Exercise 1

Consider a two-way classification with row and column factors denoted by \(A.f\) and \(B.f\) respectively and with more than one observation per cell. Using just one set of contrasts (contr.treatment, say), determine which of the following model formulae would produce:

(i) the same deviance (residual sum of squares);

(ii) the same estimates, for all of the parameters.

(a) \(A.f \times B.f\)  
(b) \(A.f : B.f\)  
(c) \(A.f + B.f + A.f : B.f\)  
(d) \(A.f : B.f - 1\)  
(e) \(A.f + A.f : B.f\)  
(f) \(B.f + A.f : B.f\)  
(g) \(A.f \times B.f - A.f : B.f\)  
(h) \(A.f + B.f\)  
(i) \(A.f + B.f - 1\)  
(j) \(B.f + A.f - 1\)

You can (should) check your answers by generating an appropriate data set (see below) and then using the model specifications listed above.

The following commands generate data and factors for a \(3 \times 4\) table with 2 observations per cell:

```r
> y <- rnorm(24, 0, 1)
> A.f <- factor(rep(rep(1:3, each = 2), 4))
> B.f <- factor(rep(1:4, each = 6))
```

Answer of Exercise 1

(i) Models with the same deviance:

• (a) & (b) & (c) & (d) & (f)  [all these include the interaction \(A.f:B.f\)]
• (g) & (h) & (i) & (j)  [all these include \(A.f + B.f\)]

(ii) Models with exactly the same parameter estimates:

• (a) & (c)
• (g) & (h)

Exercise 2

The data for this question were taken from a paper by Carter, Collier and Davis (1951), *Agronomy Journal*, and refer to an experiment to determine the effectiveness of blast furnace slags as agricultural liming materials with three different types of soils. The soil types were I: Sandy loam, II: Sandy clay loam, III: Loamy sand. The soil treatments were None, Limestone, Coarse slag, Medium slag and Agricultural slag, all applied at 4000 lbs per acre.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Soil type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
</tr>
<tr>
<td>1. None</td>
<td>11.1</td>
</tr>
<tr>
<td>2. Limestone</td>
<td>25.6</td>
</tr>
<tr>
<td>3. Coarse slag</td>
<td>15.3</td>
</tr>
<tr>
<td>4. Medium slag</td>
<td>22.7</td>
</tr>
<tr>
<td>5. Ag. slag</td>
<td>23.8</td>
</tr>
</tbody>
</table>
The values in the table are bushels of shelled corn per acre.

1. (a) Carry out an analysis of these data, including the use of an appropriate graph, an ANOVA table from R, and Fisher and Tukey tests and/or intervals, and summarise your findings. Do not transform the data, although you may choose to discuss the usefulness of a transformation for these data.

(b) Give an estimate of, and 95% confidence interval for, the difference between the average of the three “slag” treatments and “None”, i.e. the difference between the average of treatments 3 to 5, and treatment 1.

2. One of the 15 observations is very different from the others. Which one is it? Repeat both parts of (1) with this observation omitted.

**Answer of Exercise 2**

1. Read in the data

   ```r
   > corn <- data.frame(soil = factor(rep(c("I","II","III"), 5)),
   + treatment = factor(rep(c("None","Limestone","Coarse slag",
   + "Medium slag","Aggregate slag"), each = 3)),
   + bushels = c(11.1, 32.6, 63.3, 25.6, 63.1, 44.1, 15.3,
   + 40.8, 65.0, 22.7, 52.1, 58.8, 23.8, 52.8, 61.4))
   ```

   (a) Plot: An interaction plot is appropriate here.

   ```r
   > par(las = 1)
   > with(corn, interaction.plot(treatment, soil, bushels))
   ```
The plot strongly suggests that there may be a significant soil effect. There may also be a treatment effect, but it is likely to be much smaller. The plot also suggests that there may be interaction between treatment and soil. However, since there is only one observation per cell it is not possible to test for interaction and we need to ASSUME that treatment and soil do not interact in order to proceed with the analysis.

Since the design is balanced, equal number of observations (namely 1) per cell, we don’t have to worry about the order in which terms are fitted in order to test for the effects of treatment and soil.

First, fit the model and check the diagnostics.

```r
> corn.lm <- lm(bushels ~ soil + treatment, data=corn)
> opar <- par(mfrow=c(2,2), las=1, mar=c(4,4,3,2))
> plot(corn.lm)
> par(opar)
```

Ok ... maybe a log transform would be helpful. But we won’t do it. Let’s look at the `anova()` statement.

```r
> anova(corn.lm)
```

**Analysis of Variance Table**

**Response:** bushels
Soil is highly significant ($P < 0.001$), but treatment is nowhere near being significant ($P = 0.712$).

