Compulsory Exercise 3

Helge Bergo

22 April, 2020

Problem 1

a)

Using the `College` data set, the training and test data was preprocessed, by separating the response and predictors into an x-matrix and y-vector for each set, and then scaling the predictors.

```
y.train = college.train$Outstate
y.test = college.test$Outstate

x.train <- subset(college.train, select = -c(Outstate))
x.test <- subset(college.test, select = -c(Outstate))

mean <- apply(x.train, 2, mean)
std <- apply(x.train, 2, sd)

x.train <- as.array(scale(x.train, center = mean, scale = std))
x.test <- as.array(scale(x.test, center = mean, scale = std))
```

b)

The equation for the network to predict `Outstate`, using an input layer with the 17 predictors and a `relu` activation function for the hidden layers is:

$$\hat{y}_1(\mathbf{x}) = \beta_{01} + \sum_{m=1}^{64} \beta_{m1} \max(\sum_{l=1}^{64} \gamma_{lm} \cdot \max(\sum_{j=1}^{17} \alpha_{jl} x_j, 0), 0) \tag{1}$$

The activation function chosen for the output layer was the `linear` function, since this is a regression problem.

c)

(i)

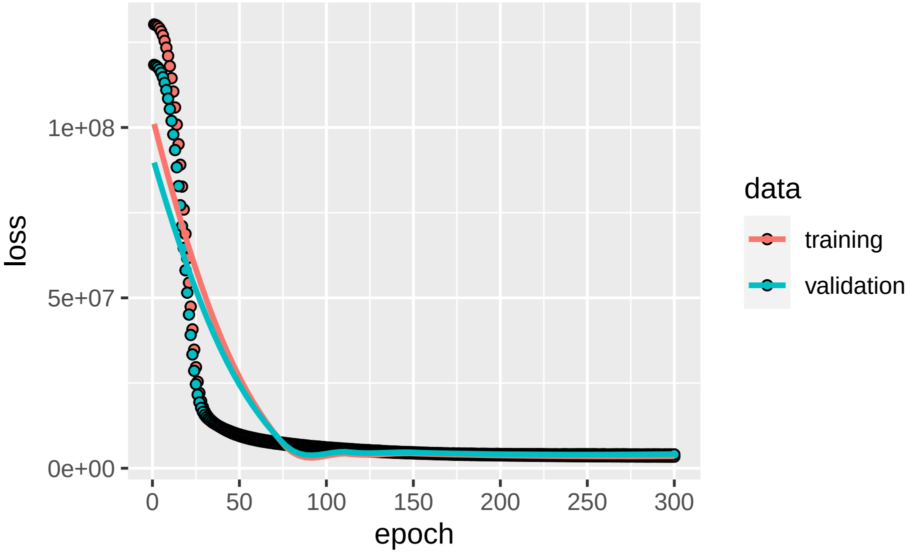
The network was trained using the `keras` library, using the chosen `linear` function for the output layer, and `mse` as the loss function.

```
model <- keras_model_sequential() %>%
  layer_dense(units = 64, activation = "relu", input_shape = c(17)) %>%
  layer_dense(units = 64, activation = "relu") %>%
  layer_dense(units = 1, activation = "linear")

model %>% compile(optimizer = "rmsprop", loss = "mse")

history <- model %>% fit(x.train, y.train, epochs = 300, batch_size = 8,
                         validation_split = 0.2)
```

(ii)

After training for 300 epochs, with 20% of the training data as the validation set, the results are plotted below.

```
plot(history)
```



As can be seen, both the training and validation error falls very quickly the first 30 epochs, and then continue to decrease slowly throughout the training.

