Notes of OLS

Yue Wu

1 Introduction

This note mainly focuses on OLS(which assumes fixed X and random Y) and population OLS(which views (x_i, y_i) as IID draws from a population).

For OLS, firstly we talk about some algebraic facts, where we regard X as a fixed covariate matrix, and Y as a constant vector or any realization of random vector. For this part, no assumption is needed and we can regard the OLS as a procedure of calculation.

Then we begin to regard Y as a random vector(thus $\hat{\beta}, \hat{Y}, \hat{\varepsilon}$ as random) and make statistical analysis. We talk about two statistical model assumptions, the Gauss-Markov Model and the Normal Linear Model. It is noted that Gauss-Markov Model does make statistical assumptions on the first two moments.

Conditional on X.

This note can be seemed as a study summary of Ding Peng's book(Chapter 3,4,5,7,12), which I learned a lot from and highly recommend.

[Variance estimation]

2 OLS

In this section, fixed X and random Y.

2.1 OLS formula and Algebraic facts

- the outcome vector

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

- covariate matrix

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} x_1^\top \\ x_2^\top \\ \vdots \\ x_n^\top \end{pmatrix} = (X_1, \ldots, X_p),$$

- OLS coefficient($X^\top X$ is non-degenerate)

$$\hat{\beta} = \left(\sum_{i=1}^{n} x_i x_i^\top \right)^{-1} \left(\sum_{i=1}^{n} x_i y_i \right) = (X^\top X)^{-1} X^\top Y$$

- other hat formulas

$$\hat{Y} = X\hat{\beta}$$
$$\hat{\varepsilon} = Y - \hat{Y}$$

Here are some algebraic facts about vector $\hat{\beta}$, vector $\hat{Y}$, vector $\hat{\varepsilon}$ which are calculated through OLS procedure with constant matrix X, and constant vector Y.

[the vector Y can be seemed as any number/realization of random variable]

- (the Normal equation) $X^\top(Y - X\hat{\beta}) = 0$
- (automatically centered residuals) When X contains a column of intercepts $\mathbf{1}_n = (1, \ldots, 1)^\top$, we have $\mathbf{1}_n^\top \hat{\varepsilon} = 0 \implies n^{-1}\sum_{i=1}^n \hat{\varepsilon}_i = 0$
- (projection matrix/hat matrix) $\hat{Y} = HY$ where $H = X(X^\top X)^{-1}X^\top$, and $H^2 = H, H^\top = H$. H also satisfies $HX = H$, $(I_n - H)X = 0$
- (orthogonal $\hat{Y}$ and $\hat{\varepsilon}$) $\hat{Y}^\top\hat{\varepsilon} = 0$ or $\hat{Y} \perp \hat{\varepsilon}$
- (best $\hat{\beta}$?) For any $b \in \mathbb{R}^p$, we have the following decomposition

$$\|Y - Xb\|^2 = \|Y - X\hat{\beta}\|^2 + \|X(\hat{\beta} - b)\|^2,$$

  which implies that $\|Y - Xb\|^2 \geq \|Y - X\hat{\beta}\|^2$ with equality holding if and only if $b = \hat{\beta}$.
- (FWL, Partitioned regression) The OLS estimator for β_2 in the long regression has the equivalent form $\hat{\beta}_2 = (\tilde{X}_2^\top\tilde{X}_2)^{-1}\tilde{X}_2^\top\tilde{Y}$ where $\tilde{X}_2 = (I_n - H_1)X_2$ and $\tilde{Y} = (I_n - H_1)Y$

2.2 Statistical properties with Gauss-Markov Model assumption

From here, we induce the statistical model and some assumptions, and regard Y as random vector.

2.2.1 Assumption of GM model

The Gauss-Markov model assumes that Y has mean $X\beta$ and covariance matrix $\sigma^2 I_n$.

Assumption 1 (Gauss-Markov model)**.** *We have*

$$Y = X\beta + \varepsilon$$

where the design matrix X is fixed with linearly independent column vectors, and the random error term ε has the first two moments

$$E(\varepsilon) = 0,$$
$$\operatorname{cov}(\varepsilon) = \sigma^2 I_n.$$

The unknown parameters are (β, σ^2).

The equivalent forms of Gauss-Markov model:

- $E(Y) = X\beta,\ \operatorname{cov}(Y) = \sigma^2 I_n$.
- $y_i = x_i^\top\beta + \varepsilon_i,\ (i = 1, \ldots, n)$ where the error terms are uncorrelated with mean 0 and variance σ^2.

Some example distribution of ε/Y that satisfying GM assumption:

- $\varepsilon_i \overset{\text{IID}}{\sim} N(0, \sigma^2), \quad (i = 1, \ldots, n)$.

