THE UNIVERSITY OF MELBOURNE

Student number	

Semester 1 Assessment, 2023

School of Mathematics and Statistics

MAST30025 Linear Statistical Models Assignment 2

Submission deadline: **Friday April 28, 5pm**

This assignment consists of 14 pages (including this page) with 5 questions and 40 total marks

Instructions to Students

Writing

- This assignment is worth 7% of your total mark.
- You may choose to either typeset your assignment in LaTeX, or handwrite and scan it to produce an electronic version.
- You may use R for this assignment, including the `lm` function unless otherwise specified. If you do, include your R commands and output.
- Write your answers on A4 paper. Page 1 should only have your student number, the subject code and the subject name. Write on one side of each sheet only. Each question should be on a new page. The question number must be written at the top of each page.

Scanning and Submitting

- Put the pages in question order and all the same way up. Use a scanning app to scan all pages to PDF. Scan directly from above. Crop pages to A4.
- Submit your scanned assignment as a single PDF file and carefully review the submission in Gradescope. Scan again and resubmit if necessary.

Question 1 (4 marks)

Prove the formula on slide 126 of chapter 4: that is,

$$\frac{y^* - (\mathbf{x}^*)^T \mathbf{b}}{s\sqrt{1 + (\mathbf{x}^*)^T (X^T X)^{-1} \mathbf{x}^*}}.$$

has a t distribution with $n - p$ degrees of freedom.

Write

$$z = \frac{y^* - (\mathbf{x}^*)^T \mathbf{b}}{s\sqrt{1 + (\mathbf{x}^*)^T (X^T X)^{-1} \mathbf{x}^*}} = \frac{(y^* - (\mathbf{x}^*)^T \mathbf{b})/\sigma\sqrt{1 + (\mathbf{x}^*)^T (X^T X)^{-1} \mathbf{x}^*}}{s/\sigma}.$$

By the results from slide 125, the numerator

$$\frac{y^* - (\mathbf{x}^*)^T \mathbf{b}}{\sigma\sqrt{1 + (\mathbf{x}^*)^T (X^T X)^{-1} \mathbf{x}^*}} \sim N(0, 1).$$

The denominator is

$$\frac{s}{\sigma} = \sqrt{\frac{s^2}{\sigma^2}} = \sqrt{\frac{SS_{Res}/\sigma^2}{n - p}},$$

and

$$\frac{SS_{Res}}{\sigma^2} \sim \chi^2_{n-p}.$$

Finally, SS_{Res} is independent of both y^* and $\mathbf{b}$, and therefore of the numerator. Therefore, z follows a t distribution with $n - p$ degrees of freedom.

Question 2 (11 marks)

We wish to predict the price of apartments in Melbourne using some of their features. Let y be the apartment price per square metre, x_1 be the apartment age (in years), x_2 be the distance (in metres) to the nearest train station, and x_3 be the number of convenience stores nearby. The following data is collected:

x_1 (years)	x_2 (meters)	x_3	y (\$, $\times 10^2$)
32	84.9	10	37.9
19.5	306.6	9	42.2
13.3	562.0	5	47.3
13.3	562.0	5	43.1
5	390.6	5	54.8
7.1	2175.0	3	47.1
34.5	623.5	7	40.3

For this question, you may not use the `lm` function in R.

(a) Fit a linear model to the data and estimate the parameters and error variance.

(b) Calculate 95% confidence intervals for the parameters.

(c) Calculate a 90% prediction interval for the price per square metre of a 5 year old apartment that is 100 meters away from the nearest train station and has 6 convenience stores nearby.

(d) Test the hypothesis that the price per square metre falls by \$100 for every year that the apartment ages, at the 5% significance level.

(e) Test for model relevance using a corrected sum of squares.

