Trial exam STK2100 - spring 2021

Geir Storvik

Spring 2021

Problem 1

In this exercise we will look at at a dataset `frogs` where the variable of interest is the presence of a specific type of frogs at different locations (0/1 variable where 1 corresponds to presence). There are 9 explanatory variables, all numerical. We denote the data set by $\{(y_i, \boldsymbol{x}_i), i = 1, ..., n\}$

A fit to a logistic regression model gave the following result:

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.635e+02  2.153e+02  -0.759  0.44764
northing     1.041e-02  1.654e-02   0.630  0.52901
easting     -2.158e-02  1.268e-02  -1.702  0.08872
altitude     7.091e-02  7.705e-02   0.920  0.35745
distance    -4.835e-04  2.060e-04  -2.347  0.01893
NoOfPools    2.968e-02  9.444e-03   3.143  0.00167
NoOfSites    4.294e-02  1.095e-01   0.392  0.69482
avrain      -4.058e-05  1.300e-01   0.000  0.99975
meanmin      1.564e+01  6.479e+00   2.415  0.01574
meanmax      1.708e+00  6.809e+00   0.251  0.80198
```

The log-likelihood value for this fit was -97.83.

The following table shows the confusion matrix for the observed y_i's:

$y_i\backslash\hat{y}_i$	0	1
0	113	20
1	24	55

(a) Based on the results from the fit for the logistic regression model, is this a model you find satisfactory? Also include an argument to your answer.

An alternative model also based on logistic regression gave the following result:

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.916e+01  1.611e+01  -4.293 1.76e-05
easting     -9.236e-03  4.479e-03  -2.062  0.03921
altitude     3.217e-02  8.049e-03   3.997 6.41e-05
distance    -5.099e-04  1.837e-04  -2.776  0.00550
NoOfPools    2.969e-02  9.091e-03   3.266  0.00109
meanmin      8.916e+00  2.030e+00   4.391 1.13e-05
```

with a log-likelihood value of -98.71. The corresponding confusion matrix was in this case

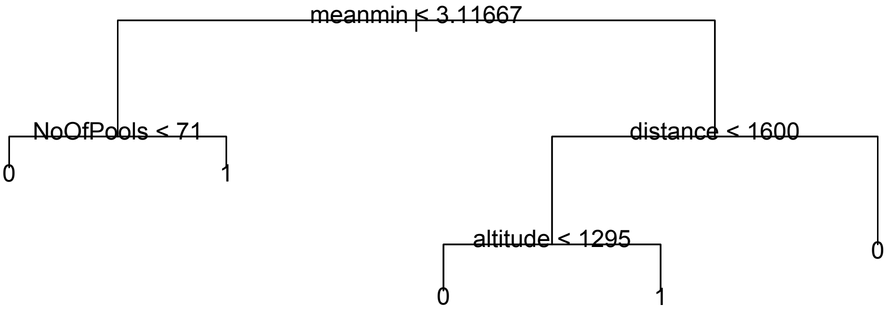
$y_i \backslash \hat{y}_i$	0	1
0	112	21
1	24	55

(*b*) Explain why the log-likelihood value for this new model is lower than for the first model.

Use these log-likelihood values for making a choice between the two models. Specify which criterion you use for this choice.

(*c*) Explain what the P-values in the last column in the two tables mean. Discuss the actual values given for the two models. Also give a possible explanation to why some of the P-values are quite different for the same explanatory variable for the two models.

We will now look at classification trees. The plot below shows an estimated tree based on 5 leaves:



The table below gives the confusion matrix obtained in this case:

$y_i \backslash \hat{y}_i$	0	1
0	117	16
1	24	55

(*d*) Explain why a likelihood function for a classification tree can be written in the form

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{1-y_i}$$

where $p_i = c_m$ for $\boldsymbol{x}_i \in R_m$.

(*e*) For the specific classification tree we obtained a log-likelihood value equal to -90.21. What is the number of parameters in this case?

Use this to calculate the AIC value for the classification tree and use this to make an evaluaton of this model compared to previous models.

Consider now instead cross-validation. The table below give the confusion matrices for logistic regression with 5 explanatory variables and for the classification tree with 5 leaves. Here, leave-one-out cross-validation is used.

	Logistic regression		Classification tree	
$y_i\backslash\hat{y}_i$	0	1	0	1
0	109	24	100	33
1	24	55	24	55

(*f*) Explain how the cross-validation method work and also why the table for logistic regression now is different from the one you saw earlier.

Based on these new confusion matrices, which method would you prefer?

Now look at bagging. The confusion matrices below are based on *out-of-bag* estimation.

	Logistic regression		Classification tree	
$y\backslash\hat{y}$	0	1	0	1
0	109	24	115	18
1	24	55	32	47

Problem 2

Consider now a linear regression model

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \varepsilon$$

Assume $\boldsymbol{\beta} = (\beta_0, ..., \beta_p)$ is estimated by the criterion

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

(*a*) What is this method called? Discuss this method compared to ordinary least squares.

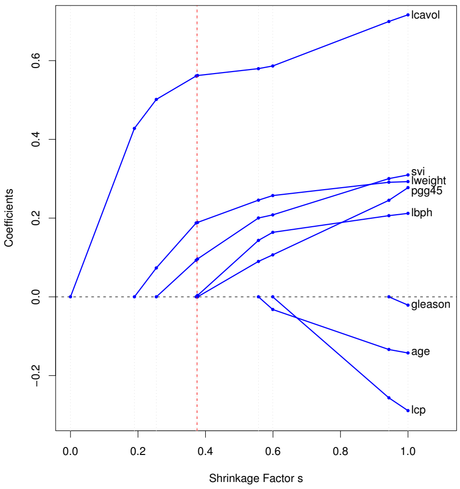
(*b*) Find an explicit expression for $\hat{\boldsymbol{\beta}}$. If you make some simplifying assumptions, clearly state these.

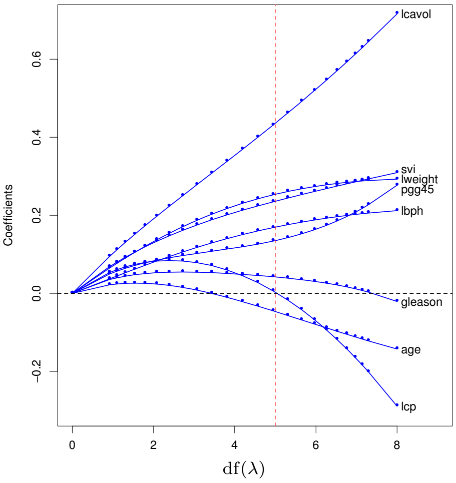
An alternative way of estimating $\boldsymbol{\beta}$ is through the criterion

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

(c) What is this method called? Discuss this method in relation to both ordinary least squares and the method above.

