Exercises

620–371 Linear Models
Department of Mathematics and Statistics
University of Melbourne
Parkville, Vic. 3010

June 10, 2008
Exercise 1

Heights of plants. Complete the ANOVA table.

```r
> plants <- read.csv("../data/plants.csv")
> str(plants)
'data.frame': 9 obs. of 3 variables:
$ moisture: int 204 121 261 460 468 299 308 235 188
$ height : int 22 13 24 35 29 27 29 18 23
$ variety : int 1 2 1 3 1 3 3 2 2
> summary(lm(height ~ moisture, data = plants))
Call:
lm(formula = height ~ moisture, data = plants)
Residuals:
   Min     1Q   Median     3Q    Max
-4.431 -3.605   1.370   1.957   3.327
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.73809   3.08227   3.484  0.01021 *
moisture     0.04849    0.01015   4.777  0.00202 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 3.378 on 7 degrees of freedom
Multiple R-squared: 0.7653,     Adjusted R-squared: 0.7317
F-statistic: 22.82 on 1 and 7 DF,  p-value: 0.002019
> anova(lm(height ~ moisture, data = plants))
Analysis of Variance Table

Response: height
          Df Sum Sq Mean Sq     F value  Pr(>F)
moisture  1 260.356 260.356 22.81900099 0.002019 **
Residuals 7  79.866  11.409
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Answer of Exercise 1

```r
> anova(lm(height ~ moisture, data = plants))
Analysis of Variance Table

Response: height
          Df Sum Sq Mean Sq     F value  Pr(>F)
moisture  1 260.356 260.356 22.81900099 0.002019 **
Residuals 7  79.866  11.409
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```
Exercise 2

Heights of plants by variety. Explain the meaning of all output.

<table>
<thead>
<tr>
<th>Variety</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>22</td>
<td>13</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>18</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>29</td>
<td>23</td>
<td>27</td>
</tr>
</tbody>
</table>

> plant.lm.1 <- lm(height ~ factor(variety), data = plants)
> summary(plant.lm.1)

Call:
  lm(formula = height ~ factor(variety), data = plants)

Residuals:
  Min 1Q Median 3Q Max
  -5  -3  -1  4  5

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                        25.000     2.480 10.082 5.53e-05 ***
factor(variety)2                   -7.000     3.507 -1.996 0.093     .
factor(variety)3                   -5.333     3.507  1.521 0.179     .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.295 on 6 degrees of freedom
Multiple R-squared: 0.6747,   Adjusted R-squared: 0.5663
F-statistic: 6.223 on 2 and 6 DF,  p-value: 0.03442

> anova(plant.lm.1)

Analysis of Variance Table

Response: height

                                   Df  Sum Sq Mean Sq F value Pr(>F)
factor(variety)                  2 229.556 114.778  6.223 0.03442 *
Residuals                         6 110.667 18.444
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Answer of Exercise 2

No answers provided.
Exercise 3

Below are three scatterplots and twelve numbers given in strictly increasing order. Use each of these numbers once only to fill in the table.

\[-24\ -7\ -0.7\ 0\ 0\ 0.9\ 2\ 4\ 6\ 9\ 13\ 47\]

<table>
<thead>
<tr>
<th>(y_1) versus (x_1)</th>
<th>(y_2) versus (x_2)</th>
<th>(y_3) versus (x_3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>slope</td>
<td></td>
<td></td>
</tr>
<tr>
<td>intercept</td>
<td></td>
<td></td>
</tr>
<tr>
<td>correlation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>residual sd</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Answer of Exercise 3
<table>
<thead>
<tr>
<th>y_1 \text{ versus } x_1</th>
<th>y_2 \text{ versus } x_2</th>
<th>y_3 \text{ versus } x_3</th>
</tr>
</thead>
<tbody>
<tr>
<td>slope</td>
<td>6</td>
<td>-7</td>
</tr>
<tr>
<td>intercept</td>
<td>-24</td>
<td>13</td>
</tr>
<tr>
<td>correlation</td>
<td>0.9</td>
<td>-0.7</td>
</tr>
<tr>
<td>residual sd</td>
<td>4</td>
<td>9</td>
</tr>
</tbody>
</table>

**Exercise 4**

If \( \text{var}(X) = 2 \), \( \text{var}(Y) = 3 \) and \( \text{cov}(X, Y) = -1 \), find:

1. \( \text{var}(2X + 3Y) \)
2. \( \text{var}(2X - 3Y) \)

**Answer of Exercise 4**

```r
> a <- 2
> b <- 3
> a^2 * 2 + b^2 * 3 + 2 * -1 * a * b
[1] 23

> a^2 * 2 + b^2 * 3 - 2 * -1 * a * b
[1] 47
```
Exercise 5

If \( \text{var}(X) = 2\), \( \text{var}(Y) = 3 \) and \( \text{correlation}(X, Y) = 0.8 \), find:

1. \( \text{var}(2X + 3Y) \)
2. \( \text{var}(2X - 3Y + 4) \)

**Answer of Exercise 5**

```r
> a <- 2
> b <- 3
> cov <- 0.8 * sqrt(2) * sqrt(3)
> a^2 * 2 + b^2 * 3 + 2 * cov * a * b
[1] 58.5151
> a^2 * 2 + b^2 * 3 - 2 * cov * a * b
[1] 11.48490
```

Exercise 6

Let \( Y = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \), with \( D(Y) = \begin{bmatrix} 2 & 3 \\ 3 & 5 \end{bmatrix} \).

Find:

(i) \( \text{var}(2Y_1 + 6Y_2) \)

(ii) \( \text{var}(2Y_1 - 6Y_2) \)

**Answer of Exercise 6**

(i) \( > \text{var.Y} <- \text{matrix}(c(2, 3, 3, 5), \text{nrow} = 2) \)

\( > \text{m.i} <- \text{c}(2, 6) \)

\( > \text{t(m.i) } \%\% \text{ var.Y } \%\% \text{ m.i} \)

\( [,1] \)

\( [1,] \ 260 \)

(ii) \( > \text{m.i} <- \text{c}(2, -6) \)

\( > \text{t(m.i) } \%\% \text{ var.Y } \%\% \text{ m.i} \)

\( [,1] \)

\( [1,] \ 116 \)

Exercise 7

If \( Y = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \), show that \( D(Y) \) cannot be \( \begin{bmatrix} 2 & 3 \\ 3 & 4 \end{bmatrix} \).

(Hint: find the value of \( a \) for which \( \text{var}(Y_1 + aY_2) \) is a minimum, and the corresponding value of the ‘variance’.)

**Answer of Exercise 7**

\( \text{var} = \text{minimum when } a = -\frac{3}{4} \Rightarrow \text{var}(Y_1 + aY_2) = -\frac{1}{4} \) which is not possible; \( \text{variance must be } \geq 0 \).

Alternatively find the correlation of \( Y_1 \) and \( Y_2 \):
\[3/(\sqrt{4} \times \sqrt{2})\]

\[1.060660\]

\[-1 \leq \rho \leq 1\), so this covariance matrix is impossible.

**Exercise 8**

If \(Y \sim \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix}\), with \(D(Y) = \begin{bmatrix} 2 & 1 & -1 \\ 1 & 3 & 2 \\ -1 & 2 & 4 \end{bmatrix}\) and \(Z = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = \begin{bmatrix} Y_1 + 2Y_2 + 3Y_3 \\ 3Y_1 - Y_2 - Y_3 \end{bmatrix}\), find \(D(Z)\).

**Answer of Exercise 8**

\[
\begin{bmatrix}
2 & 1 & -1 \\
1 & 3 & 2 \\
-1 & 2 & 4 \\
\end{bmatrix}
\]

\[
\begin{bmatrix}
72 & -25 \\
-25 & 29 \\
\end{bmatrix}
\]

**Exercise 9**

Let \(Y \sim \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix}\), with \(D(Y) = \begin{bmatrix} 5 & 2 & -2 \\ 2 & 4 & 1 \\ -2 & 1 & 3 \end{bmatrix}\).

1. Find \(\text{var}(P'Y)\) when:

(a) \(P = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}\)

(b) \(P = \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}\)
2. Find $\mathcal{D}(\bar{Z})$ where $\bar{Z} = MY$ and $M = \begin{bmatrix} -1 & 0 & 1 \\ 1 & -2 & 1 \\ 1 & 1 & -2 \\ 2 & -1 & -1 \end{bmatrix}$

Answer of Exercise 9

> (var.Y <- matrix(c(5, 2, -2, 2, 4, 1, -2, 1, 3), nrow = 3))

\[
\begin{bmatrix}
[1,] & 5 & 2 & -2 \\
[2,] & 2 & 4 & 1 \\
[3,] & -2 & 1 & 3 \\
\end{bmatrix}
\]

> P.a <- c(-1, 0, 1)
> t(P.a) %*% var.Y %*% P.a

\[
\begin{bmatrix}
[1,] \\
[1,] & 12 \\
\end{bmatrix}
\]

> P.b <- c(1, -2, 1)
> t(P.b) %*% var.Y %*% P.b

\[
\begin{bmatrix}
[1,] \\
[1,] & 8 \\
\end{bmatrix}
\]

var = 12, 8.

> M <- matrix(c(-1, 0, 1, -2, 1, 1, -2, 1, -2, 2, -1, -1), nrow = 4, + byrow = TRUE)
> M %*% var.Y %*% t(M)

\[
\begin{bmatrix}
[1,] & 12 & 0 & -18 & -18 \\
[2,] & 0 & 8 & -4 & 4 \\
[3,] & -18 & -4 & 29 & 25 \\
[4,] & -18 & 4 & 25 & 29 \\
\end{bmatrix}
\]

\[\mathcal{D}(\bar{Z}) = \begin{bmatrix} 12 & 0 & -18 & -18 \\ 0 & 8 & -4 & 4 \\ -18 & -4 & 29 & 25 \\ -18 & 4 & 25 & 29 \end{bmatrix}\]

Exercise 10

Let $\hat{B} = (X'G^{-1}X)^{-1}X'G^{-1}Y$, and $\mathcal{D}(\hat{Y}) = \sigma^2G$, where the matrices $X$ and $G$ consist of known constants, and $\bar{Y}$ is a vector of random variables. Find $\mathcal{D}(\hat{B})$.

Answer of Exercise 10

\[\mathcal{D}(\hat{B}) = \sigma^2(X'G^{-1}X)^{-1}\]
Exercise 12

Let

\[
\begin{bmatrix}
0 & 1 & -1 \\
1 & 1 & 0 \\
1 & 0 & 1 \\
1 & 2 & -1 \\
2 & 1 & 1 \\
3 & 1 & 2 \\
3 & 2 & 1 \\
\end{bmatrix}
\]

and

\[
\begin{bmatrix}
-2.2 \\
5.8 \\
6.8 \\
1.8 \\
11.7 \\
17.4 \\
17.3 \\
\end{bmatrix}
\]

1. Show that \( \text{rank}(X) < 3 \).

2. Derive the normal equations.

3. Find two (of the infinitely many) solutions to the normal equations.

4. Find \( \hat{y} = X\hat{\beta} \) for each of the solutions you found in (c).

5. For which \( P \) (one only) is the parametric function \( P^T\hat{\beta} \) estimable?

\[
\begin{align*}
(P_1) & \quad P = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \\
(P_2) & \quad P = \begin{bmatrix} 2.3 \\ 5.4 \\ -3.1 \end{bmatrix}, \\
(P_3) & \quad P = \begin{bmatrix} 6.7 \\ -4.3 \\ 4.2 \end{bmatrix}.
\end{align*}
\]

Answer of Exercise 12

1. \( X_1 - X_2 - X_3 = 0 \) so \( X \) is not of full rank.

2. \( X <- \text{matrix}(c(0, 1, -1, 1, 0, 1, 0, 1, 1, 1, -1, 2, -1, 2, 1, 1, \\
+ 3, 1, 2, 3, 2, 1), \text{ncol} = 3, \text{byrow} = \text{TRUE}) \\
Y <- \text{c}(-2.2, 5.8, 6.8, 1.8, 11.7, 17.4, 17.3) \)

\[
\begin{array}{ccc}
[1,] & 25 & 14 & 11 \\
[2,] & 14 & 12 & 2 \\
[3,] & 11 & 2 & 9 \\
\end{array}
\]

\( \text{XpY <- t(X) \%\% Y} \)

\[
\begin{array}{c}
[.1] \\
[1,] & 141.9 \\
[2,] & 70.9 \\
[3,] & 71.0 \\
\end{array}
\]

Note: row 1 = row 2 + row 3.

3. \( \hat{\beta} = \begin{bmatrix} 6.8288 \\ -2.0587 \end{bmatrix}, \begin{bmatrix} 4.7702 \\ 0 \end{bmatrix}, \begin{bmatrix} 4.7702 \\ 0 \end{bmatrix} \) or \( \hat{\beta} = \begin{bmatrix} 0 \\ 6.8288 \end{bmatrix}, \begin{bmatrix} 0 \\ 2.0587 \end{bmatrix} \) or \( \hat{\beta} = \begin{bmatrix} 4.7702 \\ 0 \end{bmatrix} \) or \( \hat{\beta} = \begin{bmatrix} 0 \\ 6.8288 \end{bmatrix} \)

4. \( \hat{\beta}' = [-2.0587, 4.7702, 6.8288, 2.7115, 11.5990, 18.4279, 16.3692] \) for all \( \hat{\beta} \).

