

A Notebook of Linear Model

Yuwen Long

2024

1 Basic Linear Model

1.1 Definition and Assumption

Consider the linear model of the form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad (1.1)$$

where $\mathbf{y} \in \mathbb{R}^{n \times 1}$ is the vector of responses and n is the number of observed samples, $\mathbf{X} \in \mathbb{R}^{n \times p'}$ is the design matrix of p explanatory variables of n observations and if the intercept is taken into consideration then $p' = p + 1$, $\boldsymbol{\beta} \in \mathbb{R}^{p' \times 1}$ is the coefficients vector and $\boldsymbol{\epsilon} \in \mathbb{R}^{n \times 1}$ is the random error vector. There are 2 common assumptions of this model.

Assumption 1.1. (*Gauss-Markov Condition*)

$$\begin{aligned} \mathbb{E}(\boldsymbol{\epsilon}) &= \mathbf{0}, \\ \text{Cov}(\boldsymbol{\epsilon}) &= \sigma^2 \mathbf{I}, \\ \text{Cov}(\mathbf{X}, \boldsymbol{\epsilon}) &= \mathbf{0}. \end{aligned} \quad (1.2)$$

Assumption 1.2. (*Normality Condition*)

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}), \quad (1.3)$$

combined with (1.1), we have

$$\mathbf{y}|\mathbf{X} \sim \mathcal{N}(\mathbf{y}|\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}) \quad (1.4)$$

1.2 Estimation

1.2.1 OLS

We can use the *Ordinary Least Square*(OLS) method to estimate $\boldsymbol{\beta}$. First, we define the *loss function* as

$$\text{Loss} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \quad (1.5)$$

Before use the OLS, we also need to assume the design matrix $n > p'$ and \mathbf{X} having rank equal to p' , i.e. \mathbf{X} has full rank.

Theorem 1.2.1. The OLS estimation of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}. \quad (1.6)$$

Lemma 1.3. If \mathbf{X} is a $m \times n$ matrix with rank equal to n , then $\mathbf{X}^\top \mathbf{X}$ is invertible.

Proof. It suffices to show that $(\mathbf{X}^\top \mathbf{X}) \mathbf{v} = \mathbf{0}$ if and only if $\mathbf{v} = \mathbf{0}$. If $\mathbf{X}\mathbf{v} = \mathbf{0}$, then $\mathbf{X}^\top \mathbf{X}\mathbf{v} = \mathbf{0}$, thus $\text{nullity}(\mathbf{X}) \subset \text{nullity}(\mathbf{X}^\top \mathbf{X})$. For the other direction, if $\mathbf{X}^\top \mathbf{X}\mathbf{v} = \mathbf{0}$, then $\mathbf{v}^\top \mathbf{X}^\top \mathbf{X}\mathbf{v} = (\mathbf{X}\mathbf{v})^\top (\mathbf{X}\mathbf{v}) = \mathbf{0}$, if and only if $\mathbf{X}\mathbf{v} = \mathbf{0}$, thus $\text{nullity}(\mathbf{X}^\top \mathbf{X}) \subset \text{nullity}(\mathbf{X})$. In conclusion, we have $\text{nullity}(\mathbf{X}^\top \mathbf{X}) = \text{nullity}(\mathbf{X})$. And notice that $\text{nullity}(\mathbf{X}) = 0$, thus $\text{nullity}(\mathbf{X}^\top \mathbf{X}) = 0$, which implies that $\mathbf{X}^\top \mathbf{X}$ is invertible. \square

Proof. Theorem 1.1

Using matrix calculus, we have

$$\frac{\partial \text{Loss}}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}. \quad (1.7)$$

Let $\frac{\partial \text{Loss}}{\partial \boldsymbol{\beta}} = 0$, then we have

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}. \quad (1.8)$$

\square

Remark 1.1. We can also use some linear algebraic techniques to prove this theorem. Minimizing the loss function is equal to find a $\boldsymbol{\beta}$ which minimizes the distance between \mathbf{y} and $\mathbf{X}\boldsymbol{\beta}$, i.e. minimize $\|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2$. Recall that only the projection of \mathbf{y} on the $\text{range}(\mathbf{X})$ can minimize this distance, so we only need to find the projection of \mathbf{y} , which implies that $\mathbf{X}\boldsymbol{\beta} - \mathbf{y}$ should be perpendicular to $\mathbf{X}\boldsymbol{\beta}$ which is in the range of \mathbf{X} . Recall that the *orthogonal complement* of $\text{range}(\mathbf{X})$ is $\text{nullity}(\mathbf{X}^\top)$, thus such $\mathbf{X}\boldsymbol{\beta} - \mathbf{y}$ must in $\text{nullity}(\mathbf{X}^\top)$, indicating $\mathbf{X}^\top (\mathbf{X}\boldsymbol{\beta} - \mathbf{y}) = \mathbf{0}$, leading to $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$.

1.3.1 MLE

We can also use the *Maximal Likelihood Estimation* to estimate $\boldsymbol{\beta}$ if the assumption of normality (1.2) holds. Under this assumption, the probability density function of \mathbf{y} conditioned on \mathbf{X} is given as (1.4), which takes the form as follows:

$$f(\mathbf{Y}|\mathbf{X}) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{\sigma^2}{2} (\mathbf{X}\boldsymbol{\beta} - \mathbf{y})^\top (\mathbf{X}\boldsymbol{\beta} - \mathbf{y}) \right\}. \quad (1.9)$$

Maximizing this likelihood function is equivalent to maximizing its log-likelihood function:

$$\log f(\mathbf{Y}|\mathbf{X}) = -\frac{1}{2} \log(2\pi\sigma^2) - \frac{\sigma^2}{2} (\mathbf{X}\boldsymbol{\beta} - \mathbf{y})^\top (\mathbf{X}\boldsymbol{\beta} - \mathbf{y}), \quad (1.10)$$

which is equivalent to the OLS.

1.4 Properties of the Estimators

We can get the estimated value of \mathbf{y} as

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}. \quad (1.11)$$

We call $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ *hat matrix* since it is a projection matrix which projects \mathbf{y} onto $\hat{\mathbf{y}}$.

Theorem 1.4.1. \mathbf{H} is an orthogonal projection.

Proof. $\mathbf{H}^2 = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top = \mathbf{H}$;
 $\mathbf{H}^\top = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top = \mathbf{H}$. \square

One can also verify that $(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H}) = \mathbf{I} - \mathbf{H}$.

Theorem 1.4.2. The OLS estimator $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$.

Proof. $\mathbb{E}(\hat{\beta}) = \mathbb{E}\left((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}\right) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbb{E}(\mathbf{y}) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X} \beta = \beta.$ \square

From Theorem 1.4.2, we have $\hat{\mathbf{y}}$ is an unbiased estimator of \mathbf{y} .

Corollary 1.4.2.1. $\hat{\mathbf{y}}$ is an unbiased estimator of \mathbf{y} and $\mathbb{E}(e_i) = 0$.

