Chapter 3

Least Squares

Notation: y_i is a scalar, and $x_i = (x_{i1}, \ldots, x_{iK})'$ is a $K \times 1$ vector. $Y = (y_1, \ldots, y_n)'$ is an $n \times 1$ vector, and

$$X = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{22} & \cdots & x_{nK} \end{bmatrix}$$

is an $n \times K$ matrix. I_n is an $n \times n$ identity matrix.

Ordinary least squares (OLS) is the most basic estimation technique in econometrics. It is simple and transparent. Understanding it thoroughly paves the way to study more sophisticated linear estimators. Moreover, many nonlinear estimators resemble the behavior of linear estimators in a neighborhood of the true value.

In this lecture, we study the finite sample properties of OLS. We first learn a series of facts from the linear algebra operation. Next, noticing that OLS coincides with the maximum likelihood estimator if the error term follows a normal distribution, we derive its finite-sample exact distribution which can be used for statistical inference. Finally, the Gauss-Markov theorem justifies the optimality of OLS under the classical assumptions.

To manipulate Leopold Kronecker's famous saying "God made the integers; all else is the works of man", I would say "Gauss made OLS; all else is the works of applied researchers." Popularity of OLS goes far beyond our dismal science. But be aware that OLS is a pure statistical or supervised machine learning method which reveals correlation instead of causality. Rather, economic theory hypothesizes causality while data are collected to test the theory or quantify the effect.

3.1 Algebra of Least Squares

3.1.1 OLS

As we have learned from the linear project model, the projection coefficient β in the regression

$$y = x'\beta + e$$

can be written as

$$\beta = \left(E\left[xx'\right]\right)^{-1} E\left[xy\right]. \tag{3.1}$$

We draw a pair of (y, x) from the joint distribution, and we mark it as (y_i, x_i) for $i = 1, \ldots, n$ repeated experiments. We possess a *sample* $(y_i, x_i)_{i=1}^n$.

Remark 3.1. Is (y_i, x_i) random or deterministic? Before we make the observation, they are treated as random variables whose realized values are uncertain. (y_i, x_i) is treated as random when we talk about statistical properties — statistical properties of a fixed number is meaningless. After we make the observation, they become deterministic values which cannot vary anymore.

Remark 3.2. In reality, we have at hand fixed numbers (more recently, words, photos, audio clips, video clips, etc., which can all be represented in digital formats with 0 and 1) to feed into a computational operation, and the operation will return one or some numbers. All statistical interpretation about these numbers are drawn from the probabilistic thought experiments. A *thought experiment* is an academic jargon for a *story* in plain language. Under the axiomatic approach of probability theory, such stories are mathematical consistent and coherent. But mathematics is a tautological system, not science. The scientific value of a probability model depends on how close it is to the *truth* or implications of the truth. In this course, we suppose that the data are generated from some mechanism, which is taken as the truth. In the linear regression model for example, the joint distribution of (y, x) is the truth, while we are interested in the linear projection coefficient β, which is an implication of the truth as in (3.1).

The sample mean is a natural estimator of the population mean. Replace the population mean $E\left[\cdot\right]$ in (3.1) by the sample mean $\frac{1}{n}\sum_{i=1}^{n}\cdot$, and the resulting estimator is

$$\begin{aligned}\widehat{\beta} &= \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} x_i y_i \\ &= \left(\frac{X'X}{n}\right)^{-1} \frac{X'y}{n} = \left(X'X\right)^{-1} X'y\end{aligned}$$

if $X'X$ is invertible. This is one way to motivate the OLS estimator.

Alternatively, we can derive the OLS estimator from minimizing the sum of squared residuals $\sum_{i=1}^{n}\left(y_i - x_i'b\right)^2$, or equivalently

$$Q\left(b\right) = \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - x_i'b\right)^2 = \frac{1}{2n}\left(Y - Xb\right)'\left(Y - Xb\right) = \frac{1}{2n}\left\|Y - Xb\right\|^2,$$

where the factor $\frac{1}{2n}$ is nonrandom and does not change the minimizer, and $\|\cdot\|$ is the Euclidean norm of a vector. Solve the first-order condition

$$\frac{\partial}{\partial b}Q\left(b\right) = \begin{bmatrix} \partial Q\left(b\right)/\partial b_1 \\ \partial Q\left(b\right)/\partial b_2 \\ \vdots \\ \partial Q\left(b\right)/\partial b_K \end{bmatrix} = -\frac{1}{n}X'\left(Y - Xb\right) = 0.$$

This necessary condition for optimality gives exactly the same $\widehat{\beta} = \left(X'X\right)^{-1} X'y$.

Moreover, the second-order condition

$$\frac{\partial^2}{\partial b \partial b'} Q(b) = \begin{bmatrix} \frac{\partial^2}{\partial b_1^2} Q(b) & \frac{\partial^2}{\partial b_2 \partial b_2} Q(b) & \cdots & \frac{\partial^2}{\partial b_K \partial b_1} Q(b) \\ \frac{\partial^2}{\partial b_1 \partial b_2} Q(b) & \frac{\partial^2}{\partial b_2^2} Q(b) & \cdots & \frac{\partial^2}{\partial b_K \partial b_2} Q(b) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial b_1 \partial b_K} Q(b) & \frac{\partial^2}{\partial b_2 \partial b_K} Q(b) & \cdots & \frac{\partial^2}{\partial b_K^2} Q(b) \end{bmatrix} = \frac{1}{n} X'X$$

shows that $Q(b)$ is convex in b due to the positive semi-definite matrix $X'X/n$. (The function $Q(b)$ is strictly convex in b if $X'X/n$ is positive definite.)