5. (Answer = (ii) )
Exercise 13

For a linear model \( y = X\beta + \varepsilon \) with \( E(\varepsilon) = N(0, \sigma^2 I) \), the deviance \( d(\hat{\beta}) = 42 \) on 13 df. In order to test a (null) hypothesis about \( \beta \), the model \( y = X_0\beta_0 + \varepsilon \), which assumes that \( H_0 \) is true, was fitted. For this model \( d(\hat{\beta}_0) = 56 \) on 15 df. Complete the test and state your conclusion.

Answer of Exercise 13

\[
> (F <- (56 - 42)/2/(42/13))
\]

[1] 2.16667

\[
> 1 - pf(F, 2, 13)
\]

[1] 0.1541339

F = 2.1667; do not reject \( H_0 \).
**Exercise 14**

For a linear model \( y = X\beta + \varepsilon \) with \( E\varepsilon = N(0, \sigma^2 I) \),

\[
\hat{\beta} = \begin{bmatrix}
\beta_1 \\
\beta_2 \\
\beta_3 \\
\beta_4
\end{bmatrix} = \begin{bmatrix}
9.90 \\
6.57 \\
5.60 \\
12.20
\end{bmatrix}
\]

and

\[
D(\hat{\beta}) = \begin{bmatrix}
5.29 & 1.51 & 2.27 & -3.02 \\
1.51 & 5.29 & 2.27 & -3.02 \\
2.27 & 2.27 & 9.06 & -4.53 \\
-3.02 & -3.02 & -4.53 & 6.04
\end{bmatrix}
\]

based on a sample of 20 observations. Carry out a test of \( H_0 : \beta_2 = (\beta_1 + \beta_3)/2 \).

**Answer of Exercise 14**

\[
> beta.hat <- c(9.9, 6.57, 5.60, 12.20) \\
> D.beta <- matrix(c(5.29, 1.51, 2.27, -3.02, 1.51, 5.29, 2.27, -3.02, \\
+ 2.27, 2.27, 9.06, -4.53, -3.02, -3.02, -4.53, 6.04), nrow=4) \\
> my.contrast <- c(0.5, -1, 0.5, 0) \\
> sum(beta.hat*my.contrast) / \\
+ sqrt(t(my.contrast) %*% D.beta %*% my.contrast)
\]

\[
[,1]
[1,] 0.4726622
\]

t = 0.47; do not reject \( H_0 \).

**Exercise 29**

Carry out a (full) analysis of the following data, and state your conclusions.

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>A</td>
<td>24</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>64</td>
</tr>
<tr>
<td>2</td>
<td>74</td>
<td>135</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>109</td>
</tr>
</tbody>
</table>

\( A \) and \( B \) are factors with 2 and 3 levels, respectively.

**Answer of Exercise 29**

\[
> y29 <- c(24, 17, 74, 80, 49, 164, 135, 109, 174, 236, 423, 388) \\
> A29 <- rep(rep(1:2,each=2),3) \\
> B29 <- rep(1:3,each=4) \\
> q29.1 <- lm(y29~factor(A29)*factor(B29)) \\
> q29.2 <- lm(sqrt(y29)~factor(A29)*factor(B29)) \\
> q29.3 <- lm(log(y29)~factor(A29)*factor(B29))
\]

Check assumptions for all three:
Square root is best.

```r
> q29.4 <- lm(sqrt(y29) ~ factor(A29) + factor(B29))
> anova(q29.4)
```

Analysis of Variance Table

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>factor(A29)</td>
<td>1</td>
<td>62.086</td>
<td>62.086</td>
<td>69.777</td>
<td>3.195e-05 ***</td>
</tr>
<tr>
<td>factor(B29)</td>
<td>2</td>
<td>241.971</td>
<td>120.986</td>
<td>135.973</td>
<td>6.669e-07 ***</td>
</tr>
<tr>
<td>Residuals</td>
<td>8</td>
<td>7.118</td>
<td>0.890</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(q29.4)$coef

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 4.367566  | 0.5446031 | 8.019724 | 4.290085e-05 |
| factor(A29)2 | 4.549208  | 0.5446031 | 8.353255 | 3.195335e-05 |
| factor(B29)2 | 2.622644  | 0.6669998 | 3.932000 | 4.344452e-03 |
| factor(B29)3 | 10.562299 | 0.6669998 | 15.835534 | 2.529556e-07 |

Summarize with interaction plots for each candidate transformation. Note use of standardised residuals.

> par(mfrow = c(3, 2), las = 1, mar = c(4, 4, 1, 1))
> plot(fitted(q29.1), rstandard(q29.1))
> interaction.plot(B29, A29, y29)
> plot(fitted(q29.2), rstandard(q29.2))
> interaction.plot(B29, A29, sqrt(y29))
> plot(fitted(q29.3), rstandard(q29.3))
> interaction.plot(B29, A29, log(y29))
Using the diagnostics, the ‘best’ model is the strictly additive model using $\sqrt{y}$ as response variable. On the square-root scale, LSD$t = LSD$q = 1.26 for factor A, while LSD$t = 1.54 and LSD$q = 1.91 for factor B. For factor A, the mean response for level 2 is significantly greater than that for level 1. For factor B, the mean response for level 3 is significantly greater than that for level 2 which, in turn, is significantly greater than that for level 1. Our conclusion looks robust across different possible transformations of $y$.

**Exercise 16**

Dotplots are given for the numbers of bugs attracted to boards of four different colours, and for the overall data set.

<table>
<thead>
<tr>
<th>colour</th>
<th>dots</th>
<th>bugs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>.....</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>. .</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td>overall</td>
<td>. .</td>
<td>.</td>
</tr>
</tbody>
</table>

The within colour and overall (sample) standard deviations are two of the following:

0.143  0.848  2.196  6.784  14.950  26.479  42.020

*Produce the ANOVA table for these data — find the F-value.*

**Answer of Exercise 16**

```r
> sd.overall <- 14.95
> sd.within <- 6.784
> n <- 24
> r <- 4
> (RSS <- sd.within^2 * (n - r))
[1] 920.4531
> (TSS <- sd.overall^2 * (n - 1))
[1] 5140.557
> (MSS <- TSS - RSS)
[1] 4220.104
> (F <- (MSS/(r - 1))/(RSS/(n - r)))
[1] 16
```
Exercise 22

For exercise 19, carry out a test of the hypothesis:

\[ \mu + \alpha_1 = 1.5(\mu + \alpha_2) \quad \text{and} \quad \mu + \alpha_3 = 2(\mu + \alpha_2) \]

simultaneously.

Answer of Exercise 22

```r
> q21 <- data.frame(t.f = factor(c(1, 2, 2, 3, 3, 3)), y = c(22, + 13, 18, 27, 29, 35))
> q21$q22.h0 <- c(1.5, 1, 1, 2, 2, 2)
> q22.1 <- lm(y ~ t.f, data = q21)
> q22.0 <- lm(y ~ q22.h0 - 1, data = q21)
> anova(q22.1, q22.0)
```

Analysis of Variance Table

<table>
<thead>
<tr>
<th>Res.Df</th>
<th>RSS</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>F</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>47.167</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>47.938</td>
<td>-2</td>
<td>-0.772</td>
<td>0.0245</td>
<td>0.976</td>
</tr>
</tbody>
</table>

Do not reject the null hypothesis.

Exercise 23

Find 95% Fisher and 95% Tukey confidence intervals for the differences between the mean mark for each pair of classes for the Maths Classes example (on page 35 of the lecture notes). Ignore the IQ values.

Answer of Exercise 23

```r
> maths <- read.csv("../data/maths.csv")
> TukeyHSD(aov(maths.y ~ factor(class), data=maths))
```

Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = maths.y ~ factor(class), data = maths)

```r
$factor(class)'
   diff  lwr  upr p adj
2-1  6.6 -0.5980652 13.79807 0.0770029
3-1  9.5  2.3019348 16.69807 0.0079455
3-2  2.9 -4.2980652 10.09807 0.5838606
```

> require(gmodels)
> estimable(lm(maths.y ~ factor(class), data=maths),
Exercise 24

Fit an appropriate model to the data below and use it to estimate the value of \( x \) for which \( \mathbb{E}(Y|x) \) is a maximum. Find also the estimated maximum value of \( \mathbb{E}(Y|x) \).

\[
\begin{array}{cc}
 x & y \\
1 & 67.0 \\
2 & 73.1 \\
3 & 78.5 \\
4 & 73.0 \\
5 & 73.3 \\
6 & 73.2 \\
7 & 64.7 \\
8 & 59.0 \\
9 & 46.2 \\
10 & 32.8 \\
\end{array}
\]

Answer of Exercise 24

‘Best’ model is a quadratic in \( x \).

```r
> q24.x <- seq(1:10)
> q24.y <- c(67, 73.1, 78.5, 73, 73.3, 73.2, 64.7, 59, 46.2, 32.8)
> anova(lm(q24.y ~ q24.x + I(q24.x^2) + I(q24.x^3) + I(q24.x^4)))
```

