Modélisation et identification causale
Séance 5 – Utiliser une variable instrumentale

Pierre Pora

2022-12-15

Objet de la séance

Retour sur les séances précédentes

- L'idée maîtresse des deux dernières séances :
  - En se plaçant dans des groupes suffisamment fins, on peut assimiler la situation empirique que l'on regarde à une **expérience aléatoire contrôlée** ou à une **expérience naturelle**
  - Il faut faire exclusivement des comparaisons **à l'intérieur de ces groupes**
  - Quand les groupes sont très fins et que l'on travaille sur un échantillon de taille finie cela peut devenir difficile mais on a des moyens de s'en sortir

L'objectif de cette séance

- Cette fois-ci on prend pour acquis le fait que l'on est dans une situation expérimentale ou quasi-expérimentale
- Mais la variable qui fait l'objet de cette variation quasi-aléatoire **n'est pas celle qui nous intéresse *a priori***
- On va montrer que, dans certaines conditions, ce n'est pas forcément si grave que ça
- Résultat pas vraiment intuitif à première vue que l'on va prendre le temps de décortiquer

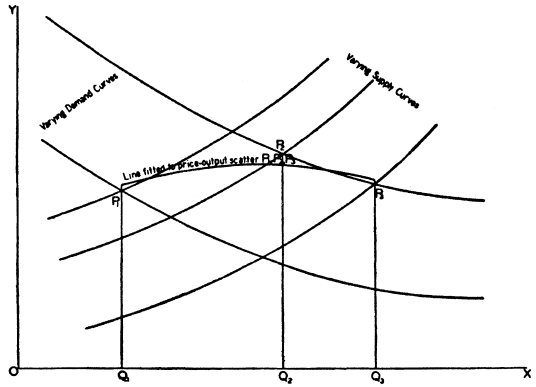
Aux origines de la méthode des variables instrumentales

La seule technique d'inférence causale inventée par un économiste

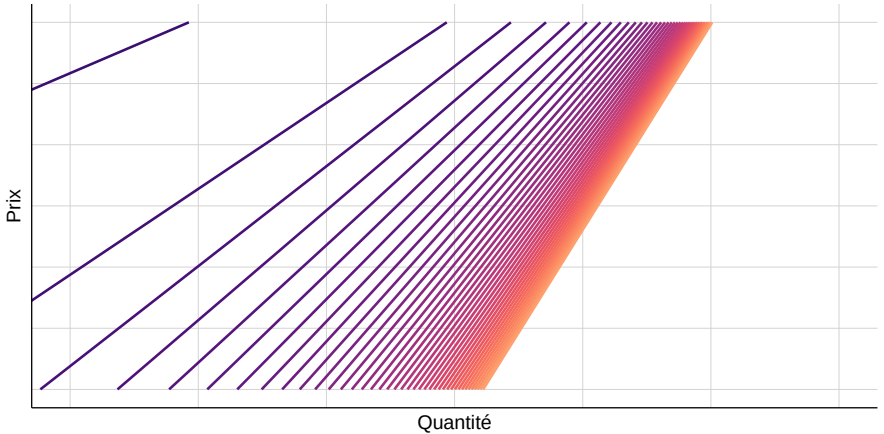
- Expérience aléatoire contrôlée $\longrightarrow$ statistique agricole puis sciences médicales
- Approches graphiques de la causalité $\longrightarrow$ informatique
- Score de propension $\longrightarrow$ sciences médicales
- Discontinuité $\longrightarrow$ psychologie
- Différence-de-différences $\longrightarrow$ épidémiologie

"If both supply and demand conditions change, price-output data yield no direct information as to either curve" (Wright (1928))

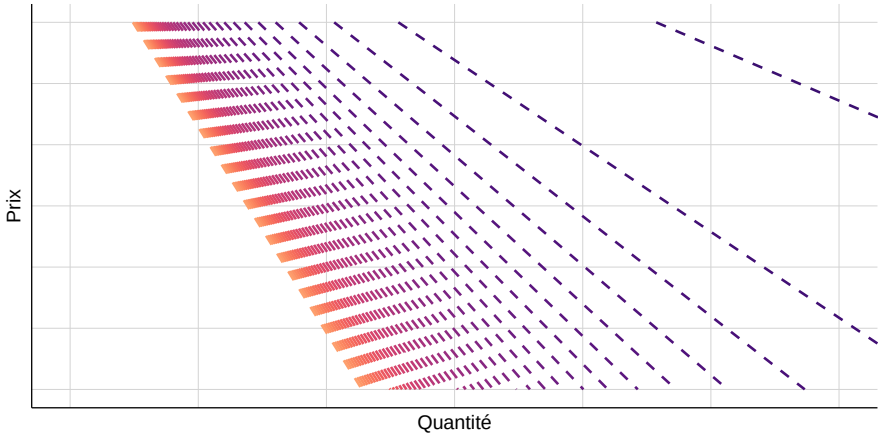
FIGURE 4. PRICE-OUTPUT DATA FAIL TO REVEAL EITHER SUPPLY OR DEMAND CURVE.



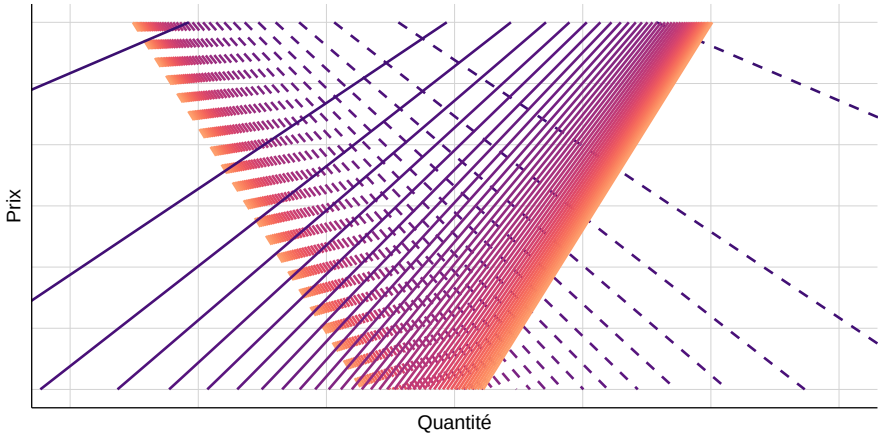
Le raisonnement de Wright : les courbes d'offre (croissantes)



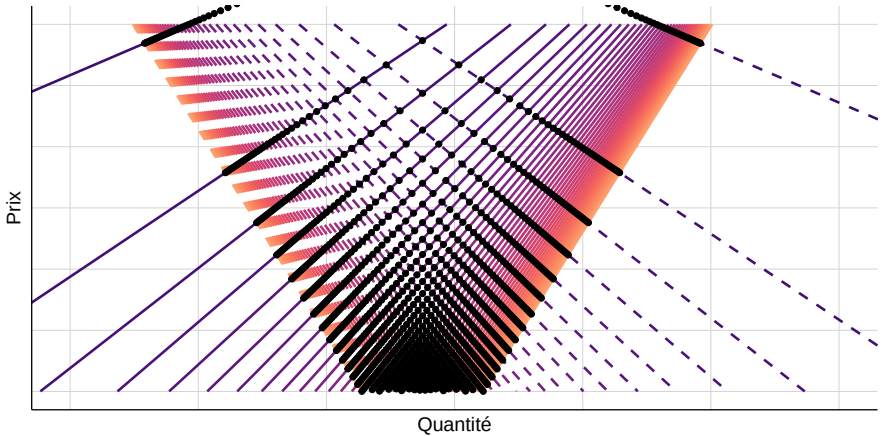
Le raisonnement de Wright : les courbes de demande (décroissantes)



