Linear Regression Q&A

Advanced Theoretical Analysis of Linear Regression

We consider the classical linear model: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

where:

- $\mathbf{X} \in \mathbb{R}^{n\mathrm{x}p}$
- $\boldsymbol{\beta} \in \mathbb{R}^{p}$
- $\boldsymbol{\varepsilon} \sim (0, \sigma^2\mathbf{I})$

Assume **X** has full column rank unless stated otherwise

1. Prove that OLS estimator is unbiased.

OLS estimator: $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

Substitute $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})\\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}\\ &= \boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}\end{aligned}$$

Take expectation: $\mathbf{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{E}[\boldsymbol{\varepsilon}]$

Since $\mathbf{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$:

$\mathbf{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$

✔ OLS is unbiased.

2. Derive covariance of OLS estimator.

From: $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}$

Variance: $\mathbf{Var}(\hat{\boldsymbol{\beta}}) = \mathbf{Var}((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon})$
$= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\ \mathbf{Var}(\boldsymbol{\varepsilon})\ \mathbf{X}\ (\mathbf{X}^T\mathbf{X})^{-1}$

Since $\mathbf{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$:

$\mathbf{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2\ (\mathbf{X}^T\mathbf{X})^{-1}$

✔ Covariance depends on $(X^T X)^{-1}$

3. When is OLS BLUE? (Gauss-Markov Theorem)

OLS is BLUE (Best Linear Unbiased Estimator) if:

- Model is linear in parameters
- $E[\varepsilon] = 0$
- $Var(\varepsilon) = \sigma^2 I$
- No perfect multicollinearity

Then among all linear unbiased estimators, OLS has minimum variance.

✔ No normality assumption required.

4. Why does multicollinearity increase variance?

Variance formula: $Var(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$

If features are highly correlated:

- $X^T X$ becomes nearly singular
- Its inverse contains very large values
- Variance becomes very large

✔ Multicollinearity → unstable coefficient estimates

5. Show ridge regression shrinks eigenvalues.

Ridge estimator: $\hat{\beta}_{ridge} = (X^T X + \lambda I)^{-1} X^T y$

Eigen-decomposition: $X^T X = Q \Lambda Q^T$

Then: $X^T X + \lambda I = Q(\Lambda + \lambda I) Q^T$

Thus: $(X^T X + \lambda I)^{-1} = Q(\Lambda + \lambda I)^{-1} Q^T$

Each eigenvalue λ_i becomes: $1 / (\lambda_i + \lambda)$

✔ Ridge increases eigenvalues
✔ Reduces variance
✔ Stabilizes matrix inversion

6. Compare ridge vs lasso geometrically.

Ridge: min $\|y - X\beta\|^2 + \lambda\|\beta\|^2$

Lasso: min $\|y - X\beta\|^2 + \lambda\|\beta\|_1$

Geometrically:

- Ridge constraint region = circle (L2 ball)
- Lasso constraint region = diamond (L1 ball)

Because the diamond has sharp corners, the solution often occurs on axes → exact zeros.

✔ Lasso performs feature selection
✔ Ridge shrinks coefficients but rarely sets them to zero

7. Derive linear regression as MAP estimate.

Assume:

Likelihood: $\mathbf{y \mid X, \beta \sim N(X\beta, \sigma^2 I)}$

Prior (Gaussian): $\mathbf{\beta \sim N(0, \tau^2 I)}$

Posterior maximization: $\log p(\beta \mid y) \propto -\|y - X\beta\|^2 - \lambda\|\beta\|^2$

This equals the Ridge objective.

✔ OLS = MLE
✔ Ridge = MAP with Gaussian prior
✔ Lasso = MAP with Laplace prior

8. Analyze condition number of $X^T X$.

Condition number: $\mathbf{\kappa(X^T X) = \lambda_{\square ax} / \lambda_{\square i \square}}$

If smallest eigenvalue ≈ 0:

$\mathbf{\kappa} \to \infty$

Large condition number ⇒ numerical instability.

Multicollinearity → small eigenvalues → large condition number.

✔ Ridge improves conditioning by adding λ.

9. Prove convexity of squared loss.

Loss: $\mathbf{L(\beta) = \|y - X\beta\|^2}$

Gradient: $\nabla \mathbf{L} = \mathbf{-2X^T(y - X\beta)}$

Hessian: $\nabla^2 \mathbf{L} = \mathbf{2X^TX}$

Since $\mathbf{X^TX}$ is positive semidefinite: $\nabla^2 \mathbf{L} \succeq \mathbf{0}$

✔ Squared loss is convex
✔ If full rank → strictly convex → unique global minimum

10. What happens when p >> n?

When number of features exceeds number of samples:

- $\mathbf{X^TX}$ is singular
- Infinite OLS solutions
- Model interpolates training data
- Severe overfitting
- High variance

Solutions:

- Ridge regression
- Lasso
- Dimensionality reduction

Modern Interpretation:

This setting is common in high-dimensional statistics and deep learning.