Analysis of Variance Table

<table>
<thead>
<tr>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.391e-05</td>
<td>***</td>
</tr>
</tbody>
</table>

```r
> anova(lm(q24.y ~ q24.x + I(q24.x^2)))
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Analysis of Variance Table

Response: q24.y

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>q24.x</td>
<td>1</td>
<td>1159.59</td>
<td>1159.59</td>
<td>252.75</td>
<td>9.450e-07 ***</td>
</tr>
<tr>
<td>I(q24.x^2)</td>
<td>1</td>
<td>678.19</td>
<td>678.19</td>
<td>147.82</td>
<td>5.823e-06 ***</td>
</tr>
<tr>
<td>Residuals</td>
<td>7</td>
<td>32.12</td>
<td>4.59</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> coef(summary(lm(q24.y ~ q24.x + I(q24.x^2))))

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 59.7666667 | 2.51925337 | 23.723960 | 6.008444e-08 |
| q24.x | 8.717576 | 1.05214641 | 8.285516 | 7.281443e-05 |
| I(q24.x^2) | -1.1333333 | 0.09321622 | -12.158114 | 5.823286e-06 |

> (xmax <- 8.717576/(2 * 1.133333))

[1] 3.845991

> (ymax <- 59.7666667 + 8.717576 * xmax - 1.1333333 * xmax^2)

[1] 76.53052

> q24.xl <- seq(1, 10, by = 0.1)
> q24.yl <- 59.7666667 + 8.717576 * q24.xl - 1.1333333 * q24.xl^2
> par(las = 1)
> plot(q24.x, q24.y)
> lines(q24.xl, q24.yl)

Maximum (is estimated to occur) when \( x = 3.846 \), when expected value of \( Y \) is estimated to be 76.531.

We can obtain estimates of the standard errors of the location of the maximum on \( x \) and the value of the maximum at \( y \) in one of the following ways:

1. delta method (620-371)
2. bootstrap (620-374) or jackknife.

3. reformulate as a non-linear model so that the quantities of interest are parameters, and then use asymptotic standard errors or small-sample asymptotics (How serious do we want to get?).

Exercise 25

To study the effect of advertising, a company runs 50 TV commercials one week, 60 the next, 70 the next and 80 the last week. The number of sales, in hundreds, is recorded for two products, $P_1$ and $P_2$. The data are shown.

<table>
<thead>
<tr>
<th>Advertising</th>
<th>$P_1$</th>
<th>$P_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>29</td>
<td>34</td>
</tr>
<tr>
<td>60</td>
<td>41</td>
<td>43</td>
</tr>
<tr>
<td>70</td>
<td>50</td>
<td>54</td>
</tr>
<tr>
<td>80</td>
<td>61</td>
<td>64</td>
</tr>
</tbody>
</table>

Let $P$ be a factor with two levels, $i = 1, 2$, for product, $A$ be a factor with four levels, $j = 1, 2, 3, 4$, for advertising, and $Adv$ be a numerical variable taking values 50, 60, 70 or 80. Let $Y = (29, 34, 41, 43, 50, 54, 61, 64)'$ in the linear model $Y = X\beta + \epsilon$. Don’t do any computations below, but a graph will help.

1. For the model $y_{i,j} = \alpha + \beta Adv_{i,j} + \epsilon_{i,j}$ write down each of the following:
   A: The X matrix.
   B: The rank of X.
   C: The residual degrees of freedom of the model.
   D: Whether the residual standard deviation $\hat{\sigma}$ is closer to 1.0, 10.0 or 100.0?

2. For the model $y_{i,j} = \alpha + \beta Adv_{i,j} + \gamma Adv_{i,j}^2 + \epsilon_{i,j}$, write down A, B and D.

3. For the model $y_{i,j} = \mu + \epsilon_{i,j}$, write down A, B, C and D.

4. For the model $y_{i,j} = \mu + p_i + a_j + \epsilon_{i,j}$, write down A, B, C and D.

5. For the model $y_{i,j} = p_i + \beta Adv_{i,j} + \epsilon_{i,j}$, write down A, B, C and D.

6. By examining the data, which model above do you expect to be best? Explain briefly.

7. It is suggested that the products might be affected differently but linearly by advertising, but if no advertising is done, no products will be sold. Write down a model and X matrix that describes this.

8. Explain briefly why this experiment is poorly designed.

9. The following output was obtained for the two-way ANOVA model $y_{i,j} = \mu + p_i + a_j + \epsilon_{i,j}$. How can you tell it is wrong, what do you think the error was, and how should it be done correctly?

| Coefficients: | Estimate | Std. Error | t value | Pr(>|t|) |
|---------------|----------|------------|---------|----------|
| (Intercept)   | 16.0000  | 0.9874     | 16.204  | 1.63e-05 *** |
| p             | 3.5000   | 0.5099     | 6.864   | 0.00100 **  |
| a             | 10.3000  | 0.2280     | 45.168  | 1.00e-07 *** |

Residual standard error: 0.7211 on 5 degrees of freedom
Multiple R-Squared: 0.9976
F-statistic: 1044 on 2 and 5 degrees of freedom,
the p-value is 2.792e-07
Answer of Exercise 25

1. \( X = \begin{bmatrix} 1 & 50 \\ 1 & 50 \\ 1 & 60 \\ 1 & 60 \\ 1 & 70 \\ 1 & 70 \\ 1 & 80 \\ 1 & 80 \end{bmatrix} \) \( \text{rank}(X) = 2 \) residual df = 6 \( \hat{\sigma} = 1 \)

\[
\begin{bmatrix} 1 & 50 & 2500 \\ 1 & 50 & 2500 \\ 1 & 60 & 3600 \\ 1 & 60 & 3600 \\ 1 & 70 & 4900 \\ 1 & 70 & 4900 \\ 1 & 80 & 6400 \\ 1 & 80 & 6400 \end{bmatrix}
\]

2. \( X = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \) \( \text{rank}(X) = 1 \) residual df = 7 \( \hat{\sigma} = 10 \)

\[
\begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}
\]

3. \( X = \begin{bmatrix} 50 \\ 0 \\ 1 & 50 \\ 0 & 50 \\ 0 & 60 \\ 1 & 60 \\ 1 & 70 \\ 0 & 70 \\ 1 & 80 \\ 0 & 80 \end{bmatrix} \) \( \text{rank}(X) = 3 \) residual df = 5 \( \hat{\sigma} = 1 \)

6. Model (e); approximately linear in adv with consistent difference between the two products.

7. \( y_{ij} = \beta_i \text{adv}_{ij} + e_{ij}; \quad X = \begin{bmatrix} 50 & 0 \\ 0 & 50 \\ 60 & 0 \\ 0 & 60 \\ 70 & 0 \\ 0 & 70 \\ 80 & 0 \\ 0 & 80 \end{bmatrix} \)
8. No randomisation; there could be carry-over effects.

9. Wrong number of parameters in output (3 instead of 5). Likely error is not specifying \(adv\) (and possibly \(p\), though it doesn’t matter with \(p\)) to be a factor.

Fix by using `lm(y ~ factor(p) + factor(adv))`.

**Exercise 30**

The information below was obtained from a study of oxygen consumption rates of 2 species of limpets, *Acmaea scabra* and *A. digitalis*, at 3 concentrations of sea water. The variable measured was \(\mu l\) O\(_2\)/mg dry body weight/min at 22\(^\circ\)C.

<table>
<thead>
<tr>
<th>Sea water Concentration</th>
<th>Species</th>
<th>Level mean</th>
<th>Level mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>100%</td>
<td><em>A. scabra</em></td>
<td>8.995</td>
<td>10.208</td>
</tr>
<tr>
<td>75%</td>
<td><em>A. digitalis</em></td>
<td>7.614</td>
<td>9.031</td>
</tr>
<tr>
<td>50%</td>
<td></td>
<td>12.250</td>
<td></td>
</tr>
</tbody>
</table>

Using R with conc.f and species.f specified as factors with 3 and 2 levels, respectively, the following deviances were obtained:

<table>
<thead>
<tr>
<th>Model</th>
<th>Deviance</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>623.41</td>
<td>47</td>
</tr>
<tr>
<td>species.f</td>
<td>606.77</td>
<td>46</td>
</tr>
<tr>
<td>species.f + conc.f</td>
<td>425.45</td>
<td>44</td>
</tr>
<tr>
<td>species.f*conc.f</td>
<td>401.52</td>
<td>42</td>
</tr>
</tbody>
</table>

1. Given that there were 8 replicates per combination of species and salinity, complete the analysis as far as you can with the information given. In particular, if possible, determine the significance of various effects and, if relevant, use Tukey’s standardised range method to determine which levels of the factors differ significantly. If it is not possible to complete any part of the analysis, explain what additional information would be needed, and how it could be obtained from R.

2. Assuming that the same information (means, deviances etc) had been obtained, but that the numbers of replicates per combination of species and salinity were as follows, repeat part(a).

<table>
<thead>
<tr>
<th>Species</th>
<th>Salinity</th>
<th><em>A. scabra</em></th>
<th><em>A. digitales</em></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100%</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>75%</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>50%</td>
<td>10</td>
<td>6</td>
</tr>
</tbody>
</table>

**Answer of Exercise 30**

1. \([(425.45 - 401.52)/2]/(401.52/42)\]

[1] 1.251569

Interaction not significant (\(F = 1.25\)).

\(> (623.41 - 606.77)/1)/(606.77/46)\)

[1] 1.261499
Species not sig. (F = 1.27).

\[ \frac{(606.77 - 425.45)}{2} \div \frac{425.45}{44} \]

\[ 9.376049 \]

\[ 3.44 \div \sqrt{2} \times \sqrt{2 \times \frac{425.45}{44} / 16} \]

\[ 2.674215 \]

Conc is significant (F = 9.38) while LSDQ for conc. = 2.67, hence no sig. difference between conc. 100% and 75%, but rate of oxygen consumption with conc = 50% is sig. higher than for the other two concentrations.

2. F-tests for interaction and for concentration would be unchanged from (a), but F-test for species would need the deviance for the model with (just) conc. Further follow-up would need the summary output for the additive model to obtain estimates of the parameters, and the estimable output (for the additive model), in addition to the summary output, to obtain all of the required standard errors for conc. Only one standard error is needed for species, and is provided in the summary output. LSDs could then be obtained ‘in the usual way’.

Exercise 31

Oxygen consumption of crabs. The data below give the oxygen consumption of male and female crabs of three species at three temperatures. Four crabs were used for each combination of sex, species and temperature, a total of 72 crabs.

<table>
<thead>
<tr>
<th>Species 1</th>
<th>Species 2</th>
<th>Species 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Low temp.</strong></td>
<td><strong>Med. temp.</strong></td>
<td><strong>High temp.</strong></td>
</tr>
<tr>
<td>Male Female</td>
<td>Male Female</td>
<td>Male Female</td>
</tr>
<tr>
<td>1.9 1.8</td>
<td>2.3 2.4</td>
<td>2.9 3.0</td>
</tr>
<tr>
<td>1.8 1.7</td>
<td>2.1 2.7</td>
<td>2.8 3.1</td>
</tr>
<tr>
<td>1.6 1.4</td>
<td>2.0 2.4</td>
<td>3.4 3.0</td>
</tr>
<tr>
<td>1.4 1.5</td>
<td>2.6 2.6</td>
<td>3.2 2.7</td>
</tr>
<tr>
<td><strong>Low temp.</strong></td>
<td><strong>Med. temp.</strong></td>
<td><strong>High temp.</strong></td>
</tr>
<tr>
<td>Male Female</td>
<td>Male Female</td>
<td>Male Female</td>
</tr>
<tr>
<td>2.1 2.3</td>
<td>2.4 2.0</td>
<td>3.6 3.1</td>
</tr>
<tr>
<td>2.0 2.0</td>
<td>2.6 2.3</td>
<td>3.1 3.0</td>
</tr>
<tr>
<td>1.8 1.9</td>
<td>2.7 2.1</td>
<td>3.4 2.8</td>
</tr>
<tr>
<td>2.2 1.7</td>
<td>2.3 2.4</td>
<td>3.2 3.2</td>
</tr>
<tr>
<td><strong>Low temp.</strong></td>
<td><strong>Med. temp.</strong></td>
<td><strong>High temp.</strong></td>
</tr>
<tr>
<td>Male Female</td>
<td>Male Female</td>
<td>Male Female</td>
</tr>
<tr>
<td>1.1 1.4</td>
<td>2.0 2.4</td>
<td>2.9 3.2</td>
</tr>
<tr>
<td>1.2 1.0</td>
<td>2.1 2.6</td>
<td>2.8 2.9</td>
</tr>
<tr>
<td>1.0 1.3</td>
<td>1.9 2.3</td>
<td>3.0 2.8</td>
</tr>
<tr>
<td>1.4 1.2</td>
<td>2.2 2.2</td>
<td>3.1 2.9</td>
</tr>
</tbody>
</table>

Carry out a complete analysis of these data, and state your conclusions.

Answer of Exercise 31
> crabs <- data.frame(o2 =
+   c(1.9, 1.8, 1.6, 1.4, 1.8, 1.7, 1.4, 1.5, 2.3, 2.1, 2.0, 2.6, 2.4,
+     2.7, 2.4, 2.6, 2.9, 2.8, 3.4, 3.2, 3.0, 3.1, 3.0, 2.7, 2.1, 2.0,
+     1.8, 2.2, 2.3, 2.0, 1.9, 1.7, 2.4, 2.6, 2.7, 2.3, 2.0, 2.3, 2.1,
+     2.4, 3.6, 3.1, 3.4, 3.2, 3.1, 3.0, 2.8, 3.2, 1.1, 1.2, 1.0, 1.4,
+     1.4, 1.0, 1.3, 1.2, 2.0, 2.1, 1.9, 2.2, 2.4, 2.6, 2.3, 2.2, 2.9,
+     2.8, 3.0, 3.1, 3.2, 2.9, 2.8, 2.9),
+   species = factor(rep(1:3,each=24)),
+   temperature = factor(rep(rep(1:3,each=8),3)),
+   sex = factor(rep(rep(1:2,each=4),9)))
> crabs.1 <- lm(o2~species*temperature*sex,data=crabs)

Examine the residuals.
These look good - no evidence of a problem.

> anova(crabs.1)

Analysis of Variance Table

Response: o2

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>species</td>
<td>2</td>
<td>1.8175</td>
<td>0.9088</td>
<td>24.4751</td>
<td>2.715e-08 ***</td>
</tr>
<tr>
<td>temperature</td>
<td>2</td>
<td>24.6558</td>
<td>12.3279</td>
<td>332.0237</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>sex</td>
<td>1</td>
<td>0.0089</td>
<td>0.0089</td>
<td>0.2394</td>
<td>0.6266</td>
</tr>
<tr>
<td>species:temperature</td>
<td>4</td>
<td>1.1017</td>
<td>0.2754</td>
<td>7.4177</td>
<td>7.752e-05 ***</td>
</tr>
<tr>
<td>species:sex</td>
<td>2</td>
<td>0.3703</td>
<td>0.1851</td>
<td>4.9863</td>
<td>0.0103 *</td>
</tr>
<tr>
<td>temperature:sex</td>
<td>2</td>
<td>0.1753</td>
<td>0.0876</td>
<td>2.3603</td>
<td>0.1041</td>
</tr>
<tr>
<td>species:temperature:sex</td>
<td>4</td>
<td>0.2206</td>
<td>0.0551</td>
<td>1.4850</td>
<td>0.2196</td>
</tr>
<tr>
<td>Residuals</td>
<td>54</td>
<td>2.0050</td>
<td>0.0371</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> # Drop non-significant interaction
> crabs.2 <- lm(o2~(species+sex+temperature)^2,data=crabs)
> anova(crabs.2)

Analysis of Variance Table

Response: o2

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>species</td>
<td>2</td>
<td>1.8175</td>
<td>0.9088</td>
<td>23.6829</td>
<td>3.028e-08 ***</td>
</tr>
<tr>
<td>sex</td>
<td>1</td>
<td>0.0089</td>
<td>0.0089</td>
<td>0.2317</td>
<td>0.6321</td>
</tr>
<tr>
<td>temperature</td>
<td>2</td>
<td>24.6558</td>
<td>12.3279</td>
<td>321.2767</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>species:sex</td>
<td>2</td>
<td>0.3703</td>
<td>0.1851</td>
<td>4.8249</td>
<td>0.0115 *</td>
</tr>
<tr>
<td>species:temperature</td>
<td>4</td>
<td>1.1017</td>
<td>0.2754</td>
<td>7.1776</td>
<td>9.136e-05 ***</td>
</tr>
<tr>
<td>sex:temperature</td>
<td>2</td>
<td>0.1753</td>
<td>0.0876</td>
<td>2.2839</td>
<td>0.1109</td>
</tr>
<tr>
<td>Residuals</td>
<td>58</td>
<td>2.2256</td>
<td>0.0384</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> crabs.3 <- lm(o2~species*temperature+species*sex,data=crabs)
> anova(crabs.3)

Analysis of Variance Table

Response: o2

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>species</td>
<td>2</td>
<td>1.8175</td>
<td>0.9088</td>
<td>22.7109</td>
<td>4.536e-08 ***</td>
</tr>
<tr>
<td>temperature</td>
<td>2</td>
<td>24.6558</td>
<td>12.3279</td>
<td>308.0909</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>sex</td>
<td>1</td>
<td>0.0089</td>
<td>0.0089</td>
<td>0.2221</td>
<td>0.6391196</td>
</tr>
<tr>
<td>species:temperature</td>
<td>4</td>
<td>1.1017</td>
<td>0.2754</td>
<td>6.8830</td>
<td>0.0001253 ***</td>
</tr>
<tr>
<td>species:sex</td>
<td>2</td>
<td>0.3703</td>
<td>0.1851</td>
<td>4.6269</td>
<td>0.0135283 *</td>
</tr>
<tr>
<td>Residuals</td>
<td>60</td>
<td>2.4008</td>
<td>0.0400</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Examine the residuals again.
Now stop and think! We have significant effects for sex, temperature, and species, and interactions between species and sex, and species and temperature. As a brief interlude, it may be worth taking the fish one species at a time, and verifying the pattern.

