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

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1.1}$$

where $\boldsymbol{y} \in \mathbb{R}^{n\times 1}$ 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*)

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

Assumption 1.2. (*Normality Condition*)

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

combined with (1.1), we have

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

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

$$Loss = \left(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\right)^\top \left(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\right). \tag{1.5}$$

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

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

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 $\left(\mathbf{X}^\top\mathbf{X}\right)\boldsymbol{v} = \mathbf{0}$ if and only if $\boldsymbol{v} = \mathbf{0}$. If $\mathbf{X}\boldsymbol{v} = \mathbf{0}$, then $\mathbf{X}^\top\mathbf{X}\boldsymbol{v} = \mathbf{0}$, thus nullity($\mathbf{X}$) $\subset$ nullity($\mathbf{X}^\top\mathbf{X}$). For the other direction, if $\mathbf{X}^\top\mathbf{X}\boldsymbol{v} = \mathbf{0}$, then $\boldsymbol{v}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{v} = (\mathbf{X}\boldsymbol{v})^\top(\mathbf{X}\boldsymbol{v}) = \mathbf{0}$, if and only if $\mathbf{X}\boldsymbol{v} = \mathbf{0}$, thus nullity($\mathbf{X}^\top\mathbf{X}$) $\subset$ nullity($\mathbf{X}$). In conclusion, we have nullity($\mathbf{X}^\top\mathbf{X}$) = nullity($\mathbf{X}$). And notice that nullity($\mathbf{X}$)= 0, thus nullity($\mathbf{X}^\top\mathbf{X}$)= 0, which implies that $\mathbf{X}^\top\mathbf{X}$ is invertible. □

Proof. Theorem 1.1

Using matrix calculus, we have

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

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

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

□

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 $\boldsymbol{y}$ and $\mathbf{X}\boldsymbol{\beta}$, i.e. minimize $\|\mathbf{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2$. Recall that only the projection of $\boldsymbol{y}$ on the range($\mathbf{X}$) can minimize this distance, so we only need to find the projection of $\boldsymbol{y}$, which implies that $\mathbf{X}\boldsymbol{\beta} - \boldsymbol{y}$ should be perpendicular to $\mathbf{X}\boldsymbol{\beta}$ which is in the range of $\mathbf{X}$. Recall that the *orthogonal complement* of range($\mathbf{X}$) is nullity($\mathbf{X}^\top$), thus such $\mathbf{X}\boldsymbol{\beta} - \boldsymbol{y}$ must in nullity($\mathbf{X}^\top$), indicating $\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta} - \boldsymbol{y}) = \mathbf{0}$, leading to $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\boldsymbol{y}$.

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 $\boldsymbol{y}$ conditioned on $\mathbf{X}$ is given as (1.4), which takes the form as follows:

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

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

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

which is equivalent to the OLS.

1.4 Properties of the Estimators

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\mathrm{Cov}\left(\hat{\boldsymbol{\beta}}, \boldsymbol{e}\right) = \mathrm{Cov}\left(\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\boldsymbol{y}, \left(\boldsymbol{I}-\boldsymbol{H}\right)\boldsymbol{y}\right) = \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\left(\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top - \boldsymbol{I}\right)^\top = \sigma^2\left(\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top - \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\right) = \mathbf{0}.$$

☐

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

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

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

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 $\sigma^2 = 1$.

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

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

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 $S_T/\sigma^2 \sim \chi^2(n-1)$. $\square$

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

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

Proof. $\mathbf{X}^\top \boldsymbol{e} = \mathbf{X}^\top \left(\mathbf{I} - \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\right)\boldsymbol{y} = \mathbf{0}$. $\square$

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 $\mathbf{1}_{1\times n}$. $\square$

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

1.7 Generalized Least Square

Suppose that the random errors now have covariance $\mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{\Sigma}$, 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 $\boldsymbol{v} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\epsilon}$, $\mathbf{U} = \boldsymbol{\Sigma}^{-1/2}\mathbf{X}$, $\boldsymbol{z} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{y}$, then we have

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

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

Theorem 1.7.1. (1) $\mathbb{E}(\boldsymbol{\beta}^*) = \boldsymbol{\beta}$,
(2) $\mathrm{Var}(\boldsymbol{\beta}^*) = \sigma^2\left(\mathbf{X}^\top\boldsymbol{\Sigma}^{-1}\mathbf{X}\right)^{-1}$,
(3) For all fixed constant vector $\boldsymbol{c}$, $\boldsymbol{c}^\top\boldsymbol{\beta}^*$ is the only linear by $\boldsymbol{y}$ unbiased estimator with minimal variance of $\boldsymbol{c}^\top\boldsymbol{\beta}$.