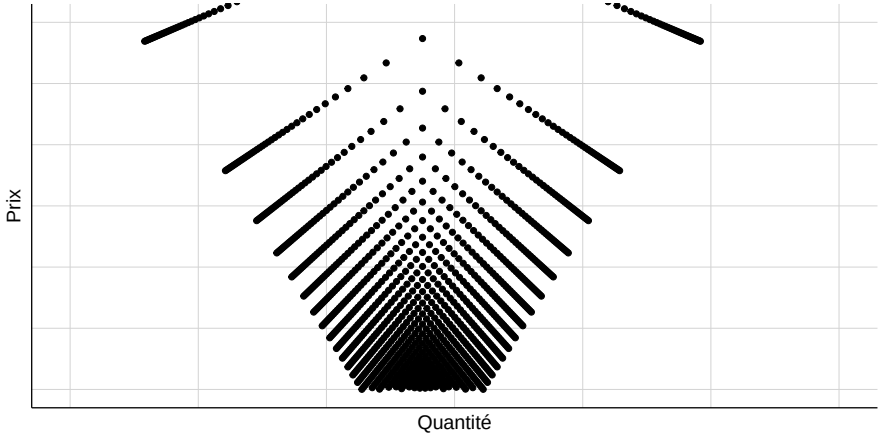
Le raisonnement de Wright : les prix et quantités d'équilibre



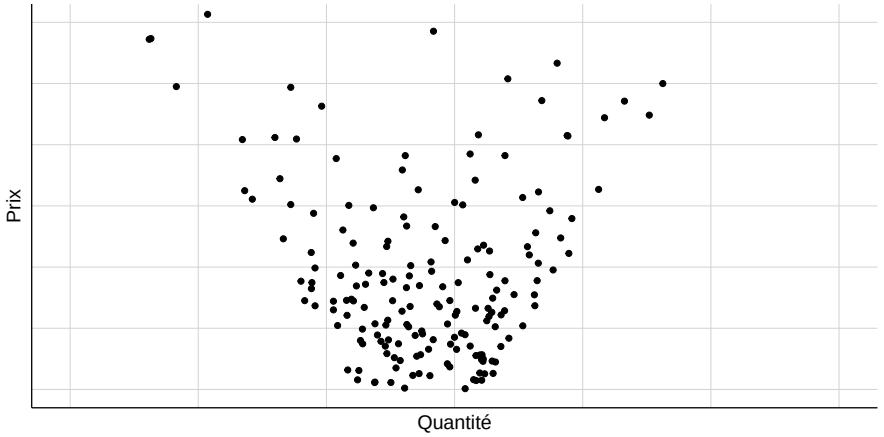
Le raisonnement de Wright : les prix et quantités d'équilibre



Les courbes d'offre et de demande ne sont jamais observées en réalité



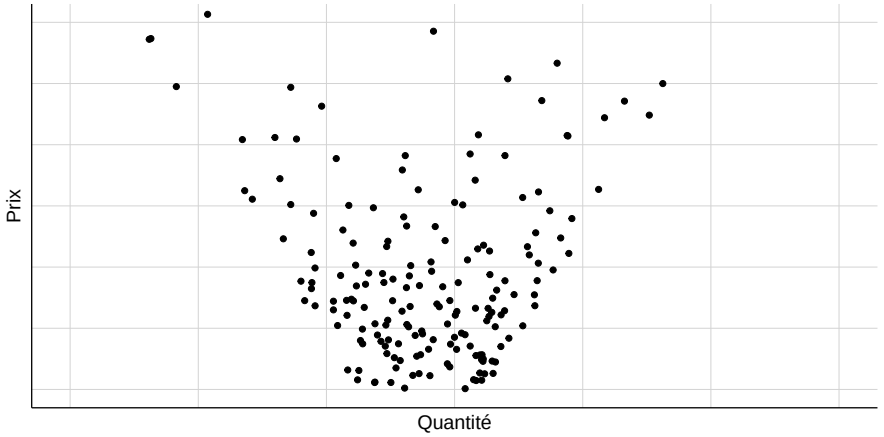
Et l'on n'observe qu'un échantillon des équilibres possibles



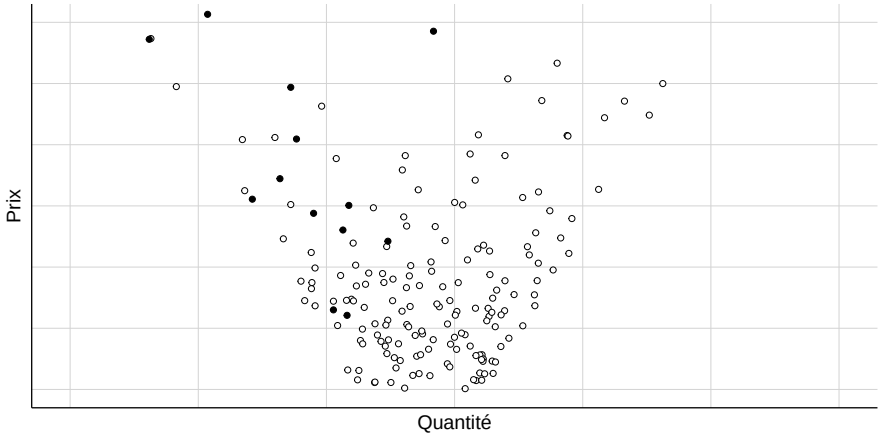
Peut-on réussir à retrouver de l'information sur les courbes de demande à partir de ce nuage de points ?

- La réponse de Wright : oui !
- Wright s'intéresse à de la production agricole : l'on peut distinguer des années au cours desquelles les rendements agricoles ont été particulièrement mauvais, par exemple du fait des conditions météorologiques
  - **Cela affecte l'offre mais (*a priori*) pas la demande**

Distinguer les années de mauvaise récolte



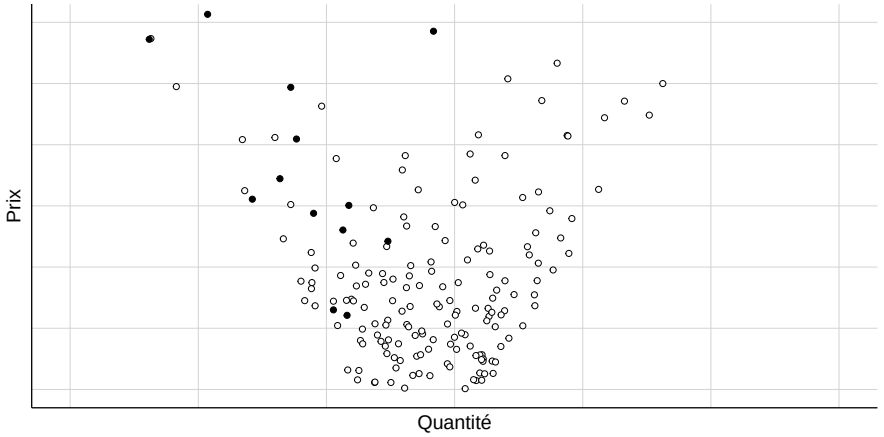
Distinguer les années de mauvaise récolte



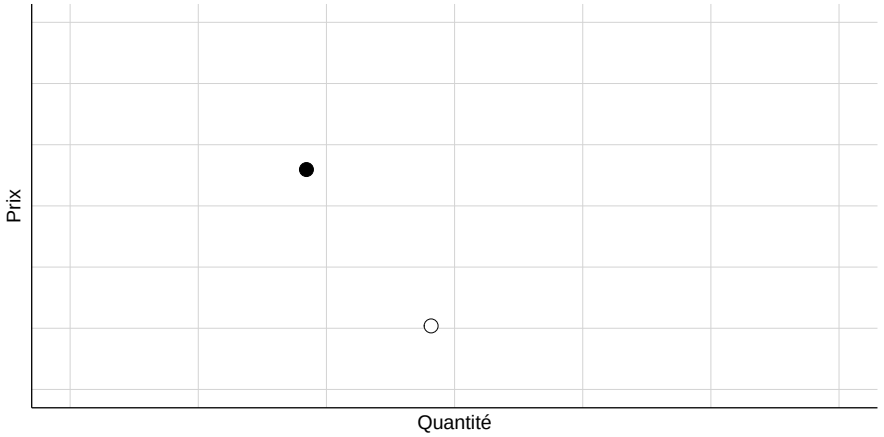
Et alors ?

