MAST30025: Linear Statistical Models

Assignment 2 Solutions

Total marks: 40

1. Consider a general full rank linear model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $p > 2$ parameters. Derive an expression for a joint $100(1-\alpha)\%$ confidence region for parameters β_i and β_j, where i and j are arbitrary.

   Solution [5 marks]: Suppose without loss of generalisation that $i = 0$ and $j = 1$ (we can re-label the parameters to achieve this). We know that $(b_0, b_1) \sim MVN((\beta_0, \beta_1), A\sigma^2)$, where A is the 2×2 principal minor of $(X^TX)^{-1}$ (suitably re-ordered). Therefore the quadratic form

   $$\frac{\begin{bmatrix} b_0 - \beta_0 & b_1 - \beta_1 \end{bmatrix} A^{-1} \begin{bmatrix} b_0 - \beta_0 \\ b_1 - \beta_1 \end{bmatrix}}{\sigma^2}$$

   has a χ^2 distribution with 2 degrees of freedom. It is also independent of s^2 as it depends only on elements of $\mathbf{b}$. Following the derivation in the notes, we get the joint confidence region

   $$\begin{bmatrix} b_0 - \beta_0 & b_1 - \beta_1 \end{bmatrix} A^{-1} \begin{bmatrix} b_0 - \beta_0 \\ b_1 - \beta_1 \end{bmatrix} \leq 2s^2 f_\alpha,$$

   where f_α is the critical value of an F distribution with 2 and $n - p$ degrees of freedom.

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 error 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) Calculate 95% confidence intervals for the model parameters.

Solution [2 marks]:

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

[1] -13.8196491  -0.9893101

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

[1] -0.1525428  0.3940720

> (ci3 <- b[3] + c(-1,1)*qt(0.975,n-p)*sqrt(s2*C[3,3]))

[1] 0.6817719 1.5531974

> (ci4 <- b[4] + c(-1,1)*qt(0.975,n-p)*sqrt(s2*C[4,4]))

[1] -0.2563181  1.0285593
```

(c) In a year with 8% unemployment rate and 3.5% interest rate, we price a car at \$12,000 and observe that 7,000 cars are sold. Is this an atypical year (according to your model)?

Solution [2 marks]: We find a prediction interval for this year:

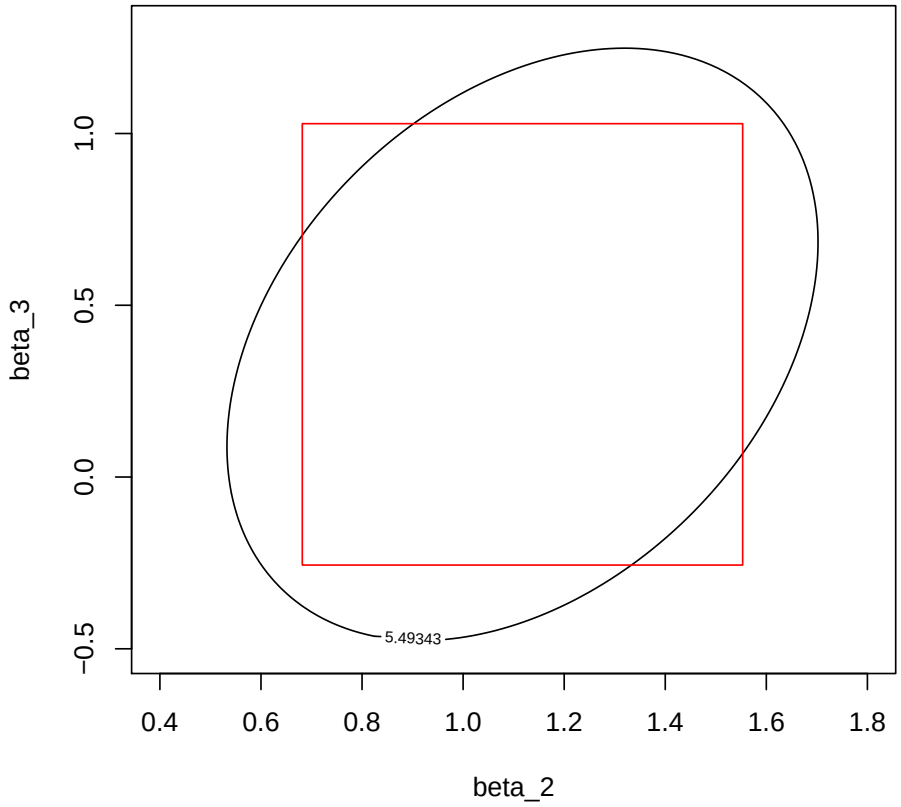
```
> xst <- c(1,12,8,3.5)
> xst%*%b + c(-1,1)*qt(0.975,n-p)*sqrt(s2*(1+t(xst)%*%C%*%xst))

