MAST30025: Linear Statistical Models

Assignment 2, 2019 Solutions

Total marks: 45

Due: 5pm Friday, May 3 (week 8)

1. Prove Theorem 4.8: show that the maximum likelihood estimator of the error variance σ^2 is

$$\hat{\sigma}^2 = \frac{SS_{Res}}{n}.$$

Solution [4 marks]: The log-likelihood is given in the lecture notes as

$$\begin{aligned}
\ln L(\boldsymbol{\beta}, \sigma^2) &= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - X\boldsymbol{\beta})^T(\mathbf{y} - X\boldsymbol{\beta}) \\
\frac{\partial}{\partial\sigma^2}\ln L(\boldsymbol{\beta}, \sigma^2) &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y} - X\boldsymbol{\beta})^T(\mathbf{y} - X\boldsymbol{\beta}) = 0 \\
\sigma^2 &= \frac{1}{n}(\mathbf{y} - X\boldsymbol{\beta})^T(\mathbf{y} - X\boldsymbol{\beta})
\end{aligned}$$

which gives the required formula on the substitution of the ML estimator $\mathbf{b}$ for $\boldsymbol{\beta}$.

2. An experiment is conducted to estimate the annual demand for cars, based on their cost, the current unemployment rate, and the current interest rate. A survey is conducted and the following measurements obtained:

Cars sold ($\times 10^3$)	Cost ($\$k$)	Unemployment rate (%)	Interest rate (%)
5.5	7.2	8.7	5.5
5.9	10.0	9.4	4.4
6.5	9.0	10.0	4.0
5.9	5.5	9.0	7.0
8.0	9.0	12.0	5.0
9.0	9.8	11.0	6.2
10.0	14.5	12.0	5.8
10.8	8.0	13.7	3.9

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 variance.

Solution [2 marks]:

```
> n <- 8
> p <- 4
> X <- matrix(c(rep(1,n),7.2,10,9,5.5,9,9.8,14.5,8,
+                       8.7,9.4,10,9,12,11,12,13.7,
+                       5.5,4.4,4,7,5,6.2,5.8,3.9),n,p)
> y <- c(5.5,5.9,6.5,5.9,8,9,10,10.8)
> (b <- solve(t(X)%*%X,t(X)%*%y))
           [,1]
[1,] -7.4044796
[2,]  0.1207646
[3,]  1.1174846
[4,]  0.3861206
> (s2 <- sum((y-X%*%b)^2)/(n-p))
```

```
[1] 0.3955368
```

(b) Which two of the parameters have the highest (in magnitude) covariance in their estimators?

```
> (C <- solve(t(X)%*%X))

             [,1]         [,2]        [,3]         [,4]
[1,] 13.49743324 -0.054817613 -0.69854293 -1.029731987
[2,] -0.05481761  0.024498395 -0.01478859 -0.001937333
[3,] -0.69854293 -0.014788594  0.06226378  0.031714790
[4,] -1.02973199 -0.001937333  0.03171479  0.135362495
```

Solution [2 marks]: Parameters β_0 (intercept) and β_3 (interest rate) have the estimators with the highest covariance in magnitude.

(c) Find a 99% confidence interval for the average number of \$8,000 cars sold in a year which has unemployment rate 9% and interest rate 5%.

Solution [2 marks]:

```
> xst <- as.vector(c(1,8,9,5))
> xst %*% b + c(-1,1)*qt(0.995,df=n-p)*sqrt(s2 * t(xst) %*% C %*% xst)

[1] 3.926075 7.173129
```

(d) A prediction interval for the number of cars sold in such a year is calculated to be $(4012, 7087)$. Find the confidence level used.