(iii)

```
score <- model %>% evaluate(x.test, y.test)
```

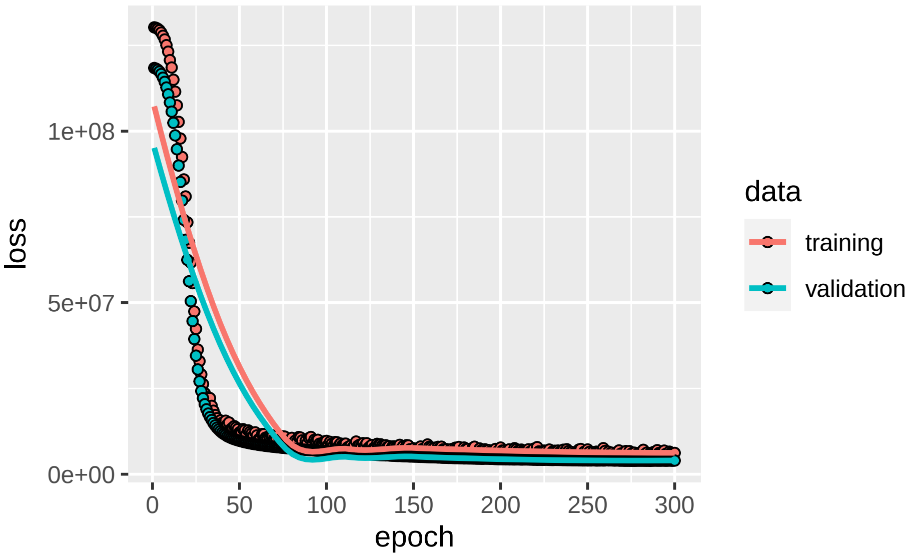
The final MSE after training the model for 300 epochs was 3.6443029×10^6. Compared to the MSE of the methods from Compulsory 2, this is a relatively good MSE score, and compares to both lasso and forward selection. It is better than polynomial regression and smoothing splines, but both bagging and random forest beats it, scoring 3.3×10^6 and 2.6×10^6 respectively.

d)

Both dropout and weight decay was tried out for improving the performance of the network.

```
model_reg <- keras_model_sequential() %>%
  layer_dense(units = 64, activation = "relu", input_shape = c(17),
              kernel_regularizer = regularizer_l2(l = 0.001)) %>%
  layer_dropout(0.3) %>%
  layer_dense(units = 64, activation = "relu") %>%
  layer_dropout(0.3) %>%
  layer_dense(units = 1, activation = "linear")

model_reg %>% compile(optimizer = "rmsprop", loss = "mse")

history_reg <- model_reg %>% fit(x.train, y.train, epochs = 300, batch_size = 8,
                                 validation_split = 0.2)
```

```
plot(history_reg)
```



```
score_reg <- model_reg %>% evaluate(x.test, y.test)
```

After implementing 30% dropout for the two hidden layers, and L2 regularization for the first hidden layer, the final MSE after training was 3.4367589×10^6, so notably lower than the unimproved network, but still not better than random forest, for example.

Problem 2 (10P)

Load your data into R using the following code:

```
id <- "1CA1RPRYqU9oTIaHfSroitnWrI6WpUeBw" # google file ID
d.corona <- read.csv(sprintf("https://docs.google.com/uc?id=%s&export=download", id), header = T)
```

a) Inspecting your data (1P)

Table 1: Number of deceased per country.

country	n
France	14
indonesia	2
japan	3
Korea	26

Table 2: Number of deceased per sex.

sex	n
female	14
male	31

Table 3: Number of deceased per country, separated by gender.

country	male	female
France	9	5
japan	3	0
indonesia	1	1
Korea	18	8

b) Multiple choice

FALSE, FALSE, , TRUE (i) Country is not a relevant variable in the model.

(ii) The slope for indonesia has a large p-value, which shows that we should remove the Indonesian population from the model, as they do not fit the model as well as the Japanese population.

