1 Prove OLS Estimator is Unbiased

Substitute model into estimator:

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

$$= (X^T X)^{-1} X^T X\beta + (X^T X)^{-1} X^T \varepsilon$$

$$= \beta + (X^T X)^{-1} X^T \varepsilon$$

Take expectation:

$$\mathbb{E}[\hat{\beta}] = \beta + (X^T X)^{-1} X^T \mathbb{E}[\varepsilon]$$

Since $\mathbb{E}[\varepsilon] = 0$,

$$\mathbb{E}[\hat{\beta}] = \beta$$

✓ **OLS is unbiased**

2 Derive Covariance of OLS Estimator

$$\hat{\beta} - \beta = (X^T X)^{-1} X^T \varepsilon$$

Covariance:

$$\mathrm{Var}(\hat{\beta}) = (X^T X)^{-1} X^T \mathrm{Var}(\varepsilon) X (X^T X)^{-1}$$

Since $\mathrm{Var}(\varepsilon) = \sigma^2 I$,

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

3 When is OLS BLUE? (Gauss–Markov Theorem)

OLS is **BLUE (Best Linear Unbiased Estimator)** if:

1. Linear model is correct
2. E[ε]=0\mathbb{E}[\varepsilon] = 0E[ε]=0
3. Var(ε)=σ2I\text{Var}(\varepsilon) = \sigma^2 IVar(ε)=σ2I (homoscedasticity)
4. No perfect multicollinearity
5. Errors uncorrelated

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

If heteroscedasticity exists → OLS not efficient.

4 Why Multicollinearity Increases Variance

From covariance:

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

If predictors are highly correlated:

- $X^T X$ becomes nearly singular
- Small eigenvalues appear
- Inverse blows up

Since:

$$(X^T X)^{-1} = Q \Lambda^{-1} Q^T$$

Small eigenvalue $\lambda_i \Rightarrow 1/\lambda_i$ very large

Thus variance explodes.

5 Show Ridge Regression Shrinks Eigenvalues

Ridge estimator:

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

Eigen-decompose:

$$X^T X = Q \Lambda Q^T$$

Then:

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

Inverse:

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

Each eigenvalue transforms:

$$\lambda_i \to \frac{1}{\lambda_i + \lambda}$$

Since $\lambda > 0$,

Denominator increases ⇒ coefficients shrink.

6 Ridge vs Lasso (Geometric View)

OLS:

- Elliptical loss contours
- No constraint

Ridge:

- Constraint: ||β||22≤t||\beta||_2^2 \le t||β||22≤t
- Circular constraint region
- Shrinks continuously
- Rarely exact zero

Lasso:

- Constraint: ||β||1≤t||\beta||_1 \le t||β||1≤t
- Diamond-shaped region
- Corners → sparsity
- Produces exact zeros

Thus:

- Ridge = shrinkage
- Lasso = feature selection

7 Linear Regression as MAP Estimate

Assume:

Likelihood:

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

Prior:

$$\beta \sim \mathcal{N}(0, \tau^2 I)$$

Posterior ∝ Likelihood × Prior

Taking negative log:

$$\frac{1}{2\sigma^2}||y - X\beta||^2 + \frac{1}{2\tau^2}||\beta||^2$$

Multiply by constant:

$$||y - X\beta||^2 + \lambda||\beta||^2$$

Where:

$$\lambda = \sigma^2/\tau^2$$

This is Ridge regression.

Thus:

- OLS = MLE
- Ridge = MAP with Gaussian prior

8 Condition Number of XTXX^T XXTX

Condition number:

$$\kappa(X^T X) = \frac{\lambda_{max}}{\lambda_{min}}$$

If $\lambda_{min} \to 0$:

- Ill-conditioned
- Numerical instability
- High variance

Multicollinearity ⇒ small eigenvalues ⇒ large condition number.

Ridge improves condition number:

$$\kappa_{ridge} = \frac{\lambda_{max} + \lambda}{\lambda_{min} + \lambda}$$

9 Prove Convexity of Squared Loss

Loss:

$$J(\beta) = ||y - X\beta||^2$$

Expand:

$$= \beta^T X^T X\beta - 2y^T X\beta + y^T y$$

Hessian:

$$\nabla^2 J(\beta) = 2X^T X$$

Since $X^T X$ is positive semidefinite:

$$z^T X^T X z = ||Xz||^2 \geq 0$$

Therefore Hessian PSD ⇒ convex.

If full rank ⇒ strictly convex.

🔟 What Happens When p>>np >> np>>n?

Then:

- XTXX^T XXTX is singular
- Infinite OLS solutions
- Overfitting guaranteed
- Interpolates data perfectly

Geometrically:

- Many hyperplanes fit data

Variance:

- Extremely high

Solution:

- Ridge
- Lasso
- Dimensionality reduction

Modern ML (deep learning) often works in p>>np >> np>>n, but classical OLS theory fails.