Theorem 1.4.3. $\text{Var}(\hat{\beta}) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}.$

Proof. $\text{Var}(\hat{\beta}) = \text{Var}\left((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}\right) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \text{Var}(\mathbf{y}) \left((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top\right)^\top = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}.$ \square

Lemma 1.5. $\text{tr}(\mathbf{H}) = p' = p + 1$

Proof. $\text{tr}(\mathbf{H}) = \text{tr}\left(\mathbf{y} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top\right) = \text{tr}\left((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}\right) = \text{tr}(\mathbf{I}) = p' = p + 1.$ \square

Further more, we estimate σ^2 as $\hat{\sigma}^2 = \frac{1}{n-p-1} (\mathbf{y} - \hat{\mathbf{y}})^\top (\mathbf{y} - \hat{\mathbf{y}}) = \frac{1}{n-p-1} \sum_{i=1}^n e_i^2$, where $e_i = y_i - \hat{y}_i$.

Theorem 1.5.1. $\hat{\sigma}^2 = \frac{1}{n-p-1} (\mathbf{y} - \hat{\mathbf{y}})^\top (\mathbf{y} - \hat{\mathbf{y}}) = \frac{1}{n-p-1} \sum_{i=1}^n e_i^2$ is an unbiased estimator of σ^2 .

Proof. $\mathbb{E}(\hat{\sigma}^2) = \frac{1}{n-p-1} \sum_{i=1}^n \mathbb{E}(e_i^2) = \frac{1}{n-p-1} \sum_{i=1}^n \left(\text{Var}(e_i) - \mathbb{E}(e_i)^2\right) = \frac{1}{n-p-1} \sum_{i=1}^n \text{Var}(e_i) = \frac{1}{n-p-1} \text{tr}(\text{Cov}(\mathbf{e})) = \frac{1}{n-p-1} \text{tr}(\text{Cov}((\mathbf{I} - \mathbf{H}) \mathbf{y})) = \frac{1}{n-p-1} \text{tr}((\mathbf{I} - \mathbf{H}) \sigma^2) = \frac{1}{n-p-1} \sigma^2 (\text{tr}(\mathbf{I}) - \text{tr}(\mathbf{H})) = \frac{1}{n-p-1} \sigma^2 (n - p - 1) = \sigma^2.$ \square

Theorem 1.5.2. The $\mathbf{c}^\top \hat{\beta}$ is the best estimator of $\mathbf{c}^\top \beta$, which has the minimal variance, among the unbiased estimators of linear combination of \mathbf{y} .

Proof. We have already known that $\hat{\beta}$ is an unbiased estimator. Let $\mathbf{a}^\top \mathbf{y}$ be another unbiased estimator, then $\mathbb{E}(\mathbf{a}^\top \mathbf{y}) = \mathbf{a}^\top \mathbf{X} \beta = \mathbf{c}^\top \beta$, which implies that $\mathbf{a}^\top \mathbf{X} = \mathbf{c}^\top$.

Then, $\text{Var}(\mathbf{c}^\top \hat{\beta}) - \text{Var}(\mathbf{a}^\top \mathbf{y}) = \sigma^2 \mathbf{c}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{c} - \sigma^2 \mathbf{a}^\top \mathbf{a} = \sigma^2 \mathbf{a}^\top \left(\mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top - \mathbf{I}\right) \mathbf{a} = \sigma^2 \mathbf{a}^\top (\mathbf{H} - \mathbf{I}) \mathbf{a}$, since $\mathbf{I} - \mathbf{H}$ is a positive semidefinite matrix, thus $\text{Var}(\mathbf{c}^\top \hat{\beta}) - \text{Var}(\mathbf{a}^\top \mathbf{y}) = \sigma^2 \mathbf{a}^\top (\mathbf{H} - \mathbf{I}) \mathbf{a} \leq 0.$ \square

Theorem 1.5.3. Under the assumption of normality, we have $\hat{\beta}$ is independent from $\hat{\sigma}^2$.

Proof. It suffice to show that $\hat{\beta}$ is independent from \mathbf{e} since $\hat{\sigma}^2 = \frac{1}{n-p-1} \mathbf{e}^\top \mathbf{e}$. What's more, we have $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1})$ and $\mathbf{e} \sim \mathcal{N}(0, \sigma^2 (\mathbf{I} - \mathbf{H}))$, thus it suffices to show that $\hat{\beta}$ and \mathbf{e} are uncorrelated.

$$\begin{aligned} \text{Cov}(\hat{\beta}, \mathbf{e}) &= \text{Cov}\left((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}, (\mathbf{I} - \mathbf{H}) \mathbf{y}\right) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \left(\mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top - \mathbf{I}\right)^\top \\ &= \sigma^2 \left((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top - (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top\right) = \mathbf{0}. \end{aligned}$$

\square

Corollary 1.5.3.1. Define $S_E = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \mathbf{e}^\top \mathbf{e}$ and $S_R = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = (\mathbf{X} \hat{\beta} - \bar{\mathbf{y}})^\top (\mathbf{X} \hat{\beta} - \bar{\mathbf{y}})$. Then, from the above theorem, we have S_E is independent from S_R .

Lemma 1.6. Assume $\mathbf{y} \sim \mathcal{N}(0, \mathbf{I}_n)$ and \mathbf{H} with $\text{tr}(\mathbf{H}) = h$ is a idempotent and positive semidefinite matrix, then we have $\mathbf{y}^\top \mathbf{H} \mathbf{y} \sim \chi^2(h)$.

Proof. Since \mathbf{H} is idempotent and positive semidefinite, we can diagonalize it as $\mathbf{H} = \mathbf{P}^\top \mathbf{\Sigma} \mathbf{P}$, where \mathbf{P} is a unitary matrix and $\mathbf{\Sigma}$ is a diagonal matrix with first h entries is 1 and else 0. Then, $\mathbf{y}^\top \mathbf{H} \mathbf{y} = \mathbf{y}^\top \mathbf{P}^\top \mathbf{\Sigma} \mathbf{P} \mathbf{y}$. Let $\mathbf{X} = \mathbf{y} \mathbf{P}$, then $\mathbf{X} \sim \mathcal{N}(0, \mathbf{P}^\top \mathbf{P}) = \mathcal{N}(0, \mathbf{I})$, thus $\mathbf{y}^\top \mathbf{H} \mathbf{y} = \mathbf{X}^\top \mathbf{\Sigma} \mathbf{X} = \sum_{i=1}^h x_i^2 \sim \chi^2(h).$ \square

Theorem 1.6.1. Under the assumption of normality, we have $S_E/\sigma^2 \sim \chi^2(n-p-1)$.