- On sait que ces mauvaises conditions météorologiques n'affectent l'offre que parce qu'elles diminuent les rendements agricoles
- **Elles ne changent donc rien à la demande**
- En moyenne les points qui correspondent à ces mauvaises conditions sont sur les mêmes courbes de demande que les autres !
- **La comparaison entre les deux groupes donne de l'information sur la pente des courbes de demande**

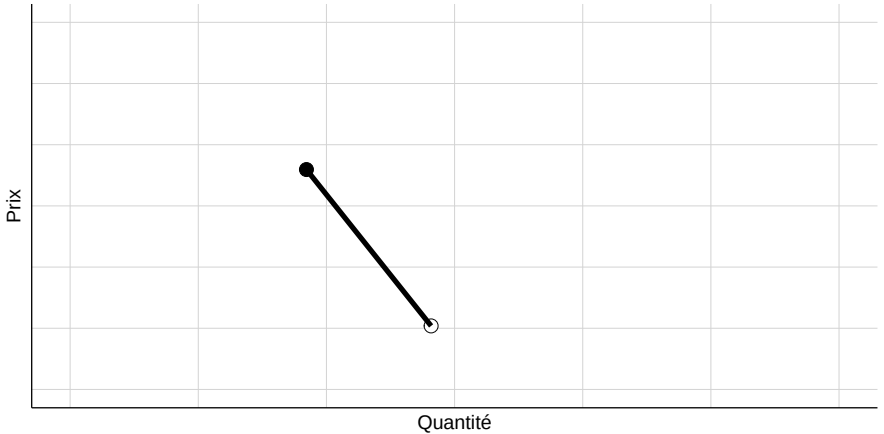
Comparer les deux groupes



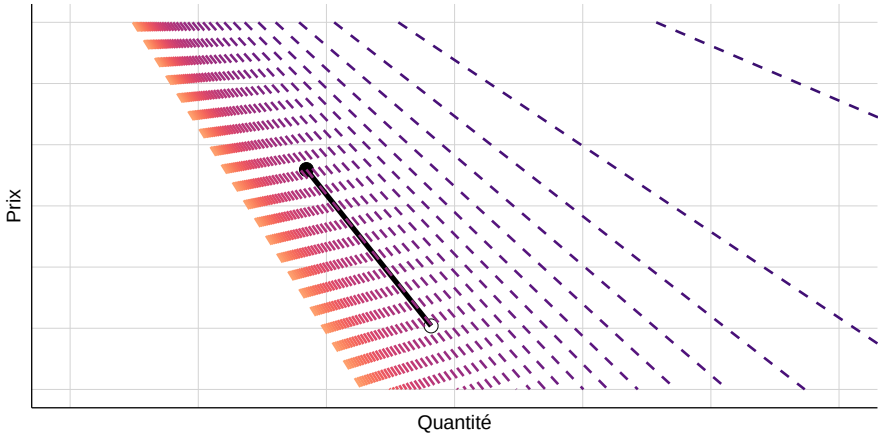
Comparer les deux groupes



Approximer la pente des courbes de demande



Approximer la pente des courbes de demande



Un peu de vocabulaire

- On dirait ici que l'on a **instrumenté** les quantités échangées par des chocs météorologiques qui ne peuvent changer les prix que parce que l'on change l'offre
  - On parle de **restriction d'exclusion**
- Comme on ne peut que changer d'offre, on est toujours sur la (les) même(s) courbe(s) de demande, et donc la variation des prix avec la quantité échangée renseigne sur la (les) courbe(s) de demande

Comparaison avec la stratégie des séances précédentes

- On fait ici quelque chose de **très différent** de la stratégie de conditionnement
- Si on conditionnait sur le fait d'être ou pas une année à mauvaise récolte, alors on comparerait entre elles :
  - les années à mauvaise récolte
  - les années à bonne récolte
  - mais jamais entre elles les années à bonne et mauvaise récolte

Comparaison avec la stratégie des séances précédentes

- On fait ici le contraire ! On compare entre elles
  - les années à bon et mauvaise récolte
  - mais jamais entre elles les années à bonne récolte ou à mauvaise récolte
- **Variable de conditionnement** $\longrightarrow$ la variabilité intéressante est *intra*
- **Variable instrumentale** $\longrightarrow$ la variabilité intéressante est *inter*

Un exemple empirique : la conscription aux États-Unis pendant la guerre du Viêtnam

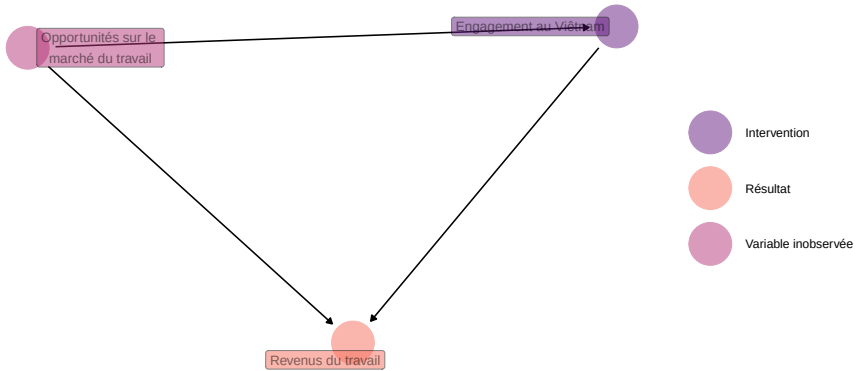
L'objet de cet exemple

- Répliquer un exemple historiquement important : Angrist (1990) sur les effets de l'engagement dans la guerre du Viêtnam sur les revenus des vétérans
  - Question pratique et politiquement sensible de l'indemnisation des anciens combattants
  - Travail qui a ensuite motivé les travaux joints d'Angrist et Imbens (Imbens and Angrist (1994)) $\longrightarrow$ la façon contemporaine d'aborder la méthode des variables instrumentales
- L'objectif :
  1. Retrouver à partir des données les résultats présentés dans Angrist (1990)
  2. Retracer le raisonnement explicité dans Imbens and Angrist (1994) à partir de cet exemple

Pourquoi ne pas simplement comparer les vétérans et les autres ?

- **Comparaison difficile** : une partie des vétérans sont des engagés volontaires
  - Si cet engagement se fait en considérant les opportunités que l'on a sur le marché du travail à l'âge de l'engagement
  - Alors il est probable que ces engagés volontaires aient eu des opportunités moins intéressantes que celles de ceux à qui on les compare
  - Et que cela soit encore le cas longtemps après la guerre
- Le contraste entre les vétérans et les autres n'a pas d'interprétation causale

Pourquoi ne pas simplement comparer les vétérans et les autres ?



Comment se passe la conscription ?

Une petite vidéo

Pour commencer

- Chargez les données du fichier `earnings.csv`
- Pour chaque cellule, les variables `real_earnings` et `nominal_earnings` représentent les revenus du travail moyen réels et nominaux, et `nonzero_earnings` la part d'hommes dont les revenus du travail sont strictement positif. Peut-on en déduire les revenus du travail moyens des hommes dont les revenus du travail sont strictement positifs ?

Pour commencer

```
#1. Télécharger les données sur les revenus

earnings_dat<-fread("./Data/Angrist1990/earnings.csv")

#2. Estimer les revenus du travail moyens des individus ayant un revenu
#du travail strictement positif

earnings_dat[,
             c("positive_real_earnings",
               "positive_nominal_earnings"):=
               list(real_earnings*sample_size/nonzero_earnings,
                    nominal_earnings*sample_size/nonzero_earnings)]
```

Vérification : la variable `positive_nominal_earning` permet de bien reproduire la table A.1

```
##    year race mean_earnings
## 1:   69    1          1473
## 2:   70    1          1977
## 3:   71    1          2581
## 4:   72    1          3614
## 5:   73    1          4738
## 6:   74    1          5726
```

