SDS 4130/5130: Linear Statistical Models

Topics 3: Multiple Linear Regression Model II

Formulas and Properties Of The OLSE

Maxine Yu[1]

[1] Department of Statistics and Data Science
https://maxineyu.github.io/personal_web/

Washington University in St. Louis

September 2, 2025

Today's Class

1. Multiple Linear Regression Model
   - Formulas For the OLSE $\hat{\beta}$
   - Existence of the OLSE $\hat{\beta}$
   - Properties of OLS Estimators when $\mathbf{X}'\mathbf{X}$ is invertible
   - Gauss-Markov Theorem

Ordinary Least Squares Estimators (OLSE) I

- Recall that the OLSE $\hat{\boldsymbol{\beta}} = [\hat{\beta}_0, \ldots, \hat{\beta}_k]'$ minimizes the SSE

$$S(\tilde{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \tilde{e}_i^2 := \sum_{i=1}^{n} \Big(y_i - (\tilde{\beta}_0 + \tilde{\beta}_1 x_{i1} + \cdots + \tilde{\beta}_k x_{ik}) \Big)^2$$

  over all possible $\tilde{\boldsymbol{\beta}} = [\tilde{\beta}_0, \ldots, \tilde{\beta}_k]'$.[1]

- $\tilde{e}_i := y_i - (\tilde{\beta}_0 + \tilde{\beta}_1 x_{i1} + \cdots + \tilde{\beta}_k x_{ik})$ represents the fitting error for the i^{th} observation point $(x_{i1}, \ldots, x_{ik}, y_i)$ incurred by the tentative regression formula $y = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \cdots + \tilde{\beta}_k x_k$.

[1] Hereafter, $\boldsymbol{A}'$ or $\boldsymbol{A}^T$ denote the transpose of a matrix or vector $\boldsymbol{A}$.

Normal Equations

From basic multivariate calculus, the OLSE $\hat{\beta}$ must satisfy the following equations (called the normal equations):

$$\frac{\partial S(\hat{\boldsymbol{\beta}})}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) = 0,$$

$$\frac{\partial S(\hat{\boldsymbol{\beta}})}{\partial \hat{\beta}_\ell} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) x_{i\ell} = 0, \quad \ell = 1, \ldots, k.$$

This is a linear system of $k + 1$ equations with $k + 1$ unknowns $(\hat{\beta}_0, \ldots, \hat{\beta}_k)$.

We can write the equations above as follows:

$$n\hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j \sum_{i=1}^{n} x_{ij} = \sum_{i=1}^{n} y_i,$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{i\ell} + \sum_{j=1}^{k} \hat{\beta}_j \sum_{i=1}^{n} x_{ij} x_{i\ell} = \sum_{i=1}^{n} y_i x_{i\ell}, \quad \ell = 1, \ldots, k.$$

Matrix Form Of Normal Equations

Note that we can write the equations above in matrix form as:

$$\begin{bmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} & \dots & \sum_{i=1}^n x_{ik} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} & \dots & \sum_{i=1}^n x_{i1}x_{ik} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum_{i=1}^n x_{ik} & \sum_{i=1}^n x_{ik}x_{i1} & \sum_{i=1}^n x_{ik}x_{i2} & \dots & \sum_{i=1}^n x_{ik}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i1}y_i \\ \vdots \\ \sum_{i=1}^n x_{ik}y_i \end{bmatrix}$$

Or, using the design matrix $\boldsymbol{X}$ (check that $\boldsymbol{X}'\boldsymbol{X}$ is the matrix on the left above),

$$\boldsymbol{X}'\boldsymbol{X}\hat{\beta} = \boldsymbol{X}'\boldsymbol{y}.$$

where recall that $$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix}.$$

Formula For The OLSE

The system of equations (called Normal Equations)

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

can be solved by pre-multiplying both sides by the inverse of the $p \times p$ matrix $\boldsymbol{X}'\boldsymbol{X}$:

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

provided that the inverse of $\boldsymbol{X}'\boldsymbol{X}$ exists!!!

