Compulsory exercise 1: Group 27

TMA4268 Statistical Learning V2021

Maren Bråthen Kristoffersen, Vilde Marie Skårdal Jensen and Viveka Priya Simhan

19 februar, 2021

Problem 1

a)

The expected value of the estimator $\tilde{\beta}$ is derived the following way

$$\begin{aligned}\mathrm{E}\left[\tilde{\beta}\right] &= \mathrm{E}\left[\left(X^TX+\lambda I\right)^{-1}X^TY\right] = \left(X^TX+\lambda I\right)^{-1}X^T\ \mathrm{E}[Y] \\ &= \left(X^TX+\lambda I\right)^{-1}X^T\ \mathrm{E}\left[X\beta+\epsilon\right] = \left(X^TX+\lambda I\right)^{-1}X^TX\beta.\end{aligned}$$

where Y is the vector of all $Y_i = x_i^T\beta + \epsilon_i$ for $i = 1, ..., p$.
The variance-covariance matrix of $\tilde{\beta}$ is

$$\begin{aligned}\mathrm{Cov}\left(\tilde{\beta}\right) &= \mathrm{Cov}\left(\left(X^TX+\lambda I\right)^{-1}X^TY\right) = \left(X^TX+\lambda I\right)^{-1}X^T\mathrm{Cov}(Y)\left(\left(X^TX+\lambda I\right)^{-1}X^T\right)^T \\ &= \left(X^TX+\lambda I\right)^{-1}X^T\sigma^2\left(\left(X^TX+\lambda I\right)^{-1}X^T\right)^T.\end{aligned}$$

b)

The expected value of $\tilde{f}(x_0) = x_0^T\tilde{\beta}$ is

$$\mathrm{E}\left[\tilde{f}(x_0)\right] = \mathrm{E}\left[x_0^T\tilde{\beta}\right] = x_0^T\mathrm{E}\left[\tilde{\beta}\right] = x_0^T\left(X^TX+\lambda I\right)^{-1}X^TX\beta.$$

The variance of $\tilde{f}(x_0) = x_0^T\tilde{\beta}$ is

$$\begin{aligned}\mathrm{Var}\left[\tilde{f}(x_0)\right] &= \mathrm{Var}\left[x_0^T\tilde{\beta}\right] = x_0^T\mathrm{Var}\left[\tilde{\beta}\right]x_0 \\ &= x_0^T\left(X^TX+\lambda I\right)^{-1}X^T\sigma^2\left(\left(X^TX+\lambda I\right)^{-1}X^T\right)^T x_0.\end{aligned}$$

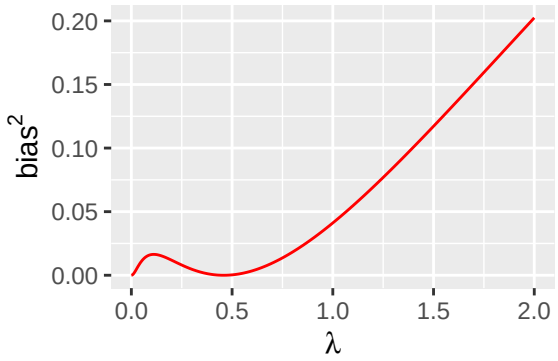
c)

The expected mean square error MSE at x_0 can be expressed the following way

$$\begin{aligned}\text{MSE} &= \text{E}\left[(y_0 - \tilde{f}(x_0))^2\right] = \left[\text{E}(\tilde{f}(x_0) - f(x_0))\right]^2 + \text{Var}(\tilde{f}(x_0)) + Var(\epsilon) \\ &= \left[\text{E}\left(\tilde{f}(x_0)\right) - \text{E}(f(x_0))\right]^2 + \text{Var}(\tilde{f}(x_0)) + Var(\epsilon) \\ &= \left[x_0^T \left(X^T X + \lambda I\right)^{-1} X^T X\beta - x_0^T \beta\right]^2 + x_0^T \left(X^T X + \lambda I\right)^{-1} X^T \sigma^2 \left(\left(X^T X + \lambda I\right)^{-1} X^T\right)^T x_0 + \sigma^2 I\end{aligned}$$