2.2.2 Properties of the OLS estimator

Under Gauss-Markov model assumption, we have:

- (the mean and covariance of $\hat{\beta}$)

$$E(\hat{\beta}) = \beta,$$
$$\operatorname{cov}(\hat{\beta}) = \sigma^2 (X^\top X)^{-1}.$$

- (the mean and covariance matrix of $(\hat{Y}, \hat{\varepsilon})$, uncorrelated $\hat{Y}$ and $\hat{\varepsilon}$)

$$E\begin{pmatrix} \hat{Y} \\ \hat{\varepsilon} \end{pmatrix} = \begin{pmatrix} X\beta \\ 0 \end{pmatrix}$$

$$\operatorname{cov}\begin{pmatrix} \hat{Y} \\ \hat{\varepsilon} \end{pmatrix} = \sigma^2 \begin{pmatrix} H & 0 \\ 0 & I_n - H \end{pmatrix}.$$

  so $\hat{Y}$ and $\hat{\varepsilon}$ are uncorrelated.

- (BLUE, Gauss-Markov Theorem) OLS estimator is optimal in the sense of having the smallest covariance matrix among all linear unbiased estimators.

2.3 Statistical properties with Normal Linear Model assumption

Recall what we learned in Mathematical Statistics, actually we did not see things like BLUE. It is noted that the Gauss-Markov model just makes assumption on the first two moments of ε/Y(thus $\hat{\beta}$), but does not fully determine its distribution. It enables us to derive some nice properties of the OLS estimator, but is inadequate for statistical inference.

With assumption of Normal linear model, the distribution of ε/Y(thus $\hat{\beta}$) is determined.

2.3.1 Assumption of NL model

Assumption 2 (Normal linear model)**.** *We have*

$$Y \sim N(X\beta, \sigma^2 I_n),$$

where the design matrix X is fixed with linearly independent column vectors. The unknown parameters are (β, σ^2).

The equivalent forms of Normal linear model:

- $y_i \overset{\text{IND}}{\sim} N(x_i^\top \beta, \sigma^2), \quad (i = 1, \ldots, n),$
- $Y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_n)$
- $y_i = x_i^\top \beta + \varepsilon_i, \quad \varepsilon_i \overset{\text{IID}}{\sim} N(0, \sigma^2), \quad (i = 1, \ldots, n).$

Assumption 2 implies Assumption 1

2.3.2 Properties of the OLS estimator

Under Normal Linear Model assumption, we have:

- the properties derived under Gauss-Markov model assumption
- (the distribution of $\hat{\beta}$)
$$\hat{\beta} \sim N(\beta, \sigma^2 (X^\top X)^{-1}),$$
- (the joint distribution of $(\hat{Y}, \hat{\varepsilon})$, independent $\hat{Y}$ and $\hat{\varepsilon}$)
$$\begin{pmatrix} \hat{Y} \\ \hat{\varepsilon} \end{pmatrix} \sim N\left\{ \begin{pmatrix} X\beta \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} H & 0 \\ 0 & I_n - H \end{pmatrix} \right\},$$
so
$$\hat{Y} \perp\!\!\!\perp \hat{\varepsilon}.$$

2.3.3 pivotal quantities

2.3.4 Prediction based on pivotal quantities

If we want to predict future outcomes using OLS, which means for the pair of future data (x_{n+1}, y_{n+1}), we observe only x_{n+1} and want to predict y_{n+1} based on (X, Y) and x_{n+1}, we need to make assumption on the relationship between y_{n+1} and x_{n+1} as

$$y_{n+1} \sim N(x_{n+1}^\top \beta, \sigma^2)$$

with the same (β, σ^2).

Then we can make predictions for

- mean of y_{n+1}, which is $x_{n+1}^\top \beta$, a one-dimensional linear function of β
- y_{n+1}, which is a random variable

3 Population OLS

For the above analysis, we regard X as a fixed design matrix. This assumption is not essential, because we can condition on X even if we think X is random, as we always learned in the regression analysis or econometrics courses(e.g. turning $E(\varepsilon) = 0$ to $E(\varepsilon|X) = 0$).

Let's think about the question: do we first set/design X and then get Y? or do we get some samples from the whole distribution of X, Y, with X, Y of equal status and we just choose one as X and another to be Y?

It is worth mentioning Freedman's classification of OLS: the regression model and the correlation model. In the regression model, we first generate X and ε under some restrictions, and then generate the outcome based on $y = x^\top \beta + \varepsilon$, a linear function of x with error ε. In the correlation model, we start with a pair (x, y), then decompose y into the best linear predictor $x^\top \beta = x^\top \{E(xx^\top)\}^{-1} E(xy)$ and the leftover residual ε. The latter ensures $E(\varepsilon x) = 0$, but the former requires $E(\varepsilon \mid x) = 0$.

Note here ε is a random scalar variable.

In this section, we view each observation (x_i, y_i) as IID draws from a population and discuss population OLS.

3.1 conditional expectation

Predictor?

Why is the Conditional Expectation Function (CEF) important? Because the conditional expectation function is the solution to the MSE minimization problem(It achieves the smallest approximation error in the MSE sense).

References

[1] Ding, P. (2025) Linear model and extensions. Chapman & Hall.