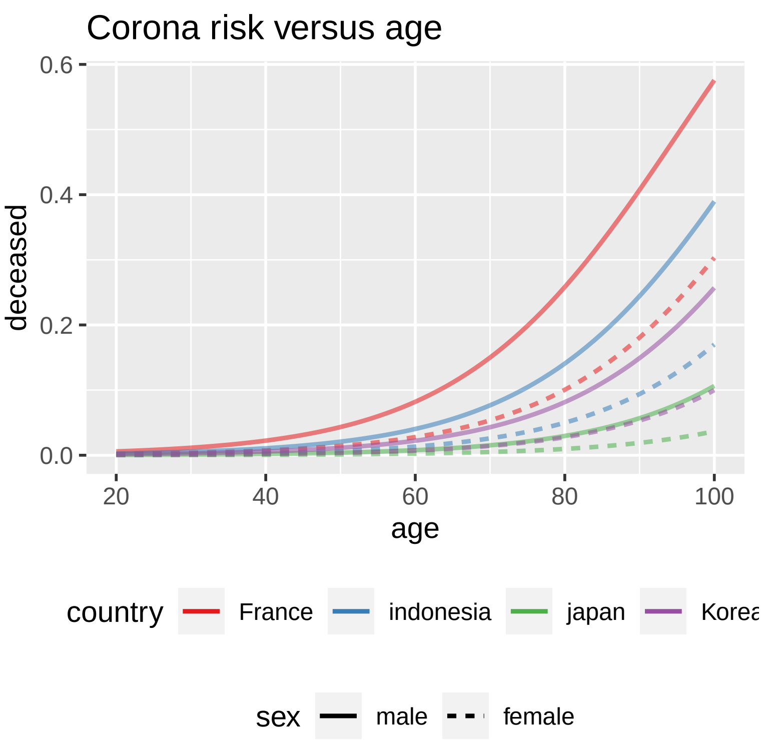
(iii) Increasing the age by 10 years, $x^*_{age} = x_{age} + 10$, and holding all other covariates constant, the odds ratio to die increases by a factor of 1.97.

(iv) The probability to die is approximately 3.12 larger for males than for females.

```
glm_model <- d.corona %>%
  glm(deceased ~., data = ., family = 'binomial')

gender_ratio <- exp(glm_model$coefficients[1] + glm_model$coefficients[2]) / glm_model$coefficients[1]
gender_ratio <- exp(glm_model$coefficients[2])

age_ratio <- mean((glm_model$coefficients[1] + glm_model$coefficients[3] *
                     seq(30,100,10)) /
                    (glm_model$coefficients[1] + glm_model$coefficients[3] *
                       seq(20,90,10)))
```

c)

```
newdat = expand.grid(age = seq(20,100,1), sex = c('male','female'), country = unique(d.corona$country))
newdat$pred = predict(glm_model, newdata = newdat, type = 'response')

d.corona %>%
  ggplot(aes(x = age, y = deceased, colour = country, lty = sex, alpha = 0.8)) +
  geom_line(data = newdat, aes(y = pred), size = 0.8) +
  labs(title = "Corona risk versus age")  +
  theme(legend.position = "bottom", legend.box = 'vertical') +
  scale_alpha(guide = 'none') +
  scale_color_brewer(palette = "Set1")
```



d) (3P)

As a statistician working on these data, you are asked the following questions:

Answer the questions by fitting appropriate models (1P each).

i)

(i) Have males generally a higher probability to die of coronavirus than females?

```
d.corona %>%
  lm(deceased ~ sex, data = .) %>%
  summary
```

```
##
## Call:
## lm(formula = deceased ~ sex, data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.03366 -0.03366 -0.01286 -0.01286  0.98714
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.012856   0.004474   2.873  0.00411 **
## sexmale     0.020803   0.006610   3.147  0.00167 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1477 on 2008 degrees of freedom
## Multiple R-squared:  0.004909,   Adjusted R-squared:  0.004413 
## F-statistic: 9.905 on 1 and 2008 DF,  p-value: 0.001672
```

ii)

(ii) Is age a greater risk factor for males than for females?

```
d.corona %>%
  lm(deceased ~ sex*age, data = .) %>%
  summary
```

```
## 
## Call:
## lm(formula = deceased ~ sex * age, data = .)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.12013 -0.03801 -0.01709  0.00596  0.98834 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -0.0262976  0.0108511  -2.423 0.015460 *  
## sexmale     -0.0261346  0.0155259  -1.683 0.092475 .  
## age          0.0007747  0.0001964   3.944 8.29e-05 ***
## sexmale:age  0.0009684  0.0002826   3.427 0.000622 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.1445 on 2006 degrees of freedom
## Multiple R-squared:  0.04728,    Adjusted R-squared:  0.04585 
## F-statistic: 33.18 on 3 and 2006 DF,  p-value: < 2.2e-16
```

iii)