Proof. (1) $\mathbb{E}(\boldsymbol{\beta}^*) = \mathbb{E}\left(\left(\mathbf{X}^\top \boldsymbol{\Sigma}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{y}\right) = \left(\mathbf{X}^\top \boldsymbol{\Sigma}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^\top\boldsymbol{\Sigma}^{-1}\mathbf{X} = 0.$
(2)

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

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

2 Linear Model with Linear Constraint

2.1 Estimation

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

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

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

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

To solve this function, let

$$\begin{cases}\dfrac{\partial\mathcal{L}}{\partial\boldsymbol{\beta}} = 2\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} - 2\mathbf{X}^\top\boldsymbol{y} + \mathbf{H}^\top\boldsymbol{\lambda} = 0\\ \dfrac{\partial\mathcal{L}}{\partial\boldsymbol{\lambda}} = \boldsymbol{c} - \mathbf{H}\boldsymbol{\beta} = 0,\end{cases} \tag{2.2}$$

Then, we have

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

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

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

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

Thus, We have

Theorem 2.1.1. $\hat{\boldsymbol{\beta}}_H = \hat{\boldsymbol{\beta}} + \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{H}^\top\left(\mathbf{H}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{H}^\top\right)^{-1}\left(\boldsymbol{c} - \mathbf{H}\hat{\boldsymbol{\beta}}\right)$. And $\hat{\boldsymbol{\beta}}_H$ 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. $S_{HE} - S_E = \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_H\right)^\top\mathbf{X}^\top\mathbf{X}\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_H\right)$.

Proof.

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

□

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

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

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

.

Proof.

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

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

□

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

Proof.

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

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

□

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

Proof.

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

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

□

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

Proof.

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

□

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: $H_0 : \mathbf{H}\boldsymbol{\beta} = \boldsymbol{c}$. 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 $S_{HE} - S_E = \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_H\right)^\top \left(\mathbf{X}^\top\mathbf{X}\right)^{-1} \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_H\right)$, which can be written as

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

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

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

.

Also, we already have $S_E \sim \chi^2\left(n-p-1\right)$ and S_E is independent from $\hat{\boldsymbol{\beta}}$, which implies $S_{HE} - S_E$ is independent form S_E, thus we have $F = \frac{S_{HE}-S_E}{S_E}\frac{n-p-1}{q} \sim \mathcal{F}\left(q, n-p-1\right)$ 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 $S_T = \sum_{i=1}^n \left(y_i - \bar{y}\right)^2$, which is $\left(n_1\right)S^2$ 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 $\frac{S_R}{S_E}\frac{p}{n-p-1} = \frac{\sum_{i=1}^n(\hat{y_i}-\bar{y_i})^2}{\sum_{i=1}^n(y_i-haty_i)^2}\frac{p}{n-p-1}$, to test the null hypothesis. Under the null hypothesis, $F \sim \mathcal{F}\left(p, n-p-1\right)$. Thus, if $F > \mathcal{F}_{1-\alpha}\left(p, n-p-1\right)$, 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 $\mathbf{H} = \boldsymbol{e}_{j+1}^\top$, where $\boldsymbol{e}_{j+1}$ 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 $\mathcal{F}\left(1, n-p-1\right)$ 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

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

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

What's more, we can use the t test to do the work. We already have $\hat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\right)$. Use the property of multivariate normal distribution, we have $\hat{\beta}_j \sim \mathcal{N}\left(\beta_j, \sigma^2 c_{jj}\right)$, 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 $\boldsymbol{x}_0^\top$ 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. $\mathbb{E}(\hat{y}_0) = \mathbb{E}\left(\boldsymbol{x}_0^\top\hat{\boldsymbol{\beta}}\right) = \mathbb{E}\left(\boldsymbol{x}_0^\top\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\boldsymbol{y}\right) = \boldsymbol{x}_0^\top\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} = \boldsymbol{x}_0^\top\boldsymbol{\beta} = y_0$. □

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

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

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

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

Property 3.3. Under the assumption of normality, we have

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

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

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

□

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

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

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

.

4 Model Selection

4.1 Some Criteria

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

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

- **BIC** $= -2\ln L + p'\ln n$
- 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 a eigenvalue of $\mathbf{X}^\top\mathbf{X}$ and $\boldsymbol{\phi}$ be its scaled eigenvector with length equal to 1. We have $\mathbf{X}^\top\mathbf{X}\boldsymbol{\phi} = \lambda\boldsymbol{\phi}$. Multiply $\boldsymbol{\phi}^\top$ to the both sides, $\boldsymbol{\phi}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\phi} = (\mathbf{X}\boldsymbol{\phi})^\top(\mathbf{X}\boldsymbol{\phi}) = \|\mathbf{X}\boldsymbol{\phi}\|_2^2 = \boldsymbol{\phi}^\top\lambda\boldsymbol{\phi} = \lambda\|\boldsymbol{\phi}\|_2^2$. If λ is close to 0, then we have $\mathbf{X}\boldsymbol{\phi} = \phi_1\boldsymbol{x}_1 + \cdots + \phi_{p'}\boldsymbol{x}_{p'} \approx 0$. 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.

