Compulsory exercise 1: Group 4

TMA4268 Statistical Learning V2021

Øystein Alvestad, Lars Evje and William Scott Grundeland Olsen

17 februar, 2021

Problem 1

We consider the regression problem

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon,$$

where $\mathrm{E}(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2$. We define $\mathbf{x}, \boldsymbol{\beta} \in \mathbb{R}^{p+1}$ such that $f(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$.

a)

We consider the estimator

$$\tilde{\boldsymbol{\beta}} = (X^\top X + \lambda I)^{-1} X^\top \mathbf{Y},$$

for $\boldsymbol{\beta}$. Here X is the design matrix, $\mathbf{Y}$ is the response vector, I is the identity matrix, and $\lambda \geq 0$ is a constant. We recall that for a constant matrix A of appropriate dimensions, and a random vector $\mathbf{Z}$, we have that

$$\mathrm{E}(A\mathbf{Z}) = A\,\mathrm{E}(\mathbf{Z}) \quad \text{and} \quad \mathrm{Var}(A\mathbf{Z}) = A\,\mathrm{E}(\mathbf{Z})A^\top. \tag{1}$$

First we now find the expected value

$$\mathrm{E}(\tilde{\boldsymbol{\beta}}) = \mathrm{E}((X^\top X + \lambda I)^{-1} X^\top \mathbf{Y}) = (X^\top X + \lambda I)^{-1} X^\top \,\mathrm{E}(\mathbf{Y}),$$

and because $\mathrm{E}(\mathbf{Y}) = X\boldsymbol{\beta}$, we get that

$$\mathrm{E}(\tilde{\boldsymbol{\beta}}) = (X^\top X + \lambda I)^{-1} X^\top X \boldsymbol{\beta}.$$

Similarly, the variance-covariance matrix is

$$\mathrm{Var}(\tilde{\boldsymbol{\beta}}) = \mathrm{Var}((X^\top X + \lambda I)^{-1} X^\top \mathbf{Y}) = (X^\top X + \lambda I)^{-1} X^\top \,\mathrm{Var}(\mathbf{Y})[(X^\top X + \lambda I)^{-1} X^\top]^\top,$$

and because $\mathrm{Var}(\mathbf{Y}) = \sigma^2 I$, we get that

$$\mathrm{Var}(\tilde{\boldsymbol{\beta}}) = \sigma^2 (X^\top X + \lambda I)^{-1} X^\top X (X^\top X + \lambda I)^{-1},$$

where we used $(B^{-1})^\top = (B^\top)^{-1}$, for an invertible matrix B.

b)

We now let $\tilde{f}(\mathbf{x}_0) = \mathbf{x}_0^\top \tilde{\boldsymbol{\beta}}$ be the prediction at a new covariate vector $\mathbf{x}_0$, and we wish the expected value and variance of this. Using what we learned in **a)** and Equation (1), we find that

$$\mathrm{E}(\tilde{f}(\mathbf{x}_0)) = \mathrm{E}(\mathbf{x}_0^\top \tilde{\boldsymbol{\beta}}) = \mathbf{x}_0^\top \,\mathrm{E}(\tilde{\boldsymbol{\beta}}) = \mathbf{x}_0^\top (X^\top X + \lambda I)^{-1} X^\top X \boldsymbol{\beta}.$$

Similarly, we find that

$$\mathrm{Var}(\tilde{f}(\mathbf{x}_0)) = \mathrm{Var}(\mathbf{x}_0^\top \tilde{\boldsymbol{\beta}}) = \mathbf{x}_0^\top \,\mathrm{Var}(\tilde{\boldsymbol{\beta}})\mathbf{x}_0 = \sigma^2 \mathbf{x}_0^\top (X^\top X + \lambda I)^{-1} X^\top X (X^\top X + \lambda I)^{-1} \mathbf{x}_0. \tag{2}$$

c)