Below are two plots showing the estimates of $\boldsymbol{\beta}$ on a prostate cancer example. Relate the two plots to the two methods discussed above. Also explain how the least squares estimates can be read from the plot(s).





(d) In both cases, the penalty variable λ needs to be specified. Discuss methods for doing that.

Problem 3

In this exercise, you will first be introduced to a problem about traffic related air pollution, and then you will be asked to interpret a GAM plot.

Particles between 2.5 and 10 micrometers in size are called coarse particles. At or beside a road in a Norwegian city, there will typically be a lot of road dust containing many coarse particles, especially in the winter when many cars are fitted with studded tyres (piggdekk). These particles are typically whirled into the air by the cars. In addition to the road dust which is re-suspended into the air, the exhaust from the vehicles also gives direct emissions of coarse particles (in addition to emissions of smaller particles and gases).

A large data set has been analysed to give a description of how the concentration of coarse particles are related to the traffic volume and meteorological conditions. This data set consists of 70 000 hourly values of the concentration of coarse particles and corresponding explanatory variables in the period 2001-2009 at Kirkeveien in Oslo. The speed limit at Kirkeveien is 50 km per hour. The variables used in this analysis includes

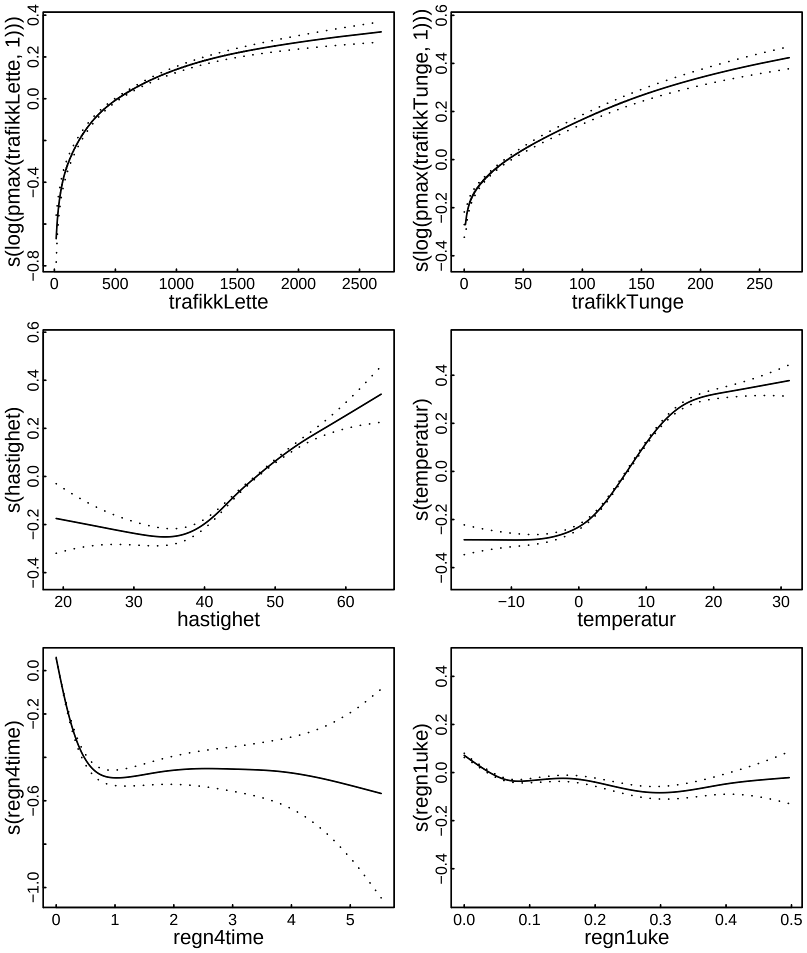
- y - the (natural) logarithm of the concentration of coarse particles in the air
- x_1 = the number of light vehicles (shorter or equal to 5.5 m) per hour = "trafikkLette"
- x_2 = the number of heavy vehicles (longer than 5.5 m) per hour = "trafikkTunge"

- x_3 = average velocity of the vehicles = "hastighet"
- x_4 = temperature in degrees Celsius = "temperatur"
- x_5 = accumulated precipitation (in mm) last 4 hours = "regn4time"
- x_6 = accumulated precipitation (in mm) last week = "regn1uke"
- other explanatory variables that you can ignore in this exercise

The following generalised additive model (GAM) has been fitted to the data

$$y = s_1(x_1) + s_2(x_2) + s_3(x_3) + s_4(x_4) + s_5(x_5) + s_6(x_6) + \cdots + \varepsilon.$$

Here, $+\ldots$ mean that also some other explanatory variables are included in the model. The figure below shows the GAM plot for x_1-x_6.



(a) Give a short interpretation of the estimated effect of each of the six explanatory variables, i.e. describe how they may affect the concentration of coarse particles, and discuss whether the results are reasonable in light of your understanding of the physical process of this type of air pollution.

(*b*) Until a few years ago, the speed limit at some of the main roads in Oslo was lowered from 80 km/hour to 60 km/hour during the winter months. Use the GAM plot above to quantify what effect this intervention might have had on the concentration on coarse particles near these roads.

Problem 4

Consider the general neural network model (with one hidden layer) given by

$$\begin{aligned} z_{im} =& f_0(\alpha_{0m} + \boldsymbol{\alpha}_m^T \boldsymbol{x}_i) \qquad m = 1, ..., M \\ \eta_i =& f_1(\gamma_0 + \boldsymbol{\gamma}^T \boldsymbol{z}_i) \\ y_i \overset{ind}{\sim}& N(\eta_i, \sigma^2) \end{aligned} \tag{*}$$

where $\overset{ind}{\sim}$ means that the observations are independent. Here $\boldsymbol{x} \in \mathcal{R}^p$ while $y \in \mathcal{R}$. We assue that estimation of parameters are based on the least squares measure $\sum_{i=1}^N (y_i - \eta_i)^2$, perhaps including some regularization term.

(*a*) Consider first the case where $M = 1$ and $f_0(x) = f_1(x) = x$.

Derive an expression for η_i in this case.

Explain why this corresponds to a linear regression model.

Is it possible to estimate all the parameters involved?

(*b*) Consider now the case where $M > 1$ but still with $f_0(x) = f_1(x) = x$. Discuss similarities and differences compared to linear regression based on principal components.