Remark 3.3. In the derivation of OLS we presume that the K columns in X are *linearly independent*, which means there is no $K \times 1$ vector b such that $b \neq 0_K$ and $Xb = 0_n$. Linear independence of the columns implies $n \geq K$ and the invertibility of $X'X/n$. Linear independence is violated when some regressors are *perfectly collinear*, for example when we use dummy variables to indicate categorical variables and put all these categories into the regression. Modern econometrics software automatically detects and reports perfect collinearity. What is treacherous is *nearly collinear*, meaning that the minimal eigenvalue of $X'X/n$ is close to 0, though not exactly equal to 0. We will talk about the consequence of near collinearity in the chapter of asymptotic theory.

Here are some definitions and properties of the OLS estimator.

- Fitted value: $\widehat{Y} = X\widehat{\beta}$.

- Projection matrix: $P_X = X\left(X'X\right)^{-1}X$; Residual maker matrix: $M_X = I_n - P_X$.

- $P_X X = X$; $X'P_X = X'$.

- $M_X X = 0_{n\times K}$; $X'M_X = 0_{K\times n}$.

- $P_X M_X = M_X P_X = 0_{n\times n}$.

- If $AA = A$, we call it an *idempotent* matrix. Both P_X and M_X are idempotent. All eigenvalues of an idempotent matrix must be either 1 or 0.

- $\text{rank}\left(P_X\right) = K$, and $\text{rank}\left(M_X\right) = n - K$ (See the Appendix of this chapter).

- Residual: $\widehat{e} = Y - \widehat{Y} = Y - X\widehat{\beta} = Y - X(X'X)^{-1}X'Y = (I_n - P_X)Y = M_X Y = M_X\left(X\beta + e\right) = M_X e$. Notice $\widehat{e}$ and e are two different objects.

- $X'\widehat{e} = X'M_X e = 0_K$.

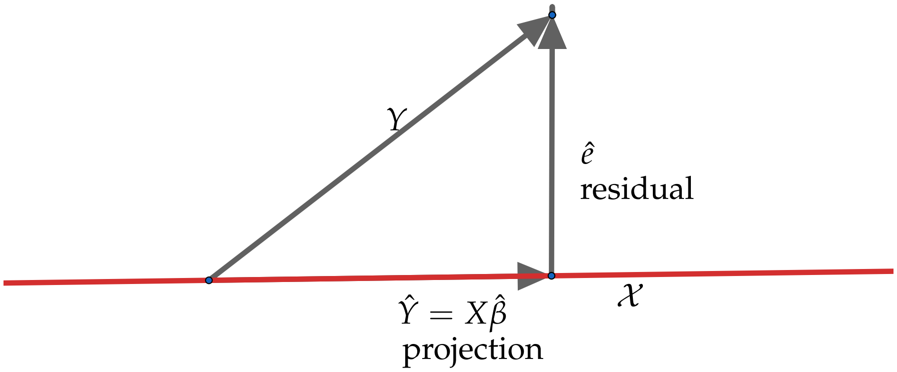
- $\sum_{i=1}^{n}\widehat{e}_i = 0$ if x_i contains a constant.

  (Because $X'\widehat{e} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \heartsuit & \heartsuit & \cdots & \heartsuit \\ \cdots & \cdots & \ddots & \vdots \\ \heartsuit & \heartsuit & \cdots & \heartsuit \end{bmatrix}\begin{bmatrix} \widehat{e}_1 \\ \widehat{e}_2 \\ \vdots \\ \widehat{e}_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ and the the first row implies $\sum_{i=1}^{n}\widehat{e}_i = 0$. "$\heartsuit$" indicates the entries irrelevant to our purpose.)

The operation of OLS bears a natural geometric interpretation. Notice $\mathcal{X} = \left\{Xb : b \in \mathbb{R}^K\right\}$ is the linear space spanned by the K columns of $X = [X_{\cdot 1}, \ldots, X_{\cdot K}]$, which is of K-dimension if the columns are linearly independent. The OLS estimator is the minimizer of $\min_{b \in \mathbb{R}^K} \|Y - Xb\|$ (Square the Euclidean norm or not does not change the minimizer because a^2 is a monotonic transformation for $a \geq 0$). In other words, $X\widehat{\beta}$ is the point in $\mathcal{X}$ such that it is the closest to the vector Y in terms of the Euclidean norm.