We want to find the expected MSE at $\mathbf{x}_0$, and this is most easily done using the relationship between the expected value and the variance-covariance matrix,

$$\mathrm{E}((y_0 - \tilde{f}(\mathbf{x}_0))^2) = \mathrm{E}(\tilde{f}(\mathbf{x}_0) - f(\mathbf{x}_0))^2 + \mathrm{Var}(\tilde{f}(\mathbf{x}_0)) + \mathrm{Var}(\varepsilon).$$

The last two terms are known to us now, so we look further at the first term, the squared bias,

$$\begin{aligned} \mathrm{E}(\tilde{f}(\mathbf{x}_0) - f(\mathbf{x}_0))^2 &= [\mathrm{E}(\tilde{f}(\mathbf{x}_0)) - \mathrm{E}(f(\mathbf{x}_0))]^2 = [\mathrm{E}(\tilde{f}(\mathbf{x}_0)) - f(\mathbf{x}_0)]^2 \\ &= [\mathbf{x}_0^\top (X^\top X + \lambda I)^{-1} X^\top X \boldsymbol{\beta} - \mathbf{x}_0^\top \boldsymbol{\beta}]^2. \end{aligned} \tag{3}$$

We then get that

$$\begin{aligned} \mathrm{E}((y_0 - \tilde{f}(\mathbf{x}_0))^2) = \sigma^2 &+ [\mathbf{x}_0^\top (X^\top X + \lambda I)^{-1} X^\top X \boldsymbol{\beta} - \mathbf{x}_0^\top \boldsymbol{\beta}]^2 \\ &+ \sigma^2 \mathbf{x}_0^\top (X^\top X + \lambda I)^{-1} X^\top X (X^\top X + \lambda I)^{-1} \mathbf{x}_0. \end{aligned}$$

d)

We start by importing the relevant quantities, as given in the project description.

```r
id <- "1X_8OKcoYbng1XvYFDirxjEWr7LtpNr1m" # Google file ID
values <- dget(sprintf("https://docs.google.com/uc?id=%s&export=download", id))

X <- values$X
x0 <- values$x0
beta <- values$beta
sigma <- values$sigma
```

We may then calculate the squared bias $\mathrm{E}(\tilde{f}(\mathbf{x}_0) - f(\mathbf{x}_0))^2$, using Equation (3). Plotting this for $\lambda \in [0, 2]$, we get the result as in Figure 1. We see that the squared bias increases with large $\lambda > 0.5$, and has a minimum when $\lambda = 0$ and $\lambda \approx 0.5$. This is expected for $\lambda = 0$, because we then get that $\tilde{\boldsymbol{\beta}}$ is equal to the OLS estimator. The bias measures how good the estimator is in estimating the real value, so this makes sense. From the figure it also appears that for $\lambda \approx 0.5$, the estimator $\tilde{\boldsymbol{\beta}}$ us estimating the real value $\boldsymbol{\beta}$ very good. For increasing λ after this, the estimator seems to do a worse job in estimating $\boldsymbol{\beta}$.

```r
library(ggplot2)
bias <- function(lambda, X, x0, beta) {
  p <- ncol(X)
  inv <- solve(t(X) %*% X + lambda * diag(p))
  value <- (t(x0) %*% inv %*% t(X) %*% X %*% beta - t(x0) %*% beta)^2
  return(value)
}

lambdas <- seq(0, 2, length.out = 500)
BIAS <- rep(NA, length(lambdas))

for (i in 1:length(lambdas)) {
  BIAS[i] <- bias(lambdas[i], X, x0, beta)
}
dfBias <- data.frame(lambdas = lambdas, bias = BIAS)

ggplot(dfBias, aes(x = lambdas, y = bias)) +
  geom_line(color = "red") +
  xlab(expression(lambda)) +
  ylab(expression(bias^2))
```