(*c*) We will now consider the more general case where $f_0(\cdot)$ is the sigmoid function ($f_0(x) = \exp(x)/(1 + \exp(x))$) while $f_1(\cdot)$ is still the identify function.

Write down η_i as a function of $\boldsymbol{x}_i$ in this case.

Argue why one can use gradient methods for minimizing the least squares measure with respect to the unknown parameters.

Will you expect some identifiability problems also in this case?

(*d*) A common way to regularize parameters in neural networks is to introduce a penalty term so that one minimizes

$$L(\boldsymbol{\theta}) = \sum_{i=1}^N (y_i - \eta_i)^2 + \lambda \left[\sum_{j=1}^q \theta_j^2 \right]$$

where q is the total number of parameters and θ_j is a specific parameter involved.

Explain why the penalty term is in particular useful for neural networks.

We will now consider the NASA data set comprising different size NACA 0012 airfoils at various wind tunnel speeds and angles of attack. The span of the airfoil and the observer position were the same in all of the experiments.

The data, which is downloaded from `https://archive.ics.uci.edu/ml/datasets/Airfoil+Self-Noise` has $N = 1503$ instances and 5 input variables:

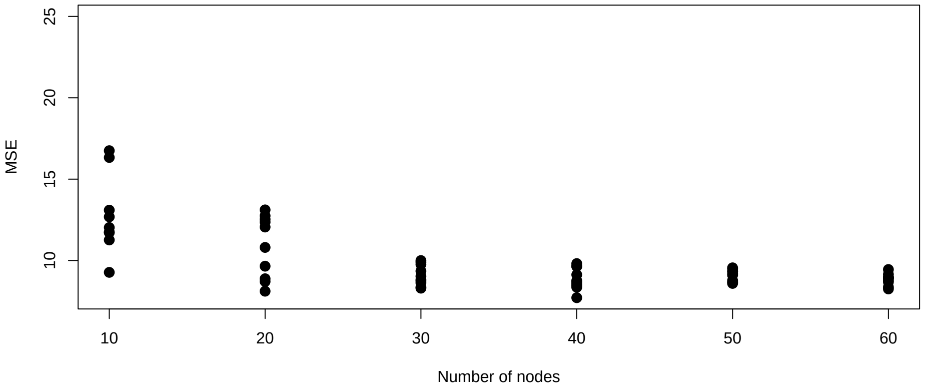
- Frequency, in Hertzs.
- Angle of attack, in degrees.
- Chord length, in meters.
- Free-stream velocity, in meters per second.
- Suction side displacement thickness, in meters.

The only output is:

- Scaled sound pressure level, in decibels.

A subset of $n_{tr} =$ is used for fitting a model, the remaining data are used for evaluation.

The plot below shows results for fitting neural networks with one hidden layer and varying number of hidden nodes. For each number, the procedure is repeated 10 times and the points shows mean squared errors on the evaluation set for all repetitions and all number of nodes. The minimum value obtain was in this case 7.71, obtained with 40 hidden nodes. Note that for 10 hidden nodes, one run gave an error of 48.7, but the y-axis is here truncated in order to visualize the rest of the points better.



(e) Why do we get different results when repeating the fit with the same number of hidden nodes?

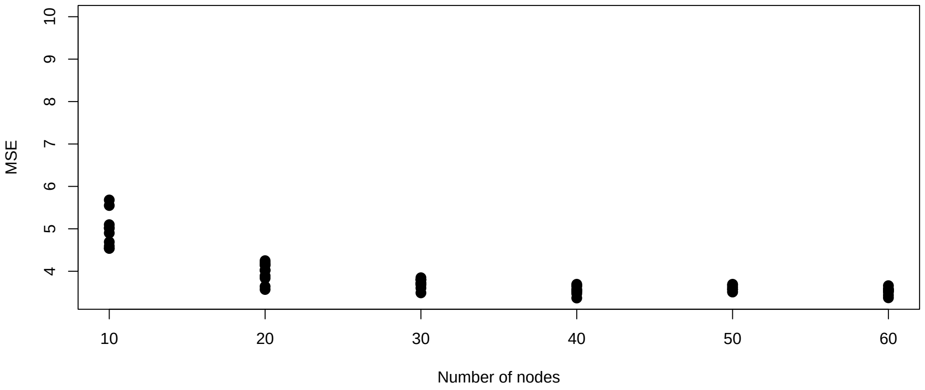
(f) Based on the plot above, how many number of nodes would you choose in your network?

If you need to report the accuracy on your final prediction method, how would you proceed for doing that?

The plot below shows similar results, but now based on that each covariate x_{ij} is transformed by

$$x_{ij} \to \frac{x_{ij} - \bar{x}_{\cdot j}}{\sqrt{N^{-1}\sum_{i'=1}^{N}(x_{i'j} - \bar{x}_{\cdot j})^2}}.$$

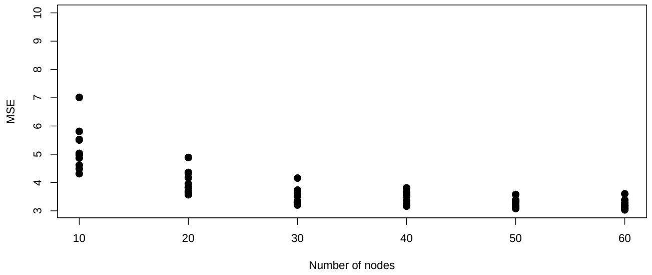
The minimum value obtained was in this case 3.37, now with 40 hidden nodes.

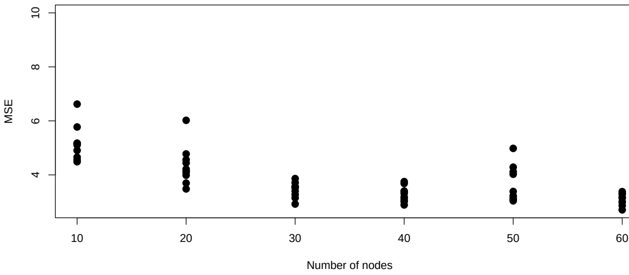


(g) Based on the equations (*), argue why the model should in principle not depend on any scaling or any linear transformations of the input variables $\boldsymbol{x}_i$.

Argue then why one *do* get different results when training on different (linear) transformations of the input variables.

The previous plots were based on a penalty parameter $\lambda = 0.5$. The two plots below shows similar results for $\lambda = 0.25$ (left) and $\lambda = 0.1$ (right). In both cases, the input variables are scaled. For $\lambda = 0.25$ the minimum MSE value 3.03, obtained with 60 hidden nodes while for $\lambda = 0.1$ the minimum MSE value 2.71, also obtained with 60 hidden nodes





The table below gives the standard deviations (over the 10 repetitions) of MSE for the different values of λ and number of hidden nodes.

