Modélisation et identification causale
Séance 7 – Comment utiliser une variable instrumentale ?

Pierre Pora

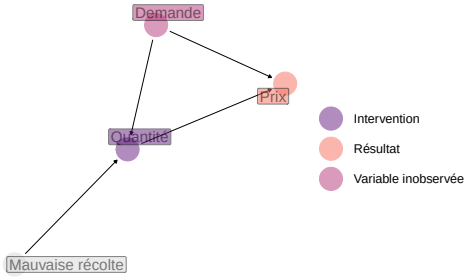
2023-01-09

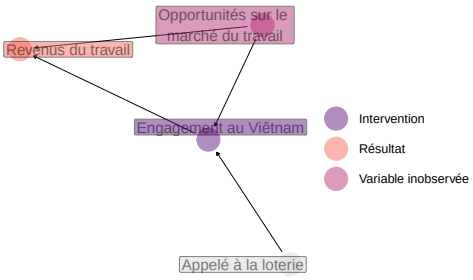
Objet de la séance

Retour sur la séance précédente

- On a introduit une **nouvelle approche** :
  - Elle repose toujours sur une **variation quasi-aléatoire** (~ expérience naturelle)
  - Cette variation quasi-aléatoire **ne concerne pas directement** l'intervention qui nous intéresse
    - La comparaison naïve des individus selon la valeur de l'intervention n'a pas d'interprétation causale
- L'approche repose sur des hypothèses relatives au **mécanisme causal** que l'on peut visualiser avec un graphe

Graphe représentant une variable instrumentale





Le principe derrière le graphe

- Tout l'effet causal de la mauvaise récolte sur le prix s'explique par le fait que les mauvaises récoltes changent la quantité produite
  - Si on divise l'effet causal de la mauvaise récolte sur le prix par l'effet causal de la mauvaise récolte sur la quantité produite, alors on récupère la pente des courbes de demande ~ l'effet causal de l'offre sur les prix

- Tout l'effet causal de la loterie sur les revenus en 1984 s'explique par le fait qu'elle conduit certains jeunes hommes américains à participer à la guerre du Viêtnam au début des années 70
  - Si on divise l'effet causal de l'appel généré par la loterie sur les revenus du travail en 1984 par l'effet causal de cet appel sur la participation effective au conflit, alors on récupère l'effet causal de la participation à la guerre du Viêtnam sur les revenus du travail ultérieurs

Une hypothèse qui ne figure pas sur le graphe : la monotonie

- **Hypothèse de monotonie** : il faut que la variable instrumentale affecte tout le monde dans le même sens
  - Ici dans le même sens inclut les individus qui ne réagissent pas à la variation contrefactuelle de la variable instrumentale

Retour sur Angrist (1990) : l'hypothèse de monotonie affirme qu'il n'y a pas de *defiers*

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 appelés (*defiers*)
$D_i(1) = 1$	Ne participent pas si pas appelés, participent si appelés (*compliers*)	Participent toujours (*always takers*)

Notion d'effet causal moyen local

- Avec cette approche **on ne peut pas identifier les effets causaux moyens pour toute la population**
  - Il y a des *always takers* et des *never takers* qui ne réagissent pas du tout aux variations de la variable instrumentale
  - Regarder les effets de la variation de la variable instrumentale ne donne aucune information sur ces sous-populations
- La quantité que l'on peut identifier est l'**effet causal moyen local** de l'intervention chez les individus qui réagissent aux variations de la variable instrumentale
  - Les *compliers* dans le tableau précédent
  - *Local Average Treatment Effect* (*LATE*) en anglais

Estimande de Wald

- Dans le cas où l'instrument et l'intervention sont binaires, les effets causaux moyens de l'intervention sur la variable d'intérêt = le rapport entre l'effet de l'instrument sur la variable d'intérêt et l'effet de l'instrument sur l'intervention
- Généralisation à n'importe quel type d'instrument Z et d'intervention D : **estimande de Wald**

$$\frac{\mathscr{C}(Z,Y)}{\mathscr{C}(Z,D)}$$

- Dans le cas général cette quantité est égale à une moyenne d'effets causaux moyens locaux avec des poids que l'on connaît et qui mettent l'accent sur les sous-populations qui réagissent le plus aux variations de l'instrument
  - **Il faut toujours que les hypothèses soient respectées** (exogénéité, restriction d'exclusion, monotonie)

Comparaison avec la stratégie de conditionnement

- Conditionner sur une variable = comparer des individus ayant (approximativement) la même valeur pour cette variable $\longrightarrow$ **variabilité *intra***
- Ici c'est tout le contraire ! On compare des individus ayant des valeurs différentes pour la variable instrumentale, sans se soucier de la valeur de la variable d'intervention $\longrightarrow$ **variabilité *inter***

L'objet de cette séance

1. Comment **trouver des instruments convaincants** ?
2. Comment **estimer les effets causaux moyens locaux** ?

Comment trouver des instruments convaincants ?

Ce que l'on veut

- **Hypothèse d'exogénéité** : les valeurs prises par la variable instrumentale s'assimilent aux résultats d'une **expérience aléatoire contrôlée** ou d'une **expérience naturelle** (quitte à conditionner sur des variables observables)
- **Restriction d'exclusion** : l'effet causal moyen de la variable instrumentale sur la variable d'intérêt **ne transite que par l'intervention** (quitte à conditionner sur des variables observables)
- **Hypothèse de monotonie** : la variable instrumentale change les valeurs de l'intervention **dans le même sens** pour tous les individus de la population (y compris pas du tout)

Qu'est-ce que cela veut dire concrètement ? Une petite heuristique

- La variable instrumentale doit :
  - **Ne pas avoir de lien causal avec l'ensemble des (nombreuses) causes de l'intervention** qui sont aussi des causes de la variable d'intérêt
    - Ces causes jointes de l'intervention et de la grandeur d'intérêt sont quantitativement importantes !
    - Sinon on pourrait faire la comparaison naïve
  - **Ne pas avoir de lien causal direct avec la grandeur d'intérêt**
    - Tout passe par l'intervention

Qu'est-ce que cela veut dire concrètement ? Une petite heuristique

- La situation idéale
  - On ne voit pas *a priori* quel est le rapport entre la variable instrumentale et la grandeur d'intérêt
  - *A posteriori* quand on explique quelle est l'intervention et pourquoi la variable instrumentale change les valeurs de la variable d'intervention on comprend pourquoi elles sont corrélées

Comment y parvenir ?

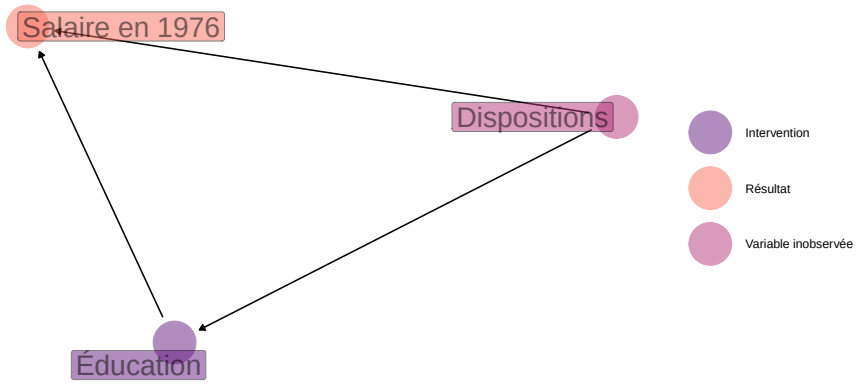
1. S'habituer au raisonnement
2. Connaître ses classiques
3. Connaître le contexte institutionnel

S'habituer au raisonnement

Un premier exemple d'instrumentation par la distance : Card (1993)

- Un exemple classique : l'effet de l'éducation (en années passées dans le système scolaire) sur le salaire futur, aux États-Unis
- Une partie des dispositions (inobservées) qui pousse à des études plus longues est également valorisée sur le marché du travail
- Travail à partir des données du *National Longitudinal Survey of Young Men* (NLSYM) une enquête longitudinale du *Bureau of Labor* étatsunien
  - Enquête qui suit de jeunes hommes (14-24 ans) à partir de 1966
  - Suivi annuel jusqu'en 1981

Le problème que l'on rencontre



La solution proposée par Card (1993)

- Card propose d'**instrumenter le niveau d'éducation par la distance entre le logement en 1966 et une université**
- Qu'est-ce que cela signifie ?
  - Quelle **comparaison** faut-il effectuer pour identifier les effets causaux de l'éducation sur le salaire ?
  - Quelles **hypothèses** cela suppose-t-il de faire ?
  - Ces hypothèses sont-elles **crédibles** ?

