Regression Analysis
Chapter 3: Multiple Regression

Shaobo Jin

Department of Mathematics

Simple Linear Regression: Matrix Notation

The simple linear regression model is

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, 2, ...n.$$

We can express it using matrix operations:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}.$$

Or simply

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}.$$

We often call the matrix $\boldsymbol{X}$ a design matrix.

OLS with Matrix Notation

The ordinary sum-of-squares becomes an Euclidean inner product:

$$\begin{aligned}\mathrm{RSS}\left(\beta_0,\beta_1\right) &= \sum_{i=1}^{n}\left[y_i-\left(\beta_0+\beta_1 x_i\right)\right]^2 \\ &= \left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right)^T\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right).\end{aligned}$$

Hence, we can say that the OLS estimator of $\boldsymbol{\beta}$ minimizes the quadratic form

$$\begin{aligned}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right)^T\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right) &= \boldsymbol{y}^T\boldsymbol{y}-2\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \\ &= \text{Constant}-\text{Linear}+\text{quadratic}.\end{aligned}$$

Gradient of Linear Form

Consider the vector $\boldsymbol{a} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$ and $\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$. Consider

$$\begin{aligned} f(\boldsymbol{x}) &= \boldsymbol{a}^T \boldsymbol{x} \\ &= a_1 x_1 + a_2 x_2. \end{aligned}$$

Its gradient is

$$\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} = \begin{bmatrix} \partial f(\boldsymbol{x}) / \partial x_1 \\ \partial f(\boldsymbol{x}) / \partial x_2 \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \boldsymbol{a}.$$

Gradient of Quadratic Form

Consider the matrix $\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$ and $\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$. Consider

$$\begin{aligned} f(\boldsymbol{x}) &= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= a_{11}x_1^2 + a_{12}x_1x_2 + a_{21}x_2x_1 + a_{22}x_2^2. \end{aligned}$$

The gradient is

$$\begin{aligned} \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} = \begin{bmatrix} \partial f(\boldsymbol{x}) / \partial x_1 \\ \partial f(\boldsymbol{x}) / \partial x_2 \end{bmatrix} &= \begin{bmatrix} 2a_{11}x_1 + a_{12}x_2 + a_{21}x_2 \\ a_{12}x_1 + a_{21}x_1 + 2a_{22}x_2 \end{bmatrix} \\ &= \left(\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} + \begin{bmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{bmatrix} \right) \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= \left(\boldsymbol{A} + \boldsymbol{A}^T \right) \boldsymbol{x}. \end{aligned}$$

OLS Estimator

Using above results,

$$\begin{aligned} \frac{\partial \boldsymbol{y}^T \boldsymbol{X} \boldsymbol{\beta}}{\partial \boldsymbol{\beta}} &= \boldsymbol{X}^T \boldsymbol{y}, \\ \frac{\partial \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\beta}}{\partial \boldsymbol{\beta}} &= 2\boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\beta}. \end{aligned}$$

Hence,

$$\frac{\partial \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)^T \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}^T \boldsymbol{y} + 2\boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\beta},$$

leading to the stationary point

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{y}.$$

OLS Estimator

In simple linear regression,

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}.$$

It can be shown that

$$\begin{aligned} \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} &= \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}^{-1} \\ &= \frac{1}{\sum_{i=1}^n \left(x_i - \bar{x}\right)^2} \begin{bmatrix} n^{-1}\sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}, \\ \boldsymbol{X}^T\boldsymbol{y} &= \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix}. \end{aligned}$$

Multiple Linear Regression

The simple linear regression model is

$$\mathrm{E}\left(Y \mid X = x\right) = \beta_0 + \beta_1 x.$$

The multiple linear regression model is

$$\begin{aligned}\mathrm{E}\left(Y \mid \boldsymbol{X} = \boldsymbol{x}\right) &= \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \\ &= \boldsymbol{x}^T \boldsymbol{\beta},\end{aligned}$$

where $\boldsymbol{\beta}$ is a column vector.

When we have observed a data set, the matrix notation is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}.$$

Notation

Consider the multiple linear regression model

$$\begin{aligned} \mathrm{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}) &= \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \\ &= \boldsymbol{x}^T \boldsymbol{\beta}. \end{aligned}$$

- The textbook treats $\boldsymbol{\beta}$ as a $(p+1) \times 1$ column vector.
- Even though the intercept is often included in the model, it can be excluded. Hence, we will treat

  $$\mathrm{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{\beta}$$

  with a $p \times 1$ vector $\boldsymbol{\beta}$.

