Séance 9 – Travailler sur des comparaisons longitudinales : questions d'estimation

Pierre Pora

Rappel des séances précédentes

- Ce que l'on a vu jusque là
  - Pour les premières séances :
    - Un outil de comparaison **entre des unités d'une même population**
    - Ces comparaisons sont essentiellement des **comparaisons de moyennes entre des groupes** définis par les variables indépendantes
    - Quand on observe seulement un **échantillon** et pas la population toute entière, les calculs que l'on fait sur l'échantillon sont **informatifs** quant à ce qui se passe dans la population

Rappel des séances précédentes

- Ce que l'on a vu jusque là
  - Pour la **dernière séance** :
    - En adaptant un tout petit peu l'outil
    - Et lorsque l'on a des **données où l'on peut suivre les mêmes individus en plusieurs points du temps**
    - On peut faire de petites adaptations du cadre pour passer de **comparaisons entre individus** (*inter* / *between* / en coupe) à des **comparaisons longitudinales** (*intra* / *within*)

Ce que l'on va voir aujourd'hui

- On sait **interpréter les coefficients** que donnent les régressions linéaires pour ces comparaisons longitudinales
- On va exprimer un tout petit peu plus abstraitement ces coefficients
- On va se pencher un peu plus précisément sur l'**estimation** et l'**inférence**
- En rentrant moins dans les détails que ce que l'on avait fait pour les comparaisons entre individus
  - Mais quand même...

Partir d'un exemple (toujours le même !)

```r
library(AER)
library(bife)

data("psid")

psid[,
     children :=
       as.numeric(
         KID1 > 0
         | KID2 > 0
         | KID3 > 0
       )]
```

Partir d'un exemple (toujours le même !)

- On veut estimer la régression
  $LFP_{it} = \beta \texttt{children}_{it} + \lambda_i + u_{it}$ avec
  - $\mathbb{E}[\texttt{children}_{it} u_{is}] = 0$ pour toutes les périodes t et s
    - Important ! Les résidus doivent ne pas être corrélés avec les covariables des autres périodes
  - $\mathbb{E}[C_i u_{it}] = 0$ pour tous les individus i
- On n'a **aucune restriction sur les** λ_i, et on sait que le coefficient β s'interprète comme une **moyenne sur la population des femmes mariées** de **comparaisons longitudinales** (au long de la vie de chaque femme) entre leur activité quand elles ont et n'ont pas d'enfants qui vivent avec elles

Peut-on avoir une écriture formalisée de β ?

- Pour chaque individu on peut construire
  - $\bar{Y}_i = \frac{1}{T}\sum_{t=1}^{T} Y_{it}$
    - C'est juste le **taux d'activité moyen** de chaque femme sur la **période observée**
  - $\ddot{Y}_{it} = Y_{it} - \bar{Y}_i$
  - Et exactement la même chose sur les variables indépendantes X_{it}
  - Et aussi sur les résidus u_{it}
    - Evidemment **inobservés** mais on peut toujours raisonner dessus !

Peut-on avoir une écriture formalisée de β ?

- Si on empile tous les Y_{it} dans un vecteur colonne Y_i de taille T
- Et qu'on définit la matrice Q comme $Q = I_T - \frac{1}{T} e_T e_T'$
  - où $e_T = (1 \dots 1)'$
    - donc $e_T e_T'$ est la matrice carrée de taille T remplie de 1

$$Q := \begin{array}{cc} \begin{pmatrix} 1-\frac{1}{T} & -\frac{1}{T} & \cdots & -\frac{1}{T} & -\frac{1}{T} \\ -\frac{1}{T} & 1-\frac{1}{T} & -\frac{1}{T} & -\frac{1}{T} & -\frac{1}{T} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -\frac{1}{T} & -\frac{1}{T} & \cdots & 1-\frac{1}{T} & -\frac{1}{T} \\ -\frac{1}{T} & -\frac{1}{T} & \cdots & -\frac{1}{T} & 1-\frac{1}{T} \end{pmatrix} & T \text{ lignes} \\ T \text{ colonnes} & \end{array}$$

Peut-on avoir une écriture formalisée de β ?

- Avec ces notations $QY_i = \ddot{Y}_i$
- Et de la même façon $QX_i = \ddot{X}_i$ et $Qu_i = \ddot{u}_i$
- On en tire $QY_i = QX_i\beta + Qu_i$ avec $\mathbb{E}[(QX_i)'Qu_i] = 0$
  - Car $\ddot{Y}_{it} = Y_{it} - \frac{1}{T}\sum_{s=1}^{s} Y_{is}$
  - Mais $Y_{it} = \lambda_i + X'_{it}\beta + u_{it}$
  - En substituant $\ddot{Y}_{it} = \ddot{X}'_{it}\beta + \ddot{u}_{it}$
    - Avec $\mathbb{E}[\ddot{X}_{it}\ddot{u}_{it}] = 0$
    - Et c'est vrai pour t variant de 1 à T

Peut-on avoir une écriture formalisée de β ?

