Ridge regression
Exercises

1 CASI

6. Verify (7.39).
7. (a) How were the columns sd(0) and sd(0.1) calculated in Table 7.3?
   (b) Calculate $\hat{\beta}(0.2)$ and sd(0.2).
8. Derive (7.43).
9. Carry out the differentiation following (7.41) to derive (7.36).

2 CASL

3.7 Exercises

1. Being careful about the handling of scale, intercepts, and penalties, verify that our function `casl_lm_ridge` produces similar results to `MASS::lm.ridge` and `glmnet::glmnet` for specific values of λ.
2. Add functionality to `casl_lm_ridge` to pick a reasonable sequence of values λ when none is supplied. Include an option `nlam`, set to 100 by default, to set the number of lambda values to be created. Reasonable values can be inferred from the range of the singular values of X as shown in Equation 3.29. Note that it makes sense to select lambda values on the log scale.
3. Now, add additional input parameters to `ridge_reg` for the validation data: `X_valid` and `y_valid`. If supplied, return only the best value of β.
4. Construct a function `cv.ridge_reg` that performs 10-fold cross-validation to select the optimal value of λ for ridge regression.
5. Modify `ridge_reg` to include an option `scale` that, when set to `TRUE`, centers and scales the columns of X before running the regression. Make sure to return the result in the original scale.

10. There is a well-known theoretical result showing that there must exist a positive λ such that the training mean squared error of ridge regression dominates that of the ordinary least squares fit. Design a simulation to test this claim empirically and describe the results.

3 Lecture notes (van Wieringen, 2015)

1.12 Exercises

Question 1.1 [†]
Consider the simple linear regression model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. The data on the covariate and response are: $\mathbf{X}^\top = (X_1, X_2, \ldots, X_8)^\top = (-2, -1, -1, -1, 0, 1, 2, 2)^\top$ and $\mathbf{Y}^\top = (Y_1, Y_2, \ldots, Y_8)^\top = (35, 40, 36, 38, 40, 43, 45, 43)^\top$, with corresponding elements in the same order.

a) Find the ridge regression estimator for the data above for a general value of λ.

b) Evaluate the fit, i.e. $\widehat{Y}_i(\lambda)$ for $\lambda = 10$. Would you judge the fit as good? If not, what is the most striking feature that you find unsatisfactory?

c) Now zero center the covariate and response data, denote it by $\tilde{X}_i$ and $\tilde{Y}_i$, and evaluate the ridge estimator of $\tilde{Y}_i = \beta_1 \tilde{X}_i + \varepsilon_i$ at $\lambda = 4$. Verify that in terms of original data the resulting predictor now is: $\widehat{Y}_i(\lambda) = 40 + 1.75X$.

Note that the employed estimate in the predictor found in part *c)* is effectively a combination of a maximum likelihood and ridge regression one for intercept and slope, respectively. Put differently, only the slope has been regularized/penalized.

[†]This exercise is inspired by one from Draper and Smith (1998)

Question 1.2
Consider the simple linear regression model $Y_i = \beta_0 + X_i\beta + \varepsilon_i$ for $i = 1, \ldots, n$ and with $\varepsilon_i \sim_{i.i.d.} \mathcal{N}(0, \sigma^2)$. The model comprises a single covariate and an intercept. Response and covariate data are: $\{(y_i, x_i)\}_{i=1}^4 = \{(1.4, 0.0), (1.4, -2.0), (0.8, 0.0), (0.4, 2.0)\}$. Find the value of λ that yields the ridge regression estimate (with an unregularized/unpenalized intercept as is done in part *c)* of Question 1.1) equal to $(1, -\frac{1}{8})^\top$.

Question 1.3
Plot the regularization path of the ridge regression estimator over the range $\lambda \in (0, 20.000]$ using the data of Example 1.2

Question 1.4 [‡]
Show that the ridge regression estimator can be obtained by ordinary least squares regression on an augmented data set. Hereto augment the matrix $\mathbf{X}$ with p additional rows $\sqrt{\lambda}\mathbf{I}_{pp}$, and augment the response vector $\mathbf{Y}$ with p zeros.

Question 1.6
The coefficients β of a linear regression model, $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, are estimated by $\hat{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$. The associated fitted values then given by $\widehat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y} = \mathbf{H}\mathbf{Y}$, where $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ referred to as the hat matrix. The hat matrix $\mathbf{H}$ is a projection matrix as it satisfies $\mathbf{H} = \mathbf{H}^2$. Hence, linear regression projects the response $\mathbf{Y}$ onto the vector space spanned by the columns of $\mathbf{Y}$. Consequently, the residuals $\hat{\varepsilon}$ and $\hat{\mathbf{Y}}$ are orthogonal. Now consider the ridge estimator of the regression coefficients: $\hat{\beta}(\lambda) = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}\mathbf{X}^\top\mathbf{Y}$. Let $\hat{\mathbf{Y}}(\lambda) = \mathbf{X}\hat{\beta}(\lambda)$ be the vector of associated fitted values.