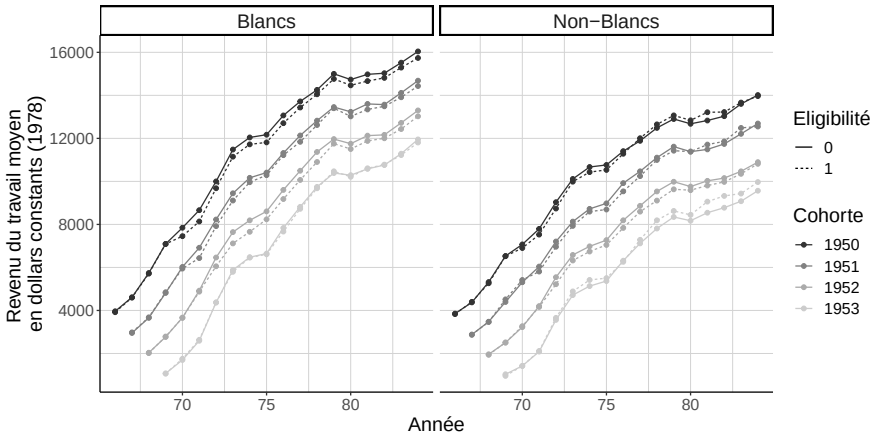
Comparer les appelés et les autres

- La variable `eligible` est une variable dichotomique qui indique si les hommes de la cellule – qui regroupe 5 numéros de loterie consécutifs – ont finalement été appelés ou pas.
- Pour chaque cohorte (`birthyear`) et chaque groupe défini par l'appel, tracez l'évolution au cours du temps (variable `year`) de la part imposable (`type=="TAXAB"`) des revenus du travail moyens des hommes dont ces revenus sont strictement positifs, séparément pour les Blancs et les non-Blancs
  - Reproduit la figure 1 de l'article (mal-numérotée, apparaît en 3)

Comparer les appelés et les autres

```
earnings_traj<-earnings_dat[,
                            list(real_earnings=
                                   sum(positive_real_earnings*
                                         nonzero_earnings)/
                                   sum(nonzero_earnings),
                                 nominal_earnings=
                                   sum(positive_nominal_earnings*
                                         nonzero_earnings)/
                                   sum(nonzero_earnings)),
                            c("birthyear",
                              "year",
                              "race",
                              "eligible",
                              "type")]
```

Représenter les trajectoire de revenus du travail



Comparer les appelés et les autres (encore !)

- Pour chaque cohorte, estimez l'évolution dans le temps de la différence entre le revenu du travail moyen des appelés et celui des autres hommes, séparément pour les Blancs et les non-Blancs, en vous restreignant aux hommes dont ces revenus sont strictement positifs
- Cette quantité a-t-elle une interprétation causale ?
- Quel est l'intérêt d'estimer cette quantité *avant* la conscription ?

Comparer les appelés et les autres (encore !)

```
earnings_traj_wide<-dcast(earnings_traj,
                          birthyear + year + race + type ~ eligible,
                          value.var=c("real_earnings",
                                      "nominal_earnings"),
                          fill=NA_real_,
                          fun.aggregate = I)[,
                                             c("real_diff",
                                               "nominal_diff"):=
                                               list(real_earnings_1-
                                                      real_earnings_0,
                                                    nominal_earnings_1-
                                                      nominal_earnings_0)]
```

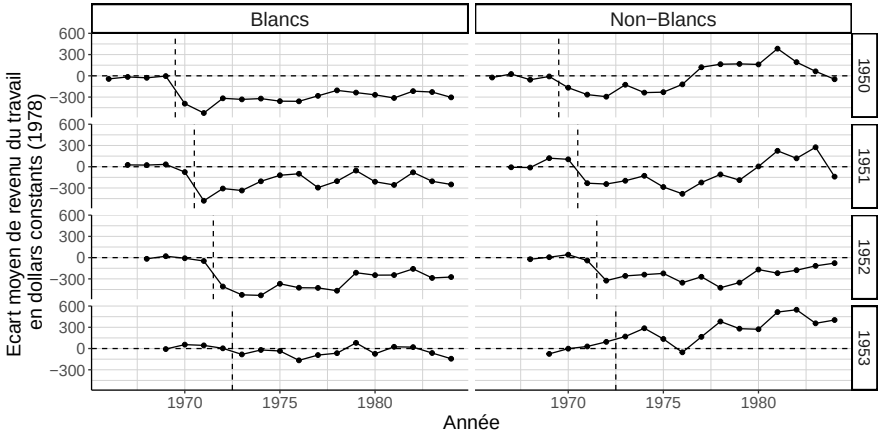
On retrouve la table 1 du papier

```
table1<-setorder(earnings_traj_wide[type=="TAXAB",
                                    c("birthyear",
                                      "race",
                                      "year",
                                      "nominal_diff")],
                 race,
                 birthyear,
                 year)
```

On retrouve la table 1 du papier

```
##       birthyear race year nominal_diff
##    1:        50    1   66 -21.8095....
##    2:        50    1   67 -8.01996....
##    3:        50    1   68 -14.9039....
##    4:        50    1   69 -2.09808....
##    5:        50    1   70 -233.867....
##   ---
## 148:        53    2   80 344.0962....
## 149:        53    2   81 717.8528....
## 150:        53    2   82 810.2171....
## 151:        53    2   83 543.6810....
## 152:        53    2   84 641.3994....
```

Pour une visualisation plus agréable (figure 2 du papier)



Interprétation causale de la **forme réduite**

- L'assignation au groupe des appelés est faite sur la base d'une loterie
  - On est dans le cas d'une **expérience naturelle** !
- La différence entre les appelés et les autres est égale aux effets causaux moyens de l'appel sur les revenus du travail plusieurs années après
- Evaluer cette différence avant l'appel est un test *placebo* qui permet de vérifier que les deux groupes sont bien équilibrés
  - Trouver des différences avant que la loterie ait effectivement lieu aurait été très inquiétant !

Tous les appelés sont-ils des vétérans ?

- Chargez les données de la table `sipp84.csv`
- Pour chaque cellule cohorte (`birthyear`) × catégorie ethno-raciale (`race`) × niveau scolaire (`education`) × avoir été appelé (`eligible`) × être un vétéran, la table renseigne sur les effectifs (`fnlwgt_5`)
- Estimez l'écart entre la part des vétérans parmi les appelés et la part des vétérans parmi les non-appelés pour chaque cohorte et séparément pour les Blancs et les non-Blancs
  - Lissez en faisant la moyenne sur 3 cohortes consécutives
- Ces quantités ont-elles une interprétation causale ?
- Que remarquez vous ?

Tous les appelés sont-ils des vétérans ?

```r
#On charge les données du fichier sipp84.csv
sipp<-fread("./Data/Angrist1990/sipp84.csv")

head(sipp)
```

```
##    V1 birthyear race eligible nvstat educ fnlwgt_5
## 1:  1      1947    1        0      0    8 42521850
## 2:  2      1954    1        0      0   12 40123011
## 3:  3      1951    1        0      0   18 42271558
## 4:  4      1953    2        0      1   14 73927797
## 5:  5      1951    1        0      0   16 42615699
## 6:  6      1943    1        0      0   16 39870633
```

Tous les appelés sont-ils des vétérans ?

```
#Cette fonction permet de calculer la part de vétérans dans chaque cohorte
# en lissant sur trois cohorts consécutives
veteran_probability_smoothed<-function(cohort){

  sipp[birthyear>=cohort-1
       & birthyear<=cohort+1,
       list(vet_prob=sum(nvstat*fnlwgt_5, na.rm=TRUE)/
              sum(fnlwgt_5, na.rm=TRUE)),
       by=c("race",
            "eligible")][,
                         birthyear:=cohort]

}
```

Tous les appelés sont-ils des vétérans ?

```
veteran_probability<-rbindlist(lapply(1950:1952,
                                      veteran_probability_smoothed))

veteran_probability[,
                    eligibility:=fcase(eligible==1, "eligible",
                                       eligible!=1, "noneligible")]
```

Tous les appelés sont-ils des vétérans ?

```
veteran_probability_wide<-
  dcast(veteran_probability,
        birthyear + race ~ eligibility,
        value.var = "vet_prob",
        fun.aggregate=I,
        fill=NA_real_)[,
                       probability_gap:=
                         round(eligible-noneligible,4)]
setorder(veteran_probability_wide, race, birthyear)
```

Tous les appelés sont-ils des vétérans ? Réplication de la table 2

```
##    birthyear race     eligible  noneligible probability_gap
## 1:      1950    1 0.352715.... 0.193355....          0.1594
## 2:      1951    1 0.283092.... 0.146897....          0.1362
## 3:      1952    1 0.231021.... 0.125739....          0.1053
## 4:      1950    2 0.195748.... 0.135472....          0.0603
## 5:      1951    2 0.201385.... 0.151400....            0.05
## 6:      1952    2 0.144885.... 0.128775....          0.0161
```