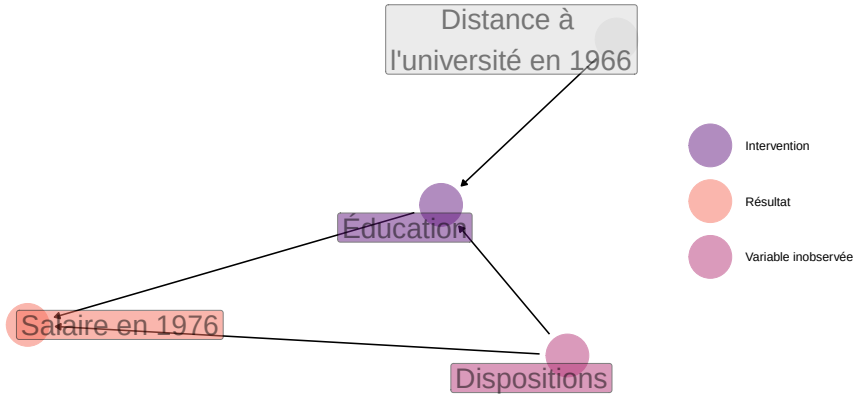
La solution proposée par Card (1993)

- La comparaison que l'on veut faire : celle des jeunes hommes qui vivaient en 1966 très près d'une université avec les autres
- L'idée est qu'indépendamment de leurs dispositions, vivre très près d'une université rend plus probable d'entrer dans l'enseignement supérieur long
  - Par exemple, pas besoin de payer pour un logement supplémentaire $\longrightarrow$ coût plus faible

Les hypothèses

- **Exogénéité** : les dispositions (inobservées) sont les mêmes que l'on habite très près d'une université ou pas
- **Restriction d'exclusion** : habiter très près d'une université en 1966 n'a d'effet sur le salaire en 1976 que dans la mesure où cela augmente le niveau d'éducation
- **Monotonie** : il n'y a pas de jeunes hommes qui réduisent leur niveau d'éducation s'ils habitent très près d'une université

Les hypothèses



Les critiques que l'on peut faire

- La distance est estimée à partir de la **résidence en 1966**
  - Une partie des enquêtés est potentiellement déjà à l'université à ce moment-là
    - Causalité inversée

- La décision de résidence est potentiellement causée par des variables qui sont cause des dispositions qui causent à la fois le salaire et l'éducation
  - Par exemple les origines sociales

  - En pratique Card conditionne par le niveau d'éducation des parents
    - Pas certain que ce soit suffisant

- Les universités peuvent être implantées dans des marchés du travail locaux très dynamiques sans que cela ne vienne de la présence de l'université, ni ne se limite aux emplois des diplômés de l'enseignement supérieur
  - En pratique Card contrôle conditionne sur la distinction urbain / rural
    - Pas certain que cela soit suffisant

Un second exemple d'instrumentation par la distance : Duflo (2001)

- Toujours sur la même question des effets de l'éducation sur les salaires futurs
  - Mais cette fois-ci dans un contexte différent : en Indonésie dans les années 70
  - Et en se concentrant sur l'éducation primaire
- La remarque initiale sur les **difficultés de l'interprétation causale de la corrélation entre éducation et salaire** tient toujours

Peut-on répondre (partiellement) aux critiques que l'on adresse à Card ?

- Les écoles (universités) peuvent être implantées dans des marchés du travail locaux plus dynamiques sans que cela ne tienne à la présence des écoles (universités)
  - Cela pourrait même plutôt être le contraire
- On pourrait regarder cela avec des données antérieures à l'implantation !
- C'est ce que fait Duflo
  - Programme indonésien de création de nouvelles écoles dans les années 70
  - On peut comparer entre eux les enfants de la même commune dont l'accès à l'école est différent parce qu'ils appartiennent à des générations différentes
  - Mais aussi à ceux de la même génération des communes sans écoles

D'une génération à l'autre, le temps passé dans le système scolaire par les enfants des communes dans lesquelles une école a été créée dans les années 70, a augmenté plus vite que dans les autres communes, et leur salaire dans les années 90 a également augmenté plus vite

FIGURE 3. COEFFICIENTS OF THE INTERACTIONS AGE IN 1974* PROGRAM INTENSITY IN THE REGION OF BIRTH IN THE WAGE AND EDUCATION EQUATIONS

Instrumenter par les ouvertures / fermetures de services locaux

- En fait on fait ici une espèce de différence-de-différence sauf que tous les enfants des communes où une école se crée ne la fréquentent pas, et tous les enfants des communes sans école ne sont pas déscolarisés
  - Si on veut vraiment pousser l'idée jusqu'au bout quelques précautions à prendre (Chaisemartin and D'Haultfœuille (2017))
- **Approche très classique** : crèches, écoles, universités, services publics, services commerciaux, presse etc.
- Il faut que l'ouverture / fermeture de service ne génère pas de changement dans la composition de la population
  - Ou ne soit pas générée par un tel changement

Un troisième exemple d'instrumentation par la distance : Pascali (2017)

- Quel est l'effet du commerce international sur la croissance ?
  - Question très difficile : ce qui est exporté d'un côté est importé de l'autre
  - Intégration dans le commerce international et croissance dépendent (notamment) de la spécialisation technologique
    - La corrélation n'a pas d'interprétation causale
- L'intensité des échanges entre deux pays dépend de la distance entre les deux
  - Mais celle-ci ne varie pas dans le temps
  - Et c'est la même dans les deux sens !
  - Difficile *a priori* d'en faire quelque chose

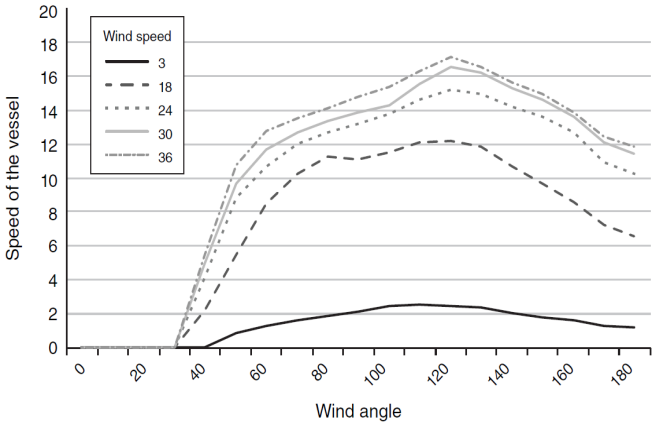
L'idée de Pascali (2017)

- Ce qui compte : pas la distance à vol d'oiseau, mais le temps de transport maritime
- Quand on se déplace avec un bateau à voile, la vitesse n'est pas la même dans toutes les directions !
- Et à la surface du globe le vent a des directions privilégiées !
- Au contraire avec un bateau à vapeur on peut aller dans toutes les directions à la même vitesse