One Predictor Case $k = 1$

- As shown on the top of p. 6,

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} n & \sum_{i=1}^n x_{i1} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 \end{bmatrix}, \qquad \boldsymbol{X}'\boldsymbol{y} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n y_i x_{i1} \end{bmatrix}$$

- Recall the following shortcut formula for the inverse of a 2×2 matrix:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

- Then, $\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ takes the form

$$\hat{\beta} = \frac{1}{n\sum_{i=1}^n x_{i1}^2 - (\sum_{i=1}^n x_{i1})^2} \begin{bmatrix} \sum_{i=1}^n x_{i1}^2 & -\sum_{i=1}^n x_{i1} \\ -\sum_{i=1}^n x_{i1} & n \end{bmatrix} \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n y_i x_{i1} \end{bmatrix}$$

- We then recover the formula for $\hat{\beta}_1$ from Lecture 1 (p. 26):

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n}}{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}},$$

Toy Example

Suppose that we want to predict Y = weight as a linear function of X_1 = height, X_2 = Age, and X_3 = gender (1=male, 0=female).

We sample $n = 6$ students and collected the data:

Height	Age	Gender	Weight
5	19	1	120
5	21	0	145
7	20	1	175
6	19	0	110
5	18	1	95
7	17	1	180

Toy Example. Cont...

Then, the design matrix and the response vector are

$$\boldsymbol{X} = \begin{bmatrix} 1 & 5 & 19 & 1 \\ 1 & 5 & 21 & 0 \\ 1 & 7 & 20 & 1 \\ 1 & 6 & 19 & 0 \\ 1 & 5 & 18 & 1 \\ 1 & 7 & 17 & 1 \end{bmatrix} \quad \boldsymbol{y} = \begin{bmatrix} 120 \\ 145 \\ 175 \\ 110 \\ 95 \end{bmatrix}$$

Toy Example. Cont...

The R code to find the OLSE is as follows:

```
(X=matrix(c(1,1,1,1,1,1,5,5,7,6,5,7,19,21,20,19,18,17,1,0,1,0,1,1),6,4))
(y=matrix(c(120,145,175,110,95,180),6,1))
(t(X)%*%X)
(C=solve(t(X)%*%X))
(b=t(X)%*%y)
(beta=C%*%b)
```

$$\hat{\beta} = \begin{bmatrix} -232.704918 \\ 29.426230 \\ 9.918033 \\ 15.163934 \end{bmatrix}$$

Some Interpretations :

$\longleftarrow$ On average, you gain 9 lb each year.

$\longleftarrow$ Males on average weight 15 lb more than females.

Example 1

The goal is to compute the OLSE using matrix multiplications and compare it to the one obtained from the function `lm()`.

Recall the Example 1 from Lect. 2 regarding sale prices of houses.

```
# Regress House Sale Price against x1=taxes (in thousand do
# and x3 lot size (in thousand of square feet)
> library(MPV)
> data(table.b4)
> mymodel<-lm(y~x1+x3,data=table.b4)
> summary(mymodel)
```

Example 1. Output of `lm()`

```
Call:
lm(formula = y ~ x1 + x3, data = table.b4)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1578 -2.5246 -0.0439  1.6875  6.1523

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.1739     2.6209   5.026 5.63e-05 ***
x1            3.0971     0.5466   5.666 1.27e-05 ***
x3            0.2656     0.4401   0.603    0.553
---
```

Example 1. Obtaining $\hat{\beta}$ by Matrix Multiplication

```
> X<-model.matrix(mymodel) # Extract Design Matrix
> C<-solve(t(X)%*% X)       # Inverse of X'X
> y<-table.b4$y             # Extract the vector y
> (OLSE<-C%*%t(X)%*%y)      # Compute the OLSE
                  [,1]
(Intercept) 13.1739426
x1           3.0970720
x3           0.2655657
```

The same values as those obtained using `lm` function shown in the previous page.

Characterization Of The OLSE

The following result states that the OLSE are always solutions of the normal equations and vice versa.

Theorem (OLSE as Solutions Of Normal Equations)

$\hat{\beta}$ *is such that* $\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{y}$ *if and only if* $\hat{\beta}$ *minimizes the SSE* $S(\tilde{\beta})$

The proof is given in Appendix A.

Existence of the OLSE $\hat{\beta}$

The previous result shows that the key for the existence of the OLSE $\hat{\beta}$ is the existence of a solution to the Normal Equations:

$$\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{y} \tag{$\star$}$$

The following result shows that such solutions always exist regardless of the design matrix $\boldsymbol{X}$:

Theorem (The Normal Equations are Consistent)

The normal equations ($\star$) are always consistent; i.e., there always exists a $\hat{\beta}$ satisfying the system of equations ($\star$) and, thus, there is always an OLSE.