The relationship $Y = X\widehat{\beta} + \widehat{e}$ decomposes Y into two orthogonal vectors $X\widehat{\beta}$ and $\widehat{e}$ as $\left\langle X\widehat{\beta}, \widehat{e} \right\rangle = \widehat{\beta} X' \widehat{e} = 0'_K$, where $\langle \cdot, \cdot \rangle$ is the *inner product* of two vectors. Therefore $X\widehat{\beta}$ is the *projection* of Y onto $\mathcal{X}$, and $\widehat{e}$ is the corresponding *projection residuals.* The Pythagorean theorem implies

$$\|Y\|^2 = \|X\widehat{\beta}\|^2 + \|\widehat{e}\|^2 .$$



Example 3.1. Here is a simple simulated example to demonstrate the properties of OLS. Given $(x_{1i}, x_{2i}, x_{3i}, e_i)' \sim N(0_4, I_4)$, the dependent variable y_i is generated from

$$y_i = 0.5 + 2 \cdot x_{1i} - 1 \cdot x_{2i} + e_i$$

The researcher does not know x_{3i} is redundant, and he regresses y_i on $(1, x_{1i}, x_{2i}, x_{3i})$.

```
library(magrittr); set.seed(2020-9-23)
n = 20 # sample size
K = 4  # number of paramters
b0 = as.matrix( c(0.5, 2, -1, 0) ) # the true coefficient
X = cbind(1, matrix( rnorm(n * (K-1)), nrow = n ) )  # the regressor matrix
e = rnorm(n) # the error term
Y = X %*% b0 + e # generate the dependent variable
bhat = solve(t(X) %*% X, t(X) %*% Y ) %>% as.vector() %>% print()

## [1]  0.3151672  1.9546647 -0.8520387  0.1508770
```

The estimated coefficient $\widehat{\beta}$ is (0.315, 1.955, -0.852, 0.151). It is close to the true value, but not very accurate due to the small sample size.

```
ehat = Y - X %*% bhat
as.vector( t(X) %*% ehat ) %>% print()

## [1]  2.775558e-15  5.285658e-15 -7.193253e-15 -2.085963e-15

MX = diag(n) - X %*% solve( crossprod(X) ) %*% t(X)
data.frame(e = e, ehat = ehat, MXY = MX%*%Y, MXe = MX%*%e ) %>% head()

##              e       ehat        MXY        MXe
## 1  0.11468775  0.2195704  0.2195704  0.2195704
## 2 -1.09300952 -0.7358326 -0.7358326 -0.7358326
## 3  1.06084816  0.7873848  0.7873848  0.7873848
## 4 -0.93399293 -0.5797384 -0.5797384 -0.5797384

## 5  0.05697917  0.3604994  0.3604994  0.3604994

## 6  0.03431877  0.1489134  0.1489134  0.1489134

cat("The mean of the residual is ", mean(ehat), ".\n")

## The mean of the residual is  1.374064e-16 .

cat("The mean of the true error term is", mean(e), ".")

## The mean of the true error term is -0.1582708 .
```

3.1.2 Frish-Waugh-Lovell Theorem

The Frish-Waugh-Lovell (FWL) theorem is an algebraic fact about the formula of a subvector of the OLS estimator. To derive the FWL theorem we need to use the inverse of partitioned matrix. For a positive definite symmetric matrix $A = \begin{pmatrix} A_{11} & A_{12} \\ A_{12}' & A_{22} \end{pmatrix}$, the inverse can be written as

$$A^{-1} = \begin{pmatrix} \left(A_{11} - A_{12}A_{22}^{-1}A_{12}'\right)^{-1} & -\left(A_{11} - A_{12}A_{22}^{-1}A_{12}'\right)^{-1} A_{12}A_{22}^{-1} \\ -A_{22}^{-1}A_{12}'\left(A_{11} - A_{12}A_{22}^{-1}A_{12}'\right)^{-1} & \left(A_{22} - A_{12}'A_{11}^{-1}A_{12}\right)^{-1} \end{pmatrix}.$$

In our context of OLS estimator, let $X = \begin{pmatrix} X_1 & X_2 \end{pmatrix}$

$$
\begin{aligned}
\begin{pmatrix}\widehat{\beta}_1\\ \widehat{\beta}_2\end{pmatrix} &= \widehat{\beta} = (X'X)^{-1}X'Y\\
&= \left(\begin{pmatrix}X_1'\\ X_2'\end{pmatrix}\begin{pmatrix}X_1 & X_2\end{pmatrix}\right)^{-1}\begin{pmatrix}X_1'Y\\ X_2'Y\end{pmatrix}\\
&= \begin{pmatrix}X_1'X_1 & X_1'X_2\\ X_2'X_1 & X_2'X_2\end{pmatrix}^{-1}\begin{pmatrix}X_1'Y\\ X_2'Y\end{pmatrix}\\
&= \begin{pmatrix}\left(X_1'M_{X_2}'X_1\right)^{-1} & -\left(X_1'M_{X_2}'X_1\right)^{-1}X_1'X_2\left(X_2'X_2\right)^{-1}\\ \heartsuit & \heartsuit\end{pmatrix}\begin{pmatrix}X_1'Y\\ X_2'Y\end{pmatrix}.
\end{aligned}
$$

The subvector

$$
\begin{aligned}
\widehat{\beta}_1 &= \left(X_1'M_{X_2}'X_1\right)^{-1}X_1'Y - \left(X_1'M_{X_2}'X_1\right)^{-1}X_1'X_2\left(X_2'X_2\right)^{-1}X_2'Y\\
&= \left(X_1'M_{X_2}'X_1\right)^{-1}X_1'Y - \left(X_1'M_{X_2}'X_1\right)^{-1}X_1'P_{X_2}Y\\
&= \left(X_1'M_{X_2}'X_1\right)^{-1}\left(X_1'Y - X_1'P_{X_2}Y\right)\\
&= \left(X_1'M_{X_2}'X_1\right)^{-1}X_1'M_{X_2}Y.
\end{aligned}
$$

