RESEARCH SCHOOL OF
FINANCE, ACTUARIAL STUDIES AND APPLIED STATISTICS
College of Business & Economics, The Australian National University

GENERALISED LINEAR MODELS

(STAT3015/STAT4030/STAT7030)

Solutions to Assignment 1 for 2017

In these model solutions, I have included the assignment questions and a few additional comments in italics. These comments are not an essential part of the solutions to the assignment, but they do make this document into a more readable, self-contained report. Even with these optional "extras" this report is well within the 10 *page limit and this result is achieved without using an unreasonably small font size or any unreadable formatting.*

The essential solutions are the parts NOT in italics, but there are also a number of "tangents" or different approaches you could have included in your analysis. I discuss a number of these "tangents" in the comments which I have included in the appendix of R *commands. Your solutions may sensibly include some of these "tangents", but if you include too many and go over the page limit as a result, you may well lose marks. On the other hand, good additional discussion, that is both concise and well expressed, may well compensate if you miss a few of the points that we were expecting to find.*

Question 1 **(20 marks)**

Neter et al in the text *Applied Linear Statistical Models* (4th edn, Irwin, 1996, p.1159) describe the results of a marketing experiment to investigate the effect of colour of paper (blue, green or orange) on the response rates for questionnaires distributed by the "windshield method" in supermarket parking lots. A representative sample of 15 supermarket car parks were chosen in a metropolitan area and 5 car parks were assigned at random to each of the three colours. The entire experiment was repeated in a different week, with the same colours assigned to the same car parks.

The data are available in the file qcolour.csv, which is available on Wattle. The variables are:

- rrate – the observed response rates (the percentage of questionnaires returned);
- colour – of the paper (blue, green or orange);
- size – of the car park (measured by counting the number of parking spaces); and
- week – week A or week B.

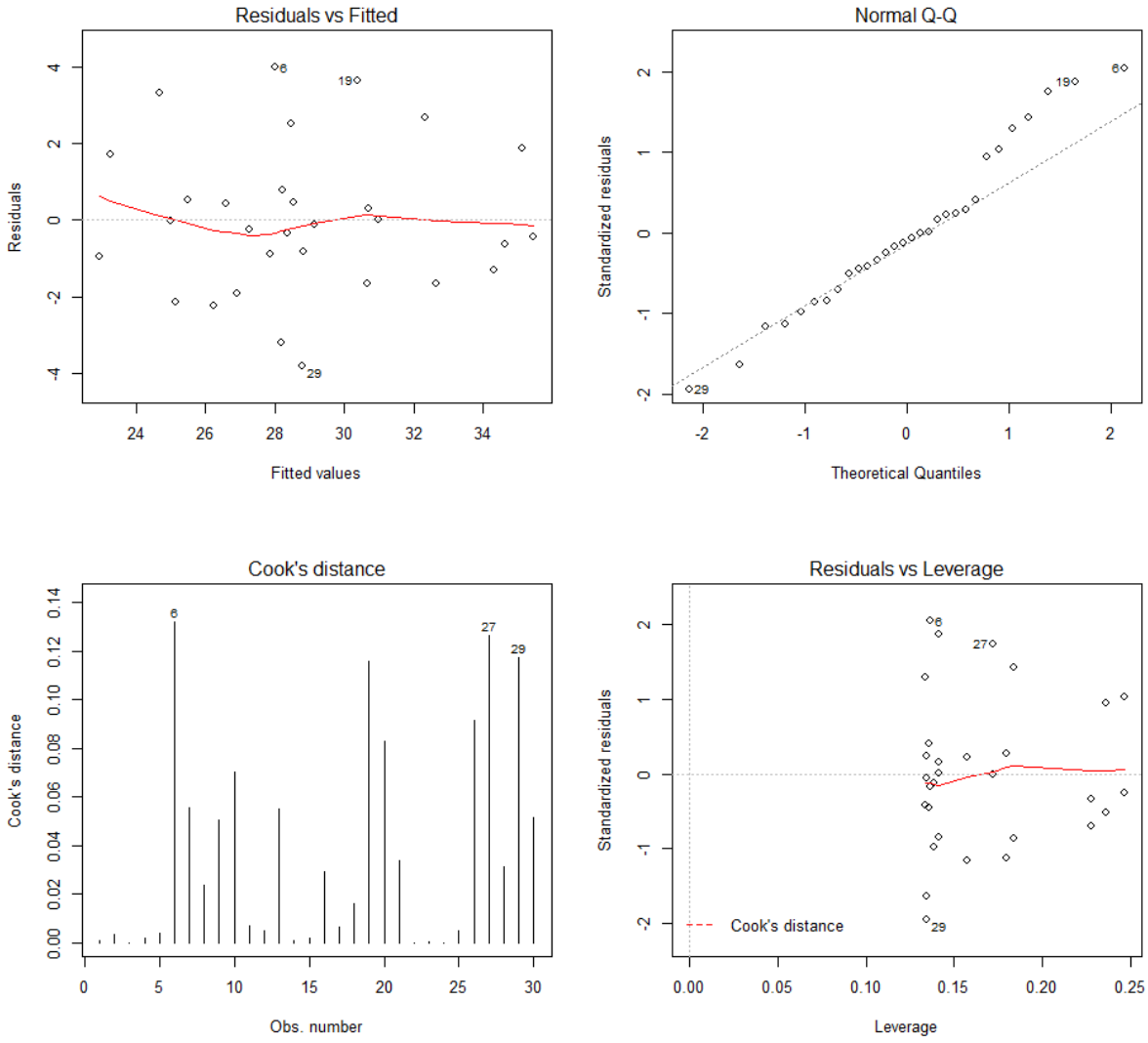
Note that the response variable rrate *is a percentage and so is really a binomial proportion, however, there are no values close to either* 0% *or* 100%*, so initially, it is not unreasonable to assume approximate normality and use an ordinary multiple regression model, rather than a more complicated generalised linear model.*

