Econometrics using STATA

Benjamin Monnery
EconomiX, Univ Paris Nanterre

M1 Economie du Droit
2017-2018

GENERAL INFO

Email : *bmonnery@parisnanterre.fr*
Office : G308A

Schedule : 30 hours

- 3-hour whiteboard classes for 7 weeks (G205)
- 2-hour computer classes on Stata during 3 weeks (G213B on 13/02 20/02 & 6/03)
- one written exam (end of March / early April)

  ↳ class **cancelled on January 30**

Class in english

Exam in french (or english) on paper

Slides will be available on my webpage

References

> Angrist & Pischke, *Mostly Harmless Econometrics*, Princeton University Press

> Cameron & Trivedi, *Microeconometrics using Stata*, Stata Press

BONUS READINGS

Bonuses are individual and non-mandatory

This week : chapters 1 & 2 of *Mostly Harmless Econometrics*
"Questions about questions" + "The experimental ideal"

> freely available on my webpage

Goal : make a 1-page critical review of the paper/chapter

- brief summary of the paper (topic, method, main points, results)
- discuss method, experimental design, interpretations, conclusions
- relate it to the class
- criticisms, shortcomings ?

... bonus based on quality/clarity/concision

Send PDF by email before next tuesday (noon)
at *bmonnery@parisnanterre.fr*

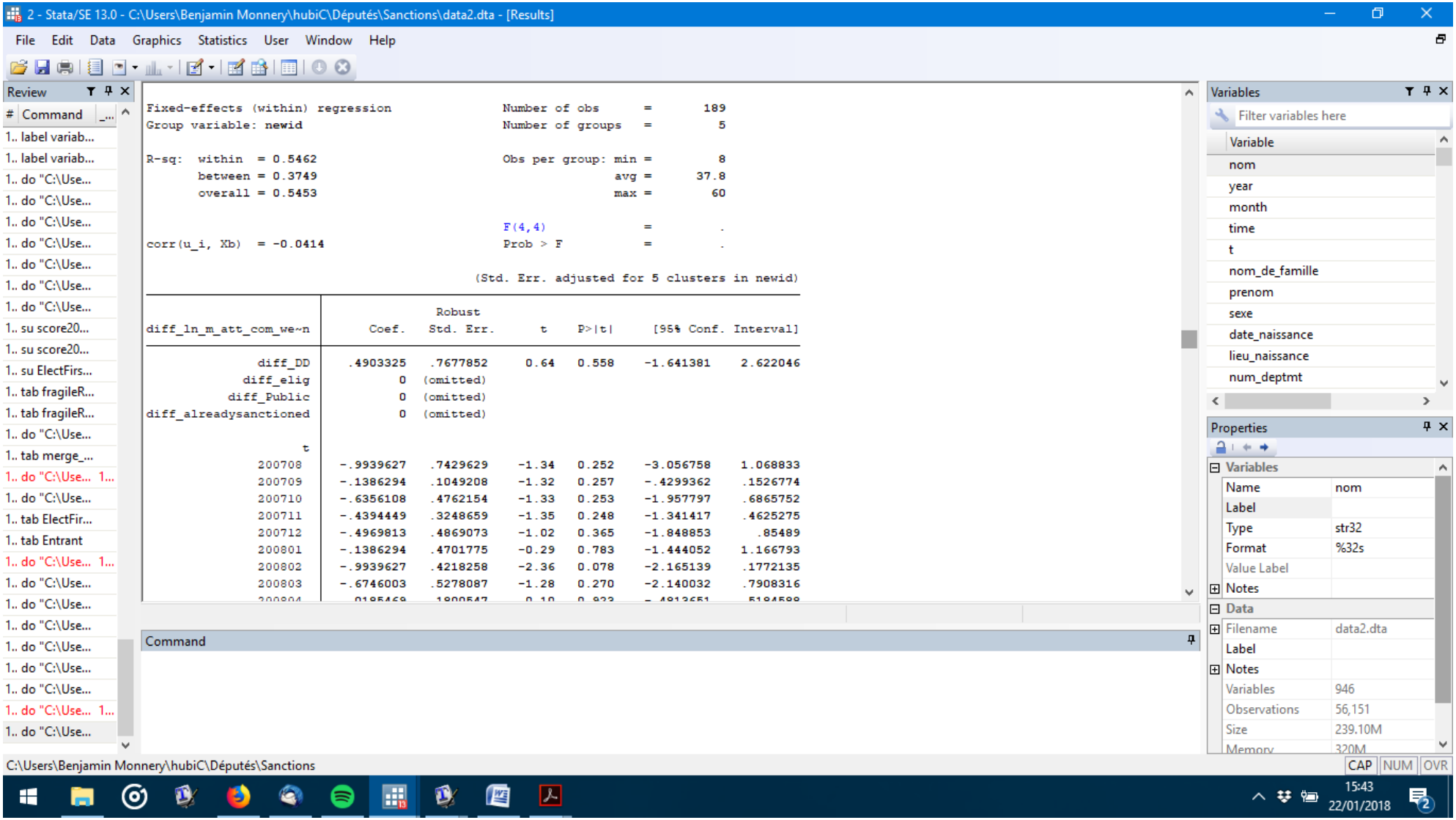
CONTENT

Econometrics using STATA

- Use econometric tools to answer policy-relevant questions
  ... of the cause-and-effect type ($\frac{\partial E[y|x]}{\partial x}$)
  ... not predictive modelling ($E[y|x]$)