(iii) Is age a greater risk factor for the French population than for the Korean population?

```
d.corona %>%
  lm(deceased ~ country*age, data = .) %>%
  summary
```

```
## 
## Call:
## lm(formula = deceased ~ country * age, data = .)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.30767 -0.03092 -0.01500  0.00221  0.99001 
## 
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)    
## (Intercept)          -0.1974974  0.0379122  -5.209 2.09e-07 ***
## countryindonesia      0.2149724  0.0607184   3.540 0.000409 ***
## countryjapan          0.1936865  0.0444551   4.357 1.39e-05 ***
## countryKorea          0.1667403  0.0388456   4.292 1.85e-05 ***
## age                     0.0051027  0.0005656   9.021  < 2e-16 ***
## countryindonesia:age -0.0048704  0.0010568  -4.609 4.30e-06 ***
## countryjapan:age     -0.0048517  0.0006862  -7.071 2.12e-12 ***
## countryKorea:age     -0.0041079  0.0005877  -6.990 3.73e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1419 on 2002 degrees of freedom
## Multiple R-squared:  0.08351,    Adjusted R-squared:  0.08031
## F-statistic: 26.06 on 7 and 2002 DF,  p-value: < 2.2e-16
```

e) Interpret your model (1P)

According to your model fitted in part b), it looks like the French population is at a much higher risk of dying from Covid-19 than the other countries. Do you trust this result? How could it be influenced by the way the data were collected?

f) Multiple choice (2P)

Which of the following statements are true, which false?

Consider the classification tree below to answer: TRUE, TRUE, ,

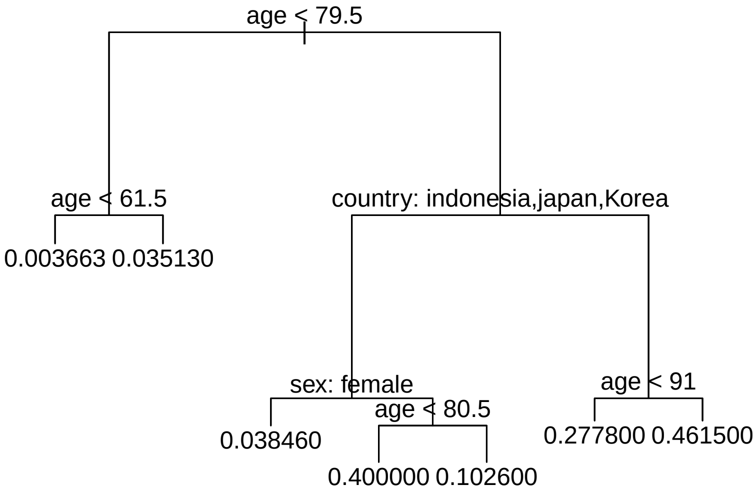
(i) The probability of dying (`deceased = 1`) is about 0.46 for a French person with age above 91.

(ii) Age seems to be a more important predictor for mortality than sex.

Consider the LDA code and output below:

(iii) The "null rate" for misclassification is 2.24%, because this is the proportion of deaths among all cases in the dataset. No classifier should have a higher misclassification rate.

(iv) LDA is not a very useful method for this dataset, among other reasons because it does not estimate probabilities, but also because the misclassification error is too high.