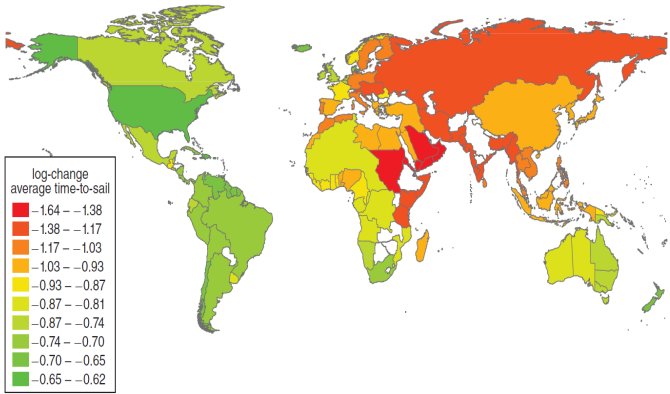
L'idée de Pascali (2017)

- Conséquence de tout cela
  1. La distance entre deux pays aux époques où le transport se fait avec un bateau à voile n'est pas la même dans les deux sens
  2. Le passage au bateau à vapeur n'a pas réduit les distances de la même façon pour tous les couples exportateur → importateur
  3. Certains pays se sont retrouvés de ce fait beaucoup plus près de leurs cibles d'export que ce qui était le cas auparavant !

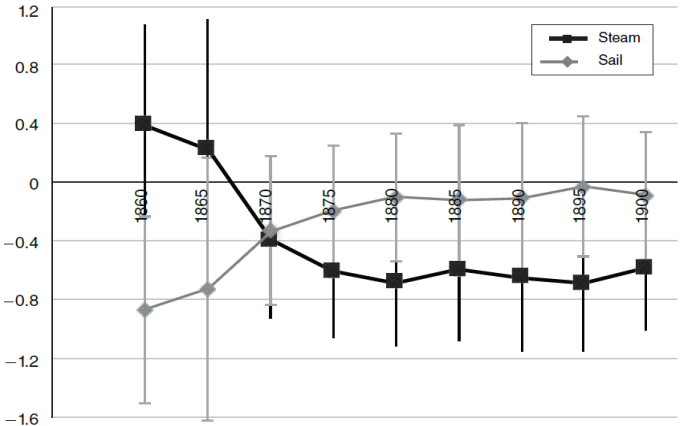
On ne va pas à la même vitesse dans toutes les directions quand on utilise un bateau à voile



Certains pays ont été beaucoup plus rapprochés de leurs importateurs potentiels que les autres



Les flux exportateur → importateur sont très corrélés (négativement) à la durée du voyage maritime avec la technologie en vigueur



Les résultats de Pascali (2017)

- Les pays qui ont été les plus rapprochés de leurs importateurs potentiels ont vu leurs exports augmenter plus vite...
- Mais leur croissance économique a été plus faible !
- **Les techniques modernes d'inférence causale ne sont pas réservées à des questions très localisées**

Encore un exemple d'instrumentation par la distance : Braun (2022)

- Travail de sociologie historique sur les sources de l'antisémitisme en Allemagne avant le génocide
  - Le caractère beaucoup plus saillant des frontières nationales après la Première Guerre mondiale a-t-il contribué à la montée de l'antisémitisme dans la République de Weimar ?
- Données recueillies par des folkloristes entre 1930 et 1935 qui permettent de repérer la présence de thèmes antisémites (croque-mitaines juifs) dans les contes pour enfants
- La variable indépendante est la distance (par la route) à la frontière physique

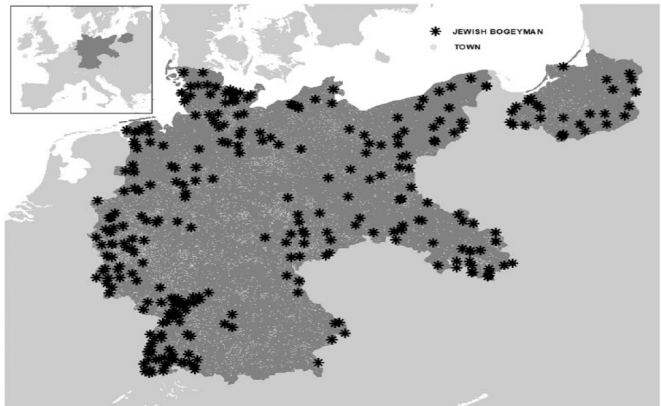
La frontière physique est devenue beaucoup plus saillante pendant et après la Première Guerre mondiale

(a) Border Crossing, 1908

(b) Border Crossing, 1915

(c) Border Crossing, 1922

Les histoires pour enfant avec des thèmes antisémites sont concentrées autour des frontières physiques de la République de Weimar



(a) Jewish Bogeymen in Localities

Cette corrélation a-t-elle une interprétation causale ?

- Les routes actuelles traduisent notamment des habitudes de voyages commerciaux...
- Auxquelles la migration de populations juives européennes a beaucoup contribué...
- Et qui peuvent donc résulter d'épisodes antisémites passés...
- Qui peuvent eux-mêmes découler de disparités géographiques persistantes en matière d'antisémitisme

La solution proposée par Braun (2022)

- Les routes au début du XXème siècle reprennent souvent le tracé de très anciennes routes médiévales
- Instrumenter la distance à la frontière par la route au début du XXème siècle par la distance par les routes existant avant la Peste noire
- Critique que l'on pourrait faire : ces routes pourraient déjà traduire des épisodes de violences antisémites antérieurs ?
  - Proposition de test : la distance à la frontière de la République de Weimar par la route antérieure à la Peste noire est-elle corrélée aux nombreux épisodes de violence antisémite (pogroms) pendant la Peste noire
  - De façon rassurante : non !
- En définitive le caractère saillant de la frontière physique a vraisemblablement contribué à la diffusion de thèmes antisémites

Connaître ses classiques

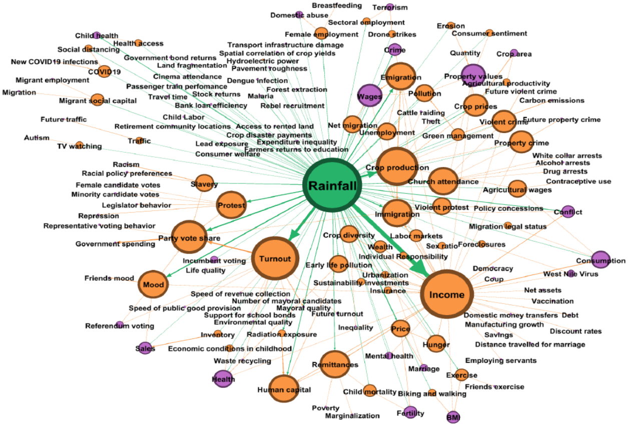
S'appuyer sur la littérature

- **Il y a des classiques dans tous les domaines !**
  - Effets de l'éducation sur les trajectoires professionnelles : distance aux établissements, trimestre de naissance (à suivre)
  - Effets de la vie familiale sur la vie professionnelle des femmes : les jumeaux, le sexe des deux premiers enfants...
  - Effet des décisions judiciaires : assignation quasi-aléatoire à des juges plus ou moins sévères
- Un classique n'est pas forcément à imiter mais
  - à améliorer
  - à critiquer
- Connaître la littérature dans laquelle on s'inscrit permet aussi de savoir quels instruments ne sont plus aujourd'hui considérés comme crédibles !