- How to answer "theoretically" or intuitively
  what "model", what estimator/technique, what *design*

- How to answer operationally on STATA
  the most popular software among (micro)econometricians
  (others include *R*, *SAS*, *SPSS*...)


2 - Stata/SE 13.0 - C:\Users\Benjamin Monnery\hubiC\Députés\Sanctions\data2.dta - [Results]
File Edit Data Graphics Statistics User Window Help
Review
Command
1.. label variab...
1.. label variab...
1.. do "C:\Use...
1.. do "C:\Use...
1.. do "C:\Use...
1.. do "C:\Use...
1.. do "C:\Use...
1.. do "C:\Use...
1.. do "C:\Use...
1.. do "C:\Use...
1.. su score20...
1.. su score20...
1.. su ElectFirs...
1.. tab fragileR...
1.. tab fragileR...
1.. do "C:\Use...
1.. tab merge_...
1.. do "C:\Use... 1...
1.. do "C:\Use...
1.. tab ElectFir...
1.. tab Entrant
1.. do "C:\Use... 1...
1.. do "C:\Use...
1.. do "C:\Use...
1.. do "C:\Use...
1.. do "C:\Use...
1.. do "C:\Use...
1.. do "C:\Use...
1.. do "C:\Use...
1.. do "C:\Use...
1.. do "C:\Use... 1...
1.. do "C:\Use...
Fixed-effects (within) regression Number of obs = 189
Group variable: newid Number of groups = 5
R-sq: within = 0.5462 Obs per group: min = 8
between = 0.3749 avg = 37.8
overall = 0.5453 max = 60
F(4,4) = .
corr(u_i, Xb) = -0.0414 Prob > F = .
(Std. Err. adjusted for 5 clusters in newid)
diff_ln_m_att_com_we~n | Coef. Robust Std. Err. t P>|t| [95% Conf. Interval]
diff_DD .4903325 .7677852 0.64 0.558 -1.641381 2.622046
diff_elig 0 (omitted)
diff_Public 0 (omitted)
diff_alreadysanctioned 0 (omitted)
t
200708 -.9939627 .7429629 -1.34 0.252 -3.056758 1.068833
200709 -.1386294 .1049208 -1.32 0.257 -.4299362 .1526774
200710 -.6356108 .4762154 -1.33 0.253 -1.957797 .6865752
200711 -.4394449 .3248659 -1.35 0.248 -1.341417 .4625275
200712 -.4969813 .4869073 -1.02 0.365 -1.848853 .85489
200801 -.1386294 .4701775 -0.29 0.783 -1.444052 1.166793
200802 -.9939627 .4218258 -2.36 0.078 -2.165139 .1772135
200803 -.6746003 .5278087 -1.28 0.270 -2.140032 .7908316
Command
Variables
Filter variables here
Variable
nom
year
month
time
t
nom_de_famille
prenom
sexe
date_naissance
lieu_naissance
num_deptmt
Properties
Variables
Name nom
Label
Type str32
Format %32s
Value Label
Notes
Data
Filename data2.dta
Label
Notes
Variables 946
Observations 56,151
Size 239.10M
Memory 320M
C:\Users\Benjamin Monnery\hubiC\Députés\Sanctions
CAP NUM OVR
15:43
22/01/2018

CONTENT

Goal of this course :
Learn how to answer empirical questions of the form "*what's the effect of ... on ... ?*" (causal effect identification) using econometrics

- first on paper : illustrated with real-life examples and studies
- then on Stata : how to answer by yourself, replicate studies

With a strong emphasis on

- law & economics topics
- endogeneity bias or selection
- actual practice (how research is done)

... and less so on more theoretical aspects (econometric theory)

Requirement : basics of Econometrics

Introduction

Most empirical questions lead to similar methodological challenges :

- Data availability
- Bias of β
- Statistical inference

Examples :

- What is the effect of gun ownership on murder rates ?
- What is the effect of legalizing marijuana on drug consumption ?
- What is the effect of incarceration on future crime ?
- What is the effect of some judicial reform on judicial outcomes ?