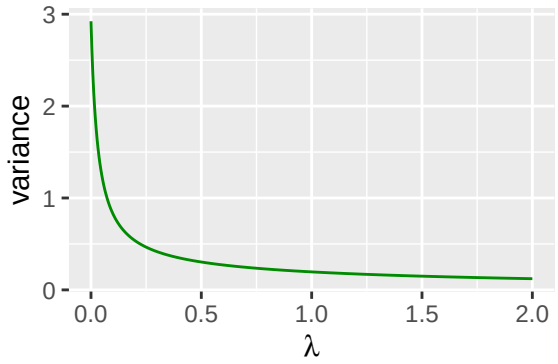
where the first term represent the squared bias, the second term the variance and the last term the irreducible error which will be used in the following tasks.

d)



The figure shows that the squared bias has a minimum value at $\lambda \approx 0.42$ and then increases with higher values of λ. This is due to the fact that using a shrinkage penalty places additional constraints on the coefficients β_i which increase with an increasing value of λ. This leads to a more rigid model and therefore an increased bias.

e)

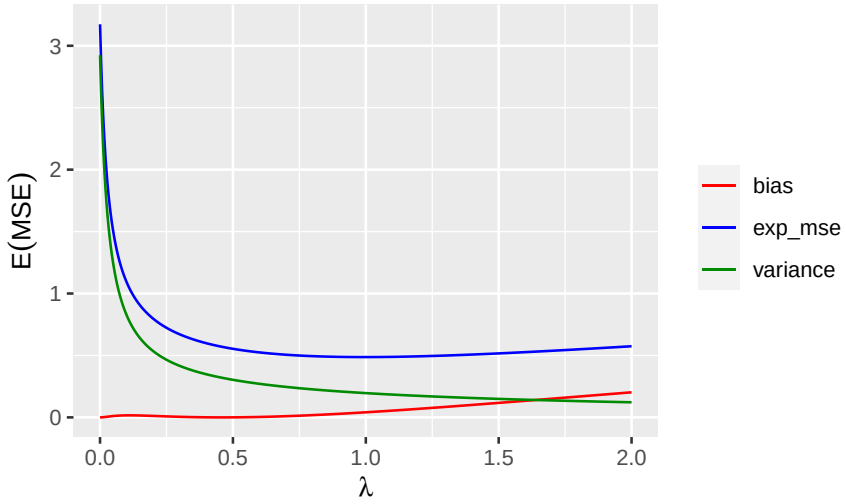


In the figure above we see that the variance decreases for increasing values of λ. Again this is due to the fact that a higher value of λ leads to a more rigid model which is less affected by changes in the data and thus has a lower variance.

f)

```
exp_mse =  BIAS + VAR + sigma^2
lambdas[which.min(exp_mse)]
```

```
## [1] 0.993988
```



The figure shows that the variance contributes significantly more to the total estimated mean squared error than the bias. As discussed earlier, the use of the shrinkage penalty controls the Bias-Variance trade-off where the bias increases and the variance decreases for higher values of λ. We can see that the variance decreases most steeply when λ is small and we have seen earlier that the bias increases for values of λ higher than approximately 0.42. Therefore, the total MSE has a minimum at $\lambda = 0.993988$.

Problem 2

a)

```
table(d.corona$deceased)
```

```
##
## Non-deceased     Deceased
##         1905          105
```

```
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$deceased, d.corona$sex)
```

```
##
##                Female Male
##   Non-deceased   1046  859
##   Deceased         43   62
```

```r
#data from France
d.corona.france = d.corona[ which(d.corona$country == "France"),]
table(d.corona.france$deceased, d.corona.france$sex)
```

```
##
##                Female Male
##   Non-deceased     55   43
##   Deceased          5   11
```

b)

```r
covid.glm = glm(deceased ~ sex + age + country, data = d.corona,
                family = "binomial")
```

The data set consists of categorical predictors with binary response where we are interested in understanding the relationship between the predictors and the probability to die of covid-19. Therefore, a logistic regression model is chosen, where it is assumed that the binary response Y_i follows a Bernoulli distribution with probability of decease p_i.

