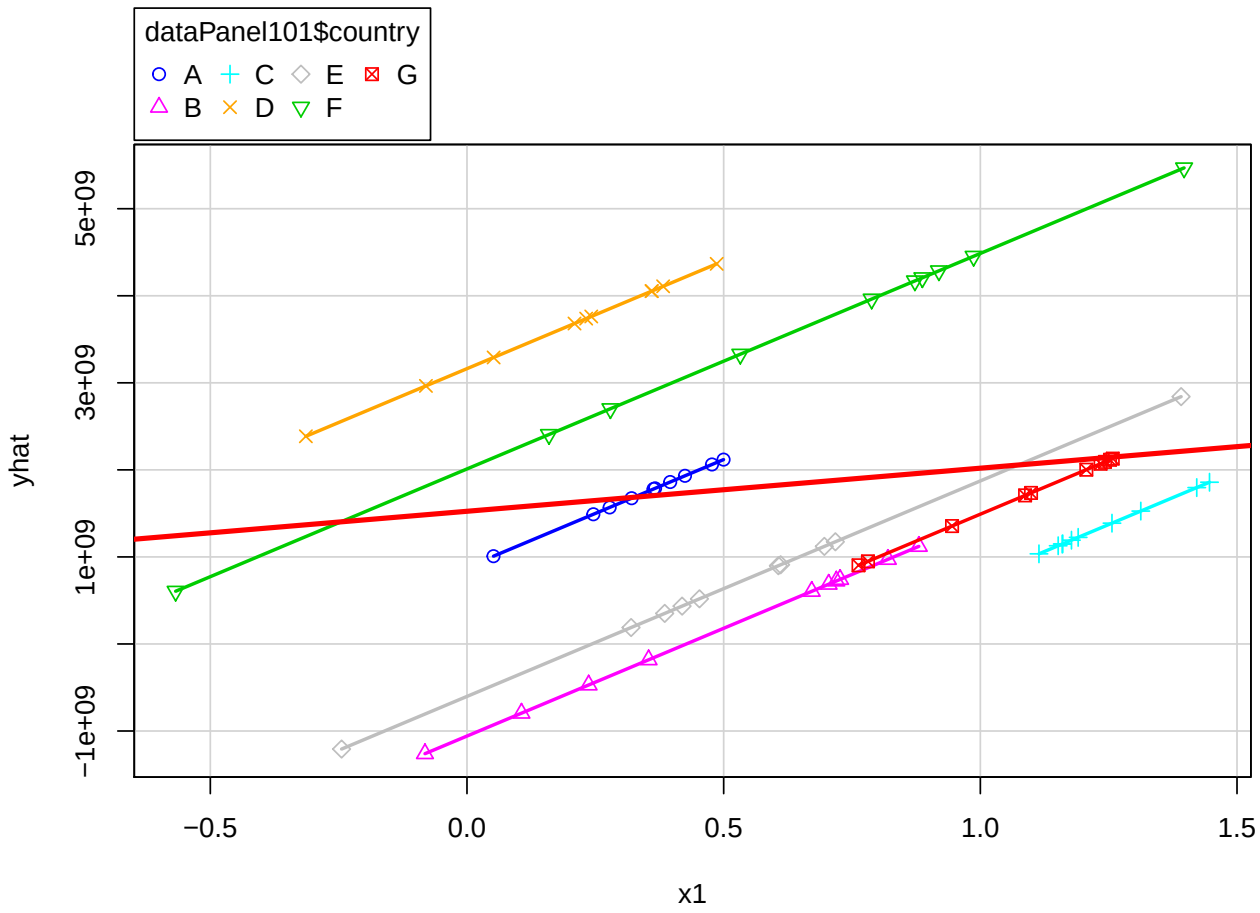
Chaque composante de la variable facteur (pays) absorbe les effets spécifiques de chaque pays. La variable x_1 n'était pas significative dans le modèle MCO. Cependant, une fois que les différences entre pays ont été maîtrisées, x_1 est devenu significatif dans le modèle à effets fixes.