There is a way to show this using linear algebra techniques. However, a more natural way is to realize that the OLSE must exist because of its geometric interpretation, as shown in the Appendix B.

Mean Vector and Covariance Matrix

- Suppose that $\boldsymbol{U} = [U_1, \ldots, U_m]'$ is an $m \times 1$ random vector. The **mean vector** of $\boldsymbol{U}$ is the $m \times 1$ vector:

$$\boldsymbol{\mu}_{\boldsymbol{U}} = E(\boldsymbol{U}) = [E(U_1), \ldots, E(U_m)]'.$$

- We can similarly define the mean $E[\boldsymbol{V}]$ of a random matrix $\boldsymbol{V}$ by taking the expectation of each entry in the matrix.
- The **covariance matrix** of $\boldsymbol{U} = [U_1, \ldots, U_m]'$ is the $m \times m$ matrix:

$$\boldsymbol{\Sigma}_{\boldsymbol{U}} = \mathrm{Cov}(\boldsymbol{U}) = E[(\boldsymbol{U} - \boldsymbol{\mu}_{\boldsymbol{U}})(\boldsymbol{U} - \boldsymbol{\mu}_{\boldsymbol{U}})'] = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{m1} & \sigma_{m2} & \cdots & \sigma_m^2 \end{pmatrix}$$

where $\sigma_{ij} = \mathrm{Cov}(U_i, U_j)$ and $\sigma_i^2 = \mathrm{Var}(U_i) = \mathrm{Cov}(U_i, U_i)$.

Properties Of The Mean and Covariance Operators

The following formulas will be needed in the future:

1. For $\boldsymbol{U} = [U_1, \ldots, U_m]' \in \mathbb{R}^{m \times 1}$, a constant matrix $\mathbf{A} \in \mathbb{R}^{\ell \times m}$, and a constant vector $\mathbf{d} \in \mathbb{R}^{\ell \times 1}$:

$$\boxed{\text{(i) } E(\mathbf{A}\boldsymbol{U} + \mathbf{d}) = \mathbf{A}E(\boldsymbol{U}) + \mathbf{d}} \qquad \boxed{\text{(ii) } \mathrm{Cov}(\mathbf{A}\boldsymbol{U} + \mathbf{d}) = \mathbf{A}\mathrm{Cov}(\boldsymbol{U})\mathbf{A}'}$$

2. Note that if $m = 1$ (i.e., $\boldsymbol{U} = U_1$),

$$\text{(iii) } \mathrm{Cov}(\boldsymbol{U}) = \mathrm{Var}(U_1).$$

3. In particular, if $\mathbf{A} = \mathbf{c} = [c_1, \ldots, c_m] \in \mathbb{R}^{1 \times m}$ (just a plain row vector), then

$$\boxed{\mathbf{c}\boldsymbol{U} = \textstyle\sum_{i=1}^{m} c_i U_i \in \mathbb{R},}$$

$$\boxed{\text{(iv) } \mathrm{Var}(\mathbf{c}\boldsymbol{U}) = \mathrm{Cov}(\mathbf{c}\boldsymbol{U}) = \mathbf{c}\mathrm{Cov}(\boldsymbol{U})\mathbf{c}'.}$$

Example 2

Suppose that

$$\mathrm{Cov}\left(\begin{bmatrix} U_1 \\ U_2 \end{bmatrix}\right) = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix}.$$

Then, the formula above says that

$$\begin{aligned}
\mathrm{Cov}\left(\begin{bmatrix} U_1 - U_2 \\ 2U_1 + U_2 \end{bmatrix}\right) &= \mathrm{Cov}\left(\begin{bmatrix} 1 & -1 \\ 2 & 1 \end{bmatrix}\begin{bmatrix} U_1 \\ U_2 \end{bmatrix}\right) \\
&= \begin{bmatrix} 1 & -1 \\ 2 & 1 \end{bmatrix}\begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix}\begin{bmatrix} 1 & 2 \\ -1 & 1 \end{bmatrix} \\
&= \begin{bmatrix} 1 & -1 \\ 2 & 1 \end{bmatrix}\begin{bmatrix} 2 & 1 \\ -3 & 0 \end{bmatrix} = \begin{bmatrix} 5 & 1 \\ 1 & 2 \end{bmatrix}
\end{aligned}$$