Proof. Without loss of generality, let $\sigma^2 = 1$.

$$S_E = \mathbf{e}^\top \mathbf{e} = \mathbf{y}^\top (\mathbf{I} - \mathbf{H}) \mathbf{y}.$$

Since $\mathbf{I} - \mathbf{H}$ is a idempotent and positive semidefinite matrix, it ends with Lemma (1.6). □

Theorem 1.6.2. Define $S_T = \sum_{i=1}^n (y_i - \bar{y})^2$, then $S_T = S_E + S_R$.

Remark 1.2. The term of S_T refers to the total deviation of $\{y_i\}$, S_E refers to the part that the fitted model cannot explain and the S_R represents the effectiveness of the fitted model. We also define $R^2 = \frac{S_R}{S_T} = 1 - \frac{S_E}{S_T} \in [0, 1]$ to imply the effectiveness of a fitted model. The closer R^2 to 1, the better the model.

Theorem 1.6.3. $S_T/\sigma^2 = \sum_{i=1}^n (y_i - \bar{y})^2 \sim \chi^2(n-1)$.

Proof. S_T is the sample standard deviation of $\{y_i\}$, thus we know from mathematical statistics that $S_T/\sigma^2 \sim \chi^2(n-1)$. □

Corollary 1.6.3.1. $S_R \sim \chi^2(p)$.

Property 1.1. $\mathbf{X}^\top \mathbf{e} = \mathbf{0}$.

Proof. $\mathbf{X}^\top \mathbf{e} = \mathbf{X}^\top (\mathbf{I} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top) \mathbf{y} = \mathbf{0}$. □

Corollary 1.6.3.2. $\sum_{i=1}^n e_i = 0$.

Proof. Take the result from property (1.1) and notice that the first row of \mathbf{X}^\top is $\mathbf{1}_{1 \times n}$. □

Remark 1.3. $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ is perpendicular to $\mathbf{X}\hat{\boldsymbol{\beta}}$, implying that it is belong to the orthogonal complement of $\text{range}(\mathbf{X})$ which is $\text{nullity}(\mathbf{X}^\top)$. Thus, \mathbf{e} is also perpendicular to the whole $\text{range}(\mathbf{X})$.

1.7 Generalized Least Square

Suppose that the random errors now have covariance $\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is a positive definite matrix. Then, we can decompose it with a unitary matrix \mathbf{P} and a diagonal matrix $\boldsymbol{\Lambda}$ with all entries positive as $\boldsymbol{\Sigma} = \mathbf{P}^\top \boldsymbol{\Lambda} \mathbf{P}$. Also, we have the principal square root of $\boldsymbol{\Sigma}$ as $\boldsymbol{\Sigma}^{-1/2} = \mathbf{P}^\top \boldsymbol{\Lambda}^{-1/2} \mathbf{P}$.

Let $\mathbf{v} = \boldsymbol{\Sigma}^{-1/2} \boldsymbol{\epsilon}$, $\mathbf{U} = \boldsymbol{\Sigma}^{-1/2} \mathbf{X}$, $\mathbf{z} = \boldsymbol{\Sigma}^{-1/2} \mathbf{y}$, then we have

$$\begin{aligned} \mathbf{z} &= \mathbf{U}\boldsymbol{\beta} + \mathbf{v}, \\ \mathbb{E}(\mathbf{v}) &= \mathbf{0}, \\ \text{Var}(\mathbf{v}) &= \sigma^2 \mathbf{I}. \end{aligned}$$

Then, the OLS of the transformed model is $\boldsymbol{\beta}^* = (\mathbf{U}^\top \mathbf{U})^{-1} \mathbf{U}^\top \mathbf{z} = (\mathbf{X}^\top \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\Sigma}^{-1} \mathbf{y}$, which is called *generalized least square estimation*(GLSE) or *Gauss-Markov estimation*.

Theorem 1.7.1. (1) $\mathbb{E}(\boldsymbol{\beta}^*) = \boldsymbol{\beta}$,

(2) $\text{Var}(\boldsymbol{\beta}^*) = \sigma^2 (\mathbf{X}^\top \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1}$,

(3) For all fixed constant vector \mathbf{c} , $\mathbf{c}^\top \boldsymbol{\beta}^*$ is the only linear by \mathbf{y} unbiased estimator with minimal variance of $\mathbf{c}^\top \boldsymbol{\beta}$.

Proof. (1) $\mathbb{E}(\beta^*) = \mathbb{E}\left((\mathbf{X}^\top \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \Sigma^{-1} \mathbf{y}\right) = (\mathbf{X}^\top \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \Sigma^{-1} \mathbf{X} = 0$.

(2)

$$\begin{aligned} \text{Var}(\beta^*) &= \text{Var}\left((\mathbf{X}^\top \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \Sigma^{-1} \mathbf{y}\right) \\ &= \sigma^2 (\mathbf{X}^\top \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \Sigma^{-1} \Sigma \Sigma^{-1} \mathbf{X} (\mathbf{X}^\top \Sigma^{-1} \mathbf{X})^{-1} = \sigma^2 (\mathbf{X}^\top \Sigma^{-1} \mathbf{X})^{-1}. \end{aligned}$$

(3) Suppose $\mathbf{a}^\top \mathbf{y}$ is any unbiased linear estimator of β , then $\mathbf{a}^\top \mathbf{y} = \mathbf{a}^\top \Sigma^{1/2} \Sigma^{-1/2} \mathbf{y} = \mathbf{a}^\top \Sigma^{1/2} \mathbf{z}$. And $\mathbf{c}^\top \beta^* = \mathbf{c}^\top (\mathbf{U}^\top \mathbf{U})^{-1} \mathbf{U}^\top \mathbf{z}$ is a OLS estimator of β and $\mathbf{a}^\top \Sigma^{1/2} \mathbf{z}$ is an unbiased linear estimator. Then, from Gauss-Markov theorem, we have $\text{Var}(\mathbf{c}^\top \beta^*) \leq \text{Var}(\mathbf{a}^\top \Sigma^{1/2} \mathbf{z})$, where the equality holds if and only if $\mathbf{a}^\top \mathbf{y} = \mathbf{c}^\top \beta^*$. \square

2 Linear Model with Linear Constraint

2.1 Estimation

Consider the cases where we have some constraints for the linear model as follow:

$$\begin{cases} \mathbf{y} = \mathbf{X}\beta + \epsilon \\ \mathbf{H}\beta = \mathbf{c}, \end{cases}$$

where \mathbf{H} is a $q \times (p+1)$ matrix and with $\text{rank}(\mathbf{H})=q$. We can use the Lagrange function to get estimation. Let the Lagrange function be

$$\mathcal{L}(\beta, \lambda) = (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta) + \lambda^\top (\mathbf{H}\beta - \mathbf{c}). \quad (2.1)$$

To solve this function, let

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial \beta} = 2\mathbf{X}^\top \mathbf{X}\beta - 2\mathbf{X}^\top \mathbf{y} + \mathbf{H}^\top \lambda = 0 \\ \frac{\partial \mathcal{L}}{\partial \lambda} = \mathbf{H}\beta - \mathbf{c} = 0, \end{cases} \quad (2.2)$$