ii. Graphique des effets fixes spécifiques à un pays utilisant des variables binaires

```
yhat = fixed.dum$fitted

scatterplot(yhat ~ dataPanel101$x1 | dataPanel101$country,  xlab ="x1",
            ylab ="yhat", boxplots = FALSE,smooth = FALSE)

abline(lm(dataPanel101$y~dataPanel101$x1),lwd=3, col="red")
```



iii. Effets fixes spécifiques à chaque pays using the package plm (Estimateur Within ou LSDV)

```
fixed.reg <- plm(y ~ x1, data = dataPanel101, model = "within")

summary(fixed.reg)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = y ~ x1, data = dataPanel101, model = "within")
##
## Balanced Panel: n = 7, T = 10, N = 70
##
## Residuals:
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -8.63e+09 -9.70e+08  5.40e+08  0.00e+00  1.39e+09  5.61e+09
##
## Coefficients:
##      Estimate Std. Error t-value Pr(>|t|)
## x1 2475617742 1106675596   2.237  0.02889 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    5.2364e+20
## Residual Sum of Squares: 4.8454e+20
## R-Squared:      0.074684
```

```
## Adj. R-Squared: -0.029788
## F-statistic: 5.00411 on 1 and 62 DF, p-value: 0.028892
```

Le coefficient de x_1 indique dans quelle mesure y modifie les heures supplémentaires, en moyenne par pays, lorsque x augmente de un. Nous affichons les effets fixes (constantes pour chaque pays):

```
fixef(fixed.reg)
```

```
##           A           B           C           D           E           F
##   880542434 -1057858320 -1722810680  3162826916  -602621958  2010731852
##           G
##  -984717393
```

On peut aussi faire un test pour comparer les modèle à effets fixes et OLS. L'hypothèse nulle dans ce cas est annocncée comme : H_0 : le modèle OLS est meilleurs que effets fixes.

```
pFtest(fixed.reg, ols)
```

```
##
##  F test for individual effects
##
## data:  y ~ x1
## F = 2.9655, df1 = 6, df2 = 62, p-value = 0.01307
## alternative hypothesis: significant effects
```

Si la valeur de p-value est inférieure à 0.05, le modèle à effets fixes sera un meilleur choix.

5.4 Estimateur à effets aléatoires

Dans la pratique standard de l'analyse économétrique, on suppose qu'il existe un grand nombre de facteurs qui peuvent affecter la valeur de la variable expliquée et qui pourtant ne sont pas introduits explicitement sous la forme de variables explicatives. Ces facteurs sont alors approximés par la structure des résidus. Le problème se pose de la façon similaire en économétrie de panel. La seule différence tient au fait que trois types de facteurs omis peuvent être envisagés. Il y a tout d'abord les facteurs qui affectent la variable endogène différemment suivant la période et l'individu considéré. Il peut en outre exister des facteurs qui affectent de façon identique l'ensemble des individus, mais dont l'influence dépend de la période considérée (effets temporel). Enfin, d'autres facteurs peuvent au contraire refléter des différences entre les individus de type structurelles, c'est-à-dire indépendantes du temps (effets individuel).

```
random.reg = plm(y ~ x1, data = dataPanel101, model = "random")

summary(random.reg)
```

```
## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora's transformation)
##
## Call:
## plm(formula = y ~ x1, data = dataPanel101, model = "random")
##
## Balanced Panel: n = 7, T = 10, N = 70
##
## Effects:
##                      var   std.dev share
## idiosyncratic 7.815e+18 2.796e+09 0.873
## individual    1.133e+18 1.065e+09 0.127
## theta: 0.3611
##
```

```
## Residuals:
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -8.94e+09 -1.51e+09  2.82e+08  0.00e+00  1.56e+09  6.63e+09
##
## Coefficients:
##               Estimate Std. Error t-value Pr(>|t|)
## (Intercept) 1037014329  790626206  1.3116   0.1941
## x1          1247001710  902145599  1.3823   0.1714
##
## Total Sum of Squares:    5.6595e+20
## Residual Sum of Squares: 5.5048e+20
## R-Squared:      0.02733
## Adj. R-Squared: 0.013026
## F-statistic: 1.91065 on 1 and 68 DF, p-value: 0.17141
```

L'interprétation des coefficients est compliquée car elle inclut les effets au sein de l'entité et entre les entités. Dans le cas des données TSCS (*time series and cross-section*), cela représente l'effet moyen de X sur Y lorsque X change dans le temps et d'un pays à l'autre.