**But which levels of soil differ significantly?**

The analyst *may* elect to refit the model, dropping treatment. We do not do so here.

```r
> with(corn, tapply(bushels, soil, mean))
I   II  III
19.70 48.28 58.52
> vcov(corn.lm)[1:3, 1:3]
     (Intercept) soilII soilIII
(Intercept)  45.55654 -19.52423 -19.52423
soilII     -19.52423  39.04847  19.52423
soilIII    -19.52423  19.52423  39.04847
```

The standard errors of the first two differences are 6.24888, the third is (keep in mind how it is computed!)

\[
\text{se}_{\hat{\beta}_3 - \hat{\beta}_2} = \sqrt{\hat{\sigma}^2_{\hat{\beta}_2 - \hat{\beta}_1} + \hat{\sigma}^2_{\hat{\beta}_3 - \hat{\beta}_1} - 2\text{Cov}(\hat{\beta}_2 - \hat{\beta}_1, \hat{\beta}_3 - \hat{\beta}_1)}
\]  

which is 6.24888.

... or ...

```r
> require(gmodels)
> (soil.1 <- estimable(corn.lm,
+   rbind("2 - 1" = c(0, 1, 0, 0, 0, 0, 0),
+   "3 - 1" = c(0, 0, 1, 0, 0, 0, 0),
+   "2 - 3" = c(0, 1, -1, 0, 0, 0, 0))))

       Estimate Std. Error   t value DF Pr(>|t|)
2 - 1     28.58   6.248877  4.573622  8 0.001817
3 - 1     38.82   6.248877  6.212316  8 0.000256
2 - 3    -10.24   6.248877 -1.638694  8 0.139910
```

then the LSDs are:

```r
> qt(0.975, df = 8) * soil.1[, 2]
> qtukey(0.95, nmeans = 3, df = 8) * soil.1[, 2]
```
... or ...

\[
\text{LSD}_t = t_{0.975}^{0.975} \times \sqrt{97.621 \left( \frac{1}{5} + \frac{1}{5} \right)} = 14.41
\]

\[
\text{LSD}_Q = \frac{Q(3.8, 0.95)}{\sqrt{2}} \times \sqrt{97.621 \left( \frac{1}{5} + \frac{1}{5} \right)} = 17.86
\]

By LSD_t and LSD_Q, no sig. difference between soils 2 and 3, but yield for soil 1 is significantly less that for the other two soils (reasonably confident that yield for soil 1 is less than that for other two soils by at least 10 units).

(b) Estimate and CI can be obtained from first principles: average of the 9 observations from treatments 3, 4 and 5, minus the average of the three observations from treatment 1.

\[
\text{Estimate} = 43.63 - 35.67 = 7.97,
\]

\[
\text{se(estimate)} = \sqrt{\hat{\sigma}^2 \left( \frac{1}{9} + \frac{1}{3} \right)} = \sqrt{97.621 \times \frac{4}{9}} = 6.59.
\]

\[
95\% \text{ CI} = 7.97 \pm t_{0.975}^{975} \times 6.59 = 7.97 \pm 15.19 = (-7.22, 23.16)
\]

2. Examine the residuals and the observations.

> corn$residuals <- residuals(corn.lm)
> corn$st.residuals <- rstandard(corn.lm)
> corn[order(corn$residuals, decreasing = TRUE),]

<table>
<thead>
<tr>
<th>soil</th>
<th>treatment</th>
<th>bushels</th>
<th>residuals</th>
<th>st.residuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>II Limestone</td>
<td>63.1</td>
<td>12.7200000</td>
<td>1.76285159</td>
</tr>
<tr>
<td>3</td>
<td>III None</td>
<td>63.3</td>
<td>11.2800000</td>
<td>1.56328348</td>
</tr>
<tr>
<td>9</td>
<td>III Coarse slag</td>
<td>65.0</td>
<td>8.2800000</td>
<td>1.14751660</td>
</tr>
<tr>
<td>4</td>
<td>I Limestone</td>
<td>25.6</td>
<td>3.8000000</td>
<td>0.52663805</td>
</tr>
<tr>
<td>11</td>
<td>II Medium slag</td>
<td>52.1</td>
<td>1.4533333</td>
<td>0.20141596</td>
</tr>
<tr>
<td>14</td>
<td>II Aggregate slag</td>
<td>52.8</td>
<td>0.6866667</td>
<td>0.09516442</td>
</tr>
<tr>
<td>10</td>
<td>I Medium slag</td>
<td>22.7</td>
<td>0.6333333</td>
<td>0.08777301</td>
</tr>
<tr>
<td>13</td>
<td>I Aggregate slag</td>
<td>23.8</td>
<td>0.2666667</td>
<td>0.03695706</td>
</tr>
<tr>
<td>15</td>
<td>III Aggregate slag</td>
<td>61.4</td>
<td>-0.9533333</td>
<td>-0.13212148</td>
</tr>
<tr>
<td>12</td>
<td>III Medium slag</td>
<td>58.8</td>
<td>-2.0866667</td>
<td>-0.28918897</td>
</tr>
<tr>
<td>1</td>
<td>I None</td>
<td>11.1</td>
<td>-2.1000000</td>
<td>-0.29103682</td>
</tr>
<tr>
<td>7</td>
<td>I Coarse slag</td>
<td>15.3</td>
<td>-2.6000000</td>
<td>-0.36033130</td>
</tr>
<tr>
<td>8</td>
<td>II Coarse slag</td>
<td>40.8</td>
<td>-5.6800000</td>
<td>-0.78718530</td>
</tr>
<tr>
<td>2</td>
<td>II None</td>
<td>32.6</td>
<td>-9.1800000</td>
<td>-1.27224666</td>
</tr>
<tr>
<td>6</td>
<td>III Limestone</td>
<td>44.1</td>
<td>-16.5200000</td>
<td>-2.28948964</td>
</tr>
</tbody>
</table>