```r
> anova(lm(o2 ~ temperature * sex, data = crabs, subset = species == + "1"))
```

Analysis of Variance Table

Response: o2

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>temperature</td>
<td>2</td>
<td>7.5833</td>
<td>3.7917</td>
<td>81.0089</td>
<td>9.991e-10 ***</td>
</tr>
<tr>
<td>sex</td>
<td>1</td>
<td>0.0037</td>
<td>0.0037</td>
<td>0.0801</td>
<td>0.7804</td>
</tr>
<tr>
<td>temperature:sex</td>
<td>2</td>
<td>0.1900</td>
<td>0.0950</td>
<td>2.0297</td>
<td>0.1604</td>
</tr>
<tr>
<td>Residuals</td>
<td>18</td>
<td>0.8425</td>
<td>0.0468</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> anova(lm(o2 ~ temperature * sex, data = crabs, subset = species == + "2"))

Analysis of Variance Table

Response: o2

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>temperature</td>
<td>2</td>
<td>5.8233</td>
<td>2.9117</td>
<td>73.8169</td>
</tr>
<tr>
<td>sex</td>
<td>1</td>
<td>0.2817</td>
<td>0.2817</td>
<td>7.1408</td>
</tr>
<tr>
<td>temperature:sex</td>
<td>2</td>
<td>0.0833</td>
<td>0.0417</td>
<td>1.0563</td>
</tr>
<tr>
<td>Residuals</td>
<td>18</td>
<td>0.7100</td>
<td>0.0394</td>
<td></td>
</tr>
</tbody>
</table>

---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> anova(lm(o2 ~ temperature * sex, data = crabs, subset = species == + "3"))

Analysis of Variance Table

Response: o2

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>temperature</td>
<td>2</td>
<td>12.3508</td>
<td>6.1754</td>
<td>245.6519</td>
</tr>
<tr>
<td>sex</td>
<td>1</td>
<td>0.0937</td>
<td>0.0937</td>
<td>3.7293</td>
</tr>
<tr>
<td>temperature:sex</td>
<td>2</td>
<td>0.1225</td>
<td>0.0613</td>
<td>2.4365</td>
</tr>
<tr>
<td>Residuals</td>
<td>18</td>
<td>0.4525</td>
<td>0.0251</td>
<td></td>
</tr>
</tbody>
</table>

---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

We see nothing there to discourage us from our earlier interpretation. We examine the cell means to prepare for multiple comparisons, and we make a few plots to reveal the structure according to the model.

> with(crabs, tapply(o2, list(species, sex), mean))

```
  1  2
1 2.333333 2.358333
2 2.616667 2.400000
3 2.058333 2.183333
```

> with(crabs, tapply(o2, list(temperature, sex), mean))

```
  1  2
1 1.625000 1.600000
2 2.266667 2.366667
3 3.116667 2.975000
```

> with(crabs, tapply(o2, list(temperature, species), mean))

```
  1  2  3
1 1.6375 2.000 1.2000
2 2.3875 2.350 2.2125
3 3.0125 3.175 2.9500
```

> opar <- par(mfrow = c(2, 2), las = 1, mar = c(4, 4, 1, 1))
> with(crabs, interaction.plot(species, temperature, o2))
> with(crabs, interaction.plot(species, sex, o2))
> with(crabs, interaction.plot(temperature, sex, o2))
> par(opar)
Planned & multiple comparisons

- Between sexes (within species, but averaged over temperatures):

```r
> (LSDt <- sqrt((1/12 + 1/12) * 0.04) * qt(0.975, 60))
[1] 0.1633236

> (LSDQ <- qtukey(0.95, 2, 60) * sqrt((1/12 + 1/12) * 0.04)/sqrt(2))
[1] 0.1633236

LSDt = LSDQ = 0.1633

- Species 1 & 3: no sig. difference between males & females by either Fisher or Tukey
Exercise 32

In autumn, small winged fruit called samara fall off maple trees, spinning as they go. A forest scientist studied the relationship between how fast they fell and their ‘disk loading’ (a quantity based on their size and weight). The samara disk loading is related to the aerodynamics of helicopters. The scientist collected a total of 35 samara from 3 trees. A plot of the data is given in the figure with different symbols indicating the different trees (Figure 1). Using these data 6 models were fitted in R. Model specifications and deviances are given below, as is the output produced by disp.e() (which is the same as that produced by ‘summary’) and disp.v() for each of the six models.

Note: ‘tree’ is a vector of tree numbers (1 – 3), ‘tree.f’ is a factor with 3 levels (so tree.f <- factor(tree)), ‘load’ is a vector of disk loads and ‘velocity’ is a vector of velocities.

<table>
<thead>
<tr>
<th>Model</th>
<th>Model Specification</th>
<th>Deviance</th>
</tr>
</thead>
<tbody>
<tr>
<td>lm.1</td>
<td>tree.f + tree.f:load</td>
<td>0.1655</td>
</tr>
<tr>
<td>lm.2</td>
<td>tree.f + load</td>
<td>0.2034</td>
</tr>
<tr>
<td>lm.3</td>
<td>tree + load</td>
<td>0.2059</td>
</tr>
<tr>
<td>lm.4</td>
<td>load</td>
<td>0.2148</td>
</tr>
<tr>
<td>lm.5</td>
<td>tree.f</td>
<td>0.5190</td>
</tr>
<tr>
<td>lm.6</td>
<td>tree</td>
<td>0.5767</td>
</tr>
</tbody>
</table>

Figure 1: Plot of samara velocity against load by tree.
```r
> coef(summary(lm.1))

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 0.5414479| 0.2632359  | 2.057    | 0.047    |
| tree.f2        | -0.8407505| 0.3356459  | -2.505    | 0.012    |
| tree.f3        | -0.2986812| 0.4454446  | -0.671    | 0.503    |
| tree.f1:load   | 3.0628684| 1.1590903  | 2.655    | 0.009    |
| tree.f2:load   | 6.7971294| 0.9516087  | 7.147    | 0.000    |
| tree.f3:load   | 3.8833635| 0.9516087  | 4.053    | 0.000    |

> vcov(lm.1)

<table>
<thead>
<tr>
<th></th>
<th>(Intercept)</th>
<th>tree.f2</th>
<th>tree.f3</th>
<th>tree.f1:load</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>6.929315e-02</td>
<td>-6.929315e-02</td>
<td>-6.929315e-02</td>
<td>-3.042782e-01</td>
</tr>
<tr>
<td>tree.f2</td>
<td>1.126581e-01</td>
<td>6.929315e-02</td>
<td>1.984209e-01</td>
<td>3.042782e-01</td>
</tr>
<tr>
<td>tree.f3</td>
<td>3.042782e-01</td>
<td>3.042782e-01</td>
<td>1.345720e+00</td>
<td>7.263256e-08</td>
</tr>
</tbody>
</table>

> coef(summary(lm.2))

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 0.0756114| 0.1687096  | 0.448    | 0.657    |
| tree.f2        | -0.0104696| 0.0343977  | -0.304   | 0.763    |
| tree.f3        | -0.0587938| 0.0462856  | -1.270   | 0.213    |
| load           | 5.12257416| 0.73875148 | 6.934    | 8.9e-08  |

> vcov(lm.2)

<table>
<thead>
<tr>
<th></th>
<th>(Intercept)</th>
<th>tree.f2</th>
<th>tree.f3</th>
<th>load</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.028462932</td>
<td>-0.0015697917</td>
<td>-0.0059572908</td>
<td>-0.123431306</td>
</tr>
<tr>
<td>tree.f2</td>
<td>0.0011831996</td>
<td>0.0007509489</td>
<td>0.0046554449</td>
<td>0.545753748</td>
</tr>
<tr>
<td>tree.f3</td>
<td>0.021423605</td>
<td>0.023922206</td>
<td>0.545753748</td>
<td></td>
</tr>
<tr>
<td>load</td>
<td>0.123431306</td>
<td>0.0046554449</td>
<td>0.545753748</td>
<td></td>
</tr>
</tbody>
</table>

> coef(summary(lm.3))

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 0.0749115| 0.1751842  | 0.423    | 0.672    |
| tree.f2        | -0.02626565| 0.02234011 | -1.1757 | 0.243    |
| load           | 5.2654560| 0.69350638 | 7.5924777 | 1.93766e-08 |

> vcov(lm.3)

<table>
<thead>
<tr>
<th></th>
<th>(Intercept)</th>
<th>tree</th>
<th>load</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.031868825</td>
<td>-0.0031954757</td>
<td>-0.12133938</td>
</tr>
<tr>
<td>tree</td>
<td>0.003195476</td>
<td>0.0004990807</td>
<td>0.1054084</td>
</tr>
<tr>
<td>load</td>
<td>-0.121339381</td>
<td>0.0105408449</td>
<td>0.4809511</td>
</tr>
</tbody>
</table>

> coef(summary(lm.4))

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | -0.09326016| 0.1074302 | -0.8680998 | 3.916909e-01 |
| load           | 5.82018989 | 0.5111904 | 11.3855613 | 5.703727e-13 |

> vcov(lm.4)

<table>
<thead>
<tr>
<th></th>
<th>(Intercept)</th>
<th>load</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.01154125</td>
<td>-0.05447311</td>
</tr>
<tr>
<td>load</td>
<td>-0.05447311</td>
<td>0.26131565</td>
</tr>
</tbody>
</table>
```
> coef(summary(lm.5))

| Estimate  | Std. Error | t value | Pr(>|t|) |
|-----------|------------|---------|----------|
| (Intercept) | 1.23416667 | 0.03676301 | 33.570879 | 1.617080e-26 |
| tree.f2 | -0.05416667 | 0.05315923 | -1.018951 | 3.158684e-01 |
| tree.f3 | -0.28333333 | 0.05199075 | -5.449687 | 5.368041e-06 |

> vcov(lm.5)

<table>
<thead>
<tr>
<th>(Intercept)</th>
<th>tree.f2</th>
<th>tree.f3</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.001351519</td>
<td>-0.001351519</td>
</tr>
<tr>
<td>tree.f2</td>
<td>-0.001351519</td>
<td>0.002825904</td>
</tr>
<tr>
<td>tree.f3</td>
<td>-0.001351519</td>
<td>0.001351519</td>
</tr>
</tbody>
</table>

> coef(summary(lm.6))

| Estimate  | Std. Error | t value | Pr(>|t|) |
|-----------|------------|---------|----------|
| (Intercept) | 1.40333333 | 0.05841345 | 24.02415 | 1.763690e-22 |
| tree | -0.1416667 | 0.02698516 | -5.24980 | 8.828638e-06 |

> vcov(lm.6)

<table>
<thead>
<tr>
<th>(Intercept)</th>
<th>tree</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.003412131</td>
</tr>
<tr>
<td>tree</td>
<td>-0.0014563973</td>
</tr>
</tbody>
</table>

1. Give the residual degrees of freedom for each of the six models.

2. Two of the six models are totally inappropriate, which are they and why are they inappropriate?

3. For each of the four ‘reasonable’ models, give a rough sketch of the nature of the relationship between velocity and disk load that is implied by the model.

4. Among the four reasonable models, is it valid to compare all of the six possible pairs using F tests? If not, say, with reasons, which models cannot be compared.

5. Using the deviances given in the table above, carry out suitable tests to determine which model seems most appropriate. Give full details of any tests used and write one or two sentences describing your conclusions.

6. Using the model found in (e):
   
   a) give an estimate of σ, the standard deviation of the errors;
   
   b) give an estimate of the velocity of a samara from tree 2 with a disk loading of 0.22;
   
   c) derive a 95% prediction interval for the velocity of a samara from tree 2 with a disk loading of 0.22.

7. For each of the six models in the table above, what would happen to the deviance if the term ‘1’ were subtracted from each of the models; would it increase, decrease or remain unchanged?