- On retombe bien sur nos pieds :
  - QY_i est un vecteur colonne de taille T
  - QX_i est une matrice de taille $T \times d$
  - β un vecteur de colonne de taille d
  - Qu_i est un vecteur de taille T

Peut-on avoir une écriture formalisée de β ?

- On retrouve le **même principe que ce qu'on faisait pour les MCO**
  - On part de $\mathbb{E}[(QX_i)'Qu_i] = 0$
  - On substitue $\mathbb{E}[(QX_i)'(QY_i - QX_i\beta)] = 0$
  - Par linéarité $\beta = \mathbb{E}[X_i'Q'QX_i]^{-1}\mathbb{E}[X_i'Q'QY_i]$
    - A condition que $\mathbb{E}[X_i'Q'QX_i]$ soit bien inversible !
  - En définitive $\beta = \mathbb{E}[X_i'QX_i]^{-1}\mathbb{E}[X_i'QY_i]$
    - Car $Q' = Q$ et $Q^2 = Q$

Que dit la condition de rang ?

- $\mathbb{E}[X_i' Q X_i]$ doit être inversible
  - Cela revient exactement à dire que $\mathbb{E}[\ddot{X}_i' \ddot{X}_i]$ est inversible
  - Il doit donc y avoir **indépendance linéaire** des différentes composantes de $\ddot{X}_i$
    - Au sens de l'algèbre linéaire une fois encore !
  - Quand on n'a qu'une dimension, cela revient simplement à dire qu'**il y a des individus qui changent de valeur de X_{it} au cours du temps**
    - Si ce n'est pas le cas alors $\ddot{X}_i = 0$
    - Distinction *stayers* / *movers* classique dans la littérature
    - Les coefficients ne portent (évidemment !) que sur les *movers*

Que dit la condition de rang ?

- Si on est à plusieurs dimensions alors il faut (en gros) que **les changements individuels n'interviennent pas au même moment pour toutes les dimensions**
  - C'est tout bête !
  - Si on suivant une population de femmes à qui peuvent arriver deux choses
    - Se marier
    - Devenir mères
  - Si toutes ont leur premier enfant l'année où elles se marient
  - Alors on ne sait pas quand on fait des comparaisons longitudinales ce qu'il faut mettre sur les différences entre non-mariées et mariées et ce qu'il faut mettre sur les différences entre sans enfants et mères

Et maintenant : l'estimation !

- On a un échantillon de 1461 femmes observées 9 années consécutives
- Intuitivement, et du point de vue de l'estimation, qu'est-ce qui va marcher différemment des cas que l'on a étudié auparavant ?

Quel est le régime asymptotique pertinent ?

- On raisonne toujours dans la **limite d'un grand nombre d'observations**
- Schématiquement on a **trois raisons d'avoir beaucoup d'observations**
  - On observe **beaucoup d'individus**
    - Ici de femmes mariéees
  - On les suit pendant **très longtemps**
    - Ici beaucoup d'années
  - **Les deux**
- Ici c'est clairement la première raison qui compte (*large* N) et pas la seconde (*large* T)

Quelles implications ?

- Chaque femme supplémentaire de l'échantillon nous donne
  - **Une comparaison longitudinale de plus à rajouter dans l'estimation de** β (pourvu que son statut change au cours de la période d'observation !)
  - Un λ_i de plus dans l'équation
  - Mais aucune information sur les valeurs de λ_i pour les autres femmes

Quelles implications ?

- On a dit la semaine dernière que la régression revenait à **régresser $\texttt{LFP}_{it}$ sur $\texttt{children}_{it}$ et les indicatrices spécifique à chaque femme**
- En augmentant la taille de l'échantillon, on rajoute des comparaisons *intra* dont β n'est que la moyenne (avec de bons poids)
  - Donc on a une **information de plus en plus pertinente sur β**
- En revanche les λ_i qui seraient juste les coefficients sur les indicatrices **ne bénéficient pas de ce gain en précision** !

Quelles implications ?