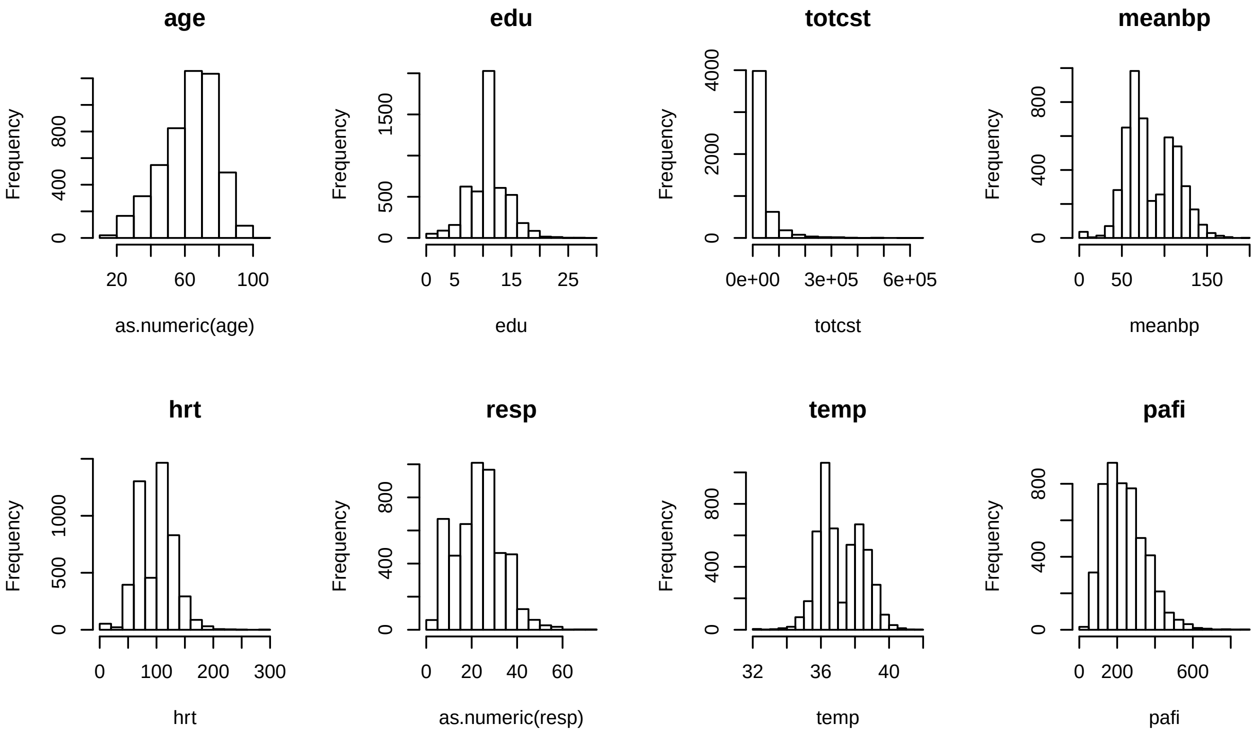


```
library(MASS)
table(predict = predict(lda(deceased ~ age + sex + country, data = d.corona))$class, true = d.corona$de
```

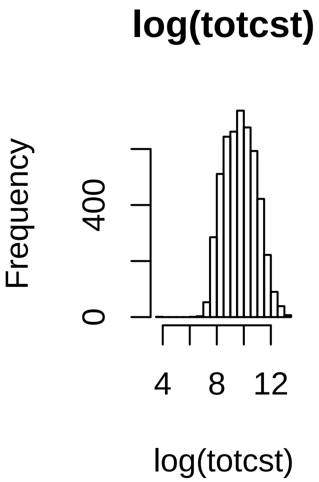
```
##        true
## predict    0    1
##        0 1926   31
##        1   39   14
```

Problem 3 (14P)

a)



A fitting transformation of the `totcst` variable is `log(totcst)`, as shown below.



b) (3P)

Fit a multiple linear regression model with the six covariates `age`, `temp`, `edu`, `resp`, `num.co` and `dzgroup` and the (transformed version of the) response `totcst`.

```
d.support$num.co <- as.integer(d.support$num.co)
mlr_model <- lm(log(totcst) ~ age + temp + edu + resp + num.co + dzgroup, data = d.support)
```

(i)

```
age_increase_cost <- exp(mean((mlr_model$coefficients[1] + mlr_model$coefficients[2] *
                      seq(30,100,10)) /
                     (mlr_model$coefficients[1] + mlr_model$coefficients[2] *
                        seq(20,90,10))))
```

The avereage total costs increase by a factor of 2.69 when the patient's age is increased by 10 years.

(ii)

(ii) Do a residual analysis using the Tukey-Anscombe plot and the QQ-diagram. Are the assumptions fulfilled? (1P)

```
par(mfrow = c(2,2))
# plot(mlr_model)
library(ggfortify)
# autoplot(lm(totcst ~ age + temp + edu + resp + num.co + dzgroup, data = d.support))
# autoplot(lm(log(totcst) ~ age + temp + edu + resp + num.co + dzgroup, data = d.support))
```

(iii)