	Number of hidden nodes					
	10	20	30	40	50	60
$\lambda = 0.5$	0.402	0.234	0.107	0.097	0.065	0.079
$\lambda = 0.25$	0.794	0.413	0.299	0.218	0.151	0.172
$\lambda = 0.1$	0.636	0.705	0.276	0.284	0.673	0.210

(h) Do you find the table of the standard errors to be reasonable?

Based on these results, but also on how the penalty term enters the loss function, discuss the role of the decay parameter here.

Problem 5

Consider a general situation where you for $i = 1, ..., N$ have observed inputs $\boldsymbol{x}_i \in \mathcal{R}^p$ and outputs y_i where y_i is either numerical or categorical. You want to use the data to fit a model that predicts future Y's.

(a) It is common to divide a dataset into a *training set* and a *test set* and sometimes also a *validation set.* Discuss the role of these sets and the advantages/disadvantages in doing such a split of the dataset.

(b) Explain what we mean about cross-validation. Discuss its use and how this method relates to the training/validation/test sets.

Problem 6

Assume $Y = f(x) + \varepsilon$ where $f(x)$ is a piecewise quadratic polynomial:

$$f(x) = \begin{cases} \beta_{0,1} + \beta_{1,1}x + \beta_{2,1}x^2 & \text{for } x < c; \\ \beta_{0,2} + \beta_{1,2}x + \beta_{2,2}x^2 & \text{for } x \geq c. \end{cases}$$

(a) Assume we want to put constraint on $f(x)$ by assuming the function both is continuous and have continuous first derivatives. What kind of constraints do this put on the $\beta_{j,m}$'s?

How many *effective* (or free) number of parameters do you then end up with?

(b) Now define

$$g(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 (x - c)_+^2$$

where $(x - c)_+^2 = (x - c)^2$ if $x > c$ and 0 otherwise.

Show that $g(x)$ is continuous, have continuous first derivatives and is quadratic within each of the intervals $(-\infty, c)$ and $[c, \infty)$.

(c) Show that we can obtain $f(x) = g(x)$ for a suitable choice of $\theta_j, j = 0, ..., M + 1$.

Assume now $Y = f(x) + \varepsilon$ where $f(x)$ is a piecewise quadratic polynomial within a set of intervals:

$$f(x) = \beta_{0,m} + \beta_{1,m}x + \beta_{2,m}x^2 \quad \text{for } c_{m-1} \leq x < c_m$$

for $m = 1, ..., M$, $c_0 = -\infty < c_1 < \cdots < c_{M-1} < c_M = \infty$

(d) Assume again we want to build in constraints on $f(x)$ in that the function is both continuous and have continuous first derivatives. What constraints do this put on the $\beta_{j,m}$'s?

How many *effective* (or free) parameters do you end up with in this case?

(e) How can estimation of the parameters be performed?

You do not need to do the actual calculations, only describe which method that can be applied

Problem 7

We will in this exercise consider a dataset which are results of a spinal operation "laminectomy" on children, to correct for a condition called "kyphosis". The dataset consists of 81 observations on the following 4 variables.

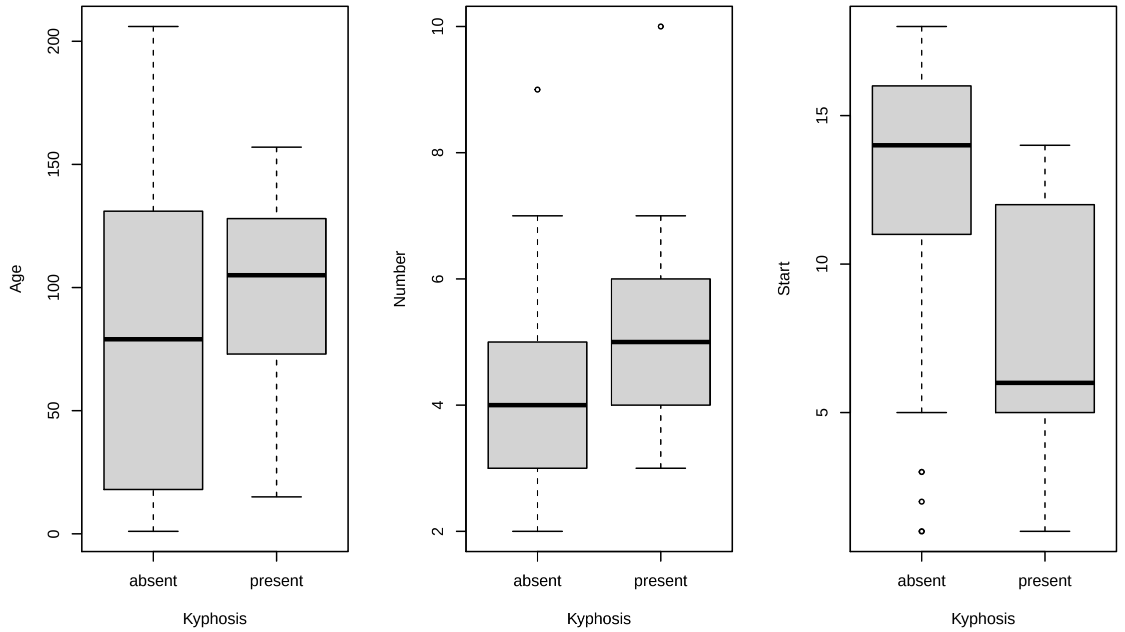
Kyphosis a response factor with levels absent (0) and present (1) (denoted by y).

Age of child in months, a numeric vector (denoted by x_1).

Number of vertebra involved in the operation,a numeric vector (denoted by x_2).

Start level of the operation, a numeric vector (denoted by x_3).

Among the 81 observations 64 have absent, 17 have present as response. The plots below shows boxplots for Age (left), Number (middle), Start (right) divided into the two classes.



Consider first a logistic regression model explanatory variables $(x_1, x_1^2, x_1^3, x_2, x_2^2, x_2^3, x_3, x_3^2, x_3^3)$.

(a) Write down the actual logistic model for this specific case.

