MAST30025: Linear Statistical Models

Assignment 2 Solutions

Total marks: 42

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: 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. *[1]*

   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. *[1]*

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:

```
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
```

[2]

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

```
(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
```

Parameters β_0 (intercept) and β_3 (interest rate) have the highest covariance in magnitude.

[2]

(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:

```
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)
```

```
## Warning in c(-1, 1) * qt(0.995, df = n - p) * sqrt(s2 * t(xst) %*% C %*% :  Recycling
array of length 1 in vector-array arithmetic is deprecated.
##  Use c() or as.vector() instead.
## Warning in xst %*% b + c(-1, 1) * qt(0.995, df = n - p) * sqrt(s2 * t(xst) %*%
:  Recycling array of length 1 in array-vector arithmetic is deprecated.
##  Use c() or as.vector() instead.
## [1] 3.926075 7.173129
```

[2]

(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: Let α be the level used. Then

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

```
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
```

The confidence level is 90%. *[2]*

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

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

```
##          [,1]
## [1,] 23.47683

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

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

We reject the null hypothesis of model irrelevance. *[2]*

3. 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:

```
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
```

[2]

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

```
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. *[2]*

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

```
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. *[2]*

(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 :

```
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"! *[2]*

(e) Test whether the parameters corresponding to father's and mother's heights are equal.

Solution: 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
```

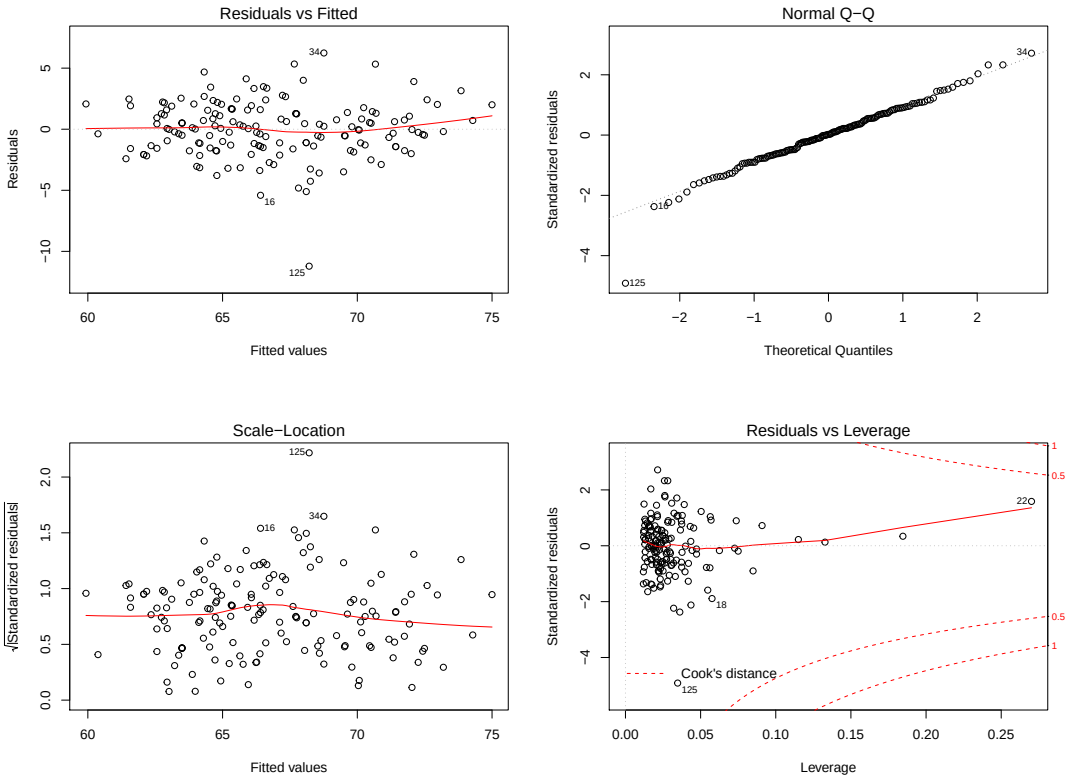
```
##
## Model 1: restricted model
## Model 2: height ~ alcohol + male + dadht + momht
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    154 831.20
## 2    153 824.74  1     6.453 1.1971 0.2756
```

We cannot reject the hypothesis that they are equal. *[2]*

(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. *[2]*

4. A study was conducted to determine the effect of the size of the root system on the growth of Douglas-fir seedlings when they are planted out. Seedlings were obtained from three seed lots, and when they were planted out their root volume was classified as small (RV1), medium (RV2), or large (RV3). The heights of the seedlings were then measured at the end of the first growing season. The data from the experiment is given in the file `douglas.csv`.

(a) Fit a linear model with interaction to the data. Calculate a confidence interval for the difference between the heights of large (RV3) and medium (RV2) seedlings in the B349 seed lot.

Solution:

```
douglas <- read.csv('../data/douglas.csv')
douglas$RootVolume <- factor(douglas$RootVolume)
douglas$SeedLot <- factor(douglas$SeedLot)
imodel <- lm(Height ~ RootVolume * SeedLot, data=douglas)
library(gmodels)
ci <- estimable(imodel, c(0,-1,1,0,0,-1,1,0,0), conf.int=0.95)
c(ci$Lower, ci$Upper)
```

```
## [1] 1.437703 5.495631
```

[2]

(b) Is it possible to estimate, from this model, an overall difference between the heights of large and medium seedlings?

Solution: No; we assume that there is interaction with the seed lots and so an "overall" difference is meaningless. *[2]*

(c) Test the hypothesis that the height of seedlings from the J052 plot has no dependence on root volume.

Solution:

```
library(car)
(C <- matrix(c(0,1,0,0,0,0,0,1,0,0,0,1,0,0,0,0,0,1), 2, 9, byrow=T))
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
## [1,]    0    1    0    0    0    0    0    1    0
## [2,]    0    0    1    0    0    0    0    0    1
```

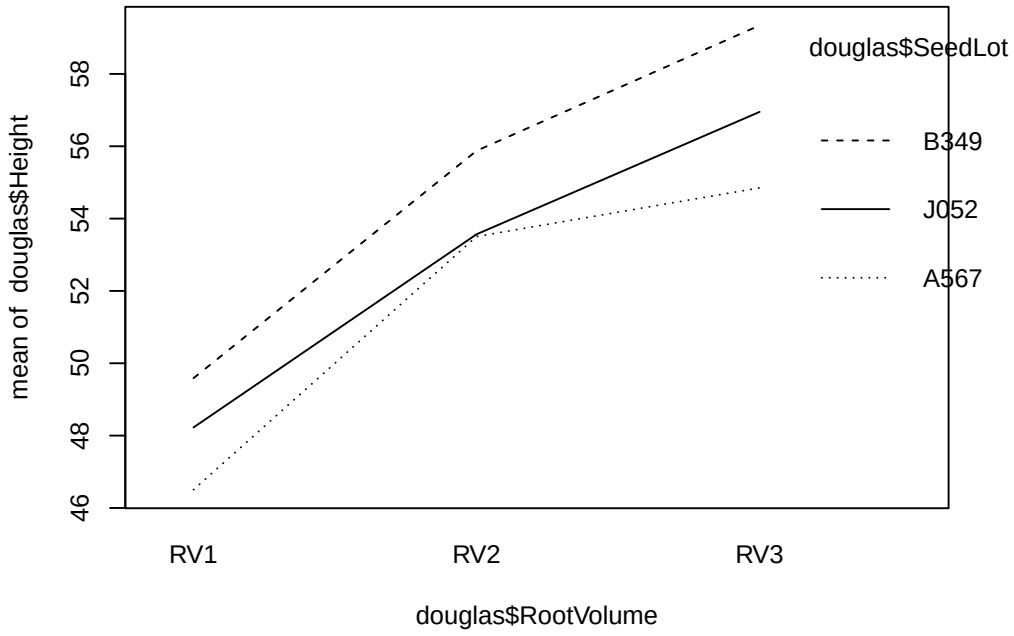
```
linearHypothesis(imodel, C)
```

```
## Linear hypothesis test
##
## Hypothesis:
## RootVolumeRV2  + RootVolumeRV2:SeedLotJ052 = 0
## RootVolumeRV3  + RootVolumeRV3:SeedLotJ052 = 0
##
## Model 1: restricted model
## Model 2: Height ~ RootVolume * SeedLot
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     47 369.12
## 2     45 137.00  2    232.12 38.122 2.065e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We reject the null hypothesis; there is a dependence for the J052 plot. *[2]*

(d) Generate an interaction plot for the data. Is there any evidence of an interaction?

```
interaction.plot(douglas$RootVolume, douglas$SeedLot, douglas$Height)
```



The lines are close to parallel, so there is very little indication of an interaction. *[2]*

(e) Test for the presence of interaction between root volume and seed lot.