A consequence is that

- if the intercept is included in the model, then we have $p - 1$ covariates.
- if the intercept is not included in the model, then we have p covariates.

Least Squares

In simple linear regression, we minimize the ordinary sum-of-squares

$$\begin{aligned} \mathrm{RSS}\left(\beta_0, \beta_1\right) &= \sum_{i=1}^{n}\left[y_i - \left(\beta_0 + \beta_1 x_i\right)\right]^2 \\ &= \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)^T\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right) \\ &= \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}. \end{aligned}$$

to estimate the regression coefficients β_0 and β_1.

The ordinary least squares (OLS) method for multiple linear regression minimizes

$$\mathrm{RSS}\left(\boldsymbol{\beta}\right) = \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)^T\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right).$$

Gradient of Linear and Quadratic Forms

Consider column vectors $\boldsymbol{a}$ and $\boldsymbol{x}$. The gradient of $f(\boldsymbol{x}) = \boldsymbol{a}^T\boldsymbol{x}$ is

$$\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} = \boldsymbol{a}.$$

Consider a square matrix $\boldsymbol{A}$ and a column vector $\boldsymbol{x}$. The gradient of $f(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}$ is

$$\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} = \left(\boldsymbol{A} + \boldsymbol{A}^T\right)\boldsymbol{x}.$$

OLS Estimator

Using above results,

$$\begin{aligned}\frac{\partial \boldsymbol{y}^T \boldsymbol{X}\boldsymbol{\beta}}{\partial \boldsymbol{\beta}} &= \boldsymbol{X}^T \boldsymbol{y},\\ \frac{\partial \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{X}\boldsymbol{\beta}}{\partial \boldsymbol{\beta}} &= 2\boldsymbol{X}^T \boldsymbol{X}\boldsymbol{\beta}.\end{aligned}$$

Hence,

$$\frac{\partial \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)^T \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}^T \boldsymbol{y} + 2\boldsymbol{X}^T \boldsymbol{X}\boldsymbol{\beta},$$

leading to the stationary point

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{y}.$$

Property of OLS Estimator

1. Under the assumption that $\mathrm{E}(Y \mid \boldsymbol{X}=\boldsymbol{x})=\boldsymbol{x}^T\boldsymbol{\beta}$ is correctly specified, the OLS estimator is unbiased as

$$\mathrm{E}\left(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) = \boldsymbol{\beta}.$$

2. The covariance matrix of the OLS estimator is

$$\mathrm{Var}\left(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\mathrm{Var}\left(\boldsymbol{y} \mid \boldsymbol{X}\right)\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}.$$

   - If we further assume (1) data are independent conditional on $\boldsymbol{x}$ and (2) $\mathrm{Var}\left(\boldsymbol{y} \mid \boldsymbol{X}\right)=\sigma^2$, we have

$$\mathrm{Var}\left(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) = \sigma^2\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}.$$

Prediction/Fitted Value

Once $\boldsymbol{\beta}$ is estimated, the fitted/estimated regression line is

$$\hat{\mathrm{E}}\left(Y \mid \boldsymbol{x} = \boldsymbol{x}\right) = \boldsymbol{x}^T \hat{\boldsymbol{\beta}}.$$

- The fitted value of $\hat{y}_i$ is

$$\hat{y}_i = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}.$$

  In matrix notation,

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

- For a new $\boldsymbol{x}_0$, the predicted y is

$$\hat{y} = \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}}.$$

Residual

The residual is

$$\hat{e}_i = y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}.$$

In matrix notation,

$$\begin{aligned} \hat{\boldsymbol{e}} &= \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} \\ &= \left[\boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\right]\boldsymbol{y}, \end{aligned}$$

where $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$ is called the hat matrix. The hat matrix is symmetric and idempotent!