(i) To find the probability that a 75 year old man from Korea who is infected with covid-19, will die from the infection, we use the intercepts in the logistic regression model to find $\eta = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$. We then find the probability using that $\eta = \log\left(\frac{p}{1-p}\right)$.

```r
person_vec <- c(1, 1, 75, 0, 0, 1) #covariate vector, the first 1 for the intercept \beta_0
eta_person <- covid.glm$coefficients %*% person_vec
p_man = exp(eta_person) / (1+ exp(eta_person))
p_man
```

```
##           [,1]
## [1,] 0.1084912
```

We find that $p =$ 0.1084912.

(ii) To investigate whether there is evidence from the data set that males have a higher probability to die than women, we look at the summary of the coefficients in the logistic regression model and observe the estimated values and p-values of the coefficients.

```r
summary(covid.glm)$coefficients
```

```
##                     Estimate  Std. Error    z value     Pr(>|z|)
## (Intercept)      -3.99348478 0.462190319 -8.6403471 5.604250e-18
## sexMale           0.62606777 0.209044668  2.9948995 2.745353e-03
## age               0.02713421 0.004736262  5.7290352 1.010034e-08
## countryIndonesia -0.41185539 0.550050523 -0.7487592 4.540024e-01
## countryJapan     -1.34338289 0.417195836 -3.2200295 1.281774e-03
## countryKorea     -0.77389515 0.307979819 -2.5128112 1.197734e-02
```

The estimate for `sexMale` is $\beta_1 =$ `0.6260678` which means that the η-value will increase with this value for a male compared to a female. This means that the probability to die if infected by covid-19 increases. Based on this estimate and the fact that the p-value is relatively low, `0.0027454`, we conclude that there is evidence that males have a higher probability to decease.

(iii) By looking at the estimates and the p-values from the summary of the coefficients printed in the previous task, we can see that the probability to decease if infected in Japan is lower than in France. We can also see that there is a similar relation for Korea as well, even though the coefficient is less significant. For Indonesia, there is no evidence to suggest a difference from France, as a p-value of `0.4540024` is too high to discard the idea that there is no relationship. Despite this, overall we conclude that there is evidence that the country of residence has an influence on the probability to decease.

(iv) To quantify how the odds to die changes, we have determined the odds ratio by using the estimates from the logistic regression line to calculate two η-values, where an arbitrary individual is compared to an individual with identical covariates except that age predictor is increased by 10 years.

```
eta_0 <- covid.glm$coefficients  %*% c(1, 1, 10, 1, 0, 0)
eta_10 <- covid.glm$coefficients %*% c(1, 1, 20, 1, 0, 0)

p_0 <- exp(eta_0) / (1 + exp(eta_0))
p_10 <- exp(eta_10) / (1 + exp(eta_10))

odds_0 <- (p_0/(1-p_0))
odds_10 <- (p_10/(1-p_10))

odds_ratio <- odds_10/odds_0
odds_ratio
```

```
##          [,1]
## [1,] 1.311724
```

The resulting odds ratio is `1.3117238`. This corresponds to changing the odds to die by a factor $e^{\beta_2 \cdot 10}$. The odds ratio means that the probability of deceasing of covid-19 increases with 57% if one is 10 years older.

c)

(i) To investigate whether age is a greater risk factor for males than for females, we fit a logistic regression model including an interaction term between the predictors age and sex.

```
covid_sex_age.glm <- glm(deceased ~ age*sex + country, data = d.corona,
                         family = "binomial")
summary(covid_sex_age.glm)$coefficients
```

```
##                     Estimate  Std. Error    z value     Pr(>|z|)
## (Intercept)      -3.95358024 0.571711957 -6.9153359 4.667559e-12
## age               0.02648543 0.007260877  3.6476902 2.646084e-04
## sexMale           0.55674536 0.624833574  0.8910298 3.729132e-01
## countryIndonesia -0.41019697 0.550304327 -0.7454002 4.560298e-01
## countryJapan     -1.34443976 0.417364271 -3.2212622 1.276273e-03
## countryKorea     -0.77259572 0.308244156 -2.5064408 1.219535e-02
## age:sexMale       0.00111088 0.009442750  0.1176437 9.063500e-01
```

We do not find evidence that age is a greater risk factor for males than females, since the interaction term `0.0011109` is close to zero, meaning that the two predictors are more or less independent of each other. More importantly, the p-value is `0.90635`, meaning that there is high chance of observing this data without there being a correlation.