Then, we have

$$\begin{cases} \beta = (\mathbf{X}^\top \mathbf{X})^{-1} \left(\mathbf{X}^\top \mathbf{y} - \frac{1}{2} \mathbf{H}^\top \lambda \right) \\ \mathbf{H}\beta = \mathbf{c} \end{cases} \quad (2.3)$$

$$\begin{cases} \beta = (\mathbf{X}^\top \mathbf{X})^{-1} \left(\mathbf{X}^\top \mathbf{y} - \frac{1}{2} \mathbf{H}^\top \lambda \right) \\ \mathbf{H} (\mathbf{X}^\top \mathbf{X})^{-1} \left(\mathbf{X}^\top \mathbf{y} - \frac{1}{2} \mathbf{H}^\top \lambda \right) = \mathbf{H} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} - \mathbf{H} (\mathbf{X}^\top \mathbf{X})^{-1} \frac{1}{2} \mathbf{H}^\top \lambda = \mathbf{c} \end{cases} \quad (2.4)$$

$$\begin{cases} \beta = (\mathbf{X}^\top \mathbf{X})^{-1} \left(\mathbf{X}^\top \mathbf{y} - \frac{1}{2} \mathbf{H}^\top \lambda \right) \\ \hat{\lambda} = 2 \left(\mathbf{H} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top \right)^{-1} \left(\mathbf{H} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} - \mathbf{c} \right) \end{cases} \quad (2.5)$$

$$\begin{cases} \beta = \hat{\beta} - (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top \left(\mathbf{H} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top \right)^{-1} \left(\mathbf{H} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} - \mathbf{c} \right) \\ = \hat{\beta} - (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top \left(\mathbf{H} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top \right)^{-1} \left(\mathbf{H} \hat{\beta} - \mathbf{c} \right) \\ \hat{\lambda} = 2 \left(\mathbf{H} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top \right)^{-1} \left(\mathbf{H} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} - \mathbf{c} \right) \end{cases} \quad (2.6)$$

Thus, We have

Theorem 2.1.1. $\hat{\beta}_H = \hat{\beta} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top \left(\mathbf{H} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top \right)^{-1} (\mathbf{c} - \mathbf{H}\hat{\beta})$. And $\hat{\beta}_H$ is the estimator such that minimizes S_{HE} under the constraints. But $S_{HE} \geq S_E$, where the S_E may not satisfy the constraints.

Property 2.1. $S_{HE} - S_E = (\hat{\beta} - \hat{\beta}_H)^\top \mathbf{X}^\top \mathbf{X} (\hat{\beta} - \hat{\beta}_H)$.

Proof.

$$\begin{aligned}
S_{HE} &= (\mathbf{y} - \mathbf{X}\hat{\beta}_H)^\top (\mathbf{y} - \mathbf{X}\hat{\beta}_H) \\
&= (\mathbf{y} - \mathbf{X}\hat{\beta} + \mathbf{X}\hat{\beta} - \mathbf{X}\hat{\beta}_H)^\top (\mathbf{y} - \mathbf{X}\hat{\beta} + \mathbf{X}\hat{\beta} - \mathbf{X}\hat{\beta}_H) \\
&= (\mathbf{y} - \mathbf{X}\hat{\beta})^\top (\mathbf{y} - \mathbf{X}\hat{\beta}) + (\hat{\beta} - \hat{\beta}_H)^\top \mathbf{X}^\top \mathbf{X} (\hat{\beta} - \hat{\beta}_H) + 2(\mathbf{y} - \mathbf{X}\hat{\beta})^\top (\mathbf{X}\hat{\beta} - \mathbf{X}\hat{\beta}_H) \\
&= S_E + (\hat{\beta} - \hat{\beta}_H)^\top \mathbf{X}^\top \mathbf{X} (\hat{\beta} - \hat{\beta}_H) + 2(\mathbf{X}^\top \mathbf{y} - \mathbf{X}^\top \mathbf{X}\hat{\beta})^\top (\hat{\beta} - \hat{\beta}_H) \\
&= S_E + (\hat{\beta} - \hat{\beta}_H)^\top \mathbf{X}^\top \mathbf{X} (\hat{\beta} - \hat{\beta}_H)
\end{aligned}$$

□

If the $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$, then we have the four following theorems.

Theorem 2.1.2. $\hat{\beta}_H \sim \mathcal{N}(\beta, \sigma^2 \mathbf{G})$, where

$$\mathbf{G} = (\mathbf{X}^\top \mathbf{X})^{-1} \left\{ \mathbf{I} - \mathbf{H}^\top \left(\mathbf{H} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top \right)^{-1} \mathbf{H} (\mathbf{X}^\top \mathbf{X})^{-1} \right\}$$

Proof.

$$\begin{aligned}
\mathbb{E}(\hat{\beta}_H) &= \mathbb{E} \left(\hat{\beta} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top \left(\mathbf{H} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top \right)^{-1} (\mathbf{c} - \mathbf{H}\hat{\beta}) \right) \\
&= \beta + \mathbf{H}^\top \left(\mathbf{H} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top \right)^{-1} (\mathbf{c} - \mathbf{H}\beta) = \beta.
\end{aligned}$$

$$\begin{aligned}
\text{Var}(\hat{\beta}_H) &= \text{Var} \left(\hat{\beta} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top \left(\mathbf{H} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top \right)^{-1} (\mathbf{c} - \mathbf{H}\hat{\beta}) \right) \\
&= \text{Var} \left(\left(\mathbf{I} - (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top \left(\mathbf{H} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top \right)^{-1} \mathbf{H} \right) \beta \right) \\
&= \sigma^2 \left(\mathbf{I} - (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top \left(\mathbf{H} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top \right)^{-1} \mathbf{H} \right) (\mathbf{X}^\top \mathbf{X})^{-1} \left(\mathbf{I} - \mathbf{H}^\top \left(\mathbf{H} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top \right)^{-1} \mathbf{H} (\mathbf{X}^\top \mathbf{X})^{-1} \right) \\
&= \sigma^2 \left((\mathbf{X}^\top \mathbf{X})^{-1} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top \left(\mathbf{H} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top \right)^{-1} \mathbf{H} (\mathbf{X}^\top \mathbf{X})^{-1} \right. \\
&\quad \left. - 2(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top \left(\mathbf{H} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top \right)^{-1} \mathbf{H} (\mathbf{X}^\top \mathbf{X})^{-1} \right) \\
&= \sigma^2 \left((\mathbf{X}^\top \mathbf{X})^{-1} - (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top \left(\mathbf{H} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top \right)^{-1} \mathbf{H} (\mathbf{X}^\top \mathbf{X})^{-1} \right) \\
&= \sigma^2 \mathbf{G}.
\end{aligned}$$

□

Theorem 2.1.3. $\hat{\lambda} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{D})$, where $\mathbf{D} = 4 \left(\mathbf{H} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top \right)^{-1}$.