There would also be problems with interpreting the response as a binomial proportion and fitting a logistic regression, as we do not know the sample sizes for each car park – using size of car park as a proxy for the number of questionnaires that were handed out assumes that each car park was full at the time of the survey, which is stretching credibility a little.

summary(lgcolour)
also tells it's a balanced design.

Question 1 continued

(a) Using R, fit an ordinary (normally distributed) additive linear model with rrate as the response variable, colour and week as exploratory factors and size as a continuous covariate. Do not vary the default contrasts used by R. For this model produce: a plot of the residuals against the fitted values for this model; a normal quantile plot of the residuals; a bar plot of Cook's Distances for each of the observations; and a plot of the standardised residuals against the leverage values. Are there any obvious problems with these plots? **(2 marks)**



The main residual plot ("Residuals vs Fitted") has no obvious problems with lack of independence or non-constant variance. The three points highlighted by default are not really outlying in the sense that they are not extreme compared to the next highest (or the next lowest) residual.

The normal quantile plot shows only a slight deviation from normality in the upper tail (probably ignorable given the relatively small sample size), and only 1 of the 30 observations (less than 4%) has a standardised residual value just outside of the 2 standard deviation range ($r_6 = 2.045$).

Finally, the two outlier plots show no real problems, with no observation having relatively high leverage or a relatively extreme value of Cook's D, with the largest Cook's D again being observation 6 ($d_6 = 0.132$); but that observation has only relatively small leverage ($h_6 = 0.136$, compared with an average of $p/n = 5/30 = 0.167$).