Notice that $\widehat{\beta}_1$ can be obtained by the following:

1. Regress Y on X_2, obtain the residual $\tilde{Y}$;

2. Regress X_1 on X_2, obtain the residual $\tilde{X}_1$;

3. Regress $\tilde{Y}$ on $\tilde{X}_1$, obtain OLS estimates $\widehat{\beta}_1$.

Similar derivation can also be carried out in the population linear projection. See Hansen (2020) [E] Chapter 2.22-23.

```
X1 = X[,1:2];X2 = X[,3:4]
PX2 = X2 %*% solve( t(X2) %*% X2) %*% t(X2)
MX2 = diag(rep(1,n)) - PX2


bhat1 <- (solve(t(X1)%*% MX2 %*% X1, t(X1) %*% MX2 %*% Y )) %>%
  as.vector() %>% print()

## [1] 0.3151672 1.9546647

ehat1 = MX2 %*% Y - MX2 %*% X1 %*% bhat1
data.frame(ehat = ehat, ehat1 = ehat1) %>% head() %>% print()

##         ehat      ehat1
## 1  0.2195704  0.2195704
## 2 -0.7358326 -0.7358326
## 3  0.7873848  0.7873848
## 4 -0.5797384 -0.5797384
## 5  0.3604994  0.3604994
## 6  0.1489134  0.1489134
```

3.1.3 Goodness of Fit

Consider the regression with the intercept $Y = X_1\beta_1 + \beta_2 + e$. The OLS estimator gives

$$Y = \widehat{Y} + \widehat{e} = \left(X_1\widehat{\beta}_1 + \widehat{\beta}_2\right) + \widehat{e}. \tag{3.2}$$

Applying the FWL theorem with $X_2 = \iota$, where ι (Greek letter, iota) is an $n \times 1$ vector of 1's. Then $M_{X_2} = M_\iota = I_n - \frac{1}{n}\iota\iota'$. Notice M_ι is the *demeaner* in that $M_\iota z = z - \bar{z}$. It subtract the vector mean $\bar{z} = \frac{1}{n}\sum_{i=1}^n z_i$ from the original vector z. The above three-step procedure becomes

1. Regress Y on ι, and the residual is $M_\iota Y$;
2. Regress X_1 on ι, and the residual is $M_\iota X_1$;
3. Regress $M_\iota Y$ on $M_\iota X_1$, and the OLS estimates is exactly the same as $\widehat{\beta}_1$ in (3.2).

The last step gives the decomposition

$$M_\iota Y = M_\iota X_1 \widehat{\beta}_1 + \tilde{e}, \tag{3.3}$$

and the Pythagorean theorem implies

$$\|M_\iota Y\|^2 = \|M_\iota X_1 \widehat{\beta}_1\|^2 + \|\widehat{e}\|^2 .$$

Exercise 3.1. Show that $\widehat{e}$ in (3.2) is exactly the same as $\tilde{e}$ in (3.3).

R-squared is a popular measure of goodness-of-fit in the linear regression. The (in-sample) R-squared

$$R^2 = \frac{\|M_\iota X_1 \widehat{\beta}_1\|^2}{\|M_\iota Y\|^2} = 1 - \frac{\|\tilde{e}\|^2}{\|M_\iota Y\|^2}.$$

is well defined only when a constant is included in the regressors.

Exercise 3.2. Show

$$R^2 = \frac{\widehat{Y}' M_\iota \widehat{Y}}{Y' M_\iota Y} = \frac{\sum_{i=1}^n (\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^n (y_i - \overline{y})^2}$$

as in the decomposition (3.2). In other words, it is the ratio between the sample variance of $\widehat{Y}$ and the sample variance of Y.

The magnitude of R-squared varies in different contexts. In macro models with the lagged dependent variables, it is not unusually to observe R-squared larger than 90%. In cross sectional regressions it is often below 20%.

Exercise 3.3. Consider a short regression "regress y_i on x_{1i}" and a long regression "regress y_i on (x_{1i}, x_{2i})". Given the same dataset (Y, X_1, X_2), show that the R-squared from the short regression is no larger than that from the long regression. In other words, we can always (weakly) increase R^2 by adding more regressors.

Conventionally we consider the regressions when the number of regressors K is much smaller the sample size n. In the era of big data, it can happen that we have more potential regressors than the sample size.

Exercise 3.4. Show $R^2 = 1$ when $K \geq n$. (When $K > n$, the matrix $X'X$ must be rank deficient. We can generalize the definition OLS fitting as any vector that minimizes $\|Y - Xb\|^2$ though the minimizer is not unique.

```
n = 5; K = 6;
Y = rnorm(n)
X = matrix( rnorm(n*K), n)
summary( lm(Y~X) )

##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
## ALL 5 residuals are 0: no residual degrees of freedom!
##
## Coefficients: (2 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.2229         NA      NA       NA
## X1           -0.6422         NA      NA       NA
## X2            0.1170         NA      NA       NA
## X3            1.1844         NA      NA       NA
## X4            0.5883         NA      NA       NA
## X5                NA         NA      NA       NA
## X6                NA         NA      NA       NA
##
## Residual standard error: NaN on 0 degrees of freedom
## Multiple R-squared:      1,Adjusted R-squared:    NaN
## F-statistic:   NaN on 4 and 0 DF,  p-value: NA
```

With a new dataset $(Y^{\text{new}}, X^{\text{new}})$, the *out-of-sample* (OOS) R-squared is

$$OOS\ R^2 = \frac{\widehat{\beta}' X^{\text{new}\prime} M_\iota X^{\text{new}} \widehat{\beta}}{Y^{\text{new}\prime} M_\iota Y^{\text{new}}}.$$

OOS R-squred measures the goodness of fit in a new dataset given the coefficient estimated from the original data. In financial market shorter-term predictive models, a person may become a billion if he can systematically achieve 2% OOS R-squared.