(a)
```
> n <- 7
> p <- 4
> X <- matrix(c(rep(1,n), 32, 19.5, 13.3, 13.3, 5, 7.1, 34.5,
+                        84.9, 306.6, 562.0, 562.0, 390.6, 2175.0, 623.5,
+                        10, 9, 5, 5, 5, 3, 7),n,p)
> y <- c(37.9, 42.2, 47.3, 43.1, 54.8, 47.1, 40.3)
> (b <- solve(t(X)%*%X,t(X)%*%y))

             [,1]
[1,] 58.369312708
[2,] -0.346291960
[3,] -0.002900359
[4,] -0.887671692

> (s2 <- sum((y-X%*%b)^2)/(n-p))

[1] 13.06871
```

(b)
```
> C <- solve(t(X)%*%X)
> b[1] + c(-1,1)*qt(0.975,n-p)*sqrt(s2*C[1,1])

[1] 34.10183 82.63680

> b[2] + c(-1,1)*qt(0.975,n-p)*sqrt(s2*C[2,2])

[1] -0.9970944  0.3045104

> b[3] + c(-1,1)*qt(0.975,n-p)*sqrt(s2*C[3,3])

[1] -0.013166703  0.007365986

> b[4] + c(-1,1)*qt(0.975,n-p)*sqrt(s2*C[4,4])

[1] -4.817973  3.042630
```

(c)
```
> C <- solve(t(X)%*%X)
> xst <- as.vector(c(1,5,100,6))
> xst %*% b + c(-1,1)*qt(0.95,df=n-p)*
+   sqrt(s2*(1+ t(xst) %*% C %*% xst))

[1] 39.86902 62.17455
```

(d) This is a general linear hypothesis with $C = \begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix}$ and $\delta^* = -1$.

```
> C <- matrix(c(0,1,0,0),1,4)
> dst <- -1
> Fstat <- (t(C%*%b-dst)%*%solve(C%*%solve(t(X)%*%X)%*%t(C))%*%
+              (C%*%b-dst)/1)/s2
> pf(Fstat, 1, n-p, lower=F)

           [,1]
[1,] 0.04945829
```

We reject the null hypothesis at the 5% significance level.

(e)
```
> SSRegc <- t(y) %*% X %*% b - sum(y)^2 / n
> SSRes <- s2*(n-p)
> ( Fstat <- (SSRegc/(p-1))/(SSRes/(n-p)) )

      [,1]
[1,] 3.819

> pf(Fstat, p-1, n-p, lower.tail = FALSE)

          [,1]
[1,] 0.1500833
```

We do not reject the null hypothesis of model irrelevance.

Question 3 (5 marks)

Show that for a full rank linear model with p parameters, the Akaike's information criterion, defined as $-2\log(\text{Likelihood}) + 2p$, can be written as

$$n\log\left(\frac{SS_{Res}}{n}\right) + 2p + \text{const.}$$

If $\mathbf{y} \sim MVN(X\boldsymbol{\beta}, \sigma^2 I_n)$, then $\mathbf{y}$ has the density

$$f(\mathbf{y}) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(\mathbf{y}-X\boldsymbol{\beta})^T(\mathbf{y}-X\boldsymbol{\beta})/(2\sigma^2)}.$$

The maximum likelihood estimates of $\boldsymbol{\beta}$ and σ^2 are $\mathbf{b}$ and SS_{Res}/n respectively.
Substituting these into the density, we get the maximised likelihood

$$\begin{aligned} L &= \frac{1}{(2\pi)^{n/2}(SS_{Res}/n)^{n/2}} e^{-n(\mathbf{y}-X\mathbf{b})^T(\mathbf{y}-X\mathbf{b})/(2SS_{Res})} \\ &= \frac{1}{(2\pi)^{n/2}(SS_{Res}/n)^{n/2}} e^{-n/2}. \end{aligned}$$

Thus the AIC is

$$-2\log L + 2p = n\log(2\pi) + n\log(SS_{Res}/n) + n + 2p.$$

The terms $n\log(2\pi)$ and n are same for any linear model fitted to a given data set, and hence play no part when you use the AIC for model selection.

Question 4 (12 marks)

In this question, we study a dataset of 50 US states. This dataset contains the variables:

- `Population`: population estimate as of July 1, 1975
- `Income`: per capita income (1974)
- `Illiteracy`: illiteracy (1970, percent of population)
- `Life.Exp`: life expectancy in years (1969–71)
- `Murder`: murder and non-negligent manslaughter rate per 100,000 population (1976)
- `HS.Grad`: percentage of high-school graduates (1970)
- `Frost`: mean number of days with minimum temperature below freezing (1931–1960) in capital or large city
- `Area`: land area in square miles

The dataset is distributed with R. Open it with the following commands:

```
> data(state)
> statedata <- data.frame(state.x77, row.names=state.abb, check.names=TRUE)
```

We wish to use a linear model to model the murder rate (`Murder`) in terms of the other variables.