**Answer of Exercise 32**

<table>
<thead>
<tr>
<th>model</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>lm.1</td>
<td>29</td>
</tr>
<tr>
<td>lm.2</td>
<td>31</td>
</tr>
<tr>
<td>lm.3</td>
<td>32</td>
</tr>
<tr>
<td>lm.4</td>
<td>33</td>
</tr>
<tr>
<td>lm.5</td>
<td>32</td>
</tr>
<tr>
<td>lm.6</td>
<td>33</td>
</tr>
</tbody>
</table>

1. lm.3 and lm.6. They treat tree as a variable, not as a factor.
model | relationship
--- | ---
lm.1 | three non-parallel straight lines
lm.2 | three parallel straight lines
lm.4 | a single straight line
lm.5 | three straight lines parallel to the x-axis

4. lm.4 and lm.5

5. > anova(lm.1, lm.2, lm.4)

Analysis of Variance Table

Model 1: velocity ~ tree.f * load - load
Model 2: velocity ~ tree.f + load
Model 3: velocity ~ load

<table>
<thead>
<tr>
<th>Res.Df</th>
<th>RSS</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>F</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.165492</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.203441</td>
<td>-2</td>
<td>-0.037949</td>
<td>3.325</td>
<td>0.05011</td>
</tr>
<tr>
<td>3</td>
<td>0.214763</td>
<td>-2</td>
<td>-0.011322</td>
<td>0.992</td>
<td>0.38306</td>
</tr>
</tbody>
</table>

---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> anova(lm.1, lm.4)

Analysis of Variance Table

Model 1: velocity ~ tree.f * load - load
Model 2: velocity ~ load

<table>
<thead>
<tr>
<th>Res.Df</th>
<th>RSS</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>F</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.165492</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.214763</td>
<td>-4</td>
<td>-0.049272</td>
<td>2.1585</td>
<td>0.09885</td>
</tr>
</tbody>
</table>

---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

lm.4 (or lm.1) is 'best'

6. > summary(lm.4)$sigma
   
   [1] 0.08067208

   lm.4: 0.081; lm.1: 0.076

   > predict(lm.4, newdata = data.frame(load = 0.22))

   1
   1.187182

   (b) lm.4: 1.187; lm.1: 1.196

   > predict(lm.4, newdata = data.frame(load = 0.22), interval = "prediction")

   fit  lwr  upr
   1 1.187182 1.020293 1.354071

   (c) lm.4: 1.187 ± 0.167; lm.1: 1.196 ± 0.161

model | change?
--- | ---
lm.1 | no change
lm.2 | no change

7. lm.3 | increase
lm.4 | increase
lm.5 | no change
lm.6 | increase
Record-winning times (in minutes) are available for 35 hill races in Scotland. The distance travelled (in miles) and the height climbed (in feet) in each race is also given. The recorded time for one of the races (Knock Hill) is known to be wrong and the aim here is to predict what it might be from the record-winning times for the other (34) races.

As part of the model fitting process, a case analysis led to observations 7 and 32 being omitted. The figures below were produced as part of the case analysis while the R output was obtained when a multiple regression model was fitted to the remaining 32 observations (34 – 2 outliers).

1. Explain briefly (one or two sentences) the meaning of each of the following in the context of linear models with numerical explanatory variables.
   (a) leverage
   (b) standardised residuals
   (c) Cook’s distance

Which, if any, of these three quantities would change if the distance travelled and the height climbed were both measured in kilometres rather than in miles and feet? [There is no need to justify your answer.]

2. Based on the (two) graphs in Figure , which refer to the 34 observations used for the first model that was fitted, answer the following questions either by giving the number of the required observation or by stating that it is not possible to say with the information provided (there is no need to justify your answers):
   (a) which of the observations 7, 11, 30, 32 or 34 has the largest leverage?
   (b) which of the observations 7 or 11 has the larger absolute value of the standardised residual?
   (c) which of the observations 30, 32 or 34 has the largest absolute value of the (raw) residual?
3. Describe briefly (two or three sentences) what was done and what was probably seen to justify the removal of observation 32. [In the original model the standardised residual for observation 32 was 0.9826.]

4. The Knock Hill race is 3 miles in distance and the height climbed is 350 feet.

   (a) Use the R output below, together with the result that the se.fit for Knock Hill is 1.2424, to find an estimate of and a 95% prediction interval for the record-winning time for this race.

   (b) Give details of how the se.fit of 1.2424 could be obtained from the R output given below. There is no need to complete the calculations.

   (c) The time recorded for the Knock Hill race is 78.65 minutes. What do you think it probably should be?

```r
> hills.lm1 <- lm(time ~ climb + dist, subset = -c(7, 32), data = hills34)
> summary(hills.lm1)
Call:
  lm(formula = time ~ climb + dist, data = hills34, subset = -c(7, 32))
Residuals:
    Min     1Q   Median     3Q    Max
  -10.79964 -2.89707 -0.08491  3.03124  9.48160
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.2303409  1.5525443 -5.301 1.10e-05 ***
climb       0.0065897  0.0008888  7.414 3.60e-08 ***
dist        6.6349395  0.1986398 33.402 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 4.718 on 29 degrees of freedom
Multiple R-squared: 0.9866,  Adjusted R-squared: 0.9857
F-statistic: 1066 on 2 and 29 DF,  p-value: < 2.2e-16
> vcov(hills.lm1)
            (Intercept)   climb      dist
(Intercept)  2.4103937609 -0.400672e-04 -0.1218662203
  climb      -0.0005400672  7.899657e-07 -0.0000977985
  dist      -0.1218662203 -9.977950e-05  0.0394577794
```

Answer of Exercise 33

1. see lecture notes, and none of them.
   (a) 11
   (b) 7
   (c) Can’t say

2. Observation 7 was omitted and the model refitted. Case analysis applied to the refitted model showed observation 32 to be ‘unusual’.

3. (a) > sum(coef(hills.lm1) * c(1, 350, 3))
   [1] 13.98089
   Can use this to check:
Exercise 34

Consider the fitting of a simple linear regression of a response variable \( y \) on a single explanatory variable \( x \).

1. Under what circumstances would you consider transforming the response variable \( y \)?

   Explain how you would determine if those circumstances applied, and explain how you would determine whether your choice of transformation was appropriate.

   (If any plots are used, clearly state what you would plot against what, and what you would expect to see.)

2. Repeat (a) for transforming the explanatory variable \( x \).

3. One of the assumptions of the linear model is that of normality. One way to check for normality is with a normal Q–Q plot. For the simple straight line model \( y_i = \alpha + \beta x_i + e_i \) state, with reasons, what conclusions, if any, could be drawn about the validity of the ‘usual’ assumptions if the normal Q–Q plot showed obvious departure from normality for:

   (a) the \( y_i \) values
   (b) the \( x_i \) values
   (c) the residuals \( r_i = (y_i - \hat{y}_i) \)
   (d) the standardised residuals \( z_i = \frac{r_i}{\hat{\sigma}\sqrt{1-h_{ii}}} \)

   If you had to choose just one of the above four Q–Q plots, which one would you choose, and why?

Answer of Exercise 34

1. \( \text{var}(Y) \) not constant. Plot of (standardised) residuals versus fitted values; look for ‘fanning’. Trial & error until plot of (standardised) residuals versus fitted values looks OK.

2. Plot of \( y \) or (standardised) residuals versus \( x \) is non-linear. Use trial & error to determine most appropriate transformation.

3. (a) No conclusions.
   (b) No conclusions.
   (c) Maybe some indication of a problem, but \( r_i \)’s do not (in general) have the same variance so that interpretation can be difficult.
(d) Indication of a problem.

4. (iv); most reliable assessment of assumptions.

Exercise 36

From *Experimental Design*, 2nd edition, by Cochran and Cox, Wiley, 1957. Several experiments have demonstrated that people find it difficult to select, by personal judgment, unbiased samples from relatively small populations that can be thoroughly inspected before selection. In this experiment, each population consisted of a small area of wheat, containing about 80 shoots, the shoots being slightly over 2 feet high. Six samplers (labelled A – F), all experienced in studying the growth of wheat, inspected each area and measured the heights of 8 shoots whose heights were to give a representative sample of the shoot heights in the area. The quantity recorded in the table below is the difference between the mean height of the 8 selected shoots and the true mean height in the corresponding area. There were 6 areas, each sampled by each of the 6 samplers. The order of sampling was also allowed for with the experiment conducted as a $6 \times 6$ latin square.

<table>
<thead>
<tr>
<th>Order</th>
<th>Areas</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>F</td>
<td>3.5</td>
<td>4.2</td>
<td>6.7</td>
<td>6.6</td>
<td>4.1</td>
<td>3.8</td>
</tr>
<tr>
<td>2</td>
<td>F</td>
<td>8.9</td>
<td>1.9</td>
<td>5.8</td>
<td>4.5</td>
<td>2.4</td>
<td>5.8</td>
</tr>
<tr>
<td>3</td>
<td>C</td>
<td>9.6</td>
<td>3.7</td>
<td>9.7</td>
<td>3.7</td>
<td>6.0</td>
<td>7.0</td>
</tr>
<tr>
<td>4</td>
<td>B</td>
<td>10.5</td>
<td>10.2</td>
<td>4.6</td>
<td>3.7</td>
<td>5.1</td>
<td>3.8</td>
</tr>
<tr>
<td>5</td>
<td>A</td>
<td>3.1</td>
<td>7.2</td>
<td>4.0</td>
<td>3.7</td>
<td>3.5</td>
<td>5.0</td>
</tr>
<tr>
<td>6</td>
<td>E</td>
<td>5.9</td>
<td>7.6</td>
<td>-0.7</td>
<td>3.0</td>
<td>4.0</td>
<td>8.6</td>
</tr>
</tbody>
</table>

1. Carry out an analysis of the data and comment on what you find.

In particular, comment on whether order and/or areas seem to affect the samplers’ errors and, if there is a significant difference between samplers, use Tukey’s multiple comparison method to determine which samplers differ significantly from each other.

2. The true mean shoot heights for each area are given below.

Carry out an analysis to determine whether there is a (statistically significant) relationship between the average of the errors for each of the areas and the true mean shoot height of the area. What do you conclude from the analysis?

<table>
<thead>
<tr>
<th>True mean shoot height</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
</tbody>
</table>

Answer of Exercise 36

1. > LS <- read.csv("../data/LS.csv")
   > LS$area <- factor(LS$area)
   > LS$order <- factor(LS$order)
   > LS$sampler <- factor(LS$sampler)
   > anova(lm(error ~ sampler + order + area, data = LS))
Analysis of Variance Table

Response: error

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>sampler</td>
<td>5</td>
<td>155.596</td>
<td>31.119</td>
<td>9.3502</td>
<td>0.0001027 ***</td>
</tr>
<tr>
<td>order</td>
<td>5</td>
<td>28.599</td>
<td>5.720</td>
<td>1.7186</td>
<td>0.1763454</td>
</tr>
<tr>
<td>area</td>
<td>5</td>
<td>78.869</td>
<td>15.774</td>
<td>4.7395</td>
<td>0.0051140 **</td>
</tr>
<tr>
<td>Residuals</td>
<td>20</td>
<td>66.563</td>
<td>3.328</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 1

Sampler and Areas are both highly significant (P < 0.01).

```r
> with(LS, tapply(error, sampler, mean))
1 2 3 4 5 6
6.066667 5.583333 6.116667 6.916667 2.666667 1.200000
```

```r
> LS.1 <- aov(error ~ sampler + order + area, data = LS)
> TukeyHSD(LS.1)$sampler

<table>
<thead>
<tr>
<th>diff</th>
<th>lwr</th>
<th>upr</th>
<th>p adj</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-1</td>
<td>-0.4833333</td>
<td>-3.794048</td>
<td>0.9970578934</td>
</tr>
<tr>
<td>3-1</td>
<td>0.0500000</td>
<td>-3.260715</td>
<td>0.9999999598</td>
</tr>
<tr>
<td>4-1</td>
<td>0.8500000</td>
<td>-2.460715</td>
<td>0.9628730613</td>
</tr>
<tr>
<td>5-1</td>
<td>-3.4000000</td>
<td>-6.710715</td>
<td>0.0419957269</td>
</tr>
<tr>
<td>6-1</td>
<td>-4.8666667</td>
<td>-8.177382</td>
<td>0.0073426295</td>
</tr>
<tr>
<td>3-2</td>
<td>0.5333333</td>
<td>-2.777382</td>
<td>0.9953326873</td>
</tr>
<tr>
<td>4-2</td>
<td>1.3333333</td>
<td>-1.977382</td>
<td>0.7993406054</td>
</tr>
<tr>
<td>5-2</td>
<td>-2.9166667</td>
<td>-6.227382</td>
<td>0.1046489234</td>
</tr>
<tr>
<td>6-2</td>
<td>-4.3833333</td>
<td>-7.694048</td>
<td>0.0055433643</td>
</tr>
<tr>
<td>4-3</td>
<td>0.8000000</td>
<td>-2.510715</td>
<td>0.9712549187</td>
</tr>
<tr>
<td>5-3</td>
<td>-3.4500000</td>
<td>-6.760715</td>
<td>0.0380506362</td>
</tr>
<tr>
<td>6-3</td>
<td>-4.9166667</td>
<td>-8.227382</td>
<td>0.0017822017</td>
</tr>
<tr>
<td>5-4</td>
<td>-4.2500000</td>
<td>-7.560715</td>
<td>0.0073478336</td>
</tr>
<tr>
<td>6-4</td>
<td>-5.7166667</td>
<td>-9.027382</td>
<td>0.0003270184</td>
</tr>
<tr>
<td>6-5</td>
<td>-1.4666667</td>
<td>-4.777382</td>
<td>0.7308921626</td>
</tr>
</tbody>
</table>
```