(iii) Does the effect of age depend on the disease group? Do a formal test and report the *p*-value. (1P)

```
age_disease_model <- lm(log(totcst) ~ age*dzgroup, data = d.support)
summary(age_disease_model)
```

```
##
## Call:
## lm(formula = log(totcst) ~ age * dzgroup, data = d.support)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7205 -0.6581 -0.0453  0.6140  3.5211
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             10.756350   0.072578 148.203  < 2e-16 ***
## age                     -0.005879   0.001148  -5.123 3.12e-07 ***
## dzgroupCHF              -1.172289   0.189892  -6.173 7.22e-10 ***
## dzgroupCirrhosis        -0.578046   0.259700  -2.226 0.026071 *
## dzgroupColon Cancer     -1.958877   0.542087  -3.614 0.000305 ***
## dzgroupComa              0.294489   0.218542   1.348 0.177876
## dzgroupCOPD             -1.648502   0.253519  -6.502 8.68e-11 ***
## dzgroupLung Cancer      -1.877570   0.300034  -6.258 4.23e-10 ***
## dzgroupMOSF w/Malig      0.398847   0.199195   2.002 0.045308 *
## age:dzgroupCHF          -0.005322   0.002794  -1.904 0.056907 .
## age:dzgroupCirrhosis    -0.007375   0.004669  -1.580 0.114285
## age:dzgroupColon Cancer  0.007503   0.008338   0.900 0.368258
## age:dzgroupComa         -0.011290   0.003352  -3.368 0.000763 ***
## age:dzgroupCOPD          0.004215   0.003635   1.159 0.246310
## age:dzgroupLung Cancer   0.002645   0.004780   0.553 0.579989
## age:dzgroupMOSF w/Malig -0.010919   0.003249  -3.361 0.000784 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9414 on 4931 degrees of freedom
## Multiple R-squared:  0.3714, Adjusted R-squared:  0.3695
## F-statistic: 194.2 on 15 and 4931 DF,  p-value: < 2.2e-16
```

```
mlr_model_age <- lm(log(totcst) ~ age + dzgroup, data = d.support)
mlr_model_age_intercept <- lm(log(totcst) ~ age + dzgroup + age*dzgroup, data = d.support)
# anova(mlr_model_age, mlr_model_age_intercept)
```

c)

The training and test set was created, and made into a data matrix, to use the `glmnet` package.

```
library(glmnet)
set.seed(12345)
d.support <- read_csv(sprintf("https://docs.google.com/uc?id=%s&export=download", id))

train.ind = sample(1:nrow(d.support), 0.8 * nrow(d.support))
d.support.train = d.support[train.ind, ]
d.support.test = d.support[-train.ind, ]
x.train = model.matrix(log(totcst) ~ ., data = d.support.train)[,-1]
y.train = log(d.support.train$totcst)
x.test = model.matrix(log(totcst) ~ ., data = d.support.test)[,-1]
y.test = log(d.support.test$totcst)
```

Cross-validation was run, to find the largest λ within 1 standard error of the smallest λ.

```
ridge_model = cv.glmnet(x.train, y.train, alpha = 0)
best_lambda = ridge_model$lambda.1se
```

The value of λ was 0.186, which was then used to find the MSE of the ridge regression.

```
ridge_pred = predict(ridge_model, s = best_lambda, newx = x.test)
ridge_MSE = mean((ridge_pred - y.test)^2)
```

The final calculated MSE is 0.922.

d) (3P)

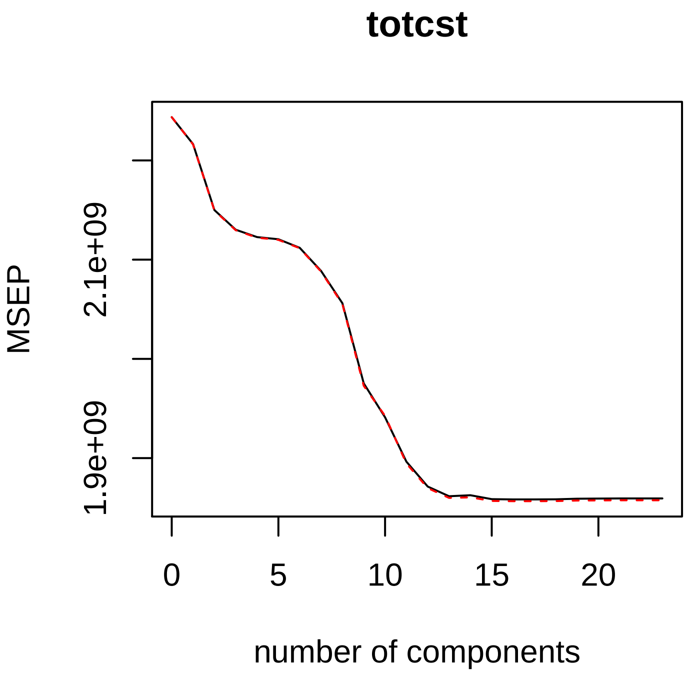
Now assume that our sole aim is prediction. In the course you heard about *partial least squares (PLS)*. It is a smart approach that uses the principal component regression idea, but finds the components that are best correlated with the response.