Interprétation causale de la **première étape**

- L'assignation à l'appel est fondé sur une loterie : c'est une **expérience naturelle** !
- La différence entre la part des vétérans parmi les appelés et la part des vétérans parmi les non-appelés est égale aux effets causaux moyens de l'appel sur la participation effective à la guerre du Viêtnam
- Cet effet n'est pas égal à 1
  - Il y a des appelés qui n'ont pas participé à la guerre
  - Et des non-appelés qui se sont engagés

Théoriser à partir de l'exemple

- Repartez du formalisme des valeurs potentielles à la Neyman-Rubin, considérant que l'intervention que l'on étudie est la loterie qui distingue les hommes appelés des autres
- Combien de groupes les valeurs potentielles de la participation effective à la guerre ?
- Y a-t-il un groupe dont on peut supposer qu'il est vide ?
- Exprimez les effets causaux moyens de l'appel sur la participation au conflit en fonction de la part des différents groupes

Revenir aux valeurs potentielles à la Neyman-Rubin

- L'intervention que l'on étudie : Z_i le fait d'être appelé
  - 0 si on n'est pas tiré à la loterie
  - 1 si on est tiré à la loterie
- La variable d'intérêt : D_i le fait de participer effectivement au conflit
  - 0 si on ne part pas au Viêtnam
  - 1 si on part effectivement au Viêtnam
- Les valeurs potentielles $D_i(z)$ avec $z \in \{0,1\}$: les réactions contrefactuelles (inobservées) au fait d'être tiré au sort

Revenir aux valeurs potentielles à la Neyman-Rubin : un petit tableau

Valeurs potentielles	$D_i(0) = 0$	$D_i(0) = 1$
$D_i(1) = 0$		
$D_i(1) = 1$		

Revenir aux valeurs potentielles à la Neyman-Rubin : un petit tableau

Valeurs potentielles	$D_i(0) = 0$	$D_i(0) = 1$
$D_i(1) = 0$	Ne participent jamais (*never takers*)	Participent si pas appelés, ne participent pas si appelé (*defiers*)
$D_i(1) = 1$	Ne participent pas si pas appelé, participent si appelés (*compliers*)	Participent toujours (*always takers*)

Hypothèse de monotonie

- La définition des *defiers*
  - $D_i(0) = 1$ ils participent s'ils ne sont pas tirés au sort
  - $D_i(1) = 0$ ils ne participent pas s'ils sont tirés au sort
- Raisonnablement on peut supposer que ce groupe est vide ou de taille négligeable
- **Hypothèse de monotonie** : être tiré au sort influence tout le monde *dans le même sens*
  - c'est-à-dire ou bien pas du tout
  - ou bien en faisant participer à la guerre

Que quantifient les effets causaux moyens de l'appel sur la participation ?

- Il faut utiliser la loi des espérances itérées
- Ce qu'elle dit : les effets causaux moyens de l'appel pour tous les hommes américains = la moyenne des effets causaux moyens dans chaque groupe × la part de ce groupe dans la population
- Quels sont les effets causaux moyens de l'appel sur la participation chez les *never takers* ? chez les *always takers* ? chez les *compliers* ?

Que quantifient les effets causaux moyens de l'appel sur la participation ? Un peu de formalisme

$$
\begin{aligned}
&\mathbb{E}[D_i(1) - D_i(0)] \\
=&\underbrace{\mathbb{P}(D_i(1) = D_i(0) = 0)}_{\text{Part de } \textit{never takers}}\underbrace{\mathbb{E}[D_i(1) - D_i(0) \mid D_i(1) = D_i(0) = 0]}_{\text{Effet chez les } \textit{never takers}} \\
&+\underbrace{\mathbb{P}(D_i(1) = 0, D_i(0) = 1)}_{\text{Part de } \textit{defiers}}\underbrace{\mathbb{E}[D_i(1) - D_i(0) \mid D_i(1) = 0, D_i(0) = 1]}_{\text{Effet chez les } \textit{defiers}} \\
&+\underbrace{\mathbb{P}(D_i(1) = 1, D_i(0) = 0)}_{\text{Part de } \textit{compliers}}\underbrace{\mathbb{E}[D_i(1) - D_i(0) \mid D_i(1) = 1, D_i(0) = 0]}_{\text{Effet chez les } \textit{compliers}} \\
&+\underbrace{\mathbb{P}(D_i(1) = D_i(0) = 1)}_{\text{Part de } \textit{always takers}}\underbrace{\mathbb{E}[D_i(1) - D_i(0) \mid D_i(1) = D_i(0) = 1]}_{\text{Effet chez les } \textit{always takers}}
\end{aligned}
$$

Que quantifient les effets causaux moyens de l'appel sur la participation ? On simplifie

$$
\begin{aligned}
&\mathbb{E}[D_i(1) - D_i(0)] \\
&= \underbrace{\mathbb{P}(D_i(1) = D_i(0) = 0)}_{\text{Part de } \textit{never takers}} \underbrace{\mathbb{E}[D_i(1) - D_i(0) \mid D_i(1) = D_i(0) = 0]}_{0} \\
&\quad + \underbrace{\mathbb{P}(D_i(1) = 0, D_i(0) = 1)}_{\text{Part de } \textit{defiers}} \underbrace{\mathbb{E}[D_i(1) - D_i(0) \mid D_i(1) = 0, D_i(0) = 1]}_{-1} \\
&\quad + \underbrace{\mathbb{P}(D_i(1) = 1, D_i(0) = 0)}_{\text{Part de } \textit{compliers}} \underbrace{\mathbb{E}[D_i(1) - D_i(0) \mid D_i(1) = 1, D_i(0) = 0]}_{1} \\
&\quad + \underbrace{\mathbb{P}(D_i(1) = D_i(0) = 1)}_{\text{Part de } \textit{always takers}} \underbrace{\mathbb{E}[D_i(1) - D_i(0) \mid D_i(1) = D_i(0) = 1]}_{0}
\end{aligned}
$$

Les effets causaux moyens de l'appel sur la participation sont égaux à la part de *compliers*

$$\begin{aligned}&\mathbb{E}[D_i(1) - D_i(0)] \\ =&\underbrace{\mathbb{P}(D_i(1) = 1, D_i(0) = 0)}_{\text{Part de } \textit{compliers}}\end{aligned}$$

Un dernier effort avant de conclure

- Exprimez les effets causaux moyens de l'appel sur les revenus en fonction des parts des différents groupes
- Quelle hypothèse peut-on faire sur les effets causaux moyens de l'appel sur les revenus du travail des *never takers* ?
- Quelle hypothèse peut-on faire sur les effets causaux moyens de l'appel sur les revenus du travail des *always takers*
- Quelle hypothèse peut-on faire sur les effets causaux moyens de l'appel sur les revenus du travail des *compliers*

Décomposer les effets moyens de l'appel sur les revenus du travail

- Il faut utiliser la loi des espérances itérées
- Ce qu'elle dit : les effets causaux moyens de l'appel pour tous les hommes américains = la moyenne des effets causaux moyens dans chaque groupe $\times$ la part de ce groupe dans la population

Restriction d'exclusion

- On va introduire une hypothèse supplémentaire par rapport à ce qu'on a déjà fait :
  - Le fait d'avoir été appelé ou non **n'a pas d'effet direct** sur les revenus du travail
- Dit autrement : le fait d'être appelé conduit à un certain nombre d'actions
- Ces actions ne changent les revenus futurs **que parce qu'elles changent le fait de participer ou pas à la guerre**

Conséquences de la restriction d'exclusion

- Que peut-on en déduire sur les effets causaux moyens de l'appel sur les revenus futurs des *never takers* ? des *always takers* ?
- Que peut-on en déduire sur les effets causaux moyens de l'appel sur les revenus futurs des *compliers* ?

