Lecture 3: OLS Properties

Big Data and Machine Learning for Applied Economics
Econ 4676

Ignacio Sarmiento-Barbieri

Universidad de los Andes

August 17, 2021

Agenda

Recap

- We began shifting paradigms
- Decision Theory:
  - The goal here is to solve something which looks like

perdida

$$f^{\star} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \{\underbrace{E[L(Y, f(\mathbf{X}))]}_{\text{Riesgo}}\} \quad (1)$$

$L = (\theta - \hat{\theta})^2 = (\hat{\theta} - \theta)^2$

  - Square error loss $\rightarrow$ MSE
  - Solution: CEF ($E[Y|X]$)
  - Linear model is a "work horse" and approximates the CEF quite well

Linear Regression Model

- If $f(X) = X\beta$, obtaining $f(.)$ boils down to obtaining β

$$y = X\beta + u \tag{2}$$

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}_{n\times 1}$$

- where
  - y is a vector $n \times 1$ with typical element y_i
  - X is a matrix $n \times k$
    - Note that we can represent it as a column vector $\underset{n\times k}{X} = [\underset{n\times 1}{X_1} \; \underset{n\times 1}{X_2} \ldots \underset{n\times 1}{X_k}]$
  - β is a vector $k \times 1$ with typical element β_j

- Thus

$$y_i = X_i\beta + u_i \qquad i = 1, \ldots, n \qquad k = 2 \tag{3}$$

$$y = \sum_{j=1}^{k} \beta_j X_{ji} + u_i = \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

Linear Regression Model

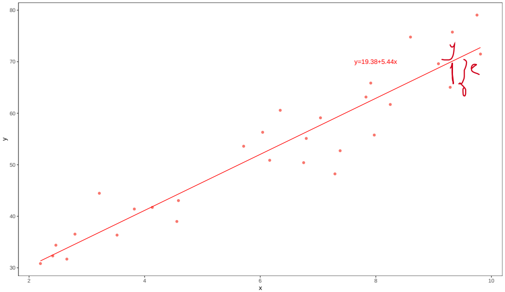
How do we obtain β?

$\hat{y} = X\hat{\beta}$

- Method of Moments (for HW)
- MLE (more on this later)
- OLS: minimize risk squared error loss $\rightarrow$ minimizes SSR $(e'e)$
  - where $e = Y - \hat{Y} = Y - X\hat{\beta}$
  - In the HW, you will show that min SSR same as max R^2 -> in sample fit

$R^2 = 1 - \frac{SSR}{TSS}$

fija



OLS

How do we obtain β?

- Consider the following loss function, where we minimize the sum of square residuals

$$SSR(\tilde{\beta}) \equiv \sum_{i=1}^{n} \tilde{e}_i^2 = \tilde{e}'\tilde{e} = (Y - X\tilde{\beta})'(Y - X\tilde{\beta}) \tag{4}$$

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

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

$$\hat{\beta} = \underset{\tilde{\beta}}{argmin}\, SSR(\tilde{\beta}) \tag{5}$$

OLS

$$SSR(\tilde{\beta}) = \tilde{e}'\tilde{e} \tag{6}$$

$$= (Y - X\tilde{\beta})'(Y - X\tilde{\beta}) \tag{7}$$

- FOC are

$$\frac{\partial \tilde{e}'\tilde{e}}{\partial \tilde{\beta}} = 0 \tag{8}$$

$$\frac{\partial \tilde{e}'\tilde{e}}{\partial \tilde{\beta}} = -2X'Y + 2X'X\tilde{\beta} = 0 \tag{9}$$

$$\cancel{-}2XY = \cancel{-}2X'X\tilde{\beta}$$

gradiente
gradient
gradient !

OLS

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

$$X'X\hat{\beta} = X'y \tag{10}$$

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

$$\hat{\beta} = (X'X)^{-1}X'y \tag{11}$$

- Note that this is a closed solution (a bonus!!)

Statistical Properties

Under certain assumptions HW Review the Assumption from Econometrics

- Small Sample (Gauss-Markov Theorem)
  - Unbiased: $E(\hat{\beta}) = \beta$
  - Minimum Variance: $\underbrace{Var(\tilde{\beta}) - Var(\hat{\beta})}_{\geq 0}$ is positive semidefinite matrix
- Large Sample
  - Consistency: $\hat{\beta} \rightarrow_p \beta$
  - Asymptotically Normal: $\underbrace{\sqrt{N}(\hat{\beta} - \beta)}_{i=1 - n} \sim_a N(0, S)$

Gauss Markov Theorem

- Gauss-Markov Theorem says that

$$\hat{\beta} = (X'X)^{-1}X'y \tag{12}$$

  - The OLS estimator ($\hat{\beta}$) is BLUE, the more efficient than any other linear unbiased estimator,
  - Efficiency in the sense that $Var(\tilde{\beta}) - Var(\hat{\beta})$ is positive semidefinite matrix.

