

Elements of Statistical Learning

Chapter 3, Exercise 29

Berit Omli Øksnes

February 2026

Suggested solution

Let $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ denote the response vector and $\mathbf{x} = (x_1, \dots, x_n)^T$ a vector of observed values of an explanatory variable X . We use $\mathbf{X}$ to denote a design/model matrix.

In Ridge regression, we seek to minimize the objective function

$$\begin{aligned} g(\boldsymbol{\beta}) &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^T \boldsymbol{\beta} \\ &= \mathbf{Y}^T \mathbf{Y} - \mathbf{Y}^T \mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{Y} + \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{X} \boldsymbol{\beta} + \lambda \boldsymbol{\beta}^T \boldsymbol{\beta}, \end{aligned}$$

with respect to $\boldsymbol{\beta} \in \mathbb{R}^{p \times 1}$. We get that

$$\begin{aligned} \frac{\partial g(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= 0 \\ \implies -2\mathbf{X}^T \mathbf{Y} + 2\mathbf{X}^T \mathbf{X} \boldsymbol{\beta} + 2\lambda \boldsymbol{\beta} &= 0, \end{aligned}$$

meaning that $\hat{\boldsymbol{\beta}}^{\text{Ridge}}$ must satisfy the equation

$$(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_p) \boldsymbol{\beta} = \mathbf{X}^T \mathbf{Y}. \quad (1)$$

Note that (1) is equivalent to a system of p equations, where p is the number of coefficients. Solving (1) with respect to $\boldsymbol{\beta}$ yields

$$\hat{\boldsymbol{\beta}}^{\text{Ridge}} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^T \mathbf{Y}.$$

First assume that $p = 1$. That is, we only include the explanatory variable X in the model (and, for simplicity, no intercept). Thus, we can write the design matrix as

$$\mathbf{X} = \mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^{n \times 1}.$$

This means that $\mathbf{X}^T \mathbf{X} = \mathbf{x}^T \mathbf{x} = \sum_{i=1}^n x_i^2$ and $\mathbf{X}^T \mathbf{Y} = \mathbf{x}^T \mathbf{Y} = \sum_{i=1}^n x_i Y_i$, and the Ridge estimator becomes

$$\hat{\beta}^{\text{Ridge}} = \frac{\mathbf{x}^T \mathbf{Y}}{\mathbf{x}^T \mathbf{x} + \lambda} = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2 + \lambda}.$$

Next, assume that we include another variable in the model, $X^* = X$. We show that the estimated coefficients become the same. The new design matrix is given by

$$\mathbf{X} = (\mathbf{x} \quad \mathbf{x}^*) = (\mathbf{x} \quad \mathbf{x}),$$

such that

$$\mathbf{X}^T \mathbf{X} = \begin{pmatrix} \mathbf{x}^T \\ \mathbf{x}^T \end{pmatrix} (\mathbf{x} \quad \mathbf{x}) = \begin{pmatrix} \mathbf{x}^T \mathbf{x} & \mathbf{x}^T \mathbf{x} \\ \mathbf{x}^T \mathbf{x} & \mathbf{x}^T \mathbf{x} \end{pmatrix},$$

and

$$\mathbf{X}^T \mathbf{Y} = \begin{pmatrix} \mathbf{x}^T \\ \mathbf{x}^T \end{pmatrix} \mathbf{Y} = \begin{pmatrix} \mathbf{x}^T \mathbf{Y} \\ \mathbf{x}^T \mathbf{Y} \end{pmatrix}.$$

Inserting this into (1) gives as system of 2 equations for which β_1 and β_2 must satisfy:

$$\begin{aligned} \begin{pmatrix} \mathbf{x}^T \mathbf{x} + \lambda & \mathbf{x}^T \mathbf{x} \\ \mathbf{x}^T \mathbf{x} & \mathbf{x}^T \mathbf{x} + \lambda \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} &= \begin{pmatrix} \mathbf{x}^T \mathbf{Y} \\ \mathbf{x}^T \mathbf{Y} \end{pmatrix} \\ \Rightarrow \begin{pmatrix} (\mathbf{x}^T \mathbf{x} + \lambda)\beta_1 + \mathbf{x}^T \mathbf{x}\beta_2 \\ \mathbf{x}^T \mathbf{x}\beta_1 + (\mathbf{x}^T \mathbf{x} + \lambda)\beta_2 \end{pmatrix} &= \begin{pmatrix} \mathbf{x}^T \mathbf{Y} \\ \mathbf{x}^T \mathbf{Y} \end{pmatrix}. \end{aligned}$$