Conséquences de la restriction d'exclusion

- Que peut-on en déduire sur les effets causaux moyens de l'appel sur les revenus futurs des *never takers* ? des *always takers* ?
  - Dans ces groupes l'appel ne change rien à la participation
  - Donc les effets causaux de l'appel sur les revenus futurs sont nuls

Conséquences de la restriction d'exclusion

- Que peut-on en déduire sur les effets causaux moyens de l'appel sur les revenus futurs des *compliers* ?
  - La seule chose qui change :
    - un *complier* appelé participe toujours
    - un *complier* pas appelé ne participe jamais
  - Les effets causaux moyens de l'appel sur les revenus futurs des *compliers* sont égaux aux effets causaux moyens de la participation sur les revenus futurs des *compliers*

On récapitule : un peu de formalisme (de nouveau)

$$
\begin{aligned}
&\mathbb{E}[Y_i(D_i(1)) - Y_i(D_i(0))] \\
=&\underbrace{\mathbb{P}(D_i(1) = D_i(0) = 0)}_{\text{Part de } \textit{never takers}}\underbrace{\mathbb{E}[Y_i(D_i(1)) - Y_i(D_i(0)) \mid D_i(1) = D_i(0) = 0]}_{\text{Effet chez les } \textit{never takers}} \\
&+\underbrace{\mathbb{P}(D_i(1) = 0, D_i(0) = 1)}_{\text{Part de } \textit{defiers}}\underbrace{\mathbb{E}[Y_i(D_i(1)) - Y_i(D_i(0)) \mid D_i(1) = 0, D_i(0) = 1]}_{\text{Effet chez les } \textit{defiers}} \\
&+\underbrace{\mathbb{P}(D_i(1) = 1, D_i(0) = 0)}_{\text{Part de } \textit{compliers}}\underbrace{\mathbb{E}[Y_i(D_i(1)) - Y_i(D_i(0)) \mid D_i(1) = 1, D_i(0) = 0]}_{\text{Effet chez les } \textit{compliers}} \\
&+\underbrace{\mathbb{P}(D_i(1) = D_i(0) = 1)}_{\text{Part de } \textit{always takers}}\underbrace{\mathbb{E}[Y_i(D_i(1)) - Y_i(D_i(0)) \mid D_i(1) = D_i(0) = 1]}_{\text{Effet chez les } \textit{always takers}}
\end{aligned}
$$

On récapitule : on simplifie

$$\mathbb{E}[Y_i(D_i(1)) - Y_i(D_i(0))]$$

$$= \underbrace{\mathbb{P}(D_i(1) = D_i(0) = 0)}_{\text{Part de } \textit{never takers}} \underbrace{\mathbb{E}[Y_i(D_i(1)) - \cancel{Y_i(D_i(0))} \mid D_i(1) = D_i(0) = 0]}_{0}$$

$$+ \underbrace{\cancel{\mathbb{P}(D_i(1) = 0, D_i(0) = 1)}}_{\text{Part de } \textit{defiers}} \underbrace{\mathbb{E}[Y_i(D_i(1)) - Y_i(D_i(0)) \mid D_i(1) = 0, D_i(0) = 1]}_{\text{Effet chez les } \textit{defiers}}$$

$$+ \underbrace{\mathbb{P}(D_i(1) = 1, D_i(0) = 0)}_{\text{Part de } \textit{compliers}} \underbrace{\mathbb{E}[Y_i(D_i(1)) - Y_i(D_i(0)) \mid D_i(1) = 1, D_i(0) = 0]}_{\text{Effet chez les } \textit{compliers}}$$

$$+ \underbrace{\mathbb{P}(D_i(1) = D_i(0) = 1)}_{\text{Part de } \textit{always takers}} \underbrace{\mathbb{E}[Y_i(D_i(1)) - \cancel{Y_i(D_i(0))} \mid D_i(1) = D_i(0) = 1]}_{0}$$

On récapitule : on conclut

$$\begin{aligned}
&\mathbb{E}[Y_i(D_i(1)) - Y_i(D_i(0))] \\
&= \underbrace{\mathbb{P}(D_i(1) = 1, D_i(0) = 0)}_{\text{Part de } \textit{compliers}} \underbrace{\mathbb{E}[Y_i(1) - Y_i(0) \mid D_i(1) = 1, D_i(0) = 0]}_{\text{Effet chez les } \textit{compliers}}
\end{aligned}$$

Pour conclure

- Les données dont on dispose renseignent-elles sur les effets causaux moyens de l'appel sur les revenus futurs ?
- Les données dont on dispose renseignent-elles sur la part de *compliers* ?
- Que peut-on en conclure ?
- Estimez les effets causaux moyens de la participation à la guerre du Viêtnam sur les revenus futurs des *compliers*

Quelles quantités connaît-on déjà ?

- Effets causaux moyens de l'appel sur les revenus futurs
  - L'appel résulte d'une loterie $\longrightarrow$ expérience naturelle
  - Il suffit de faire la différence entre les appelés et les autres
  - C'est ce que l'on a fait en premier !
- Part de *compliers*
  - C'est la même chose que les effets causaux moyens de l'appel sur la participation
  - On l'a déjà estimée !
- **Les effets causaux moyens de la participation sur les revenus futurs des *compliers* s'estime simplement comme le quotient des deux termes**

Estimateur de Wald

```
Wald_estimate<-
  merge(earnings_traj_wide[birthyear>=50
                           & birthyear<=52
                           & type=="ADJ"],
        veteran_probability_wide[,
                                 birthyear:=birthyear-1900],
        by=c("birthyear",
             "race"),
        all.x=TRUE,
        all.y=TRUE)[,
                    Wald_estimate_real:=
                      (real_earnings_1-real_earnings_0)/probability_gap]
```

Estimation des effets causaux moyens de la participation à la guerre du Viêtnam sur les revenus du travail des *compliers* (table 3 du papier)

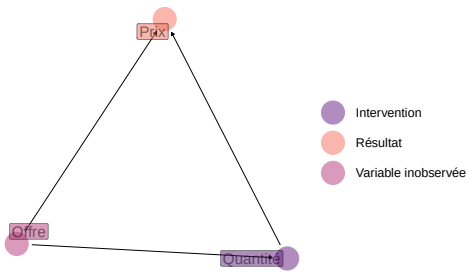
```
##     birthyear year Wald_estimate_real
##  1:        50   81       -2195.33....
##  2:        50   82       -1679.02....
##  3:        50   83       -1849.28....
##  4:        50   84       -2517.11....
##  5:        51   81       -2258.31....
##  6:        51   82       -1382.06....
##  7:        51   83       -2174.36....
##  8:        51   84       -2644.27....
##  9:        52   81       -2675.13....
## 10:        52   82       -1638.16....
## 11:        52   83       -3109.98....
## 12:        52   84       -3340.92....
```

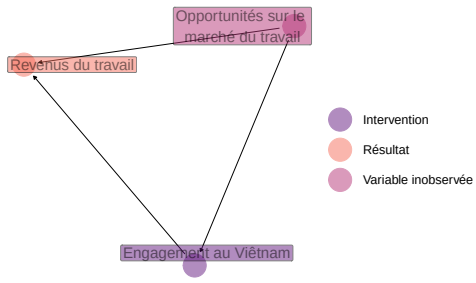
Généraliser à partir de ces exemples

Qu'est ce que ces exemples ont de commun ? Le problème

- La corrélation entre l'intervention et la variable d'intérêt n'informent pas sur les effets causaux moyens de l'intervention
  - Les prix varient avec les quantité échangés en partie à cause de variations de l'offre
  - Les jeunes hommes s'engagent dans le conflit au Viêtnam en partie parce que les opportunités qui s'ouvrent à eux sur le marché du travail sont insatisfaisantes
- La stratégie de conditionnement en bloquant les portes de sortie n'est pas raisonnablement appliquable sans avoir des données sur des variables qui sont difficiles à observer
  - Très difficile de mesurer en toute généralité les variations de l'offre
  - Pareil pour les opportunités sur le marché du travail

Qu'est-ce que ces exemples ont de commun ? Le problème

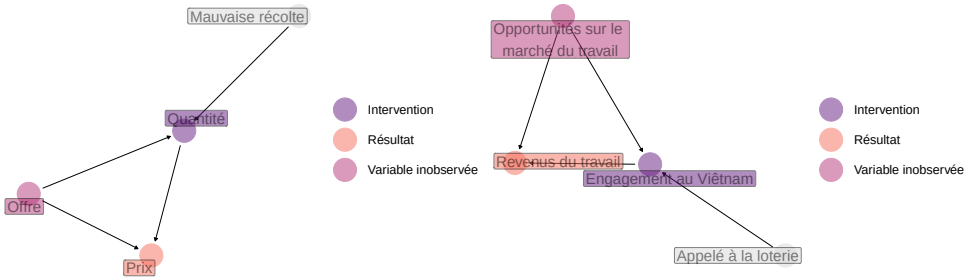




Qu'est-ce que ces exemples ont de commun ? La solution

- L'intervention résulte partiellement dans les deux cas d'un choc quasi-aléatoire
- Et qui n'a pas d'effet direct sur la variable d'intérêt
  - De mauvais rendements agricoles diminuent les quantités échangés mais ne changent rien à la demande
  - La loterie conduit à être appelé mais ne change directement rien aux revenus du travail

Qu'est-ce que ces exemples ont de commun ? La solution



Qu'est-ce que tout cela veut dire ?