Proof.

$$\begin{aligned}\mathbb{E}(\hat{\lambda}) &= \mathbb{E}\left(2\left(\mathbf{H}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top\right)^{-1} \left(\mathbf{H}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} - \mathbf{c}\right)\right) \\ &= 2\left(\mathbf{H}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top\right)^{-1} \left(\mathbf{H}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X} \boldsymbol{\beta} - \mathbf{c}\right) \\ &= 2\left(\mathbf{H}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top\right)^{-1} (\mathbf{H} \boldsymbol{\beta} - \mathbf{c}) = 0.\end{aligned}$$

$$\begin{aligned}\text{Var}(\hat{\lambda}) &= \text{Var}\left(2\left(\mathbf{H}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top\right)^{-1} \left(\mathbf{H}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} - \mathbf{c}\right)\right) \\ &= 4\sigma^2 \left(\mathbf{H}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top\right)^{-1} \mathbf{H}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top \left(\mathbf{H}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top\right)^{-1} \\ &= 4\sigma^2 \left(\mathbf{H}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top\right)^{-1}.\end{aligned}$$

□

Theorem 2.1.4. Let $\hat{\mathbf{y}}_H = \mathbf{X}\hat{\boldsymbol{\beta}}_H$, $\hat{\mathbf{e}}_H = \mathbf{y} - \hat{\mathbf{y}}_H$, then $\hat{\mathbf{e}}_H \sim \mathcal{N}(\mathbf{0}, \sigma^2 (\mathbf{I} - \mathbf{XGX}^\top))$.

Proof.

$$\mathbb{E}(\hat{\mathbf{e}}_H) = \mathbb{E}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_H) = \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta} = \mathbf{0}.$$

$$\begin{aligned}\text{Var}(\hat{\mathbf{e}}_H) &= \text{Var}(\mathbf{y} - \hat{\mathbf{y}}_H) = \text{Var}\left(\mathbf{y} - \mathbf{X}\left(\hat{\boldsymbol{\beta}} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H} \left(\mathbf{H}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top\right)^{-1} (\mathbf{c} - \mathbf{H}\hat{\boldsymbol{\beta}})\right)\right) \\ &= \text{Var}\left(\mathbf{y} - \left(\mathbf{X} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H} \left(\mathbf{H}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top\right)^{-1} \mathbf{H}\right) (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}\right) \\ &= \sigma^2 \left(\mathbf{I} - \left(\mathbf{X} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H} \left(\mathbf{H}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top\right)^{-1} \mathbf{H}\right) (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top\right) \\ &\quad \left(\mathbf{I} - \left(\mathbf{X} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H} \left(\mathbf{H}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top\right)^{-1} \mathbf{H}\right) (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top\right)^\top \\ &= \sigma^2 (\mathbf{I} - \mathbf{XGX}^\top).\end{aligned}$$

□

Theorem 2.1.5. $\mathbb{E}(S_{HE} = (n - p - 1 + q) \sigma^2)$, where $S_{HE} = (\mathbf{y} - \hat{\mathbf{y}}_H)^\top (\mathbf{y} - \hat{\mathbf{y}}_H)$.

Proof.

$$\begin{aligned}\mathbb{E}(S_{HE}) &= \sum_{i=1}^n \mathbb{E}(e_{Hi}^2) = \sum_{i=1}^n \text{Var}(e_{Hi}) = \text{tr}(\mathbf{I} - \mathbf{XGX}^\top) \\ &= \text{tr}\left(\mathbf{I} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top + \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top \left(\mathbf{H}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top\right)^{-1} \mathbf{H}\right) \\ &= n - p - 1 + \text{tr}\left(\left(\mathbf{H}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top\right)^{-1} \mathbf{H}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top \mathbf{X}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1}\right) \\ &= n - p - 1 + q.\end{aligned}$$

□

Corollary 2.1.5.1. From theorem (2.1.5), an unbiased estimator of σ^2 is $\hat{\sigma}^2 = \frac{1}{n-p-1+q} S_{HE}$.

3 Hypothesis Testing and Interval Estimation

3.1 General Hypothesis Testing

Consider the hypothesis testing: $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{c}$. We can use content of section 2 to test this hypothesis. That is, we can fit a new linear model under the constraint of H_0 and then calculate its S_{HE} which we know is always no less than S_E . And if the S_{HE} is larger than the S_E too much, we can reject the null hypothesis.

From property (2.1), we have $S_{HE} - S_E = (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_H)^\top (\mathbf{X}^\top \mathbf{X})^{-1} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_H)$, which can be written as

$$\begin{aligned} S_{HE} - S_E &= \left((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H} \left(\mathbf{H} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top \right)^{-1} (\mathbf{H} \hat{\boldsymbol{\beta}} - \mathbf{c}) \right)^\top (\mathbf{X}^\top \mathbf{X})^{-1} \left((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H} \left(\mathbf{H} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top \right)^{-1} (\mathbf{H} \hat{\boldsymbol{\beta}} - \mathbf{c}) \right) \\ &= (\mathbf{H} \hat{\boldsymbol{\beta}} - \mathbf{c})^\top \left(\mathbf{H} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top \right)^{-1} (\mathbf{H} \hat{\boldsymbol{\beta}} - \mathbf{c}). \end{aligned} \quad (3.1)$$

Under the null hypothesis, we have $\mathbf{H} \hat{\boldsymbol{\beta}} - \mathbf{c} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{H} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top)$. Since $\mathbf{H} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top$ is positive definite, thus we further have $\left(\mathbf{H} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top \right)^{-1/2} (\mathbf{H} \hat{\boldsymbol{\beta}} - \mathbf{c}) / \sigma \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, which indicates

$$(\mathbf{H} \hat{\boldsymbol{\beta}} - \mathbf{c})^\top \left(\mathbf{H} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top \right)^{-1} (\mathbf{H} \hat{\boldsymbol{\beta}} - \mathbf{c}) \sim \chi^2(q)$$

Also, we already have $S_E \sim \chi^2(n - p - 1)$ and S_E is independent from $\hat{\boldsymbol{\beta}}$, which implies $S_{HE} - S_E$ is independent from S_E , thus we have $F = \frac{S_{HE} - S_E}{S_E} \frac{n - p - 1}{q} \sim \mathcal{F}(q, n - p - 1)$ under the null hypothesis. That is, we can use F to test the null hypothesis.

3.2 Goodness of Fit

We consider the following hypothesis testing in this section:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0,$$

which is a special case of the general hypothesis testing in section 3.1. The null hypothesis can also be

written as $\mathbf{H}\boldsymbol{\beta} = \mathbf{0}$, where $\mathbf{H} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}_{p \times (p+1)}$.