Application to the Linear Regression Model

Recall that for the linear regression model, we have

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon} = [\varepsilon_1, \ldots, \varepsilon_m]'$ is assumed to have the following properties:

(i) $\mathbb{E}(\varepsilon_i) = 0$, (ii) $\mathrm{Var}(\varepsilon_i) = \sigma^2$, and (iii) $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$, for $i \neq j$.

In particular, from (i)-(iii) and the definitions in p. 19, we have:

$$E(\boldsymbol{\varepsilon}) = \mathbf{0} \quad \text{(the } n \times 1 \text{ vector of } 0's\text{)},$$

$$\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_{n\times n} \quad (\boldsymbol{I}_{n\times n} = n \times n \text{ identity matrix}).$$

Then, using the formulas (i)-(ii) of page 20, we deduce that

$$E(\boldsymbol{y}) = E(\boldsymbol{X}\beta + \boldsymbol{\varepsilon}) = \boldsymbol{X}\beta + E(\boldsymbol{\varepsilon}) = \boldsymbol{X}\beta + \mathbf{0} = \boldsymbol{X}\beta,$$

$$\mathrm{Cov}(\boldsymbol{y}) = \mathrm{Cov}(\boldsymbol{X}\beta + \boldsymbol{\varepsilon}) = \mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_{n\times n}.$$

Bias of $\hat{\beta}$

Suppose that $\mathbf{X}'\mathbf{X}$ is invertible so that $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Then, applying the formula (i) of page 20 and the formula $E(\boldsymbol{y})$ of the previous page:

$$\begin{aligned} E(\hat{\beta}) &= E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}) \\ &= ((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')E(\mathbf{y}) \\ &= ((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')(\mathbf{X}\beta) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta \\ &= \beta. \end{aligned}$$

Therefore, $\underline{\hat{\beta} \text{ is an unbiased estimator of } \beta}$.

Variance-Covariance of $\hat{\beta}$

Recall from page 19 that

$$\text{Cov}(\boldsymbol{y}) = \text{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_{n\times n}$$

Therefore, applying the formula (ii) of page 20,

$$\begin{aligned}
\text{Cov}(\hat{\beta}) &= \text{Cov}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}) \\
&= ((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\text{Cov}(\mathbf{y})((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')' \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2 \boldsymbol{I}_{n\times n})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.
\end{aligned}$$

One Predictor Case

Let us consider the case of $k = 1$:

- As shown on top of p. 6,

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} n & \sum_{i=1}^n x_{i1} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 \end{bmatrix}.$$

- We can then compute the covariance matrix $\mathrm{Cov}(\hat{\beta}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$:

$$\mathrm{Cov}(\hat{\beta}) = \frac{\sigma^2}{n\sum_{i=1}^n x_{i1}^2 - (\sum_{i=1}^n x_{i1})^2} \begin{bmatrix} \sum_{i=1}^n x_{i1}^2 & -\sum_{i=1}^n x_{i1} \\ -\sum_{i=1}^n x_{i1} & n \end{bmatrix}.$$

- Hence, we obtain the formulas (check):

$$\mathrm{Var}\left(\hat{\beta}_1\right) = \frac{\sigma^2}{S_{xx}}, \quad \mathrm{Var}\left(\hat{\beta}_0\right) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right), \quad \mathrm{Cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) = -\frac{\bar{x}\sigma^2}{S_{xx}}.$$

Gauss-Markov Theorem

Let us assume that $\mathbf{X}'\mathbf{X}$ is invertible so that

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

The Gauss-Markov Theorem states that the OLSE $\hat{\beta}$ is BLUE (the Best Linear Unbiased Estimator) when estimating linear combinations of the parameters:

- That is, suppose we want to estimate $\boldsymbol{\gamma} = \boldsymbol{c}'\boldsymbol{\beta} = \sum_{i=0}^{k} c_i\beta_i$ for some given vector $\boldsymbol{c} = [c_0, \ldots, c_k]' \in \mathbb{R}^{p\times 1}$.
- Then, $\hat{\boldsymbol{\gamma}} := \boldsymbol{c}'\hat{\boldsymbol{\beta}} = \sum_{i=0}^{k} c_i\hat{\beta}_i$ is unbiased for γ and

  $$\mathrm{Var}(\hat{\gamma}) \leq \mathrm{Var}(\tilde{\gamma}),$$

  for any other linear unbiased estimator $\tilde{\gamma}$ of γ; i.e., for any $\tilde{\gamma}$ which is of the form $\tilde{\gamma} = \mathbf{d}'\mathbf{y}$, for some vector $\mathbf{d} \in \mathbb{R}^{p\times 1}$, and which satisfies $E[\tilde{\gamma}] = \boldsymbol{\gamma}$.

The proof is given in Appendix C.

Example 2

Consider again the data of the Toy Example if page 9. Suppose that we have an estimate for σ of $\hat{\sigma} = .1$. Given the matrix below, give estimates for the variances of the LSE $\hat{\beta}_j$.

```
(X=matrix(c(1,1,1,1,1,1,5,5,7,6,5,7,19,21,20,19,18,17,1,0,1,0,1,1),6,4))
(y=matrix(c(120,145,175,110,95,180),6,1))
(C=solve(t(X)%*%X))
          [,1]        [,2]        [,3]        [,4]
[1,] 73.672131 -1.91803279 -3.13114754 -4.23770492
[2,] -1.918033  0.22950820  0.03278689 -0.06557377
[3,] -3.131148  0.03278689  0.14754098  0.20491803
[4,] -4.237705 -0.06557377  0.20491803  1.09016393
```

Example 2. Solution

Per the formula on page 24 and the diagonal entries in the previous page, we have:

$$\widehat{\mathrm{Var}(\hat{\beta}_0)} = .1^2 \times 73.6721 = .736721$$
$$\widehat{\mathrm{Var}(\hat{\beta}_1)} = .1^2 \times .2295 = .002295$$
$$\widehat{\mathrm{Var}(\hat{\beta}_2)} = .1^2 \times .1475 = .001475$$
$$\widehat{\mathrm{Var}(\hat{\beta}_3)} = .1^2 \times 1.0901 = .010901.$$

Appendix A: Proof of Theorem 1 p. 15

Proof: $\Longleftarrow$) This direction is proved in pages 5-6 above.

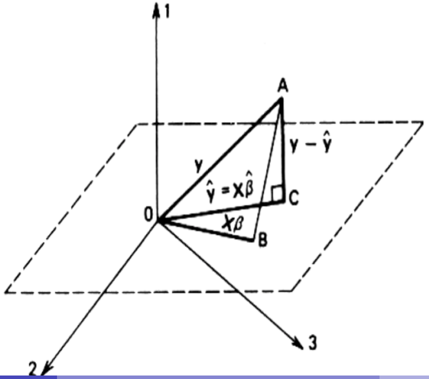
$\Longrightarrow$) Conversely, let $\hat{\beta}$ be a solution of the normal equations; i.e., $\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{y}$. First, note that $S(\tilde{\beta}) = \|\mathbf{y} - \mathbf{X}\tilde{\beta}\|^2 = (\mathbf{y} - \mathbf{X}\tilde{\beta})'(\mathbf{y} - \mathbf{X}\tilde{\beta})$. Then,

$$
\begin{aligned}
S(\tilde{\beta}) &= (\mathbf{y} - \mathbf{X}\tilde{\beta})'(\mathbf{y} - \mathbf{X}\tilde{\beta}) \\
&= (\mathbf{y} - \mathbf{X}\hat{\beta} + \mathbf{X}\hat{\beta} - \mathbf{X}\tilde{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta} + \mathbf{X}\hat{\beta} - \mathbf{X}\tilde{\beta}) \\
&= S(\hat{\beta}) + \|\mathbf{X}(\hat{\beta} - \tilde{\beta})\|^2,
\end{aligned}
$$

where the last equality follows since $(\mathbf{X}\hat{\beta} - \mathbf{X}\tilde{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta}) = (\hat{\beta} - \tilde{\beta})'(\mathbf{X}'\mathbf{y} - \mathbf{X}\mathbf{X}'\hat{\beta}) = \mathbf{0}$. The previous identity shows that $S(\hat{\beta}) \leq S(\tilde{\beta})$ for any other $\tilde{\beta}$.