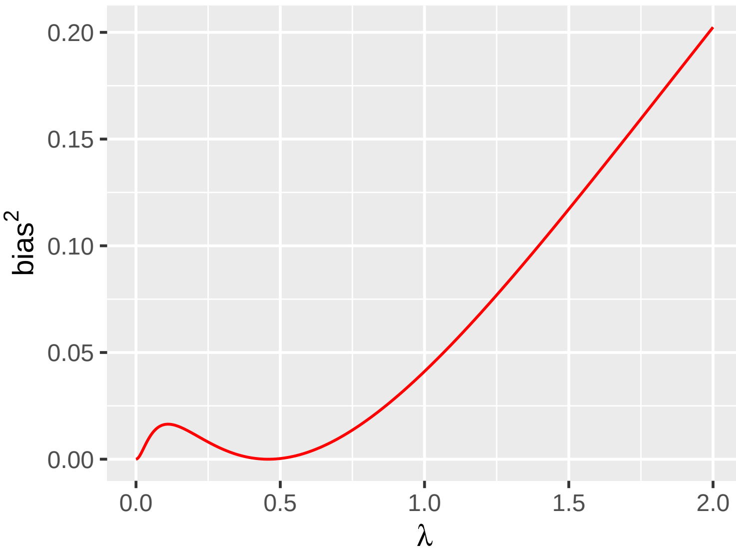

Figure 1: The squared bias $\mathrm{E}(\tilde{f}(\mathbf{x}_0) - f(\mathbf{x}_0))^2$ as a function of λ.

e)

We are now interested in $\mathrm{Var}(\tilde{f}(\mathbf{x}_0))$, which can be calculated using Equation (2). Using $\lambda \in [0, 2]$, the result is plotted in Figure 2. We notice that the variance is decreasing for increasing values of λ. That is, as long as λ increases, the model is less prone to overfitting. Letting $\lambda \to \infty$, we also see that the variance tends to zero.

```
variance <- function(lambda, X, x0, sigma) {
  p <- ncol(X)
  inv <- solve(t(X) %*% X + lambda * diag(p))
  value <- sigma^2 * t(x0) %*% inv %*% t(X) %*% X %*% inv %*% x0
  return(value)
}

lambdas <- seq(0, 2, length.out = 500)
VAR <- rep(NA, length(lambdas))

for (i in 1:length(lambdas)) {
  VAR[i] <- variance(lambdas[i], X, x0, sigma)
}
dfVar <- data.frame(lambdas = lambdas, var = VAR)

ggplot(dfVar, aes(x = lambdas, y = var)) +
  geom_line(color = "green4") +
  xlab(expression(lambda)) +
  ylab("variance") +
  theme(plot.title = element_text(hjust = 0.5))
```

f)

Lastly, we are interested in the expected MSE at $\mathbf{x}_0$, $\mathrm{E}((y_0 - \tilde{f}(\mathbf{x}_0))^2)$, which, when we know the squared bias and the variance from **e)**, is given as $(\text{bias})^2 + \mathrm{Var}(\tilde{f}(\mathbf{x}_0)) + \sigma^2$, because the irreducible error is σ^2. The plot of the expected MSE, the squared bias and the variance is shown in Figure 3, and by using `lambdas[which.min(exp_mse)]` we find that the minimal expected MSE is found when $\lambda \approx 0.993988$. That

