Homework 3

Will Edwards

Due @ 11:59pm on September 30, 2019

Part 1. Let $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{w}$, where $\mathbf{y} \in \mathbb{R}^n, \mathbf{X} \in \mathbb{R}^{n \times p}$, $\boldsymbol{\beta} \in \mathbb{R}^p$, and w_i are i.i.d. random vectors with zero mean and variance σ^2. Recall that the ridge regression estimate is given by

$$\hat{\boldsymbol{\beta}}_\lambda = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \frac{\lambda}{2}\|\boldsymbol{\beta}\|_2^2$$

1. Show that the variance of $\hat{\boldsymbol{\beta}}_\lambda$ is given by

$$\sigma^2 \mathbf{W}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{W},$$

where $\mathbf{W} = (\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I})^{-1}$. To get full credit, you need to argue why $\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I}$ is invertible.

Answer:

$$\hat{\beta} = (X^T X + \lambda I)^{-1} X^T Y$$

We let $W = (X^T X + \lambda I)^{-1}$ Then

$$Var(\hat{\beta}) = Var(W X^T Y) = W X^T Var(Y)(W X^T)^T$$

$$Var(\hat{\beta}) = W X^T \sigma^2 X W^T$$

And because W is symmetric we get the following:

$$Var(\hat{\beta}) = \sigma^2 W X^T X W$$

And we have shown the result.

We know that $\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}$ is invertible because first, for a matrix $X \in \mathbb{R}^{n \times p}$, $X^T X$ is positive semi-definite. Let $z \in \mathbb{R}^p$ Then

$$z^T X^T X z = \|Xz\|_2^2$$

And so it is positive semi-definite. The diagonals of

$$X^T X \geq 0$$

Adding λ to all the diagonals will ensure that the matrix is positive definite, because $\lambda > 0$. And it is a well-known fact that positive definite matrices are invertible. Hence, W is invertible. Q.E.D.

2. Show that the bias of $\hat{\boldsymbol{\beta}}_\lambda$ is given by

$$-\lambda \mathbf{W} \boldsymbol{\beta}$$

Answer:

$$E(\hat{\beta}_\lambda) = E((X^T X + \lambda I)^{-1} X^T Y) = (X^T X + \lambda I)^{-1} X^T X B$$

$$(X^T X + \lambda I)^{-1} X^T X B = (X^T X + \lambda I)^{-1} (X^T X + \lambda I - \lambda I) B = (X^T X + \lambda I)^{-1} (-\lambda I) B$$
$$(X^T X + \lambda I)^{-1} (-\lambda I) B = -(\lambda I) W B = -\lambda W B$$

Because λ is just a constant and we can pull it out front. Q.E.D.

3. A natural question is how to choose the tuning parameter λ. There are several classes of solutions. For example given a collection of linear estimators $\hat{\mathbf{y}} = \mathbf{S}_\lambda \mathbf{y}$, we can choose the $\mathbf{S}_\lambda$ that minimizes the generalized cross validation (GCV) criterion:

$$\mathrm{GCV}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}i}{1 - \frac{\mathrm{dof}(\mathbf{S}_\lambda)}{n}}\right)^2,$$

where the degrees of freedom $\mathrm{dof}(\mathbf{S}_\lambda)$ of a linear estimator $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$ is given by $\mathrm{tr}(\mathbf{S})$. For other criteria for performing model selection, see Efron's work "The Estimation of Prediction Error."

Ridge regression provides a linear estimator of the observed response $\mathbf{y}$ where $\mathbf{S}_\lambda = \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\mathsf{T}$. Show that the degrees of freedom of the ridge estimator is given by

$$\sum_i \frac{\sigma_i^2}{\sigma_i^2 + \lambda},$$

where σ_i is the ith singular value of $\mathbf{X}$.

Answer: Starting from

$$\mathrm{dof}(S)_\lambda = tr(S),$$

where $\mathbf{S}_\lambda = \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\mathsf{T}$.

Then we have that

$$\mathbf{S}_\lambda = \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\mathsf{T} = UDV^T(VD^TU^TUDV^T + \lambda I)^{-1}VD^TU^T$$

Recall that U^TU is the identity. Also that $D = D^T$. Finally, apply the fact that VD^TDV^T is commutable with λI. And so applying those facts gives the following:

$$UDV^T(VD^TU^TUDV^T + \lambda I)^{-1}VD^TU^T = UDV^TV(D^TD + \lambda I)^{-1}V^TVDU^T = UD(D^TD + \lambda I)^{-1}DU^T$$

Recall that D is diagonal and that each value of the diagonal is the ith singular value. So D^TD is just D^2 and so adding λ to the diagonal and finding the inverse results in inverse being the following:

$$(D^TD + \lambda I)^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2+\lambda} & 0 & 0 \\ 0 & ..... & 0 \\ 0 & 0 & \frac{1}{\sigma_n^2+\lambda} \end{bmatrix}$$

And so multiplying by UD on the left and DU^T on the right results in a diagonal matrix with diagonal values of $\frac{\sigma_i^2}{\sigma_i^2+\lambda}$. Because we have been told that degrees of freedom $\mathrm{dof}(\mathbf{S}_\lambda)$ of a linear estimator $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$ is given by $\mathrm{tr}(\mathbf{S})$ where $\mathbf{S}_\lambda = \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\mathsf{T}$, we can use the well-known fact that the trace of a matrix is the sum of its eigenvalues to get that the trace of $\mathbf{S}_\lambda\mathbf{y}$ is

$$\sum_i \frac{\sigma_i^2}{\sigma_i^2 + \lambda}$$

Q.E.D.