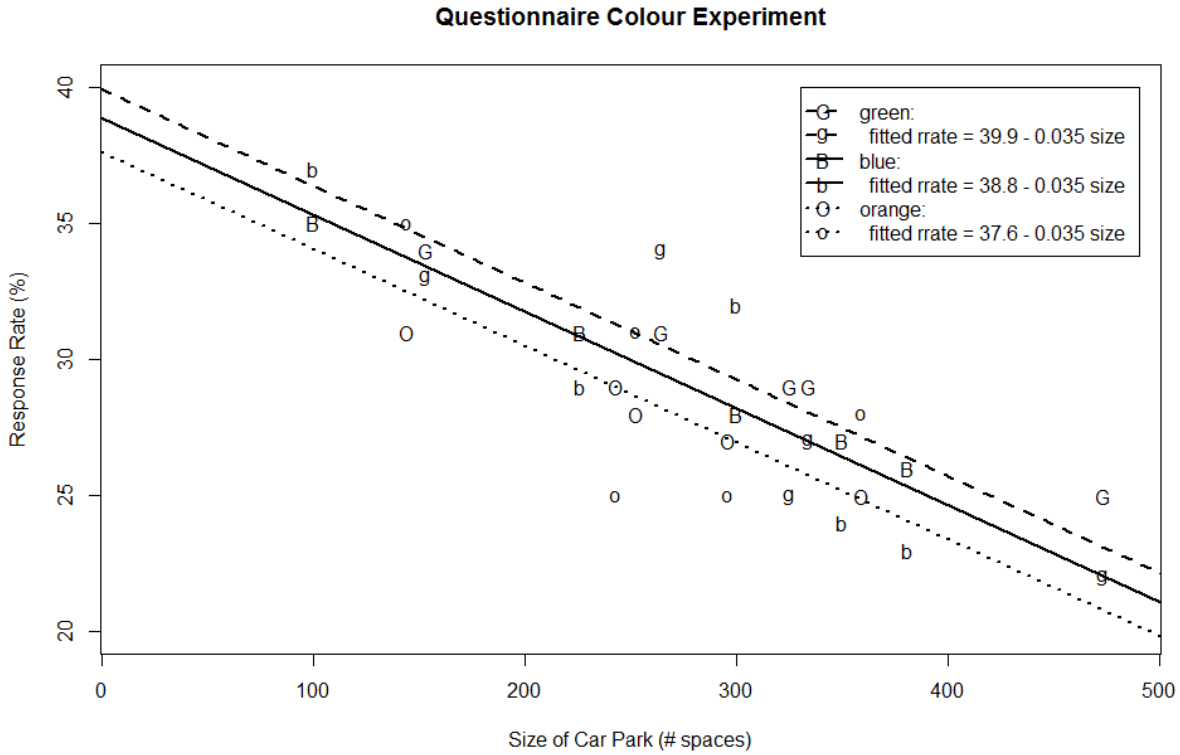
Question 1 continued

(b) Is week an important factor in the model in part (a)? Present appropriate R output to support your conclusion. Re-fit the model in part (a), without week as a factor. Produce a plot of the data with rrate on the vertical axis and size on the horizontal axis, using plotting symbols as follows: B for blue questionnaires in week A; b for blue questionnaires in week B; G for green questionnaires in week A; g for green questionnaires in week B, O for orange questionnaires in week A and o for orange questionnaires in week B. What are the fitted regression lines from the reduced model (excluding week) for each of the three colours? Include these lines on the plot and also include an appropriate legend. **(2 marks)**

```
> anova(qcolour.lma)
Analysis of Variance Table

Response: rrate
          Df Sum Sq Mean Sq F value    Pr(>F)
colour     2   3.27    1.63  0.3720    0.6931
week       1   0.83    0.83  0.1898    0.6668
size       1 326.29  326.29 74.3075 5.866e-09 ***
Residuals 25 109.78    4.39
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

No, week is not an important factor ($F_{1,25} = 0.19$, $p = 0.67 < \alpha = 0.05$), suggesting that there were no significant differences in the mean response rates between the two different weeks during which the experiment was conducted.
Quoting a t statistic rather than an F statistic is only worth 0.5 *of the* 1 *mark available here; think about what would have happened if I had asked about* colour *instead?*



This is clearly an example of a “parallel lines” analysis of covariance (ANCOVA) model.

Question 1 continued

(c) For the reduced model in part (b), give the algebraic equation for the underlying population model, including any assumptions about the error distribution, details of transformations (if any) applied to the variables and the constraints applied to any factor variables. Present appropriate R output that gives a summary of the fitted coefficients of this model and interpret the significance of these coefficients. Are the contrasts that have been used in the model a good choice to address the research question for this experiment? **(4 marks)**

The underlying population model is: $Y_{ij} = \beta_0 + \tau_j + \beta_1 X_{ij} + \varepsilon_{ij}$

where Y is the response rate (rrate),

X is the size of the car park

j indicates the levels of colour = {"blue", "green", "orange"} = {1, 2, 3}, and

i = 1, 2, … 10 for all 3 levels of j, for a total of 3 x 10 = 30 observations,

with the constraint: $\tau_1 = \tau_{blue} = 0$ and

assuming the errors (ε_{ij}) are independently and identically distributed N(0, σ^2).

A summary of the model fitted to the sample data using R is:

```
> summary(qcolour.lmb)
Call:
lm(formula = rrate ~ colour + size)
Residuals:
    Min      1Q  Median      3Q     Max
-3.9608 -1.3761 -0.0202  0.8919  3.8152
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  38.833718   1.278859  30.366  < 2e-16 ***
colourgreen   1.063061   0.935453   1.136    0.266
colourorange -1.247254   0.923827  -1.350    0.189
size         -0.035496   0.004053  -8.758 3.11e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.063 on 26 degrees of freedom
Multiple R-squared:  0.7487,     Adjusted R-squared:  0.7197
F-statistic: 25.82 on 3 and 26 DF,  p-value: 5.838e-08
```

The obvious contrasts to use in fitting the model would be sum contrasts, as there is no obvious control group in this experiment. However, the research question that the researchers are most likely to be interested in: "is there a difference in the response rates due to the different questionnaire colours?"; requires a direct comparison of the three colours and so treatment contrasts will probably be more useful. To do this, we need to arbitrarily pick one of the three colours to be the control or reference group.

A good choice of reference group is often to pick the group closest to the overall mean response (i.e. the middle group). In this case, that is the "blue" group, which luckily turns out to the one that R uses by default (i.e. using treatment contrasts, the default R reference group is the first group in alphabetical order).

The above coefficients for the "green" group (t_{26} = 1.14, p = 0.27) and the "orange" group (t_{26} = –1.35, p = 0.19) can be used to decide if either group differs significantly from the "blue" group. We can infer from the large p-values for these coefficients and from the F statistic ($F_{2,25}$ = 0.37, p = 0.69) from the ANOVA table for the closely related model in part (b) (or from the ANOVA table for this model, included in the R appendix), that there are no significant differences between the three colours.

However, response rates do decrease significantly (β_{size} = –0.0355, t_{26} = –8.76, p = 0.00) as the size of the car park increases.

Question 1 continued

(d) Use the reduced model in part (b) to estimate the response rate for questionnaires of all three different colours which have been distributed in a car park with 250 spaces and find 95% confidence intervals for these estimates. **(2 marks)**

```
> newcarpark <- data.frame(colour=c("blue","green","orange"),
 size=c(250,250,250))
> row.names(newcarpark) <- c("blue","green","orange")
> newcarpark
       colour size
blue     blue  250
green   green  250
orange orange  250
>
> predict(qcolour.lmb, newdata=newcarpark, interval="confidence")
            fit      lwr      upr
blue   29.95962 28.60711 31.31213
green  31.02268 29.59240 32.45297
orange 28.71237 27.36966 30.05508
```

Note that the three confidence intervals all overlap, reflecting the fact that there are not significant differences between the three colours.

(e) Compare the reduced model in part (b) with a multiplicative model that includes an interaction term between the factor variable (colour) and the covariate (size). Describe how this changes the algebraic equation in part (c). Is this additional term a significant improvement to the model? Present some R output and give full details of an appropriate hypothesis test. What do your results suggest about the relationship between the response rates and the explanatory variables and factors? **(2 marks)**

If we add an interaction term to the model in part (b), to allow the slopes to differ for each of the three colours (as well as the intercepts), we get the following model:

$$Y_{ij} = \beta_0 + \tau_j + \beta_1 X_{ij} + \delta_j X_{ij} + \varepsilon_{ij}$$

where Y is the response rate (rrate),

X is the size of the car park

j indicates the levels of colour = {"blue", "green", "orange"} = {1, 2, 3}, and

i = 1, 2, …, 10 for all 3 levels of j; for a total of 10 x 3 = 30 observations,

with the constraints: $\tau_1 = \tau_{blue} = 0$, $\delta_1 = \delta_{blue} = 0$ and

assuming the errors (ε_{ij}) are independently and identically distributed N(0, σ^2).

The analysis of variance table for this expanded model is:

```
> qcolour.lmb_int <- lm(rrate ~ colour * size)
> anova(qcolour.lmb_int)
Analysis of Variance Table