- Pour le dire de façon plus savante
  - Dans le régime *large* N qui est le plus pertinent ici (et dans la grande majorité des applications en micro-économétrie appliquée)
  - On va facilement réussir à construire un **estimateur** $\hat{\beta}$ qui **convergera en probabilité vers** β
  - Et on aura encore des résultats de **normalité asymptotique** (et des tests etc.)
  - Mais on ne saura jamais rien dire d'intelligent sur les λ_i
    - **Même quand ce serait intéressant**
    - Même quand les logiciels de calcul statistique nous fournissent sournoisement des estimations…
    - Pour ça il faut être en *large* T (c'est logique !)

Comment construire l'estimateur *within* ?

- Il n'y a rien de difficile
- Mais la formalisation est quand même un peu pénible !
- On part d'un **échantillon** de N individus indexés par i, tirés de façon **équiprobable** dans une **grande population (infinie)**
  - Pour chacun d'entre eux on observe $((Y_{i1}, X_{i1}), \dots, (Y_{iT}, X_{iT}))$
  - Comme on tire de façon équiprobable dans la même population infinie, les grosses v.a. qui empilent toutes les trajectoires des individus i entre les instants 1 et T sont **indépendantes** et **identiquement distribuées**

Comment construire l'estimateur ? Un petit formalisme matriciel (encore !)

- Si on repart des grosses écritures matricielles où on empile toutes les observations : $\mathbf{X}$ et $\mathbf{Y}$ qui représentent les observations empilées
  - D'abord l'observation de l'individu tiré en 1 à la période 1, puis période 2, etc. puis T, puis l'individu tiré en 2 à la période 1 etc.
- L'équivalent de la matrice Q c'est de récupérer les résidus de la régression sur les indicatrices d'être une observations relative à chaque femme
- On peut noter $\mathbf{C}$ la matrice dans laquelle on empile les observations de ces indicatrices, avec les observations mises dans le bon ordre

Comment construire l'estimateur ? Un petit formalisme matriciel (encore !)

$$\mathbf{C} := \begin{array}{cl} \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 1 & \cdots & 0 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 1 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix} & NT \text{ lignes} \\ N \text{ colonnes} & \end{array}$$

Comment construire l'estimateur ? Un petit formalisme matriciel (encore !)

- La matrice $\mathbf{C}$ est la matrice des régresseurs dans les régression sur toutes les indicatrices
  - Les estimateurs associés à ces régressions sont $(\mathbf{C}'\mathbf{C})^{-1}\mathbf{C}'\mathbf{Y}$ et $(\mathbf{C}'\mathbf{C})^{-1}\mathbf{C}'\mathbf{X}$
  - Donc les matrices qui empilent les résidus de ces régressions s'écrivent $\ddot{\mathbf{Y}} = \left(I_{NT} - \left(\mathbf{C}'\mathbf{C}\right)^{-1}\mathbf{C}'\right)\mathbf{Y}$ et $\ddot{\mathbf{X}} = \left(I_{NT} - \left(\mathbf{C}'\mathbf{C}\right)^{-1}\mathbf{C}'\right)\mathbf{X}$

Comment construire l'estimateur ? Un petit formalisme matriciel (encore !)

- On sait que l'on a $\ddot{Y}_i = \ddot{X}_i\beta + \ddot{u}_i$ avec $\mathbb{E}[\ddot{X}_i'\ddot{u}_i] = 0$
  - Et donc $\beta = \mathbb{E}[\ddot{X}_i'\ddot{X}_i]^{-1}\mathbb{E}[\ddot{X}_i'\ddot{Y}_i]$

- On n'a plus qu'à remplacer les espérances par les moyennes empiriques
  - $\hat{\beta} = (\ddot{\mathbf{X}}'\ddot{\mathbf{X}})^{-1}\ddot{\mathbf{X}}'\ddot{\mathbf{Y}}$

- Est-ce bien ce qu'on veut ?
  - $\ddot{\mathbf{X}}'\ddot{\mathbf{X}}$ est une matrice carrée de taille d
  - $\ddot{\mathbf{X}}'\ddot{\mathbf{X}}_{lm} = \sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{X_{it}^{l}}X_{it}^{m}$
  - Donc $\frac{1}{N}\ddot{\mathbf{X}}'\ddot{\mathbf{X}}$ est l'équivalent en moyenne empirique de $\mathbb{E}[\ddot{X}_i{}'\ddot{X}_i]$
  - Même raisonnement pour l'autre terme

Est-ce que ça converge vers ce qu'on veut ?