- L'information encodée dans ces graphes :
  - **Hypothèse d'exogénéité** : une variable instrumentale s'assimile à l'assignation lors d'une expérience aléatoire contrôlée ou d'une expérience naturelle
    - Il faut **éventuellement conditionner** au préalable pour en arriver là
  - **Restriction d'exclusion** : une variable instrumentale n'a d'effet sur la variable d'intérêt que dans la mesure où elle a un effet sur la variable qui représente l'intervention qui nous intéresse
    - Même remarque sur le conditionnement

Cela suffit-il ? Presque !

- Il faut une hypothèse supplémentaire
  - Hypothèse de monotonie : tout le monde réagit à la variable instrumentale dans le même sens
    - C'est-à-dire pour la loterie, soit parce que l'appel encourage la participation
    - soit parce que l'appel ne change rien !
  - Hypothèse nécessaire dés lors que l'on ne suppose pas que les effets causaux de l'intervention sont constants (Imbens and Angrist (1994))
- Difficile à représenter graphiquement, une des explications à la relative hostilité de certains économètres aux approches graphiques
- Ce n'est pas toujours évidente $\longrightarrow$ un exemple la semaine prochaine

L'idée générale

- Quand on est dans cette situation
  - Les variations moyennes de la variable d'intervention avec la variable instrumentale s'interprètent comme les effets causaux moyens de la variable instrumentale sur la variable d'intervention
  - Les variations moyennes de la variable d'intérêt avec la variable instrumentale s'interprète comme les effets causaux moyens de la variable instrumentale sur la variable d'intérêt
  - Le quotient des deux s'interprète comme **les effets causaux moyens de la variable d'intervention sur la variable d'intérêt dans une certaine sous-population**
    - Les *compliers* : ceux qui réagissent à la variation de la variable instrumentale
- On parle d'**effets causaux moyens locaux** (en anglais *LATE*, *Local Average Treatment Effects*)

Peut-on apprendre des choses sur les *compliers* ?

- Dans les données du fichier `sipp84.csv`, la variable `educ` représente le niveau d'éducation (en années passées dans le système scolaire)
- Calculez le niveau d'éducation moyen dans chaque groupe défini à la fois par le fait d'avoir été appelé et le fait d'avoir effectivement participé au conflit
  - séparément pour les Blancs et les non-Blancs
  - en lissant sur 3 cohortes consécutives
- On va supposer que le niveau d'éducation atteint ne varie pas en réaction au fait d'avoir été tiré
  - Peut-on vérifier cette hypothèse ?
- Comment s'interprète le niveau d'éducation moyen dans le groupe des appelés ?
- Que peut-on en conclure ?

Peut-on apprendre des choses sur les *compliers* ?

```
#Cette fonction permet de calculer le niveau d'éducation moyen dans chaque
# cohorte en lissant sur trois cohorts consécutives
education_smoothed<-function(cohort,
                             group=NULL){
  sipp$effectif<-sipp$fnlwgt_5/10^6
  sipp[birthyear>=cohort-1
       & birthyear<=cohort+1,
       list(education=sum(educ*effectif, na.rm=TRUE)/
              sum(effectif, na.rm=TRUE)),
       by=c("race",
            "eligible",
            group)][,
                    birthyear:=cohort]

}
```

Peut-on apprendre des choses sur les *compliers* ?

```
education_reaction<-rbindlist(lapply(1950:1952,
                                     education_smoothed))

education_reaction[,
                   eligibility:=fcase(eligible==1, "eligible",
                                      eligible!=1, "noneligible")]
```

Peut-on apprendre des choses sur les *compliers* ?

```
education_reaction_wide<-
  dcast(education_reaction,
        birthyear + race ~ eligibility,
        value.var = "education",
        fun.aggregate=I,
        fill=NA_real_)[,
                       education_gap:=
                         round(eligible-noneligible,4)]
setorder(education_reaction_wide, race, birthyear)
```

En moyenne l'effet causal de la loterie sur l'éducation est proche de 0

```
##    birthyear race     eligible  noneligible education_gap
## 1:      1950    1 13.59349.... 13.68488....       -0.0914
## 2:      1951    1 13.66649.... 13.71028....       -0.0438
## 3:      1952    1 13.46819.... 13.42756....        0.0406
## 4:      1950    2 12.35349.... 12.99647....        -0.643
## 5:      1951    2 12.68049.... 12.79023....       -0.1097
## 6:      1952    2 12.96571.... 12.95315....        0.0126
```

On calcule le niveau d'éducation moyen par groupe

```
education_group<-
  rbindlist(lapply(1950:1952,
                   FUN=function(x){
                     education_smoothed(cohort = x,
                                        group = "nvstat")}))
```

On calcule le niveau d'éducation moyen par groupe

```
##    race eligible nvstat education birthyear
## 1:    1        0      0  13.84333      1950
## 2:    1        1      0  13.77660      1950
## 3:    1        1      1  13.25747      1950
## 4:    1        0      1  13.02386      1950
## 5:    2        0      0  13.00281      1950
## 6:    2        0      1  12.95608      1950
```

Il va de nouveau falloir formaliser un peu ! Un petit tableau

Valeurs observées	$D_i = 0$	$D_i = 1$
$Z_i = 0$		
$Z_i = 1$		

Il va de nouveau falloir formaliser un peu ! Un petit tableau

Valeurs observées	$D_i = 0$	$D_i = 1$
$Z_i = 0$	*never takers* + *compliers*	*always takers*
$Z_i = 1$	*never takers*	*always takers* + *compliers*

On récapitule : la loi des espérances itérées

$$\underbrace{\mathbb{E}[X_i \mid Z_i = 1]}_{\substack{\text{Education moyenne} \\ \text{chez les appelés}}}$$

$$= \underbrace{\mathbb{E}[X_i \mid D_i(1) = D_i(0) = 0, Z_i = 1]}_{\text{Moyenne des } \textit{never takers} \text{ dans le groupe}} \underbrace{\mathbb{P}(D_i(1) = D_i(0) = 0 \mid Z_i = 1)}_{\text{Part des } \textit{never takers} \text{ dans le groupe}}$$

$$+ \underbrace{\mathbb{E}[X_i \mid D_i(1) = 1, D_i(0) = 0, Z_i = 1]}_{\text{Moyenne des } \textit{compliers} \text{ dans le groupe}} \underbrace{\mathbb{P}(D_i(1) = 1, D_i(0) = 0 \mid Z_i = 1)}_{\text{Part des } \textit{compliers} \text{ dans le groupe}}$$

$$+ \underbrace{\mathbb{E}[X_i \mid D_i(1) = D_i(0) = 0, Z_i = 1]}_{\text{Moyenne des } \textit{always takers} \text{ dans le groupe}} \underbrace{\mathbb{P}(D_i(1) = D_i(0) = 0 \mid Z_i = 1)}_{\text{Part des } \textit{always takers} \text{ dans le groupe}}$$

La loterie est purement aléatoire !

$$\underbrace{\mathbb{E}[X_i \mid Z_i = 1]}_{\substack{\text{Education moyenne} \\ \text{chez les appelés}}}$$

$$= \underbrace{\mathbb{E}[X_i \mid D_i(1) = D_i(0) = 0]}_{\text{Moyenne des } \textit{never takers}} \underbrace{\mathbb{P}(D_i(1) = D_i(0) = 0)}_{\text{Part des } \textit{never takers}}$$

$$+ \underbrace{\mathbb{E}[X_i \mid D_i(1) = 1, D_i(0) = 0]}_{\text{Moyenne des } \textit{compliers}} \underbrace{\mathbb{P}(D_i(1) = 1, D_i(0) = 0)}_{\text{Part des } \textit{compliers}}$$

$$+ \underbrace{\mathbb{E}[X_i \mid D_i(1) = D_i(0) = 0]}_{\text{Moyenne des } \textit{always takers}} \underbrace{\mathbb{P}(D_i(1) = D_i(0) = 0 \mid Z_i = 1)}_{\text{Part des } \textit{always takers}}$$

Quels éléments connaît-on déjà ?