Response: rrate
            Df Sum Sq Mean Sq F value    Pr(>F)
colour       2   3.27    1.63  0.3628    0.6995
size         1 326.29  326.29 72.4698 1.032e-08 ***
colour:size  2   2.55    1.28  0.2834    0.7557
Residuals   24 108.06    4.50
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As $F_{2,24} = 0.28$ ($p = 0.76 > \alpha = 0.05$), the hypothesis test associated with the interaction term is not significant, so we would accept the following null hypothesis and conclude that this term is not a significant addition to the model:

$H_0: \quad \delta_{blue} = \delta_{green} = \delta_{orange} = 0 \quad$ vs $\quad H_A: \quad$ not all $\delta = 0$ (at least one $\delta \neq 0$)

So, we do not need to move to a "separate lines" model with different slopes as well as different intercepts, nor do we really require different intercepts for the three colours.

Question 1 continued

(f) Fit the reduced model in part (b) to the data for week A only and repeat the analysis in part (c). What would you conclude about the effect of questionnaire colour on response rates, if there had only been one round of this experiment? **(3 marks)**

Fitting the same model as described in parts (b) and (c) to the reduced data set gives the following analysis of variance table and table of coefficients:

```
> qcolour.lmf <- lm(rrate ~ colour + size, data=qcolour[week=="A",])
> anova(qcolour.lmf)
Analysis of Variance Table
Response: rrate
          Df  Sum Sq Mean Sq F value    Pr(>F)
colour     2   7.600    3.80  31.758 2.693e-05 ***
size       1 115.084  115.08 961.809 4.645e-12 ***
Residuals 11   1.316    0.12
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> summary(qcolour.lmf)$coefficients
                Estimate   Std. Error    t value     Pr(>|t|)
(Intercept)  37.49122495 0.3033119764 123.606148 1.216239e-18
colourgreen   1.34481591 0.2218648844   6.061418 8.174650e-05
colourorange -1.77564272 0.2191074932  -8.103980 5.776210e-06
size         -0.02981291 0.0009613022 -31.013051 4.644962e-12
```

The very small *p*-value associated with the colour term in the ANOVA table is considerably smaller than $\alpha = 0.05$ ($F_{2,11} = 31.8$, $p = 0.00003$), which indicates that there are significant differences in the response rates between the three colours, i.e. that we should reject the following null hypothesis in favour of the alternative:

$$\mathrm{H}_0: \quad \tau_{blue} = \tau_{green} = \tau_{orange} = 0 \quad \text{vs} \quad \mathrm{H}_\mathrm{A}: \quad \text{not all } \tau = 0 \text{ (at least one } \tau \neq 0)$$

The *t* test associated with the colourgreen coefficient provides a test for the difference in the intercepts between the green and blue groups:

$$\mathrm{H}_0: \quad \tau_{green} = 0 \quad \text{vs} \quad \mathrm{H}_\mathrm{A}: \quad \tau_{green} \neq 0$$

So, as $t_{11} = 6.1$ ($p = 0.00008 < \alpha = 0.05$), we should reject the null hypothesis in favour of the alternative and conclude that there is a significant difference.

Note that as the colourgreen *coefficient is positive, we can infer that green questionnaires produce, on average, a* 1.3% *significantly higher response rate than blue questionnaires. Using a* 95% *confidence interval (and rounding outwards to ensure at least the required confidence), the expected increase in response rates is between:*

$$\tau_{green} \pm t_{11}(0.975) \cdot se(\tau_{green}) = 1.345 \pm (2.201) \cdot (0.222) = (0.8\%, 1.9\%)$$

Similarly, the *t* test associated with the colourorange coefficient provides a test for the difference in the intercepts between the orange and blue groups:

$$\mathrm{H}_0: \quad \tau_{orange} = 0 \quad \text{vs} \quad \mathrm{H}_\mathrm{A}: \quad \tau_{orange} \neq 0$$

Again, as $t_{11} = -8.1$ ($p = 0.000006 < \alpha = 0.05$), we should reject the null hypothesis in favour of the alternative and conclude that there is a significant difference.