Solution [3 marks]: Let α be the level used. Then

$$\begin{aligned}(\mathbf{x}^*)^T\mathbf{b} - t_{\alpha/2}s\sqrt{1 + (\mathbf{x}^*)^T(X^TX)^{-1}\mathbf{x}^*} &= 4.012 \\ t_{\alpha/2} &= \frac{(\mathbf{x}^*)^T\mathbf{b} - 4.012}{s\sqrt{1 + (\mathbf{x}^*)^T(X^TX)^{-1}\mathbf{x}^*}}\end{aligned}$$

The confidence level is 90%.

```
> talph <- (t(xst) %*% b - 4.012) / sqrt(s2) / sqrt(1 + t(xst) %*% C %*% xst)
> 1-2*pt(talph, n-p, lower.tail=FALSE)

          [,1]
[1,] 0.9000747
```

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

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

         [,1]
[1,] 23.47683

> Fstat

         [,1]
[1,] 23.47683

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

            [,1]
[1,] 0.005317255
```

Solution [2 marks]: We reject the null hypothesis of model irrelevance.

3. Consider two full rank linear models $\mathbf{y} = X_1\boldsymbol{\gamma}_1 + \boldsymbol{\varepsilon}_1$ and $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}_2$, where all predictors in the first model ($\boldsymbol{\gamma}_1$) are also contained in the second model ($\boldsymbol{\beta}$). Show that the SS_{Res} for the first model is at least the SS_{Res} for the second model.

Solution [5 marks]: Let $\hat{\gamma}_1$ be the least squares estimates for $\boldsymbol{\gamma}_1$ in the first model. Then $\begin{bmatrix} \hat{\gamma}_1 \\ \mathbf{0} \end{bmatrix}$ is a (not necessarily optimal) estimate for $\boldsymbol{\beta}$ in the second model, with residual sum of squares

$$\begin{aligned} \left(\mathbf{y} - X \begin{bmatrix} \hat{\gamma}_1 \\ \mathbf{0} \end{bmatrix}\right)^T \left(\mathbf{y} - X \begin{bmatrix} \hat{\gamma}_1 \\ \mathbf{0} \end{bmatrix}\right) &= (\mathbf{y} - X_1\hat{\gamma}_1)^T(\mathbf{y} - X_1\hat{\gamma}_1) \\ &= SS_{Res} \text{ (first model)}. \end{aligned}$$

But the least squares estimates $\mathbf{b}$ of $\boldsymbol{\beta}$ minimise the residual sum of squares for the second model, so we get

$$SS_{Res} \text{ (second model)} \leq SS_{Res} \text{ (first model)}.$$

4. 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 in terms of the other variables.

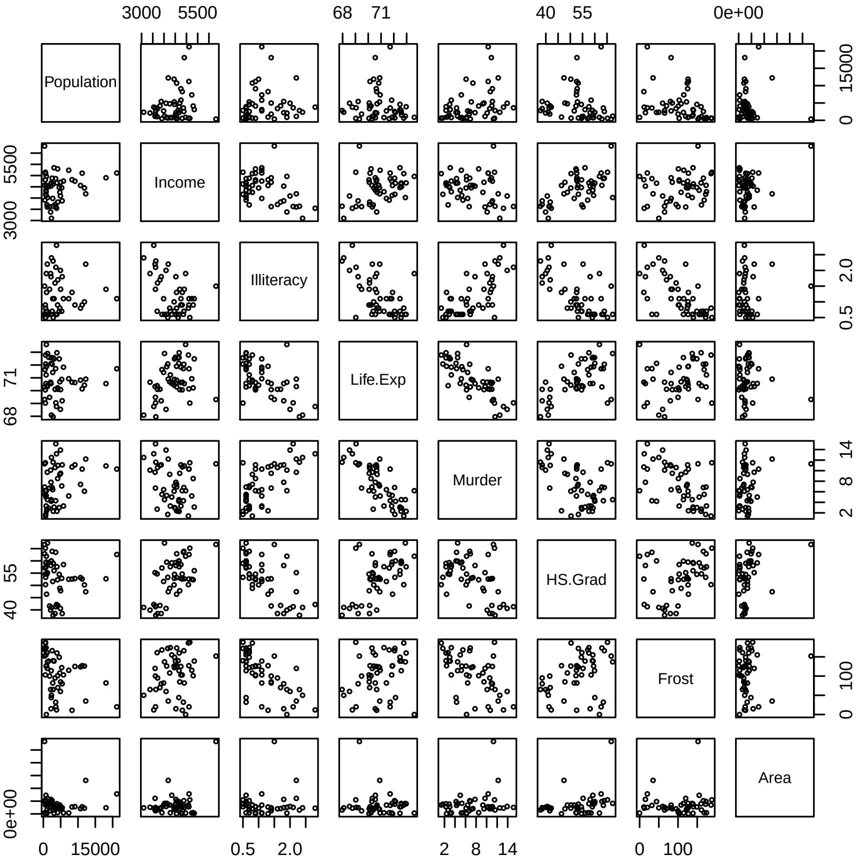
(a) Plot the data and comment. Should we consider any variable transformations?

Solution [3 marks]: Looking at murder rate against the other variables, there is evidence of a linear relationship with income, illiteracy, life expectancy, high school grad 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 high school grad, and non-linearity in illiteracy, but neither enough for immediate concern.

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



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