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

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

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

We have already known that $\mathbf{X}^\top\mathbf{X}$ is positive definite. Thus there exits unitary matrix $\mathbf{P}$ and diagonal matrix $\boldsymbol{\Lambda} = \text{diag}\left(\lambda_1, \ldots, \lambda_p\right)$ such that $\mathbf{X}^\top\mathbf{X} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}$ 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, $\text{MSE}\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2\sum_{i=1}^p\frac{1}{\lambda_i}$. If there exists multicollinearity, then one of the λ_i would be close to zero, leading to the MSE of $\hat{\boldsymbol{\beta}}$ expanding to large values and unstable estimator of $\boldsymbol{\beta}$. 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:

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

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

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

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

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

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

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

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

$$\begin{aligned}
\left\|\hat{\boldsymbol{\beta}}(k)\right\|^2 &= \left(\left(\mathbf{X}^\top\mathbf{X} + k\mathbf{I}\right)^{-1}\mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}}\right)^\top\left(\left(\mathbf{X}^\top\mathbf{X} + k\mathbf{I}\right)^{-1}\mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}}\right)\\
&= \left(\left(\mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^\top + k\mathbf{I}\right)^{-1}\mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}}\right)^\top\left(\left(\mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^\top + k\mathbf{I}\right)^{-1}\mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}}\right)\\
&= \left(\left(\mathbf{P}\boldsymbol{\Lambda}(k)\mathbf{P}^\top\right)^{-1}\mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}}\right)^\top\left(\left(\mathbf{P}\boldsymbol{\Lambda}(k)\mathbf{P}^\top\right)^{-1}\mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}}\right)\\
&= \left(\mathbf{P}\boldsymbol{\Lambda}^{-1}(k)\mathbf{P}^\top\mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}}\right)^\top\left(\mathbf{P}\boldsymbol{\Lambda}^{-1}(k)\mathbf{P}^\top\mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}}\right)\\
&= \left(\mathbf{P}\boldsymbol{\Lambda}^{-1}(k)\mathbf{P}^\top\mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^\top\hat{\boldsymbol{\beta}}\right)^\top\left(\mathbf{P}\boldsymbol{\Lambda}^{-1}(k)\mathbf{P}^\top\mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^\top\hat{\boldsymbol{\beta}}\right)\\
&= \left(\mathbf{P}\boldsymbol{\Lambda}^{-1}(k)\boldsymbol{\Lambda}\mathbf{P}^\top\hat{\boldsymbol{\beta}}\right)^\top\left(\mathbf{P}\boldsymbol{\Lambda}^{-1}(k)\boldsymbol{\Lambda}\mathbf{P}^\top\hat{\boldsymbol{\beta}}\right)\\
&= \hat{\boldsymbol{\beta}}^\top\mathbf{P}\boldsymbol{\Lambda}^2\boldsymbol{\Lambda}^{-2}(k)\mathbf{P}^\top\hat{\boldsymbol{\beta}}\\
&= \left(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{-1}(k)\hat{\boldsymbol{\beta}}\right)^\top\mathbf{P}\mathbf{P}^\top\left(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{-1}(k)\hat{\boldsymbol{\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{\boldsymbol{\beta}}\right\|^2 \leq \left\|\hat{\boldsymbol{\beta}}\right\|^2.
\end{aligned} \tag{5.4}$$

☐

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

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

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

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

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

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

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

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

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

$$\begin{aligned}
g'(k) &= g_1'(k)+g_2'(k) \\
&= -2\sigma^2\sum_{i=1}^{p}\frac{\lambda_i}{\left(\lambda_i+k\right)^3} + \sum_{i=1}^{p}\frac{2\alpha_i^2k-2\alpha_i^2k^2\left(\lambda_i+k\right)}{\left(\lambda_i+k\right)^4} \\
&= -2\sigma^2\sum_{i=1}^{p}\frac{\lambda_i}{\left(\lambda_i+k\right)^3} + \sum_{i=1}^{p}\frac{2\alpha_i^2k-2\alpha_i^2k^2\lambda_i-2\alpha_i^2k^3}{\left(\lambda_i+k\right)^4},
\end{aligned}$$

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 $\operatorname{MSE}\left(\hat{\boldsymbol{\beta}}(k)\right) \le \operatorname{MSE}\left(\hat{\boldsymbol{\beta}}\right)$. $\square$