Prediction and Linear Regression

Big Data y Machine Learning para Economía Aplicada

Ignacio Sarmiento-Barbieri

Universidad de los Andes

Agenda

1. Linear Regression as Best Predictor
   - Statistical Properties
   - Numerical Properties
2. Calculating the OLS coefficients
3. Big N: Parallel vs Distributed
4. Review

Agenda

1. Linear Regression as Best Predictor
   - Statistical Properties
   - Numerical Properties

2. Calculating the OLS coefficients

3. Big N: Parallel vs Distributed

4. Review

Best Predictor

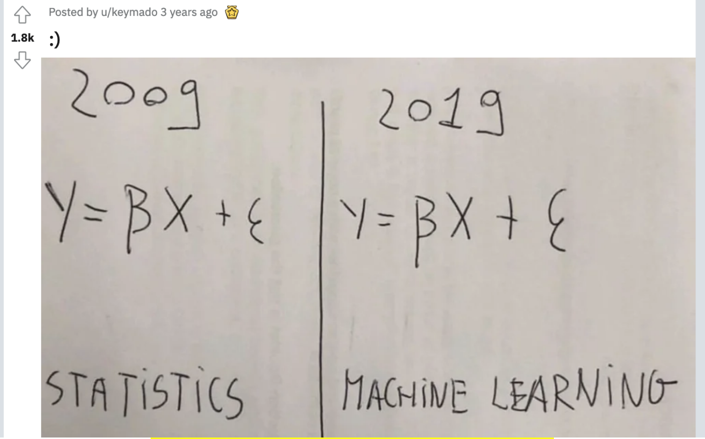
- In the population, the best predictor of Y given X

$$f(X) = E[Y|X] \tag{1}$$

Best Predictor

- In the population, the best predictor of Y given X

$$f(X) = E[Y|X] \tag{1}$$



Posted by u/keymado 3 years ago

1.8k

:)

2009

$Y = \beta X + \varepsilon$

STATISTICS

2019

$Y = \beta X + \varepsilon$

MACHINE LEARNING

Agenda

1. Linear Regression as Best Predictor
   - Statistical Properties
   - Numerical Properties
2. Calculating the OLS coefficients
3. Big N: Parallel vs Distributed
4. Review

Statistical Properties

Under certain assumptions HW Review the Assumption from Econometrics

- Small Sample (Gauss-Markov Theorem)
  - Unbiased: $E(\hat{\beta}) = \beta$
  - Minimum Variance: $Var(\tilde{\beta}) - Var(\hat{\beta})$ is positive semidefinite matrix Proof: HW. Remember: a matrix $M_{p\times p}$ is positive semi-definite iff $c'Mc \geq 0 \; \forall c \in \mathbb{R}^p$
- Large Sample
  - Consistency: $\hat{\beta} \to_p \beta$
  - Asymptotically Normal: $\sqrt{n}(\hat{\beta} - \beta) \sim_a N(0, \Gamma)$

Gauss Markov Theorem

- Gauss Markov Theorem that says OLS is BLUE is perhaps one of the most famous results in statistics.
  - $E(\hat{\beta}) = \beta$
  - $Var(\hat{\beta}) = \sigma^2(X'X)^{-1}$
- and implies that $\hat{y}$ is an unbiased predictor and minimum variance, from the class of unbiased linear predictors (BLUP) H.W. proof

Gauss Markov Theorem

- Gauss Markov Theorem that says OLS is BLUE is perhaps one of the most famous results in statistics.
  - $E(\hat{\beta}) = \beta$
  - $Var(\hat{\beta}) = \sigma^2(X'X)^{-1}$
- and implies that $\hat{y}$ is an unbiased predictor and minimum variance, from the class of unbiased linear predictors (BLUP) H.W. proof
- However, it is essential to note the limitations of the theorem.
  - Correctly specified with exogenous Xs,
  - The term error is homoscedastic
  - No serial correlation.
  - Nothing about the OLS estimator being the more efficient than any other estimator one can imagine.

Statistical Properties

- The fundamental statistical issue is that we are trying to estimate k parameters
- We need many observations per parameter
- n/k should be large, or, equivalently that k/n should be small

Agenda

1 Linear Regression as Best Predictor
   - Statistical Properties
   - Numerical Properties

2 Calculating the OLS coefficients

3 Big N: Parallel vs Distributed

4 Review

Numerical Properties

- Numerical properties have nothing to do with how the data was generated
- These properties hold for every data set, just because of the way that $\hat{\beta}$ was calculated
- Davidson & MacKinnon, Greene y Ruud have nice geometric interpretations

Projections

Applications

- Why FWL is useful in the context of big volume of data?
- An computationally inexpensive way of
  - Removing nuisance parameters
    - E.g. the case of multiple fixed effects. The traditional way is either apply the within transformation with respect to the FE with more categories then add one dummy for each category for all the subsequent FE
    - Not feasible in certain instances.
  - Computing certain diagnostic statistics: Leverage, R^2, LOOCV.
  - Helps with online updating