Sampler 6 is significantly better than samplers 1 - 4, but not sig. better than sampler 5. Sampler 5 is sig. better than samplers 1, 3 and 4, but not sig. better than sampler 2, and not sig. worse than sampler 6. There are no sig. differences between samplers 1 - 4.

```r
> T.height <- c(59, 66.2, 76.4, 74.5, 76, 72.3)
> A.error <- tapply(LS$error, LS$area, mean)
> coef(summary(lm(A.error ~ T.height)))

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|---------|
| (Intercept) | 19.3431782 | 4.12337427 | 4.691104 | 0.009370058 |
| T.height | -0.2061948 | 0.05806807 | -3.550915 | 0.023779042 |
```

Exercise 11

Discuss the case study on Page 123 (section 7.6) of the lecture notes.
Exercise 15

The following output was obtained when the (simple) regression model \( y_i = \alpha + \beta x_i + e_i \) was fitted to a sample of observations (Figure 2).

\[
> \text{long.lm} <- \text{lm(Employed} \sim \text{Population, data} = \text{longley})
> \text{coef(summary(long.lm))}
\]

\begin{verbatim}
     Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.3806742 4.42243393 1.895037 7.892728e-02
Population  0.4848781  0.03760029 12.895593 3.693245e-09
\end{verbatim}

\[
> \text{anova(long.lm)}
\]

Analysis of Variance Table

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt; F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>1</td>
<td>170.643</td>
<td>170.643</td>
<td>166.30</td>
<td>3.693e-09 ***</td>
</tr>
<tr>
<td>Residuals</td>
<td>14</td>
<td>14.366</td>
<td>1.026</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

\[
> \text{vcov(long.lm)}
\]

\begin{verbatim}
              (Intercept) Population
(Intercept) | 19.5579219 -0.166011959
Population   | -0.1660120  0.001413782
\end{verbatim}

Find (an estimate of) the variance appropriate for finding:

1. a (95%) confidence interval for \( \mathbb{E}(Y) \) when \( x = 120 \);
2. a (95%) prediction interval for an observed value of $Y$ when $x = 120$.

**Answer of Exercise 15**

1. > t(c(1, 120)) %*% vcov(long.lm) %*% c(1, 120)  
   [,1] 0.07351514  
2. > t(c(1, 120)) %*% vcov(long.lm) %*% c(1, 120) + summary(long.lm)$sigma^2  
   [,1] 1.099653

**Exercise 17**

From the 1991 exam.

1. Consider $n$ observations from a linear model $Y = X\beta + \varepsilon$ with $\varepsilon \overset{d}{=} N(0, \sigma^2I)$.

   (a) If $X$ has full rank, say whether each of the following statements is True or False; no explanations are necessary.
   
   i. There is some linear relation among the columns of $X$.
   
   ii. There is more than one solution to the normal equations.
   
   iii. The least squares solution $\hat{\beta}$ satisfies the normal equations.
   
   iv. The least squares solution can be calculated as $(X'X)^{-1}X'y$.
   
   v. Every solution to the normal equations results in the same residuals.
   
   vi. The number of columns of $X$ is the same as the dimension of the estimation space $\mathcal{M}(X)$.

   (b) If $X$ does not have full rank, say whether each of the above statements is True or False; no explanations are necessary.

2. An experiment is conducted using 6 batches of pulp for making paper. The purpose is to compare the effect, on the brightness of the paper, of three pulp additives ($S$, the standard, $N_1$ and $N_2$).

   Three of the six batches are made using the standard additive and $N_1$ together, and the other three batches are made using the standard additive and $N_2$ together.

   (a) The investigator tries to use the results of the experiment to estimate the brightness of paper made using only additive $N_1$ without the standard or $N_2$ additive. Is this possible? Give an appropriate $X$ matrix for the experiment and use it to justify your answer.

   (b) The investigator tries to use the results of the experiment to estimate the difference in brightness of paper made using the standard and $N_1$ together versus the standard and $N_2$ together. Is this possible? Justify your answer.

**Answer of Exercise 17**

1. (a) (F; F; T; T; T; T)  
   (b) (T; T; T; F; T; F)
Exercise 18

Consider two linear models for the same data set:

• Model 1: \( \tilde{y} = X_1 \tilde{\beta}_1 + \tilde{e} \)
• Model 2: \( \tilde{y} = X_2 \tilde{\beta}_2 + \tilde{e} \)

1. What are the most general circumstances under which models 1 and 2 are equivalent (\( X_1 = X_2 \) is only one special case), and which of the following are the same for equivalent models: \( \tilde{\beta}, X \tilde{\beta}, \text{Deviance}(\tilde{\beta}), \sigma \).

2. Under what circumstances can models 1 and 2 be compared using an \( F \)-test?

3. Under what condition on the design matrix \( X_1 \) is \( \tilde{\beta}_1 \) unique?

4. What happens when the condition in (c) is not satisfied?

5. Explain how an estimate of \( \tilde{\beta}_1 \) can be found in practice when the condition in (c) is not satisfied?

6. Under what circumstances is the linear combination of the parameters \( P' \tilde{\beta}_1 \) said to be estimable?

Answer of Exercise 18

1. \( M(X_1) = M(X_2) \) same: \( X \hat{\beta}, \text{Deviance}(\hat{\beta}), \text{and } \hat{\sigma} \).

2. \( M(X_1) \subset M(X_2), \) \( \) [or \( M(X_2) \subset M(X_1) \) ]

3. \( X'_1 X_1 \) is non-singular.

4. There exists infinitely many solutions (\( \tilde{\beta} \)) to the normal equations.

5. Need to impose (\( k - r \)) constraints on the parameters.

6. \( P' \tilde{\beta} \) is unique

Exercise 19

Produce a summary table like that given on page 26 of the notes using the following:

Model \( y_{ij} = \mu + \alpha_i + e_{ij} \)

\[
X = \begin{bmatrix}
1 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 \\
1 & 0 & 0 & 1 \\
1 & 0 & 0 & 1 \\
1 & 0 & 0 & 1
\end{bmatrix}
\]

\[
\begin{bmatrix}
22 \\
13 \\
18 \\
27 \\
29 \\
35
\end{bmatrix}
\]
The required matrices can be produced using either the `lm()` function in R, together with functions that operate on an `lm` object, or from first principles using (some of) the following matrix operations that are available in R:

- Transpose (of a matrix $X$) $t(X)$
- Matrix multiplication (of $A$ and $B$) $A \times B$
- Multiplication (of $A$ by a constant $c$) $c \times A$
- Inverse (of $A$, $A^{-1}$) `solve(A)`
- Matrix addition (of $A$ and $B$) $A + B$

**Answer of Exercise 19**

Model: $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$

$$X = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \sim \begin{bmatrix} 22 \\ 13 \\ 18 \\ 27 \\ 29 \end{bmatrix}, \quad X'X = \begin{bmatrix} 6 & 1 & 2 & 3 \\ 1 & 1 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 3 & 0 & 0 & 3 \end{bmatrix}, \quad X'y = \begin{bmatrix} 144 \\ 31 \end{bmatrix}$$

<table>
<thead>
<tr>
<th>R specification</th>
<th>g.fact contr.sum</th>
<th>g.fact contr.treatment</th>
<th>g.fact-1 contr.sum</th>
<th>g.fact-1 contr.treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>constraint</td>
<td>$\alpha_1 + \alpha_2 + \alpha_3 = 0$</td>
<td>$\alpha_1 = 0$</td>
<td>$\mu = 0$</td>
<td>$\mu = 0$</td>
</tr>
</tbody>
</table>

$$\beta^\sim = \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix}$$

$$X^\sim = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

$$X'^\sim X'^\sim = \begin{bmatrix} 6 & -2 & -1 \\ -2 & 4 & 3 \\ -1 & 3 & 5 \end{bmatrix}, \quad X'^\sim y = \begin{bmatrix} 144 \\ 31 \end{bmatrix}$$

$$\ddot{\beta}^\sim = (X'^\sim X'^\sim)^{-1} X'^\sim y$$

$(\ast) \Rightarrow$ reduced vector/matrix after allowing for constraints; $(full) \Rightarrow$ full vector, including constraints.

For all $(full)$ and all $(\ast)$, $\ddot{y} = X \ddot{\beta}_{(full)} = X^\sim \ddot{\beta}^\sim = \begin{bmatrix} 22.000 \\ 15.500 \\ 15.500 \end{bmatrix}$

**Exercise 20**

41
Using R, and the dataset on the class website, produce (all of) the plots referred to in Chapter 4 of the notes.

**Answer of Exercise 20**
See lecture notes.

**Exercise 21**

For exercise 19, use `vcov` and `estimable` (for $\hat{\beta}$) when using each of the three constraints ($\sum \alpha_i = 0$, $\alpha_1 = 0$, and $\mu = 0$). Also, for “contr.sum” only ($\sum \alpha_i = 0$), verify that the output of `estimable` can be derived from those of `vcov` using the formula

$$sd(Y - Z) = \sqrt{\text{var}(Y) + \text{var}Z - 2\text{cov}(Y,Z)}$$

for arbitrary random variables $Y$ and $Z$.

**Answer of Exercise 21**

```r
> require(gmodels)
> q21 <- data.frame(t.f = factor(c(1,2,2,3,3,3)),
+ y = c(22, 13, 18, 27, 29, 35))
> options(contrasts=c("contr.sum","contr.poly"))
> vcov(q21a <- lm(y ~ t.f, data = q21))

(Intercept)  t.f1   t.f2
(Intercept)  3.2026749 2.038066 -0.5823045
t.f1         2.0380658 8.443416 -4.6584362
t.f2         -0.5823045 -4.658436  5.8230453

> estimable(q21a, rbind('f1-f2'=c(1, -1, 0),
+ 'f1-f3'=c(1, 0, -1),
+ 'f2-f3'=c(0, 1, -1)))

                     Estimate  Std. Error     t value DF  Pr(>|t|)
 f1-f2        23.22222       2.751356   8.440283  3 0.003490414
 f1-f3        29.72222       3.192230   9.310804  3 0.002622782
 f2-f3         6.50000       4.856267   1.338477  3 0.273138626

> options(contrasts=c("contr.treatment","contr.poly"))
> vcov(q21b <- lm(y ~ t.f, data = q21))

(Intercept)  t.f2   t.f3
(Intercept)  15.72222 -15.72222 -15.72222
t.f2        -15.72222  23.58333  15.72222
t.f3        -15.72222  15.72222   20.96296

> estimable(q21b, rbind('f1-f2'=c(1, -1, 0),
+ 'f1-f3'=c(1, 0, -1),
+ 'f2-f3'=c(0, 1, -1)))

                     Estimate  Std. Error     t value DF  Pr(>|t|)
 f1-f2        28.50000       8.411302   3.388298  3 0.04282852
 f1-f3        13.66667       8.254067   1.655749  3 0.19634868
 f2-f3       -14.83333       3.619648  -4.098004  3 0.02628409

> vcov(q21c <- lm(y ~ t.f - 1, data = q21))
```

42
Exercise 26

The previous question (ie Question 25) refers to six models [in parts (1) to (5) and (7)]. For each of the possible 15 pairs of these six models, say whether or not they can be formally compared using an F-test. (Assume that assumptions such as normality are OK.)

**Answer of Exercise 26**

<table>
<thead>
<tr>
<th></th>
<th>(b)</th>
<th>(c)</th>
<th>(d)</th>
<th>(e)</th>
<th>(g)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>(b)</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>(c)</td>
<td></td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>(d)</td>
<td></td>
<td>Y</td>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(e)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>N</td>
</tr>
</tbody>
</table>

Exercise 27

Show that fitting a 1-way ANOVA model for a factor with three levels is equivalent to fitting a quadratic regression model.

**Answer of Exercise 27**

For one-way ANOVA model

\[
X = \begin{bmatrix} y_1 & y_2 & y_3 & y_4 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}
\]
with replicates of each of the three rows, whereas for the quadratic regression model
\[ X = [v_1, v_2, v_3] = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix} \]
again with replicates of each of the three rows.

Now \[ v_1 = u_1 \]
\[ v_2 = u_2 + 2u_3 + 3u_4 \]
\[ v_3 = u_2 + 4u_3 + 9u_4 \]
while \[ y_1 = v_1 \]
\[ y_2 = 3v_1 - 2.5v_2 + 0.5v_3 \]
\[ y_3 = -3v_1 + 4v_2 - v_3 \]
\[ y_4 = v_1 - 1.5v_2 + 0.5v_3 \]

hence the two linear manifolds (estimation spaces) are equivalent.

Exercise 28
For a (numerical) explanatory variable \((x)\) with \(m\) equally-spaced levels (spacing can be taken to be ‘1’ without loss of generality), derive the form of the first two orthogonal polynomials, i.e. polynomials of the form \(x + a\) and \(x^2 + bx + c\) where \(a, b\) and \(c\) are constants whose values (may) depend on \(m\).