The largest residual is –16.52 (observation 6).

(a) With this one observation removed the design is no longer balanced so that the order in which terms are added becomes important.

> anova(corn.lm.1a <- lm(bushels ~ soil + treatment, data = corn, + subset = -6))

Analysis of Variance Table

Response: bushels

<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>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
soil  2 4303.7 2151.8 55.9415 4.954e-05 ***
treatment 4 462.4 115.6 3.0051 0.09713 .
Residuals 7 269.3 38.5
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> anova(corn.lm.1b <- lm(bushels ~ treatment + soil, data = corn, + subset = -6))

Analysis of Variance Table

Response: bushels

<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>treatment</td>
<td>4</td>
<td>206.6</td>
<td>51.7</td>
<td>1.3429</td>
</tr>
<tr>
<td>soil</td>
<td>2</td>
<td>4559.4</td>
<td>2279.7</td>
<td>59.2658</td>
</tr>
<tr>
<td>Residuals</td>
<td>7</td>
<td>269.3</td>
<td>38.5</td>
<td></td>
</tr>
</tbody>
</table>
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Test for soils (allowing for treatments): \( P = 0.00004 \), reject \( H_0 \).
Test for treatments (allowing for soils): \( P = 0.097 \), the test is not quite significant (at the 5% level), but is much closer to being significant compared to part (a).

Which levels of soil differ significantly?
The analyst may elect to refit the model, dropping treatment. We do not do so here.

> with(corn[-6, ], tapply(bushels, soil, mean))

    I   II   III
19.700 48.280 62.125

> with(corn[-6, ], table(soil))

soil

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>II</th>
<th>III</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>

> vcov(corn.lm.1a)[1:3, 1:3]

   (Intercept)  soilII  soilIII
(Intercept)  18.27137 -7.69321 -8.65486
soilII     -7.69321 15.38642  7.69321
soilIII    -8.65486  7.69321 18.27137

> require(gmodels)
> (soil.2 <- estimable(corn.lm.1a, rbind("2 - 1" = c(0, + 1, 0, 0, 0, 0, 0), "3 - 1" = c(0, 0, 1, 0, 0, 0, + 0), "2 - 3" = c(0, 1, -1, 0, 0, 0, 0)))

    Estimate Std. Error   t value DF Pr(>|t|)
2 - 1   28.580     3.922553  7.286072  7  1.647232e-04
3 - 1  45.015     4.274503 10.531050  7  1.519581e-05
2 - 3  -16.435     4.274503 -3.844892  7  6.334235e-03

> qt(0.975, df = 7) * soil.2[, 2]
Using computations similar to those outlined above, 
\[ \text{LSD}_t = t_{0.975} \times \text{se(estimate)} = 10.11 \text{ or } 9.28 \text{ according to whether soil 3 is or is not involved.} \]
\[ \text{LSD}_Q = \frac{Q(3,7,0.95)}{\sqrt{2}} \times \text{se(estimate)} = 12.56 \text{ or } 11.55 \text{ according to whether soil 3 is or is not involved.} \]
Differences between pairs of soil types are all significant by both Fisher’s and Tukey’s methods. Highest yield obtained with soil 3; yield estimated (on average) to be 16.4 units higher than for soil 2 and 45 higher than for soil 1.

(b) Because the omitted value is from treatment 2, which is not involved in the comparison of interest, it turns out that the only change from part (a) ii. is the \( \text{se(estimate)} \) and the \( \text{df} \) for the \( t \)-value. So
\[ \text{Estimate} = 43.63 - 35.67 = 7.97, \]
\[ \text{se(estimate)} = \sqrt{\frac{s^2(\frac{1}{g} + \frac{1}{3})}{\frac{4}{g}}} = 4.13. \]
\[ 95\% \text{ CI} = 7.97 \pm t_{0.975} \times 4.13 = 7.97 \pm 9.77 = (-1.80, 17.74) \]

**Exercise 3**

The data below were obtained from a study of the effect of oven temperature and baking time on the life (in hours) of an electrical component.