[1] 1.444686 7.227303
```

7 lies within the boundaries of our prediction interval, so it is not an atypical year (with 95% confidence).

(d) Using your answer from question 1, find and draw a joint 95% confidence region for the parameters corresponding to unemployment rate and interest rate. Superimpose a rectangle corresponding to the confidence intervals found in (b).

Solution [3 marks]:

```
> b2 <- seq(0.4,1.8,length=100)
> b3 <- seq(-0.5,1.3,length=100)
> A <- solve(t(X)%*%X)[3:4,3:4]
> f <- function(beta2, beta3) {
+     f.out <- rep(0, length(beta2))
+     for (i in 1:length(beta2)) {
+         beta <- matrix(c(beta2[i], beta3[i]), 2, 1)
+         f.out[i] <- t(c(b[3],b[4]) - beta) %*% solve(A) %*% (c(b[3],b[4]) - beta)
+     }
+     return(f.out)
+ }
> z <- outer(b2, b3, f)
> contour(b2, b3, z, levels=2*s2*qf(0.95, 2, n-p),xlab='beta_2',ylab='beta_3')
> rect(ci3[1],ci4[1],ci3[2],ci4[2],border='red')
```



(e) Do you expect the confidence region to be larger or smaller than the rectangle? Justify your answer.

Solution [2 marks]: In general you would expect the region to be larger; assuming the parameter estimators are independent, you would expect the true parameters to lie in the rectangle $0.95^2 \approx 90.2\%$ of the time, while they should lie in the region 95% of the time. The estimators are *not* independent, but the correction factor involved is usually quite small.

(f) (Bonus) What is the probability that the true parameters for unemployment rate and interest rate (jointly) lie in the rectangle you drew in (d)?

Solution [2 marks]: It is not immediately obvious what the distribution of the estimators is, since we do not know the true σ^2. A good estimate would be to replace σ^2 by s^2 and use a multivariate t-distribution:

```
> library(mvtnorm)
> pmvt(lower=c(ci3[1],ci4[1])-b[3:4],upper=c(ci3[2],ci4[2])-b[3:4],sigma=A*s2,df=n-p)
[1] 0.9134779
attr(,"error")
[1] 1e-15
attr(,"msg")
[1] "Normal Completion"
```

Thus we expect the true parameters to lie in the rectangle with about 91.3% probability.

3. 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

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

Solution [5 marks]: 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.)

4. For this question we use the data set `UCD.csv` (available on the LMS). This data set, collected on 158 UC Davis students (self-reported), includes the following variables:

   ID = the ID for that student

   alcohol = average number of alcoholic drinks consumed per week

   exercise = average hours per week the student exercises

   height = the student's height (in inches)

   male = indicator variable, 1 if male and 0 if female

   dadht = the student's father's height

   momht = the student's mother's height

   We seek to predict a person's height, based on the given data.

   (a) Fit a linear model using all of the variables (except ID).
   Solution [2 marks]:

```
> UCD <- read.csv('../data/UCD.csv')
> model <- lm(height ~ . - ID, data=UCD)
> summary(model)

Call:
lm(formula = height ~ . - ID, data = UCD)

Residuals:
     Min       1Q   Median       3Q      Max
-10.2166  -1.4627   0.0494   1.4502   6.2616

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 20.94900    5.08126   4.123 6.14e-05 ***
alcohol      0.05068    0.02616   1.938 0.054524 .
exercise    -0.05442    0.04501  -1.209 0.228506
male         5.16073    0.39794  12.969  < 2e-16 ***
dadht        0.38182    0.05394   7.079 5.01e-11 ***
momht        0.27060    0.07061   3.832 0.000185 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.318 on 152 degrees of freedom
Multiple R-squared:  0.6708,        Adjusted R-squared:  0.6599
F-statistic: 61.94 on 5 and 152 DF,  p-value: < 2.2e-16
```

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

Solution [2 marks]:

```
> nullmodel <- lm(height ~ 1, data=UCD)
> anova(nullmodel, model)
Analysis of Variance Table

Model 1: height ~ 1
Model 2: height ~ (ID + alcohol + exercise + male + dadht + momht) - ID
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1    157 2481.22
2    152  816.89  5    1664.3 61.937 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We strongly reject the null hypothesis of model irrelevance.

(c) Use forward selection with F tests to select variables for your model.

