Séance 8 – Que faire des données de panel ?

Pierre Pora

Rappel des séances précédentes

- On dispose d'un outil qui :
  - Génère des **coefficients** qui permettent de **comparer des moyennes** entre des groupes au sein d'une population
  - Donnes des valeurs prédites qui **approximent une variable dépendante** (par exemple le salaire) comme une **fonction linéaire des variables indépendantes** (par exemple l'éducation)
- On s'est dotés d'un **cadre probabiliste** qui permet de justifier qu'on peut utiliser un **échantillon** beaucoup plus petit que la population entière, mais quand même assez grand
  - Pour un assez grand échantillon, ce qu'on calcule dedans ressemble à ce qu'on aurait en population entière
  - Et on a des résultats qui permettent de quantifier l'**erreur** qu'on commet en confondant les deux

Et maintenant ?

- On revient à des questions au niveau de la **population**
- Jusque là implicitement on imaginait regarder une population **à un moment donné**
  - Sans préjugé de la nature des entités qui composent la population (personnes, entreprises, postes de travail etc.)
- La dimension temporelle importe et l'on n'est pas obligés de se restreindre à comparer entre eux des individus de la même population observés au même moment
  - Comparaisons transversales / en couple / *cross-section* vs. comparaisons longitudinales
- L'objet de la séance est de réfléchir un peu aux **comparaisons longitudinales**
  - En restant dans le cadre où on regarde une grosse population $\neq$ suivre un seul individu au cours du temps (par exemple le PIB français sur 50 ans)

Partir d'un exemple

```
library(bife)

data("psid")
```

Partir d'un exemple

- On suit le comportement d'activité des femmes américaines mariées pendant 9 années consécutives
- Il y a donc deux dimensions : si on repense à l'expérience aléatoire qui consiste à tirer un individu dans la population : - On tire un individu i : ici une femme - On la suit au cours du temps indicé par t : les années au cours desquelles on s'intéresse à ses décisions d'offre de travail

Partir d'un exemple

```
psid
```

	ID	LFP	KID1	KID2	KID3	INCH	AGE	TIME
	<int>	<int>	<int>	<int>	<int>	<num>	<num>	<int>
1:	1	1	1	1	1	58807.81	26	1
2:	1	1	1	0	2	41741.87	27	2
3:	1	1	0	1	2	51320.73	28	3
4:	1	1	0	1	2	48958.58	29	4
5:	1	1	0	1	2	53634.62	30	5

13145:	6365	1	0	0	0	19066.25	27	5
13146:	6365	1	0	0	0	16286.32	28	6
13147:	6365	1	0	0	0	18769.85	29	7
13148:	6365	1	0	0	0	18763.11	30	8
13149:	6365	1	0	0	0	20617.02	31	9

Partir d'un exemple

- Ici `ID` permet de suivre **la même femme** au cours du temps
- Et `TIME` précise **la vague (année) d'observation**

Une petite vérification

```
length(unique(psid$ID))
```

```
[1] 1461
```

```
length(unique(psid$TIME))
```

```
[1] 9
```

```
all.equal(
  nrow(psid),
  length(unique(psid$ID)) * length(unique(psid$TIME))
)
```

```
[1] TRUE
```

Une petite manipulation de données

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

Partir d'un exemple

- La variable `children` indique qu'une femme est mère de famille à un moment donné
- On se pose des questions sur les liens entre parentalité et offre de travail (variable `LFP` : activité)
- Comment peut-on les éclairer ?

Les comparaisons en coupe

- Si on refait seulement ce qu'on a déjà fait plein de fois

```
reg_coupe_naive <-
  lm(LFP ~children,
     data = psid)

reg_coupe_naive$coefficients
```

```
(Intercept)    children
 0.74456008 -0.02748251
```

Les comparaisons en coupe

- Mais ici on compare des femmes avec et sans enfant probablement observées à des moments différents !

```
psid[,
     mean(children),
     by = c("TIME")]
```

```
    TIME        V1
   <int>     <num>
1:     1 0.7809719
2:     2 0.7926078
3:     3 0.7871321
4:     4 0.7816564
5:     5 0.7665982
6:     6 0.7590691
7:     7 0.7453799
8:     8 0.7255305
9:     9 0.6906229
```

Les comparaisons en coupe

- Il y a bien une solution !

```
reg_coupe_repetee <-
  lm(LFP ~ children + factor(TIME),
     data = psid)

reg_coupe_repetee$coefficients[1:2]
```

```
(Intercept)    children
 0.72648723 -0.02488856
```