```
> 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
```

Solution [3 marks]: The final variables are life expectancy, frost, log(area), population, and illiteracy.

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

```
> 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
```

Solution [3 marks]: The model is the same as that found by forward selection.

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

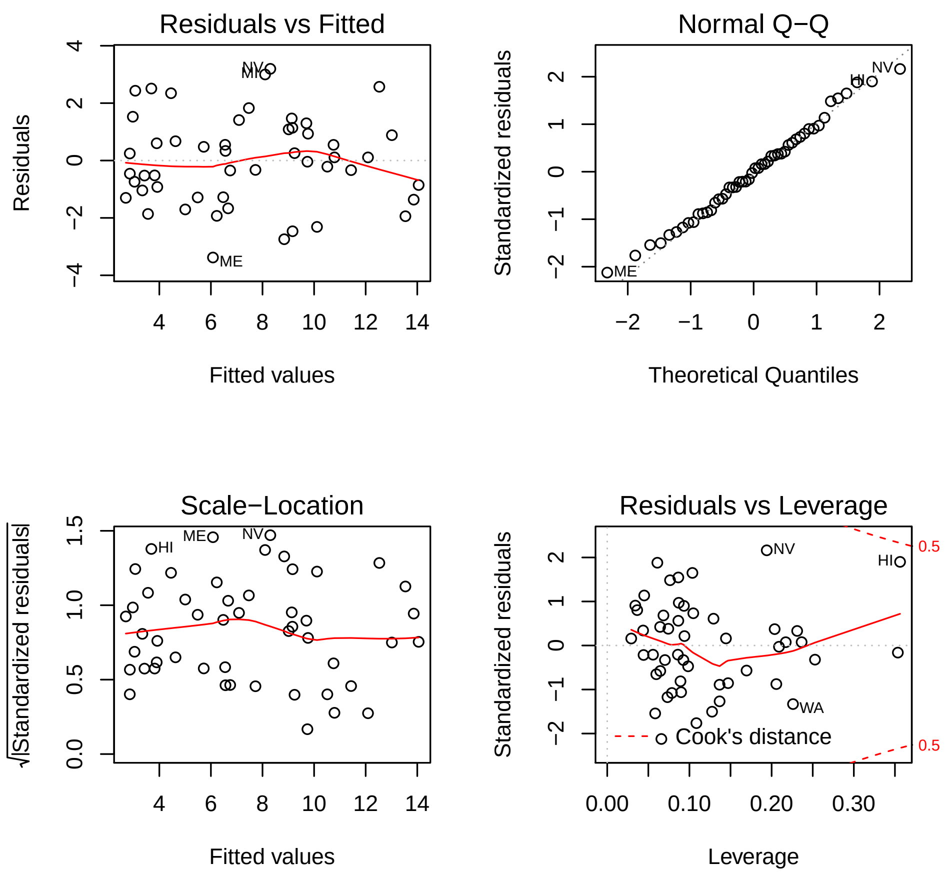
```
> model

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
```

Solution [1 mark]: The final model is

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

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

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



Solution [2 marks]: 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.

5. 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

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

Solution [4 marks]: We have

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

(b) Calculate the ridge regression estimates for the data from Q2 with penalty parameter $\lambda = 0.5$. 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).

Solution [3 marks]:

```
> Xs <- scale(X[,-1],center=T,scale=T)
> ys <- scale(y,center=T,scale=F)
> p <- p-1
> solve(t(Xs)%*%Xs + diag(rep(0.5,p)),t(Xs)%*%ys)
          [,1]
[1,] 0.3494789
[2,] 1.7899861
[3,] 0.3432961
```

(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.

Solution [5 marks]:

```
> lambda <- seq(0,1,0.001)
> 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] 0.136
```