(ii) To investigate whether age is a greater risk factor for the French population than for the Indonesian population, we fit a logistic regression model including an interaction term between the predictors age and country.

```
covid_country_age.glm <- glm(deceased ~ age*country + sex, data = d.corona,
                             family = "binomial")
summary(covid_country_age.glm)$coefficients
```

```
##                         Estimate Std. Error   z value     Pr(>|z|)
## (Intercept)          -7.04271750 1.72789145 -4.075903 4.583618e-05
## age                   0.06692562 0.02124167  3.150676 1.628933e-03
## countryIndonesia      4.34249345 2.16593765  2.004902 4.497349e-02
## countryJapan          2.13091215 2.03299249  1.048165 2.945625e-01
## countryKorea          2.37161898 1.75357306  1.352449 1.762316e-01
## sexMale               0.62777239 0.21055725  2.981481 2.868581e-03
## age:countryIndonesia -0.07189023 0.03310051 -2.171877 2.986496e-02
## age:countryJapan     -0.04630417 0.02667635 -1.735776 8.260341e-02
## age:countryKorea     -0.04141745 0.02189491 -1.891648 5.853792e-02
```

The interaction term between age and residence in Indonesia has estimate `-0.0718902`, with p-value `0.029865`. This estimate will be multiplied with a number between 2 and 99 or with zero, so that the resulting term in η can reach a similar magnitude as the intercept term. Additionally, the low p-value suggests it is plausible to discard the idea that there is no relationship. So there is evidence that age is a greater risk factor for the French population than for the Indonesian population.

d) Multiple choice

(i) TRUE
(ii) TRUE
(iii) TRUE
(iv) FALSE

Problem 3

a)

(i)

$$\begin{aligned}
\log\left(\frac{p_i}{1-p_i}\right) &= \log\left(\frac{\frac{\exp(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_7 x_{i7})}{1+\exp(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_7 x_{i7})}}{1-\frac{\exp(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_7 x_{i7})}{1+\exp(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_7 x_{i7})}}\right) \\
&= \log\left(\frac{e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_7 x_{i7}}}{1+e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_7 x_{i7}}-e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_7 x_{i7}}}\right) \\
&= \log(e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_7 x_{i7}}) \\
&= \beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_7 x_{i7}
\end{aligned}$$

(ii)

```
prediction = predict(logReg, newdata = test, type = "response")
prediction = ifelse(prediction > 0.5, 1, 0) #using 0.5 as the cut-off probability

table(predicted = prediction, test = test$diabetes)
```

```
##           test
## predicted   0   1
##         0 137  29
##         1  18  48
```

The sensitivity is $\frac{\#\text{True Positive}}{\#\text{Positive}} = \frac{48}{48+29} \approx 0.62$ and the specificity is $\frac{\#\text{True Negative}}{\#\text{Negative}} = \frac{137}{137+18} \approx 0.88$.

b)

(i) π_k is the prior probability for the class k, i.e. the probability that a random observation belongs to class k. We do not know the prior probabilities, but we can estimate them by $\hat{\pi}_k = \frac{n_k}{n}$, the number of training observations belonging to class k,n_k, divided by the total number of the training data, n. In this data set we have to classes, has diabetes (1) and does not have diabetes (0). Using the training data we find the following estimates for the prior probabilities: $\hat{\pi}_0 = \frac{n_0}{n} = \frac{200}{300} = \frac{2}{3}$ and $\hat{\pi}_1 = \frac{n_1}{n} = \frac{100}{300} = \frac{1}{3}$.

μ_k is the mean value vector for class k, with μ_{ki} being the mean value of the i^{th} covariate for observations belonging to class k. Again we do not know μ_k, but we can estimate it using the observations in the training data, so that $\frac{1}{n_k}\sum_{i:y_i=k} X_i$ where X_i^T is the i^{th} row in the design matrix. For this particular data set, we will have two mean value vectors, μ_0 and μ_1, with 7 elements each.