EVALUATIONS

Public policies employ costly resources to reach social goals

Ex : the CICE tax cut in France
↳ *boost employment, increase competitivity*

Ex : cutting class size by two in CP
↳ *improve learning*

Types of questions :

- Was the policy effective in achieving its goals ?
- Did the policy had unintented consequences (positive and negative) ?
- Was the policy cost-effective (Benefit $>$ Cost) ?
- Was the policy more efficient than oths (large return per euro $\frac{B}{C}$) ?

Answering those questions is called ex-post evaluations
↳ key parameter is the causal effect of the treatment

OTHER TYPES OF EVALUATION

Ex-post evaluation can complement -but is not synonym for- other types of evaluations :

- Audit : was the policy implemented as planned ?
  ↳ ... in terms of target population, take-up, cost, etc.
  $\Rightarrow$ project management, accounting (not this course)

- Ex-ante evaluation : what effects can we expect if implemented ?
  ↳ more difficult, more uncertain (stronger assumptions)
  $\Rightarrow$ structural econometrics (not this course)

STEPS TO ANSWER EMPIRICAL QUESTIONS

- Define your research question precisely
  - what parameter β do you want to estimate ?
  - is this estimate useful for citizens, policymakers, other researchers, your company, your customer ?
- Anticipate the key methodological challenges
- Find the appropriate data
- Adress the challenges and answer your question
- Remain cautious about your findings
  - internal validity : is my estimate really robust, unbiased ?
  - external validity : is my estimate relevant for other contexts, countries, periods... ?
  - other relevant aspects that you missed, did not measure ?

The Problem(s) With Causality

CORRELATION VS CAUSALITY

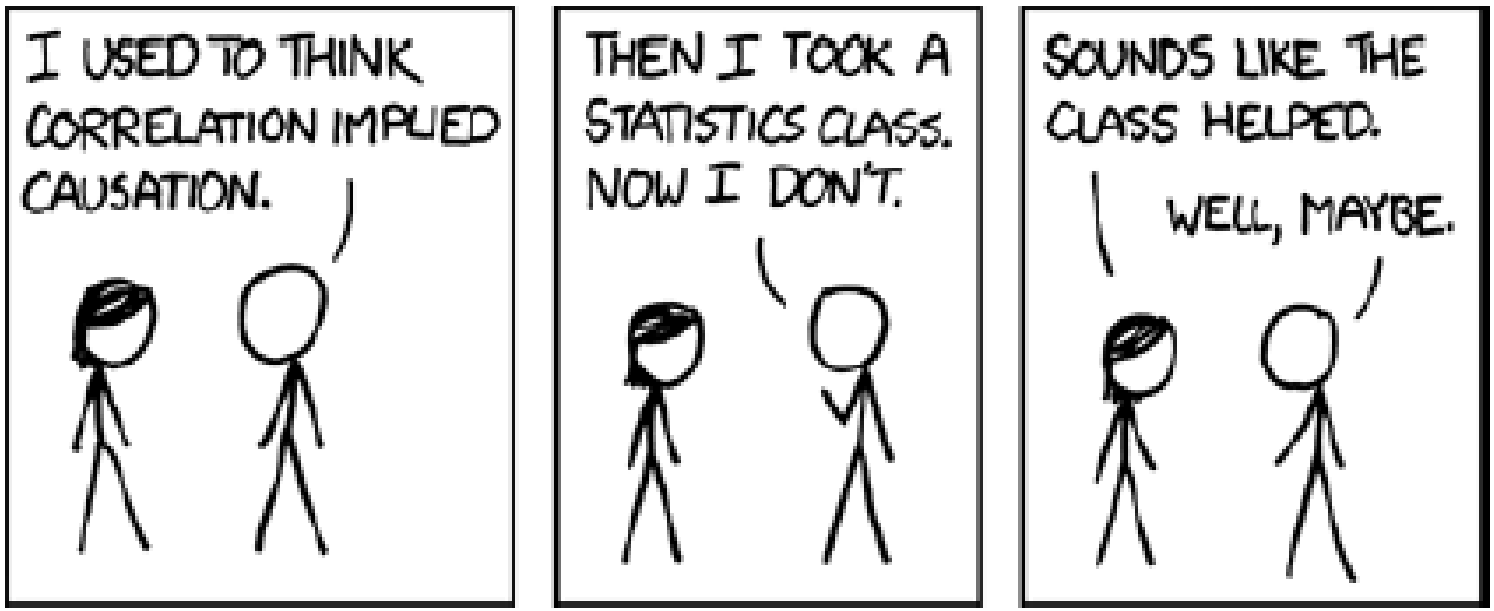


Figure: xkcd

> is the treatment responsible for the changes that occured ?