If we subtract the second equation from the first, we get that

$$\begin{aligned} (\mathbf{x}^T \mathbf{x} + \lambda)\beta_1 + \mathbf{x}^T \mathbf{x}\beta_2 - \mathbf{x}^T \mathbf{x}\beta_1 - (\mathbf{x}^T \mathbf{x} + \lambda)\beta_2 &= 0 \\ \Rightarrow \mathbf{x}^T \mathbf{x}\beta_1 + \lambda\beta_1 - \mathbf{x}^T \mathbf{x}\beta_1 &= \mathbf{x}^T \mathbf{x}\beta_2 + \lambda\beta_2 - \mathbf{x}^T \mathbf{x}\beta_2 \\ \Rightarrow \lambda\beta_1 &= \lambda\beta_2 \\ \Rightarrow \beta_1 &= \beta_2. \end{aligned}$$

This means that the estimated coefficients $\hat{\beta}_1^{\text{Ridge}}, \hat{\beta}_2^{\text{Ridge}}$ must be equal, which was what we wanted to show. We also derive an expression for the estimators. From the first equation, we get

$$\begin{aligned} (\mathbf{x}^T \mathbf{x} + \lambda)\beta_1 + \mathbf{x}^T \mathbf{x}\beta_2 &= \mathbf{x}^T \mathbf{Y} \\ \Rightarrow \mathbf{x}^T \mathbf{x}\beta_1 + \lambda\beta_1 + \mathbf{x}^T \mathbf{x}\beta_1 &= \mathbf{x}^T \mathbf{Y} \\ \Rightarrow (2\mathbf{x}^T \mathbf{x} + \lambda)\beta_1 &= \mathbf{x}^T \mathbf{Y} \\ \Rightarrow \beta &= \frac{\mathbf{x}^T \mathbf{Y}}{2\mathbf{x}^T \mathbf{x} + \lambda}, \end{aligned}$$

showing that $\hat{\beta}_1^{\text{Ridge}} = \hat{\beta}_2^{\text{Ridge}} = \frac{\mathbf{x}^T \mathbf{Y}}{2\mathbf{x}^T \mathbf{x} + \lambda}$.

For the general case with p explanatory variables, all being equal, the equations take the form

$$\begin{pmatrix} \mathbf{x}^T \mathbf{x} + \lambda & \mathbf{x}^T \mathbf{x} & \dots & \mathbf{x}^T \mathbf{x} \\ \mathbf{x}^T \mathbf{x} & \mathbf{x}^T \mathbf{x} + \lambda & \dots & \mathbf{x}^T \mathbf{x} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}^T \mathbf{x} & \mathbf{x}^T \mathbf{x} & \dots & \mathbf{x}^T \mathbf{x} + \lambda \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} = \begin{pmatrix} \mathbf{x}^T \mathbf{Y} \\ \mathbf{x}^T \mathbf{Y} \\ \vdots \\ \mathbf{x}^T \mathbf{Y} \end{pmatrix}.$$

For any i, j , subtract equation j from equation i to conclude that $\hat{\beta}_i^{\text{Ridge}} = \hat{\beta}_j^{\text{Ridge}}$, and use the first equation to derive the expression:

$$\begin{aligned} \mathbf{x}^T \mathbf{x}\beta_1 + \lambda\beta_1 + \mathbf{x}^T \mathbf{x}\beta_1 + \dots + \mathbf{x}^T \mathbf{x}\beta_1 &= \mathbf{x}^T \mathbf{Y} \\ \Rightarrow \hat{\beta}_1^{\text{Ridge}} &= \frac{\mathbf{x}^T \mathbf{Y}}{p\mathbf{x}^T \mathbf{x} + \lambda}. \end{aligned}$$