Les comparaisons en coupe

- Rappel des séances précédentes : le coefficient sur `children` s'interprète ici comme la **moyenne sur les années** des **écarts de taux d'activité entre femmes avec et sans enfants** mesurés chaque année, avec des **poids proportionnels** à la taille de l'année dans la population $(i \times t)$ et à la **variance de** `children`
  - On a la même taille pour chaque année $\rightarrow$ panel cylindré (*balanced panel*)
  - On met juste plus de poids sur les années où on est plus proches de 50% de femmes avec des enfants

Une petite vérification

```
psid_agrege_annne <-
  psid[,
       list(
         ecart_lfp = sum(LFP * children)/
           sum(children) -
           sum(LFP * (1 - children)) /
           sum(1 - children),
         var_children = var(children)),
       by = c("TIME")]

ecart_agrege_coupe_repetee <-
  psid_agrege_annne[,
                    sum(ecart_lfp * var_children) /
                      sum(var_children)]
ecart_agrege_coupe_repetee
```

```
[1] -0.02488856
```

Une petite vérification

```
all.equal(
  as.numeric(ecart_agrege_coupe_repetee),
  as.numeric(reg_coupe_repetee$coefficients["children"])
)
```

```
[1] TRUE
```

Et les comparaisons longitudinales ?

- Jusque là, on compare des mères et des femmes sans enfant, au même moment
  - **Comparaisons transversales / en coupe (*cross-section*)**
- Mais comme on sait suivre la même femme au cours de sa vie on pourrait faire d'autres comparaisons !
  - **Comparaisons longitudinales**
  - Comment procéderiez-vous ici ?

Et les comparaisons longitudinales ?

- Quelques idées :
  - Comparer les années avec et sans enfants de la même femme, puis moyenne sur les femmes
    - Peut se faire à la main
    - Mais aussi avec une régression !
  - Comparer les observations avec et sans enfants de femmes qui passent le même temps avec des enfants, sans se soucier des années d'observation et des individus
  - Différencier les données d'une année sur l'autre pour savoir quand les femmes deviennent mères de famille (ou cessent de l'être) et regarder la corrélation entre les changeements d'activité et ces changements de statut
  - Différencier les données par rapport au nombre d'année passées dans l'état de mère, et au nombre d'années passées en activité pour chaque mère, et s'intéresser à la régression de l'activité différenciée sur la parentalité différenciée

Et les comparaisons longitudinales ?

- En fait **tout ça revient à peu près au même !**
- En se souvenant bien des séances sur l'**interprétation des coefficients des régressions linéaires**, on peut le retrouver rapidement
- On va quand même le revoir ensemble

Comparer les années avec et sans enfant de la même femme

```
ecarts_activite_individuels <-
  psid[,
       list(
         ecart_activite =
           sum(children * LFP) /
           sum(children) -
           sum((1 - children) * LFP) /
           sum(1 - children),
         var_children = var(children)),
     by = c("ID")]
```

Comparer les années avec et sans enfants de la même femme

```
ecarts_activite_individuels
```

```
         ID ecart_activite var_children
      <int>          <num>        <num>
   1:     1            NaN    0.0000000
   2:    19            NaN    0.0000000
   3:    21            NaN    0.0000000
   4:    22            NaN    0.0000000
   5:    25            NaN    0.0000000
  ---
1457:  6293            NaN    0.0000000
1458:  6321            NaN    0.0000000
1459:  6331              0    0.1111111
1460:  6363            NaN    0.0000000
1461:  6365            NaN    0.0000000
```

Comparer les années avec et sans enfants de la même femme

- Il faut choisir les poids...
  - Beaucoup de solutions possibles, on peut penser à deux diférentes

```
ecarts_individuels_agreges <-
  ecarts_activite_individuels[,
    list(agreg_poids_unif =
           mean(ecart_activite,
                na.rm = TRUE),
         agreg_var =
           sum(ecart_activite * var_children,
               na.rm = TRUE) /
           sum(var_children)
         )]
```

Comparer les années avec et sans enfants de la même femme

```
ecarts_individuels_agreges
```

```
   agreg_poids_unif  agreg_var
              <num>      <num>
1:      -0.05702467 -0.0522696
```

Faire la même chose avec une régression

- Rappel de la séance sur l'**interprétation des régressions linéaires multiples**
  - Si on régresse l'activité sur la parentalité
  - Et l'indicatrice d'être une observation de chacune des femmes possibles dans la population suivie
  - Alors le **coefficient sur** `children` s'interprète comme
    - l'**écart d'activité moyenne, pour chaque femme**, entre les périodes avec et sans enfant
    - avec des **poids proportionnels à la variance de l'indicatrice d'avoir des enfants** $\rightarrow$ plus de poids pour celles qui passent la moitié du temps sans enfant, la moitié avec, et **poids nul pour celles dont le statut relativement à la maternité ne varie jamais**