CORRELATION VS CAUSALITY

A statistical correlation between two variables, $Corr(X, Y) \neq 0$, can imply many things :

- Causality : $X \rightarrow Y$
- Reverse causality : $X \leftarrow Y$
- Simultaneity : $X \leftrightarrow Y$
- Omitted variable : $Z \rightarrow X$ and $Z \rightarrow Y$
- Spurious correlation (by chance)

$\Rightarrow$ Causal effect identification : exclude all other possibilities

MEDIATION AND MECHANISMS

A causal effect between X and Y doesn't mean that no other variables play a role :

- Other causes : $X \rightarrow Y \leftarrow C$
- Mediators : $X \rightarrow M \rightarrow Y$
  ↳ mediators inform on the mechanisms explaining causal effect

Ex : the causal effect of income (X) on life expectancy (Y)

$Income \rightarrow LifeExp \leftarrow Genetics$

$Income \rightarrow Sport \rightarrow LifeExp$

CAUSALITY AND COUNTERFACTUALS

The quest for causality requires to answer :

What would have happen to Y had the treatment X not occured ?

$\Rightarrow$ refers to some unobservable situation called "counterfactual" (or "potential outcome")

$\Rightarrow$ Causal effect identification is the art or craft of choosing or constructing credible counterfactuals (credible estimates of what would have happened instead)...

... using experimental or quasi-experimental data

CAUSALITY AND COUNTERFACTUALS

Imagine you observe two groups of people

- treated group with $T = 1$
- control group with $T = 0$ (not treated for some reason)

and you want to measure the impact of T on an outcome Y

Easy solution = Outcome of Treated − Outcome of Controls

= Outcome of Treated − Outcome of Counterfactuals
+ Outcome of Counterfactuals − Outcome of Controls

= True causal effect β + Bias due to bad controls

$\Rightarrow$ Using bad controls as counterfactuals is called selection bias
↳ leads to overestimate or underestimate the true effect β

Selection bias is fundamentally a problem of endogeneity of T

$$Y_i = \beta_0 + \beta_1 T_i + \epsilon_i$$

In Stata, you type `regress Y T`

and get $Y_i = 2.1 + 4.5 T_i + u_i$

Key assumption : $Cov(T, \epsilon) = 0$ (exogeneity)

There is endogeneity if the treatment variable T is correlated with Y through unobserved variables captured in ϵ

$\Rightarrow Cov(T, \epsilon) \neq 0$ (endogeneity)

$\Rightarrow$ when T refers to some "treatment" (categorical) or "treatment intensity" (continuous), we usually call this problem selection bias or omitted variable bias

THE RUBIN CAUSAL MODEL

Donald Rubin (statistician at Harvard) helped formalize the problem of causal identification using simple notations

↳ *Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies*, Journal of Educational Psychology, 1974.

Example : the effect of hospitalization on health

- Hospitalization is our treatment (T) : $Hospit = \{0, 1\}$
  - ↳ some people are hospitalized, most others aren't
- Reported health is our outcome (Y) : $Health = \{0, 1, 2, 3, 4, 5\}$

THE RUBIN CAUSAL MODEL

Counterfactual : each individual has his own counterfactual (potential) outcome that can't be observed :

$$\text{Counterfactual}_i = \begin{cases} Y_i^1 & \text{if } Hospit_i = 1 \quad \text{(health of } i \text{ if hospitalized)} \\ Y_i^0 & \text{if } Hospit_i = 0 \quad \text{(health of } i \text{ if not hospitalized)} \end{cases}$$

- Individual Treatment Effect : $\delta_i = Y_i^1 - Y_i^0$
- Average Treatment Effect (ATE) : $E[\delta_i] = E[Y_i^1] - E[Y_i^0]$

- Fundamental problem of causal identification :

  It is impossible to observe both Y^1 and Y^0 for the same individual and so individual causal effects δ_i and average effects $E[\delta_i]$ are unknowable

HETEROGENEITY IN THE RUBIN MODEL

- Average Treatment Effect on the Treated (ATT or ATET) :
  $E[\delta_i | T = 1] = E[Y_i^1 | T = 1] - E[Y_i^0 | T = 1]$
- Average Treatment Effect on the Unteated (ATU) :
  $E[\delta_i | T = 0] = E[Y_i^1 | T = 0] - E[Y_i^0 | T = 0]$
- Local Average Treatment Effect (LATE) :
  $E[\delta_i | i \in G] = E[Y_i^1 - Y_i^0 | i \in G]$ (G is a certain subpop)

$\Rightarrow$ If everybody reacts exactly similarly to treatment, then $ATT = ATU = ATE$ (homogeneous treatment effect)

↳ very unlikely regarding hospitalization : $ATT > 0, ATU = 0$

$\Rightarrow$ If treatment effect are heterogeneous in the population, then $ATT \neq ATU \neq ATE$