Le danger des instruments classiques

- Il faut que la restriction d'exclusion soit crédible
- **Le même choc quasi-aléatoire ne peut pas être un instrument crédible pour toutes les interventions possibles**
- Exemple de Mellon (2022) : les chocs météorologiques

Toutes les restrictions d'exclusion ne peuvent pas être vraies simultanément



Connaître le contexte institutionnel

La crédibilité d'un instrument n'est (presque pas) une question statistique !

- Presque tout tient au fait que l'on a une bonne histoire et que l'on explique bien les comparaisons que l'on fait
- Tout ce qui peut veut soutenir l'argument est bon à prendre si cela permet d'éclairer le mécanisme qui justifie l'usage de l'instrument
  - Y compris des données qualitatives
- Ne pas oublier de discuter l'hypothèse de monotonie
- Le contenu statistique de la méthode est (pour l'instant) très faible
  - Ce qui compte c'est avant tout une bonne connaissance du contexte institutionnel et des faits sociaux que l'on étudie

L'estimation en pratique

Un exemple classique : Angrist and Krueger (1991)

- Encore un papier sur l'effet de l'éducation sur le salaire...
- Stratégie de variable instrumentale :
  - l'âge de fin de la scolarité obligatoire est défini à partir de l'âge exact (16 ans révolus en France)
  - l'entrée dans le système scolaire se fait sur la base de l'année de naissance
    - les enfants nés en fin d'année entrent dans le système scolaire plus jeunes en moyenne que les autres
- Si la part d'élèves qui restent dans le système scolaire seulement du fait de l'obligation légale de scolarité ne dépend pas du jour de naissance...
- alors les personnes nées en fin d'année passent en moyenne plus de temps dans le système scolaire que les personnes nées en début d'année

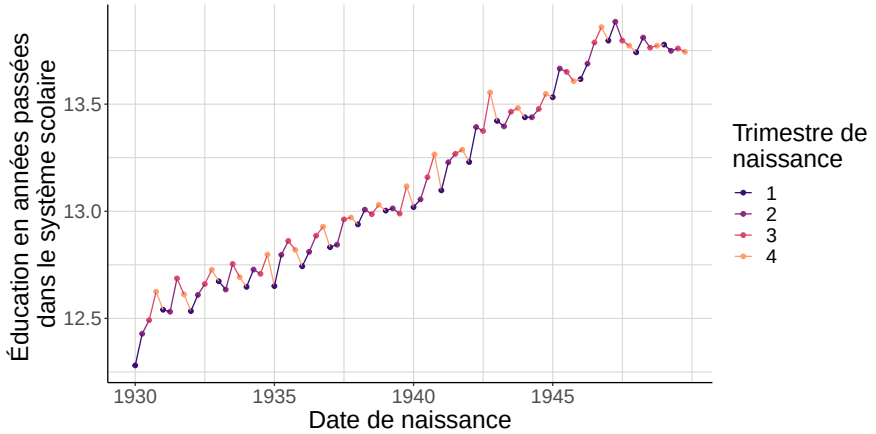
L'éducation dépend-elle du trimestre de naissance ?

- Chargez les données du fichier `ak80.csv` qui contient un extrait de l'information issue du *Census* étatsunien de 1980 exploité par Angrist et Krueger
- Pour chaque année de naissance de 1930 à 1949 (variable `yob`), représentez le niveau d'éducation moyen atteint en 1980 (variable `educ`) en fonction du trimestre de naissance (variable `qob`) pour reproduire les figures I à III de l'article. Que peut-on remarquer ?
- Pour chaque année de naissance de 1930 à 1949 (variable `yob`), représentez le salaire (en logarithme) moyen en 1980 (variable `lwage`) en fonction du trimestre de naissance (variable `qob`) pour reproduire la figure IV de l'article. Que peut-on remarquer ?
- Sous quelles hypothèses peut-on utiliser ce fait comme instrument pour identifier les effets causaux moyens de l'éducation sur le salaire ?
- Sur quelle population portent ces effets causaux moyens ?

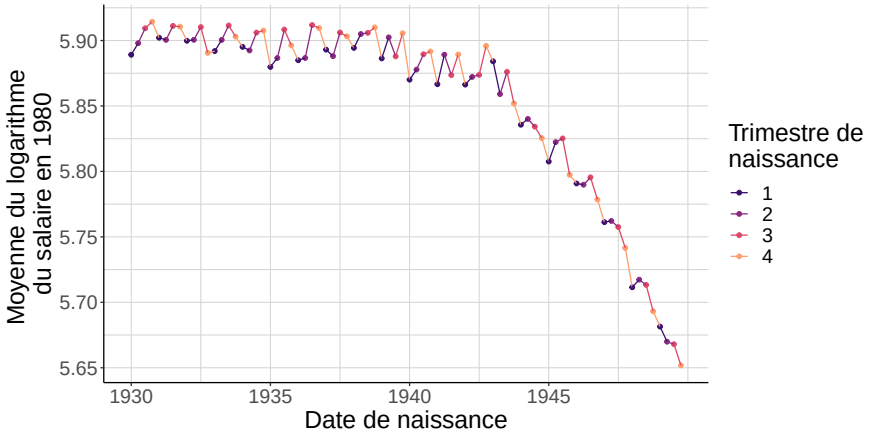
On charge les données

```
ak80<-fread("./Data/AngristKrueger1991/ak80.csv")

ak80[,
     dob:=yob+0.25*(qob-1)]

average_quarter<-ak80[,
                      list(education=mean(educ),
                           lwage=mean(lwage),
                           sample_size=.N),
                      by=c("yob",
                           "qob",
                           "dob")]
```

En moyenne les personnes nées plus tard dans l'année ont passé plus de temps dans le système scolaire que celles nées plus tôt la même année



En moyenne les personnes nées plus tard dans l'année ont passé plus de temps dans le système scolaire que celles nées plus tôt la même année



Vérifier que le mécanisme est bien celui que l'on pense

- Retranchez au niveau moyen d'éducation le niveau d'éducation moyen estimé sur les 2 cohortes qui précèdent et les deux cohortes qui suivent, et représentez la différence (figure IV du papier)
- Reproduisez l'exercice sur la probabilité d'atteindre l'équivalent du baccalauréat (12 années dans le système scolaire), la probabilité d'atteindre l'équivalent de la licence (16 années dans le système scolaire), la probabilité d'atteindre le niveau master (18 années dans le système scolaire) et d'atteindre le niveau doctorat (20 années dans le système scolaire).
- Utilisez une régression linéaire pour estimer l'effet d'être né le premier, le deuxième ou le troisième trimestre plutôt que le dernier (table I du papier). Que peut-on en déduire ?