3.2 Statistical Properties of Least Squares

In this section we return to the classical statistical framework under restrictive distributional assumption

$$y_i | x_i \sim N\left(x_i' \beta, \gamma\right) \tag{3.4}$$

, where $\gamma = \sigma^2$ to ease the differentiation. This assumption is equivalent to $e_i|x_i = (y_i - x_i'\beta)\,|x_i \sim N(0,\gamma)$. Because the distribution of e_i is invariant to x_i, the error term $e_i \sim N(0,\gamma)$ and is statistically independent of x_i. This is a very strong assumption.

3.2.1 Maximum Likelihood Estimation

The likelihood of observing a pair (y_i, x_i) is

$$\begin{aligned} f_{yx}(y_i, x_i) &= f_{y|x}(y_i|x_i)\, f_x(x) \\ &= \frac{1}{\sqrt{2\pi\gamma}} \exp\left(-\frac{1}{2\gamma}\left(y_i - x_i'\beta\right)^2\right) \times f_x(x), \end{aligned}$$

where f_{yx} is the joint pdf, $f_{y|x}$ is the conditional pdf and f_x is the marginal pdf of x, and the second equality holds under the assumption (3.4). The likelihood a random sample $(y_i, x_i)_{i=1}^n$ is

$$\begin{aligned} \prod_{i=1}^n f_{yx}(y_i, x_i) &= \prod_{i=1}^n f_{y|x}(y_i|x_i)\, f_x(x) \\ &= \prod_{i=1}^n f_{y|x}(y_i|x_i) \times \prod_{i=1}^n f_x(x) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\gamma}} \exp\left(-\frac{1}{2\gamma}\left(y_i - x_i'\beta\right)^2\right) \times \prod_{i=1}^n f_x(x). \end{aligned}$$

The parameters of interest (β, γ) are irrelevant to the second term $\prod_{i=1}^{n} f_x(x)$ for they appear only in the conditional likelihood

$$\prod_{i=1}^{n} f_{y|x}(y_i|x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\gamma}} \exp\left(-\frac{1}{2\gamma}\left(y_i - x_i'\beta\right)^2\right).$$

We focus on the conditional likelihood. To facilitate derivation, we work with the conditional log-likelihood function

$$L(\beta, \gamma) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\gamma - \frac{1}{2\gamma}\sum_{i=1}^{n}\left(y_i - x_i'\beta\right)^2,$$

for $\log(\cdot)$ is a monotonic transformation that does not change the maximizer. The maximum likelihood estimator $\widehat{\beta}_{MLE}$ can be found using the FOC:

$$\frac{\partial}{\partial\beta}L(\beta, \gamma) = \frac{1}{2\gamma}\sum_{i=1}^{n} 2x_i\left(y_i - x_i'\beta\right)^2 = \frac{1}{\gamma}\sum_{i=1}^{n} x_i\left(y_i - x_i'\beta\right)^2 = 0.$$

Rearranging the above equation in matrix form $X'Y = X'X\widehat{\beta}_{MLE}$, we explicitly solve

$$\widehat{\beta}_{MLE} = (X'X)^{-1}X'Y.$$

The maximum likelihood estimator (MLE) coincides with the OLS estimator. Similarly, the other FOC with respect to γ gives $\widehat{\gamma}_{\mathrm{MLE}} = \widehat{e}'\widehat{e}/n$.

3.2.2 Classical Finite Sample Distribution

We can show the finite-sample exact distribution of $\widehat{\beta}$ assuming the error term follows a Gaussian distribution. *Finite sample distribution* means that the distribution holds for any n; it is in contrast to *asymptotic distribution*, which is a large sample approximation to the finite sample distribution. We first review some properties of a generic jointly normal random vector.

Fact 3.1. *Let $z \sim N(\mu, \Omega)$ be an $l \times 1$ random vector with a positive definite variance-covariance matrix Ω. Let A be an $m \times l$ non-random matrix where $m \leq l$. Then $Az \sim N(A\mu, A\Omega A')$.*

Fact 3.2. *If $z \sim N(0,1)$, $w \sim \chi^2(d)$ and z and w are independent. Then $\frac{z}{\sqrt{w/d}} \sim t(d)$.*

The OLS estimator

$$\widehat{\beta} = \left(X'X\right)^{-1} X'Y = \left(X'X\right)^{-1} X'\left(X'\beta + e\right) = \beta + \left(X'X\right)^{-1} X'e,$$

and its conditional distribution can be written as

$$\begin{aligned}
\widehat{\beta}|X &= \beta + \left(X'X\right)^{-1} X'e|X \\
&\sim \beta + \left(X'X\right)^{-1} X' \cdot N\left(0_n, \sigma^2 \cdot I_n\right) \\
&\sim N\left(\beta, \sigma^2 \left(X'X\right)^{-1} X'X \left(X'X\right)^{-1}\right) \sim N\left(\beta, \sigma^2 \left(X'X\right)^{-1}\right)
\end{aligned}$$

by Fact 3.1. The k-th element of the vector coefficient

$$\widehat{\beta}_k|X = \eta_k'\widehat{\beta}|X \sim N\left(\beta_k, \sigma^2\eta_k'\left(X'X\right)^{-1}\eta_k\right) \sim N\left(\beta_k, \sigma^2\left(X'X\right)_{kk}^{-1}\right),$$

where $\eta_k = \left(1\left\{l = k\right\}\right)_{l=1,\dots,K}$ is the selector of the k-th element.

