Compulsory exercise 1: Group 5

TMA4268 Statistical Learning V2021

Hans Røhjell Odland and Aksel Haugen Madslien

2/9/2021

For some problems you will need to include some LaTex code. Please install latex on your computer and then consult Compulsory1.Rmd for hints how to write formulas in LaTex.

An example:

$$Y_i = f(x_i) + \varepsilon_i \ ,$$

Or the same formula $Y_i = f(x_i) + \varepsilon_i$ in-line.

Problem 1

a)

We consider $Y = f(\mathbf{x}) + \varepsilon,$ where $\text{E}(\varepsilon) = 0$ and $\text{Var}(\varepsilon) = \sigma^2$.

We find the expected value for $\tilde{\beta}$ as

$$E(\tilde{\beta}) = E[(\mathbf{x}^T\mathbf{x}+\lambda\mathbf{I})^{-1}\mathbf{x}^T\mathbf{y}] = (\mathbf{x}^T\mathbf{x}+\lambda\mathbf{I})^{-1}\mathbf{x}^T E[\mathbf{y}] = (\mathbf{x}^T\mathbf{x}+\lambda\mathbf{I})^{-1}\mathbf{x}^T\mathbf{x}\beta+\lambda\mathbf{I}\beta-\lambda\mathbf{I}\beta = (\mathbf{x}^T\mathbf{x}+\lambda\mathbf{I})^{-1}(-\lambda\mathbf{I})\beta+\mathbf{I}\beta = \beta-\lambda(\mathbf{x}^T\mathbf{x}-$$

b)

We let $\widetilde{f}(\mathbf{x}_0) = \mathbf{x}_0^T\widetilde{\boldsymbol{\beta}}$ The variance for $\widetilde{f}(\mathbf{x}_0)$ then becomes

$$E[\widetilde{f}(\mathbf{x}_0)] = E[\mathbf{x}_0^T\widetilde{\boldsymbol{\beta}}] = \mathbf{x}_0^T E[\widetilde{\boldsymbol{\beta}}] = \mathbf{x}_0^T(\beta - \lambda(\mathbf{x}^T\mathbf{x} + \lambda\mathbf{I})\beta)$$

For the variation we get

$$Var[\widetilde{f}(\mathbf{x}_0)] = \mathbf{x}_0^T Var[\widetilde{\boldsymbol{\beta}}]\mathbf{x} = \mathbf{x}_0^T(\beta - \lambda(\mathbf{x}^T\mathbf{x} + \lambda\mathbf{I})\beta)\mathbf{x}$$

c)

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

```
## [1] 100  81
```

```
x0 = values$x0
dim(x0)
```

```
## [1] 81  1
```

```
beta=values$beta
dim(beta)
```

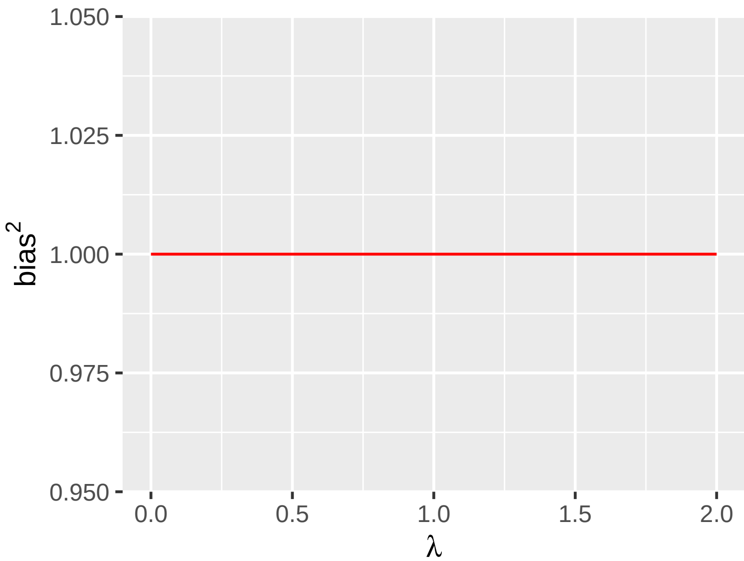
```
## [1] 81  1
```

```
sigma=values$sigma
sigma
```

```
## [1] 0.5
```

d)

```
library(ggplot2)
bias = function(lambda, X, x0, beta) {
    p = ncol(X)
    value = 1
    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))
```