- **Loi faible des grands nombres** :
  - $\frac{1}{N}\ddot{\mathbf{X}}'\ddot{\mathbf{X}}$ converge en probabilité vers $\mathbb{E}[\ddot{X}_i'\ddot{X}_i]$
  - $\frac{1}{N}\ddot{\mathbf{X}}'\ddot{\mathbf{Y}}$ converge en probabilité vers $\mathbb{E}[\ddot{X}_i'\ddot{Y}_i]$
  - donc $\hat{\beta}$ converge en probabilité vers $\beta = \mathbb{E}[\ddot{X}_i'\ddot{X}_i]^{-1}\mathbb{E}[\ddot{X}_i'\ddot{Y}_i]$

Comportement asymptotique

- On ne va pas refaire tout le détail mais (en gros) on retombe sur le calcul du **comportement asymptotique pour les données groupées !**
  - Avec des *clusters* définis au niveau de chaque femme
  - Et c'est juste complètement logique !
    - **On est bien sur des données groupées** puisqu'on commence par tirer des femmes et ensuite on observe toutes les périodes sur lesquelles on les voit...
- Remarque : beaucoup d'ouvrages traitent aussi du cas où on ne prend pas en compte le caractère groupé des données mais c'est très peu crédible...

Un résumé rapide

- Avec un petite réarrangement : $\hat{\beta} - \beta = (\ddot{\mathbf{X}}'\ddot{\mathbf{X}})^{-1}\ddot{\mathbf{X}}'\ddot{\mathbf{u}}$
- Donc $\sqrt{N}\left\{\hat{\beta} - \beta\right\} = \left(\frac{1}{N}\ddot{\mathbf{X}}'\ddot{\mathbf{X}}\right)^{-1}\sqrt{N}\left(\frac{1}{N}\ddot{\mathbf{X}}'\ddot{\mathbf{u}}\right)$
  - **Loi faible des grands nombres** : le premier terme converge en probabilité vers $\mathbb{E}[\ddot{X}_i'\ddot{X}_i]^{-1} = \mathbb{E}[X_i'QX_i]^{-1}$
  - **TCL** : le second terme converge en loi vers $\mathcal{N}\left(0, \mathbb{E}\left[X_i'Qu_iu_i'QX_i\right]\right)$
  - **Théorème de Slutsky** : le produit converge en loi vers $\mathcal{N}\left(0, \mathbb{E}[X_i'QX_i]^{-1}\mathbb{E}\left[X_i'Qu_iu_i'QX_i\right]\mathbb{E}[X_i'QX_i]^{-1}\right)$
  - L'expression est très laide mais on peut **estimer simplement la matrice de variance-covariance asymptotique** par $(\ddot{\mathbf{X}}'\ddot{\mathbf{X}})^{-1}(\ddot{\mathbf{X}}'\ddot{\mathbf{u}}\ddot{\mathbf{u}}'\ddot{\mathbf{X}})(\ddot{\mathbf{X}}'\ddot{\mathbf{X}})^{-1}$

Une toute petite extension

- Jusque là on a traité le cas où on suit **tous les individus sur la même période** et où on observe chaque individu autant de fois que les autres
  - Panel cylindré / *balanced panel*
- Ca marche exactement pareil pour le cas où le panel n'est pas cylindré
  - Quand on veut interpréter économiquement / causalement les comparaisons longitudinales, il faut juste parfois prendre garde aux raisons pour lesquelles certains individus sont observés moins longtemps que les autres...

Et en pratique avec R ?

- C'est la seule question réellement pertinente !
- Dans la plupart des cas, les *packages* qui font ces estimations proposent par défaut des matrices de variance-covariance
  - **Au pire** sous hypothèse d'homoscédasticité et sans prise en compte du caractère groupé des données
  - **Souvent** sans prise en compte du caractère groupé des données
  - **Ces matrices ne sont pas adaptées à la situation**
    - On regarde peu d'observations répétées de beaucoup d'individus
    - La population dans laquelle on tire est définie au niveau des individus pas au niveau des individus $\times$ périodes

Le cas catastrophique

```r
reg_longitudinale <-
  lm(LFP ~ children + factor(ID),
     data = psid)

reg_longitudinale$coefficients["children"]
```

```
  children
-0.0522696
```

```r
sqrt(vcov(reg_longitudinale)["children","children"])
```

```
[1] 0.01183346
```

Une petite amélioration, et pour finir le bon écart-type

```r
sqrt(vcovHC(reg_longitudinale,
            type = "HC0")["children","children"])
```

```
[1] 0.01123627
```

```r
sqrt(vcovCL(reg_longitudinale,
            cluster = psid$ID,
            type = "HC0"))["children","children"]
```

```
[1] 0.0156519
```