In reality, σ^2 is an unknown parameter, and

$$s^2 = \widehat{e}'\widehat{e}/\left(n-K\right) = e'M_Xe/\left(n-K\right)$$

is an unbiased estimator of σ^2. Consider the t-statistic

$$\begin{aligned} T_k &= \frac{\widehat{\beta}_k - \beta_k}{\sqrt{s^2\left[\left(X'X\right)^{-1}\right]_{kk}}} = \frac{\widehat{\beta}_k - \beta_k}{\sqrt{\sigma^2\left[\left(X'X\right)^{-1}\right]_{kk}}} \cdot \frac{\sqrt{\sigma^2}}{\sqrt{s^2}} \\ &= \frac{\left(\widehat{\beta}_k - \beta_k\right)/\sqrt{\sigma^2\left[\left(X'X\right)^{-1}\right]_{kk}}}{\sqrt{\frac{e'}{\sigma}M_X\frac{e}{\sigma}/\left(n-K\right)}}. \end{aligned}$$

The numerator follows a standard normal, and the denominator follows $\frac{1}{n-K}\chi^2\left(n-K\right)$. Moreover, the numerator and the denominator are statistically independent (See Section 3.4.2). As a result, we conclude $T_k \sim t\left(n-K\right)$ by Fact 3.2. This finite sample distribution allows us to conduct statistical inference.

3.2.3 Mean and Variance

Now we relax the normality assumption and statistical independence. Instead, we represent the regression model as $Y = X\beta + e$ and

$$\begin{aligned} E[e|X] &= 0_n \\ \text{var}\left[e|X\right] &= E\left[ee'|X\right] = \sigma^2 I_n. \end{aligned}$$

where the first condition is the *mean independence* assumption, and the second condition is the *homoskedasticity* assumption. These assumptions are about the first and second *moments* of e_i conditional on x_i. Unlike the normality assumption, they do not restrict the distribution of e_i.

- Unbiasedness:

$$\begin{aligned} E\left[\widehat{\beta}|X\right] &= E\left[\left(X'X\right)^{-1} XY|X\right] = E\left[\left(X'X\right)^{-1} X\left(X'\beta + e\right)|X\right] \\ &= \beta + \left(X'X\right)^{-1} XE\left[e|X\right] = \beta. \end{aligned}$$

  Unbiasedness does not rely on homoskedasticity.

- Variance:

$$\begin{aligned}
\operatorname{var}\left[\widehat{\beta}|X\right] &= E\left[\left(\widehat{\beta}-E\widehat{\beta}\right)\left(\widehat{\beta}-E\widehat{\beta}\right)'|X\right] \\
&= E\left[\left(\widehat{\beta}-\beta\right)\left(\widehat{\beta}-\beta\right)'|X\right] \\
&= E\left[\left(X'X\right)^{-1}X'ee'X\left(X'X\right)^{-1}|X\right] \\
&= \left(X'X\right)^{-1}X'E\left[ee'|X\right]X\left(X'X\right)^{-1}
\end{aligned}$$

  where the second equality holds as $E\left[\widehat{\beta}\right]=E\left[E\left[\widehat{\beta}|X\right]\right]=\beta$. Under the assumption of homoskedasticity, it can be simplified as

$$\operatorname{var}\left[\widehat{\beta}|X\right]=\left(X'X\right)^{-1}X'\left(\sigma^2 I_n\right)X\left(X'X\right)^{-1}=\sigma^2\left(X'X\right)^{-1}.$$

Example 3.2. (Heteroskedasticity) If $e_i = x_i u_i$, where x_i is a scalar random variable, u_i is statistically independent of x_i, $E\left[u_i\right]=0$ and $E\left[u_i^2\right]=\sigma^2$. Then $E\left[e_i|x_i\right]=0$ but $E\left[e_i^2|x_i\right]=\sigma^2 x_i^2$ is a function of x_i. We say e_i^2 is a heteroskedastic error.