a) Show that the ridge hat matrix $\mathbf{H}(\lambda) = \mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_{pp})^{-1}\mathbf{X}^\top$, associated with ridge regression, is not a projection matrix (for any $\lambda > 0$), i.e. $\mathbf{H}(\lambda) \neq [\mathbf{H}(\lambda)]^2$.

b) Show that for any $\lambda > 0$ the 'ridge fit' $\widehat{\mathbf{Y}}(\lambda)$ is not orthogonal to the associated 'ridge residuals' $\hat{\varepsilon}(\lambda)$, defined as $\varepsilon(\lambda) = \mathbf{Y} - \mathbf{X}\hat{\beta}(\lambda)$.

Question 1.9 *(Numerical inaccuracy)*
The linear regression model, $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ with $\varepsilon \sim \mathcal{N}(\mathbf{0}_n, \sigma^2\mathbf{I}_{nn})$, is fitted by to the data with the following response, design matrix, and relevant summary statistics:

$$\mathbf{X} = \begin{pmatrix} 0.3 & -0.7 \end{pmatrix}, \, \mathbf{Y} = \begin{pmatrix} 0.2 \end{pmatrix}, \, \mathbf{X}^\top\mathbf{X} = \begin{pmatrix} 0.09 & -0.21 \\ -0.21 & 0.49 \end{pmatrix}, \text{ and } \mathbf{X}^\top\mathbf{Y} = \begin{pmatrix} 0.06 \\ -0.14 \end{pmatrix}.$$

Hence, $p = 2$ and $n = 1$. The fitting uses the ridge regression estimator.

[‡]This exercise is freely rendered from Hastie *et al.* (2009), but can be found in many other places. The original source is unknown to the author.

a) Section 1.4.1 states that the regularization path of the ridge regression estimator, i.e. $\{\hat{\beta}(\lambda) : \lambda > 0\}$, is confined to a line in $\mathbb{R}^2$. Give the details of this line and draw it in the (β_1, β_2)-plane.
b) Verify numerically, for a set of penalty parameter values, whether the corresponding estimates $\hat{\beta}(\lambda)$ are indeed confined to the line found in part *a)*. Do this by plotting the estimates in the (β_1, β_2)-plane (along with the line found in part *a)*. In this use the following set of λ's:

Listing 1.6 R code

```
lambdas <- exp(seq(log(10^(-15)), log(1), length.out=100))
```

Question 1.17
Consider the standard linear regression model $Y_i = \mathbf{X}_{i,*}\beta + \varepsilon_i$ for $i = 1, \ldots, n$ and with the ε_i i.i.d. normally distributed with zero mean and a common but unknown variance. Information on the response, design matrix and relevant summary statistics are:

$$\mathbf{X}^\top = \begin{pmatrix} 2 & 1 & -2 \end{pmatrix}, \, \mathbf{Y}^\top = \begin{pmatrix} -1 & -1 & 1 \end{pmatrix}, \, \mathbf{X}^\top\mathbf{X} = \begin{pmatrix} 9 \end{pmatrix}, \text{ and } \mathbf{X}^\top\mathbf{Y} = \begin{pmatrix} -5 \end{pmatrix},$$

from which the sample size and dimension of the covariate space are immediate.
a) Evaluate the ridge regression estimator $\hat{\beta}(\lambda)$ with $\lambda = 1$.
b) Evaluate the variance of the ridge regression estimator, i.e. $\widehat{\text{Var}}[\hat{\beta}(\lambda)]$, for $\lambda = 1$. In this the error variance σ^2 is estimated by $n^{-1}\|\mathbf{Y} - \mathbf{X}\hat{\beta}(\lambda)\|_2^2$.
c) Recall that the ridge regression estimator $\hat{\beta}(\lambda)$ is normally distributed. Consider the interval

$$\mathcal{C} \;=\; \big(\hat{\beta}(\lambda) - 2\{\widehat{\text{Var}}[\hat{\beta}(\lambda)]\}^{1/2}, \, \hat{\beta}(\lambda) + 2\{\widehat{\text{Var}}[\hat{\beta}(\lambda)]\}^{1/2}\big).$$

Is this a genuine (approximate) 95% confidence interval for β? If so, motivate. If not, what is the interpretation of this interval?

Question 1.22 *(LOOCV)*
The linear regression model, $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ with $\varepsilon \sim \mathcal{N}(\mathbf{0}_n, \sigma^2\mathbf{I}_{nn})$ is fitted by means of the ridge regression estimator. The design matrix and response are:

$$\mathbf{X} = \begin{pmatrix} 2 & -1 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{Y} = \begin{pmatrix} 1 \\ \frac{1}{2} \end{pmatrix}.$$

The penalty parameter is chosen as the minimizer of the leave-one-out cross-validated squared error of the prediction (i.e. Allen's PRESS statistic). Show that $\lambda = \infty$.