(a) Plot the data and comment. Should we consider any variable transformations?

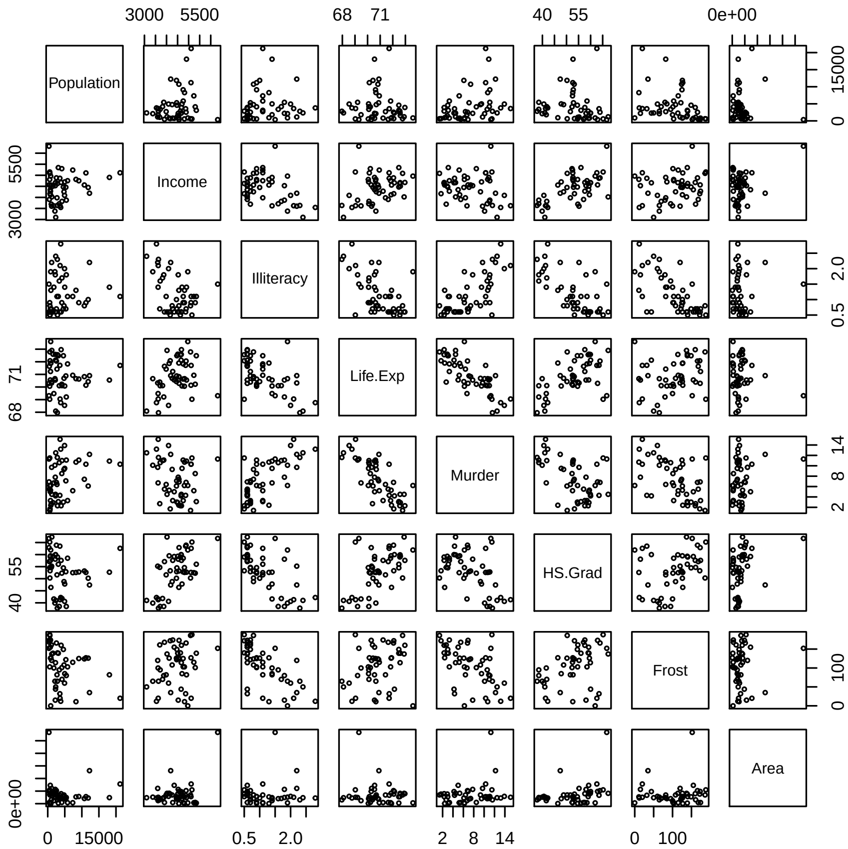
(b) Perform model selection using forward selection, using all variable transformations that may be relevant.

(c) Starting from the full model, perform model selection using stepwise selection with the AIC.

(d) Write down your final fitted model (including any variable transformations used).

(e) Produce diagnostic plots for your final model and comment.

(a)
```
> pairs(statedata,cex=0.5)
> statedata$logPopulation <- log(statedata$Population)
> statedata$logArea <- log(statedata$Area)
```



Looking at murder rate against the other variables, there is evidence of a linear relationship with income, illiteracy, life expectancy, percentage of high school graduates and frost. There is no obvious relationship with population and area.

Population and area both have distributions heavily skewed to the right. log(population) and log(area) would be less skewed and might fit better with the other variables.

There is potential heteroskedasticity in the relationship with high school grad, and non-linearity in illiteracy, but neither enough for immediate concern.