Σ is the pooled covariance matrix for both classes, which we assume to be equal for both classes. This is estimated by first estimating the covariance matrices Σ_k for each class separately, using a weighted average of the sample variance for class k so that $\hat{\Sigma}_k = \frac{1}{n_k-1}\sum_{i:y_i=k}(X_i-\hat{\mu}_k)(X_i-\hat{\mu}_k)^T$, and then estimate Σ by $\hat{\Sigma} = \sum_{k=1}^{K}\frac{n_k-1}{n-K}\cdot\hat{\Sigma}_k$.

$f_k(x)$ is the density function of the covariates belonging to class k, that is $f_k(x) = \Pr\{X = x|Y = k\}$.It is generally difficult to estimate $f_k(x)$ so we normally tend to make assumptions about its form. In this particular problem we are given that the covariates belonging to each class are normally distributed.

(ii) The difference between LDA and QDA is that LDA assumes that all the K classes share the same covariance matrix Σ while QDA allows for different covariance matrices for the different classes. This makes for a more flexible model which might be better to describe a more complex relationship, but also has greater risk of overfitting.

```
#LDA
diabetes.lda = lda(diabetes ~., data = train)
diabetes.lpred = predict(diabetes.lda, test)
table(predicted = diabetes.lpred$class, test = test$diabetes)
```

```
##          test
## predicted   0   1
##         0 138  30
##         1  17  47
```

```
#QDA
diabetes.qda = qda(diabetes~., data = train)
diabetes.qpred = predict(diabetes.qda, test)
table(predicted = diabetes.qpred$class, test = test$diabetes)
```

```
##          test
## predicted   0   1
##         0 131  32
##         1  24  45
```

c)

(i) In the KNN approach one uses the classification of the K nearest neighboring points, measured by Euclidian distance, to classify a new observation. The distribution is estimated non-parametrically and the new observation is classified to the most occurring class among its neighbors.

(ii) To choose the optimal value for the tuning parameter, one would have to test for different values of K and then choose the value that results in the lowest test-error. This could for example be done using cross-validation.

(iii)

```
trainMatrix = cbind(train$npreg, train$glu, train$bp, train$skin, train$bmi,
                    train$ped, train$age)
testMatrix = cbind(test$npreg, test$glu, test$bp, test$skin, test$bmi,
                   test$ped, test$age)
knn.predict = knn(train = trainMatrix, test = testMatrix, cl = train$diabetes,
                  k = 25, prob = T)
table(predicted = knn.predict, test = test$diabetes)
```

```
##          test
## predicted   0   1
##         0 144  36
##         1  11  41
```

The sensitivity is $\frac{\#\text{True Positive}}{\#\text{Positive}} = \frac{41}{41+36} \approx 0.53$ and the specificity is $\frac{\#\text{True Negative}}{\#\text{Negative}} = \frac{144}{144+11} \approx 0.93$.

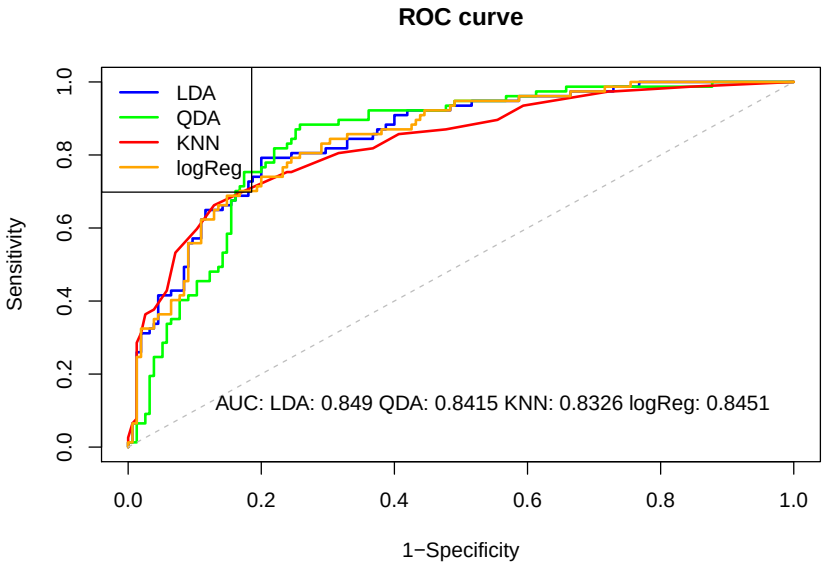
d)

```
prob.lda = diabetes.lpred$posterior[,2]
lda_roc = roc(response = test$diabetes, predictor = prob.lda, legacy.axes = TRUE)

prob.qda = diabetes.qpred$posterior[,2]
qda_roc = roc(response = test$diabetes, predictor = prob.qda, legacy.axes = TRUE)

prob.KNN = ifelse(knn.predict == 0, 1 - attributes(knn.predict)$prob,
                  attributes(knn.predict)$prob)
KNN_roc = roc(response = test$diabetes, predictor = prob.KNN, legacy.axes = TRUE)

prob.logReg = predict(logReg, newdata = test, type = "response")
logReg_roc = roc(response = test$diabetes, predictor = prob.logReg, legacy.axes = TRUE)
```