En pratique avec R

- Utiliser `lm` en introduisant explicitement les effets fixes dans la régression n'est pas une bonne solution
  - Pousse à calculer des λ_i qui sont inutilisables et prennent de la capacité de calcul
  - Les calculs sont vite assez longs
- **Ma recommandation : utiliser le package `lfe`**
  - Explicitement fait pour estimer ça !
  - Peut gérer des effets fixes sur plusieurs dimensions $\rightarrow$ utile pour vous à terme (cf conclusion de la séance)
  - On peut préciser qu'on veut la **matrice de variance-covariance groupée** dés le début

Avec le package lfe

```
library(lfe)

reg_longitudinale_lfe <-
  felm(LFP ~ children | ID | 0 | ID,
       data = psid)

coeftest(reg_longitudinale_lfe)
```

```
t test of coefficients:

          Estimate Std. Error t value  Pr(>|t|)
children -0.052270   0.015652 -3.3394 0.0008608 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Une petite vérification

```r
all.equal(
  as.numeric(reg_longitudinale$coefficients["children"]),
  as.numeric(reg_longitudinale_lfe$coefficients[1]))
```

```
[1] TRUE
```

```r
all.equal(
  sqrt(vcovCL(reg_longitudinale,
              cluster = psid$ID, type = "HC0"))["children"
  as.numeric(sqrt(vcov(reg_longitudinale_lfe))),
  scale = 1)
```

```
[1] "Mean absolute difference: 5.952536e-07"
```

Avec le package lfe

```
reg_longitudinale_lfe_mauvaise_vcov <-
  felm(LFP ~ children | ID | 0 | 0,
       data = psid)

coeftest(reg_longitudinale_lfe_mauvaise_vcov)
```

```
t test of coefficients:

          Estimate Std. Error t value  Pr(>|t|)
children -0.052270   0.011833 -4.4171 1.009e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Une autre solution

```
psid[,
     mean_indiv_children := mean(children),
     by = c("ID")]

reg_mundlak <-
  lm(LFP ~ children + mean_indiv_children,
     data = psid)
```

Une autre solution

```r
coeftest(reg_mundlak,
         vcov = vcovCL(reg_mundlak,
                       cluster = psid$ID))
```

```
t test of coefficients:

                     Estimate Std. Error t value  Pr(>|t|)
(Intercept)          0.737654   0.022529 32.7425 < 2.2e-16
children            -0.052270   0.015653 -3.3393 0.0008424
mean_indiv_children  0.033888   0.030033  1.1284 0.2591834
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Une dernière vérification

```
all.equal(
  sqrt(vcovCL(reg_longitudinale,
              cluster = psid$ID,
              type = "HC0"))["children","children"],
  sqrt(vcovCL(reg_mundlak,
              cluster = psid$ID,
              type = "HC0"))["children","children"])
```

```
[1] TRUE
```

En guise de conclusion

- Avec des **adaptations minimales**, on peut utiliser des régressions linéaires pour faire des **comparaisons longitudinales** et les considérer **en moyenne dans une certaine population**
- Ce n'est pas très compliqué d'utiliser R pour ça !
  - Mais il faut bien se demander si on estime la **bonne matrice de variance-covariance** !
  - Au total ce n'est pas trop compliqué
    - Si on a réussi à avoir une population où on observe le même individu plusieurs fois
    - C'est bien que l'on a commencé par tirer des individus puis par les suivre
    - Par parce que l'on tire des individus $\times$ périodes au hasard...
    - **Les écarts-types pour les données groupées sont donc assez naturels à utiliser**

En guise de conclusion

- On peut bien entendu combiner et raffiner tout cela :
  - **Comparer entre des groupes de la population des évolutions individuelles** $\rightarrow$ modèles à doubles effets fixes, différences-de-différences etc.
    - Vous verrez tout ça dans la suite de votre parcours
    - Il y a évidemment à chaque fois des subtilités !
- On n'a pratiquement pas parlé de l'**interprétation causale des coefficients**
  - Vous avez vu que **tous les coefficients s'interprètent simplement comme des comparaisons de moyenne**
  - La question n'est donc jamais de savoir si vous avez utilisé une régression ou une autre façon de faire une comparaison
  - Mais seulement de savoir **à quelle condition ces comparaisons ont une interprétation causale !**

Ce qu'il reste pour ce semestre

- Une petite séance de TD sur les données de panel
- Une séance d'examen
  - Ce sera facile !
  - Sur ordinateur
  - Quelques manipulations de données et estimer quelques régressions avec R
  - Lire une table de régression issue d'un article
  - Quelques questions théoriques (faciles !!)