Element of Statistical Learning Note

Zichong Wang

December 30, 2025

Contents

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 $k = \text{rank}(\boldsymbol{\Sigma})$, 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

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

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

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

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. — [1]

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

Any real squared matrix $\boldsymbol{A}$ can be decomposed as $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$, where $\boldsymbol{Q}$ is orthogonal matrix ($\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}$), and $\boldsymbol{R}$ is upper triangular matrix.

Any rectangular matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ $(m \geq n)$, we can decompose it as $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$, where $\boldsymbol{Q}$ is $m \times m$ orthogonal matrix, and $\boldsymbol{R}$ is $m \times n$ upper triangular matrix (the last $m - n$ rows are all zero). It can be regarded as $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R} = [\boldsymbol{Q}_1 \quad \boldsymbol{Q}_2] \begin{bmatrix} \boldsymbol{R}_1 \\ \boldsymbol{0} \end{bmatrix} = \boldsymbol{Q}_1\boldsymbol{R}_1$, where $\boldsymbol{Q}_1 \in \mathbb{R}^{m\times n}, \boldsymbol{Q}_2 \in \mathbb{R}^{m\times(m-n)}, \boldsymbol{R}_1 \in \mathbb{R}^{n\times n}$ which is upper triangular.

If $\boldsymbol{A}$ have k linearly independent columns, then first k columns of $\boldsymbol{Q}$ form an orthonormal basis of the column space of $\boldsymbol{A}$. The fact that any column k of $\boldsymbol{A}$ only depends on the first k columns of $\boldsymbol{Q}$ corresponds to the triangular form of $\boldsymbol{R}$.

QR decomposition can be calculated using Gram-Schmidt process, or using Householder reflections. In practice, Householder reflections are more stable and efficient.

1.4.1 Application to Least Squared

In linear least squares problems, we aim to find a vector $\boldsymbol{x}$ that minimizes the Euclidean norm of the residual for an overdetermined system $\boldsymbol{A}\boldsymbol{x} \approx \boldsymbol{b}$, where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $m \geq n$. The goal is to solve:

$$\min_{\boldsymbol{x}} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2$$

Using the Full QR Decomposition, we substitute $\boldsymbol{A} = \boldsymbol{Q}\begin{bmatrix} \boldsymbol{R}_1 \\ \boldsymbol{0} \end{bmatrix}$:

$$\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2 = \left\| \boldsymbol{Q}\begin{bmatrix} \boldsymbol{R}_1 \\ \boldsymbol{0} \end{bmatrix}\boldsymbol{x} - \boldsymbol{b} \right\|_2$$

Since multiplying by an orthogonal matrix $\boldsymbol{Q}$ preserves the Euclidean norm, we can left-multiply the entire expression by $\boldsymbol{Q}^T$:

$$\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_2 = \left\| \begin{bmatrix} \boldsymbol{R}_1 \\ \boldsymbol{0} \end{bmatrix} \boldsymbol{x} - \boldsymbol{Q}^T\boldsymbol{b} \right\|_2$$

If we partition $\boldsymbol{Q}^T\boldsymbol{b}$ into two components—$\boldsymbol{c}_1 \in \mathbb{R}^n$ and $\boldsymbol{c}_2 \in \mathbb{R}^{m-n}$:

$$\left\| \begin{bmatrix} \boldsymbol{R}_1\boldsymbol{x} - \boldsymbol{c}_1 \\ -\boldsymbol{c}_2 \end{bmatrix} \right\|_2^2 = \|\boldsymbol{R}_1\boldsymbol{x} - \boldsymbol{c}_1\|_2^2 + \|\boldsymbol{c}_2\|_2^2$$

To minimize the total error, we must make the first term zero. The least squares solution is found by solving the square, upper-triangular system:

$$\boldsymbol{R}_1\boldsymbol{x} = \boldsymbol{c}_1$$

Since $\boldsymbol{R}_1$ is upper triangular, we can efficiently solve this system using back substitution, **that's how we solve linear equations manually in algebra class**! The remaining term $\|\boldsymbol{c}_2\|_2$ represents the minimum residual norm (the "error" of the fit).