Pour décider entre effets fixes et effets aléatoires, on peut exécuter un test de *Hausman* dans lequel l'hypothèse nulle affirme que le modèle le plus approprié est celui à effets aléatoires; l'alternative serait à effets fixes (voir Green, 2008, chapitre 9). Essentiellement, on teste si les erreurs sont corrélées avec les régresseurs, l'hypothèse nulle est qu'elles ne les sont pas. Si la valeur p est significative (< 0.05, par exemple), effets fixes sont utilisés, sinon on utilise à effets aléatoires.

Le test de spécification d'Hausman (1978) est un test général qui peut être appliqué à des nombreux problèmes de spécification en économétrie. Mais son application la plus répandue est celle des tests de spécification des effets individuels en panel. Il sert ainsi à discriminer les effets fixes et aléatoires.

```
phtest(fixed.reg, random.reg)
```

```
##
##  Hausman Test
##
## data:  y ~ x1
## chisq = 3.674, df = 1, p-value = 0.05527
## alternative hypothesis: one model is inconsistent
```

Dans ce cas, il faut utiliser le modèle à effets aléatoires.

6 Diagnostic de régression

6.1 Test à effets fixes dans le temps

```
fixed.time = plm(y ~ x1 + factor(year), data = dataPanel101, model = "within")

summary(fixed.time)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = y ~ x1 + factor(year), data = dataPanel101, model = "within")
##
## Balanced Panel: n = 7, T = 10, N = 70
##
```

```
## Residuals:
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -7.92e+09 -1.05e+09 -1.40e+08  0.00e+00  1.63e+09  5.49e+09
##
## Coefficients:
##                     Estimate Std. Error t-value Pr(>|t|)
## x1                1389050208 1319849568  1.0524  0.29738
## factor(year)1991   296381592 1503368532  0.1971  0.84447
## factor(year)1992   145369724 1547226550  0.0940  0.92550
## factor(year)1993  2874386825 1503862558  1.9113  0.06138 .
## factor(year)1994  2848156370 1661498931  1.7142  0.09233 .
## factor(year)1995   973941363 1567245752  0.6214  0.53698
## factor(year)1996  1672812635 1631539257  1.0253  0.30988
## factor(year)1997  2991770146 1627062033  1.8388  0.07156 .
## factor(year)1998   367463673 1587924443  0.2314  0.81789
## factor(year)1999  1258751990 1512397631  0.8323  0.40898
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    5.2364e+20
## Residual Sum of Squares: 4.0201e+20
## R-Squared:      0.23229
## Adj. R-Squared: 0.0005285
## F-statistic: 1.60365 on 10 and 53 DF, p-value: 0.13113
```

Le test à effets fixes dans le temps: L'hypothèse nulle est qu'aucun effet fixe n'est nécessaire dans le temps.

```
pFtest(fixed.time, fixed.reg)
```

```
##
##  F test for individual effects
##
## data:  y ~ x1 + factor(year)
## F = 1.209, df1 = 9, df2 = 53, p-value = 0.3094
## alternative hypothesis: significant effects
```

Pour le test du Multiplicateur de Lagrange (Breusch-Pagan) l'hypothèse nulle est qu'aucun effet fixe n'est nécessaire dans le temps.

```
plmtest(fixed.reg, c("time"), type = "bp")
```

```
##
##  Lagrange Multiplier Test - time effects (Breusch-Pagan) for
##  balanced panels
##
## data:  y ~ x1
## chisq = 0.16532, df = 1, p-value = 0.6843
## alternative hypothesis: significant effects
```

Si la valeur p est inférieure à 0.05, des effets fixes sont utilisés. Dans cet exemple, il n'est pas nécessaire d'utiliser des effets fixes au fil du temps.