- Il n'y a pas de *defiers* !
- La moyenne des *never takers* : on l'estime dans la cellule des appelés non-vétérans
- La moyenne des *always takers* : on l'estime dans la cellule des vétérans non-appelés
- La part des *never takers* dans la population : c'est la part des non-vétérans parmi les appelés
- La part des *always takers* dans la population : c'est la part des vétérans parmi les non-appelés
- La part des *compliers* : c'est l'effet causal moyen de l'appel sur la participation au conflit

On recombine tout

$$\underbrace{\mathbb{E}[D_i \mid Z_i = 1] - \mathbb{E}[D_i \mid Z_i = 0]}_{\text{Ecart de participation entre les appelés et les autres}} \quad \underbrace{\mathbb{E}[X_i \mid D_i(1) = 1, D_i(0) = 0]}_{\text{Moyenne chez les } \textit{compliers}}$$

$$= \underbrace{\mathbb{E}[X_i \mid Z_i = 1]}_{\text{Moyenne des appelés}}$$

$$- \underbrace{\mathbb{E}[X_i \mid D_i = 0, Z_i = 1]}_{\text{Moyenne chez les } \textit{never takers}} \quad \underbrace{(1 - \mathbb{E}[D_i \mid Z_i = 1])}_{\text{Part de } \textit{never takers}}$$

$$- \underbrace{\mathbb{E}[X_i \mid D_i = 1, Z_i = 0]}_{\text{Moyenne chez les } \textit{always takers}} \quad \underbrace{\mathbb{E}[D_i \mid Z_i = 0]}_{\text{Part de } \textit{always takers}}$$

Il n'y a plus qu'à estimer : on récupère les moyennes dont on a besoin

```
moyenne_appeles<-education_reaction[eligible==1]
moyenne_appeles
```

```
##    race eligible education birthyear eligibility
## 1:    1        1  13.59349      1950    eligible
## 2:    2        1  12.35349      1950    eligible
## 3:    1        1  13.66649      1951    eligible
## 4:    2        1  12.68049      1951    eligible
## 5:    1        1  13.46819      1952    eligible
## 6:    2        1  12.96572      1952    eligible
```

Il n'y a plus qu'à estimer : on récupère les moyennes dont on a besoin

```
moyenne_nevertakers<-education_group[eligible==1
                                     & nvstat==0]
moyenne_nevertakers
```

```
##    race eligible nvstat education birthyear
## 1:    1        1      0  13.77660      1950
## 2:    2        1      0  12.20169      1950
## 3:    1        1      0  13.91570      1951
## 4:    2        1      0  12.58175      1951
## 5:    1        1      0  13.59568      1952
## 6:    2        1      0  12.86145      1952
```

Il n'y a plus qu'à estimer : on récupère les moyennes dont on a besoin

```
moyenne_alwaystakers<-education_group[eligible==0
                                      & nvstat==1]
moyenne_alwaystakers
```

```
##    race eligible nvstat education birthyear
## 1:    1        0      1  13.02386      1950
## 2:    2        0      1  12.95608      1950
## 3:    1        0      1  13.05508      1951
## 4:    2        0      1  12.87752      1951
## 5:    2        0      1  12.67354      1952
## 6:    1        0      1  12.77019      1952
```

Il n'y a plus qu'à estimer : on récupère la part de *never takers*

```
#Cette fonction permet de calculer la part de never takers dans chaque
# cohorte en lissant sur trois cohorts consécutives
nevertakers_smoothed<-function(cohort){

  sipp$effectif<-sipp$fnlwgt_5/10^6
  sipp[birthyear>=cohort-1
       & birthyear<=cohort+1
       & eligible==1,
       list(share=sum((1-nvstat)*effectif, na.rm=TRUE)/
              sum(effectif, na.rm=TRUE)),
       by=c("race")][,
                          birthyear:=cohort]

}
```

Il n'y a plus qu'à estimer : on récupère la part de *never takers*

```
nevertakers<-rbindlist(lapply(1950:1952,
                              nevertakers_smoothed))
nevertakers
```

```
##    race     share birthyear
## 1:    1 0.6472841      1950
## 2:    2 0.8042515      1950
## 3:    1 0.7169074      1951
## 4:    2 0.7986142      1951
## 5:    1 0.7689783      1952
## 6:    2 0.8551143      1952
```

Il n'y a plus qu'à estimer : on récupère la part de *always takers*

```
#Cette fonction permet de calculer la part de never takers dans chaque
# cohorte en lissant sur trois cohorts consécutives
alwaystakers_smoothed<-function(cohort){

  sipp$effectif<-sipp$fnlwgt_5/10^6
  sipp[birthyear>=cohort-1
       & birthyear<=cohort+1
       & eligible==0,
       list(share=sum((nvstat)*effectif, na.rm=TRUE)/
              sum(effectif, na.rm=TRUE)),
       by=c("race")][,
                          birthyear:=cohort]

}
```

Il n'y a plus qu'à estimer : on récupère la part de *always takers*

```
alwaystakers<-rbindlist(lapply(1950:1952,
                               alwaystakers_smoothed))
alwaystakers
```

```
##    race     share birthyear
## 1:    1 0.1933552      1950
## 2:    2 0.1354721      1950
## 3:    1 0.1468978      1951
## 4:    2 0.1514002      1951
## 5:    1 0.1257396      1952
## 6:    2 0.1287759      1952
```

Il n'y a plus qu'à estimer : on recombine tout

```
education_compliers<-
  merge(merge(moyenne_nevertakers[,c("race",
                                     "birthyear",
                                     "education")],
              nevertakers,
              by=c("race","birthyear")),
        merge(moyenne_alwaystakers[,c("race",
                                     "birthyear",
                                     "education")],
              alwaystakers,
              by=c("race","birthyear")))
```

Il n'y a plus qu'à estimer : on recombine tout

```
education_compliers<-
  merge(education_compliers,
        moyenne_appeles,
        by=c("race","birthyear"))
```

Il n'y a plus qu'à estimer : on recombine tout

```
education_compliers<-merge(education_compliers[,birthyear:=
                                                 birthyear-1900],
                           veteran_probability_wide[,
                                                    c("race",
                                                      "birthyear",
                                                      "probability_gap")],
                           by=c("race",
                                "birthyear"))
```

Il n'y a plus qu'à estimer : on recombine tout

```
education_compliers[,
                    education_compliers:=
                      1/probability_gap*
                      (education-
                         education.x*share.x-
                         education.y*share.y)]
```

Les caractéristiques observables des *compliers*

```
##    race birthyear probability_gap education_compliers
## 1:    1        50          0.1594        13.53757....
## 2:    1        51          0.1362        13.01370....
## 3:    1        52          0.1053        13.36843....
## 4:    2        50          0.0603        13.01956....
## 5:    2        51            0.05        13.65735....
## 6:    2        52          0.0161        20.84853....
```

Pour conclure

- On a une technique qui permet d'identifier les effets causaux de l'intervention
- Pour une sous-population qu'on sait caractériser du point de vue de ses caractéristiques observables
- Mais souvent ce sont des caractéristiques inobservables qui importent !
  - Raison pour laquelle l'approche en termes de LATE (*Local Average Treatment Effect*) est parfois critiquée

La prochaine séance

- On a juste explicité le principe de la méthode
- Ce qu'il nous faut encore discuter :
  - Problèmes pratiques d'estimation
  - Comment quantifier l'incertitude ?
  - Comment trouver des instruments convaincants ?

Bibliographie

Bibliographie I

Angrist, Joshua D. 1990. "Lifetime Earnings and the Vietnam Era Draft Lottery : Evidence from Social Security Administrative Records." *The American Economic Review* 80 (3) : 313–36. http://www.jstor.org/stable/2006669.

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62 (2) : 467–75.

Wright, Philip G. 1928. *Tariff on Animal and Vegetable Oils*. Macmillan Company, New York.