Big Data: Perspectives for Economics and Applied Econometrics

Walter Sosa-Escudero

Universisad de San Andrés and CONICET, Argentina

wsosa@udesa.edu.ar • @wsosaescudero • waltersosa.weebly.com

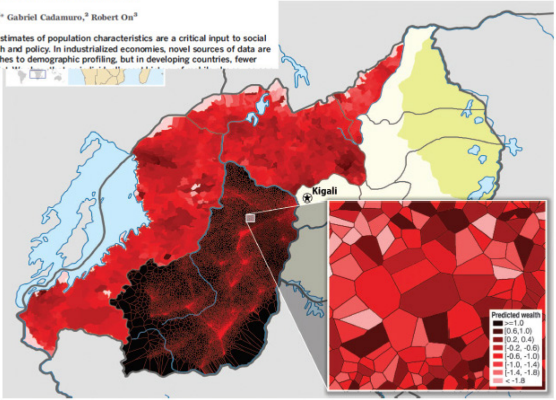
Poverty in Rwanda (predict)

ECONOMICS

Predicting poverty and wealth from mobile phone metadata

Joshua Blumenstock,[1*] Gabriel Cadamuro,[2] Robert On[3]

Accurate and timely estimates of population characteristics are a critical input to social and economic research and policy. In industrialized economies, novel sources of data are enabling new approaches to demographic profiling, but in developing countries, fewer

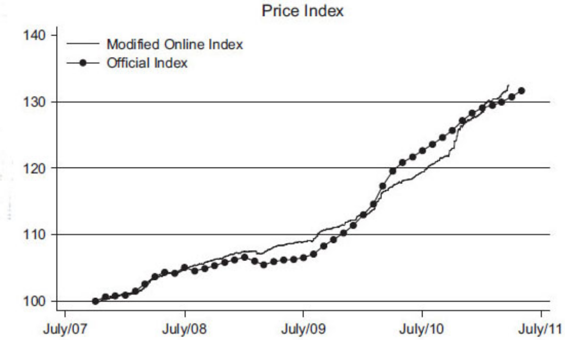


Prices in Argentina (measurement)

Contents lists available at SciVerse ScienceDirect
Journal of Monetary Economics
journal homepage: www.elsevier.com/locate/jme

Online and official price indexes: Measuring Argentina's inflation

Alberto Cavallo *
Massachusetts Institute of Technology, Sloan School of Management, 77 Massachusetts Ave E62-512, Cambridge, MA 02139, USA



Sales tax in the US (causal effect)

Sales Taxes and Internet Commerce

Liran Einav

Dan Knoepfle

Jonathan Levin

Neel Sundaresan

AMERICAN ECONOMIC REVIEW
VOL. 104, NO. 1, JANUARY 2014
(pp. 1-26)

Big data vs. standard statistics

Standard statistics

- How to get the most out of few data?
- Solution: structured data (survey sampling)
- Approach: complex sampling to approximate random sampling (slow and expensive).

Big Data

- Lots of data (Volume)
- Lots of unstructured data (Variety)
- Lots of unstructured, immediate data (Velocity)
- Cheap

Econometrics

$$Y = f(X) + u$$

- Interest in $f(.)$ (marginal effects).
- Model? Theory, well designed experiment.
- Mostly causal effects.
- Probabilistic (standard errors, test)
- Good? Unbiased/consistent/valid inference.

Machine/Statistical Learning

$$Y = f(X) + u$$

- Interest in Y: predict, classify, measure.
- Model? No model. We learn it.
- Point estimate (no inference).
- Good? Predictive performance. Out of sample.

Example: LASSO

$$L(\beta) = \sum (y_i - x_i\beta)^2 + \lambda\, \beta^2$$

- $\lambda = 0$ back to OLS.
- $\lambda \neq 0$ biased but....
- ... can always outperform OLS in prediction.
- Bias is capital sin in econometrics (not in ML).
- ML Idea: bias can reduce variance dramatically.

This is ridge regression. LASSO replaces β^2 with $|\beta|$.

Model assessment

- Econometric etiquette: ex-ante. Good theory, clean identification (consistency). Robust inference.
- Machine learning: ex-post, iterative. Cross validation.
- Cross validation: keep data out to 'test' the accuracy of the model. Switch roles. 'Learn' model to maximize predictive accuracy through cross validation.
- Machine learning builds rather than estimates a model, guided by out of sample predictive/measurement ability.