↳ the effect on "actually treated individuals" does not predict what would happen to other people

$\rightarrow$ most methods only provide ATT or $LATE$

SELECTION BIAS IN THE RUBIN MODEL

Since we don't observe the same individuals in the two conditions ($T = 1; T = 0$), we can only compute estimates of the form :

$$\begin{aligned} E[Y^1|T=1] - E[Y^0|T=0] &= E[Y^1|T=1] - E[Y^0|T=1] \\ &\quad + E[Y^0|T=1] - E[Y^0|T=0] \\ &= \text{ATT} + \text{Selection bias} \end{aligned}$$

For hospitalization :
Comparing mean health of hospitalized people $E[Y^1|T=1]$ and mean health of people not in hospital $E[Y^0|T=0]$ mixes :

- the true effect (ATT > 0)
- the fact that people in hospital are sick ($E[Y^0|T=1] - E[Y^0|T=0] << 0$)

$\Rightarrow$ This estimator is biased negatively : underestimates the positive effect of hospitalization

When facing a question of the form "*what's the effect of ... on ... ?*", first reaction should be :

- Can I suspect selection bias ? (i.e. are simple group comparisons bad estimators ?)
- Can I anticipate the sign of the bias ? (positive, negative, unknown ?)
- Can I anticipate the magnitude of the bias ? (large, very small ?)
- How can I control for the bias ?

AN EXAMPLE WITH PRISONERS AND RECIDIVISM

Le Monde WEEK-END

Samedi 15 octobre 2011 - 67e année - N°20756 - www.lemonde.fr Fondateur : Hubert Beuve-Méry - Directeur : Erik Izraelewicz

Comment les prisons françaises fabriquent de la récidive

- Près de 60 % des sortants sont recondamnés dans les cinq ans
- Reportage dans un quartier d'Orléans avec des familles de détenus

La récidive est à la fois le tourment et l'obsession de la majorité : sept lois ont été votées depuis 2004, visant toutes à durcir les peines ; Eric Ciotti, le député UMP de Nice et bras armé du chef de l'Etat à l'Assemblée, ne jure d'ailleurs que sur *« le caractère dissuasif de la sanction »*.

La récidive est pourtant un phénomène assez mal connu, et il est douteux que l'alourdissement des peines puisse la réduire. Une passionnante étude de la direction de l'administration pénitentiaire, passée assez inaperçue cet été dans les *Cahiers d'études pénitentiaires et criminologiques*, vient utilement recadrer le débat, et indirectement proposer des solutions.

Le chiffre, d'abord, est énorme : 59 % des détenus sont de nouveau condamnés dans les cinq ans qui suivent leur libération, et 46 % d'entre eux à de la prison ferme.

FRANCK JOHANNÈS

▶ Lire la suite page 10

10 | SOCIÉTÉ

Le Monde
Samedi 15 octobre 2011

L'absence d'aménagement de peine aggrave le risque de récidive des sortants de prison

Selon une étude inédite de l'administration pénitentiaire, 60 % des détenus sont condamnés dans les cinq ans qui suivent leur libération

►►► Suite de la première page

Les mineurs sont les plus exposés à la récidive, surtout dans les deux premières années de liberté, mais l'aménagement des peines et la liberté conditionnelle font chuter les taux dans des proportions spectaculaires : pour éviter la récidive, mieux vaut préparer la sortie que condamner lourdement.

L'étude que publient les démographes Annie Kensey et Abdelmalik Benaouda, du bureau des études et de la prospective de l'administration pénitentiaire, est l'une des plus complètes qui soient : 7 000 dossiers de détenus libérés entre juin et décembre 2002 ont été comparés cinq ans plus tard, c'est-à-dire dans les années 2007-2008, à leur casier judiciaire.

Il ne s'agit pas de la récidive légale, qui ne s'intéresse qu'aux condamnations pour une même infraction ou une même famille d'infractions, mais du *« devenir judiciaire d'anciens condamnés »*, c'est-à-dire de la récidive quel que soit le motif de la nouvelle condamnation.

La récidive n'est évidemment pas la même selon la nature de l'infraction initiale. Les voleurs sont 74 % à être à nouveau condamnés cinq ans plus tard, les violeurs d'enfants 19 %, et encore, pas pour ce crime : ils ne sont que 0,6 % à être condamnés à de la réclusion criminelle. 32 % des meurtriers sont à nouveau condamnés, mais pour 19 % à de la prison ferme, et 0,7 % seulement pour un nouveau crime – et pas forcément un meurtre : il n'y a guère, en France, de tueurs en série. Avoir été condamné pour homicide volontaire diminue ainsi de moitié le risque de recondamnation ou de retour en prison par rapport aux voleurs ou aux receleurs. Les condamnés pour viol ou agression sexuelle ont une probabilité trois fois moindre d'avoir une nouvelle condamnation dans les cinq ans que les condamnés pour vols.

