Lecture 5: Intro to Big Data, OLS Numerical Properties

Big Data and Machine Learning for Applied Economics
Econ 4676

Ignacio Sarmiento-Barbieri

Universidad de los Andes

August 25, 2020

Recap

- Prediction vs Estimation
- Train and Test Samples
- Linear Regression
- Example in R
- Machine Learning → build robust predictions from complex data
- Problem Set 1 is up

Agenda

1. Motivation
2. Statistical Properties
3. Numerical Properties
4. Further Readings

Motivation

- *Big Data* has its origin in computer engineering
  - Data so large that cannot be loaded into memory or even stored on a single machine.
  - We need tools, distributed algorithms, that can compute data summaries across multiple independent machines.
  - The bigness comes from *"n"*
- But volume is only a part of the story
- This is data of high complexity: anarchic and spontaneous
- They are the by product of an action: pay with a credit card, browse the web, tweet, etc...

Motivation

- We are going to start exploring tools to
  - Handle
  - Acquire
  - Model
- Today we move in familiar waters
- We are going to explore some properties of linear regression to handle *"Big Data"*

Motivation

- Model $f(x)$ we begin in familiar waters:

$$y = X\beta + u \tag{1}$$

  where
  - y is the dependent variable, a vector $n \times 1$
  - X is a matrix of $k-1$ explanatory variables and an intercept, so its dimension is $n \times k$
  - β is a vector of coefficients $k \times 1$
  - u is the error term, a vector $n \times 1$

- How do we estimate β?

OLS

How do we estimate β?

- Least squares → most widely used
- Consider the following loss function, where we minimize the sum of square residuals

$$SSR(\tilde{\beta}) \equiv \sum_{i=1}^{n} \tilde{e}_i^2 = \tilde{e}'\tilde{e} = (Y - X\tilde{\beta})'(Y - X\tilde{\beta}) \tag{2}$$

  - $SSR(\tilde{\beta})$ is the aggregation of squared errors if we choose $\tilde{\beta}$ as an estimator.

- The **least squares estimator** $\hat{\beta}$ will be

$$\hat{\beta} = \underset{\tilde{\beta}}{argmin}\, SSR(\tilde{\beta}) \tag{3}$$

OLS

$$\begin{aligned} SSR(\tilde{\beta}) &= \tilde{e}'\tilde{e} && (4) \\ &= (Y - X\tilde{\beta})'(Y - X\tilde{\beta}) && (5) \end{aligned}$$

FOC are

$$\frac{\partial \tilde{e}'\tilde{e}}{\partial \tilde{\beta}} = 0 \tag{6}$$

$$\frac{\partial \tilde{e}'\tilde{e}}{\partial \tilde{\beta}} = -2X'Y + 2X'X\tilde{\beta} = 0 \tag{7}$$

OLS

Let $\hat{\beta}$ be the solution.Then $\hat{\beta}$ satisfies the following normal equation

$$X'X\hat{\beta} = X'y \tag{8}$$

If the inverse of $X'X$ exists, then

$$\hat{\beta} = (X'X)^{-1}X'y \tag{9}$$

Statistical Properties

Under certain assumptions HW Review the Assumption from Econometrics

- Small Sample
  - Unbiased: $E(\hat{\beta}) = \beta$
  - Minimum Variance: $Var(\tilde{\beta}) - Var(\hat{\beta})$ is positive semidefinite matrix
- Large Sample
  - Consistency: $\hat{\beta} \rightarrow_p \beta$
  - Asymptotically Normal: $\sqrt{N}(\hat{\beta} - \beta) \sim_a N(0, S)$

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
- Helps in computing with big data

Projection

OLS Residuals:

$$
\begin{aligned}
e &= y - \hat{y} && (10)\\
&= y - X\hat{\beta} && (11)
\end{aligned}
$$

replacing $\hat{\beta}$

$$
\begin{aligned}
e &= Y - X(X'X)^{-1}X'y && (12)\\
&= (I - X(X'X)^{-1}X')y && (13)
\end{aligned}
$$