Part 2. Ridge Regression.

You will next add an implementation of the ridge regression to your R package.

Please complete the following steps.

Step 0: Make a file called `ridge.R` in your R package. Put it in the R subdirectory, namely we should be able to see the file at github.ncsu.edu/unityidST758/unityidST758/R/ridge.R

Step 1: Write a function `ridge_regression` that computes the ridge regression coefficient estimates for a sequence of regularization parameter values λ.

It should return an error message

- if the response variable $\mathbf{y} \in \mathbb{R}^n$ and the design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ are not conformable
- if the tuning parameters are negative

Please use the `stop` function.

- Your function should return a matrix of regression coefficients $\mathbf{B} \in \mathbb{R}^{p \times n_\lambda}$ whose columns are regression coefficient vectors for each value of λ in the vector `lambda` and n_λ is `length(lambda)`.

Step 2: Write a unit test function `test-ridge` that

- checks the error messages for your `ridge_regression` function
- checks the correctness of the estimated regression coefficients produced by `ridge_regression` function. Given data $(\mathbf{y}, \mathbf{X})$, recall that $\mathbf{b}$ is the ridge estimate with regularization parameter λ if and only if

$$(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I})\mathbf{b} = \mathbf{X}^\mathsf{T}\mathbf{y}.$$

```r
library(devtools)
```

```
## Loading required package: usethis
```

```r
setwd('/Users/bulldogwill/Documents/ST 758/twedwar2ST758/twedwar2ST758')
test()
```

```
## Loading twedwar2ST758
```

```
##
## Attaching package: 'testthat'
```

```
## The following object is masked from 'package:devtools':
##
##     test_file
```

```
## Testing twedwar2ST758
```

```
## v |  OK F W S | Context
##
/ |   0       | test_ridge
v |   4       | test_ridge
##
## == Results =====================================================================
## OK:       4
## Failed:   0
## Warnings: 0
## Skipped:  0
```

Step 3: Write a function `gcv` that computes the GCV criterion of your ridge regression model prediction error estimate.

Step 4: Write a function `leave_one_out` that computes the following leave-one-out (LOO) prediction error estimate:

$$\text{LOO}(\lambda) = \frac{1}{n}\sum_{k=1}^{n}(y_k - \hat{y}_k^{-k}(\lambda))^2,$$

where

$$y_k - \hat{y}_k^{-k}(\lambda) = \frac{y_k - \hat{y}_k(\lambda)}{1 - h_k(\lambda)},$$

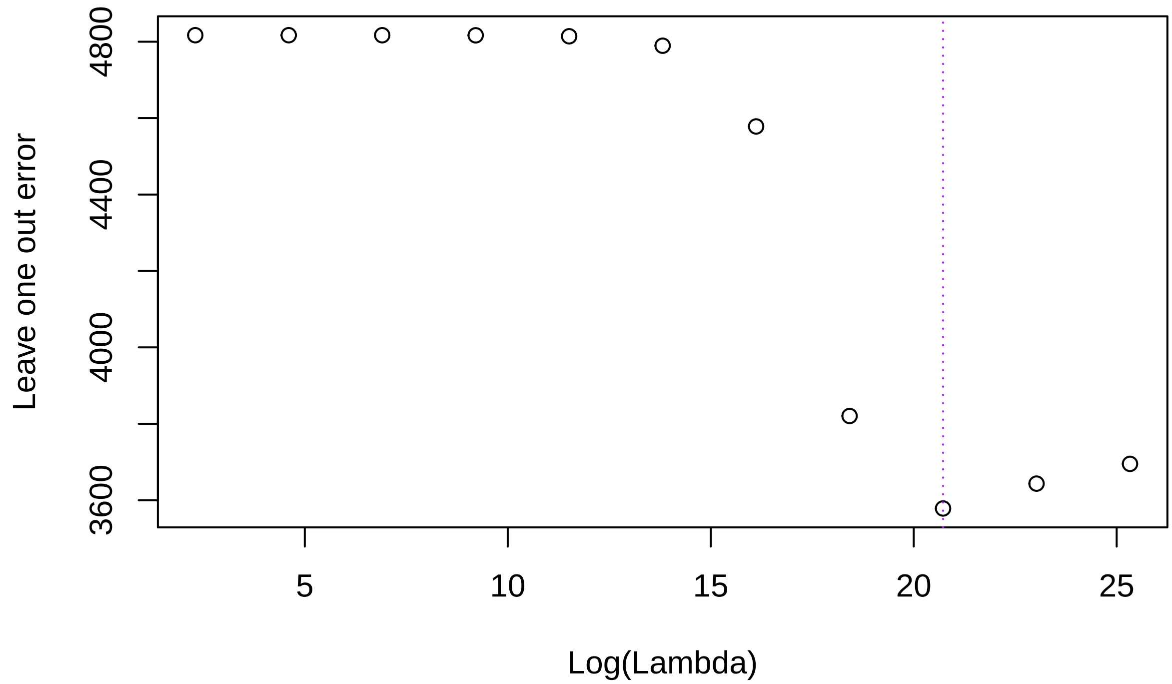
and $h_k(\lambda)$ is the kth diagonal entry of the matrix $\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top$.