> Les récidivistes sont plutôt les condamnés pour les délits les moins graves

Les récidivistes sont plutôt les condamnés pour les délits les moins graves. Les détenus condamnés à des peines de moins de douze mois sont 61 % à récidiver cinq ans plus tard, les condamnés à cinq ans et plus sont 33 % à récidiver.

Plus on a été condamné, plus on récidive : les libérés qui avaient déjà une condamnation antérieure avant d'être incarcérés en 2002 sont 34 % à recommencer. Ceux qui avaient deux condamnations ou plus sont 70 % : plus on a fait de prison, plus on en fera. Ce n'est pas une surprise, les hommes, plus délinquants, sont aussi plus récidivistes que les femmes. La probabilité de recondamnation est deux fois plus faible pour les femmes que pour les hommes.

En revanche, le risque est trois fois plus important pour les mineurs à la libération que pour les jeunes majeurs de moins de 30 ans. Ne pas être marié multiplie même par 1,5 le risque de retourner en prison, les détenus chômeurs récidivent à hauteur de 61 %, ceux qui ont un emploi à 55 %. Les populations à risque sont bien les mineurs. 78 % des mineurs ont de nouveaux ennuis avec la justice dans les cinq ans, les plus de 50 ans sont, eux, 29 %.

Statistiquement, la récidive est plus forte dans les premiers mois après la sortie : plus de la moitié des récidivistes (54,6 %) ont été à nouveau condamnés au cours de la première année de leur sortie, les trois quarts dans les deux ans. C'est encore plus vrai pour les condamnés à la prison ferme : le taux de récidive est de 62 % la première année, 81 % dans les deux ans. Après la quatrième année de liberté, la courbe de récidive se tasse. L'urgence est donc bien d'accompagner le mineur à la sortie de prison et pendant les deux années qui suivent, sinon il rechute.

La variable la plus intéressante et la plus encourageante est sans doute le mode d'exécution de la peine : plus les condamnés restent enfermés, plus ils récidivent en sortant. *« Les risques de recondamnation des libérés n'ayant bénéficié d'aucun aménagement de peine demeurent 1,6 fois plus élevés que ceux des bénéficiaires d'une libération conditionnelle »*, notent les démographes.

Les libérés qui n'ont pas bénéficié d'aménagements de peine ont été 63 % à être recondamnés au bout de cinq ans (contre 39 % pour les sortants en libération conditionnelle). *« Il y a effectivement des populations plus fragiles,* indique Annie Kensey, *des personnes qu'il faut accompagner, c'est tout l'intérêt de l'individualisation des peines et du suivi des conseillers d'insertion et de probation. »* ■

FRANCK JOHANNÈS

Les mineurs sont les plus exposés à la récidive

TAUX DE RÉCIDIVE DES LIBÉRÉS DE 2002 AU BOUT DE CINQ ANS SELON LEUR SITUATION, en %

SOURCE : ADMINISTRATION PÉNITENTIAIRE

Hidden assumption :

$$E[Recid^{NoPrison}|Prison = 1] - E[Recid^{NoPrison}|Prison = 0] = 0$$

Regarding prison and recidivism, it is very likely that :

- There is selection bias :
  $E[Recid^{NoPrison}|Prison = 1] - E[Recid^{NoPrison}|Prison = 0] \neq 0$
- Selection bias is *positive* :
  $E[Recid^{NoPrison}|Prison = 1] > E[Recid^{NoPrison}|Prison = 0]$

$\Rightarrow$ offenders sent to prison by judges are intrinsically more crime-prone than offenders not sent to prison

$\Rightarrow$ comparing recidivism rates overestimates the criminogenic effect of incarceration

- How large is selection bias ? : depends on the precise context governing treatment assignment :
  - assignment is blind, constrained (small to no bias)
  - assignment is discretionary, subjective, based on rich, qualitative info (possibly large bias)

$\Rightarrow$ plausible that the true effect of incarceration may actually be beneficial, not criminogenic

KEY ASSUMPTION : SUTVA

SUTVA refers to a key assumption in most evaluations, whatever the method used

$\Rightarrow$ *Stable Unit Treatment-Value Assumption*

Incorporates two assumptions :

- there is only one treatment : same type, intensity for everybody
- assignment of other people to treatment does not affect *i*'s potential outcomes

  there is no contagion of treatment effects from treated to untreated people (no peer effects) orgeneral equilibrum effects

  probably ok for hospitalization (though treatment quality can vary)

  probably no ok for a large cash transfer program in a village

There are solutions to avoid SUTVA when it doesn't hold (e.g. look at more aggregate effects, compare with non-contaminated people...) but we assume SUTVA always holds in this course