The table below shows the results by fitting such a model to the data:

	Estimate	Std. Error	z value	Pr(> $\|z\|$)
(Intercept)	-9.25	7.02	-1.32	0.19
Age	0.06	0.08	0.73	0.46
I(Age^2)	-0.00	0.00	-0.06	0.95
I(Age^3)	-0.00	0.00	-0.32	0.75
Number	2.06	3.81	0.54	0.59
I(Number^2)	-0.21	0.71	-0.29	0.77
I(Number^3)	0.01	0.04	0.13	0.90
Start	0.40	0.96	0.41	0.68
I(Start^2)	-0.03	0.14	-0.19	0.85
I(Start^3)	-0.00	0.01	-0.16	0.87

Discuss possible weaknesses with this fit.

(b) The results below shows the fit obtained after model selection using a stepwise AIC procedure.

Explain what such a stepwise AIC procedure actually do. Also comment on the differences between this model and the one obtained in (a).

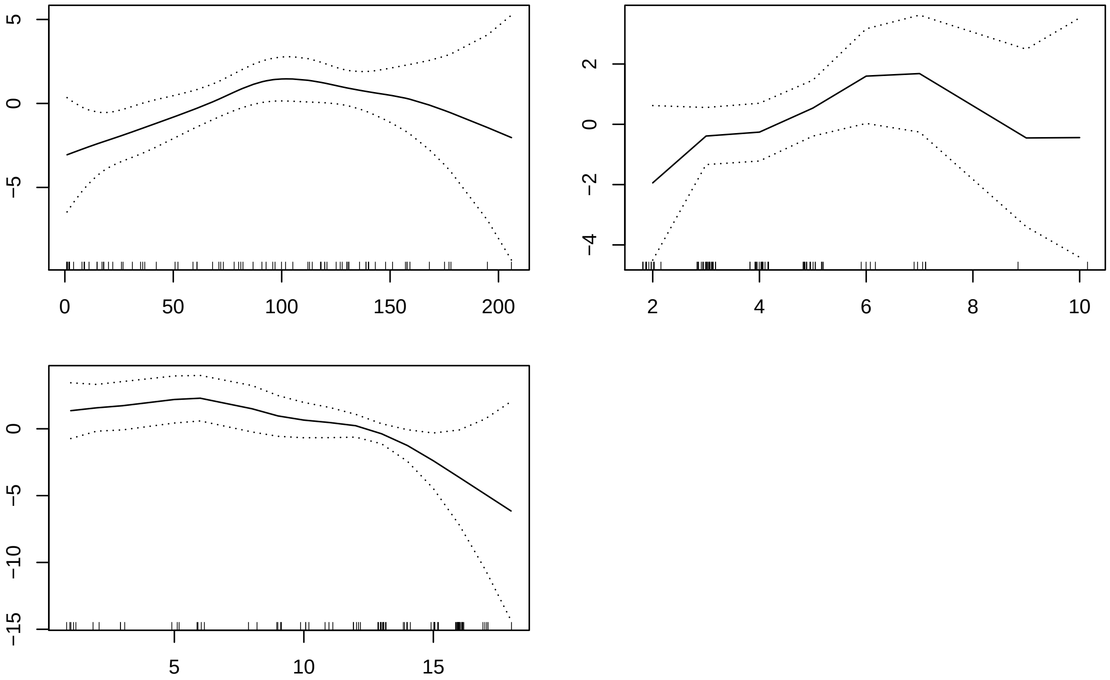
	Estimate	Std. Error	z value	Pr(> $\|z\|$)
(Intercept)	-5.26	2.09	-2.52	0.01
Age	0.05	0.02	2.38	0.02
I(Age^3)	-0.00	0.00	-1.88	0.06
Number	0.32	0.24	1.34	0.18
Start	0.48	0.33	1.48	0.14
I(Start^2)	-0.04	0.02	-2.06	0.04

(*c*) An alternative to the AIC criterion is the BIC criterion. A stepwise procedure using BIC resulted in the following model and table:

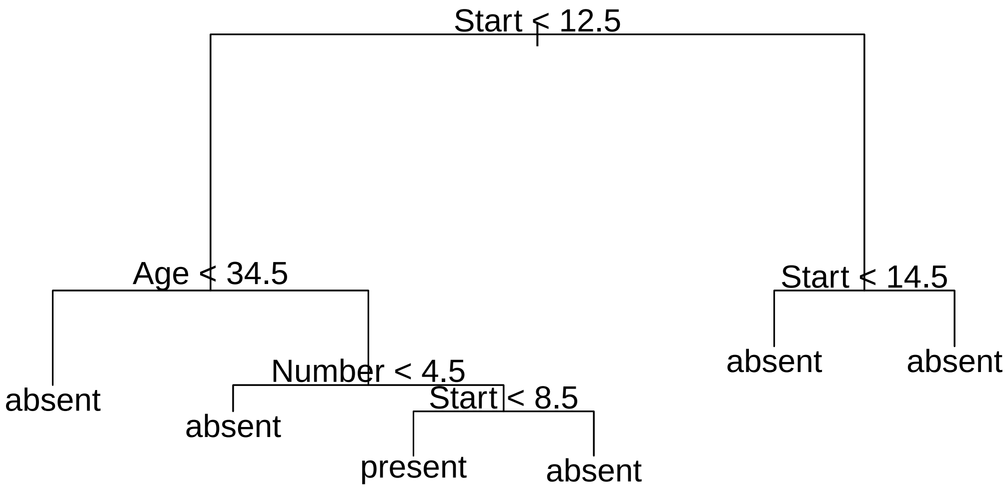
	Estimate	Std. Error	z value	Pr(> $\|z\|$)
(Intercept)	-1.84	1.09	-1.69	0.09
Age	0.05	0.02	2.40	0.02
I(Age^3)	-0.00	0.00	-1.88	0.06
I(Start^2)	-0.02	0.00	-3.58	0.00

Explain the differences between AIC and BIC. Why do we obtain a smaller model in this case?

(*d*) An alternative to polynomial regression is GAM. The plots below shows the estimated nonlinear functions for Age (top left), Number (top right) and Start (bottom left). Based on the results obtained earlier with polynomial regression, do you find these results reasonable?



(*e*) Yet another fit is based on trees. The plot below shows a fitted tree based on first building a large tree and then pruning.



Discuss whether this tree is reasonable, both based on the boxplots shown earlier and the fits obtained for the other methods.

Assume you have an individual of Age 18 with Number=5 and Start=2, what will be the prediction in this case?

(f) The table below shows the log-likelihood values for the different methods discussed.

Method	Log-likelihood
Logistic (a)	-23.83
Logistic select AIC (b)	-24.77
Logistic select BIC (c)	-27.44
GAM (d)	-29.26
Tree (e)	-23.94

Based on this table, which model would you prefer? Specify which criteria you are using for this choice.