The AUC-value for the four different methods are fairly similar, but the value for LDA is the largest. The larger AUC, the better the model performs. Hence, LDA performs better than QDA, KNN and logistic regression.

An interpretable model would tell us something about how the response variable reacts to changes in the covariates and the relationship between them. In the KNN we do not obtain any information about how the covariates affect the response variable. Hence, we want to use one of the other methods. The discriminant methods are generally preferred for more than two classes, or if the classes are well separated which will cause the estimates in the logistic regression model to be unstable. This is not the case here, so then we would prefer to use logistic regression as this is the most simple model and the AUc-values are virtually the same.

Problem 4

a)

In this task, we will show that the linear regression model $Y = X\beta$ for the LOOCV statistic can be computed by the formula

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

We start with reformulation $y_{(-i)}$ in terms of $h_i = \mathbf{x}_i^T(X^TX)^{-1}\mathbf{x}_i$.

$$\begin{aligned}
&\hat{y}_{(-i)} = \mathbf{x}_i^T\hat{\beta}_{(-i)} = \mathbf{x}_i^T\left(X_{(-i)}^TX_{(-i)}\right)^{-1}X_{(-i)}^T\mathbf{y}_{(-i)} = \mathbf{x}_i^T\left(X^TX - \mathbf{x}_i\mathbf{x}_i^T\right)^{-1}\left(X^T\mathbf{y} - \mathbf{x}_iy_i\right)\\
&= \mathbf{x}_i^T\left((X^TX)^{-1} + \frac{(X^TX)^{-1}\mathbf{x}_i\mathbf{x}_i^T(X^TX)^{-1}}{1-\mathbf{x}_i^T(X^TX)^{-1}\mathbf{x}_i}\right)\left(X^T\mathbf{y} - \mathbf{x}_iy_i\right)\\
&= \left(\frac{(1-h_i)\mathbf{x}_i^T(X^TX)^{-1} + \mathbf{x}_i^T(X^TX)^{-1}\mathbf{x}_i\mathbf{x}_i^T(X^TX)^{-1}}{1-h_i}\right)\left(X^T\mathbf{y} - \mathbf{x}_iy_i\right)\\
&= \left(\frac{\mathbf{x}_i^T(X^TX)^{-1} - \mathbf{x}_i^T(X^TX)^{-1}\mathbf{x}_i\mathbf{x}_i^T(X^TX)^{-1} + \mathbf{x}_i^T(X^TX)^{-1}\mathbf{x}_i\mathbf{x}_i^T(X^TX)^{-1}}{1-h_i}\right)\left(X^T\mathbf{y} - \mathbf{x}_iy_i\right)\\
&= \frac{\mathbf{x}_i^T(X^TX)^{-1}\left(X^T\mathbf{y} - \mathbf{x}_iy_i\right)}{1-h_i} = \frac{\mathbf{x}_i^T(X^TX)^{-1}X^T\mathbf{y} - \mathbf{x}_i^T(X^TX)^{-1}\mathbf{x}_iy_i}{1-h_i}\\
&= \frac{\mathbf{x}_i^T\hat{\beta} - h_iy_i}{1-h_i} = \frac{\hat{y}_i - h_iy_i}{1-h_i}.
\end{aligned}$$

