Element of Statistical Learning Note

Zichong Wang

December 29, 2025

1 Chap 3: Linear Methods for Regression

1.1 Confidence region for $\hat{\beta}$

We know that $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$, thus $\hat{\beta} - \beta \sim \mathcal{N}(0, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$.

Since for $\boldsymbol{z} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, we have $\boldsymbol{z}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{z} \sim \chi^2_k$, where $k = \mathrm{rank}(\boldsymbol{\Sigma})$, we have

$$(\hat{\beta} - \beta)^T(\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})^{-1}(\hat{\beta} - \beta) \sim \chi^2_{p+1}$$

Thus,

$$\frac{1}{\sigma^2}(\hat{\beta} - \beta)^T\boldsymbol{X}^T\boldsymbol{X}(\hat{\beta} - \beta) \sim \chi^2_{p+1}$$

And have approx confidence region

$$(\hat{\beta} - \beta)^T\boldsymbol{X}^T\boldsymbol{X}(\hat{\beta} - \beta) \le \hat{\sigma}^2\chi^2_{p+1,1-\alpha}$$

Actually, $\hat{\sigma}^2 = \frac{1}{N-p-1}\sum_{i=1}^N(y_i - \hat{y}_i)^2 = \frac{1}{N-p-1}\mathrm{RSS}$, and $\mathrm{RSS}/\sigma^2 \sim \chi^2_{N-p-1}$,

$$\frac{(\hat{\beta} - \beta)^T\boldsymbol{X}^T\boldsymbol{X}(\hat{\beta} - \beta)/(p+1)}{\hat{\sigma}^2} \sim F_{p+1,N-p-1}$$

1.2 What is Linear

In the context of **linear model**, we are talking about linearity in parameters, meaning that the prediction $\hat{y}$ is a linear combination of the parameters β_j. The $\boldsymbol{X}$ itself can be non-linear transformations of the original features, e.g., polynomial terms, interaction terms, etc. $y = 1/(\beta_0 + \beta_1 x)$ and $y = \beta_0 e^{\beta_1 x}$ are not linear models, since they're not linear in parameters.

In the context of **Linear estimators**, we are talking about the estimator $\hat{\theta}$ (e.g. $\hat{\beta}$) can be written as a linear combination of the observed response values y_i, i.e. $\hat{\theta} = \boldsymbol{c}^T\boldsymbol{y}$. The weight $\boldsymbol{c}$ depends only on $\boldsymbol{X}$, not on $\boldsymbol{y}$. A linear estimator **can be** a prediction at a new point, or the estimated coefficients $\hat{\beta}$ themselves.

1.3 Gauss-Markov Theorem

Why assume only know $\boldsymbol{X}$, but not $\boldsymbol{y}$?

> Note that though y_i as sample responses, are observable, the following statements and arguments including assumptions, proofs and the others assume under the only condition of knowing $\boldsymbol{X}_{i,j}$ but not y_i. — [?]

We have a *challenger* linear estimator $\tilde{\beta} = \boldsymbol{C}\boldsymbol{y}$, where $\boldsymbol{C} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T + \boldsymbol{D}$, a modification of OLS estimator. Ensure it's unbiased:

$$
\begin{aligned}
\mathbb{E}(\tilde{\beta}) &= \mathbb{E}(\boldsymbol{C}\boldsymbol{y}) \\
&= \boldsymbol{C}\mathbb{E}(\boldsymbol{y}) = ((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T + \boldsymbol{D})\boldsymbol{X}\beta \\
&= \beta + \boldsymbol{D}\boldsymbol{X}\beta \\
&= \beta
\end{aligned}
$$

Meaning $\boldsymbol{D}\boldsymbol{X} = 0$.

Now, compute the variance:

$$
\begin{aligned}
\mathrm{Var}(\tilde{\beta}) &= \mathrm{Var}(\boldsymbol{C}\boldsymbol{y}) \\
&= \boldsymbol{C}\mathrm{Var}(\boldsymbol{y})\boldsymbol{C}^T \\
&= \sigma^2\boldsymbol{C}\boldsymbol{C}^T \\
&= \sigma^2((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T + \boldsymbol{D})((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T + \boldsymbol{D})^T \\
&= \sigma^2((\boldsymbol{X}^T\boldsymbol{X})^{-1} + \boldsymbol{D}\boldsymbol{D}^T) \\
&= \mathrm{Var}(\hat{\beta}) + \sigma^2\boldsymbol{D}\boldsymbol{D}^T && \geq \mathrm{Var}(\hat{\beta})
\end{aligned}
$$

1.4 QR decomposition

References

[1] Wikipedia contributors, "Gauss–Markov theorem," *Wikipedia, The Free Encyclopedia*, `https://en.wikipedia.org/wiki/Gauss%E2%80%93Markov_theorem` (accessed Dec 29, 2025).