Proof: HW. Remember: a matrix $M_{p\times p}$ is positive semi-definite iff $c'Mc \geq 0 \; \forall \overset{c}{\cancel{x}} \in \mathbb{R}^p$

forma cuadrática

Gauss Markov Theorem

- Gauss Markov Theorem that says OLS is BLUE is perhaps one of the most famous results in statistics.

  supuestos

  - $E(\hat{\beta}) = \beta$ — $E(u|X) = 0$
  - $V(\hat{\beta}) = \sigma^2 (X'X)^{-1}$ — $V(u|X) = \sigma^2 I$

- 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.

Prediction vs Estimation

$f() = X\beta$

$\boxed{f = I}$

- **Predicting well in this context $\rightarrow$ estimating well**
  - Note that the prediction of y will be given by $\hat{y} = X\hat{\beta}$
  - Under Gauss-Markov framework

    $\hat{y} \rightarrow$ predictor

    $\hat{\beta} \rightarrow$ estimator

    - $E(\hat{y}) = X\beta$
    - $V(\hat{y}) = \boxed{\sigma^2 X'(X'X)^{-1}X}$

    $= V(\hat{\beta}) = \sigma^2 (X'X)^{-1}$

    $\hat{y} = X\hat{\beta} \rightarrow V(\hat{y}) = V(X\hat{\beta}) = X' V(\hat{\beta}) X$

- Then if $\hat{\beta}$ is unbiased and of minimum variance,
- $\hat{y}$ is an unbiased predictor and minimum variance, from the class of unbiased linear predictors (BLUP)
  - Proof: for HW, see proof for $\hat{\beta}$

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:

$$e = y - \hat{y} \tag{13}$$

$$= y - X\hat{\beta} \tag{14}$$

replacing $\hat{\beta}$

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

$$e = y - X\overbrace{(X'X)^{-1}X'y}^{\hat{\beta}} \tag{15}$$

$$= \underbrace{(I - \underbrace{X(X'X)^{-1}X'}_{P_X})}y \tag{16}$$

Define two matrices

$$\boxed{e} = M_X y$$

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

Projection

$X_{n \times k}$ $\quad P_{X\, n \times n}$

- $P_X = X(\underbrace{X'X}_{k \times k})^{-1}X'_{k \times n}$
- $M_X = (I - P_X)$
- Both are symmetric $\quad A' = A$
- Both are idempotent $(A'A) = A$
- $P_X X = X$ hence projection matrix
- $M_X X = 0$ hence annihilator matrix

$A'A \to AA = A \qquad P_X P_X = P_X$

$\underbrace{X(X'X)^{-1}X'X}_{I}(X'X)^{-1}X' = XI(X'X)^{-1}X' = X(X'X)^{-1}X'$

We can write

$$SSR = e'e = u'M_X u \quad \to \quad \underbrace{u'M_X}_{e'}\ \underbrace{M_X u}_{e} \tag{17}$$

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 $n \times k_2$, $k = k_1 + k_2$
  - $\beta = [\beta_1\ \beta_2]$

$X = [X_1, X_2]$

Theorem

1. The OLS estimates of β_2 from these equations

$$y = X_1\beta_1 + X_2\beta_2 + u \tag{18}$$

$$M_{X_1}y = M_{X_1}X_2\beta_2 + residuals \tag{19}$$

are numerically identical

2. the OLS residuals from these regressions are also numerically identical $\rightarrow V(\hat{\beta})$

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.
  - Way to add more data without having to compute everything again

Applications: Fixed Effects

reghdfe

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

$$\ln w_{ijft} = x_{it}\beta + \lambda_i + \theta_j + \gamma_f + u_{ijft} \tag{20}$$

Areg or ?