The residual sum-of-squares is

$$\hat{\boldsymbol{e}}^T\hat{\boldsymbol{e}} = \boldsymbol{y}^T\left(\boldsymbol{I} - \boldsymbol{H}\right)\boldsymbol{y}.$$

That is,

$$\mathrm{RSS}\left(\hat{\boldsymbol{\beta}}\right) = \hat{\boldsymbol{e}}^T\hat{\boldsymbol{e}} = \boldsymbol{y}^T\left(\boldsymbol{I} - \boldsymbol{H}\right)\boldsymbol{y}.$$

Properties of Residuals

1. We always have $\sum_i \hat{e}_i = 0$ in models where the intercept is included.
2. Sample correlation between residual and regressors is always zero, if the intercept is included in the model.
3. Under the assumption that $\mathrm{E}\left(Y \mid \boldsymbol{X} = \boldsymbol{x}\right) = \boldsymbol{x}^T \boldsymbol{\beta}$ is correctly specified,

$$\mathrm{E}\left(\hat{\boldsymbol{e}} \mid \boldsymbol{X}\right) = \mathbf{0}.$$

4. Under the assumption that $\mathrm{Var}\left(\boldsymbol{y} \mid \boldsymbol{X}\right) = \sigma^2 \boldsymbol{I}$, conditional on $\boldsymbol{X}$,

$$\begin{aligned} \mathrm{Cov}\left(\hat{\boldsymbol{e}}, \hat{\boldsymbol{\beta}}\right) &= \mathbf{0}. \\ \mathrm{Cov}\left(\hat{\boldsymbol{e}}, \hat{\boldsymbol{y}}\right) &= \mathbf{0}. \end{aligned}$$

Illustration (2D)

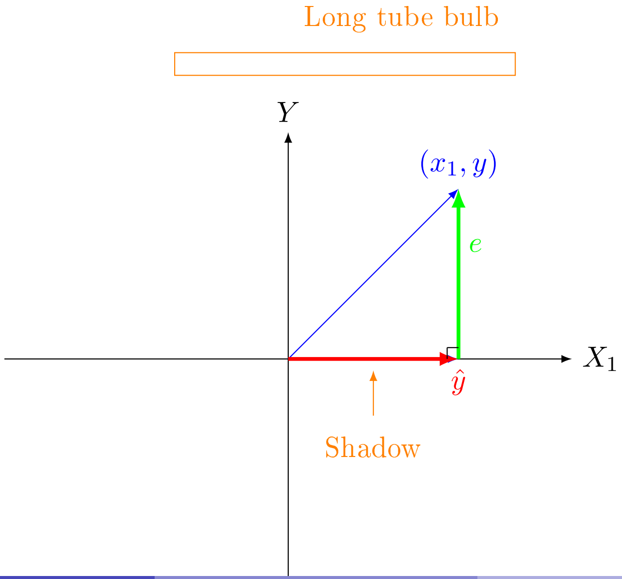
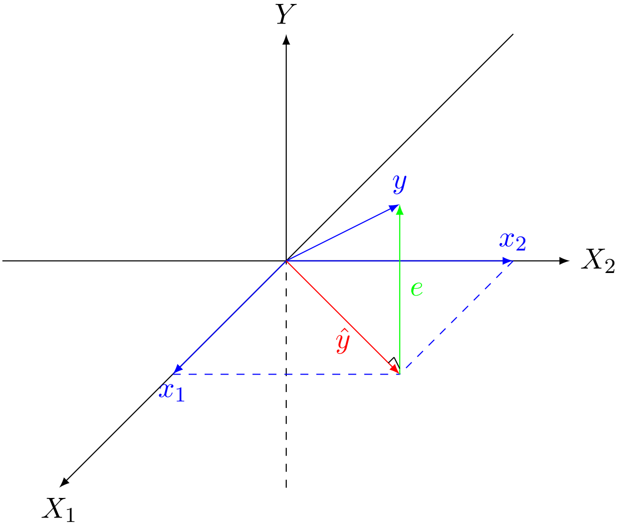


Illustration (3D)



Gauss-Markov Theorem

Theorem (Gauss-Markov Theorem)

Suppose that $E(\boldsymbol{y} \mid \boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}$ *and* $Var(\boldsymbol{y} \mid \boldsymbol{X}) = \sigma^2 \boldsymbol{I}$. *Then the best linear unbiased estimator (BLUE) of* $\boldsymbol{\beta}$ *is*

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{y}.$$

That is, for any linear unbiased estimator $\tilde{\boldsymbol{\beta}}$ *of* $\boldsymbol{\beta}$, $Var\left(\tilde{\boldsymbol{\beta}}\right) - Var\left(\hat{\boldsymbol{\beta}}\right) \geq 0$ *(positive semi-definite).*

Equivalently, let $\boldsymbol{a}^T \boldsymbol{y}$ *be any linear unbiased estimator of* $\boldsymbol{a}^T \boldsymbol{\beta}$ *for fixed vector* $\boldsymbol{a}$, *then* $Var\left(\boldsymbol{a}^T \boldsymbol{y}\right) - Var\left(\boldsymbol{a}^T \hat{\boldsymbol{\beta}}\right) \geq 0$.