(b)

```
> model0 <- lm(Murder ~ 1, data=statedata)
> add1(model0, scope= ~ . + Population + Income + Illiteracy + Life.Exp + HS.Grad
+       + Frost + Area + logPopulation + logArea, test="F")

Single term additions

Model:
Murder ~ 1
              Df Sum of Sq    RSS     AIC F value    Pr(>F)
<none>                     667.75 131.594
Population     1     78.85 588.89 127.311  6.4273 0.0145504 *
Income         1     35.35 632.40 130.875  2.6829 0.1079683
Illiteracy     1    329.98 337.76  99.516 46.8943 1.258e-08 ***
Life.Exp       1    407.14 260.61  86.550 74.9887 2.260e-11 ***
HS.Grad        1    159.00 508.75 119.996 15.0017 0.0003248 ***
Frost          1    193.91 473.84 116.442 19.6433 5.405e-05 ***
Area           1     34.83 632.91 130.916  2.6416 0.1106495
logPopulation  1     86.37 581.37 126.668  7.1313 0.0103090 *
logArea        1     58.63 609.12 128.999  4.6201 0.0366687 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> model1 <- lm(Murder ~ Life.Exp, data=statedata)
> add1(model1, scope= ~ . + Population + Income + Illiteracy + HS.Grad
+       + Frost + Area + logPopulation + logArea, test="F")

Single term additions

Model:
Murder ~ Life.Exp
              Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>                     260.61 86.550
Population     1    56.615 203.99 76.303 13.0442 0.0007374 ***
Income         1     0.958 259.65 88.366  0.1733 0.6790605
Illiteracy     1    60.549 200.06 75.329 14.2249 0.0004533 ***
HS.Grad        1     1.124 259.48 88.334  0.2035 0.6539823
Frost          1    80.104 180.50 70.187 20.8575 3.576e-05 ***
Area           1    14.121 246.49 85.764  2.6926 0.1074933
logPopulation  1    50.862 209.75 77.694 11.3972 0.0014838 **
logArea        1    30.223 230.38 82.386  6.1656 0.0166517 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> model2 <- lm(Murder ~ Life.Exp + Frost, data=statedata)
> add1(model2, scope= ~ . + Population + Income + Illiteracy + HS.Grad
+          + Area + logPopulation + logArea, test="F")

Single term additions

Model:
Murder ~ Life.Exp + Frost
              Df Sum of Sq    RSS    AIC F value   Pr(>F)
<none>                     180.50 70.187
Population     1   23.7098 156.79 65.146  6.9559 0.011358 *
Income         1    5.5598 174.94 70.622  1.4619 0.232807
Illiteracy     1    6.0663 174.44 70.477  1.5997 0.212315
HS.Grad        1    2.0679 178.44 71.610  0.5331 0.469015
Area           1   21.0840 159.42 65.976  6.0837 0.017430 *
logPopulation  1   12.2130 168.29 68.684  3.3382 0.074179 .
logArea        1   30.9733 149.53 62.774  9.5283 0.003422 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> model3 <- lm(Murder ~ Life.Exp + Frost + logArea, data=statedata)
> add1(model3, scope= ~ . + Population + Income + Illiteracy + HS.Grad
+          + Area + logPopulation, test="F")

Single term additions

Model:
Murder ~ Life.Exp + Frost + logArea
              Df Sum of Sq    RSS    AIC F value  Pr(>F)
<none>                     149.53 62.774
Population     1   16.3474 133.18 58.985  5.5235 0.02321 *
Income         1    4.7860 144.75 63.147  1.4879 0.22889
Illiteracy     1    8.7371 140.79 61.764  2.7925 0.10165
HS.Grad        1    0.1900 149.34 64.710  0.0572 0.81200
Area           1    1.2394 148.29 64.358  0.3761 0.54278
logPopulation  1    9.1315 140.40 61.623  2.9268 0.09401 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> model4 <- lm(Murder ~ Life.Exp + Frost + logArea + Population, data=statedata)
> add1(model4, scope= ~ . + Income + Illiteracy + HS.Grad
+          + Area + logPopulation, test="F")

Single term additions

Model:
Murder ~ Life.Exp + Frost + logArea + Population
              Df Sum of Sq    RSS    AIC F value  Pr(>F)
<none>                     133.18 58.985
Income         1    0.9201 132.26 60.639  0.3061 0.58289
Illiteracy     1   13.9190 119.26 55.466  5.1351 0.02842 *
HS.Grad        1    0.0829 133.10 60.954  0.0274 0.86929
Area           1    2.0911 131.09 60.194  0.7019 0.40668
logPopulation  1    0.5229 132.66 60.789  0.1734 0.67911
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> model5 <- lm(Murder ~ Life.Exp + Frost + logArea + Population
+          + Illiteracy, data=statedata)
> add1(model5, scope= ~ . + Income + HS.Grad + Area + logPopulation, test="F")

Single term additions

Model:
Murder ~ Life.Exp + Frost + logArea + Population + Illiteracy
              Df Sum of Sq    RSS    AIC F value Pr(>F)
<none>                     119.26 55.466
Income         1    3.7237 115.54 55.880  1.3858 0.2456
HS.Grad        1    2.0218 117.24 56.611  0.7415 0.3940
Area           1    0.4459 118.82 57.279  0.1614 0.6899
logPopulation  1    0.4628 118.80 57.272  0.1675 0.6844
```