Estimer les probabilités d'atteindre les différents niveaux

```
average_quarter_education<-
  ak80[,
       list(educ=mean(educ),
            highschoolgrad=mean(educ>=12),
            bachelor=mean(educ>=16),
            master=mean(educ>=18),
            phd=mean(educ>=20),
            lwage=mean(lwage),
            sample_size=.N),
       by=c("yob",
            "qob",
            "dob")]
average_quarter_education<-setorder(average_quarter_education,
                                    dob)
```

Estimer la moyenne mobile

```
trend_education<-
  average_quarter_education[,
                            lapply(X=.SD,
                                   FUN=function(x){
                                     rollmean(x,
                                              5,
                                              align=
                                                "center")
                                     }),
                            .SDcols=c("dob",
                                      "educ",
                                      "highschoolgrad",
                                      "bachelor",
                                      "master",
                                      "phd")]
```

Restructurer et retrancher la moyenne mobile

```
long_average_quarter_education<-
  melt(average_quarter_education[,c("dob",
                                    "educ",
                                    "highschoolgrad",
                                    "bachelor",
                                    "master",
                                    "phd")],
       id.vars=c("dob"),
             variable.name="level",
             value.name="mean_quarter")
```

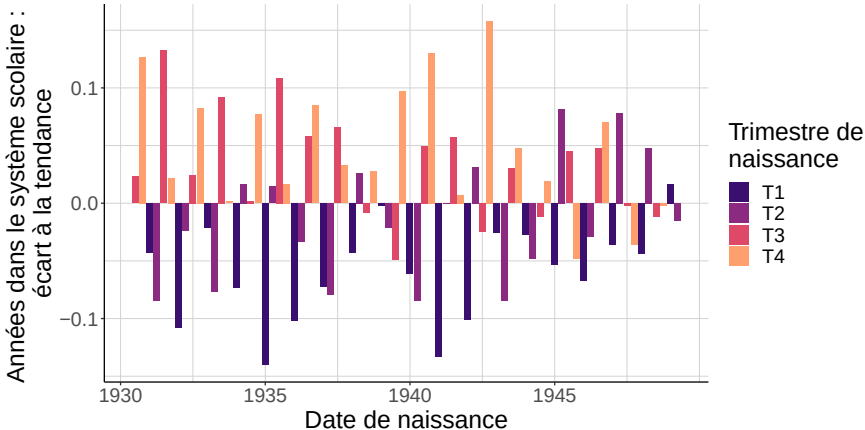
Restructurer et reformater la moyenne mobile

```
long_trend_education<-
  melt(trend_education[,c("dob",
                          "educ",
                          "highschoolgrad",
                          "bachelor",
                          "master",
                          "phd")],
       id.vars=c("dob"),
             variable.name="level",
             value.name="trend")
```

Restructurer et reformater la moyenne mobile

```
long_average_quarter_education<-
  merge(long_average_quarter_education,
        long_trend_education,
        by=c("dob",
             "level"))[,
                       residual:=5/4*mean_quarter-5/4*trend]
```

Représenter le résidu



Cette disparité apparaît-elle pour tous les niveaux d'éducation ?

```
long_average_quarter_education[,
                               quarter:=
                                 as.factor(paste0("T",
                                                  (4*(dob-floor(dob))+
                                                     1)))]
reg_quarter_education<-function(educlevel,
                                cohort){

  reg<-lm(residual ~ relevel(quarter,
                             ref="T4"),
          data=long_average_quarter_education[level==educlevel
                                              & dob>=min(cohort)
                                              & dob<max(cohort)+1])

  coeftest(reg,
           vcov=vcovHC(reg))

}
```

Le trimestre de naissance est corrélé au temps total passé dans le système scolaire

```
reg_quarter_education(educlevel = "educ", cohort=30:39)
```

```
##
## t test of coefficients:
##
##                                 Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                     0.056795   0.013932  4.0766 0.0002598 ***
## relevel(quarter, ref = "T4")T1 -0.124028   0.021011 -5.9031 1.151e-06 ***
## relevel(quarter, ref = "T4")T2 -0.085940   0.020724 -4.1468 0.0002121 ***
## relevel(quarter, ref = "T4")T3 -0.011871   0.023463 -0.5059 0.6161629
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Mais à mesure que l'on considère des niveaux d'éducation plus élevés, la corrélation s'affaiblit

```
reg_quarter_education(educlevel = "highschoolgrad", cohort=30:39)
```

```
##
## t test of coefficients:
##
##                                  Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                     0.0107285  0.0022109  4.8525 2.666e-05 ***
## relevel(quarter, ref = "T4")T1 -0.0189992  0.0026059 -7.2909 1.919e-08 ***
## relevel(quarter, ref = "T4")T2 -0.0196566  0.0030124 -6.5252 1.806e-07 ***
## relevel(quarter, ref = "T4")T3 -0.0036935  0.0029412 -1.2558    0.2178
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Mais à mesure que l'on considère des niveaux d'éducation plus élevés, la corrélation s'affaiblit

```
reg_quarter_education(educlevel = "bachelor", cohort=30:39)
```

```
##
## t test of coefficients:
##
##                                  Estimate  Std. Error t value Pr(>|t|)
## (Intercept)                     1.4225e-05  1.8479e-03  0.0077   0.9939
## relevel(quarter, ref = "T4")T1 -4.9793e-03  3.0047e-03 -1.6572   0.1067
## relevel(quarter, ref = "T4")T2  2.8586e-03  2.3288e-03  1.2275   0.2281
## relevel(quarter, ref = "T4")T3  2.4122e-03  3.3864e-03  0.7123   0.4811
```

Mais à mesure que l'on considère des niveaux d'éducation plus élevés, la corrélation s'affaiblit

```
reg_quarter_education(educlevel = "master", cohort=30:39)
```

```
##
## t test of coefficients:
##
##                                  Estimate  Std. Error t value Pr(>|t|)
## (Intercept)                     9.8745e-05  1.5318e-03  0.0645   0.9490
## relevel(quarter, ref = "T4")T1 -1.0097e-03  1.8290e-03 -0.5521   0.5845
## relevel(quarter, ref = "T4")T2  1.8599e-03  2.0304e-03  0.9160   0.3661
## relevel(quarter, ref = "T4")T3 -7.9870e-04  1.8594e-03 -0.4296   0.6702
```

Mais à mesure que l'on considère des niveaux d'éducation plus élevés, la corrélation s'affaiblit

```
reg_quarter_education(educlevel = "phd", cohort=30:39)
```

```
##
## t test of coefficients:
##
##                                  Estimate  Std. Error t value Pr(>|t|)
## (Intercept)                    -0.00105835  0.00075272 -1.4060  0.16879
## relevel(quarter, ref = "T4")T1  0.00160222  0.00140051  1.1440  0.26060
## relevel(quarter, ref = "T4")T2  0.00245529  0.00111105  2.2099  0.03395 *
## relevel(quarter, ref = "T4")T3  0.00042309  0.00098552  0.4293  0.67041
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Estimer l'estimande de Wald

Quel est l'effet de l'éducation sur le salaire ?

- Créez une nouvelle variable dichotomique qui vaut 1 pour les individus nés le premier trimestre et 0 sinon
- En assimilant le trimestre de naissance à une expérience naturelle, estimez pour la cohorte 1930-1939 :
  - le rapport entre l'effet d'être né le premier trimestre sur le salaire et l'effet d'être né le premier trimestre sur l'éducation
  - le rapport entre la covariance de la nouvelle variable et le salaire et la covariance de la nouvelle variable et de l'éducation
  - le rapport entre le coefficient de la nouvelle variable, dans une régression linéaire par les MCO du salaire sur cette variable, et le coefficient de la nouvelle variable, dans une régression linéaire par les MCO de l'éducation sur cette variable
  - le coefficient sur la valeur prédite par la régression de l'éducation sur la nouvelle variable, dans une régression du salaire sur cette valeur prédite

Créer la variable indiquant la naissance le premier trimestre

```
ak80[,
     q1:=as.numeric(qob==1)]
```

Estimer le rapport des contrastes

```
contrast_ratio<-
  ak80[yob %in% 30:39,
       list(estimate=
              (sum(lwage*q1)/sum(q1)-
                 sum(lwage*(1-q1))/sum(1-q1))/
              (sum(educ*q1)/sum(q1)-
                 sum(educ*(1-q1))/sum(1-q1)))]
```

Estimer le rapport des covariances

```
cov_ratio<-
  ak80[yob %in% 30:39,
       list(estimate=
              cov(lwage,
                  q1)/
              cov(educ,
                  q1))]
```

Estimer le rapport des coefficients

```
reg_wage<-lm(lwage ~q1,
             data=ak80[yob %in% 30:39])
reg_educ<-lm(educ ~q1,
             data=ak80[yob %in% 30:39])

coeff_ratio<-
  reg_wage$coefficients["q1"]/
  reg_educ$coefficients["q1"]
```

Estimer le coefficient sur la valeur prédite de l'éducation