Applications: Fixed Effects

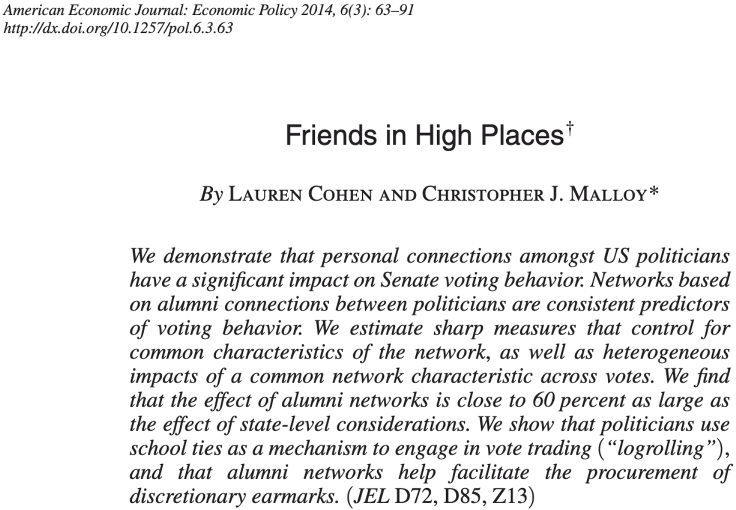
- For example: Carneiro, Guimarães, & Portugal (2012) *AEJ: Macroeconomics*

$$\ln w_{ijft} = x_{it}\beta + \lambda_i + \theta_j + \gamma_f + u_{ijft} \tag{2}$$

$$Y = X\beta + D_1\lambda + D_2\theta + D_3\gamma + u \tag{3}$$

  - Data set 31.6 million observations, with 6.4 million individuals (i), 624 thousand firms (f), and 115 thousand occupations (j), 11 years (t).
  - Storing the required indicator matrices would require 23.4 terabytes of memory
  - From their paper
    "In our application, we first make use of the Frisch-Waugh-Lovell theorem to remove the influence of the three high- dimensional fixed effects from each individual variable, and, in a second step, implement the final regression using the transformed variables. With a correction to the degrees of freedom, this approach yields the exact least squares solution for the coefficients and standard errors"

Applications: Fixed Effects



American Economic Journal: Economic Policy 2014, 6(3): 63–91
http://dx.doi.org/10.1257/pol.6.3.63

Friends in High Places[†]

By Lauren Cohen and Christopher J. Malloy*

We demonstrate that personal connections amongst US politicians have a significant impact on Senate voting behavior. Networks based on alumni connections between politicians are consistent predictors of voting behavior. We estimate sharp measures that control for common characteristics of the network, as well as heterogeneous impacts of a common network characteristic across votes. We find that the effect of alumni networks is close to 60 percent as large as the effect of state-level considerations. We show that politicians use school ties as a mechanism to engage in vote trading ("logrolling"), and that alumni networks help facilitate the procurement of discretionary earmarks. (*JEL* D72, D85, Z13)

Applications: Fixed Effects

photo from https://www.dailydot.com/parsec/batman-1966-labels-tumblr-twitter-vine/

Applications: Outliers and High Leverage Data

- App

Agenda

1. Linear Regression as Best Predictor
   - Statistical Properties
   - Numerical Properties
2. Calculating the OLS coefficients
3. Big N: Parallel vs Distributed
4. Review

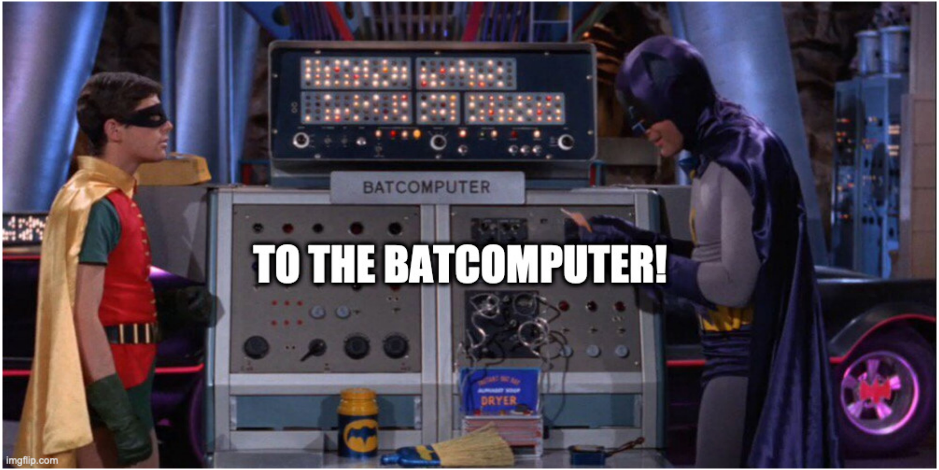


photo from https://www.dailydot.com/parsec/batman-1966-labels-tumblr-twitter-vine/