Define two matrices

- Projection matrix $P_X = X(X'X)^{-1}X'$
- Annihilator (residual maker) matrix $M_X = (I - P_X)$

Projection

- $P_X = X(X'X)^{-1}X'$
- $M_X = (I - P_X)$
- Both are symmetric
- Both are idempotent $(A'A) = A$
- $P_X X = X$ hence projection matrix
- $M_X X = 0$ hence annihilator matrix

We can write

$$SSR = e'e = u'Mu \tag{14}$$

So we can relate SSR to the true error term u

Frisch-Waugh-Lovell (FWL) Theorem

- Lineal Model: $Y = X\beta + u$
- Split it: $Y = X_1\beta_1 + X_2\beta_2 + u$
  - $X = [X_1\ X_2]$, X is $n \times k$, X_1 $n \times k_1$, X_2 $n \times k_2$, $k = k_1 + k_2$
  - $\beta = [\beta_1\ \beta_2]$

Theorem

1. The OLS estimates of β_2 from these equations

$$y = X_1\beta_1 + X_2\beta_2 + u \tag{15}$$

$$M_1 y = M_1 X_2 \beta_2 + residuals \tag{16}$$

   are numerically identical

2. the OLS residuals from these regressions are also numerically identical

Applications

- Why FWL is useful in the context of Big 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.

Applications: Fixed Effects

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

$$\ln w_{ijft} = x_{it}\beta + \lambda_i + \theta_j + \gamma_f + u_{ijft} \tag{17}$$

$$Y = X\beta + D_1\lambda + D_2\theta + D_3\gamma + u \tag{18}$$

  - 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: Leverage

Note the following

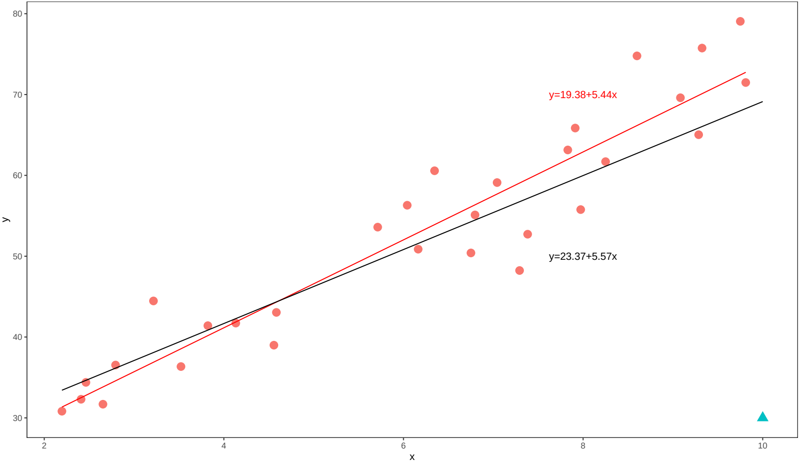
$$\hat{\beta} = (X'X)^{-1}X'y \tag{19}$$

each element of the vector of parameter estimates $\hat{\beta}$ is simply a weighted average of the elementes of the vector y

Let's call c_i the i-th row of the matrix $(X'X)^{-1}X'$ then

$$\hat{\beta}_i = c_i y \tag{20}$$

Applications: Leverage



Applications: Leverage

Consider a dummy variable e_j which is an $n - vector$ with element j equal to 1 and the rest is 0. Include it as a regressor

$$y = X\beta + \alpha e_j + u \tag{21}$$

using FWL we can do

$$M_{e_j} y = M_{e_j} X\beta + r \tag{22}$$

- β and *residuals* from both regressions are identical
- Same estimates as those that would be obtained if we deleted observation j from the sample we are going to denote this as $\beta^{(j)}$

Note:

- $M_{e_j} = I - e_j(e_j'e_j)^{-1}e_j'$
- $M_{e_j}y = y - e_j(e_j'e_j)^{-1}e_j'y = y - y_je_j$
- $M_{e_j}X$ is X with the j *column* replaced by zeros