Problem 8

Consider a somewhat more general K-means algorithm described below:

1. Choose K and initial arbitrary centroids $\boldsymbol{m}_1, ..., \boldsymbol{m}_k$

2. Cycle for $r = 1, 2, ...$

   a. for $i = 1, ..., n$, assign $\boldsymbol{x}_i$ to group k so that $d(\boldsymbol{x}_i, \boldsymbol{m}_k)$ is the minimum

b. for $k = 1, ..., K$, let $\boldsymbol{m}_k$ be the vector which minimizes $\sum_{c_i==k} d(\boldsymbol{x}_i, \boldsymbol{m}_k)$ where c_i is the group which was assigned in the previous step

(a) Show that if $d(\boldsymbol{x}_i, \boldsymbol{m}_k) = ||\boldsymbol{x}_i - \boldsymbol{m}_k||^2$, that is standard Euclidian distance, we obtain the ordinary K-means algorithm.

(b) An alternative would be to choose $d(\boldsymbol{x}_i, \boldsymbol{m}_k) = (\boldsymbol{x}_i - \boldsymbol{m}_k)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{m}_k)$ for some appropriate chosen matrix $\boldsymbol{\Sigma}$.

Based on the similarities between ordinary K-means clustering and model based clustering, what kind of model for $\boldsymbol{x}$ given class membership would this correspond to?

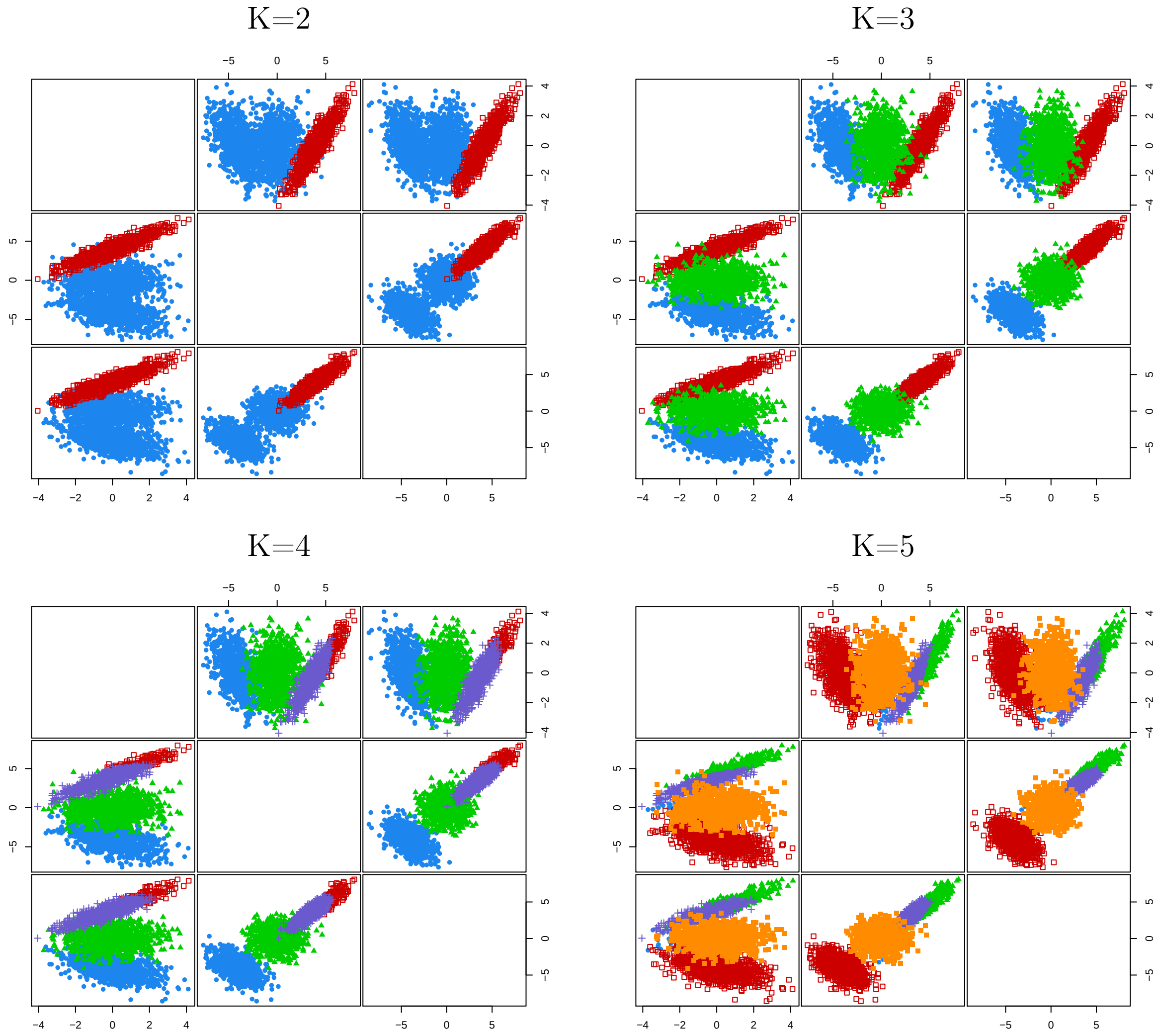
Discuss possible advantages in this more general procedure.

(c) An even more general procedure would be to use $d(\boldsymbol{x}_i, \boldsymbol{m}_k) = (\boldsymbol{x}_i - \boldsymbol{m}_k)\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i - \boldsymbol{m}_k)$, that is group-specific matrices $\boldsymbol{\Sigma}_k$. Discuss possible advantages in this procedure.

Also try to suggest some ways of specifying the $\boldsymbol{\Sigma}_k$'s within the algorithm.

(d) The four plots below shows clustering using $K = 2, K = 3, K = 4$ and $K = 5$, respectively on $n = 3000$ simulated data in 3 dimensions. Each subplot show cross-plot of corresponding variables and colours correspond to different groups the observations are allocated to in the final run.

Based on these plots, explain why it was reasonable to use a group dependent $\boldsymbol{\Sigma}_k$ in this case.