Under the null hypothesis, we have all coefficients equal to 0 except the intercept β_0 , which is the mean of $\{y_i\}$ in this case. Thus, the S_{HE} is simply equal to $S_T = \sum_{i=1}^n (y_i - \bar{y})^2$, which is $(n_1) S^2$ where S^2 is the sample variation.

Recall the equation $S_T = S_E + S_R$, then we can construct the F statistic, which is $\frac{S_R}{S_E} \frac{p}{n - p - 1} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^n (y_i - \hat{y}_i)^2} \frac{p}{n - p - 1}$, to test the null hypothesis. Under the null hypothesis, $F \sim \mathcal{F}(p, n - p - 1)$. Thus, if $F > \mathcal{F}_{1-\alpha}(p, n - p - 1)$, then we can reject the null hypothesis.

3.3 Testing the Individual Coefficient

We now consider the individual testing: $H_{0j} : \beta_j = 0$. We can also use the general hypothesis to test this one. Let $\mathbf{H} = \mathbf{e}_{j+1}^\top$, where \mathbf{e}_{j+1} is vector with 1 in its $j + 1$ th entry and all other entries equal to 0. Then, we use $F = \frac{S_{HE} - S_E}{S_E} \frac{1}{n - p - 1}$, which follows $\mathcal{F}(1, n - p - 1)$ under the null hypothesis, to test the null

hypothesis. In this case, the $S_{HE} - S_E$ is called *partial correlation* of x_j . We can show that

$$\begin{aligned} S_{HE} - S_E &= (\mathbf{H}\hat{\boldsymbol{\beta}} - \mathbf{c})^\top (\mathbf{H}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{H}^\top)^{-1} (\mathbf{H}\hat{\boldsymbol{\beta}} - \mathbf{c}) \\ &= \beta_j^2 / c_{jj}, \end{aligned} \quad (3.2)$$

where c_{jj} is the j th entry in the diagonal of $(\mathbf{X}^\top \mathbf{X})^{-1}$.

What's more, we can use the t test to do the work. We already have $\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1})$. Use the property of multivariate normal distribution, we have $\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2 c_{jj})$, which implies $\frac{\hat{\beta}_j - \beta_j}{\sigma \sqrt{c_{jj}}} \sim \mathcal{N}(0, 1)$. Let $t_j = \frac{\hat{\beta}_j}{\hat{\sigma} \sqrt{c_{jj}}}$, which follows $t(n-p-1)$ under the null hypothesis (recall that $\hat{\sigma}^2 = \frac{S_E}{n-p-1}$). Then, if $|t_j| > t_{1-\alpha/2}(n-p-1)$, we can reject the null hypothesis.

3.4 Forecast and Interval Estimation

Consider the case where we have a *new* sample with \mathbf{x}_0^\top and we want to forecast the y_0 with our fitted model. Denote the fitted value as \hat{y}_0 . We have the following property:

Property 3.1. \hat{y}_0 is an unbiased estimator of y_0 .

Proof. $\mathbb{E}(\hat{y}_0) = \mathbb{E}(\mathbf{x}_0^\top \hat{\boldsymbol{\beta}}) = \mathbb{E}(\mathbf{x}_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}) = \mathbf{x}_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X} \boldsymbol{\beta} = \mathbf{x}_0^\top \boldsymbol{\beta} = y_0.$ \square

Property 3.2. Among all the unbiased linear by \mathbf{y} estimators of y_0 , \hat{y}_0 has the least variance.

Proof. Let $\mathbf{c}^\top \mathbf{y}$ be an unbiased estimator of y_0 . Then, we have $\mathbb{E}(\mathbf{c}^\top \mathbf{y}) = \mathbf{c}^\top \mathbf{X} \boldsymbol{\beta} = \mathbf{x}_0^\top \boldsymbol{\beta}$, which implies that $\mathbf{c}^\top \mathbf{X} = \mathbf{x}_0^\top$. We have

$$\begin{aligned} \text{Var}(\mathbf{c}^\top \mathbf{y}) - \text{Var}(\hat{y}_0) &= \sigma^2 \mathbf{c}^\top \mathbf{c} - \text{Var}(\mathbf{x}_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}) \\ &= \sigma^2 (\mathbf{c}^\top \mathbf{c} - \mathbf{x}_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_0) \\ &= \sigma^2 (\mathbf{c}^\top (\mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top) \mathbf{c}) \\ &= \sigma^2 \mathbf{c}^\top (\mathbf{I} - \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top) \mathbf{c} \\ &= \sigma^2 \mathbf{c}^\top (\mathbf{I} - \mathbf{H}) \mathbf{c} \\ &\geq 0 \end{aligned}$$

since $\mathbf{I} - \mathbf{H}$ is positive semidefinite. \square

Property 3.3. Under the assumption of normality, we have

$$\hat{y}_0 - y_0 \sim \mathcal{N}\left(0, \sigma^2 \left(1 + \mathbf{x}_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_0\right)\right).$$

Proof. We have $y_0 \sim \mathcal{N}(\mathbf{x}_0 \boldsymbol{\beta}, \sigma^2)$ and $\hat{y}_0 \sim \mathcal{N}(\mathbf{x}_0 \boldsymbol{\beta}, \sigma^2 \mathbf{x}_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_0)$, thus

$$\hat{y}_0 - y_0 \sim \mathcal{N}\left(0, \sigma^2 \left(1 + \mathbf{x}_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_0\right)\right).$$

\square

Property 3.4. $\frac{\hat{y}_0 - y_0}{\hat{\sigma} \sqrt{1 + \mathbf{x}_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_0}} \sim t(n-p-1).$

Property 3.5. The $1 - \alpha$ forecast interval of y_0 is

$$\left[\hat{y}_0 - t_{1-\alpha/2}(n-p-1) \hat{\sigma} \sqrt{1 + \mathbf{x}_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_0}, \hat{y}_0 + t_{1-\alpha/2}(n-p-1) \hat{\sigma} \sqrt{1 + \mathbf{x}_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_0}, \right]$$

4 Model Selection

4.1 Some Criteria

- adjusted $R^2 = 1 - \frac{n-1}{n-p'} (1 - R^2)$
- $\text{AIC} = -2 \ln L + 2p'$, under the normal linear model, we have

$$-2 \ln L = n \ln S_E$$

- $\text{BIC} = -2 \ln L + p' \ln n$
- $\text{RES} = \frac{S_E}{n-p'}$
- $C_p = \frac{S_E}{\hat{\sigma}^2} - (n - 2p')$, where $\hat{\sigma}^2$ is estimated under the full model.

5 Ridge Regression

The design matrix \mathbf{X} in this section is centered, which means it has no intercept term.

