Linear Regression-Q&A

Ordinary Least Squares (OLS)

Ordinary Least Squares (OLS) is a fundamental statistical technique used to estimate the relationship between one or more independent variables (predictors) and a dependent variable (outcome).it is one of the most broadly used methods for linear regression analysis.

Goal of OLS

- Minimize the sum of squared differences between the observed values and the predicted values.
- Find the best-fitting line (or hyperplane for multiple variables) by reducing the residuals.

Consider the linear regression model:

$$y = X\beta + \varepsilon$$

where
$y \in \mathbb{R}^n$is the response vector,
$X \in \mathbb{R}^{n\times p}$is the design matrix,
$\beta \in \mathbb{R}^p$is the parameter vector,
$\varepsilon \in \mathbb{R}^n$is the error vector.
The Ordinary Least Squares (OLS) estimator minimizes the squared error:

$$\widehat{\beta} = \arg\min_{\beta} \parallel y - X\beta \parallel^2$$

Expanding the loss function:

$$L(\beta) = (y - X\beta)^T(y - X\beta)$$

Taking derivative with respect to β:

$$\nabla_{\beta} L(\beta) = -2X^T(y - X\beta)$$

Setting derivative equal to zero:

$$X^T X\beta = X^T y$$

These are called the normal equations.
If $X^T X$is invertible:

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

This is the OLS estimator.

1. Prove that OLS Estimator is Unbiased

Assume that,

1. $y = X\beta + \varepsilon$
2. $E[\varepsilon] = 0$
3. $X$is fixed

The OLS estimator is:

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

Substitute $y = X\beta + \varepsilon$:

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

Taking expectation:

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

Since $E[\varepsilon] = 0$:

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

Hence, OLS estimator is unbiased.

3. Derive Covariance of OLS Estimator

We have,

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

Variance:

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

Using property:

$$Var(A\varepsilon) = A Var(\varepsilon) A^T$$

Assume:

$$Var(\varepsilon) = \sigma^2 I$$

Then:

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

4. When is OLS BLUE? (Gauss–Markov Theorem)

OLS is BLUE (Best Linear Unbiased Estimator) if:

1. Linear model: $y = X\beta + \varepsilon$
2. $E[\varepsilon] = 0$
3. $Var(\varepsilon) = \sigma^2 I$(Homoscedasticity)
4. No perfect multicollinearity (X has full column rank).

Then among all linear unbiased estimators:

$$Var(\hat{\beta}_{OLS}) \le Var(\tilde{\beta})$$

for any other linear unbiased estimator $\tilde{\beta}$.
Note: Normality of errors is NOT required.

5. Why Multicollinearity Increases Variance

Variance of OLS:

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

If predictors are highly correlated:

- Columns of X are nearly linearly dependent
- $X^T X$becomes nearly singular
- Small eigenvalues appear

Since:

$$(X^T X)^{-1}$$

contains reciprocals of eigenvalues, small eigenvalues lead to very large variances.
Thus, multicollinearity increases estimator variance.

6. Show Ridge Regression Shrinks Eigenvalues

Ridge estimator:

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

Eigen-decompose:

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

Then:

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

Inverse:

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

Eigenvalues become:

$$\frac{1}{\lambda_j + \lambda}$$

Since $\lambda > 0$:

$$\frac{1}{\lambda_j + \lambda} < \frac{1}{\lambda_j}$$

Thus ridge shrinks eigenvalues and reduces variance.

7. Derive Linear Regression as MAP Estimate

Assume:
Likelihood:

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

Prior:

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

Posterior ∝ Likelihood × Prior.
Log-posterior:

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

Maximizing posterior is equivalent to minimizing:

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

where $\lambda = \frac{\sigma^2}{\tau^2}$.
This is ridge regression.

8. Analyze the Condition Number of $X^T X$

The condition number of a symmetric positive definite matrix $A$is defined as:

$$\kappa(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$$

where
$\lambda_{\max}(A)$= largest eigenvalue
$\lambda_{\min}(A)$= smallest eigenvalue

For the matrix $X^T X$:

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

Relation with Singular Values of *X*

Let the singular values of Xbe:

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p > 0$$

Then eigenvalues of $X^T X$are:

$$\lambda_i = \sigma_i^2$$

Therefore:

$$\kappa(X^T X) = \frac{\sigma_1^2}{\sigma_p^2}$$

Since:

$$\kappa(X) = \frac{\sigma_1}{\sigma_p}$$

we get:

$$\boxed{\kappa(X^T X) = \kappa(X)^2}$$

forming $X^T X$ squares the condition number and makes instability worse.

Why Condition Number Matters

OLS estimator:

$$\widehat{\beta} = (X^T X)^{-1} X^T y$$

If $\lambda_{\min}(X^T X)$is very small:

- $(X^T X)^{-1}$contains very large values
- Small changes in data produce large changes in $\widehat{\beta}$
- The system becomes numerically unstable

Thus:

- Small κ→ stable solution

- Large κ→ unstable solution

Interpretation

1. If predictors are nearly linearly dependent,
$$\lambda_{\min} \approx 0$$
2. Then:
$$\kappa(X^T X) \to \infty$$
3. This implies severe multicollinearity.

Practical Thresholds

There is no strict rule, but commonly:

- $\kappa < 10$→ well-conditioned
- $10 < \kappa < 100$→ moderate multicollinearity
- $\kappa > 1000$→ severe instability

Effect on Variance

Recall:

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

If $\lambda_{\min}$is small:

$$\frac{1}{\lambda_{\min}} \text{ becomes very large}$$

So the variance of OLS explodes in directions corresponding to small eigenvalues.

How Ridge Fixes This

Ridge replaces: $X^T X$ with: $X^T X + \lambda I$

New eigenvalues:

$$\lambda_i + \lambda$$

So smallest eigenvalue becomes:

$$\lambda_{\min} + \lambda$$

Thus condition number becomes:

$$\frac{\lambda_{\max} + \lambda}{\lambda_{\min} + \lambda}$$

This reduces instability.

Final Conclusion

The condition number of $X^T X$:

- Measures numerical stability
- Is directly tied to multicollinearity
- Controls variance inflation
- Explains why OLS becomes unstable
- Is squared relative to the condition number of X

$$\boxed{\kappa(X^T X) = \kappa(X)^2}$$

That squared relationship is why forming normal equations is numerically dangerous in ill-conditioned problems.

9. Prove Convexity of Squared Loss

Loss function:

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

Hessian:

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

Since:

$$v^T X^T X v = \| Xv \|^2 \geq 0$$

$X^T X$ is positive semi-definite.
Therefore squared loss is convex.
If X has full rank, it is strictly convex.

10. What Happens When p >> n?

If number of predictors exceeds observations:

- Rank(X) ≤ n
- $X^T X$ is singular
- Inverse does not exist

Therefore OLS has infinitely many solutions.
Regularization (Ridge/Lasso) is required to obtain a unique solution.