Applications: Leverage

Let's define a new matrix $Z = [X, e_j]$

$$y = X\beta + \alpha e_j + u \tag{23}$$

$$y = Z\theta + u \tag{24}$$

then the fitted values

$$y = P_Z y + M_z y \tag{25}$$

$$= X\hat{\beta}^{(j)} + \hat{\alpha} e_j + M_Z y \tag{26}$$

Pre-multiply by P_X (remember $M_Z P_X = 0$)

$$P_X y = X\hat{\beta}^{(j)} + \hat{\alpha} P_X e_j \tag{27}$$

$$X\hat{\beta} = X\hat{\beta}^{(j)} + \hat{\alpha} P_X e_j \tag{28}$$

$$X(\hat{\beta} - \beta^{(j)}) = \hat{\alpha} P_X e_j \tag{29}$$

Applications: Leverage

How to calculate α? FWL once again

$$M_X y = \hat{\alpha} M_X e_j + res \tag{30}$$

$$\hat{\alpha} = (e_j' M_X e_j)^{-1} e_j' M_X y \tag{31}$$

- $e_j' M_X y$ is the j element of $M_X y$ is the vector of residuals from the regression including all observations
- $e_j' M_x e_j$ is just a scalar, the diagonal element of M_X
  Then

$$\hat{\alpha} = \frac{\hat{u}_j}{1 - h_j} \tag{32}$$

where h_j is the j diagonal element of P_X

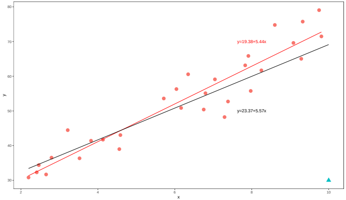
Applications: Leverage

Finally we get

$$(\hat{\beta}^{(j)} - \hat{\beta}) = -\frac{1}{1-h_j}(X'X)^{-1}X_j'\hat{u}_j \tag{33}$$

Influence depends on two factors

- $\hat{u}_j$ large residual → related to y coordinate
- $\hat{h}_t$ related to x coordinate → if h_j is large, we have a high leverage



Applications: Leverage

Case of $y = \alpha + \beta x_u$ (ISLR)

$$h_j = e_j' P_X e_j \tag{34}$$

$$. \tag{35}$$

$$\text{(steps as HW)} \tag{36}$$

$$. \tag{37}$$

$$h_j = \frac{1}{n} + \frac{(x_j - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2} \tag{38}$$

- Then h_j is always between $\frac{1}{n}$ and 1
- The average $\sum_j h_j/n$ is always equal to $(k+1)/n$

Goodness of Fit

R^2: the fraction of the variation of the dependent variable that is attributable to the variation in the explanatory variables

$$R^2 = \frac{ESS}{TSS} = \frac{||P_X y||^2}{||y||^2} = 1 - \frac{||M_X y||^2}{||y||^2} \tag{39}$$

- Problem: not invariant to changes in units, can be negative
- In practice we use the centered version:

$$R_c^2 = \frac{||P_X M_\iota y||^2}{||M_\iota y||^2} \tag{40}$$

R_c^2: is a measure of the explanatory power of the nonconstant regressors.

Review & Next Steps

- What is Big Data?
- Quick Review of Statistical Properties
- Numerical Properties
- FWL
  - Fixed Effects
  - Leverage
  - Goodness of Fit
- **Next Class:** OLS Computation, Scraping Data
- Questions? Questions about software?

Further Readings

- Carneiro, A., Guimarães, P., & Portugal, P. (2012). Real Wages and the Business Cycle: Accounting for Worker, Firm, and Job Title Heterogeneity. American Economic Journal: Macroeconomics, 4 (2): 133-52.
- Davidson, R., & MacKinnon, J. G. (2004). Econometric theory and methods (Vol. 5). New York: Oxford University Press.
- Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, No. 10). New York: Springer series in statistics.
- Greene, W. H. (2003). Econometric analysis fifth edition. New Yersey: Prentice Hall.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.
- Ruud, P. A. (2000). An introduction to classical econometric theory. OUP Catalogue
- Taddy, M. (2019). Business data science: Combining machine learning and economics to optimize, automate, and accelerate business decisions. McGraw Hill Professional.