```
ak80[yob %in% 30:39,
     pred_educ:=reg_educ$fitted.values]

reg_wage_pred_educ<-lm(lwage ~ pred_educ,
                       data=ak80[yob %in% 30:39])

coeff_pred_educ<-reg_wage_pred_educ$coefficients["pred_educ"]
```

Toutes ces techniques renvoient la même quantité

```
all(all.equal(as.numeric(contrast_ratio),
              as.numeric(cov_ratio)),
    all.equal(as.numeric(contrast_ratio),
              as.numeric(coeff_ratio)),
    all.equal(as.numeric(contrast_ratio),
              as.numeric(coeff_pred_educ)))
```

```
## [1] TRUE
```

Faut-il utiliser ces techniques pour l'estimation ?

- La réponse est négative !
- Aucune de ces techniques ne permet en soi de récupérer une **estimation satisfaisante de l'incertitude** sur l'effet
- La dernière doit être envisagée avec méfiance
  - De façon automatique les logiciels statistiques vont renvoyer une estimation de l'incertitude
  - Cette estimation traite la valeur prédite comme une valeur observée sur laquelle l'incertitude est nulle
- Il faut privilégier les fonctions pré-implémentées de R, comme `ivreg` (package `AER`) ou `felm` (package `lfe`)
- Ne pas oublier l'hétéroscédasticité ou le *clustering* des observations

Un peu de vocabulaire

- L'estimateur usuel est dit **estimateur moindres carrés en deux étapes** (*Two-Stage Least Squares* en anglais, 2SLS)
- La régression par les MCO de l'éducation sur le trimestre de naissance est la **première étape**
- La régression par les MCO du salaire sur la valeur prédite de l'éducation est la **seconde étape**
- La régression par les MCO du salaire directement sur le trimestre de naissance est la **forme réduite**

Estimer les effets causaux moyens de l'éducation sur le salaire

- Utilisez la fonction `ivreg` du package `AER` pour estimer les effets causaux moyens locaux de l'éducation sur le salaire
- Utilisez la fonction `felm` du package `lfe` pour estimer les effets causaux moyens locaux de l'éducation sur le salaire
- Sur quelle population portent ces effets ?

La fonction `ivreg` permet de récupérer les effets causaux moyens locaux de l'éducation pour la sous-population des hommes qui, sans la scolarité obligatoire quitteraient le système scolaire plus tôt

```
iv_estimates<-ivreg(lwage ~ educ | q1,
                    data=ak80[yob %in% 30:39])
coeftest(iv_estimates,
         vcov=vcovHC(iv_estimates,
                     type = "HC0"))
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 4.597477   0.306889 14.9809 < 2.2e-16 ***
## educ        0.101995   0.024032  4.2442 2.194e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La fonction `lfe` permet de récupérer les effets causaux moyens locaux de l'éducation pour la sous-population des hommes qui, sans la scolarité obligatoire quitteraient le système scolaire plus tôt

```
iv_estimates_felm<-felm(lwage ~ 1 | 0 |  (educ ~ q1) | 0,
                        data=ak80[yob %in% 30:39])
```

La fonction `lfe` permet de récupérer les effets causaux moyens locaux de l'éducation pour la sous-population des hommes qui, sans la scolarité obligatoire quitteraient le système scolaire plus tôt

```
summary(iv_estimates_felm,
        robust=TRUE)$coefficients
```

```
##               Estimate Robust s.e   t value     Pr(>|t|)
## (Intercept) 4.5974757 0.30689000 14.980859 1.017914e-50
## `educ(fit)` 0.1019951 0.02403159  4.244208 2.194265e-05
```

Problème des instruments faibles

Peut-on multiplier les instruments ? Faut-il le faire ?

- Il y a des situations dans lesquelles on pourrait imaginer utiliser **plusieurs instruments**
- Des raisons conceptuelles d'éviter de le faire
  - Si les effets sont hétérogènes alors chaque instrument renvoie à une sous-population différente de *compliers*
  - Ces sous-populations ne sont *a priori* pas les mêmes et la sous-population composite est encore plus mal définie
  - Cependant quand ces sous-populations sont supposées très comparables cela peut permettre de tester statistiquement la crédibilité des instruments (test de suridentification)
    - Il faut justifier cette comparabilité qui n'a rien d'évident !

Peut-on multiplier les instruments ? Faut-il le faire ?

- Ici on va s'intéresser à une raison **moins conceptuelle** et **plus statistique** d'éviter cela
- C'est aussi une raison importante de **s'assurer que l'instrument a un effet suffisamment fort sur l'intervention** qui nous intéresse

D'où vient le problème ?

- L'estimande de Wald a la forme d'un r**apport entre les coefficients de deux régressions**
- Pour Angrist and Krueger (1991)
  - Celle du salaire sur le trimestre de naissance
  - Celle de l'éducation sur le trimestre de naissance
- **Il faut que le dénominateur soit différent de 0 !**

D'où vient le problème ?

- On n'a accès qu'à des estimations du dénominateur !
  - On ne voit jamais de vrai 0 quand on estime à partir de données en quantité finie
- Quand faut-il se méfier ?
  - Dés lors que le dénominateur estimé et l'incertitude sur l'estimation sont du même ordre de grandeur
  - Cela génère de très grandes instabilités dans l'estimation
- On parle d'**instrument faible**
- Avec un seul instrument faible on arrive en général à repérer qu'il y a un problème
  - Ce n'est pas du tout évident quand on a de nombreux instruments faibles qui donnent l'illusion de stabilité et de précision

Le résultat théorique

- L'usage d'un grand nombre d'instruments faibles **n'est pas informatif sur les effets causaux moyens** de l'intervention et équivaut à une **régression linéaire par les MCO**
- Résultat pas du tout intuitif $\longrightarrow$ **petite simulation** pour se convaincre

Simuler les instruments faibles : on crée de nombreux instruments et des variables latentes indépendantes les unes des autres

```
GenerateDataWeakInstruments_LatVar<-function(nobs,
                                             nbinstruments,
                                             seed){
  set.seed(seed = seed)
  dat<-
    data.table(matrix(c(rbinom(nobs*nbinstruments, size=1, prob=0.5),
                        rnorm(nobs, mean=0, sd=0.5),
                        rnorm(nobs, mean=0, sd=1),
                        rnorm(nobs, mean=0, sd=0.5)),
                      nobs))

  colnames(dat)<-c(paste0("Z", 1:nbinstruments), "u", "v", "w")

  dat

  }
```

Simuler les instruments faibles : on crée les variables observées (intervention et variable d'intérêt)

```
GenerateDataWeakInstruments_ObsVar<-function(nobs,
                                             nbinstruments,
                                             seed,
                                             effetinstrument,
                                             effetintervention){

  dat<-GenerateDataWeakInstruments_LatVar(nobs=nobs,
                                          nbinstruments=nbinstruments,
                                          seed=seed)
  dat[, D:=eval(parse(text=paste0("1+", effetinstrument, "*(",
                                  paste(paste0("Z", 1:nbinstruments),
                                        collapse = "+"),
                                  ") + u + v")))]
  dat[, Y:=1+effetintervention*D+3*v+w]

}
```

Simuler les instruments faibles : on fait l'estimation sur les données simulées

```
ComputeEstimatesWeakInstrumentsSimulation<-function(dat,
                                                    nbinstruments){
        OLS_estimate<-lm(Y~D, data=dat)$coefficient["D"]
        TwoSLS_estimate<-ivreg(as.formula(paste0("Y ~ D |",
                                                 paste(paste0("Z",
                                                        1:nbinstruments),
                                                       collapse="+"))),
                               data=dat)$coefficients["D"]

        estimates<-as.data.table(matrix(c(OLS_estimate,
                                          TwoSLS_estimate),
                                        1))
        colnames(estimates)<-c("OLS", "TwoSLS")

        return(estimates)
}
```