Agenda

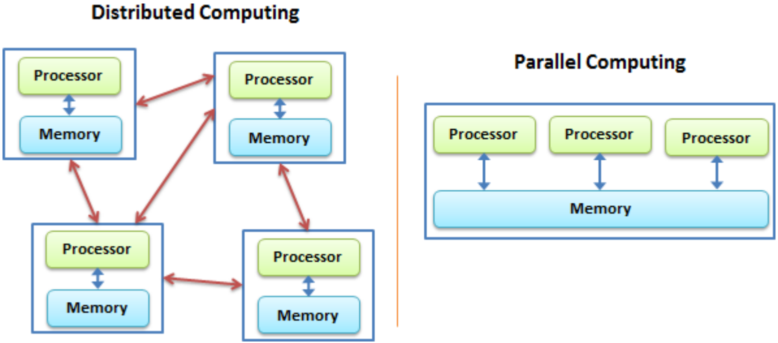
1. Linear Regression as Best Predictor
   - Statistical Properties
   - Numerical Properties
2. Calculating the OLS coefficients
3. Big N: Parallel vs Distributed
4. Review

Big N

- Hemos visto cómo computar $\hat{\beta}$ vía:
  - Descomposición QR
  - SVD
  - Gradiente descendente
- Estos métodos asumen que los datos caben en memoria
- ¿Y si no caben? $\Rightarrow$ Computación distribuida

Parallel vs Distributed

- An algorithm is parallel if it does many computations at once.
  - It needs to see all of the data
- It is distributed if you can work with subsets of data
  - `Stata-mp` is parallel. (license charges by core)
  - `R` and `Python` can be parallel **and** distributed

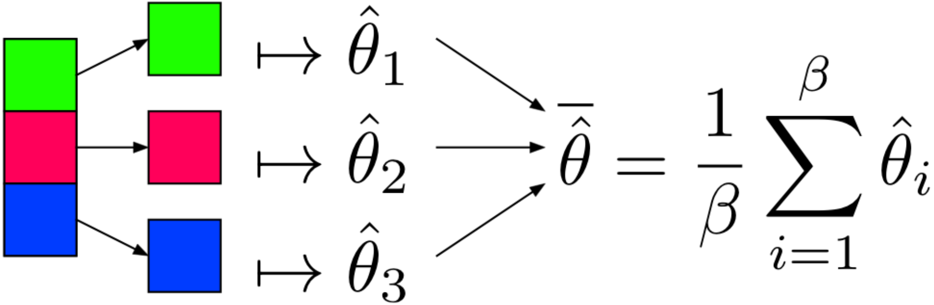


https://tinyurl.com/y3nzvkwh

Map Reduce

- Original Paper *MapReduce: Simplified data processing on large clusters (2004) Dean and Ghemawat*
- It is of the most popular frameworks
- Basic Idea:
  1. Define a key that identifies subgroups of the data which can be processed independently.
  2. **Map**: For each data item, compute the relevant statistics and emit them as '(key, value)' pairs.
  3. Shuffle/Partition: Redistribute the '(key, value)' pairs so that all values with the same key are sent to the same machine.
  4. **Reduce**: On each machine, aggregate or summarize the values for each key into the final result.

Example: Mean by groups



https://datascienceguide.github.io/map-reduce

Spark

- The tools facilitating distributed computing are rapidly improving.
- One prominent system is Spark, that is quickly replacing MapReduce
- Seamlessly integration with `R` and `Python` and has it's own `MLlib`
  - E.g. Spark uses distributed version of stochastic gradient descent to compute OLS
- One of the key differences with MapReduce is how they load data
  - MapReduce has to read from and write to a disk
  - Spark loads it in-memory (can get 100x faster)

Agenda

1. Linear Regression as Best Predictor
    - Statistical Properties
    - Numerical Properties

2. Calculating the OLS coefficients

3. Big N: Parallel vs Distributed

4. Review

¿Por qué no usamos Spark ni Hadoop en este curso?

- Aunque Spark y Hadoop son herramientas populares para procesar grandes volúmenes de datos en entornos distribuidos, este curso no las utilizará directamente. Esto se debe a varias razones prácticas y pedagógicas:
  - Enfoque aplicado desde la economía: En nuestra disciplina, muchas veces el reto no está en el volumen masivo de datos, sino en la complejidad: datos de alta dimensión (k), provenientes de múltiples fuentes, con estructuras no convencionales.
  - Recursos disponibles: El curso está diseñado para que puedas trabajar con tu propia laptop. Spark y Hadoop requieren entornos de cómputo distribuido (clusters), lo cual no es realista en este contexto.
- Lo importante en este curso no es aprender una herramienta específica, sino desarrollar una mentalidad escalable saber cómo estructurar un flujo de trabajo robusto, identificar cuellos de botella y aplicar técnicas de modelado apropiadas cuando el volumen de datos o la complejidad lo exige.