UNIDENTIFIED QUESTIONS

Some cause-and-effect questions are fundamentally unidentified questions (FUQ in *Angrist&Pischke*)

FUQ relate to situations where one variable ("treatment") can not be manipulated or manipulated alone. Examples :

1. race, gender, age... (in discrimination studies)
2. time spent in treatment and aging

Imperfect solutions :

1. discrimination : testing method (make people react to fake profiles, one white, one black)
2. try to control statistically for the influence of the other mechanism (like aging)

Solutions for Identification of Causal Effects

FIXING SELECTION BIAS

The solution is to find treated and untreated people who have similar potential outcomes

i.e. people with $E[Y^0|T=1] - E[Y^0|T=0] = 0$ (no selection bias)

For hospitalization :

Find similar people in terms of initial health and other determinants of health (wealth...) where :

- some are hospitalized
- some are not
- for exogeneous reasons (good luck/bad luck)

  ↳ random assignment, beds available, strike of nurses, etc.

MAIN METHODS TO FIX SELECTION BIAS

1. Randomized Experiments (aka field experiments, clinical trials, randomized controlled experiments RCT)
   - ↳ make the exog. assumption $Cov(Hospit, \epsilon) = 0$ hold *by design*

2. Covariate-adjustment : control statistically for ex-ante differences
   - ↳ make exog. assumption more credible with more X

   $Cov(Hospit, \epsilon | PreviousHealth) = 0$

3. Matching : compare matched/twin individuals (one treated, one untreated) with similar objective proba of treatment

   $Cov(Hospit, \epsilon | Pr(Hospit)) = 0$

Methods 2 and 3 only work with selection on observables

MAIN METHODS TO FIX SELECTION BIAS

In many cases (law & econ), covariate adjustment or matching is not enough and RCT is not possible in practice (ethical issues)

4. Panel data methods : Diff-in-Diff, fixed effects, event studies, synthetic control...
   - ↳ follow the same individuals or similar cohorts over time
5. Instrumental Variable methods : find a credible Z that affects T but not directly Y
   - ↳ exploit arbitrary specificities of the context / legislation...
6. Regression Discontinuity Designs : exploit sharp or fuzzy cutoffs in treatment assignment

Methods 4, 5, 6 also work with selection on unobservables (participation is discretionary, based on unobserved variables)

Main caveat : often identify local, group-specific effects

Randomized Experiments

RANDOMIZED EXPRIMENTS

Randomized experiments (or RCT) are the gold standard of policy evaluations

- massively used in medicine : clinical / pharmaceutical trials
- now often used in development economics
- not so much in law & economics...
- great internal validity

- 2 main critiques : weak external validity ?
- Hawthorne effect ?

Random assignment to treatment implies by construction that treated and controls are similar on average :

$$E[Y^0|T=1] = E[Y^0|T=0] \quad \text{or} \quad Cov(T,\epsilon) = 0$$

IMPLEMENTATION OF RCTS

Best case scenario :

- there is perfect enforcement of the randomization
- all those drawn (D=1) are effectively treated (T=1) ; all those not drawn (D=0= are not treated (T=0)
- and SUTVA is credible

$\Rightarrow$ Then, comparing mean group outcomes gives the causal effect :

$$\frac{1}{N}\sum(Y^{T=1} - Y^{T=0}) = ATT = ATU = ATE$$

IMPLEMENTATION OF RCTS

Usual case :

- the randomization is not perfectly followed, either because field operators have discretion or because participants can opt out/in
- potential bias from selection or self-selection after randomization

$$\frac{1}{N}\sum(Y^{T=1} - Y^{T=0}) \neq ATE$$

- here, what you can get with certainty is the average causal effect of the "intention-to-treat" (ITT effect)

$$\text{ITT} = \frac{1}{N}\sum(Y^{D=1} - Y^{D=0})$$

- You can also get the ATT by instrumenting T with D (IV)

  by Two-Stage-Least-Squares (2SLS) :

$$T_i = \alpha_0 + \alpha_1 D_i + \alpha X + e_i$$
$$Y_i = \beta_0 + \beta_1 \hat{T}_i + \beta X + u_i$$

IMPLEMENTATION OF RCTS

Imbens (2009) : *... in a situation where one has control over the assignment mechanism, there is little to gain, and much to lose, by giving that up through allowing individuals to choose their own treatment regime. Randomization ensures exogeneity of key variables, where in a corresponding observational study one would have to worry about their endogeneity.*

When individuals can self-select (or be selected) after randomization, the RCT is still very useful but data analysis requires precaution and regression techniques (no simple mean comparisons)

In any case, need to check balancing of key variables across the two groups : in expectations, treated and controls should be similar on each observed variable

BONUS 2

Bonuses are individual and non-mandatory

This week : Killias et al. (2010) *How damaging is imprisonment in the long-term ? A controlled experiment comparing long-term effects of community service and short custodial sentences on re-offending and social integration*

> available on my webpage

Goal : make a 1-page critical review of the paper/chapter

- brief summary of the paper (topic, method, main points, results)
- discuss method, experimental design, interpretations, conclusions
- relate it to the class
- criticisms, shortcomings ?