Appendix B: Geometric Interpretation of the OLSE

The collection $\mathcal{C}(\mathbf{X})$ of all vectors $\{\mathbf{X}\beta \in \mathbb{R}^n : \beta \in \mathbb{R}^p\}$ form a linear space in $\mathbb{R}^n$, called the column space of the matrix $\mathbf{X}$ (think of a plane in $\mathbb{R}^3$). $\|\mathbf{y} - \mathbf{X}\beta\|$ is the distance between $\mathbf{X}\beta$ and $\mathbf{y}$. So, the vector $\mathbf{X}\hat{\beta}$ is the closest vector in $\mathcal{C}(\mathbf{X})$ to $\mathbf{y}$: The orthogonal projection of $\mathbf{y}$ onto $\mathcal{C}(\mathbf{X})$. Then, $\hat{\beta}$ always exists.



Appendix C: Sketch Of The Proof I

- Let us first compute $\mathrm{Var}(\hat{\gamma})$:

$$\mathrm{Var}(\hat{\gamma}) = \mathrm{Var}(\boldsymbol{c}'\hat{\beta}) = \boldsymbol{c}'\mathrm{Var}(\hat{\beta})\boldsymbol{c} = \sigma^2\boldsymbol{c}'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{c}$$

- Recall that we are assuming that $\tilde{\gamma} = \mathbf{d}'\mathbf{y}$, for some vector $\mathbf{d} \in \mathbb{R}^{p\times 1}$.
- Let us now find conditions on $\mathbf{d}'$ for $\tilde{\gamma}$ to be unbiased for $\gamma = \boldsymbol{c}'\beta$:

$$E[\tilde{\gamma}] = E[\mathbf{d}'\mathbf{y}] = \mathbf{d}'E[\mathbf{y}] = \mathbf{d}'\mathbf{X}\beta,$$

since recall that $\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon$, $E[\varepsilon] = \mathbf{0}$, and thus, $E[\mathbf{y}] = \mathbf{X}\beta$.

- Therefore, $E[\tilde{\gamma}] = \boldsymbol{c}'\beta$, for all β, if and only if

$$\mathbf{d}'\mathbf{X} = \mathbf{c}'$$

Appendix C: Sketch Of The Proof II

- We are now ready to show the result. Note that:

$$\mathrm{Var}(\tilde{\gamma}) = \mathrm{Var}(\mathbf{d}'\mathbf{y}) = \mathbf{d}'\mathrm{Cov}(\mathbf{y})\mathbf{d}' = \sigma^2\mathbf{d}'\mathbf{d}.$$

- Next, we write $\mathrm{Var}(\tilde{\gamma})$ as

$$\mathrm{Var}(\tilde{\gamma}) = \sigma^2(\mathbf{d} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c} + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c})'(\mathbf{d} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c} + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c})$$

- Next, introduce $\boldsymbol{q} = \mathbf{d} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}$ and note that

$$\begin{aligned}\mathrm{Var}(\tilde{\gamma}) &= \sigma^2(\boldsymbol{q} + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c})'(\boldsymbol{q} + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}) \\ &= \sigma^2\boldsymbol{q}'\boldsymbol{q} + \sigma^2\boldsymbol{c}'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{c},\end{aligned}$$

  because (as seen below) $\boldsymbol{q}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c} = 0$ and $\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{q} = 0$.

- We conclude the result since $\boldsymbol{q}'\boldsymbol{q} \geq 0$ and $\mathrm{Var}(\hat{\gamma}) = \sigma^2\boldsymbol{c}'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{c}$

Appendix C: Sketch Of The Proof III

- It remains to show that $\boldsymbol{q}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c} = 0$ and $\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{q} = 0$. Indeed,

$$\begin{aligned}
\boldsymbol{q}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c} &= (\mathbf{d} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c})'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c} \\
&= (\mathbf{d}' - \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c} \\
&= (\mathbf{d}'\mathbf{X} - \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c} \\
&= (\mathbf{d}'\mathbf{X} - \mathbf{c}')(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c} = 0
\end{aligned}$$

  because of the unbiasedness condition.

- We also deduce that $\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{q} = (\boldsymbol{q}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c})' = 0$.