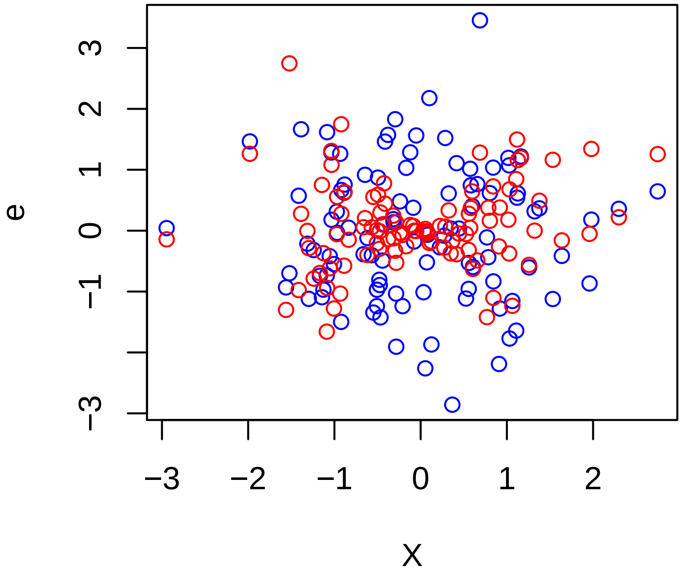
```r
n = 100; X = rnorm(n)

e1 = rnorm(n);

plot( y = e1, x = X, col = "blue", ylab = "e")

e2 = X * rnorm(n);

points( y = e2, x = X, col = "red")
```



It is important to notice that independently and identically distributed sample (iid) (y_i, x_i) does not imply homoskedasticity. Homoskedasticity or heterskdasticity is about the relationship between $(x_i, e_i = y_i - \beta x)$, whereas iid is about the relationship between (y_i, x_i) and (y_j, x_j) for $i \neq j$.

3.2.4 Gauss-Markov Theorem

Gauss-Markov theorem is concerned about the optimality of OLS. It justifies OLS as the efficient estimator among all linear unbiased ones. *Efficient* here means that it enjoys the smallest variance in a family of estimators.

We have shown that OLS is unbiased in that $E\left[\widehat{\beta}\right] = \beta$. There are numerous linearly unbiased estimators. For example, $\left(Z'X\right)^{-1} Z'y$ for $z_i = x_i^2$ is unbiased because $E\left[\left(Z'X\right)^{-1} Z'y\right] = E\left[\left(Z'X\right)^{-1} Z'\left(X\beta + e\right)\right] = \beta$. We cannot say OLS is better than those other unbiased estimators because they are equally good in this aspect. Thus, we move to the second order property of variance: an estimator is better if its variance is smaller.

Fact 3.3. *For two generic random vectors X and Y of the same size, we say X's variance is smaller or equal to Y's variance if $\left(\Omega_Y - \Omega_X\right)$ is a positive semi-definite matrix. The comparison is defined this way because for any non-zero constant vector c, the variance of the linear combination of X*

$$\mathrm{var}\left(c'X\right) = c'\Omega_X c \leq c'\Omega_Y c = \mathrm{var}\left(c'Y\right)$$

is no bigger than the same linear combination of Y.

Let $\tilde{\beta} = A'y$ be a generic linear estimator, where A is any $n \times K$ functions of X. As

$$E\left[A'y|X\right] = E\left[A'\left(X\beta + e\right)|X\right] = A'X\beta.$$

So the linearity and unbiasedness of $\tilde{\beta}$ implies $A'X = I_n$. Moreover, the variance

$$\mathrm{var}\left(A'y|X\right) = E\left[\left(A'y - \beta\right)\left(A'y - \beta\right)'|X\right] = E\left[A'ee'A|X\right] = \sigma^2 A'A.$$

Let $C = A - X\left(X'X\right)^{-1}$.

$$\begin{aligned} A'A - \left(X'X\right)^{-1} &= \left(C + X\left(X'X\right)^{-1}\right)'\left(C + X\left(X'X\right)^{-1}\right) - \left(X'X\right)^{-1} \\ &= C'C + \left(X'X\right)^{-1}X'C + C'X\left(X'X\right)^{-1} \\ &= C'C, \end{aligned}$$

where the last equality follows as

$$\left(X'X\right)^{-1}X'C = \left(X'X\right)^{-1}X'\left(A - X\left(X'X\right)^{-1}\right) = \left(X'X\right)^{-1} - \left(X'X\right)^{-1} = 0.$$

Therefore $A'A - \left(X'X\right)^{-1}$ is a positive semi-definite matrix. The variance of any $\tilde{\beta}$ is no smaller than the OLS estimator $\widehat{\beta}$. The above derivation shows OLS achieves the smallest variance among all linear unbiased estimators.

Homoskedasticity is a restrictive assumption. Under homoskedasticity, $\mathrm{var}\left[\widehat{\beta}\right] = \sigma^2\left(X'X\right)^{-1}$. Popular estimator of σ^2 is the sample mean of the residuals $\widehat{\sigma}^2 = \frac{1}{n}\widehat{e}'\widehat{e}$ or the unbiased one $s^2 = \frac{1}{n-K}\widehat{e}'\widehat{e}$. Under heteroskedasticity, Gauss-Markov theorem does not apply.

3.3 Summary

The linear algebraic properties holds in finite sample no matter the data are taken as fixed numbers or random variables. The exact distribution under the normality assumption of the error term is the classical statistical results. The Gauss Markov theorem holds under two crucial assumptions: linear CEF and homoskedasticity.

Historical notes: Carl Friedrich Gauss (1777–1855) claimed he had come up with the operation of OLS in 1795. With only three data points at hand, Gauss successfully applied his method to predict the location of the dwarf planet Ceres in 1801. While Gauss did not publish the work on OLS until 1809, Adrien-Marie Legendre (1752–1833) presented this method in 1805. Today people tend to attribute OLS to Gauss, assuming that a giant like Gauss had no need to tell a lie to steal Legendre's discovery.

