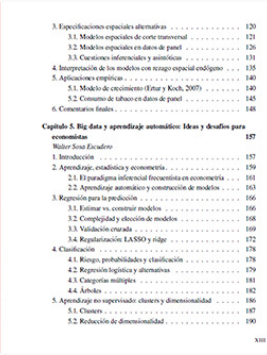
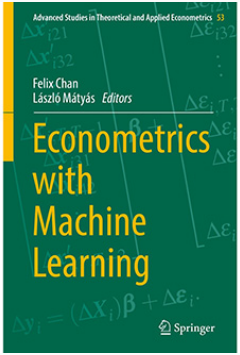
Big Data, Mineria y Aprendizaje: Conceptos y Herramientas para Economistas

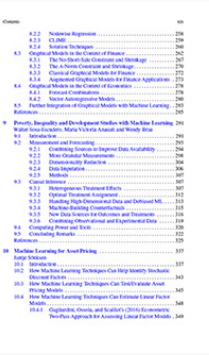
Walter Sosa Escudero
wsosa@udesa.edu.ar

10 de noviembre de 2022


walter sosa escudero
¿qué es (y qué no es) la estadística?
cómo se construyen las predicciones y los datos que más influyen en nuestras vidas (en medio de la revolución de big data)
siglo veintiuno editores
ciencia que ladra... serie mayor
big data
walter sosa escudero
breve manual para conocer la ciencia de datos que ya invadió nuestras vidas
siglo veintiuno editores
ciencia que ladra... serie mayor
walter sosa escudero
borges, big data y yo
guía nerd (y un poco real) para perderse en el laberinto borgeano
siglo veintiuno editores
ciencia que ladra... serie mayor







La grieta...

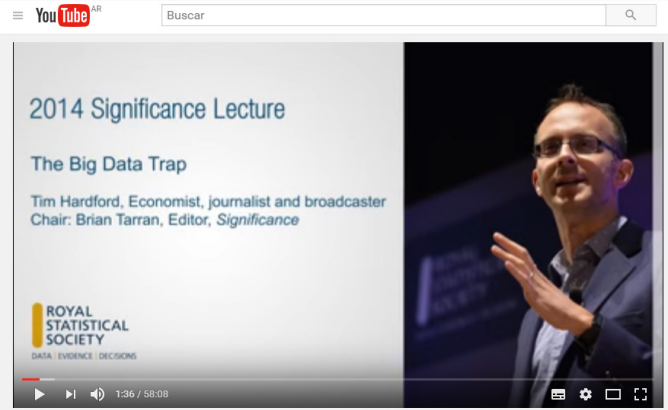
WIRED GEAR SCIENCE ENTERTAINMENT BUSINESS SECURITY DESIGN OPINION

SCIENCE : DISCOVERIES

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

By Chris Anderson 06.23.08

Illustration: Marian Bantjes



RSS 2014 Significance Lecture - The Big Data trap


Clarín
Noticias
Deportes
¡ExtraShow!
Estilo
iEco
Mercados · Indicadores · Empresas y Negocios
Economía
Clarin.com › iEco › Economía › 06/04/14
Big data: ¿Otra vez arroz?
DEBATE
2 opiná
142 shares
11
131
Walter Sosa Escudero
Profesor Asociado, Udesa
Creo conservar, en algún recóndito lugar de mi casa, mi paleta de paddle de cuando en los noventa pensaba que el juego del presente se transformaría en el deporte del futuro. También disfruto de los vinilos de mi adolescencia que escucho casi a diario. Y por alguna razón exótica guardo celosamente una caja de diskettes de mis comienzos con la computación personal, allá en los ochenta.

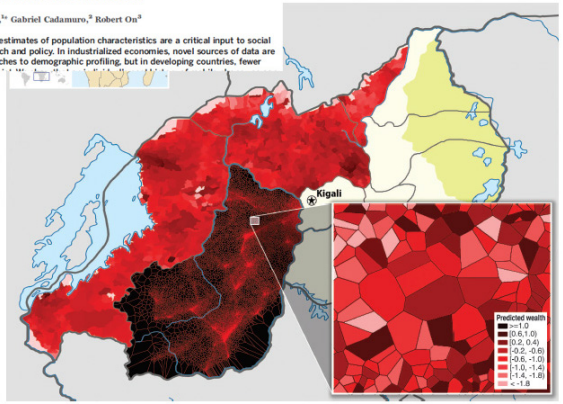
Pobreza en Rwanda (predecir)