The final variables are life expectancy, frost, log(area), population, and illiteracy.

(c)

```
> fullmodel <- lm(Murder ~ ., data = statedata)
> model <- step(fullmodel, scope = ~ .)

Start:  AIC=61.22
Murder ~ Population + Income + Illiteracy + Life.Exp + HS.Grad +
    Frost + Area + logPopulation + logArea

                Df Sum of Sq    RSS    AIC
- HS.Grad        1     0.105 114.14 59.269
- logPopulation  1     0.282 114.31 59.346
- Area           1     1.342 115.37 59.808
- Income         1     3.202 117.23 60.607
<none>                       114.03 61.223
- Population     1     5.575 119.61 61.609
- Frost          1     5.712 119.74 61.667
- logArea        1    13.175 127.21 64.690
- Illiteracy     1    15.379 129.41 65.548
- Life.Exp       1   114.344 228.38 93.948

Step:  AIC=59.27
Murder ~ Population + Income + Illiteracy + Life.Exp + Frost +
    Area + logPopulation + logArea

                Df Sum of Sq    RSS    AIC
- logPopulation  1     0.559 114.70 57.513
- Area           1     1.330 115.47 57.848
- Income         1     4.504 118.64 59.204
<none>                       114.14 59.269
- Population     1     6.314 120.45 59.961
- Frost          1     6.688 120.82 60.116
+ HS.Grad        1     0.105 114.03 61.223
- logArea        1    14.655 128.79 63.309
- Illiteracy     1    16.934 131.07 64.186
- Life.Exp       1   131.265 245.40 95.544

Step:  AIC=57.51
Murder ~ Population + Income + Illiteracy + Life.Exp + Frost +
    Area + logArea

                Df Sum of Sq    RSS    AIC
- Area           1     0.845 115.54 55.880
- Income         1     4.123 118.82 57.279
<none>                       114.70 57.513
- Frost          1     6.223 120.92 58.155
+ logPopulation  1     0.559 114.14 59.269
+ HS.Grad        1     0.382 114.31 59.346
- Population     1    11.770 126.47 60.398
- logArea        1    14.310 129.01 61.392
- Illiteracy     1    16.384 131.08 62.189
- Life.Exp       1   131.158 245.85 93.636

Step:  AIC=55.88
Murder ~ Population + Income + Illiteracy + Life.Exp + Frost +
    logArea

                Df Sum of Sq    RSS    AIC
- Income         1     3.724 119.26 55.466
<none>                       115.54 55.880
- Frost          1     7.953 123.49 57.209
+ Area           1     0.845 114.70 57.513
+ HS.Grad        1     0.159 115.38 57.811
+ logPopulation  1     0.074 115.47 57.848
- Population     1    15.280 130.82 60.090
- Illiteracy     1    16.723 132.26 60.639
- logArea        1    26.376 141.92 64.161
- Life.Exp       1   130.757 246.30 91.726

Step:  AIC=55.47
Murder ~ Population + Illiteracy + Life.Exp + Frost + logArea

                Df Sum of Sq    RSS    AIC
<none>                       119.26 55.466
+ Income         1     3.724 115.54 55.880
- Frost          1     7.639 126.90 56.570
+ HS.Grad        1     2.022 117.24 56.611
+ logPopulation  1     0.463 118.80 57.272
+ Area           1     0.446 118.82 57.279
- Illiteracy     1    13.919 133.18 58.985
- Population     1    21.529 140.79 61.764
- logArea        1    25.704 144.97 63.225
- Life.Exp       1   127.359 246.62 89.792
```