Solution [3 marks]:

```
> add1(nullmodel, scope ~ .+alcohol+exercise+male+dadht+momht, test='F')
Single term additions

Model:
height ~ 1
         Df Sum of Sq    RSS    AIC  F value    Pr(>F)
<none>                2481.2 437.12
alcohol   1    469.91 2011.3 405.94  36.4470 1.103e-08 ***
exercise  1     33.51 2447.7 436.97   2.1358    0.1459
male      1   1054.11 1427.1 351.73 115.2266 < 2.2e-16 ***
dadht     1    407.47 2073.8 410.77  30.6525 1.279e-07 ***
momht     1    287.01 2194.2 419.69  20.4056 1.230e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> add1(lm(height~male,data=UCD), scope ~ .+alcohol+exercise+dadht+momht, test='F')
Single term additions

Model:
height ~ male
         Df Sum of Sq     RSS    AIC F value    Pr(>F)
<none>                1427.11 351.73
alcohol   1    152.46 1274.65 335.88 18.5391 2.941e-05 ***
exercise  1      0.67 1426.44 353.65  0.0732    0.7871
dadht     1    491.25  935.86 287.06 81.3632 6.773e-16 ***
momht     1    254.99 1172.12 322.63 33.7192 3.492e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> add1(lm(height~male+dadht,data=UCD), scope ~ .+alcohol+exercise+momht, test='F')
Single term additions

Model:
height ~ male + dadht
         Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>                935.86 287.06
alcohol   1    29.862 905.99 283.94  5.0758   0.02567 *
exercise  1     3.866 931.99 288.41  0.6387   0.42540
momht     1    95.253 840.60 272.10 17.4505 4.923e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> add1(lm(height~male+dadht+momht,data=UCD), scope ~ .+alcohol+exercise, test='F')

Single term additions

Model:
height ~ male + dadht + momht
         Df Sum of Sq    RSS    AIC F value  Pr(>F)
<none>                840.60 272.10
alcohol   1   15.8607 824.74 271.09  2.9424 0.08831 .
exercise  1    3.5405 837.06 273.43  0.6471 0.42239
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We select the `male`, `dadht` and `momht` variables.

(d) Starting from a full model, use stepwise selection with AIC to select variables for your model. Use this as your final model; comment briefly on the variables included.

Solution [3 marks]:

```
> finalmodel <- step(model, scope = ~ .)

Start:  AIC=271.58
height ~ (ID + alcohol + exercise + male + dadht + momht) - ID

           Df Sum of Sq     RSS    AIC
- exercise  1      7.86  824.74 271.09
<none>                   816.89 271.58
- alcohol   1     20.18  837.06 273.43
- momht     1     78.93  895.82 284.15
- dadht     1    269.34 1086.22 314.60
- male      1    903.86 1720.74 387.29

Step:  AIC=271.09
height ~ alcohol + male + dadht + momht

           Df Sum of Sq     RSS    AIC
<none>                   824.74 271.09
+ exercise  1      7.86  816.89 271.58
- alcohol   1     15.86  840.60 272.10
- momht     1     81.25  905.99 283.94
- dadht     1    270.17 1094.91 313.86
- male      1    897.14 1721.88 385.40
```

We select the previous variables, together with the `alcohol` variable. It makes complete sense that gender and parents' heights affect one's height. We also have the humorous inference that drinking more makes you taller, which means it is time to pull out that old maxim "correlation does not equal causation"!

(e) Using your final model, test whether the parameters corresponding to father's and mother's heights are equal.

Solution [2 marks]: We can test this with a general linear hypothesis.

```
> library(car)
> linearHypothesis(finalmodel, c(0,0,0,1,-1), 0)

