Linear Regression Summary Sheet

What is Linear Regression?

Linear regression models the relationship between a dependent variable y and one or more independent variables X using a linear approach.

Simple Linear Regression Equation

$$y = \beta_0 + \beta_1 x + \epsilon$$

- y: dependent variable
- x: independent variable
- β_0: intercept
- β_1: slope
- ϵ: error term

Deriving Parameters in Simple Linear Regression

We want to find the values of β_0 and β_1 that minimize the sum of squared errors:

$$SSE = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

This is a convex optimization problem. We solve it by taking partial derivatives with respect to β_0 and β_1, and setting them to zero:

Step 1: Partial Derivatives

$$\frac{\partial SSE}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial SSE}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0$$

Step 2: Solve the System

From the first equation:

$$\sum y_i = n\beta_0 + \beta_1 \sum x_i$$

From the second equation:

$$\sum x_i y_i = \beta_0 \sum x_i + \beta_1 \sum x_i^2$$

Solving for β_1:

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Then:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Deriving Parameters in Multiple Linear Regression (Matrix Form)

In matrix form, the model is:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Where:

- $\mathbf{y}$: $n \times 1$ vector of responses
- $\mathbf{X}$: $n \times p$ matrix of features (with a column of 1's for the intercept)
- $\boldsymbol{\beta}$: $p \times 1$ vector of coefficients
- $\boldsymbol{\epsilon}$: $n \times 1$ vector of residuals

The loss function is the sum of squared errors:

$$L(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Taking the Gradient

We take the derivative with respect to $\boldsymbol{\beta}$:

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Set this equal to 0:

$$\mathbf{X}^\top \mathbf{y} = \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}$$

Solving for $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

This is known as the Normal Equation. It gives the best linear unbiased estimate under the Gauss-Markov assumptions.

Error Metrics and Decomposition

Total Sum of Squares (TSS)

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Explained Sum of Squares (ESS)

$$ESS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Residual Sum of Squares (RSS)

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

R-Squared

$$R^2 = 1 - \frac{RSS}{TSS} = \frac{ESS}{TSS}$$

Confidence Intervals for Predictions

Point Prediction

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

Confidence Interval for Mean Prediction

$$\hat{y}_0 \pm t^* \cdot SE(\hat{y}_0)$$

Prediction Interval (New Observation)

$$\hat{y}_0 \pm t^* \cdot \sqrt{SE(\hat{y}_0)^2 + \sigma^2}$$

Standard Error of the Fit

$$SE(\hat{y}_0) = \sqrt{s^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right)} \quad \text{where } s^2 = \frac{RSS}{n-2}$$

Confidence Intervals for Coefficients

$$\hat{\beta}_j \pm t^* \cdot SE(\hat{\beta}_j) \quad \text{with } SE(\hat{\beta}_j) \text{ from diag of } \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$$

t-Tests for Coefficients

$$t_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \quad \text{used to test } H_0 : \beta_j = 0$$

Assumptions of Linear Regression

1. Linearity
2. Independence of errors
3. Homoscedasticity (constant variance of errors)
4. Normality of residuals
5. No multicollinearity (for multiple regression)

Model Selection

- Stepwise selection (forward/backward)
- Cross-validation
- Information criteria: AIC, BIC

Regularization

Regularization techniques help combat overfitting by penalizing large coefficients in linear regression models.

Ridge Regression

$$\hat{\boldsymbol{\beta}}^{ridge} = \arg\min_{\beta} \left\{ \sum (y_i - \hat{y}_i)^2 + \lambda \sum \beta_j^2 \right\}$$

Ridge Regression (L2 penalty):
Adds a penalty proportional to the square of the coefficients. It shrinks all coefficients but never reduces them exactly to zero. Ideal when all predictors are believed to contribute to the response, even if weakly.

Lasso Regression

$$\hat{\boldsymbol{\beta}}^{lasso} = \arg\min_{\beta} \left\{ \sum (y_i - \hat{y}_i)^2 + \lambda \sum |\beta_j| \right\}$$

Lasso Regression (L1 penalty):
Adds a penalty proportional to the absolute value of the coefficients. It can force some coefficients to be exactly zero, effectively performing feature selection. Great when you suspect many predictors are irrelevant.

Elastic Net:
A combination of Ridge and Lasso. Useful when you have high-dimensional data (many predictors, possibly correlated) and want both shrinkage and sparsity.

Conclusion

Linear regression is your modeling "starter car" — simple, classic, and eventually something you trade in for better tools. Learn the math, the limits, and when to move on.