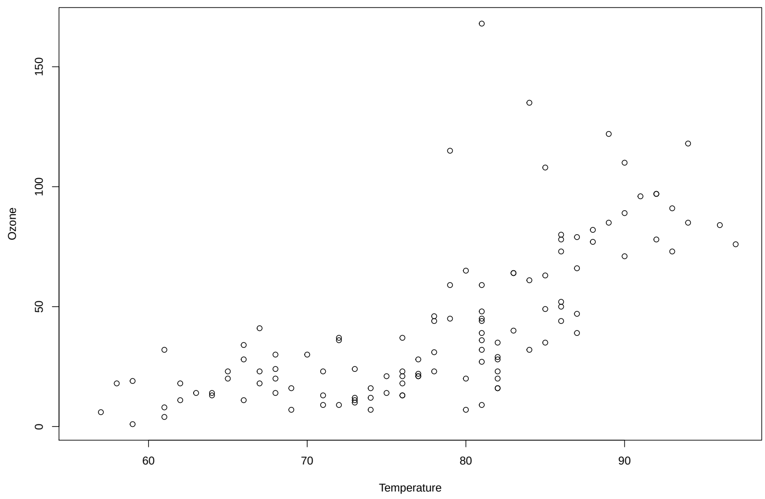
(*e*) The plots above were actually obtained using model based clustering with an assumption of Gaussian distributions within each group. The log-likelihood values was then -16752.96 for $K = 2$, 762.39 for $K = 3$, 766.40 for $K = 4$ and 775.22 for $K = 5$. Based on these values, which number of clusters would you prefer?

Problem 9

We will in this exercise look at a dataset on air quality. The dataset is from May to September 1973 and measures ozon level (the scale is ppb=parts per billion) in New York together with several other (explanatory) variables. We will in the start concentrate on temperature (in Fahrenheit). The figure below shows a plot of ozon against temperature. We will assume the model

$$Y = f(x) + \varepsilon$$

where x is temperature and Y is ozon level.



Consider first local regression in one dimension. Let $K(x_i, x_0)$ be the weight function which specifies how much weight we put on observation i when we want to predict x_0. Mathematically the method can be described through minimization of

$$\sum_{i=1}^{n} K(x_i, x_0)(y_i - \beta_0(x_0) - \sum_{j=1}^{d} \beta_j(x_0)x_i^j)^2$$

with respect to $\beta_0(x_0), ..., \beta_d(x_0)$. We obtain a prediction in point x_0 given by

$$\hat{f}(x_0) = \hat{\beta}_0(x_0) + \sum_{j=1}^{d} \hat{\beta}_j(x_0)x_i^j$$

(a) For $d = 1$, derive the optimal estimates for $\beta_0(x_0), \beta_1(x_0)$ (it is enough to write down an equation system that the estimates needs to satisify).

Will $\hat{f}(x_0)$ be an unbiased estimate for $f(x_0)$? Include an argument for your answer.

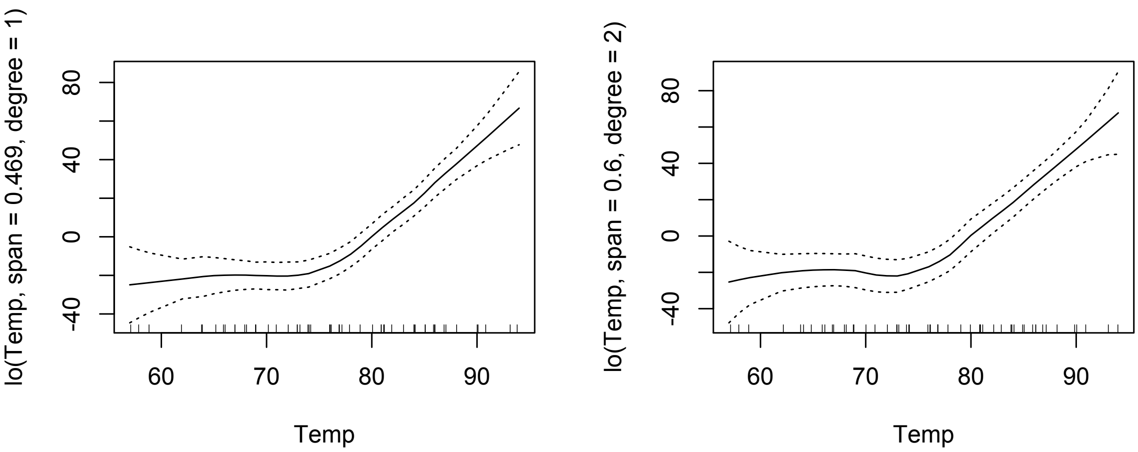
Show that $\hat{y}_i = \hat{f}(x_i) = \sum_{j=1}^{n} S_{ij}y_j$ for all i.

Argue why this also is true for $d = 2$.

(b) Below is a plot of estimated relationship between temperature and ozon based on local regression with $d = 1$ (left) and $d = 2$ (right).

For these two regressions, the corresponding kernel functions $K(x_i, x_0)$ are chosen such that $\sum_{i=1}^{n} S_{ii}$ are about similar for $d = 1$ og $d = 2$. Why is this a reasonable choise?

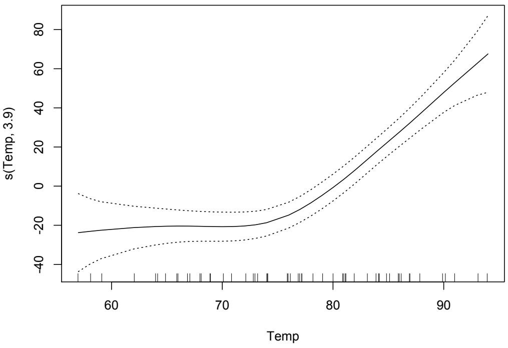
The estmated mean square error (based on a separate test set) was 688.96 for $d = 1$ and 698.71 for $d = 2$. Based on the plot below, argue why this is reasonable in this case.



(c) An alternative to local regression is splines. Below is an estimate of f based on smoothing splines. Also here $\hat{y}_i = \hat{f}(x_i) = \sum_{j=1}^n S_{ij}y_j$ (this you do not need to show) and $\sum_{i=1}^n S_{ii}$ is approximately equal to what was used for local regression.

The estimated function seems to be more smooth in this case compared to what we obtain for local regression. Argue why this is reasonable.

The estimated mean square error in this case was 694.66. Based on these results, which method would you prefer?