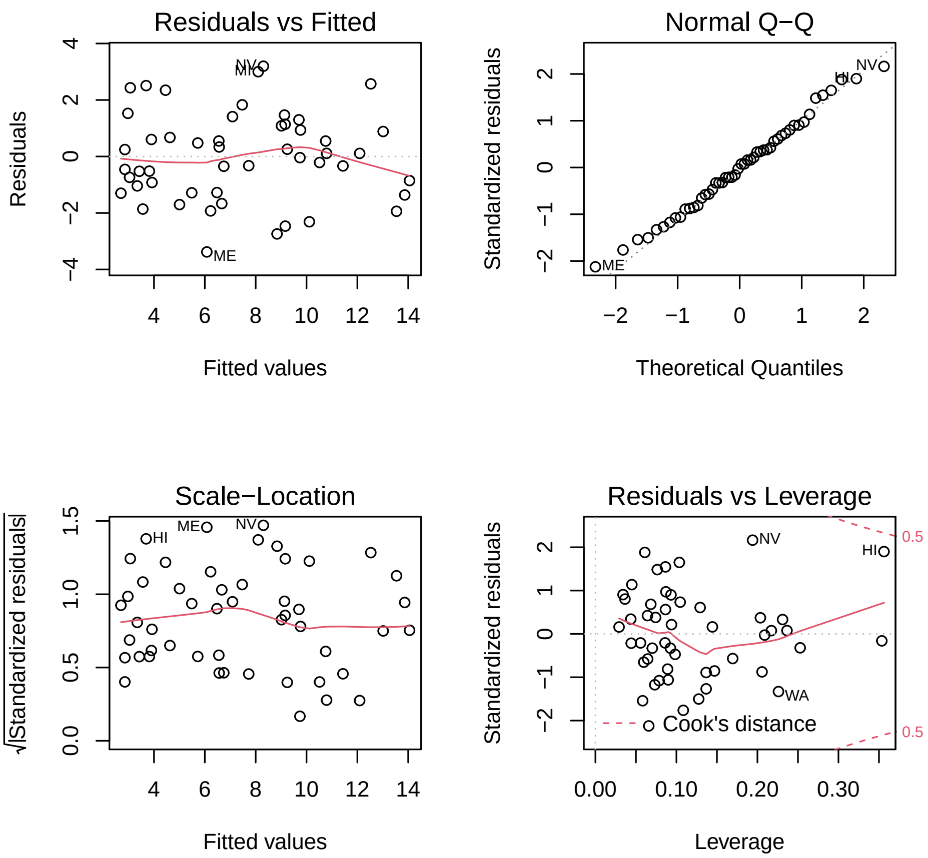
The model is the same as that found by forward selection.

(d)
```
> model

Call:
lm(formula = Murder ~ Population + Illiteracy + Life.Exp + Frost +
    logArea, data = statedata)

Coefficients:
(Intercept)   Population   Illiteracy     Life.Exp        Frost      logArea
 108.713249     0.000162     1.474305    -1.542284    -0.011293     0.632740
```

The final model is

$$\begin{aligned}\texttt{Murder} &= 108.71 + 0.00016\,\texttt{Population} + 1.47\,\texttt{Illiteracy} \\ &\quad -1.54\,\texttt{Life.Exp} - 0.011\,\texttt{Frost} + 0.63\,\ln(\texttt{Area}).\end{aligned}$$

(e)
```
> par(mfrow=c(2,2))
> plot(model, which=1)
> plot(model, which=2)
> plot(model, which=3)
> plot(model, which=5)
```



Diagnostic plots show a reasonable fit to linear model assumptions. About the only area of concern is a slight negative trend for higher fitted values and moderate leverages, but this does not appear to be too alarming.

Question 5 (8 marks)

For ridge regression, we choose parameter estimators $\mathbf{b}$ which minimise

$$\sum_{i=1}^{n} e_i^2 + \lambda \sum_{j=0}^{k} b_j^2,$$

where λ is a constant penalty parameter.