5.1 Multicollinearity

Let λ be an eigenvalue of $\mathbf{X}^\top \mathbf{X}$ and ϕ be its scaled eigenvector with length equal to 1. We have $\mathbf{X}^\top \mathbf{X} \phi = \lambda \phi$. Multiply ϕ^\top to the both sides, $\phi^\top \mathbf{X}^\top \mathbf{X} \phi = (\mathbf{X} \phi)^\top (\mathbf{X} \phi) = \|\mathbf{X} \phi\|_2^2 = \phi^\top \lambda \phi = \lambda \|\phi\|_2^2$. If λ is close to 0, then we have $\mathbf{X} \phi = \phi_1 \mathbf{x}_1 + \dots + \phi_{p'} \mathbf{x}_{p'} \approx 0$. That is, if there is an eigenvalue of $\mathbf{X}^\top \mathbf{X}$ close to 0, then $\mathbf{X}^\top \mathbf{X}$ may have *multicollinearity* which may result the variance of the OLS estimators very large.

We introduce a criteria called *MSE* (mean squared error) here.

$$\begin{aligned} \text{MSE} &= \mathbb{E} \left\| \hat{\theta} - \theta \right\|^2 = \mathbb{E} \left(\hat{\theta} - \theta \right)^\top \left(\hat{\theta} - \theta \right) = \mathbb{E} \left(\hat{\theta}^\top \hat{\theta} - 2 \hat{\theta}^\top \theta + \theta^\top \theta \right) \\ &= \mathbb{E} \left(\hat{\theta}^\top \hat{\theta} - 2 \hat{\theta}^\top \theta + \theta^\top \theta \right) + \mathbb{E} \left(\hat{\theta} \right)^\top \mathbb{E} \left(\hat{\theta} \right) - \mathbb{E} \left(\hat{\theta} \right)^\top \mathbb{E} \left(\hat{\theta} \right) \\ &= \mathbb{E} \left(\hat{\theta}^\top \hat{\theta} \right) - \mathbb{E} \left(\hat{\theta} \right)^\top \mathbb{E} \left(\hat{\theta} \right) + \mathbb{E} \left(\hat{\theta} \right)^\top \mathbb{E} \left(\hat{\theta} \right) - 2 \mathbb{E} \left(\hat{\theta}^\top \right) \theta + \theta^\top \theta \\ &= \text{tr} \left(\text{Cov} \left(\hat{\theta} \right) \right) + \left(\mathbb{E} \hat{\theta} - \theta \right)^\top \left(\mathbb{E} \hat{\theta} - \theta \right) \\ &= \text{tr} \left(\text{Cov} \left(\hat{\theta} \right) \right) + \left\| \mathbb{E} \hat{\theta} - \theta \right\|^2. \end{aligned} \tag{5.1}$$

Then, we can derive the MSE of $\hat{\beta}$:

$$\text{MSE} \left(\hat{\beta} \right) = \text{tr} \left(\sigma^2 \left(\mathbf{X}^\top \mathbf{X} \right)^{-1} \right) + \left\| \mathbb{E} \hat{\beta} - \beta \right\|^2 = \sigma^2 \text{tr} \left(\left(\mathbf{X}^\top \mathbf{X} \right)^{-1} \right). \tag{5.2}$$

We have already known that $\mathbf{X}^\top \mathbf{X}$ is positive definite. Thus there exists unitary matrix \mathbf{P} and diagonal matrix

$$\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_p) \text{ such that } \mathbf{X}^\top \mathbf{X} = \mathbf{P} \mathbf{\Lambda} \mathbf{P} \text{ which implies that } \left(\mathbf{X}^\top \mathbf{X} \right)^{-1} = \mathbf{P} \begin{bmatrix} \frac{1}{\lambda_1} & & \\ & \frac{1}{\lambda_2} & \\ & & \ddots \\ & & & \frac{1}{\lambda_p} \end{bmatrix} \mathbf{P}^\top.$$

Thus, $\text{MSE} \left(\hat{\beta} \right) = \sigma^2 \sum_{i=1}^p \frac{1}{\lambda_i}$. If there exists multicollinearity, then one of the λ_i would be close to zero, leading to the MSE of $\hat{\beta}$ expanding to large values and unstable estimator of β . To deal with this problem, *Ridge regression* is proposed.

5.2 Ridge Regression

Let k be any non-negative scalar. The OLS is modified as the following:

$$\hat{\beta}(k) = (\mathbf{X}^\top \mathbf{X} + k\mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}. \quad (5.3)$$

The $k\mathbf{I}$ may decrease the variance of the estimator.

Property 5.1. $\hat{\beta}(k)$ is a biased estimator of β .

Property 5.2. $\hat{\beta}(k)$ is a linear transformation of the OLS estimator.

Proof. $\hat{\beta}(k) = (\mathbf{X}^\top \mathbf{X} + k\mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y} = (\mathbf{X}^\top \mathbf{X} + k\mathbf{I})^{-1} \mathbf{X}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = (\mathbf{X}^\top \mathbf{X} + k\mathbf{I})^{-1} \mathbf{X}^\top \mathbf{X} \hat{\beta}$. \square

Property 5.3. For all $k > 0$, $\|\hat{\beta}\| \neq 0$, we always have

$$\|\hat{\beta}(k)\| \leq \|\hat{\beta}\|.$$

Proof. First, we denote $\Lambda(k) = \Lambda + k\mathbf{I} = \text{diag}(\lambda_1 + k, \dots, \lambda_p + k)$.

$$\begin{aligned} \|\hat{\beta}(k)\|^2 &= \left((\mathbf{X}^\top \mathbf{X} + k\mathbf{I})^{-1} \mathbf{X}^\top \mathbf{X} \hat{\beta} \right)^\top \left((\mathbf{X}^\top \mathbf{X} + k\mathbf{I})^{-1} \mathbf{X}^\top \mathbf{X} \hat{\beta} \right) \\ &= \left((\mathbf{P}\Lambda\mathbf{P}^\top + k\mathbf{I})^{-1} \mathbf{X}^\top \mathbf{X} \hat{\beta} \right)^\top \left((\mathbf{P}\Lambda\mathbf{P}^\top + k\mathbf{I})^{-1} \mathbf{X}^\top \mathbf{X} \hat{\beta} \right) \\ &= \left((\mathbf{P}\Lambda(k)\mathbf{P}^\top)^{-1} \mathbf{X}^\top \mathbf{X} \hat{\beta} \right)^\top \left((\mathbf{P}\Lambda(k)\mathbf{P}^\top)^{-1} \mathbf{X}^\top \mathbf{X} \hat{\beta} \right) \\ &= \left(\mathbf{P}\Lambda^{-1}(k)\mathbf{P}^\top \mathbf{X}^\top \mathbf{X} \hat{\beta} \right)^\top \left(\mathbf{P}\Lambda^{-1}(k)\mathbf{P}^\top \mathbf{X}^\top \mathbf{X} \hat{\beta} \right) \\ &= \left(\mathbf{P}\Lambda^{-1}(k)\mathbf{P}^\top \mathbf{P}\Lambda\mathbf{P}^\top \hat{\beta} \right)^\top \left(\mathbf{P}\Lambda^{-1}(k)\mathbf{P}^\top \mathbf{P}\Lambda\mathbf{P}^\top \hat{\beta} \right) \\ &= \left(\mathbf{P}\Lambda^{-1}(k)\Lambda\mathbf{P}^\top \hat{\beta} \right)^\top \left(\mathbf{P}\Lambda^{-1}(k)\Lambda\mathbf{P}^\top \hat{\beta} \right) \\ &= \hat{\beta}^\top \mathbf{P}\Lambda^2\Lambda^{-2}(k)\mathbf{P}^\top \hat{\beta} \\ &= \left(\Lambda\Lambda^{-1}(k)\hat{\beta} \right)^\top \mathbf{P}\mathbf{P}^\top \left(\Lambda\Lambda^{-1}(k)\hat{\beta} \right) \\ &= \left\| \begin{bmatrix} \frac{\lambda_1}{\lambda_1+k} & & \\ & \frac{\lambda_2}{\lambda_2+k} & \\ & & \ddots & \\ & & & \frac{\lambda_p}{\lambda_p+k} \end{bmatrix} \hat{\beta} \right\|^2 \leq \|\hat{\beta}\|^2. \end{aligned} \quad (5.4)$$