Answer of Exercise 28
For \(x\) taking values 1, 2, \ldots, \(m\), ie, \(x_i = i\), \(i = 1, \ldots, m\),
\[ P_0(x_i) = 1 \]
\[ P_1(x_i) = x_i - \bar{x} = x_i - \frac{m+1}{2} \]
\[ P_2(x_i) = (x_i - \bar{x})^2 - \frac{m^2 - 1}{12} = x_i^2 - (m+1)x_i + \frac{(m+1)(m+2)}{6} \]

Exercise 35
A study of air pollution (SO\(_2\) content of air) in 41 US cities considered six environmental variables. The variables considered were:

\(Y\) SO\(_2\) content of air in micrograms per cubic meter.
\(X_1\) Average annual temperature in °F.
\(X_2\) Number of manufacturing enterprises employing 20 or more workers.
\(X_3\) Population size (1970 census); in thousands.
\(X_4\) Average annual wind speed in miles per hour.
\(X_5\) Average annual precipitation in inches.
\(X_6\) Average number of days with precipitation per year.

An initial analysis did not indicate a need for any transformations though a case analysis resulted in three observations being omitted. Model selection was then carried using the remaining 38 observations. The table
below gives the deviances and degrees of freedom obtained when all possible combinations of the 6 predictors were fitted. (The model 25 refers to the regression of $Y$ on $X_2$ and $X_5$, etc.)

1. List the order in which variables would be added using the forward selection procedure, and the order in which variables would be deleted using the backward elimination procedure.

2. Show that if $\text{Fenter} = \text{Fremove} = 4.0$ is used, then forward selection results in a final model with 2 predictor variables whereas backward elimination results in a model with 5 predictor variables.

3. What model would be selected using the $C_p$ criterion?

$$C_p = \frac{\text{deviance(model)}}{\hat{\sigma}^2(\text{full model})} + 2p - n$$

where $p = \#$ parameters, including the constant, and $n = \#$ observations.

<table>
<thead>
<tr>
<th>model</th>
<th>deviance</th>
<th>df</th>
<th>model</th>
<th>deviance</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>const only</td>
<td>16440.8</td>
<td>37</td>
<td>2</td>
<td>13958.2</td>
<td>36</td>
</tr>
<tr>
<td>2</td>
<td>6806.5</td>
<td>36</td>
<td>23</td>
<td>14731.0</td>
<td>35</td>
</tr>
<tr>
<td>3</td>
<td>9667.2</td>
<td>36</td>
<td>24</td>
<td>5803.6</td>
<td>35</td>
</tr>
<tr>
<td>4</td>
<td>16440.8</td>
<td>36</td>
<td>25</td>
<td>7809.4</td>
<td>35</td>
</tr>
<tr>
<td>5</td>
<td>16424.4</td>
<td>36</td>
<td>26</td>
<td>5951.6</td>
<td>35</td>
</tr>
<tr>
<td>6</td>
<td>14731.0</td>
<td>36</td>
<td>27</td>
<td>13761.0</td>
<td>35</td>
</tr>
<tr>
<td>12</td>
<td>5803.6</td>
<td>35</td>
<td>28</td>
<td>4998.0</td>
<td>35</td>
</tr>
<tr>
<td>13</td>
<td>7809.4</td>
<td>35</td>
<td>29</td>
<td>6181.7</td>
<td>35</td>
</tr>
<tr>
<td>14</td>
<td>13761.0</td>
<td>35</td>
<td>30</td>
<td>6806.5</td>
<td>35</td>
</tr>
<tr>
<td>15</td>
<td>12873.1</td>
<td>35</td>
<td>31</td>
<td>5951.6</td>
<td>35</td>
</tr>
<tr>
<td>16</td>
<td>12873.1</td>
<td>35</td>
<td>32</td>
<td>9174.0</td>
<td>35</td>
</tr>
<tr>
<td>23</td>
<td>4998.0</td>
<td>35</td>
<td>33</td>
<td>16407.9</td>
<td>35</td>
</tr>
<tr>
<td>24</td>
<td>6181.7</td>
<td>35</td>
<td>34</td>
<td>14731.0</td>
<td>35</td>
</tr>
<tr>
<td>25</td>
<td>6806.5</td>
<td>35</td>
<td>35</td>
<td>12873.1</td>
<td>35</td>
</tr>
<tr>
<td>26</td>
<td>5951.6</td>
<td>35</td>
<td>36</td>
<td>9650.8</td>
<td>35</td>
</tr>
<tr>
<td>34</td>
<td>9174.0</td>
<td>35</td>
<td>37</td>
<td>16407.9</td>
<td>35</td>
</tr>
<tr>
<td>35</td>
<td>9650.8</td>
<td>35</td>
<td>38</td>
<td>14731.0</td>
<td>35</td>
</tr>
<tr>
<td>36</td>
<td>8319.0</td>
<td>35</td>
<td>39</td>
<td>12873.1</td>
<td>35</td>
</tr>
<tr>
<td>45</td>
<td>14731.0</td>
<td>35</td>
<td>40</td>
<td>12873.1</td>
<td>35</td>
</tr>
<tr>
<td>46</td>
<td>14731.0</td>
<td>35</td>
<td>41</td>
<td>12873.1</td>
<td>35</td>
</tr>
<tr>
<td>56</td>
<td>12873.1</td>
<td>35</td>
<td>42</td>
<td>12873.1</td>
<td>35</td>
</tr>
<tr>
<td>123</td>
<td>4669.2</td>
<td>34</td>
<td>43</td>
<td>12873.1</td>
<td>35</td>
</tr>
<tr>
<td>124</td>
<td>4735.0</td>
<td>34</td>
<td>44</td>
<td>12873.1</td>
<td>35</td>
</tr>
<tr>
<td>125</td>
<td>5063.8</td>
<td>34</td>
<td>45</td>
<td>12873.1</td>
<td>35</td>
</tr>
<tr>
<td>126</td>
<td>5359.7</td>
<td>34</td>
<td>46</td>
<td>12873.1</td>
<td>35</td>
</tr>
<tr>
<td>134</td>
<td>6707.8</td>
<td>34</td>
<td>47</td>
<td>12873.1</td>
<td>35</td>
</tr>
<tr>
<td>135</td>
<td>6888.7</td>
<td>34</td>
<td>48</td>
<td>12873.1</td>
<td>35</td>
</tr>
<tr>
<td>136</td>
<td>7217.5</td>
<td>34</td>
<td>49</td>
<td>12873.1</td>
<td>35</td>
</tr>
<tr>
<td>145</td>
<td>12741.6</td>
<td>34</td>
<td>50</td>
<td>12873.1</td>
<td>35</td>
</tr>
<tr>
<td>146</td>
<td>13103.3</td>
<td>34</td>
<td>51</td>
<td>12873.1</td>
<td>35</td>
</tr>
<tr>
<td>156</td>
<td>12856.7</td>
<td>34</td>
<td>52</td>
<td>12873.1</td>
<td>35</td>
</tr>
</tbody>
</table>

Answer of Exercise 35

1. (a) $2 \rightarrow 23 \rightarrow 234 \rightarrow 1234 \rightarrow 12345 \rightarrow 123456$
Exercise 37

When using questionnaires, the wording of the questions is important. To study this effect, a researcher designs a questionnaire with 20 questions, 10 worded positively (for example, ‘Computers save time’) and 10 worded negatively (for example, ‘Mobile phones waste time’). These 20 questions are not of much interest in their own right, only that they express positive or negative sentiments. The questionnaire is then given to 50 randomly selected subjects. For each question, the subjects give a number between 1 and 10 expressing their agreement (1 = strong disagreement, 10 = strong agreement).

1. For the three factors, $T =$ Type of question (Positive/Negative), $S =$ Subject, and $Q =$ Question:
   
   (a) Give the pattern of nesting and crossing.
   (b) Say which factors you would treat as fixed, and which as random.
   (c) Below is an appropriate ANOVA table for this study, along with Expected Mean Square values.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>E(MS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 $T$</td>
<td></td>
<td>$(6) + (5) + 10(4) + 50(2) + Q[1]$</td>
</tr>
<tr>
<td>2 $Q(T)$</td>
<td></td>
<td>$(6) + (5) + 50(2)$</td>
</tr>
<tr>
<td>3 $S$</td>
<td></td>
<td>$(6) + (5) + 20(3)$</td>
</tr>
<tr>
<td>4 $TS$</td>
<td></td>
<td>$(6) + (5) + 10(4)$</td>
</tr>
<tr>
<td>5 $QS(T)$</td>
<td></td>
<td>$(6) + (5)$</td>
</tr>
<tr>
<td>6 Error</td>
<td></td>
<td>$(6)$</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

   i. Give the degrees of freedom for each term.
   ii. Explain briefly how you would construct the F ratio for testing the $TS$ interaction.
   iii. Explain briefly why you cannot, with the usual method, construct an F ratio for testing the $T$ effect.

2. Suppose the study had been slightly different. Say there are many positive and negative questions to choose from, and each subject was given a different set of 10 positive and 10 negative questions chosen from this set (say there are enough questions so that each question is used only once). For the three factors, $T =$ Type of question (Positive/Negative), $S =$ Subject, and $Q =$ Question:

   (a) Give the pattern of nesting and crossing.
   (b) Say which factors you would treat as fixed, and which as random.
   (c) Give the form (Source and Degrees of freedom columns) of an appropriate ANOVA table.

Answer of Exercise 37

1. Question is nested within Type, Subject is crossed with both Type and Question.

2. Treat Type as fixed, Questions and Subjects as random.

3. (a) 1, 18, 49, 49, 882, 0, 999
(b) Use $F = \frac{TS[MS]}{QS(T)[MS]}$

(c) There is no term with $E(MS) = (6) + (5) + 10(4) + 50(2)$.

4. (a) Question is nested within Type and Subject; Subject is crossed with Type.
   (b) Treat Type as fixed, Questions and Subjects as random.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>1</td>
</tr>
<tr>
<td>S</td>
<td>49</td>
</tr>
<tr>
<td>TS</td>
<td>49</td>
</tr>
<tr>
<td>Q(T,S)</td>
<td>900</td>
</tr>
<tr>
<td>Error</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>999</td>
</tr>
</tbody>
</table>

**Exercise 38**

Give the sources of variation and degrees of freedom for each of the following situations involving 3 factors, $A$, $B$ and $C$, with $a$, $b$ and $c$ levels, respectively, and $n$ observations for each of the $a \times b \times c$ combinations of $A$, $B$, and $C$.

1. All factors crossed.
2. $A$ crossed with $B$, $C$ nested within (each of the $ab$ combinations of) $A$ and $B$.
3. $B$ nested within $A$, $C$ crossed with both $A$ and $B$.
4. $B$ and $C$ nested within $A$, and $B$ crossed with $C$ within $A$.
5. $B$ nested within $A$, $C$ nested within $B$.

**Answer of Exercise 38**

1. | Source       | df   |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>a–1</td>
</tr>
<tr>
<td>B</td>
<td>b–1</td>
</tr>
<tr>
<td>C</td>
<td>c–1</td>
</tr>
<tr>
<td>AB</td>
<td>(a–1)(b–1)</td>
</tr>
<tr>
<td>AC</td>
<td>(a–1)(c–1)</td>
</tr>
<tr>
<td>BC</td>
<td>(b–1)(c–1)</td>
</tr>
<tr>
<td>ABC</td>
<td>(a–1)(b–1)(c–1)</td>
</tr>
<tr>
<td>residual</td>
<td>abc(n–1)</td>
</tr>
<tr>
<td>total</td>
<td>abcn–1</td>
</tr>
</tbody>
</table>

2. | Source                  | df   |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>a–1</td>
</tr>
<tr>
<td>B</td>
<td>b–1</td>
</tr>
<tr>
<td>AB</td>
<td>(a–1)(b–1)</td>
</tr>
<tr>
<td>C%in%(A*B)</td>
<td>ab(c–1)</td>
</tr>
<tr>
<td>residual</td>
<td>abc(n–1)</td>
</tr>
<tr>
<td>total</td>
<td>abcn–1</td>
</tr>
</tbody>
</table>
Tutorial question 38 — in R

```r
> y <- rnorm(120)
> A <- factor(rep(1:2,each=60))
> B <- factor(rep(rep(1:3,each=20),2))
> C <- factor(rep(rep(1:4,each=5),6))
> B6 <- factor(rep(1:6,each=20))
> C8 <- factor(c(rep(rep(1:4,each=5),3),rep(rep(5:8,each=5),3)))
> C24 <- factor(rep(1:24,each=5))
> ## (1)
> anova(lm(y~A*B*C))
```

Analysis of Variance Table

```
Response: y

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>a-1</td>
<td>0.020</td>
<td>0.020</td>
<td>0.0186</td>
<td>0.89193</td>
</tr>
<tr>
<td>B</td>
<td>a(b-1)</td>
<td>5.343</td>
<td>2.671</td>
<td>2.4866</td>
<td>0.08854 .</td>
</tr>
<tr>
<td>C</td>
<td>c-1</td>
<td>1.889</td>
<td>0.630</td>
<td>0.5861</td>
<td>0.62557</td>
</tr>
<tr>
<td>A:B</td>
<td>a-1</td>
<td>2.425</td>
<td>1.212</td>
<td>1.1287</td>
<td>0.32772</td>
</tr>
<tr>
<td>A:C</td>
<td>a-1</td>
<td>4.079</td>
<td>1.360</td>
<td>1.2655</td>
<td>0.29062</td>
</tr>
<tr>
<td>B:C</td>
<td>a-1</td>
<td>3.931</td>
<td>0.655</td>
<td>0.6099</td>
<td>0.72188</td>
</tr>
<tr>
<td>A:B:C</td>
<td>a-1</td>
<td>6.772</td>
<td>1.129</td>
<td>1.0507</td>
<td>0.39776</td>
</tr>
<tr>
<td>Residuals</td>
<td>abc(n-1)</td>
<td>103.129</td>
<td>1.074</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
> ## (2)
> anova(lm(y~A*B+C%in%(A*B)))
```