This result in $y_{(-i)} = \frac{\hat{y}_i - h_iy_i}{1-h_i}$. Hence, the mean square error for iteration i can be expressed as follows

$$\mathrm{MSE}_i = \left(y_i - \hat{y}_{(-i)}\right)^2 = \left(y_i - \frac{\hat{y}_i - h_iy_i}{1-h_i}\right)^2 = \left(\frac{(1-h_i)y_i - \hat{y}_i + h_iy_i}{1-h_i}\right)^2 = \left(\frac{y_i - \hat{y}_i}{1-h_i}\right)^2. \tag{2}$$

Thus, the LOOCV in case of linear regression can be formulated as

$$\mathrm{CV} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{MSE}_i = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i - \hat{y}_i}{1-h_i}\right)^2. \tag{3}$$

b)

i) FALSE
ii) TRUE
iii) TRUE
iv) FALSE

Problem 5

a)

```
bodyfat.lm <- lm(bodyfat ~ age + weight + bmi, data = d.bodyfat)
summary(bodyfat.lm)
```

```
##
## Call:
## lm(formula = bodyfat ~ age + weight + bmi, data = d.bodyfat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.0307  -3.8921  -0.1454   3.8896  12.6272
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -31.272668   2.807764 -11.138  < 2e-16 ***
## age           0.133170   0.028282   4.709 4.23e-06 ***
## weight        0.004075   0.058732   0.069    0.945
## bmi           1.739406   0.216723   8.026 4.54e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.34 on 239 degrees of freedom
## Multiple R-squared:  0.5803, Adjusted R-squared:  0.575
## F-statistic: 110.2 on 3 and 239 DF,  p-value: < 2.2e-16
```

The linear regression model `bodyfat.lm` has here been fitted with age, weight and BMI as predictor variables. The coefficient of determination R^2 of the model is `0.5803041`, stating that this model explains about 58% of the response's variance.

b)

(i) In the following code 1000 bootstrap samples of the R^2 is generated.

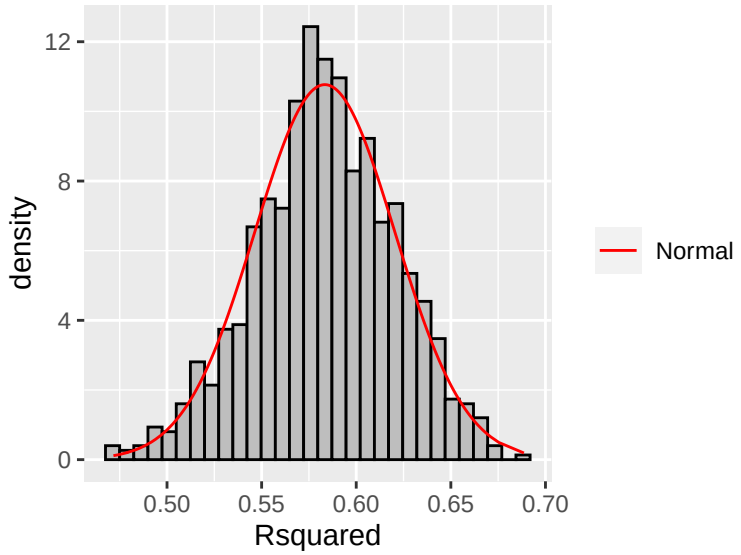
```
set.seed(4268)
B = 1000

Rsquared_func <- function(data, index){
  return(summary(lm(bodyfat ~ age + weight + bmi, data = data[index,]))$r.squared)
}

estimates = rep(NA, B)

for (b in 1:B){
  indices = sample(1:243, 243, replace = TRUE)
  thisboot = Rsquared_func(d.bodyfat, indices)
  estimates[b] = thisboot
}
```

(ii) A plot of the respective distribution of the bootstrapped R^2-values.



(iii)

```
sd(estimates)
```

```
## [1] 0.03705002
```

```
c(mean(estimates)-qnorm(0.975)*sd(estimates), mean(estimates)+qnorm(0.975)*sd(estimates))
```

```
## [1] 0.5107799 0.6560133
```

The 95% confidence interval of the R^2 is [0.5107799, 0.6560133] and the standard error is 0.03705

(iv) We observe that the confidence interval is of length 0.157. Since R^2 is a measure on the proportion of variance explained by the model, $R^2 \in [0, 1]$. This means that the confidence interval contains 15.7% of the possible values for R^2. This is a rather large interval which shows that the value of R^2 found in a) is more uncertain than immediately apparent.