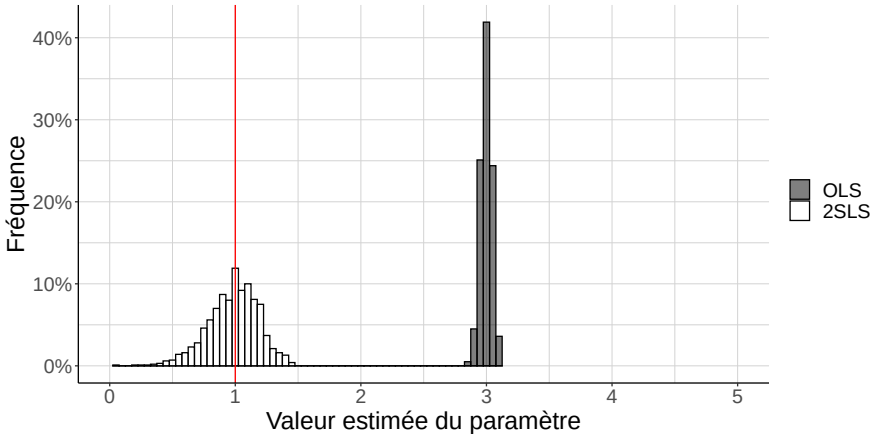
Simuler les instruments faibles : la fonction qui génère une simulation entière

```
WeakInstrumentsSimulation<-function(nobs,
                                    nbinstruments,
                                    seed,
                                    effetinstrument,
                                    effetintervention){
  simuldat<-
    GenerateDataWeakInstruments_ObsVar(nobs = nobs,
                                       nbinstruments = nbinstruments,
                                       seed =seed,
                                       effetinstrument = effetinstrument,
                                       effetintervention = effetintervention)

  estimates<-
    ComputeEstimatesWeakInstrumentsSimulation(dat = simuldat,
                                              nbinstruments = nbinstruments)
}
```

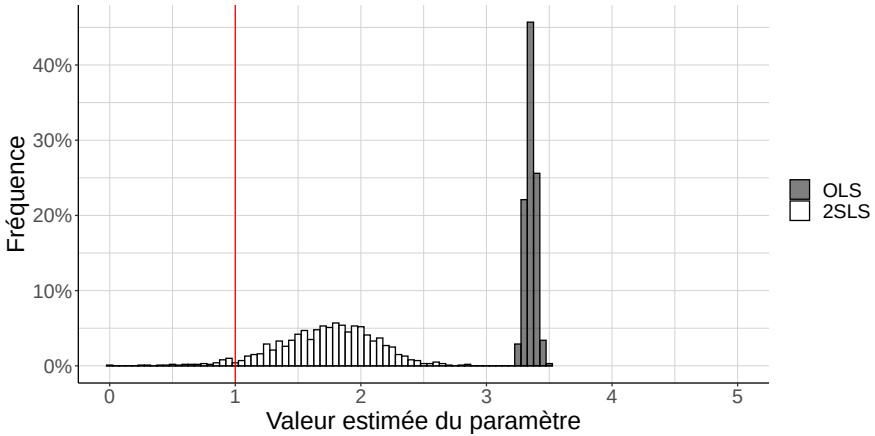
Simuler les instruments faibles : on fait tourner la simulation un grand nombre de fois

```
MonteCarloWeakInstrumentsSimulation<-function(nbobs,
                                              nbinstruments,
                                              nbsimulations,
                                              seed.init,
                                              effetinstrument,
                                              effetintervention){
  estimates<-rbindlist(lapply(1:nbsimulations,
                              FUN=function(x){
                                WeakInstrumentsSimulation(
                                  nbobs = nbobs,
                                  nbinstruments = nbinstruments,
                                  seed = seed.init + x - 1,
                                  effetinstrument =effetinstrument,
                                  effetintervention = effetintervention)
                                }),
                       idcol = "sim.index")
}
```

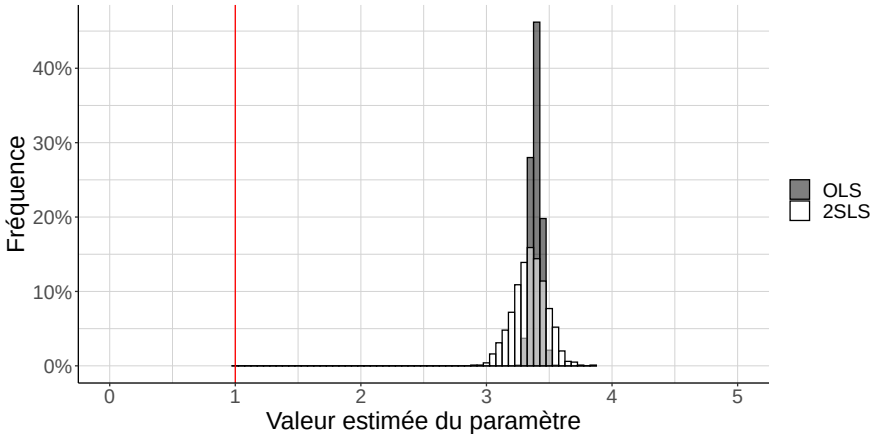
Il n'y a plus qu'à visualiser : 1000 observations, 1 instrument, effet de l'instrument = 1, effet de l'intervention = 1



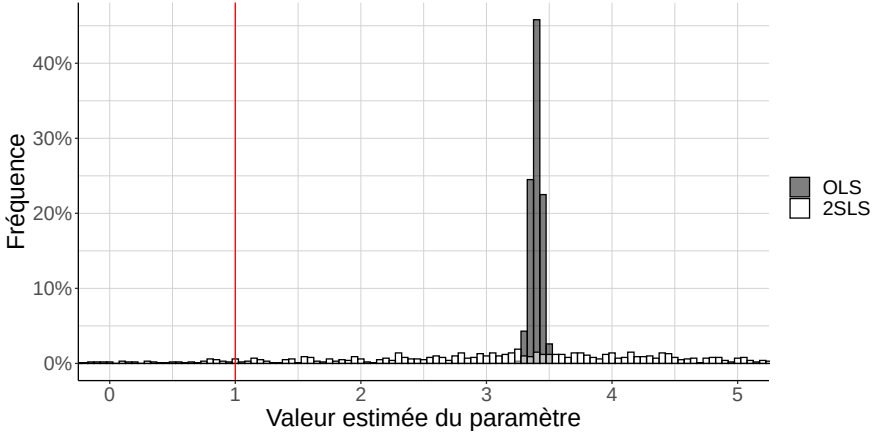
Il n'y a plus qu'à visualiser : 1000 observations, 10 instruments, effet des instruments = 0.1, effet de l'intervention = 1