1.4.2 About orthogonal matrix

Let $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ be an orthogonal matrix. Write $\boldsymbol{Q}$ in terms of its column vectors:

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{q}_1 & \boldsymbol{q}_2 & \cdots & \boldsymbol{q}_n \end{bmatrix}, \qquad \boldsymbol{q}_i^\top \boldsymbol{q}_j = \delta_{ij}.$$

Then

$$\boldsymbol{Q}^\top = \begin{bmatrix} \boldsymbol{q}_1^\top \\ \boldsymbol{q}_2^\top \\ \vdots \\ \boldsymbol{q}_n^\top \end{bmatrix}.$$

Computation of $\boldsymbol{Q}^\top\boldsymbol{Q}$.

$$\boldsymbol{Q}^\top\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{q}_1^\top \\ \boldsymbol{q}_2^\top \\ \vdots \\ \boldsymbol{q}_n^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{q}_1 & \boldsymbol{q}_2 & \cdots & \boldsymbol{q}_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{q}_1^\top\boldsymbol{q}_1 & \boldsymbol{q}_1^\top\boldsymbol{q}_2 & \cdots & \boldsymbol{q}_1^\top\boldsymbol{q}_n \\ \boldsymbol{q}_2^\top\boldsymbol{q}_1 & \boldsymbol{q}_2^\top\boldsymbol{q}_2 & \cdots & \boldsymbol{q}_2^\top\boldsymbol{q}_n \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{q}_n^\top\boldsymbol{q}_1 & \boldsymbol{q}_n^\top\boldsymbol{q}_2 & \cdots & \boldsymbol{q}_n^\top\boldsymbol{q}_n \end{bmatrix}.$$

Using $\boldsymbol{q}_i^\top \boldsymbol{q}_j = \delta_{ij}$,

$$\boldsymbol{Q}^\top\boldsymbol{Q} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = \boldsymbol{I}.$$

Computation of $\boldsymbol{Q}\boldsymbol{Q}^\top$.

$$\boldsymbol{Q}\boldsymbol{Q}^\top = \begin{bmatrix} \boldsymbol{q}_1 & \boldsymbol{q}_2 & \cdots & \boldsymbol{q}_n \end{bmatrix} \begin{bmatrix} \boldsymbol{q}_1^\top \\ \boldsymbol{q}_2^\top \\ \vdots \\ \boldsymbol{q}_n^\top \end{bmatrix} = \boldsymbol{q}_1\boldsymbol{q}_1^\top + \boldsymbol{q}_2\boldsymbol{q}_2^\top + \cdots + \boldsymbol{q}_n\boldsymbol{q}_n^\top = \sum_{k=1}^{n} \boldsymbol{q}_k\boldsymbol{q}_k^\top.$$

For any $\boldsymbol{x} \in \mathbb{R}^n$,

$$(\boldsymbol{Q}\boldsymbol{Q}^\top)\boldsymbol{x} = \sum_{k=1}^{n} \boldsymbol{q}_k(\boldsymbol{q}_k^\top \boldsymbol{x}) = \boldsymbol{x},$$

since $\{\boldsymbol{q}_1, \dots, \boldsymbol{q}_n\}$ is an orthonormal basis of $\mathbb{R}^n$. Hence

$$\boldsymbol{Q}\boldsymbol{Q}^\top = \boldsymbol{I}.$$

Therefore, for an orthogonal matrix $\boldsymbol{Q}$,

$$\boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{Q}^\top = \boldsymbol{I}.$$

1.5 Multiple testing in forward selection

ESL page 60:

> Other more traditional packages base the selection on F -statistics, adding "significant" terms, and dropping "non-significant" terms. These are out of fashion, since they do not take proper account of the multiple testing issues.