(a) Show that these estimators are given by

$$\mathbf{b} = (X^T X + \lambda I)^{-1} X^T \mathbf{y}.$$

(b) Calculate the ridge regression parameter estimates for the data from Q2 with penalty parameter $\lambda = 1.5$. In order to avoid penalising some parameters unfairly, we must first scale every predictor variable so that it is standardised (mean 0, variance 1), and centre the response variable (mean 0), in which case an intercept parameter is not used. (*Hint:* This can be done with the `scale` function).

(c) One way to calculate the optimal value for the penalty parameter is to minimise the AIC. Since the number of parameters p does not change, we use a slightly modified version:

$$AIC = n \ln \frac{SS_{Res}}{n} + 2\, df,$$

where df is the "effective degrees of freedom" defined by

$$df = tr(H) = tr(X(X^T X + \lambda I)^{-1} X^T).$$

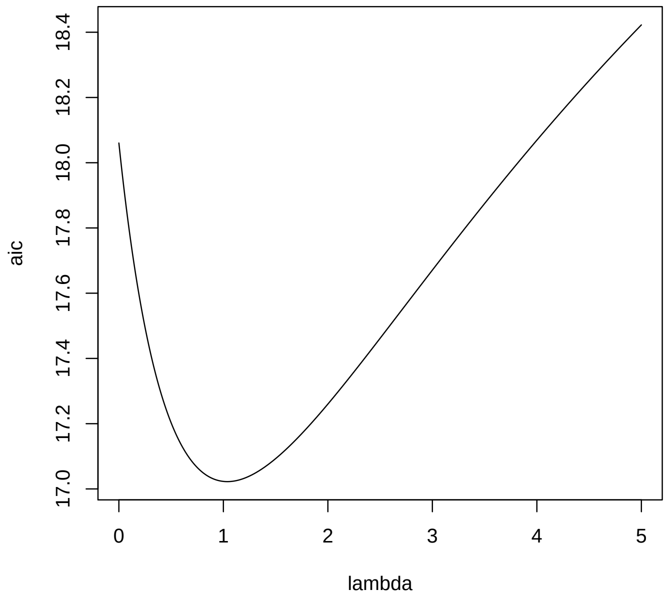
For the data from Q2, construct a plot of λ against AIC. Thereby find the optimal value for λ.

(a) We have

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{b}} \left[\sum_{i=1}^{n} e_i^2 + \lambda \sum_{j=0}^{k} b_j^2 \right] &= \frac{\partial}{\partial \mathbf{b}} \left[(\mathbf{y} - X\mathbf{b})^T (\mathbf{y} - X\mathbf{b}) + \lambda \mathbf{b}^T \mathbf{b} \right] \\
&= \frac{\partial}{\partial \mathbf{b}} \left[\mathbf{y}^T \mathbf{y} - 2\mathbf{y}^T X \mathbf{b} + \mathbf{b}^T X^T X \mathbf{b} + \lambda \mathbf{b}^T \mathbf{b} \right] \\
&= -2X^T \mathbf{y} + 2(X^T X + \lambda I)\mathbf{b} = 0 \\
(X^T X + \lambda I)\mathbf{b} &= X^T \mathbf{y} \\
\mathbf{b} &= (X^T X + \lambda I)^{-1} X^T \mathbf{y}.
\end{aligned}$$

(b)
```
> Xs <- scale(X[,-1],center=T,scale=T)
> ys <- scale(y,center=T,scale=F)
> p <- p-1
> solve(t(Xs)%*%Xs + diag(rep(1.5,p)),t(Xs)%*%ys)

          [,1]
[1,] -3.157565
[2,] -1.002633
[3,] -1.713340
```

(c)
```
> lambda <- seq(0,5,0.001)
> aic <- c()
> for (l in lambda) {
+         b <- solve(t(Xs)%*%Xs + diag(rep(l,p)),t(Xs)%*%ys)
+         ssres <- sum((ys-Xs%*%b)^2)
+         H <- Xs %*% solve(t(Xs)%*%Xs + diag(rep(l,p))) %*% t(Xs)
+         aic <- c(aic, n*log(ssres/n) + 2*sum(diag(H)))
+ }
> plot(lambda,aic,type='l')
> lambda[which.min(aic)]

[1] 1.038
```



End of Assignment — Total Available Marks = 40