Faire la même chose avec une régression

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

reg_longitudinale$coefficients["children"]
```

```
  children
-0.0522696
```

Faire la même chose avec une régression

```
all.equal(
  as.numeric(ecarts_individuels_agreges$agreg_var),
  as.numeric(reg_longitudinale$coefficients["children"]))
```

```
[1] TRUE
```

Comparer en contrôlant par le temps moyen passé avec des enfants

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

reg_mundlak <-
  lm(LFP ~ children + children_moy,
     dat = psid)

reg_mundlak$coefficients[1:2]
```

```
(Intercept)    children
  0.7376540  -0.0522696
```

Comparer en contrôlant par le temps moyen passé avec des enfants

```
all.equal(
  reg_longitudinale$coefficients["children"],
  reg_mundlak$coefficients["children"]
)
```

```
[1] TRUE
```

Pourquoi c'est la même chose

- L'objet `reg_longitudinale` fait la moyenne sur les femmes des écarts d'activité *intra* avec des poids proportionnels à la variance *intra* de `children`
- Ce poids est **le même pour toutes les femmes qui passent le même nombre d'années avec des enfants**
  - C'est donc la même chose de calculer les écarts femme par femme et d'agréger avec ces poids
  - Et de calculer l'écart à l'intérieur de la strate de toutes les femmes qui passent le même temps dans l'état $\texttt{children} == 1$
  - Se rapporte au concept de **score de propension** qui sera abordée dans la suite de vos études

Différencier d'une année sur l'autre

```
data.table::setorder(
  psid,
  ID,
  TIME
)
```

Différencier d'une année sur l'autre

```
psid[,c("ID","TIME")]
```

```
           ID  TIME
        <int> <int>
     1:     1     1
     2:     1     2
     3:     1     3
     4:     1     4
     5:     1     5
    ---
13145:  6365     5
13146:  6365     6
13147:  6365     7
13148:  6365     8
13149:  6365     9
```

Différencier d'une année sur l'autre

```
psid[,
     paste0(
       c("children",
       "LFP"),
       "_diff") :=
       lapply(X = c("children","LFP"),
              FUN = function(var)
                data.table::fcase(
                  ID == data.table::shift(ID),
                  as.numeric(
                    get(var) -
                      data.table::shift(get(var))),
                  default = NA_real_)
              )]
```

Différencier d'une année sur l'autre

```
psid[ID == 6274,
     c("TIME","children","LFP",
       "children_diff","LFP_diff")]
```

```
    TIME children   LFP children_diff LFP_diff
   <int>    <num> <int>          <num>    <num>
1:     1        0     1             NA       NA
2:     2        0     1              0        0
3:     3        0     1              0        0
4:     4        0     1              0        0
5:     5        0     1              0        0
6:     6        1     0              1       -1
7:     7        1     1              0        1
8:     8        1     1              0        0
9:     9        1     0              0       -1
```

Différencier d'une année sur l'autre

```
reg_first_diff <-
  lm(LFP_diff ~ children_diff - 1,
     dat = psid)

reg_first_diff$coefficients
```

```
children_diff
  -0.04524887
```

Différencier d'une année sur l'autre

- En fait ici ça ne revient au même que si on est dans le cas où **on ne suit la population que sur deux périodes** (et qu'on a bien enlevé la constante)

```
reg_longitudinale_2periodes <-
  lm(LFP ~ children + factor(ID),
     data = psid[TIME %in% c(1,2)])

reg_first_diff_2periodes <-
  lm(LFP_diff ~ children_diff - 1,
     data = psid[TIME %in% c(2)])
```

Différencier d'une année sur l'autre

```
all.equal(
  as.numeric(
    reg_longitudinale_2periodes$coefficients["children"]),
  as.numeric(
    reg_first_diff_2periodes$coefficients["children_diff"])
)
```

```
[1] TRUE
```

Différencier par rapport aux moyennes individuelles

- En fait on l'a déjà fait... Pourquoi ?

Différencier par rapport aux moyennes individuelles‘

- **Théorème de Frisch-Waugh-Lovell**
  - C’est la même chose de régresser `LFP` sur `children` et les indicatrices d’être une observation de chacune des femmes possibles
  - Et de régresser d’abord séparément `LFP` et `children` sur toutes ces indicatrices, puis les résidus de `LFP` sur les résidus de `children`
  - Comme on est dans le cas de la **régression saturée**, les valeurs prédites sont juste les **moyennes de** `LFP` et `children` sur la période suivie pour chaque femme
  - Et **les résidus sont les écarts à ces moyennes**

Différencier par rapport aux moyennes individuelles

```
psid[,
     paste0(
       c("children",
       "LFP"),
       "_diff_moy") :=
       lapply(X = c("children","LFP"),
              FUN = function(var)
                get(var) - mean(get(var))),
     by = c("ID")]
```