ECONOMICS

Predicting poverty and wealth from mobile phone metadata

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

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

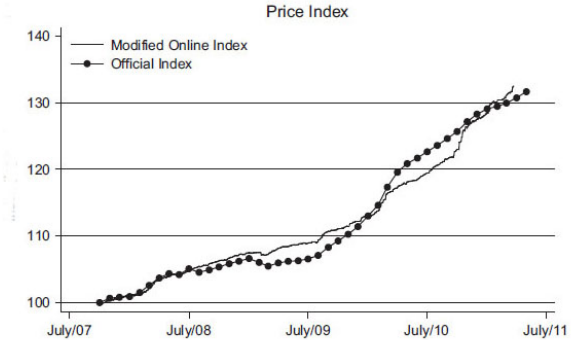


Precios en Argentina (medir)

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 3ID-512, Cambridge, MA 02139, USA



Impuesto a las ventas en EEUU (efecto causal)

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)

Small vs. Big data

Small data (estadistica clasica)

- Extraer lo maximo de pocos datos
- Solucion: estructura (muestreo, modelo)
- Enfoque: muestreo complejo aproxima muestreo al azar (lento y caro, pero bueno). Teoria, experimentos.

Big data

- Muchos datos (Volumen)
- Muchos datos no estructurados (Variedad)
- Muchos datos no estructurados e immediatos (Velocidad)
- 'Condicionalmente baratos'.

Diferencias: Enfoque clasico

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

- Interes en $f(.)$. Efecto causal
- Modelo: Teoria, experimento.
- Probabilidades (error estandar, tests)
- Bueno?: insesagdo, varianza minima, inferencia valida.

Diferencias: Machine learning

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

- Interes en Y: predecir, clasificar, medir.
- Modelo: modelo?. Lo aprendemos.
- Prediccion puntual (no inferencia).
- Bueno?: Performance predictiva fuera de la muestra.

Ejemplo: Ridge / LASSO

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

- $\lambda = 0$: MCO.
- $\lambda \neq 0$: sesgado pero...
- ... siempre le puede ganar a MCO en terminos de prediccion.
- Sesgo: pecado mortal.
- ML: sesgo puede reducir la varianza dramaticamente.
- Puede lidiar con $K > n$.

Esto es ridge. LASSO reemplaza β^2 por $|\beta|$.

LASSO

Para $\lambda \geq 0$ dado, consideremos la siguiente funcion objetivo (a minimizar):

$$R_l(\beta) = \sum_{i=1}^{n}(y_i - x_i'\beta)^2 + \lambda \sum_{s=2}^{p} |\beta_s|$$

(el primer coeficiente corresponde al intercepto).

- ¿Si $\lambda = 0$?
- ¿Si $\lambda = \infty$?

- $\sum_{i=1}^{n}(y_i - x_i'\beta)^2$ penaliza falta de ajuste.
- ¿ $\sum_{s=2}^{p} |\beta_s|$?

$$R_l(\beta) = \sum_{i=1}^{n}(y_i - x_i'\beta)^2 + \lambda \sum_{s=2}^{p} |\beta_s|$$

- LASSO magic: automaticamente elige que variables entran ($\beta_s \neq 0$) y cuales no ($\beta_s = 0$)
- Por que? Coeficientes anulados como 'soluciones de esquina'
- $R_l(\beta)$ es una funcion no diferenciable.

Caso super simple

$$R_l(b) = \sum_{i=1}^{n}(y_i - x_i b)^2 + \lambda|b|$$

$$R_l(b) = SRC(b) + \lambda|b|$$

Supongamos $\hat{b} > 0$. Empecemos en $b = 0$

- Ingreso Mg: $-\frac{d\ SRC(b)}{db}|_{b=0}$ (positivo y decreciente)
- Costo Mg: λ (positivo y constante)

$$R_l(b) = SRC(b) + \lambda|b|$$

- LASSO: poner variables solo si son suficientemente relevantes.
- En general, $\hat{b}_l = 0$ para variables irrelevantes, y $\hat{b}_l$ esta 'corrido hacia cero' para las relevantes.
- Shrinkage: estimacion *sesgada* por la regularizacion.
- Regularizar: 'disciplinar' el modelo hacia la hipotesis nula de no significatividad.
- Por que? Eliminar variables induce sesgo pero puede bajar dramaticamente la varianza, mejora ECM.

Ejemplo: Hitters

`Hitters`

Format

A data frame with 322 observations of major league players on the following 20 variables.