<table>
<thead>
<tr>
<th>Oven Temperature (°F)</th>
<th>Baking Time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5</td>
</tr>
<tr>
<td>600</td>
<td>165</td>
</tr>
<tr>
<td></td>
<td>220</td>
</tr>
<tr>
<td></td>
<td>212</td>
</tr>
<tr>
<td>620</td>
<td>170</td>
</tr>
<tr>
<td></td>
<td>192</td>
</tr>
<tr>
<td></td>
<td>181</td>
</tr>
<tr>
<td>640</td>
<td>126</td>
</tr>
<tr>
<td></td>
<td>160</td>
</tr>
<tr>
<td></td>
<td>149</td>
</tr>
</tbody>
</table>

1. For each of the following state, with reasons, whether or not it would be a reasonable study design to justify use of a standard two-way ANOVA to analyse the data.

(a) A total of three firings of the oven, one firing at each of the three temperatures with nine components per firing and three components removed after 5, a three (of the remaining six components) after 10 and the final three components after 15 minutes.

(b) A total of nine firings of the oven, three firings at each of the three temperatures with three components per firing and one component removed after each of 5, 10 and 15 minutes.
(c) A total of 27 firings of the oven, nine at each temperature, with one component per firing.

2. Show that the strictly additive model is reasonable for these data.

3. Using the strictly additive model, obtain an estimate of the baking conditions (temperature and time), within the ranges of temperature and time used in the study, for which the expected life of a component is a maximum.

### Answer of Exercise 3

```r
> ovens <- data.frame(time = rep(c(5, 10, 15), each = 9),
+     temp = rep(rep(c(600, 620, 640), each = 3), 3), life = c(165,
+     220, 212, 170, 192, 181, 126, 160, 149, 196,
+     228, 232, 176, 224, 217, 156, 138, 157, 170,
+     203, 198, 168, 189, 195, 116, 142, 159))
```

Use an interaction plot to display the data.

```r
> par(las = 1)
> with(ovens, interaction.plot(time, temp, life))
```

1. Design (iii) is appropriate. Observations within the same firing might be correlated (if one firing is bad, then all observations within that firing will be bad).

2. The test for interaction between temperature and time is not significant (P = 0.891).
> ovens.lm.1 <- lm(life ~ factor(temp) * factor(time), + data = ovens)
> opar <- par(mfrow = c(2, 2), las = 1, mar = c(4, 4, 3, + 2))
> plot(ovens.lm.1)
> par(opar)

> anova(ovens.lm.1)

Analysis of Variance Table

Response: life

<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>factor(temp)</td>
<td>2</td>
<td>16713.6</td>
<td>8356.8</td>
<td>21.6913</td>
</tr>
<tr>
<td>factor(time)</td>
<td>2</td>
<td>2121.6</td>
<td>1060.8</td>
<td>2.7534</td>
</tr>
<tr>
<td>factor(temp):factor(time)</td>
<td>4</td>
<td>422.9</td>
<td>105.7</td>
<td>0.2744</td>
</tr>
<tr>
<td>Residuals</td>
<td>18</td>
<td>6934.7</td>
<td>385.3</td>
<td></td>
</tr>
</tbody>
</table>

---

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

3. Fit what amounts to the additive model using linear and quadratic terms in temperature and time. Fit the model $y_{ijk} = \beta_0 + \beta_1 temp + \beta_2 temp^2 + \beta_3 time + \beta_4 time^2 + e_{ijk}$ and find estimates of the $\beta$s.
> ovens.lm.2 <- lm(life ~ temp + I(temp^2) + time + I(time^2),
+     data = ovens)
> coef(summary(ovens.lm.2))

| Estimate   | Std. Error | t value | Pr(>|t|) |
|------------|------------|---------|----------|
| (Intercept)| -14826.7778| 7.171008e+03| -2.067600 | 0.05063846 |
| temp       | 49.70278   | 2.314517e+01 | 2.147436 | 0.04303391 |
| I(temp^2)  | -0.04125   | 1.866465e-02 | -2.210060 | 0.03780798 |
| time       | 14.41111   | 6.034582e+00 | 2.388088 | 0.02595175 |
| I(time^2)  | -0.74000   | 2.986344e-01 | -2.477946 | 0.02137038 |

Use calculus to find the location of the turning point.
For temperature: maximum (is estimated to occur) at \( \frac{49.70278}{2 \times 0.04125} = 602.5^{\circ} \text{ F}. \)
For time: maximum (is estimated to occur) at \( \frac{14.41111}{2 \times 0.74} = 9.7 \text{ min}. \)