Step 5: Compute ridge regression estimates of the data in `homework3_x.csv.gz` (design matrix) and `homework3_y.csv` (response) for several values of λ. This is semi-synthetic data derived from the radiation sensitivity data set available at https://web.stanford.edu/~hastie/ElemStatLearn/. Plot the LOO prediction error as a function of λ and highlight the one that minimizes the LOO prediction error (plot a vertical line at λ_{LOO}). Plot the GCV criterion as a function of λ and highlight the one that minimizes the GCV criterion (plot a vertical line at λ_{LOO}). Discuss what you would do with the information conveyed in the two plots.

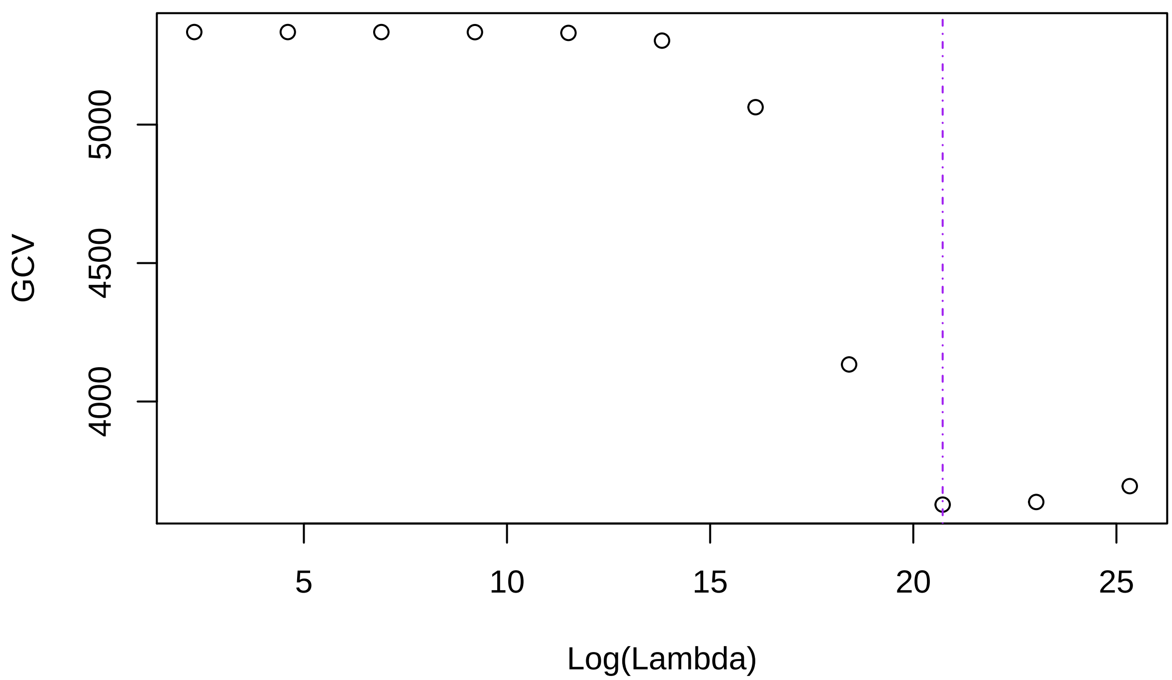
```r
library(twedwar2ST758)
library(readr)
design_mat <- as.matrix(read.csv('/Users/bulldogwill/Documents/ST 758/twedwar2ST758/HW3/homework3_x.csv
y <- read.csv('/Users/bulldogwill/Documents/ST 758/twedwar2ST758/HW3/homework3_y.csv')$V1
tuning <- c(10 , 100 ,1000 , 10^4, 10^5,10^6,10^7,10^8,10^9,10^10,10^11)


betas <- ridge_regression(y,design_mat,tuning)
loos <-leave_one_out(y, design_mat, tuning)
criterion <- gcv(y, design_mat, tuning)


plot(log(tuning), loos, xlab = 'Log(Lambda)', ylab = 'Leave one out error')
abline(v=log(10^9), col = 'purple',lty = 3)
```




```
plot(log(tuning),criterion, xlab = 'Log(Lambda)', ylab = 'GCV')
abline(v = log(10^9), col = 'purple',lty = 4)
```



Based on this information we should select $\lambda = 10^9$. It may be a good idea to search within an interval around $\lambda = 10^9$ to find an even more optimal λ, that can minimize LOO and GCV.