Linear hypothesis test

Hypothesis:
dadht - momht = 0

Model 1: restricted model
Model 2: height ~ alcohol + male + dadht + momht
```

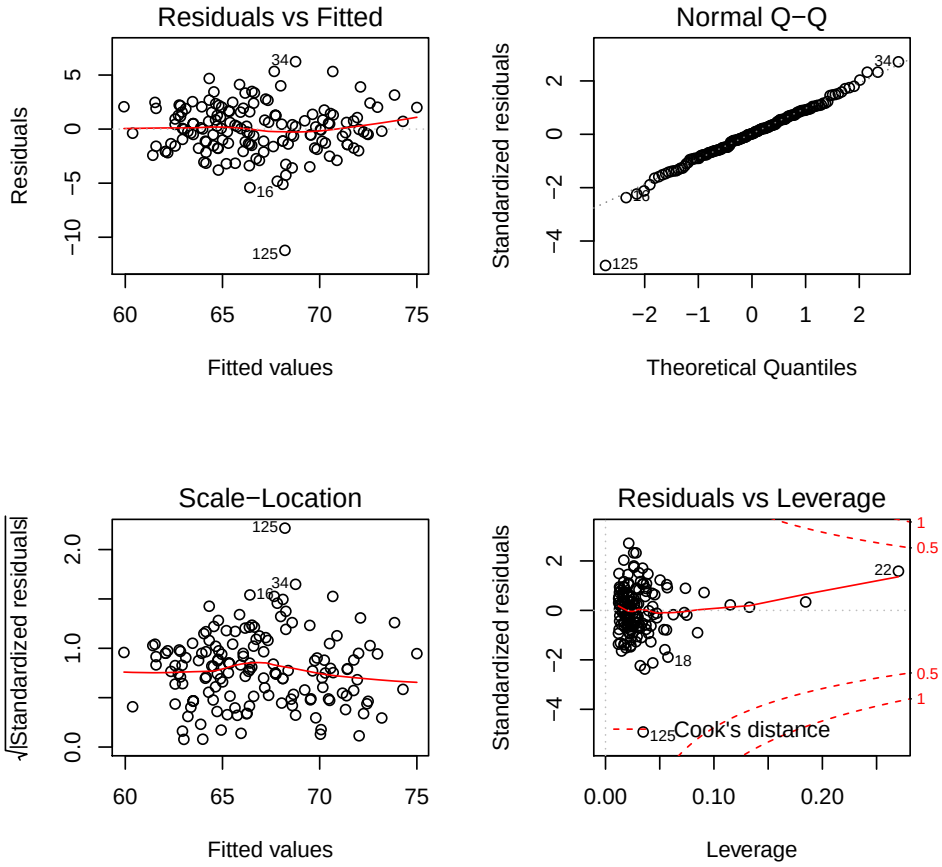
```
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    154 831.20
2    153 824.74  1     6.453 1.1971 0.2756
```

We cannot reject the hypothesis that they are equal.

(f) Comment on the suitability of your final model, using diagnostic plots.

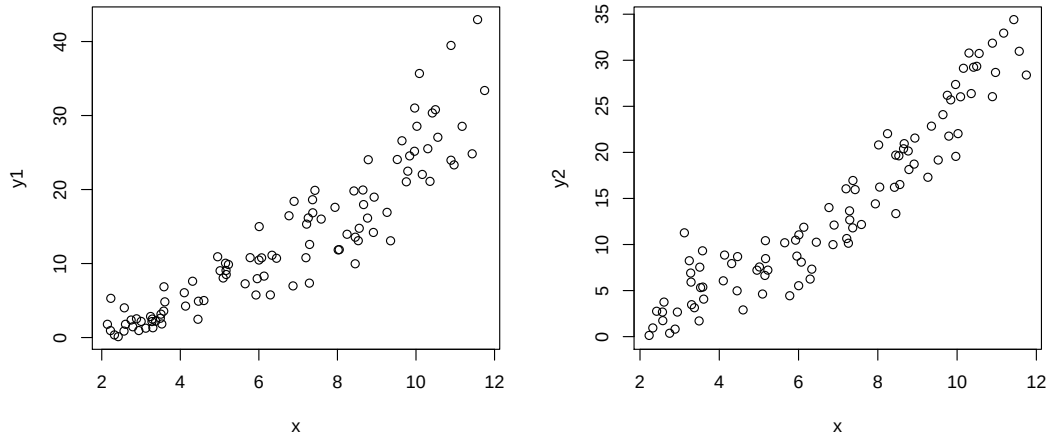
Solution [2 marks]:

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



Apart from point 125, which looks somewhat like an outlier and a potential candidate for removal, the model assumptions seem well satisfied.

5. Suppose that we have a response variable y which is known to have a quadratic relationship with a predictor variable x. Explain all of the differences between fitting a linear model of y against x and x^2, versus a linear model of $\sqrt{y}$ against x. Which would you use for the two datasets shown below?



Solution [5 marks]: When we fit a model of y against x and x^2, the model is

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon,$$

whereas when we fit a model of $\sqrt{y}$ against x, the model is

$$\begin{aligned}\sqrt{y} &= \beta_0 + \beta_1 x + \varepsilon \\ y &= \beta_0^2 + 2\beta_0\beta_1 x + \beta_1^2 x^2 + 2\beta_0\varepsilon + 2\beta_1 x\varepsilon + \varepsilon^2.\end{aligned}$$

The principal differences are (a) the first model has an extra degree of freedom, which allows us to represent a greater range of quadratic relationships; and (b) the error term in the first model is additive and normal, whereas the error term in the second model has a multiplicative component and is not normal. For the data shown, y_2 has an additive (constant variance) error, while the error in y_1 clearly scales with x (or y). Thus we would use a model of $\sqrt{y_1}$ against x, and a model of y_2 against x and x^2.