Creating a Registry Level Dataset for Catalonia: some comments

José García Montalvo

Barcelona GSE Trobada
October 16, 2015

Summary

- Big data and economics
- Administrative data versus surveys
- My own experience:
  - Finding the “real” price of housing in Spain
  - Men, women and scoring
  - Getting data on public housing lotteries
- Danger of *big data:* confidentiality and correlations
- Concluding remarks

Big data

- Massive amount of information: “the sample is the population” -> anything is potentially useful
- Huge data format heterogeneity: sensors, GPS locations, clicks, logs of servers, emails, images, voice, etc.
- Reutilization of data gather/constructed for other purposes and merge of databases from different sources
- Low level of signal to noise ratios
- Explanations do not pretend to be causal but merely predictive -> causality is irrelevant, only correlation matters: Amazon does not really care why costumers who bought a toaster also buy the book Redeployment… they only know that recommending that book works…

Big data and econometrics

- Important differences between the econometrics before and after big data:
  - Supervised learning (inputs and outputs) uses classification methods, decision trees and neural networks when we use to call that regression
  - Unsupervised learning (only inputs) by shrinkage or dimensionality reduction while we use to call that non parametric estimation
  - Traditional econometrics stop explaining ridge regressions long time ago… to go back to it
  - Stata/Mata/Gauss versus Hadoop/MapReduce/Pig/Mahout /OpenRefine/ Hive/Hbase/ZooKeeper and R
  - Importance of over-fitting and cross validation

Big data and econometrics

- Important differences between the econometrics before big data and after (classical econometrics):
  - Large dimensionality: k>>n: need to regularize and use methods to reduce the dimensionality of the models like LASSO (least absolute shrinkage and Selection Operator any other that penalizes the number of parameters)
  - The path towards causality (using randomized experiments or natural experiments) has taken a detour towards machine learning techniques where prediction is emphasized over causality
  - Big data techniques are subject to the Lucas critique: social reputation and scoring: credit card fraud and testing of 99 cents operations

Administrative data in economics

- STAR + IRS = big success
- IRS + IRS = big success

- SOI Stanford projects (Stanford Center for Poverty and Inequality)
  - Intergenerational persistence
  - Exploiting the occupational field in form 1040
  - Building a new intergenerational panel

- Denmark
- Scandinavian countries: Finland; Norway

- Moving west

Administrative data versus surveys

- For many economic indicators based on surveys there is a more complete data of administrative data
- There is need to move to the generalized used of administrative data for statistical purposes because more a more respondent are refusing to answer surveys, making survey based data less reliable and more expensive to collect
- US has a long history (struggle) on this issue: in 1965 the Johnson administration already tried to create a national data center which fail for fear of "Big brother" -> not only that but it promoted the Privacy Act of 1974

Administrative data versus surveys

- The Privacy Acts limits the use of administrative records even to the government agencies. Later laws were even more stricter than the PA -> 1998 Workforce Investment Act would have prevented LaLonde of writing his seminal paper on the randomized evaluation of training programs
- This situation may be changing: 2016 US Budget moves the agenda for investment in evidence building -> making better use of administrative data

Administrative data versus surveys

7. BUILDING EVIDENCE WITH ADMINISTRATIVE DATA

Introduction

> *"We've got Democratic and Republican elected officials across the country who are ready to roll up their sleeves and get to work. And this should be a challenge that unites us all. I don't care whether the ideas are Democrat or Republican. I do care that they work. I do care that they are subject to evaluation. . ."*
>
> -- President Obama, *"Remarks on Promise Zones,"* January 9, 2014

The Administration is committed to living up to this principle through a broad-based set of activities to better integrate evidence and rigorous evaluation in budget, management, and policy decisions, including through: (1) making better use of already-collected data within government agencies; (2) promoting the use of high-quality, low-cost evaluations and rapid, iterative experimentation; (3) adopting more evidence-based structures for grant programs; and (4) building agency evaluation capacity and developing tools to better communicate what works.

Several Administration documents lay out this "evidence agenda," including previous versions of this trative data are already influencing education and health policy, among other areas. Access to administrative data has been pivotal in some of the most innovative Federal grant reforms and in increasing accountability and transparency across a range of programs; it has also played an important role in innovation and experimentation at the State and local levels. Meanwhile, as the evidence agenda matures, lack of access to appropriate data is increasingly a key obstacle to progress along a number of dimensions. Whether the objective is to facilitate more rapid, low-cost evaluations, to base more grant decisions on strong evidence, to adopt program structures that permit greater innovation and flexibility in exchange for greater accountability for results, or to provide more and better performance information to the public, administrative data are often a crucial untapped resource.

A significant focus in this year's Budget is improving access to administrative data for purposes of evaluation, accountability and transparency, performance management, and other research and analytic purposes. (While not discussed in this chapter, the Budget also includes separate proposals to improve the use of administrative data to protect program integrity, for example to combat identity theft.) The Budget proposes a number of specific access and infrastructure improvements across multiple

Administrative data versus surveys

- Advantages of administrative data:
  - They can be obtained at much lower cost than fielding a new survey
  - They are sometimes more accurate than survey self-reports
  - They are often available for long periods of time permitting the study of the long run impact that would be prohibitively expensive with surveys
  - They can reduce the number of underpowered studies that misdiagnose programs as not working when the problem is the small sample of the study and not the program
  - They also allow for quasiexperimental studies that would be impossible with most survey datasets (especially if the research design depends on detecting small differences in outcomes)

Administrative data versus surveys

- Administrative data is not panacea:
  - The data are collected for a particular purpose (the needs of a program) and not the research design
  - They provide information only on participants and not on those eligible but not participating
  - It may be costly to make administrative data usable for statistical purposes (original data maybe incomplete, inconsistent or poorly documented)

Denmark versus US

- Centralized agency versus decentralized ad-hoc projects:
  - In some countries, like the US, the government is very decentralized (same in many other countries)
  - Different agencies are covered by different privacy laws and many different requirements for access and protection of data (same as CCAA in Spain: Basque Country example)
  - It will take a very long time to formalize and complete a centralized agency

In Spain: some examples

- The Social Security file
- Opening the gates of the Bank of Spain
- Currently the Spanish IRS is so powerful in terms of administrative data that it collects data monthly and, since the beginning of October, in real time
- My own experience:
  - Finding the “real” price of housing in Spain
  - Men, women and scoring
  - Getting data on public housing lotteries

Basic principles for success

- Need to be entrepreneurial -> “You need energy and perseverance and connections… and luck” (Gary Solon)
- No free lunch
- Sometimes you have to pay real money… and perhaps is not a bad solution
- Need cooperative effort to sensitize government agencies about the importance of sharing administrative data with researchers: Krueger in the Treasury; Card, Chetty, Feldstein and Saez’s paper;

Concluding remarks

- The issue of anonymized data -> several papers have shown how to find out the name of the anonymous person to whom the data refer to -> recurrent topic in the latest meeting of the American Statistical Association
- The general rule is that de-identification can be undone
- Careful with ethical forms in applying for research projects that are going to use administrative data (especially if they are financed by the EU)

UNIVERSITAT POMPEU FABRA