Solution:

```
amodel <- lm(Height ~ RootVolume + SeedLot, data=douglas)
anova(amodel, imodel)

## Analysis of Variance Table
##
## Model 1: Height ~ RootVolume + SeedLot
## Model 2: Height ~ RootVolume * SeedLot
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     49 147.67
## 2     45 137.00  4    10.674 0.8765 0.4855
```

There is no evidence that there is interaction. *[2]*

(f) Perform forwards selection to determine a final model.

Solution:

```
nullmodel <- lm(Height ~ 1, data=douglas)
add1(nullmodel, scope = ~ . + RootVolume + SeedLot, test='F')

## Single term additions
##
## Model:
## Height ~ 1
##            Df Sum of Sq     RSS     AIC F value    Pr(>F)
## <none>                  1003.83 159.820
## RootVolume  2    755.68  248.16  88.354 77.6509 3.336e-16 ***
## SeedLot     2    100.48  903.35 158.125  2.8365   0.06791 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

model2 <- lm(Height ~ RootVolume, data=douglas)
add1(model2, scope = ~ . + SeedLot, test='F')
```

```
## Single term additions
##
## Model:
## Height ~ RootVolume
##         Df Sum of Sq    RSS    AIC F value Pr(>F)
## <none>               248.16 88.354
## SeedLot  2    100.48 147.67 64.325  16.671  3e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model3 <- lm(Height ~ RootVolume + SeedLot, data=douglas)
add1(model3, scope = ~ . + RootVolume*SeedLot, test='F')
```

```
## Single term additions
##
## Model:
## Height ~ RootVolume + SeedLot
##                     Df Sum of Sq    RSS    AIC F value Pr(>F)
## <none>                           147.67 64.325
## RootVolume:SeedLot  4    10.674 137.00 68.274  0.8765 0.4855
```

We use the full additive model with no interaction. *[2]*

(g) Is it possible to estimate, from this model, an overall difference between the heights of large and medium seedlings?

Solution: Yes; there is no interaction in this model and we assume that we can identify effects corresponding to root volume. *[2]*

5. You wish to perform a study to determine if 3 treatments each produce no effect using a completely randomised design. To do this, you will test the hypothesis $H_0 : \mu + \tau_1 = \tau_1 - \tau_2 = \tau_2 - \tau_3 = 0$. You are given resources to study 50 sample units.

(a) Determine the optimal allocation of the number of units to assign to each treatment.

Solution: As in the lecture notes we have

$$\text{var } \hat{\mu_i} = \sigma^2 \frac{1}{n_i}$$

and

$$\text{var } \widehat{\tau_j - \tau_i} = \sigma^2 \left(\frac{1}{n_j} + \frac{1}{n_i} \right),$$

and so we want to minimise

$$f(n_1, n_2, n_3, \lambda) = \sigma^2 \left(\frac{2}{n_1} + \frac{2}{n_2} + \frac{1}{n_3} \right) + \lambda \left(\sum_{i=1}^{3} n_i - n \right).$$

This gives

$$\begin{aligned}
\frac{\partial f}{\partial n_1} &= -2\frac{\sigma^2}{n_1^2} + \lambda = 0 \\
\frac{\partial f}{\partial n_2} &= -2\frac{\sigma^2}{n_2^2} + \lambda = 0 \\
\frac{\partial f}{\partial n_3} &= -\frac{\sigma^2}{n_3^2} + \lambda = 0 \\
n_1^2 &= 2\frac{\sigma^2}{\lambda} = n_2^2 = 2n_3^2 \\
n_1 &= n_2 = \sqrt{2} n_3.
\end{aligned}$$

This gives $n_1 = 19, n_2 = 18, n_3 = 13$ — due to rounding, one of n_1 or n_2 has to be larger. *[2]*

(b) Perform the random allocation. You must use R for randomisation and include your R commands and output.

Solution:

```
x <- sample(50, 50)
x[1:19]

##  [1] 45 46 10 17  8 43 41  1 33 36 13  2 25 50 32 28 37 23  7

x[20:37]

##  [1] 11 42  9 40 47 39 44  6 22 15 16 21  5 49 18 35 30 34

x[38:50]

##  [1] 38 12 31 26 27  4 14 48 19 24 29 20  3
```

[2]