MLE was promulgated and popularized by Ronald Fisher (1890–1962). He was a major contributor of the frequentist approach which dominates mathematical statistics today, and he sharply criticized the Bayesian approach. Fisher collected the iris flower dataset of 150 observations in his biological study in 1936, which can be displayed in R by typing `iris`. Fisher invented the many concepts in classical mathematical statistics, such as sufficient statistic, ancillary statistic, completeness, and exponential family, etc.

3.4 Appendix

3.4.1 Idempotent Matrix

Let A be any $n \times K$ generic real matrix. *Singular value decomposition* (SVD) factorizes $A = USV'$, where U is an $n \times n$ real unitary matrix (A real unitary matrix is invertible and $U'U = UU' = I$, which implies $U^{-1} = U'$), $S = \begin{bmatrix} S_1 \\ 0_{(n-K)\times K} \end{bmatrix}$ is an $n \times K$ rectangular diagonal matrix with S_1 a $K \times K$ diagonal matrix of non-negative real elements (called *singular values*), and V is a $K \times K$ real unitary matrix.

We apply SVD to the projection matrix $P_X = X\left(X'X\right)^{-1} X$, where X is an $n \times K$ data matrix with K linearly independent columns. Substitute $X = USV'$ into P_X:

$$
\begin{aligned}
P_X &= USV'\left(VS'U'USV'\right)^{-1} VS'U' = USV'\left(VS'SV'\right)^{-1} VS'U' \\
&= USV'V'^{-1}\left(S'S\right)^{-1} V^{-1}VS'U' = US\left(S'S\right)^{-1} S'U' \\
&= U \begin{bmatrix} S_1 \\ 0 \end{bmatrix} S_1^{-1}S_1^{-1} \begin{bmatrix} S_1 & 0 \end{bmatrix} U' = U \begin{bmatrix} I_K & 0_{K\times(n-K)} \\ 0_{(n-K)\times K} & 0_{(n-K)\times(n-K)} \end{bmatrix} U' \\
&= U \operatorname{diag}\left(\iota_K, 0_{n-K}\right) U'.
\end{aligned}
$$

All real symmetric matrices are diagonalizable, and the the last expression is the diagonalization of P_X. The projection matrix P_X has K repeated eigenvalues of 1 and $(n-K)$ repeated eigenvalues of 0, and obviously rank$(P_X) = K$.

Two generic square matrices A and B are *similar* if there exists an invertible matrix Q such that $A = Q^{-1}BQ$. By this definition, P_X is similar to the diagonal matrix diag $(\iota_K, 0_{n-K})$, and $M_X = I_n - P_X$ is similar to diag $(0_K, \iota_{n-K})$ because

$$\begin{aligned} U' M_X U &= U' \ (I_n - P_X) \, U = U'U - U' P_X U \\ &= I_n - \text{diag}\,(\iota_K, 0_{n-K}) = \text{diag}\,(0_K, \iota_{n-K}) \, . \end{aligned}$$

It implies that rank$(M_X) = n - K$.

Both P_X and M_X are symmetric idempotent matrices. For a general idempotent matrices C which does not have to be symmetric,

- C is diagonalizable (See Horn and Johnson (1985, p.148)).

This fact immediately implies that

- All eigenvalues of C are either 0 and 1;
- rank (C) = trace (C).

3.4.2 Basu's Theorem

$Y = (y_1, \ldots, y_n)$ consists of n iid observations. We say $T\,(Y)$ is a sufficient statistic for a parameter θ if the conditional probability $f\,(Y|T\,(Y))$ does not depend on θ. For example, for $y_i \sim N\left(\mu, \sigma^2\right)$ with known σ^2 and unknown μ, We verify that the sample mean $\bar{y} = n^{-1}\sum_{i=1}^{n} y_i$ is a sufficient statistic for μ. Notice that the joint density of Y is

$$\begin{aligned} f\left(Y\right) &= \left(2\pi\sigma\right)^{-n/2}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left(y_i-\mu\right)^2\right) \\ &= \left(2\pi\sigma^2\right)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2\right)\exp\left(-\frac{n}{2\sigma^2}\left(\bar{y}-\mu\right)^2\right). \end{aligned}$$

Because $\bar{y} \sim N\left(\mu, \sigma^2/n\right)$, the marginal density is

$$f\left(\bar{y}\right) = \left(2\pi\sigma^2/n\right)^{-1/2}\exp\left(-\frac{1}{2\sigma^2/n}\left(\bar{y}-\mu\right)^2\right).$$

The conditional density is

$$f\left(Y|\bar{y}\right) = \frac{f\left(Y\right)}{f\left(\bar{y}\right)} = \frac{\left(2\pi\sigma^2\right)^{-n/2}}{\left(2\pi\sigma^2/n\right)^{-1/2}}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2\right)$$

is independent of μ, and thus $\bar{y}$ is a sufficient statistic for μ.

In the meantime, the sample standard deviation $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i-\bar{y}\right)$ is an *ancillary statistic* for μ , because the distribution of s^2 does not depend on μ.

Basu's theorem says that a *complete* sufficient statistic is statistically independent from any ancillary statistic. For a normal distribution with unknown mean and known variance, the sample mean $\bar{y}$ is the sufficient statistic and the sample standard deviation s^2 is an ancillary statistic.

`Zhentao Shi. Oct 3.`

Bibliography

Horn, R. A. and C. R. Johnson (1985). *Matrix analysis*. Cambridge University Press.