Analysis of Variance Table
```
Response: y
    Df  Sum Sq Mean Sq  F value  Pr(>F)
A     1  0.020  0.020  0.0186 0.89193
B     2  5.343  2.671  2.4866 0.08854 .
A:B   2  2.425  1.212  1.1287 0.32772
A:B:C 18 16.671  0.926  0.8621 0.62398
Residuals 96 103.129 1.074
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> anova(lm(y~A*B+C24%in%(A:B)))
Analysis of Variance Table

Response: y
    Df  Sum Sq Mean Sq  F value  Pr(>F)
A     1  0.020  0.020  0.0186 0.89193
B     2  5.343  2.671  2.4866 0.08854 .
A:B   2  2.425  1.212  1.1287 0.32772
A:B:C 18 16.671  0.926  0.8621 0.62398
Residuals 96 103.129 1.074
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> anova(lm(y~A*C+C*(B%in%A)))
Analysis of Variance Table

Response: y
    Df  Sum Sq Mean Sq  F value  Pr(>F)
A     1  0.020  0.020  0.0186 0.8919
C     3  1.889  0.630  0.5861 0.6256
A:C   3  4.079  1.360  1.2655 0.2906
A:B   4  7.767  1.942  1.8076 0.1336
A:C:B 12 10.703  0.892  0.8303 0.6192
Residuals 96 103.129 1.074
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> ## (3)
> anova(lm(y~A*B+C24%in%(A:B)))
Analysis of Variance Table

Response: y
    Df  Sum Sq Mean Sq  F value  Pr(>F)
A     1  0.020  0.020  0.0186 0.8919
C     3  1.889  0.630  0.5861 0.6256
A:C   3  4.079  1.360  1.2655 0.2906
A:B   4  7.767  1.942  1.8076 0.1336
A:C:B 12 10.703  0.892  0.8303 0.6192
Residuals 96 103.129 1.074
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> ## (3)
> anova(lm(y~A*B+C24%in%(A:B)))
Analysis of Variance Table

Response: y
    Df  Sum Sq Mean Sq  F value  Pr(>F)
A     1  0.020  0.020  0.0186 0.8919
C     3  1.889  0.630  0.5861 0.6256
A:C   3  4.079  1.360  1.2655 0.2906
A:B   4  7.767  1.942  1.8076 0.1336
A:C:B 12 10.703  0.892  0.8303 0.6192
Residuals 96 103.129 1.074
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> anova(lm(y~A*C+B6*C))
Analysis of Variance Table
```
<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>0.020</td>
<td>0.020</td>
<td>0.0186</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>1.889</td>
<td>0.630</td>
<td>0.5861</td>
</tr>
<tr>
<td>B6</td>
<td>4</td>
<td>7.767</td>
<td>1.942</td>
<td>1.8076</td>
</tr>
<tr>
<td>A:C</td>
<td>3</td>
<td>4.079</td>
<td>1.360</td>
<td>1.2655</td>
</tr>
<tr>
<td>C:B6</td>
<td>12</td>
<td>10.703</td>
<td>0.892</td>
<td>0.8303</td>
</tr>
</tbody>
</table>

Residuals 96 103.129 1.074

> `anova(lm(y~A+B6%in%A+C*A+C*B6))`

Analysis of Variance Table

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>0.020</td>
<td>0.020</td>
<td>0.0186</td>
</tr>
<tr>
<td>A:B</td>
<td>4</td>
<td>7.767</td>
<td>1.942</td>
<td>1.8076</td>
</tr>
<tr>
<td>A:C</td>
<td>6</td>
<td>5.967</td>
<td>0.995</td>
<td>0.9258</td>
</tr>
<tr>
<td>A:B:C</td>
<td>12</td>
<td>10.703</td>
<td>0.892</td>
<td>0.8303</td>
</tr>
</tbody>
</table>

Residuals 96 103.129 1.074

> `anova(lm(y~A+(B*C)%in%A))`

Analysis of Variance Table

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>0.020</td>
<td>0.020</td>
<td>0.0186</td>
</tr>
<tr>
<td>B6</td>
<td>4</td>
<td>7.767</td>
<td>1.942</td>
<td>1.8076</td>
</tr>
<tr>
<td>C8</td>
<td>6</td>
<td>5.967</td>
<td>0.995</td>
<td>0.9258</td>
</tr>
<tr>
<td>B6:C8</td>
<td>12</td>
<td>10.703</td>
<td>0.892</td>
<td>0.8303</td>
</tr>
</tbody>
</table>

Residuals 96 103.129 1.074

> `anova(lm(y~A+(B6*C8)%in%A))`

Analysis of Variance Table

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>0.020</td>
<td>0.020</td>
<td>0.0186</td>
</tr>
<tr>
<td>B6</td>
<td>4</td>
<td>7.767</td>
<td>1.942</td>
<td>1.8076</td>
</tr>
<tr>
<td>C8</td>
<td>6</td>
<td>5.967</td>
<td>0.995</td>
<td>0.9258</td>
</tr>
<tr>
<td>B6:C8</td>
<td>12</td>
<td>10.703</td>
<td>0.892</td>
<td>0.8303</td>
</tr>
</tbody>
</table>

Residuals 96 103.129 1.074
Exercise 39

The effect of four levels of nitrogen (0, 0.2, 0.4 and 0.6 cwt of manure per acre) on the yield of three varieties of oats was studied using a total of 72 plots in six blocks.

<table>
<thead>
<tr>
<th>Block</th>
<th>Variety</th>
<th>Nitrogen level</th>
<th>Block</th>
<th>Variety</th>
<th>Nitrogen level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>V1</td>
<td>0  0.2  0.4 0.6</td>
<td></td>
<td>V1</td>
<td>74 89 81 122</td>
</tr>
<tr>
<td>I</td>
<td>V2</td>
<td>111 130 157 174</td>
<td>IV</td>
<td>V2</td>
<td>64 103 132 133</td>
</tr>
<tr>
<td></td>
<td>V3</td>
<td>117 114 161 141</td>
<td></td>
<td>V3</td>
<td>70 89 104 117</td>
</tr>
<tr>
<td></td>
<td>V1</td>
<td>105 140 118 156</td>
<td></td>
<td>V1</td>
<td>62 90 100 116</td>
</tr>
<tr>
<td>II</td>
<td>V2</td>
<td>61 91 97 100</td>
<td></td>
<td>V2</td>
<td>80 82 94 126</td>
</tr>
<tr>
<td></td>
<td>V3</td>
<td>70 108 126 149</td>
<td></td>
<td>V3</td>
<td>63 70 109 99</td>
</tr>
<tr>
<td>III</td>
<td>V1</td>
<td>96 124 121 144</td>
<td></td>
<td>V1</td>
<td>53 74 118 113</td>
</tr>
<tr>
<td></td>
<td>V2</td>
<td>68 64 112 86</td>
<td></td>
<td>V2</td>
<td>89 82 86 104</td>
</tr>
<tr>
<td></td>
<td>V3</td>
<td>60 102 89 96</td>
<td></td>
<td>V3</td>
<td>97 99 119 121</td>
</tr>
</tbody>
</table>

Using B, V and N to denote blocks, varieties and nitrogen levels, respectively, the following (incomplete and possibly inappropriate) ANOVA was obtained.
The study was conducted as a split-plot design where the rows within each of the six blocks in the above table correspond to ‘whole plots’ (three per block), each of which was divided into four subplots for the application of the manure.

1. Describe, briefly, how randomisation should have been implemented here.

2. Complete the analysis to the extent of determining which effects are significant. (Follow-up analysis using multiple comparisons is desirable but is not required.)

3. Obtain estimates of the subplot and mainplot error variances.

**Answer of Exercise 39**

1. Within each of the six blocks, allocate one mainplot to each of the three varieties, at random, using a separate randomisation for each block.

Within each mainplot, allocate one of the four subplots to each of the four levels of nitrogen, at random, using a separate randomisation for each mainplot.

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>5</td>
<td>15875.3</td>
<td>3175.1</td>
<td>5.280</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>V</td>
<td>2</td>
<td>1786.4</td>
<td>893.2</td>
<td>1.485</td>
<td>ns</td>
</tr>
<tr>
<td>BV</td>
<td>10</td>
<td>6013.3</td>
<td>601.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>3</td>
<td>20020.5</td>
<td>6673.5</td>
<td>41.042</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>VN</td>
<td>6</td>
<td>321.7</td>
<td>53.6</td>
<td>0.303</td>
<td>ns</td>
</tr>
<tr>
<td>residual</td>
<td>45</td>
<td>7968.8</td>
<td>177.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>71</td>
<td>51985.9</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

BV is the mainplot (or wholeplot) error term, while the subplot error term (residual) consists of BN + BVN.

Since VN is not significant it is OK to remove it from the model. The revised anova then looks as follows:

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>5</td>
<td>15875.3</td>
<td>3175.1</td>
<td>5.280</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>V</td>
<td>2</td>
<td>1786.4</td>
<td>893.2</td>
<td>1.485</td>
<td>ns</td>
</tr>
<tr>
<td>BV</td>
<td>10</td>
<td>6013.3</td>
<td>601.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>3</td>
<td>20020.5</td>
<td>6673.5</td>
<td>41.042</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>residual</td>
<td>51</td>
<td>8290.5</td>
<td>162.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>71</td>
<td>51985.9</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

BV is the mainplot (or wholeplot) error term, while the subplot error term (residual) consists of BN + BVN.

Since VN is not significant it is OK to remove it from the model. The revised anova then looks as follows:

3. $\hat{\sigma}_e^2 = 162.6$ (or 177.1 if VN is retained in the model)

$\hat{\sigma}_m^2 = \frac{601.3 - 162.6}{4} = 107.7$ (or 106.1 if VN is retained in the model)
Exercise 40

Consider a $2^4$ experiment, with factors $A$, $B$, $C$ and $D$. Main interest is in the main effects of the four factors and the two-factor interactions between $A$ and the other factors (ie $AB$, $AC$, and $AD$); the experimenter is willing to assume that all other interactions are negligible. Give a design (ie the allocation of treatments to blocks), the form of the ANOVA table (sources of variability and df) and describe, briefly, how the (four) main effects and (three) interactions can be tested, if at all, for each of the following situations:

i. 2 blocks of 8 plots are available;

ii. 4 blocks of 4 plots are available.

Answer of Exercise 40

i. Two blocks of 8 plots = 16 observations ⇒ one replicate.
Need to confound one effect with blocks; choose ABCD.

<table>
<thead>
<tr>
<th>(1)</th>
<th>ab</th>
<th>ac</th>
<th>ad</th>
<th>bc</th>
<th>bd</th>
<th>cd</th>
<th>abcd</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>b</td>
<td>c</td>
<td>d</td>
<td>abc</td>
<td>abd</td>
<td>acd</td>
<td>bcd</td>
</tr>
</tbody>
</table>

ANOVA
Assuming that all interactions apart from $AB$, $AC$, and $AD$ are negligible,

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blocks</td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
</tr>
<tr>
<td>AB</td>
<td>1</td>
</tr>
<tr>
<td>AC</td>
<td>1</td>
</tr>
<tr>
<td>AD</td>
<td>1</td>
</tr>
<tr>
<td>residual</td>
<td>7</td>
</tr>
<tr>
<td>total</td>
<td>15</td>
</tr>
</tbody>
</table>

Main effects are two-factor interactions involving $A$ can then be tested in ‘the usual way’, using an $F$-test with 1 & 7 df

ii. Four blocks of 4 plots = 16 observations ⇒ one replicate.
Need to confound three effect with blocks, two effects plus their ‘interaction’; choose $ABC$, $ABD$ and (hence) $CD$ ($ABC \times ABD = A^2B^2CD = CD$).

<table>
<thead>
<tr>
<th>(1)</th>
<th>ab</th>
<th>acd</th>
<th>bcd</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>b</td>
<td>cd</td>
<td>abcd</td>
</tr>
<tr>
<td>c</td>
<td>abc</td>
<td>ad</td>
<td>bd</td>
</tr>
<tr>
<td>d</td>
<td>abd</td>
<td>ac</td>
<td>bc</td>
</tr>
</tbody>
</table>

ANOVA
Assuming that all interactions apart from $AB$, $AC$, and $AD$ are negligible,
Main effects are two-factor interactions involving $A$ can then be tested in ‘the usual way’, using an $F$–test with 1 & 5 df.

Alternatively, significant effects can be identified using an ‘effects plot’, in which case the ANOVA would retain the terms $BC$, $BD$, $ACD$, $BCD$ and $ABCD$, each with one df, with zero df for residual. Effects whose points do not lie close to a line through the origin will be deemed to be significant.

Given that there are only 5 df for residual, the effects plot is probably more appropriate.