6.2 Effets aléatoires vs OLS groupés (empilés):

Le test du Multiplicateur de Lagrange (LM) permet de choisir entre une régression à effets aléatoires et une régression simple par MCO.

L'hypothèse nulle dans le test de LM est que les variations entre les entités sont nulles. C'est-à-dire sans différence significative entre les unités (c'est-à-dire sans effet de panel).

```
pool = plm(y ~ x1, data = dataPanel101, model = "pooling")

summary(pool)
```

```
## Pooling Model
##
## Call:
## plm(formula = y ~ x1, data = dataPanel101, model = "pooling")
##
## Balanced Panel: n = 7, T = 10, N = 70
##
## Residuals:
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -9.55e+09 -1.58e+09  1.55e+08  0.00e+00  1.42e+09  7.18e+09
##
## Coefficients:
##               Estimate Std. Error t-value Pr(>|t|)
## (Intercept) 1524319101  621072623  2.4543  0.01668 *
## x1           494988866  778861258  0.6355  0.52722
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    6.2729e+20
## Residual Sum of Squares: 6.2359e+20
## R-Squared:      0.0059046
## Adj. R-Squared: -0.0087145
## F-statistic: 0.403897 on 1 and 68 DF, p-value: 0.52722
```

Le test du Multiplicateur de Lagrange de Breusch-Pagan pour effets aléatoires est donnée pour : l'hypothèse nulle est qu'il n'y a pas d'effet de panel (c'est-à-dire une meilleure méthode MCO)

```
plmtest(pool, type = c("bp"))
```

```
##
##  Lagrange Multiplier Test - (Breusch-Pagan) for balanced panels
##
## data:  y ~ x1
## chisq = 2.6692, df = 1, p-value = 0.1023
## alternative hypothesis: significant effects
```

Ici, nous ne pouvons pas rejeter l'hypothèse nulle, nous concluons donc que l'estimateur à effets aléatoires n'est pas approprié. Autrement dit, il n'existe aucune preuve de différences significatives entre les pays. Par conséquent, il est possible d'effectuer une simple régression par la méthode de moindre carré ordinaire (MCO).

6.3 Test de dépendance en coupe transversale

Pour tester la dépendance en coupe transversale de chaque variable, il est fort récommandé d'utiliser le test d'indépendance de Breusch-Pagan (LM) et le test de Pesaran, 2004 (statistique CD). L'hypothèse nulle H_0 est qu'il y a indépendance en coupe transversale. La statistique CD de Pesaran est basée sur la moyenne des coefficients de corrélation entre les différents pays pris deux-à-deux pour chaque période de temps. Sous l'hypothèse nulle, cette statistique est asymptotiquement distribuée selon une normale standard $N(0,1)$. Ce test peut être basé également sur les coefficients de corrélation des résidus obtenus par MCO (De Hoyos et Sarafidis, 2006).

Selon Baltagi, la dépendance transversale est un problème dans les macro-panels à longue série chronologique. Ce n'est pas un gros problème dans les micro-panels (quelques années et un grand nombre de cas).

```
fixed = plm(y ~ x1, data = dataPanel101, model = "within")

pcdtest(fixed, test = c("lm"))
```

```
##
##  Breusch-Pagan LM test for cross-sectional dependence in panels
##
## data:  y ~ x1
## chisq = 28.914, df = 21, p-value = 0.1161
## alternative hypothesis: cross-sectional dependence
```

```
pcdtest(fixed, test = c("cd"))
```

```
##
##  Pesaran CD test for cross-sectional dependence in panels
##
## data:  y ~ x1
## z = 1.1554, p-value = 0.2479
## alternative hypothesis: cross-sectional dependence
```

Une fois que les tests montrent que $p - value > 0.05$, on conclut qu'il n'y a pas de dépendance transversale comme le cas pour notre exemple.