Proceed as follows:

(i)

(i) Run a PLS regression (don't forget to scale the variables, `scale=TRUE`) (1P).

```
library(pls)
# mean <- apply(x.train, 2, mean)
# std <- apply(x.train, 2, sd)
#
# x.train <- as.array(scale(x.train, center = mean, scale = std))
plsr_model <- plsr(totcst ~ ., data = d.support.train, scale = FALSE, validation = "CV")
# plsr_model <- plsr(totcst~., data = d.support.train, scale = TRUE, validation = "CV")
validationplot(plsr_model, val.type = "MSEP")
```



```
plsr_predictions = predict(plsr_model, d.support.test, ncomp = 3)
plsr_square_errors <- as.numeric((plsr_predictions - d.support.test$totcst)^2)
squared_errors <- data.frame(plsr_square_errors = plsr_square_errors)
mean(plsr_square_errors)
```

```
## [1] NA
```

(ii)

(ii) Choose an optimal number of principal components (PCs) using cross-validation (1P).

(ii)

(iii) Report the MSE of the test set when using the respective set of PCs and compare to the result from ridge regression. Conclusion? (1P)

e) (4P)

Now choose two other methods that you know from the course and try to build models with even lower test MSEs than those found so far (imagine that this is a competition where the lowest test MSE wins). Use the same training and test dataset as generated above. And remember that we are still *always* working with the transformed version of the response variable (`totcst`). In particular, use

(i) One model that involves non-linear transformations of the covariates (e.g., splines, natural splines, polynomials etc) that are combined to a GAM (2P).
(ii) One model/method based on regression trees (2P).

Very briefly discuss or explain your choices (1-2 sentences each).

Problem 4 (Mixed questions; 6P)

a) 2P

We look at the following cubic regression spline model:

$$Y = \begin{cases} \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon \ , & \text{if } \ x \leq 1 \ , \\ \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x-1)^3 + \epsilon \ , & \text{if } \ 1 < x \leq 2 \ , \\ \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x-1)^3 + \beta_5 (x-2)^3 + \epsilon \ , & \text{if } \ x > 2 \ . \end{cases}$$

Write down the basis functions (1P) and the design matrix (1P) of this model.

b) Multiple choice - 2P

Inference vs prediction: Which of the following methods are suitable when the aim of your analysis is inference?

(i) Lasso and ridge regression
(ii) Multiple linear regression with interaction terms
(iii) Logistic regression
(iv) Support Vector Machines

c) Multiple choice - 2P

We again look at the Covid-19 dataset from Problem 2 to study some properties of the bootstrap method. Below we estimated the standard errors of the regression coefficients in the logistic regression model with `sex`, `age` and `country` as predictors using 1000 bootstrap iterations (column `std.error`). These standard errors can be compared to those that we obtain by fitting a single logistic regression model using the `glm()` function. Look at the R output below and compare the standard errors that we obtain from these two approaches (note that the `t1*` to `t6*` variables are sorted in the same way as for the `glm()` output).

```
# library(boot)
# boot.fn <- function(data,index){
#   return(coefficients(glm(deceased ~ sex + age + country, family="binomial",data=data,subset=index)))
# }
# boot(d.corona,boot.fn,1000)
```

```
# Logistic regression
# r.glm <- glm(deceased ~ sex + age + country, d.corona,family="binomial")
# summary(r.glm)$coef
```

Which of the following statements are true?

(i) There are large differences between the estimated standard errors, which indicates a problem with the bootstrap.
(ii) The differences between the estimated standard errors indicate a problem with the assumptions taken about the distribution of the estimated parameters in logistic regression.
(iii) The `glm` function leads to too small p-values for the differences between countries, in particular for the differences between Indonesia and France and between Japan and France.
(iv) The bootstrap relies on random sampling the same data without replacement.