As the colourorange *coefficient is negative, we can infer that orange questionnaires produce, on average, a* 1.8% *significantly lower response rate than blue questionnaires. Using a* 95% *confidence interval the expected decrease in response rates is between:*

$$\tau_{orange} \pm t_{11}(0.975) \cdot se(\tau_{orange}) = -1.776 \pm (2.201) \cdot (0.219) = (-2.3\%, -1.2\%)$$

Question 1, part (f) continued

As you can see from the plot in part (b), blue lies in the middle of the three groups and as green is significantly higher than blue and orange is significantly lower than blue, then green and orange will also differ significantly and therefore all three colours differ significantly, with green giving the best response rate. This last piece of information is probably what the researchers really want to know!

The big problem here is that when we included the data from week B and analysed the combined results in the earlier parts of this question, we did not get the same results and the same significant p-values. In short, we were unable to successfully replicate the results achieved in week A, when we repeated the entire experiment in week B. This casts some serious doubt on the above results for week A and might even suggest that they may have just been an artefact.

Note that the real point of this experiment is whether or not the response rates differ for the three colours. In order to address this underlying research question, the model must include colour *as a factor variable. The simple linear regression model of* rrate *on* size *(a "single regression model") would not be an adequate model for the data, given this research question, as it would not address the research question.*

The one consistent result is the small p-value associated with the size term in the earlier results and in the above ANOVA table ($F_{1,11} = 961.8$, $p = 0.00000$) which does indicate that there is a significant linear relationship between the response rates and the continuous covariate (size), i.e. that we should reject the following null hypothesis in favour of the alternative:

$$H_0: \quad \beta_1 = 0 \quad \text{vs} \quad H_A: \quad \beta_1 \neq 0$$

The t test associated with the size *coefficient also tests the above hypotheses.* I*n fact, if we square* $t_{11} = -31.0131$, *we get* $F_{1,11} = 961.8$; *so the two tests are equivalent and have identical p-values (p = 0) and the same conclusion (reject the null hypothesis).*

As the size *coefficient is negative, we can infer that the larger the car park, the lower the response rate. For each increase of* 100 *parking spaces in the size of the car park, the expected response rate decreases by almost* 3% (2.98%). *Using* 100 *times both the estimated coefficient and standard error, we can calculate a* 95% *confidence interval for the expected decrease per each additional* 100 *spaces:*

$$100\beta_1 \pm t_{11}(0.975) \cdot 100se(\beta_1) = 100(-0.0298) \pm (2.201) \cdot 100(0.001) = (-3.2\%, -2.7\%)$$

The significant relationship between the response rates and the covariate size *implies that we do need to control for the effects of this covariate in the model, and we cannot remove the corresponding term from the model, so the one-way analysis of variance model of* rrate *on* colour *would also not be an adequate model for these data.*

Arguably the model fitted in parts (b) and (c) to the full data and above to just the data for week A, is the right model for these data, given the underlying research question. The problem is that it gives a positive answer to the research question when fitted to the data for week A, but a negative answer when fitted to the combined data for the two weeks combined. I will discuss a better approach to conducting this experiment in part (h) below.

Question 1 continued

(g) Now modify the reduced model in part (b) to include week as a random effect in an additive mixed effects model for the full data (not just week A). Describe the changes to the underlying population model described in part (c). Present and examine the summary output (analysis of variance table and table of coefficients) for the new mixed effects model. How has this changed from the summary output presented in part (c)? Calculate the intra-class correlation coefficient for the mixed effects model and comment on the results. **(3 marks)**

Adding week as a random effect to the model described in part (b), changes the description given in part (c) as follows:

$$Y_{ijk} = \beta_0 + \tau_j + \eta_k + \beta_1 X_{ijk} + \varepsilon_{ijk}$$

Where Y, X and j are as before (as is the constraint applied to the fixed factor $\tau_{blue} = 0$), however the new subscript k indicates the two weeks k = {"A", "B"} = {1, 2} and the other subscript i is now i = 1, 2, 3, 4, 5 observations for all combinations of j and k.

The variance model now has two independent components:

$\eta_k \sim i.i.d.\ N\left(0, \sigma_\eta^2\right)$ and $\varepsilon_{ijk} \sim i.i.d.\ N\left(0, \sigma_\varepsilon^2\right)$.