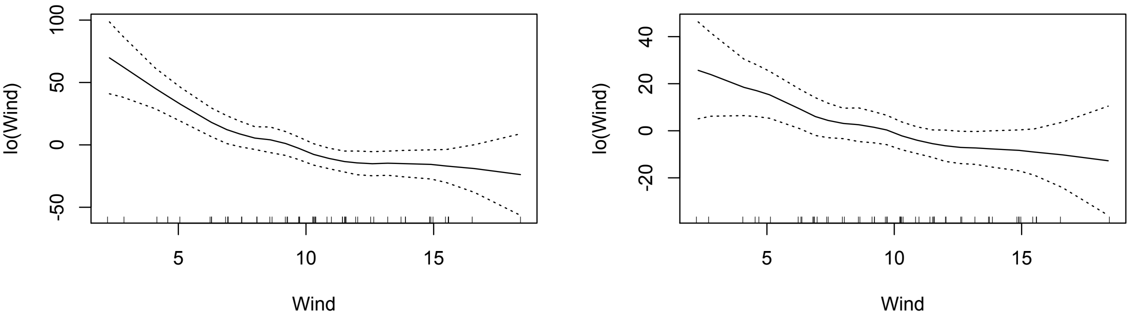
(d) We will now extend the model to also include wind. We will assume a model of the form

$$Y = f_1(x_1) + f_2(x_2) + \varepsilon$$

where x_1 corresponds to temperature and x_2 to wind. $f_1(\cdot)$ are $f_2(\cdot)$ are smooth functions.

Which class of methods to this model belong to?

Below are fits of $f_2(x_2)$ with $f_1(x_1) = 0$ (left) and $f_1(x_1)$ fitted simultaneously (right) shown. Discuss similarities/differences between the two estimates of $f_2(x_2)$.



(*e*) Assume you have a good method for fitting a model $Y = f(x) + \varepsilon$ where $f(\cdot)$ is a smooth function. Explain how you can use this method to fit a model with *two* smooth functions (as in point (*d*)).

Problem 10 (This exercise is somewhat more difficult mathematically)
Assume a linear regression model

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

We can write this model in vector/matrix form:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

Let $\hat{\boldsymbol{\beta}}$ be the least squares estimates for $\boldsymbol{\beta}$. We will in this exercise look at the quantity $\text{RSS} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ where $\widehat{Y}_i = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}$.

(*a*) Show that

$$\boldsymbol{E} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = [\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\boldsymbol{\varepsilon}$$

where $\boldsymbol{I}_n$ is a diagonal matrix of size $n \times n$.

What is the expectation vector and covariance matrix for $\boldsymbol{E}$?

(*b*) Show that $\text{RSS} = (\boldsymbol{Y} - \widehat{\boldsymbol{Y}})^T(\boldsymbol{Y} - \widehat{\boldsymbol{Y}})$ and

$$E[\text{RSS}] = \sigma^2(n - p - 1)$$

Hint: Show that RSS = trace($\boldsymbol{E}\boldsymbol{E}^T$) where the trace of a matrix is the sum of the diagonal elements. You can further use that trace($\boldsymbol{AB}$) = trace($\boldsymbol{BA}$) for $\boldsymbol{A}$ and $\boldsymbol{B}$ matrices with matching sizes.

(c) Show that $\text{Cov}(\hat{y}_i, E_j) = 0$ for all i, j. Discuss this result.

Hint: It can be easier to work directly with the vectors $\hat{\boldsymbol{y}}$ and $\boldsymbol{E}$. The cross-covariance matrix between two stochastic vectors $\boldsymbol{U}$ and $\boldsymbol{V}$ can be written as $\text{Cov}(\boldsymbol{U}, \boldsymbol{V})$ and we have the general result $\text{Cov}(\boldsymbol{AU}, \boldsymbol{BV}) = \boldsymbol{A}\text{Cov}(\boldsymbol{U}, \boldsymbol{V})\boldsymbol{B}^T$.

Problem 11

We will here look at a regression setting where we have some explanatory variables $\boldsymbol{x}$ and assume

$$Y = f(\boldsymbol{x}) + \varepsilon$$

We wish to predict Y where we use as evaluation criterion $E[(Y - \widehat{Y})^2]$. As usual, we have data $\{(y_i, \boldsymbol{x}_i), i = 1, ..., n\}$.

(a) Assume that $\widehat{Y}(\boldsymbol{x}_0) = \hat{f}(\boldsymbol{x}_0)$ for a new point $\boldsymbol{x}_0$. Show that expected loss can be written as

$$E[(Y - \widehat{Y}(\boldsymbol{x}_0))^2|\boldsymbol{x}_0] = (f(\boldsymbol{x}_0) - E[\widehat{f}(\boldsymbol{x}_0)])^2 + E[(\widehat{f}(\boldsymbol{x}_0) - E[\widehat{f}(\boldsymbol{x}_0)|\boldsymbol{x}_0])^2|\boldsymbol{x}_0] + \text{Var}(\varepsilon)$$

Give an interpretation of the different terms on the left hand side.

(b) Now let $\hat{f}_1(\boldsymbol{x})$ be a predictor based on a very restrictive method/model while $\hat{f}_2(\boldsymbol{x})$ is based on a more flexible approach. Discuss the different terms in the equation above in this setting.

(c) Discuss different methods for estimation of $E[(Y - \widehat{Y})^2]$. In particular discuss advantages and disadvantages with the different methods.

Problem 12 (Note: This problem is somewhat more difficult mathematically)

Spam filters are often based on statistical classification methods in order to distinguish between real and spam mails. Such methods can rely on frequencies of different words in the mail. If we let W be a word which we believe occur more often in a spam than in a real email, a possible procedure is based on

$$\text{Pr}(V|S) = q > p = \text{Pr}(V|R)$$

where V is the event that W is a word in the mail, S is the event that the mail is a spam and R is the event that the mail is real.

Assume further that r is the fraction of mails that are real.

(a) Derive a classification rule where you classify to spam if the probability for a spam-mail is larger than 0.5.

(An expression which is a function of q, p, r. You will need to consider both the event V and the complementary V^c)

(b) Assume that we consider classification of real mails to spam as a more serious error than classification of spam mail as real mail. Explain how this can be formulated mathematically and derive the optimal classification rule in this case.

Assume now that you have a set of words $W_1, ..., W_M$ which frequently occur. Let V_m be the event that the word W_m occur in a mail. $\boldsymbol{V} = (V_1, ..., V_M)$ will then be a vector of binary variables (where 1 corresponds to that the word occur in the mail). Let further $\Pr(V_m|\text{S}) = q_m$ and $\Pr(V_m|\text{R}) = p_m$.

(c) Assume now that $\Pr(\boldsymbol{V}|S) = \prod_{m=1}^{M} q_m^{V_m}(1-q_m)^{1-V_m}$ and $\Pr(\boldsymbol{V}|R) = \prod_{m=1}^{M} p_m^{V_m}(1-p_m)^{1-V_m}$. What kind of assumptions do these statements rely on?

Derive the classification rule also in this case. (You will get a rather complicated formulae in this case.)

(d) Often, one not only look at single words but also pair of words. Let $V_{m,m'}$ denote the event that both words W_m and $W_{m'}$ occur.

Discuss advantages and disadvantages with such an approach.