`AtBat` Number of times at bat in 1986

`Hits` Number of hits in 1986

`HmRun` Number of home runs in 1986

`Runs` Number of runs in 1986

`RBI` Number of runs batted in in 1986

`Walks` Number of walks in 1986

`Years` Number of years in the major leagues

`CAtBat` Number of times at bat during his career

`CHits` Number of hits during his career

`CHmRun` Number of home runs during his career

`CRuns` Number of runs during his career

`CRBI` Number of runs batted in during his career

`CWalks` Number of walks during his career

`League` A factor with levels `A` and `N` indicating player's league at the end of 1986

`Division` A factor with levels `E` and `W` indicating player's division at the end of 1986

`PutOuts` Number of put outs in 1986

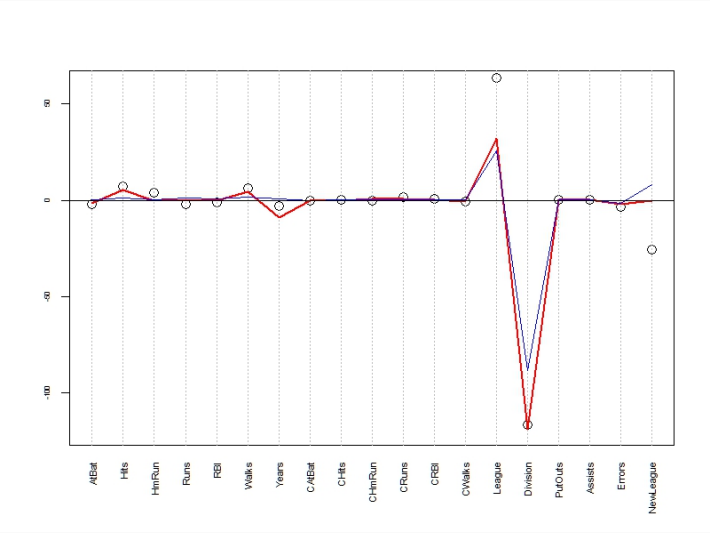
`Assists` Number of assists in 1986

`Errors` Number of errors in 1986

`Salary` 1987 annual salary on opening day in thousands of dollars

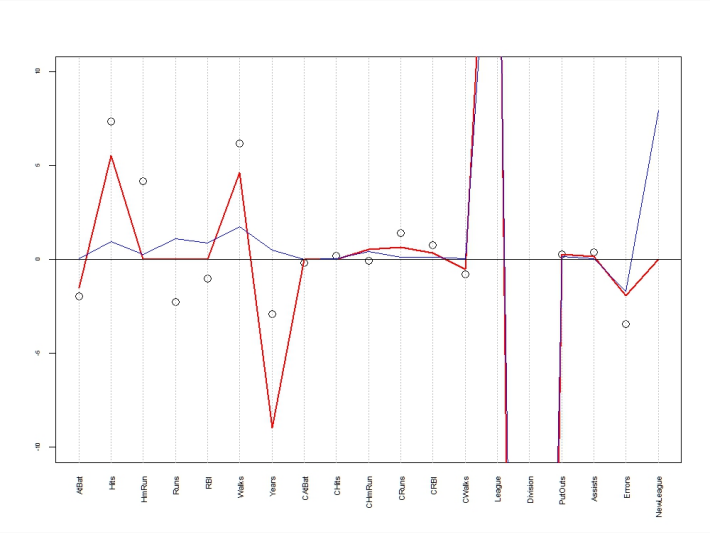
`NewLeague` A factor with levels `A` and `N` indicating player's league at the beginning of 1987

Ejemplo: Hitters



Puntos: OLS, rojo: LASSO, azul: ridge

Ejemplo: Hitters



Puntos: OLS, rojo: LASSO, azul: ridge

Evaluacion de modelos

- Etiqueta estadistica: ex-ante. Teoria, identificacion 'limpia'(consistencia). Inferencia robusta.
- Machine learning: ex-post, iterativa. Cross validation.
- Machine learning construye el modelo mas que lo estima, en base a la performance predictiva fuera de la muestra. Adios al R^2 (y a MCO? Mmm...).

Desafios

- Dependencias (realmente tenemos big data?. Trump effect)
- Choice based sampling.
- Contracticos (podemos tener *todos* los datos?).
- Falacia de la correlacion.
- Transparencia / privacidad.
- Comunicabilidad. Caja negra (deep learning, forests, etc.)
- Consenso social/politico.

Oportunidades

- More mas que big.
- Complejidad, heterogeneidad. No linealidades. Maldicion de la dimensionalidad.
- Rapido (crucial para la politica). Google Flu Trends. Price scrapping.
- Oportunidad para diseño de experimentos.
- Complementa a las estadisticas oficiales (no reemplaza).
- Cobertura. Rural vs. urbano, etc..