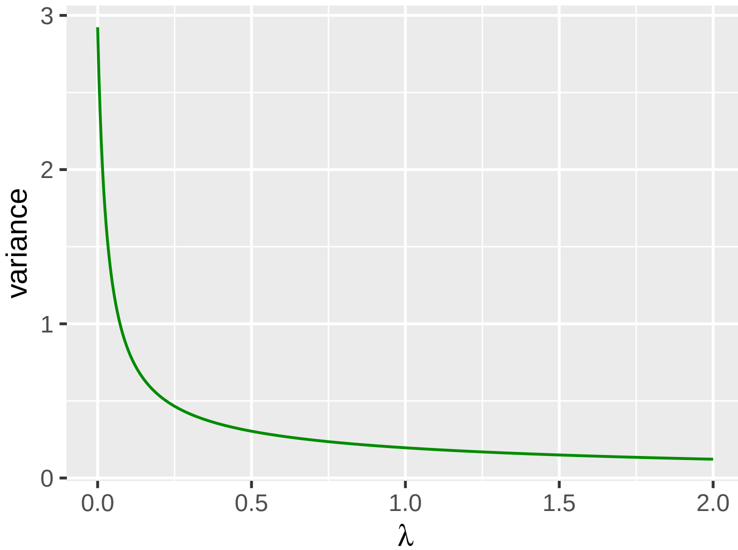

Figure 2: The variance $\mathrm{Var}(\tilde{f}(\mathbf{x}_0))$ as a function of λ.

is, it is possbile to move to $\lambda > 0$, and taking a little bias, and reducing the variance, leading to a reduction in the expected MSE.

```
exp_mse <- BIAS + VAR + sigma^2

cols <- c("exp_mse" = "blue", "bias" = "red", "variance" = "green4")
dfAll <- data.frame(lambda = lambdas, bias = BIAS, var = VAR, exp_mse = exp_mse)

ggplot(dfAll)+
  geom_line(aes(x = lambda, y = exp_mse, color = "exp_mse")) +
  geom_line(aes(x = lambda, y = bias, color = "bias")) +
  geom_line(aes(x = lambda, y = var, color = "variance")) +
  xlab(expression(lambda)) +
  ylab(expression(E(MSE))) +
  theme(legend.title = element_blank()) +
  scale_color_manual(values = cols)
```

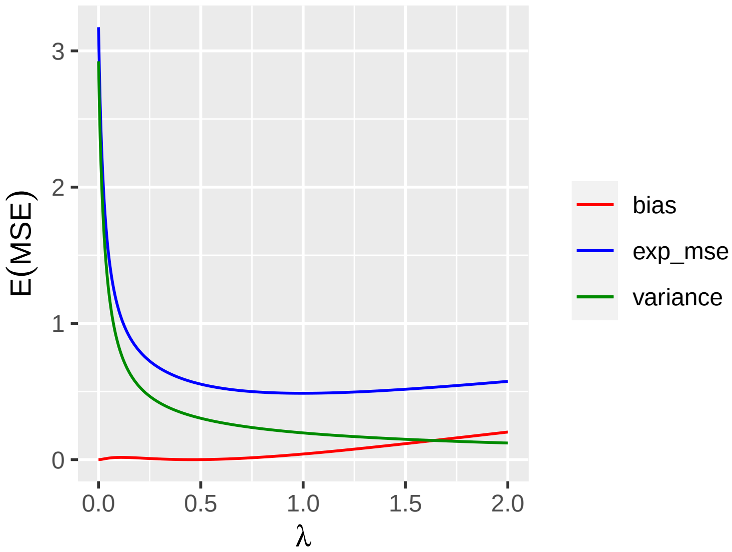

Figure 3: The expected MSE $\mathrm{E}((y_0 - \tilde{f}(\mathbf{x}_0))^2)$ as a function of λ, together with the squared bias and the variance.

Problem 2

a)

b)

c)

d)

Problem 3

a)

b)

c)

d)

Problem 4

a)

We wish to show that for the linear regression model $Y = X\beta + \varepsilon$, the LOOCV statistic is given by

$$\mathrm{CV} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2,$$

where $h_i = \mathbf{x}_i^\top (X^\top X)^{-1}\mathbf{x}_i$, and $\mathbf{x}_i^\top$ is the i-th row of X.