Assume we have p candidate features, and already selected k features. When considering adding a new feature, we are actually performing $p-k$ hypothesis tests (each test for one feature). Even the rest $p-k$ features are all noise, with significance level α, we still have a probability of $1-(1-\alpha)^{p-k}$ to incorrectly add at least one noise feature.

1.6 Ridge regression

Answer questions: why two forms are equivalent? Why not equivariant under scaling of the inputs? What is a good practice for it? df of ridge? In the case of orthogonal inputs, why $\hat{\beta}^{\text{ridge}} = \hat{\beta}/(1+\lambda)$?

Ridge regression shrinks the regression coefficients by imposing a penalty on their size:

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

and is equivalent to

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$
$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 \le t$$

And there is a one-to-one correspondence between λ and t.

1.6.1 Equivalence of two forms

First, take a review of **KKT conditions**.

Condider a minimization problem with both equality and inequality constraints:

$$\begin{aligned} &\min_{\boldsymbol{x}} f(\boldsymbol{x}) \\ \text{subject to } &g_i(\boldsymbol{x}) \leq 0, && i = 1, \ldots, m \\ &h_j(\boldsymbol{x}) = 0, && j = 1, \ldots, l \end{aligned}$$

And the lagrangian function is:

$$L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i g_i(\boldsymbol{x}) + \sum_{j=1}^{l} \mu_j h_j(\boldsymbol{x})$$

If $\boldsymbol{x}^*$ is a local minimum, then there exist multipliers $\lambda_i^* \geq 0$ and μ_j^* such that the following conditions hold:

- **Stationary**
  $$\nabla_x L(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = 0$$
- **Primal feasibility**
  $$g_i(\boldsymbol{x}^*) \leq 0, \quad i = 1, \ldots, m$$
  $$h_j(\boldsymbol{x}^*) = 0, \quad j = 1, \ldots, l$$
  The gradient can be regarded as the force to push a particle, primal stationary means the force of $\partial f(\boldsymbol{x}^*)$ is balanced by a linear sum of forces from constraints.
- **Dual feasibility**
  $$\lambda_i^* \geq 0, \quad i = 1, \ldots, m$$
  All the $\partial g_i(\boldsymbol{x}^*)$ forces must be one-sided, pointing inwards into the feasible set for $\boldsymbol{x}$.
- **Complementary slackness**
  $$\lambda_i^* g_i(\boldsymbol{x}^*) = 0, \quad i = 1, \ldots, m$$
  The force only activated when the particle is on the boundary of feasible set.

In ridge regression, we have:

$$\mathcal{L} = \|\boldsymbol{Y} - \boldsymbol{X}\beta\|_2^2 + \alpha(\|\beta\|_2^2 - t)$$

According to stationary condition:

$$\nabla_\beta \mathcal{L} = -2\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\beta) + 2\alpha\beta = 0$$

On the other hand, solving the unconstrained form is

$$\nabla_\beta \left(\|\boldsymbol{Y} - \boldsymbol{X}\beta\|_2^2 + \lambda\|\beta\|_2^2\right) = -2\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\beta) + 2\lambda\beta = 0$$

Thus, if we set $\lambda = \alpha$, the two forms are equivalent.

One step further, according to complementary slackness:

$$\alpha(\|\beta\|_2^2 - t) = 0$$

From unconstrained form, given λ, we can solve β, denote $\beta(\lambda)$, and define $t(\lambda) = \|\beta(\lambda)\|_2^2$. Then, $\beta(\lambda)$ is also the solution of constrained form with $t = t(\lambda)$ (apparently it's on the boundary, and the coefficient $\alpha = \lambda$).

From the constrained form, given t, we can also solve β, denote $\beta(t)$.

- If $\|\beta(t)\|_2^2 < t$, then according to complementary slackness, $\alpha = 0$, which means no penalty, $\lambda = 0$, back to OLS.
- If $\|\beta(t)\|_2^2 = t$, the boundary is effective, correspond to some $\alpha > 0$, and $\lambda = \alpha$.

1.7 Bayesian view

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).