$$Y = X\boxed{\beta} + D_1\lambda + D_2\theta + D_3\gamma + u \tag{21}$$

  - 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

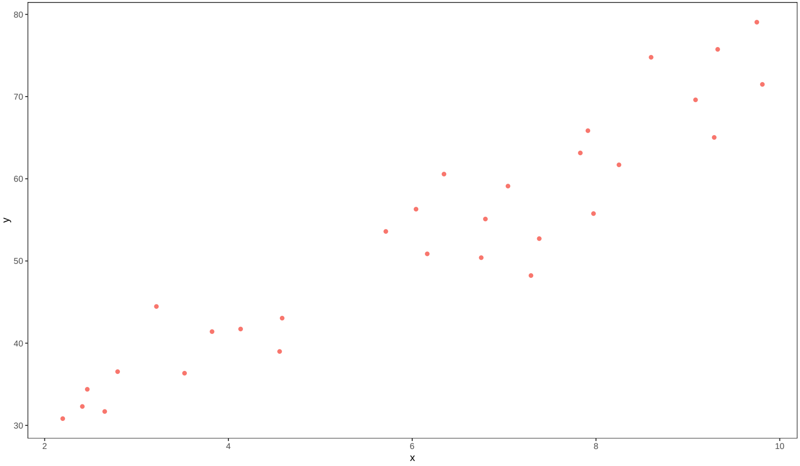
$$\hat{\beta} = (X'X)^{-1}X'y \qquad \hat{\beta} = wy \tag{22}$$

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

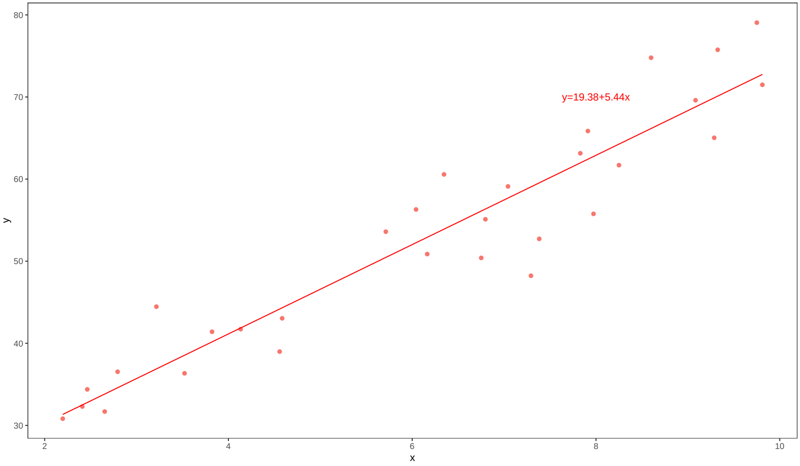
Let's call c_i the i-th row of the matrix $(X'X)^{-1}X'$ then