Using the lme() function from the nlme library to fit this model:

```
> qcolour.lme <- lme(rrate ~ colour + size, random=~1|factor(week))
>
> anova(qcolour.lme)
            numDF denDF  F-value p-value
(Intercept)     1    25 5862.587  <.0001
colour          2    25    0.384  0.6851
size            1    25   76.698  <.0001
>
> summary(qcolour.lme)
Linear mixed-effects model fit by REML
 Data: NULL
       AIC      BIC   logLik
  142.8028 150.3514 -65.4014

Random effects:
 Formula: ~1 | factor(week)
         (Intercept) Residual
StdDev: 7.935918e-05  2.06258

Fixed effects: rrate ~ colour + size
                Value Std.Error DF   t-value p-value
(Intercept)  38.83372 1.2788593 25 30.365904  0.0000
colourgreen   1.06306 0.9354526 25  1.136414  0.2666
colourorange -1.24725 0.9238265 25 -1.350096  0.1891
size         -0.03550 0.0040532 25 -8.757718  0.0000
 Correlation:
             (Intr) clrgrn clrrng
colourgreen  -0.212
colourorange -0.408  0.483
size         -0.860 -0.166  0.055

Standardized Within-Group Residuals:
         Min           Q1          Med           Q3          Max
-1.920334074 -0.667184839 -0.009788086  0.432400283  1.849720430

Number of Observations: 30
Number of Groups: 2
```

Note that using the other function discussed in lectures (the lmer() *function from the* lme4 *library) produces slightly different results – see the* R *appendix for details.*

Question 1, part (g) continued

There has been almost NO real change from the model in parts (b) and (c), the coefficients for the fixed effects and the various *F* and *t* statistics are almost identical. The variance component associated with the additional random effects term is almost 0 and the residual standard error is unchanged (to 3 decimal places) at 2.063.

The intra-class correlation coefficient is calculated as follows:

$$\frac{\hat{\sigma}_{\eta}^2}{\hat{\sigma}_{\eta}^2 + \hat{\sigma}_{\varepsilon}^2} = \frac{\left(7.935918 \times 10^{-5}\right)^2}{\left(7.935918 \times 10^{-5}\right)^2 + \left(2.06258\right)^2} \approx 0$$

So, there does not appear to have been any real additional variability between the two weeks (the first and second repeats of the experiment).

(h) Finally, discuss the results of all the above analysis. You might decide to discuss the fit of the various models, whether or not there has been an appropriate treatment of each of the variables and/or aspects of the experimental design. (Hint – we are after a good concise discussion of the important issues. This part of the question is worth as many marks as most of the other parts, so very short discussions will not get full marks; though long and winding discussions that miss the important points and even worse, cause you to exceed the overall page limit, will also not get full marks). **(2 marks)**

If there is a genuine difference in the mean response rates for the different colours of questionnaire, then it is a subtle effect which has become lost in this experiment, in the other sources of uncontrolled variation in the response rates.

The largest source of uncontrolled variation is variability between the different car parks. The fact the researchers did replicate the entire experiment in the same 15 randomly selected car parks suggests that it is possible to repeatedly sample the same car parks; though as this may be repeatedly sampling the same customers, there may be some effect the second or third time you sample the same car park (though the results of parts (b) and (g) suggest this is probably not the case in this instance). You could control for this by increasing the interval between the sampled weeks or by offering some real incentive for returning questionnaires (such as a discount voucher – the relatively high response rates achieved suggests there probably was some incentive).

The researchers could have used the different car parks as a (random) blocking variable, i.e. controlled for the differences between car parks by repeating the important experimental treatments within each selected car park. This would have involved repeatedly testing all three colours in the same car parks over three well-spaced weeks (rather than using the same colours in the same car parks).

There are 6 possible orders in which you can arrange 3 colours: (b, g, o); (b, o, g); (g, b, o); (g, o, b); (o, b, g); and (o, g, b). If you had the resources to conduct 2 runs of the experiment in 15 randomly selected car parks, you certainly could have randomly selected 6 car parks, randomly assigned one of the above orders to each car park and sampled each of the car parks with each of the three colours, for a total of $6 \times 3 = 18$ observations. You could randomly select a different 6 car parks and repeat the entire experiment for only marginally more resources than the current experiment (36 samples rather than 30). Such an experimental design would effectively control for differences between the car parks and for any effects due to the order in which you use the three colours.