Estimating σ^2

The estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{e}}^T \hat{\boldsymbol{e}}}{n-p},$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector. We can show that

$$E\left(\hat{\sigma}^2\right) = \sigma^2,$$

under the assumptions that

1. the model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ is correctly specified,
2. $\mathrm{E}\left(\boldsymbol{e} \mid \boldsymbol{X}\right) = \mathbf{0}$,
3. $\mathrm{Var}\left(\boldsymbol{e} \mid \boldsymbol{X}\right) = \sigma^2 \boldsymbol{I}$.

R^2: Coefficient of Determination

Definition (R^2)

The R^2, defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \in [0,1],$$

is to measure how much variation in Y has been explained by our model.

We can rewrite R^2 as

$$R^2 = 1 - \frac{\boldsymbol{y}^T\left[\boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\right]\boldsymbol{y}}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}.$$

The positive square root of R^2 is called the multiple correlation coefficient.

A Pitfall of R^2

Suppose that we have fitted a model with $\boldsymbol{x}_1$ as $\boldsymbol{x}_1^T \boldsymbol{\beta}_1$. The OLS estimator minimizes

$$\left(\boldsymbol{y} - \boldsymbol{X}_1 \boldsymbol{\beta}_1\right)^T \left(\boldsymbol{y} - \boldsymbol{X}_1 \boldsymbol{\beta}_1\right),$$

and

$$\text{RSS}\left(\hat{\boldsymbol{\beta}}_1\right) = \boldsymbol{y}^T \left[\boldsymbol{I} - \boldsymbol{X}_1 \left(\boldsymbol{X}_1^T \boldsymbol{X}_1\right)^{-1} \boldsymbol{X}_1^T\right] \boldsymbol{y}.$$

Now we want to add $\boldsymbol{x}_2$ into the model and consider $\boldsymbol{x}_1^T \boldsymbol{\beta}_1 + \boldsymbol{x}_2^T \boldsymbol{\beta}_2$. The OLS estimator minimizes

$$\left(\boldsymbol{y} - \boldsymbol{X}_1 \boldsymbol{\beta}_1 - \boldsymbol{X}_2 \boldsymbol{\beta}_2\right)^T \left(\boldsymbol{y} - \boldsymbol{X}_1 \boldsymbol{\beta}_1 - \boldsymbol{X}_2 \boldsymbol{\beta}_2\right),$$

and

$$\text{RSS}\left(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2\right) = \boldsymbol{y}^T \left[\boldsymbol{I} - \boldsymbol{X} \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T\right] \boldsymbol{y}.$$

More Regressors, Larger R^2

We should have

$$\text{RSS}\left(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2\right) \;\leq\; \text{RSS}\left(\hat{\boldsymbol{\beta}}_1\right).$$

Hence,

$$1 - \frac{\text{RSS}\left(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2\right)}{\sum_{i=1}^n \left(y_i - \bar{y}\right)^2} \;\geq\; 1 - \frac{\text{RSS}\left(\hat{\boldsymbol{\beta}}_1\right)}{\sum_{i=1}^n \left(y_i - \bar{y}\right)^2}.$$

When we add more regressors to the model, the R^2 will never decrease!

Adjusted R^2

The adjusted R^2 is

$$R^2_{\text{adjusted}} = 1 - \frac{n-1}{n-p-1}\left(1 - R^2\right).$$

When p increases, $1 - R^2$ decreases and $n - p - 1$ decreases. Hence, it attempts to adjusted for the number of covariates in the model.

Normally Distributed e

It is also common to assume that the error is normally distributed as

$$\boldsymbol{e} \mid \boldsymbol{X} \quad \sim \quad N\left(0, \sigma^2 \boldsymbol{I}\right).$$

1. Under the independence assumption, the log-likelihood function of $\boldsymbol{\beta}$ and σ^2 is

$$\ell\left(\boldsymbol{\beta}, \sigma^2\right) \quad = \quad \sum_{i=1}^{n} \left\{ -\frac{1}{2} \log\left(2\pi\right) - \frac{1}{2} \log\left(\sigma^2\right) - \frac{1}{2\sigma^2} \left(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}\right)^2 \right\}.$$

2. The maximum likelihood estimator (MLE) is given by

$$\hat{\boldsymbol{\beta}} \quad = \quad \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{y},$$

the same as their OLS estimator! Hence, they are still unbiased.