6.4 Test de corrélation sérielle

Les tests de corrélation sérielle s'appliquent aux données de macro-panels avec des séries chronologiques longues. Ils ne posent pas problème aussi dans les micro-panels (avec quelques années). L'hypothèse nulle H_0 dit qu'il n'y a pas corrélation sérielle.

```
pbgtest(fixed)
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel
##  models
##
## data:  y ~ x1
## chisq = 14.137, df = 10, p-value = 0.1668
## alternative hypothesis: serial correlation in idiosyncratic errors
```

On remarque que la valeur statistique de p est supérieure à 0.05, on conclut que l'hypothèse nulle ne peut pas être rejetée.

6.5 Tests de racine unitaire

Rappelons que les tests de racine unitaire individuels sont reconnus pour leur faible puissance (Cochrane, 1991) spécialement dans le cas de petites valeurs de T. En général, les tests de racine unitaire utilisés ont comme hypothèse nulle H_0 que la variable testée possède une racine unitaire. La faible puissance du test signifie qu'on est souvent dans l'impossibilité de rejeter l'hypothèse nulle et on conclut incorrectement que la variable possède une racine unitaire. Les tests de racine unitaire de données panel présentent l'avantage d'être plus puissants et de remédier ainsi à ce problème (Baltagi et Kao, 2000 et Baltagi, 2013).

```
adf.test(dataPanel101$y, k = 2)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  dataPanel101$y
## Dickey-Fuller = -3.9051, Lag order = 2, p-value = 0.0191
## alternative hypothesis: stationary
```

Par la valeur statistique de p qui est supérieure à 0.05, nous concluons que la série n'a pas de racine unitaire. En d'autres termes, la série est stationnaire. L'hypothèse nulle peut être alors rejetée.

6.6 Test d'hétéroscédasticité

L'hypothèse nulle H_0 est que le modèle est homoscédastique. Nous observons la présence de l'hétéroscédasticité, il est possible de considérer la statistique robuste.

```
bptest(y ~ x1 + factor(country), data = dataPanel101, studentize = F)
```

```
##
##  Breusch-Pagan test
##
## data:  y ~ x1 + factor(country)
## BP = 14.606, df = 7, p-value = 0.04139
```

Si l'hétéroscédasticité est détectée, il est nécessaire d'utiliser une matrice de covariance robuste (estimateur en sandwich) pour en rendre compte.

6.6.1 Contrôler l'hétéroscédasticité: effets aléatoires

La fonction *vcovHC* combine trois estimateurs de covariance compatibles avec l'hétéroscédasticité:

White1: pour l'hétéroscédasticité générale, mais pas de corrélation en série. Le test est recommandé pour les effets aléatoires.

White2: white1 est limité à une variation commune au sein des groupes. Ici aussi, le test est plus recommandé pour les effets aléatoires.

Arellano: L'hétéroscédasticité et la corrélation sérielle. Ce test est recommandé pour les effets fixes. Pour ainsi, les options suivantes s'appliquent:

HC0 - hétéroscédasticité cohérente. Il s'agit donc de la norme.

HC1, HC2, HC3 - sont recommandés pour les petits échantillons. Il est important de noter que *HC3* donne moins de poids aux observations influentes.

HC4 - sont recommandés pour les petits échantillons avec des observations d'affluents *HAC* - hétéroscédasticité et autocorrélation cohérente (type VcovHAC pour plus de détails).

```
coeftest(random.reg) # Coefficients originals
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1037014329  790626206  1.3116   0.1941
## x1          1247001710  902145599  1.3823   0.1714
```

```
coeftest(random.reg, vcovHC) # Coefficients consistents-heteroscedasticite
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1037014329  907983024  1.1421   0.2574
## x1          1247001710  828970258  1.5043   0.1371
```

```
coeftest(random.reg, vcovHC(random.reg, type = "HC3")) # Coefficients consistents:
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1037014329  943438278  1.0992   0.2756
## x1          1247001710  867137595  1.4381   0.1550
```

```
                                                    # heteroscedasticite, type 3

t(sapply(c("HC0", "HC1", "HC2", "HC3", "HC4"),
         function(x) sqrt(diag(vcovHC(random.reg, type = x)))))
```

```
##     (Intercept)        x1
## HC0   907983024 828970258
## HC1   921238952 841072654
## HC2   925403814 847733484
## HC3   943438278 867137595
## HC4   941376025 866024042
```

```
# Affiche les erreurs HC standard des coefficients
```