\square

Property 5.4. There must exist a $k > 0$ such that $\text{MSE}(\hat{\beta}(k)) \leq \text{MSE}(\hat{\beta})$.

Proof. First, we transform the ordinary form of linear model

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

to $\mathbf{y} = \mathbf{X}\mathbf{P}\mathbf{P}^\top\beta + \epsilon = \mathbf{Z}\alpha + \epsilon$, where $\mathbf{Z} = \mathbf{X}\mathbf{P}$, $\alpha = \mathbf{P}^\top\beta$. Then, we have $\mathbf{Z}^\top\mathbf{Z} = \Lambda$.

The OLS estimator of α is $\hat{\alpha} = (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{y} = \mathbf{\Lambda}^{-1} \mathbf{Z}^\top \mathbf{y}$. And the ridge estimator of α is $\hat{\alpha}(k) = (\mathbf{Z}^\top \mathbf{Z} + k\mathbf{I})^{-1} \mathbf{Z}^\top \mathbf{y} = \mathbf{\Lambda}^{-1}(k) \mathbf{Z}^\top \mathbf{y}$. Also, we have

$$\begin{aligned} \|\hat{\beta}(k) - \beta\|^2 &= \left(\hat{\beta}(k) - \beta\right)^\top \left(\hat{\beta}(k) - \beta\right) \\ &= (\mathbf{P}\hat{\alpha}(k) - \mathbf{P}\alpha)^\top (\mathbf{P}\hat{\alpha}(k) - \mathbf{P}\alpha) \\ &= (\hat{\alpha}(k) - \alpha)^\top (\hat{\alpha}(k) - \alpha) \\ &= \|\hat{\alpha}(k) - \alpha\|^2. \end{aligned}$$

Specially, when $k = 0$, we have $\|\hat{\beta} - \beta\| = \|\hat{\alpha} - \alpha\|$. Then, we have

$$\begin{aligned} \text{MSE}(\hat{\alpha}(k)) &= \text{tr}(\text{Cov}(\hat{\alpha}(k))) + \|\mathbb{E}(\hat{\alpha}(k) - \alpha)\|^2 \\ &= \sigma^2 \text{tr}(\mathbf{\Lambda}^{-2}(k) \mathbf{\Lambda}) + (\mathbf{\Lambda}^{-1}(k) \mathbf{\Lambda} \alpha - \alpha)^\top (\mathbf{\Lambda}^{-1}(k) \mathbf{\Lambda} \alpha - \alpha) \\ &= \sigma^2 \sum_{i=1}^p \frac{\lambda_i}{(\lambda_i + k)^2} + \alpha^\top (\mathbf{\Lambda}^{-1}(k) \mathbf{\Lambda} - \mathbf{I})^\top (\mathbf{\Lambda}^{-1}(k) \mathbf{\Lambda} - \mathbf{I}) \alpha \\ &= \sigma^2 \sum_{i=1}^p \frac{\lambda_i}{(\lambda_i + k)^2} + \alpha^\top (\mathbf{\Lambda}^{-2}(k) \mathbf{\Lambda}^2 - 2\mathbf{\Lambda}^{-1}(k) \mathbf{\Lambda} + \mathbf{I}) \alpha \\ &= \sigma^2 \sum_{i=1}^p \frac{\lambda_i}{(\lambda_i + k)^2} + \sum_{i=1}^p \alpha_i^2 \left(\frac{\lambda_i^2}{(\lambda_i + k)^2} - 2\frac{\lambda_i}{\lambda_i + k} + 1 \right) \\ &= \sigma^2 \sum_{i=1}^p \frac{\lambda_i}{(\lambda_i + k)^2} + \sum_{i=1}^p \alpha_i^2 \left(\frac{\lambda_i^2 - 2\lambda_i^2 - 2k\lambda_i + \lambda_i^2 + 2k\lambda_i + k^2}{(\lambda_i + k)^2} \right) \\ &= \sigma^2 \sum_{i=1}^p \frac{\lambda_i}{(\lambda_i + k)^2} + k^2 \sum_{i=1}^p \left(\frac{\alpha_i}{\lambda_i + k} \right)^2. \end{aligned}$$

Let $g_1(k) := \sigma^2 \sum_{i=1}^p \frac{\lambda_i}{(\lambda_i + k)^2}$, $g_2(k) = k^2 \sum_{i=1}^p \left(\frac{\alpha_i}{\lambda_i + k} \right)^2 = \sum_{i=1}^p \frac{\alpha_i^2 k^2}{(\lambda_i + k)^2}$, then $g(k) = \text{MSE}(\hat{\beta}(k)) = \text{MSE}(\hat{\alpha}(k)) = g_1(k) + g_2(k)$. And we have

$$\begin{aligned} g'(k) &= g'_1(k) + g'_2(k) \\ &= -2\sigma^2 \sum_{i=1}^p \frac{\lambda_i}{(\lambda_i + k)^3} + \sum_{i=1}^p \frac{2\alpha_i^2 k - 2\alpha_i^2 k^2 (\lambda_i + k)}{(\lambda_i + k)^4} \\ &= -2\sigma^2 \sum_{i=1}^p \frac{\lambda_i}{(\lambda_i + k)^3} + \sum_{i=1}^p \frac{2\alpha_i^2 k - 2\alpha_i^2 k^2 \lambda_i - 2\alpha_i^2 k^3}{(\lambda_i + k)^4}, \end{aligned}$$

and when $k = 0$, we have $g'(0) = -2\sigma^2 \sum_{i=1}^p \frac{1}{\lambda_i^2} < 0$, thus in a neighbourhood of 0, there exists $k > 0$ such that $\text{MSE}(\hat{\beta}(k)) \leq \text{MSE}(\hat{\beta})$. \square