Distribution of $\hat{\boldsymbol{\beta}}$

Under the normality assumption, we can obtain

$$\begin{aligned} \hat{\boldsymbol{\beta}} &\sim N_p\left(\boldsymbol{\beta}, \quad \sigma^2 \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}\right), \\ \text{and } \hat{\beta}_j &\sim N\left(\beta_j, \quad \sigma^2 \left[\left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}\right]_{jj}\right). \end{aligned}$$

The standard error of $\hat{\beta}_j$ is

$$\hat{\sigma}\sqrt{\left[\left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}\right]_{jj}}.$$

Student t-Distribution

It can be shown that, conditional on $\boldsymbol{X}$,

$$\frac{\hat{\boldsymbol{e}}^T \hat{\boldsymbol{e}}}{\sigma^2} \sim \chi^2 (n-p) .$$

Then,

$$\frac{\left(\hat{\beta}_j - \beta_j\right) / \sqrt{\left[\left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}\right]_{jj}}}{\sqrt{\hat{\boldsymbol{e}}^T \hat{\boldsymbol{e}} / (n-p)}} \sim t(n-p) .$$

We can use it to test H_0: $\beta_j = 0$.
A $1-\alpha$ confidence interval for β_j is

$$\hat{\beta}_j \pm t_{1-\alpha/2} (n-p) \sqrt{\hat{\sigma}^2 \left[\left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}\right]_{jj}}.$$

Prediction of Regression Function

Suppose that a new subject has the covariate value $\boldsymbol{x}_0$ and we want to predict the mean response $\mathrm{E}\left(Y \mid \boldsymbol{X} = \boldsymbol{x}\right)$.

- The predicted mean response is $\hat{\mathrm{E}}\left(Y \mid \boldsymbol{X} = \boldsymbol{x}\right) = \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}}$.
- Under the normality assumption,

$$\boldsymbol{x}_0^T \hat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{x}_0^T \boldsymbol{\beta},\ \sigma^2 \boldsymbol{x}_0^T \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{x}_0\right).$$

Confidence Interval For Regression Function

Hence,

$$\frac{\frac{\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}}-\boldsymbol{x}_0^T\boldsymbol{\beta}}{\sqrt{\sigma^2\boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0}}}{\sqrt{\frac{\hat{\boldsymbol{e}}^T\hat{\boldsymbol{e}}}{\sigma^2}/\left(n-p\right)}} = \frac{\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}}-\boldsymbol{x}_0^T\boldsymbol{\beta}}{\sqrt{\hat{\sigma}^2\left[\boldsymbol{x}_0^T\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{x}_0\right]}} \sim t\left(n-p\right).$$

A $1-\alpha$ confidence interval for $\boldsymbol{x}_0^T\boldsymbol{\beta}$ is

$$\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}} \pm t_{1-\alpha/2}\left(n-p\right)\sqrt{\hat{\sigma}^2\left[\boldsymbol{x}_0^T\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{x}_0\right]}.$$

Forecast New Response

Now we want to forecast the new response Y_0 using $\boldsymbol{x}_0$.

- Under the independence and normality assumption, given $\boldsymbol{X}$ and $\boldsymbol{x}_0$,

$$\begin{bmatrix} Y_0 \\ \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{x}_0^T \boldsymbol{\beta} \\ \boldsymbol{x}_0^T \boldsymbol{\beta} \end{bmatrix}, \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \boldsymbol{x}_0^T \left(\boldsymbol{X}^T \boldsymbol{X} \right)^{-1} \boldsymbol{x}_0 \end{bmatrix} \right).$$

- Hence,

$$Y_0 - \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}} = \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} Y_0 \\ \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}} \end{bmatrix} \sim N\left(0,\ \sigma^2 \left(1 + \boldsymbol{x}_0^T \left(\boldsymbol{X}^T \boldsymbol{X} \right)^{-1} \boldsymbol{x}_0 \right) \right).$$

- A $1-\alpha$ prediction interval for Y_0 is

$$\boldsymbol{x}_0^T \hat{\boldsymbol{\beta}} \pm t_{\alpha/2}(n-p) \sqrt{\hat{\sigma}^2 \left[1 + \boldsymbol{x}_0^T \left(\boldsymbol{X}^T \boldsymbol{X} \right)^{-1} \boldsymbol{x}_0 \right]},$$

always wider than the confidence interval.