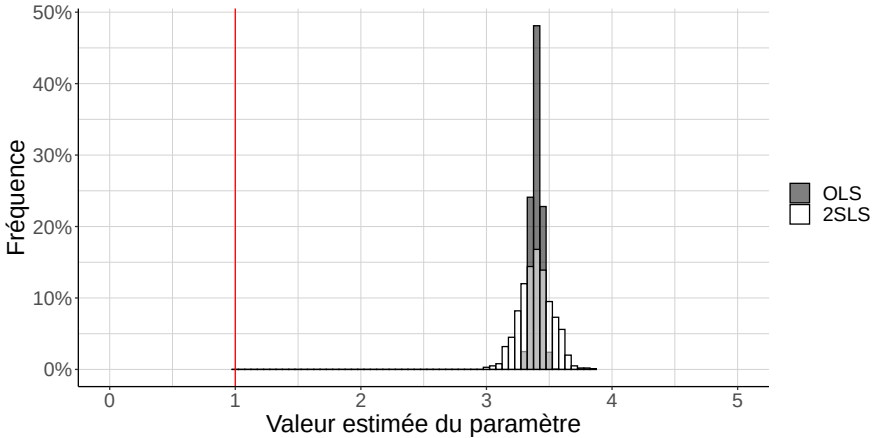
Il n'y a plus qu'à visualiser : 1000 observations, 100 instruments, effet des instruments $= 0.01$, effet de l'intervention $= 1$



Alors qu'avec un seul instrument faible l'instabilité est telle qu'on doit réussir à la repérer (1000 observations, 1 instrument, effet de l'instrument = 0, effet de l'intervention = 1)



Avec de nombreux instruments, même sans aucun rapport avec l'intervention, c'est plus difficile à voir (1000 observations, 100 instruments, effet des instrument = 0, effet de l'intervention = 1)



Quelles solutions à ce problème ?

- Les instruments faibles génèrent des résultats biaisés et les intervalles de confiance usuels renseignent mal sur l'instabilité de l'estimation
- Que faire ?
  - **Règle du pouce** : il faut considérer la F-statistique de nullité jointe des coefficients sur les instruments en première étape → elle doit être supérieure à 10
  - Course aux armements : des papiers récents argumentent en faveur de seuils plutôt à 23.1 ou 104 (!)
  - Utiliser les **intervalles de confiance à la Anderson and Rubin (1949)**, valides même en cas de faiblesse de l'instrument si on est dans le cas une variable endogène et un instrument
  - Considérer la **forme réduite**, qui n'utilise pas la division problématique

Retour sur Angrist and Krueger (1991) et Card (1993)

- Estimez la F-statistique de nullité des coefficients sur l'indicatrice d'être né le premier trimestre dans la régression de première étape
- Utilisez la fonction `ivmodel` du package `ivmodel` pour estimer les intervalles de confiance à la Anderson and Rubin (1949) sur l'exemple de Card (1993)
  - Données récupérées à partir du package `ivmodel` : `data("card.data")`
  - Il faut inclure les variables de conditionnement dans la régression : `c("exper", "expersq", "black", "south", "smsa", "reg661", "reg662", "reg663", "reg664", "reg665", "reg666", "reg667", "reg668", "smsa66")`

Calcul de la F-statistique avec `ivreg`

```
iv_estimates<-ivreg(lwage ~ educ | q1,
                    data=ak80[yob %in% 30:39])
summary(iv_estimates,
        vcov=vcovHC(iv_estimates,
                    type="HC0"),
        diagnostic=TRUE)$waldtest
```

```
## [1] 1.801339e+01 2.194165e-05 1.000000e+00 3.295070e+05
```

Calcul de la F-statistique avec `felm`

```
iv_estimates_felm<-felm(lwage ~ 1 | 0 |  (educ ~ q1) | 0,
                        data=ak80[yob %in% 30:39])

summary(iv_estimates_felm,
        vcov=vcovHC(iv_estimates_felm,
                    type="HC0"),
        diagnostic=TRUE)$P.fstat
```

```
##            p         chi2          df1          p.F            F          df2
## 2.054841e-05 1.813778e+01 1.000000e+00 2.055410e-05 1.813778e+01 3.295070e+05
## attr(,"formula")
## ~educ
## <environment: 0x000000006aec4a38>
```

Estimer les intervalles de confiance à la Anderson and Rubin (1949) avec `ivmodel` sur l'exemple de Card (1993)

```
data(card.data)
outcome<-card.data[,"lwage"]
intervention<-card.data[,"educ"]
instrument<-card.data[,"nearc4"]
conditionningvariables<-c("exper", "expersq", "black", "south", "smsa",
                          "reg661", "reg662", "reg663", "reg664", "reg665",
                          "reg666", "reg667", "reg668", "smsa66")
conditionning<-card.data[,conditionningvariables]
iv_estimates_ivmodel<-ivmodel(Y=outcome,D=intervention,
                              Z=instrument,X=conditionning)
AR.test(iv_estimates_ivmodel)$ci
```

```
##            lower     upper
## [1,] 0.02480484 0.2848236
```

L'intervalle de confiance à la Anderson and Rubin (1949) est plus large que l'intervalle de confiance usuel

```
card_iv_estimates<-ivreg(lwage ~ educ + exper + expersq + black + south +
                           smsa + reg661 + reg662 + reg663 + reg664 + reg665 +
                           reg666 + reg667 + reg668 + smsa66 | nearc4,
                         data=card.data)
```

```
## Warning in ivreg.fit(X, Y, Z, weights, offset, ...): more regressors than
## instruments
```

```
confint(card_iv_estimates,
        vcov = vcovHC(card_iv_estimates,
                      type="HC0"))["educ",]
```

```
##     2.5 %    97.5 %
## 0.1365325 0.2395927
```

Pour conclure

- Un arsenal de techniques quantitatives pour estimer les effets causaux d'une intervention sur une grandeur à laquelle on s'intéresse, pour une population bien définie
- Cadre conceptuel et méthodologique très flexible qui n'a pas lieu d'être limité à certains champs particuliers
- Il y a des difficultés statistiques sur certaines techniques mais
  - La principale difficulté reste toujours de formuler des arguments sociologiques convaincants quant à la crédibilité de l'approche que l'on utilise
  - Ces difficultés ont des solutions pratiques

La suite à court terme pour vous et pour moi

- Mise en ligne du support écrit sur les variables instrumentales : je m'y active
- Envoi d'une liste de suggestions de sujets
- N'hésitez pas à m'écrire si :
  - Vous avez la moindre question sur le contenu de l'enseignement
  - Vous souhaitez choisir un des sujets proposés, ou un autre qui vous intéresserait particulièrement

Bibliographie

Bibliographie I

Anderson, T. W., and Herman Rubin. 1949. "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations." *The Annals of Mathematical Statistics* 20 (1) : 46–63. http://www.jstor.org/stable/2236803.

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.

Angrist, Joshua D., and Alan B. Krueger. 1991. "Does Compulsory School Attendance Affect Schooling and Earnings ?" *The Quarterly Journal of Economics* 106 (4) : 979–1014.

Braun, Robert. 2022. "Bloodlines : National Border Crossings and Antisemitism in Weimar Germany." *American Sociological Review* 87 (2) : 202–36. https://doi.org/10.1177/00031224211071145.

Card, David. 1993. "Using Geographic Variation in College Proximity to Estimate the Return to Schooling." Working Paper 4483. Working Paper Series. National Bureau of Economic Research. https://doi.org/10.3386/w4483.

Bibliographie II

Chaisemartin, C de, and X D'Haultfœuille. 2017. "Fuzzy Differences-in-Differences." *The Review of Economic Studies* 85 (2) : 999–1028. https://doi.org/10.1093/restud/rdx049.

Duflo, Esther. 2001. "Schooling and Labor Market Consequences of School Construction in Indonesia : Evidence from an Unusual Policy Experiment." *American Economic Review* 91 (4) : 795–813. https://doi.org/10.1257/aer.91.4.795.

Mellon, Jonathan. 2022. "Rain, Rain, Go Away : 192 Potential Exclusion-Restriction Violations for Studies Using Weather as an Instrumental Variable." mimeo.

Pascali, Luigi. 2017. "The Wind of Change : Maritime Technology, Trade, and Economic Development." *American Economic Review* 107 (9) : 2821–54. https://doi.org/10.1257/aer.20140832.