Policy evaluation requires causal and predictive analysis

Kleinberg, Ludwig, Mullainathan and Obermeyer (2015):

- Hip replacement surgery: effectiveness of surgery (causal) and life expectancy (predictive).
- Crime: release policies depend on effectiveness of mechanism (Shargrodsky and Di Tella, 2013) and predicted probability of commiting a crime.
- Unemployment: training depends on effectiveness of program (causal) and expected unemployment (prediction)

Warnings

- Dependencies (do we really get big data?)
- Choice based sampling (i.e., informal markets)
- Non observed counterfactuals (can we really get *all* data?).
- Correlation fallacy.
- Transparency / privacy.
- Black box (deep learning, forests, etc.)
- Social/political consensus.

Opportunities

- Big data is not just lots but more data (small or big).
- Can identify non-linearities and heterogeneities. Bypass the 'curse of dimensionality'
- Fast (crucial for policy). Goggle Flu Trends. Price scrapping.
- Easy and inexpensive experimentation. Crucial for causal uses.
- Supplements standard official statistics (not replace).
- Coverage (Rwanda). Rural vs. urban.

Links


Sales Taxes and Internet Commerce
Liran Einav
Dan Knoepfle
Jonathan Levin
Neel Sundaresan
AMERICAN ECONOMIC REVIEW
VOL. 104, NO. 1, JANUARY 2014
(pp. 1-26)
the review of income and wealth
Explore this journal >
Original Article
Deprivation and the Dimensionality of Welfare: A Variable-Selection Cluster-Analysis Approach
Germán Caruso, Walter Sosa-Escudero, Marcela Svarc
First published: 6 May
DOI: 10.1111/roiw.121
Machine Learning Methods for Estimating Heterogeneous Causal Effects
Susan Athey and Guido W. Imbens
DISCUSSION PAPER SERIES
IZA DP No. 4201
Google Econometrics and Unemployment Forecasting
Nikolaos Askitas
Klaus F. Zimmermann
Machine Learning Methods for Demand Estimation
Patrick Bajari
Denis Nekipelov
Stephen P. Ryan
Miaoyu Yang
the NATIONAL BUREAU of ECONOMIC RESEARCH
Summer Institute Econometric Lectures
Summer Institute 2015
Lectures on Machine Learning
Susan Athey, Stanford University and NBER and Guido Imbens, Stanford University and NBER
Summer Institute 2014
Theory and Application of Network Models
Daron Acemoglu, MIT and NBER and Matthew O. Jackson, Stanford University
Summer Institute 2013
Econometric Methods for High-Dimensional Data
HOME PROSPECTO NEXOS TUTORIALES PROFESORES
BIG DATA Y MINERIA DE DATOS
PERSPECTIVAS, IDEAS Y HERRAMIENTAS PARA ECONOMISTAS
WALTER SOSA ESCUDERO
ASISTENTE: NOELIA ROMERO

References

- Survey I: Varian, H. R., 2014, Big data: New tricks for econometrics, Journal of Economic Perspectives, 28(2), 3-27.
- Survey II: Einav, L., and Levin, J., 2014, Economics in the age of big data. Science, 346(6210).
- Simple book: James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An introduction to statistical learning (Vol. 6). New York: Springer.
- Book: Murphy, K., 2012, Machine Learning: a Probabilistic Perspective, , MIT Press, Cambridge.
- Predictive: Blumenstock, J., Cadamuro, G., and On, R., 2015, Predicting poverty and wealth from mobile phone metadata. Science, 350(6264), 1073-1076.

- Measurement: Cavallo, A. and Rigobon, R. 2016, The Billion Prices Project: Using Online Prices for Measurement and Research, Journal of Economic Perspectives, Vol. 30, 2, pp.151-78.
- Causal: Athey, S., and G. Imbens (2015), Machine Learning Methods for Estimating Heterogeneous Causal Effects, NBER working paper.
- Policy: Kleinberg, J., Ludwig, J., Mullainathan, S., and Obermeyer, Z., 2015, Prediction policy problems. The American Economic Review, 105(5), 491-495.
- Econometrics: Belloni, A., V. Chernozhukov, and C. Hansen, 2014, High-Dimensional Methods and Inference on Structural and Treatment Effects, Journal of Economic Perspectives, 28(2): 29-50.