Problem 5 (Multiple and single choice questions; 11P)

a) Multiple choice - 2P

Which of the following are techniques for regularization?

(i) Lasso
(ii) Ridge regression
(iii) Forward and backward selection
(iv) Stochastic gradient descent

b) Multiple choice - 2P

Which of the following statements about principal component regression (PCR) and partial least squares (PLS) are correct?

(i) PCR involves the first principal components that are most correlated with the response.
(ii) PLS involves the first principal components that are most correlated with the response.
(iii) The idea in PLS is that we choose the principal components that explain most variation among all covariates.
(iv) The idea in PCR is that we choose the principal components that explain most variation among all covariates.

c) Single choice - 1P

In ridge regression, we estimate the regression coefficients in a linear regression model by minimizing

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j-1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \ .$$

What happens when we increase λ from 0? Choose the single correct statement:

(i) The training RSS will steadily decrease.
(ii) The test RSS will steadily decrease.
(iii) The test RSS will steadily increase.
(iv) The bias will steadily increase.
(v) The variance of the estimator will steadily increase.

d) Single choice - 1P

Which statement about the *curse of dimensionality* is correct?

(i) It means that we have a bias-variance tradeoff in K-nearest neighbor regression, where large K leads to more bias but less variance for the predictor function.
(ii) It means that the performance of the K-nearest neighbor classifier gets worse when the number of predictor variables p is large.
(iii) It means that the K-means clustering algorithm performs bad if the datapoints lie in a high-dimensional space.
(iv) It means that support vector machines with radial kernel function should be avoided, because radial kernels correspond to infinite-dimensional polynomial boundaries.
(v) It means that we should never measure too many covariates when we want to do classification.

e) Single choice - 1P

Now assume you have 10 covariates, X_1 to X_{10}, each of them uniformly distributed in the interval $[0, 1]$. To predict a new test observation $(X_1^{(0)}, \ldots, X_{10}^{(0)})$ in a K-nearest neighbor (KNN) clustering approach, we use all observations within 20% of the range closest to each of the covariates (that is, in each dimension). Which proportion of available (training) observations can you expect to use for prediction?

(i) $1.02 \cdot 10^{-7}$
(ii) $2.0 \cdot 10^{-3}$
(iii) 0.20
(iv) 0.04
(v) 10^{-10}

f) Multiple choice - 2P

This example is taken from a real clinical study by *Ikeda, Matsunaga, Irabu, et al. Using vital signs to diagnose impaired consciousness: cross sectional observational study. BMJ 2002;325:800.* Researchers investigated the use of vital signs as a screening test to identify brain lesions in patients with impaired consciousness. The setting was an emergency department in Japan. The study included 529 consecutive patients that arrived with consciousness. Patients were followed until discharge. The vital signs of systolic and diastolic blood pressure and pulse rate were recorded on arrival. The aim of this study was to find a quick test for assessing whether the newly arrived patient suffered from a brain lesion. While vital signs can be measured immediately, the actual diagnosis of a brain lesion can only be determined on the basis of brain imaging and neurological examination at a later stage, thus the quick measurements of blood pressure and heart rate are important to make a quick assessment. In total, 312 patients (59%) were diagnosed with a brain lesion.

The performance of each vital sign (systolic blood pressure, diastolic blood pressure and heart rate) was separately evaluated as a screening test to quickly diagnose brain lesions. To assess the quality of each of these vital signs, different thresholds were taken successively to discriminate between “negative” and “positive” screening test result. For each vital sign and each threshold the sensitivity and specificity were derived and used to plot a receiver operating characteristic (ROC) curve for the vital sign (Figure 1):

Which of the following statements are true?

(i) The value of 1-specificity represents the proportion of patients without a diagnosed brain lesion identified as positive on screening.
(ii) When we use different cut-offs, sensitivity increases at the cost of lower specificity, and vice versa.
(iii) A perfect diagnostic test has an AUC of 0.5.
(iv) The vital sign that is most suitable to distinguish between patients with and without brain lesion is systolic blood pressure.

g) Multiple choice

TRUE, FALSE, TRUE, TRUE