Problem 2

a)

```r
id <- "1yYlEl5gYY3BEtJ4d7KWaFGIOEweJIn__" # google file ID
d.corona <- read.csv(sprintf("https://docs.google.com/uc?id=%s&export=download", id),header=T)


table(d.corona$deceased)
```

```
##
##    0    1
## 1905  105
```

```r
table(d.corona$country, d.corona$sex)
```

```
##
##             female male
##   France        60   54
##   indonesia     30   39
##   japan        120  174
##   Korea        879  654
```

```r
table(d.corona$sex, d.corona$deceased)
```

```
##
##             0    1
##   female 1046   43
##   male    859   62
```

```r
francedf <- subset(d.corona, country == "France")
table(francedf$sex, francedf$deceased)
```

```
##
##           0  1
##   female 55  5
##   male   43 11
```

b)

i)

```r
# Multiple Linear Regression Example
fit <- lm(as.numeric(deceased) ~ sex + country + age, data=d.corona)
summary(fit)$coef # show results
```

```
##                      Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept)       0.043861643 0.0252284684  1.738577 8.226271e-02
## sexmale           0.030814710 0.0099018024  3.112030 1.884264e-03
## countryindonesia -0.053478088 0.0335835996 -1.592387 1.114555e-01
## countryjapan     -0.097524595 0.0242695290 -4.018397 6.076062e-05
## countryKorea     -0.071966403 0.0215415300 -3.340821 8.506420e-04
## age               0.001304608 0.0002180126  5.984098 2.571005e-09
```

```r
#predict(fit)

deceasedmale <- fit$coefficients[6]*75 + fit$coefficients[2] + fit$coefficients[5]

deceasedmale
```

```
##        age
## 0.05669394
```

The probability of dying of Covid-19 for a male age in Korea is found to be 5.7%.

ii)

Does males have higher probability to die than females?

```r
fit <- glm(deceased ~ sex, data = d.corona, family = binomial)
summary(fit)$coef
```

```
##              Estimate Std. Error    z value     Pr(>|z|)
## (Intercept) -3.191529  0.1556015 -20.510908 1.720507e-93
## sexmale      0.562894  0.2037278   2.762971 5.727782e-03
```

The estimate readings for men dying of corona is positive, which means that we can conclude that men have a higher probability of dying of corona than women.

###iii)

```r
fit <- glm(deceased ~ country, data = d.corona, family = 'binomial')
summary(fit)$coef
```

```
##                    Estimate Std. Error   z value     Pr(>|z|)
## (Intercept)      -1.8123788  0.2696369 -6.721552 1.797990e-11
## countryindonesia -0.7370664  0.5369628 -1.372658 1.698586e-01
## countryjapan     -1.4351729  0.4088358 -3.510389 4.474509e-04
## countryKorea     -1.1833535  0.2951062 -4.009925 6.073807e-05
```

From these readings we can conclude that there is not enough evidence to say that there is a higher risk of dying of corona in Indonesia than in France, since the p-value is not significant. For Japan and Korea the p-value is much more significant, and also has a negative estimate, which means that there is a higher risk of dying of corona in Japan and Korea than in France.

###iv) A person is 10 years older than another person. The probability of dying is linear in terms of age, so we can see the probability of a person dying at an age of 65 and an age of 75 from task i.

```r
fit <- lm(as.numeric(deceased) ~ sex + country + age, data=d.corona)
summary(fit)$coef
```

```
##                     Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept)       0.043861643 0.0252284684  1.738577 8.226271e-02
## sexmale           0.030814710 0.0099018024  3.112030 1.884264e-03
## countryindonesia -0.053478088 0.0335835996 -1.592387 1.114555e-01
## countryjapan     -0.097524595 0.0242695290 -4.018397 6.076062e-05
## countryKorea     -0.071966403 0.0215415300 -3.340821 8.506420e-04
## age               0.001304608 0.0002180126  5.984098 2.571005e-09
```

```r
deceasedmale75 <- fit$coefficients[6]*75 + fit$coefficients[2] + fit$coefficients[5]
deceasedmale65 <- fit$coefficients[6]*65 + fit$coefficients[2] + fit$coefficients[5]
deceasedmale55 <- fit$coefficients[6]*55 + fit$coefficients[2] + fit$coefficients[5]
```