6.6.2 Contrôler l'hétéroscédasticité: effets fixes

```
coeftest(fixed)
```

```
##
## t test of coefficients:
##
##      Estimate Std. Error t value Pr(>|t|)
## x1 2475617742 1106675596   2.237  0.02889 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
coeftest(fixed, vcovHC(fixed, method = "arellano"))
```

```
##
## t test of coefficients:
##
##      Estimate Std. Error t value Pr(>|t|)
## x1 2475617742 1358388924  1.8225  0.07321 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
coeftest(fixed, vcovHC(fixed, type = "HC3"))
```

```
##
## t test of coefficients:
##
##      Estimate Std. Error t value Pr(>|t|)
```

```
## x1 2475617742 1439083498  1.7203  0.09037 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
t(sapply(c("HC0", "HC1", "HC2", "HC3", "HC4"),
         function(x) sqrt(diag(vcovHC(fixed, type = x)))))
```

```
##          HC0.x1     HC1.x1     HC2.x1     HC3.x1     HC4.x1
## [1,] 1358388924 1368196913 1397037348 1439083498 1522166001
```

Textes sources

Arellano, M., and Bond, S. Some Tests of Specification for Panel Data : Monte Carlo Evidence and an Application to Employment Equations. *The Review of Economic Studies* 58, 2 (1991), 277-297.

Baltagi, B. Econometric Analysis of Panel Data. *John Wiley & Sons*, 2013.

Baltagi, B., and Kao, C. Nonstationary Panels, Cointegration in Panels and Dynamic Panels : A Survey. *Syracuse University Center for Policy Research Working Paper*, 16 (2000).

Baum, C.F., Schaffer, M.E.; Stillman, S. Instrumental variables and GMM: Estimation and Testing. *Boston College*. Working Paper No. 545, 2003.

Baum, C.F. Residual Diagnostics for Cross-Section Time Series Regression Models. *The Stata Journal* 1, 1 (2001), 101-104.

Blundell, R., and Bond, S. Initial Conditions and Moment Restrictions in Dynamic Panel Data Models. *Journal of Econometrics* 87, 1 (1998), 115-143.

Bond, S., Leblebicioglu, A., and Schiantarelli, F. Capital Accumulation and Growth : a New Look at the Empirical Evidence. *Journal of Applied Econometrics* 25, 7 (2010), 1073-1099.

Cameron, C. and Trivedi, P. Microeconometrics: Methods and Applications. *Stata Coorp*, LP, 2010.

Cochrane, J. H. A Critique of the Application of Unit Root Tests. *Journal of Economic Dynamics and Control* 15, 2 (1991), 275-284.

Greene, W. H. Econometric Analysis, 7th ed. *Prentice Hall*, 2012.

Hamilton, J. D. Time Series Analysis. *Princeton University Press*, Princeton, 1994.

Im, K.S., Pesaran, M. H., and Shin, Y. Testing for Unit Roots in Heterogeneous Panels. *Journal of Econometrics* 115, 1 (2003), 53-74.

Pesaran, M. H. General Diagnostic Tests for Cross Section Dependence in Panels. CESifo Working Paper no. 1229, *CESifo Group Munich*, 2004.

Pesaran, M. H. Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure. *Econometrica* 74, 4 (2006), 967-1012.

Pesaran, M.H. A Simple Panel Unit Root Test in the Presence of Cross-Section Dependence. *Journal of Applied Econometrics* 22, 2 (2007), 265-312.

Pesaran, M. H., and Smith, R. Estimating Long-Run Relationships from Dynamic Heterogeneous Panels. *Journal of Econometrics* 68, 1 (1995), 79-113.

Wooldridge, J.M. Econometric Analysis of Cross Section and Panel Data, 2nd ed. *The MIT Press*, 2010.

Wooldridge, J.M. Introductory Econometrics: a Modern Approach. *Stata Press College Station*, USA, (2006).