Derivation of Ridge Regression Coefficients

Step 1: Ridge Regression Objective Function

Start with the ridge regression objective function:

$$\min_{\beta} \left\{ (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta) + \lambda \beta^\top \beta \right\}$$

where:

- $\mathbf{y}$ is an $n \times 1$ vector of observed values,
- $\mathbf{X}$ is an $n \times p$ matrix of predictors,
- β is a $p \times 1$ vector of coefficients,
- λ is the regularization parameter.

Step 2: Expand the Objective Function

Expand and simplify the ridge regression objective function:

$$(\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta) + \lambda \beta^\top \beta$$

$$= \mathbf{y}^\top \mathbf{y} - 2\beta^\top \mathbf{X}^\top \mathbf{y} + \beta^\top \mathbf{X}^\top \mathbf{X}\beta + \lambda \beta^\top \beta$$

$$= \mathbf{y}^\top \mathbf{y} - 2\mathbf{y}^\top \mathbf{X}\beta + \beta^\top \mathbf{X}^\top \mathbf{X}\beta + \lambda \beta^\top \beta$$

Step 3: Take the Derivative

To find the ridge regression coefficients, take the derivative of the objective function with respect to β and set it to zero:

$$\frac{\partial}{\partial \beta} \left\{ \mathbf{y}^\top \mathbf{y} - 2\mathbf{y}^\top \mathbf{X}\beta + \beta^\top \mathbf{X}^\top \mathbf{X}\beta + \lambda \beta^\top \beta \right\} = 0$$

$$-2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\beta + 2\lambda\beta = 0$$

Step 4: Solve for β

Now, solve the equation for β:

$$\mathbf{X}^\top \mathbf{X}\beta + \lambda\beta = \mathbf{X}^\top \mathbf{y}$$

$$(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})\beta = \mathbf{X}^\top \mathbf{y}$$

$$\beta = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}$$

where $\mathbf{I}$ is the $p \times p$ identity matrix.

Interpretation

The formula $\beta = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}$ gives us the coefficients for ridge regression.

This equation adjusts the coefficients β to minimize the sum of squared residuals while penalizing them by the regularization parameter λ. This penalty helps prevent overfitting by shrinking the coefficients, particularly useful when predictors are correlated (multicollinearity).