Generally we can write $\mathrm{CV} = \sum_{i=1}^{N} e_{(-i)}^2 / N$, where $e_{(-i)} = y_i - \hat{y}_{(-i)}$. We use the notation $A_{(-i)}$ to symbolize that element i is removed from A if it is a vector, and row i is removed from A if it is a matrix. For a linear regression model $\mathbf{Y} = X\boldsymbol{\beta} + \varepsilon$, the estimate of $\boldsymbol{\beta}$ without the i-th case is

$$\hat{\boldsymbol{\beta}}_{(-i)} = (X_{(-i)}^\top X_{(-i)})^{-1} X_{(-i)}^\top \mathbf{Y}_{(-i)}.$$

From what we then know, we may write $e_{(-i)} = y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_{(-i)}$. We want another expression for $\hat{\boldsymbol{\beta}}_{(-i)}$, and using the Sherman-Morrison formula,

$$(X_{(-i)}^\top X_{(-i)})^{-1} = (X^\top X - \mathbf{x}_i \mathbf{x}_i^\top)^{-1} = (X^\top X)^{-1} - \frac{(X^\top X)^{-1}\mathbf{x}_i\mathbf{x}_i^\top (X^\top X)^{-1}}{1 + \mathbf{x}_i (X^\top X)^{-1} \mathbf{x}_i^\top}.$$

By the definition of h_i, we then get that

$$(X_{(-i)}^\top X_{(-i)})^{-1} = (X^\top X)^{-1} + \frac{(X^\top X)^{-1}\mathbf{x}_i\mathbf{x}_i^\top (X^\top X)^{-1}}{1 - h_i}.$$

It is also clear that $X_{(-i)}^\top \mathbf{Y}_{(-i)} = X^\top \mathbf{Y} - \mathbf{x}_i y_i$, and thus

$$\hat{\boldsymbol{\beta}}_{(-i)} = (X_{(-i)}^\top X_{(-i)})^{-1} X_{(-i)}^\top \mathbf{Y}_{(-i)} = \left[(X^\top X)^{-1} + \frac{(X^\top X)^{-1}\mathbf{x}_i\mathbf{x}_i^\top (X^\top X)^{-1}}{1 - h_i}\right](X^\top \mathbf{Y} - \mathbf{x}_i y_i).$$

Multiplying out this expression we then get

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{(-i)} &= (X^\top X)^{-1} X^\top \mathbf{Y} + \frac{(X^\top X)^{-1}\mathbf{x}_i\mathbf{x}_i^\top (X^\top X)^{-1}}{1 - h_i} X^\top \mathbf{Y} - (X^\top X)^{-1}\mathbf{x}_i y_i - \frac{(X^\top X)^{-1}\mathbf{x}_i\mathbf{x}_i^\top (X^\top X)^{-1}}{1 - h_i}\mathbf{x}_i y_i \\
&= \hat{\boldsymbol{\beta}} + \frac{(X^\top X)^{-1}\mathbf{x}_i\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}}{1 - h_i} - (X^\top X)^{-1}\mathbf{x}_i y_i - \frac{(X^\top X)^{-1}\mathbf{x}_i h_i}{1 - h_i} y_i \\
&= \hat{\boldsymbol{\beta}} + \frac{(X^\top X)^{-1}\mathbf{x}_i}{1 - h_i}\left[\mathbf{x}_i^\top \hat{\boldsymbol{\beta}} - y_i(1 - h_i) - h_i y_i\right] \\
&= \hat{\boldsymbol{\beta}} + \frac{(X^\top X)^{-1}\mathbf{x}_i}{1 - h_i}(\hat{y}_i - y_i) = \hat{\boldsymbol{\beta}} - \frac{(X^\top X)^{-1}\mathbf{x}_i}{1 - h_i} e_i,
\end{aligned}$$

where we let $e_i = y_i - \hat{y}_i$. This allows us to find

$$e_{(-i)} = y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_{(-i)} = y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}} + \frac{\mathbf{x}_i^\top (X^\top X)^{-1}\mathbf{x}_i}{1 - h_i} e_i = e_i + \frac{h_i}{1 - h_i} e_i = \frac{e_i}{1 - h_i}.$$

It then follows that

$$\mathrm{CV} = \frac{1}{N}\sum_{i=1}^{N} e_{(-i)}^2 = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{e_i}{1 - h_i}\right)^2 = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2,$$

which was what to be shown. ***Q.E.D.***

b)

False, True, True, False.

Problem 5

a)

b)