... bonus based on quality/clarity/concision

Send PDF by email before next monday (noon)
at *bmonnery@parisnanterre.fr*

FROM EXPERIMENTS TO REGRESSIONS

Even perfectly enforced RCTs can / should be analysed using econometrics. Why ?

1. natural link between “potential outcomes” and regressions
2. regressions can easily include covariates
   - to account for the design of the experiment (stratification)
   - to make sub-group analysis, interaction effects
   - to increase precision of the causal estimate
3. regressions can deal with inference issues
   - heteroscedasticity, autocorrelation of errors...

FROM EXPERIMENTS TO REGRESSIONS

Remember the notation of counterfactuals...

$$Y_i = \begin{cases} Y_i^1 & \text{if } T_i = 1 \\ Y_i^0 & \text{if } T_i = 0 \end{cases}$$

... can be rewritten as the switching equation :

$$\begin{aligned} Y_i &= T_i Y_i^1 + (1 - T_i) Y_i^0 \\ Y_i &= Y_i^0 + T_i Y_i^1 - T_i Y_1^0 \\ Y_i &= Y_i^0 + (Y_i^1 - Y_i^0) T_i \\ Y_i &= E[Y_i^0] + (Y_i^1 - Y_i^0) T_i + Y_i^0 - E[Y_i^0] \\ Y_i &= \quad \alpha \quad + \quad \beta T_i \quad + \quad \epsilon_i \end{aligned}$$

with α the baseline outcome (without treatment), β the causal ATE, and ϵ the random part of Y_1^0

FROM EXPERIMENTS TO REGRESSIONS

Regressions can include covariates (independent variables, control variables) :

$$Y_i = \alpha + \beta T_i + \theta X + \epsilon_i$$

Including covariates X can be useful :

- to account for the stratification of the experiment : conditional random assignment (some units / groups / villages have more probability to participate)
- to increase precision, i.e. reduce the variance of ϵ and thus the standard errors of β
- to make sub-group analysis, look at interaction effects

x all covariates should be “predetermined” (not posterior to treatment, not possible mediators)

Example : in the STAR experiment on class size and test scores, you can control for pupils’ sociodemographic background but not for the number of incidents during class (a potential mediator)

EXAMPLE : THE STAR EXPERIMENT

(heavily borrows from Scott Cunningham)

Krueger (1999 AER) analyzes a randomized experiment to determine the causal effect of class size on student achievement

Experiment : Tennessee Student/Teacher Achievement Ratio (STAR) experiment in the 1980s

- 11,600 students and their teachers were randomly assigned to one of the following three groups :
  - Small class of 13-17 students
  - Regular class of 22-25 students
  - Regular class of 22-25 students with a full-time teacher's aide
- After the assignment, the design called for students to remain in the same class type for four years
- Randomization occurred within schools (and difference btw urban & rural areas)

Krueger estimates the following regression :

$$Y_{ics} = \beta_0 + \beta_1 SMALL_{cs} + \beta_2 AIDE_{cs} + \alpha_s + \epsilon_{ics}$$

where :

- i is the pupil, c his class, s his school
- Y is pupils' test score at end of the year
- SMALL takes 1 for the small-class group and 0 otherwise
- AIDE takes 1 for the regular-class group with aide and 0 otherwise
- baseline category : regular-class without aide
- α_s are school fixed effects (intercepts)

Regression results from OLS on kindergarden test scores (percentiles) :

Explanatory variable	OLS: actual class size			
	(1)	(2)	(3)	(4)
			A. Kindergarten	
Small class	4.82	5.37	5.36	5.37
	(2.19)	(1.26)	(1.21)	(1.19)
Regular/aide class	.12	.29	.53	.31
	(2.23)	(1.13)	(1.09)	(1.07)
White/Asian (1 = yes	—	—	8.35	8.44
			(1.35)	(1.36)
Girl (1 = yes)	—	—	4.48	4.39
			(.63)	(.63)
Free lunch (1 = yes)	—	—	−13.15	−13.07
			(.77)	(.77)
White teacher	—	—	—	−.57
				(2.10)
Teacher experience	—	—	—	.26
				(.10)
Master's degree	—	—	—	−.51
				(1.06)
School fixed effects	No	Yes	Yes	Yes
R^2	.01	.25	.31	.31

Replication on Stata...

Econometrics using STATA

Benjamin Monnery
EconomiX, Univ Paris Nanterre

M1 Economie du Droit
2017-2018