Si risponda alle seguenti domande.

1. Si consideri il modello lineare semplice con intercetta e una covariata x_i. I dati sono

$$\{(y_i, x_i)_{i=1}^4\} = \{(1.4, 0), (1.4, -2), (0.8, 0), (0.4, 2)\}$$

   Riportare il valore di λ che corrisponde alla stima ridge $\hat{\beta}_\lambda = (1, -1/24)^t$ senza penalizzare l'intercetta.

2. Calcolare la stima ridge $\hat{\beta}(\lambda) = (\hat{\beta}_1(\lambda), \hat{\beta}_2(\lambda))^t$ con $\lambda = 0$ per i seguenti dati con $n = 1$ e $p = 2$:

$$X = [0.3, 0.7], y = [0.2], X^tX = \begin{bmatrix} 0.09 & -0.21 \\ -0.21 & 0.49 \end{bmatrix}, X^ty = \begin{bmatrix} 0.06 \\ -0.14 \end{bmatrix}$$

   Riportare il valore $\hat{\beta}_2(\lambda)$.

3. Sia

$$X = \begin{bmatrix} -1 & 2 \\ 0 & 1 \\ 2 & -1 \\ 1 & 0 \end{bmatrix}$$

   Calcolare $\mathbb{V}\mathrm{ar}(\hat{\beta}_2(\lambda))$ con λ pari al valore trovato al primo punto dell'esercizio, ipotizzando $\sigma^2 = 40$.

Ridge regression
Exercises

1 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 1.17
Consider the standard linear regression model $Y_i = \mathbf{X}_{i,*}\boldsymbol{\beta} + \varepsilon_i$ 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?