$$\hat{\beta}_i = c_i y \tag{23}$$

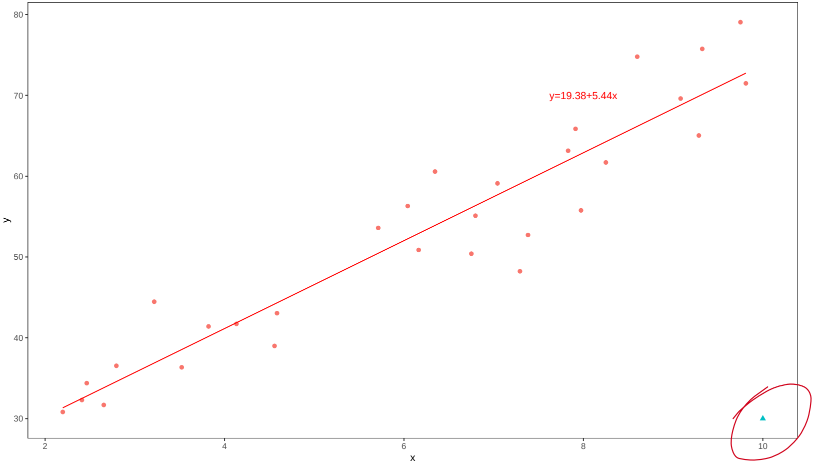
Applications: Leverage



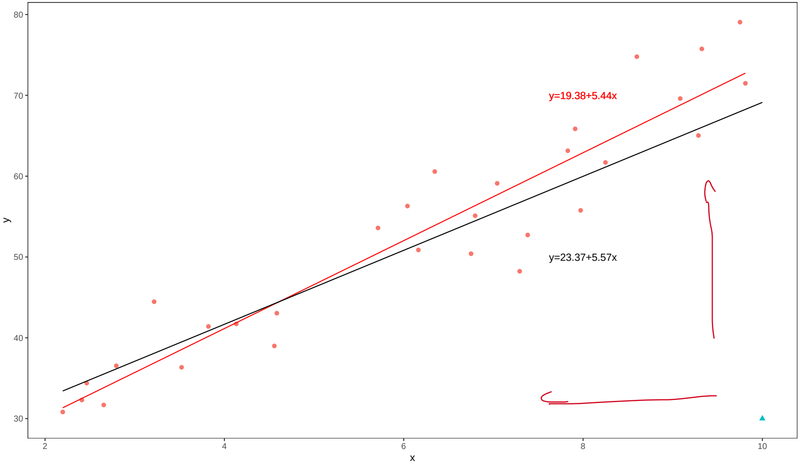
Applications: Leverage



Applications: Leverage



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

$$e_j = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \leftarrow n\ j$$

$$y = X\beta + \alpha e_j + u \tag{24}$$

using FWL we can do

$$M_{e_j}y = M_{e_j}X\beta + r \tag{25}$$

$$M_{e_j} y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} - \begin{pmatrix} 0 \\ y_2 \\ 0 \end{pmatrix} = \begin{pmatrix} y_1 \\ 0 \\ y_3 \end{pmatrix}$$

- 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 $\beta^{(j)}$

$n=3 \quad e_{j=2} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \qquad \underbrace{\begin{pmatrix} 0 & 1 & 0 \end{pmatrix}}_{e_j'} \underbrace{\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}}_{e_j} = 1$

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 *row* replaced by zeros

$$P_{e_j} = e_j (e_j'e_j)^{-1} e_j' = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} y_2 = \begin{pmatrix} 0 \\ y_2 \\ 0 \end{pmatrix}$$

Applications: Leverage

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

$$\cancel{y = X\beta + \alpha e_j + u} \tag{26}$$

$$\boxed{y = Z\theta + u} \tag{27}$$

~~then the fitted values~~

$$P_Z y + (I - P_Z) y = y$$

$$y = P_Z y + M_z y \tag{28}$$

$$= X\hat{\beta}^{(j)} + \hat{\alpha} e_j + M_Z y \tag{29}$$

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

$$P_X X = X$$

$$\underbrace{X(X'X)^{-1}X'y}_{\beta} \qquad (X'X)^{-1}$$

$$P_X y = X\hat{\beta}^{(j)} + \hat{\alpha} P_X e_j \tag{30}$$

$$X\hat{\beta} = X\hat{\beta}^{(j)} + \hat{\alpha} P_X e_j \tag{31}$$

$$\boxed{X(\hat{\beta} - \beta^{(j)}) = \hat{\alpha} P_X e_j} \tag{32}$$

Applications: Leverage

How to calculate α? FWL once again

$$M_X y = \hat{\alpha} M_X e_j + res \tag{33}$$

$(X'X)^{-1}X'y$

$$\hat{\alpha} = (e_j' M_X e_j)^{-1} e_j' M_X y \tag{34}$$

(X)

H/W

- $e_j' M_X y$ is the j element of $M_X y$ is the vector of residuals from the regression including all observations

$e_j' X e_j$

- $e_j' M_x e_j$ is just a scalar, the diagonal element of M_X

  Then

$$\hat{\alpha} = \frac{\hat{u}_j}{1 - h_j} \tag{35}$$

where h_j is the j diagonal element of P_X

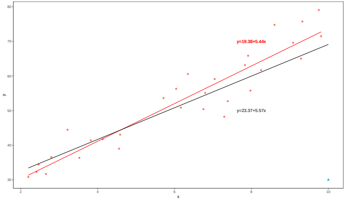
Applications: Leverage

Finally we get

$$(\hat{\beta}^{(j)} - \hat{\beta}) = -\frac{1}{1-h_j}(X'X)^{-1}X_j'\hat{u}_j \tag{36}$$

Influence depends on two factors

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



Applications: Leverage

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

$$h_j = e_j' P_X e_j \tag{37}$$

$$. \tag{38}$$

$$\text{(steps as HW)} \tag{39}$$

$$. \tag{40}$$

$$h_j = \frac{1}{n} + \frac{(x_j - \bar{x})^2}{\sum_{i'=1}^{n} (x_{i'} - \bar{x})^2} \tag{41}$$

- 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

$$R^2 = \frac{ESS}{TSS} = \frac{||P_X y||^2}{||y||^2} = 1 - \frac{||M_X y||^2}{||y||^2} \tag{42}$$

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

$$R_c^2 = \frac{||P_X M_\iota y||^2}{||M_\iota y||^2} \tag{43}$$

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

Review & Next Steps

- OLS
- Quick Review of Statistical Properties
- Numerical Properties
- FWL
  - Fixed Effects
  - Leverage
  - Goodness of Fit

- **Next Class:** SSR Computation

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.