Différencier par rapport aux moyennes individuelles

```
psid[ID == 6274,
     c("TIME","children","LFP",
       "children_diff_moy","LFP_diff_moy")]
```

	TIME	children	LFP	children_diff_moy	LFP_diff_moy
	<int>	<num>	<int>	<num>	<num>
1:	1	0	1	-0.4444444	0.2222222
2:	2	0	1	-0.4444444	0.2222222
3:	3	0	1	-0.4444444	0.2222222
4:	4	0	1	-0.4444444	0.2222222
5:	5	0	1	-0.4444444	0.2222222
6:	6	1	0	0.5555556	-0.7777778
7:	7	1	1	0.5555556	0.2222222
8:	8	1	1	0.5555556	0.2222222
9:	9	1	0	0.5555556	-0.7777778

Différencier par rapport aux moyennes individuelles

```
reg_within <-
  lm(LFP_diff_moy ~children_diff_moy,
      data = psid)
reg_within$coefficients
```

```
      (Intercept) children_diff_moy
     1.044355e-17     -5.226960e-02
```

DIfférencier par rapport aux moyennes individuelles

```
all.equal(
  as.numeric(reg_within$coefficients["children_diff_moy"]),
  as.numeric(reg_longitudinale$coefficients["children"])
)
```

```
[1] TRUE
```

Un peu de vocabulaire et de formalisation

- On a vu que toutes ces **comparaisons longitudinales** se ramenaient à une **régression de** `LFP` sur `children` et toutes les indicatrices d'être une observation relative à chacune des femmes possibles
- En général on fait le choix de noter
  $\texttt{LFP}_{it} = \beta \texttt{children}_{it} + \lambda_i + \nu_{it}$
  - On ne fait **aucune hypothèse particulière sur** λ_i : ce sont juste les coefficients sur les indicatrices
  - Tant que l'on n'a pas de **problèmes de colinéarité**, on peut toujours construire β, les λ_i et ν_{it} qui vérifient :
    - $\mathbb{E}[\texttt{children}_{it}\nu_{is}] = 0$ pour toutes les périodes t et s
    - $\mathbb{E}[C_i\nu_{it}] = 0$ pour toutes les périodes i et tous les individus i, où C_i représente l'indicatrice de l'individu i
    - C'est juste répéter ce que l'on a dit depuis le début

Un peu de vocabulaire et de formalisation

- On dit que les λ_i sont des **effets fixes individuels**
- Le reste ne change pas
  - β est le **coefficient de la régression** de `LFP` sur `chidren` avec des effets fixes individuels
  - ν_{it} est le **résidu / terme d'erreur / choc idiosyncratique**
    - Je continuerai toujours à dire résidu pour les mêmes raisons qu'auparavant
- La dernière technique est dite **transformation *within***
  - C'est en pratique ce que l'on fait le plus souvent
- L'avant-dernière est appelée **premières différences**

Quelques remarques sur l'interprétation causale

- **Pas de raison *a priori*** de dire que les comparaisons que l'on fait maintenant, i.e. les comparaisons longitudinales permises par l'inclusion d'effets fixes, ont une **interprétation causale** plus facile ou plus fréquemment plausible que les comparaisons en coupe

Quelques remarques sur l'interprétation causale

- L'**interprétation causale de la régression en coupe** suppose qu'une année donnée, la **comparaison entre les femmes avec des enfants et celles sans enfants** est aussi bonne que si le fait d'avoir des enfants était le produit d'une **expérience aléatoire avec une probabilité de traitement qui est la même pour chaque femme**

Quelques remarques sur l'interprétation causale

- L'**interprétation causale de la régression avec effets fixes individuels** suppose que pour chaque femme, la **comparaison entre les moments où elle a et n'a pas d'enfants** est aussi bonne que si ces périodes étaeient déterminées par **une expérience aléatoire avec une probabilité de traitement variable d'une femme à une autre**

Quelques remarques sur l'interprétation causale

- Ce sont des **hypothèses différentes**
- Quand on les discute il faut le faire sérieusement
- Ne pas penser que passer de comparaisons transversales à des comparaisons longitudinales règle définitivement la question
  - Il reste toujours l'idée qu'il y a quelque part un **choc aléatoire** !

Une conclusion provisoire

- On a vu que **les outils de régression peuvent être facilement adaptés pour faire des comparaisons** longitudinales lorsque l'on dispose de données adaptées
- Pour l'instant, on n'a discuté que de l'**inteprétation des coefficients obtenus** sans se soucier une fois encore de l'estimation en tant que telle et du fait qu'on travaille sur un **échantillon**
  - On y reviendra la prochaine fois
  - Il faut un peu adapter mais ce n'est pas pire que ce qu'on a déjà vu ensemble