STAT3015/4030/7030 Generalised Linear Modelling

Tutorial 11

1. The data file `Sprngs.txt` on Wattle contains data regarding the uncompressed height in inches of truck springs manufactured under various conditions. Five covariates define the manufacturing conditions, and each are two-level factor variables:

   - `ftmp` - furnace temperature (0 = low, 1 = high)
   - `htme` - heating time (0 = short, 1 = long)
   - `ttmp` - transfer time (0 = short, 1 = long)
   - `hdtmp` - hold-down time (0 = short, 1 = long)
   - `qot` - quench oil temperature (0 = low, 1 = high)

   (a) Fit an analysis of variance model to this data and determine the significant factors and interactions. The experiment was intended to identify conditions yielding springs having heights of exactly 8 inches. Which factor settings seem the best ones to achieve this goal?

   Solution: The necessary *R* commands are:

```
> spg <- read.table("Sprngs.txt",header=T)
> attach(spg)
> head(spg)

  ftmp htme ttme hdtme qot hght
1    0    0    0     0   0 7.78
2    0    0    0     0   0 7.78
3    0    0    0     0   0 7.81
4    1    0    0     1   0 8.15
5    1    0    0     1   0 8.18
6    1    0    0     1   0 7.88

> spg.aov <- aov(hght~ftmp+htme+ttme+hdtme+qot)
> anova(spg.aov)

Analysis of Variance Table

Response: hght
```

```
          Df  Sum Sq Mean Sq F value    Pr(>F)
ftmp       1 0.58742 0.58742 23.3175 1.852e-05 ***
htme       1 0.37277 0.37277 14.7970 0.0004005 ***
ttme       1 0.00992 0.00992  0.3937 0.5337453
hdtme      1 0.12917 0.12917  5.1273 0.0287751 *
qot        1 0.80860 0.80860 32.0974 1.201e-06 ***
Residuals 42 1.05807 0.02519
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> spg.aov2 <- aov(hght~ftmp+htme+qot+hdtme+ftmp*htme+
+                   ftmp*qot+ftmp*hdtme+htme*qot+htme*hdtme+qot*hdtme)
> anova(spg.aov2)

Analysis of Variance Table

Response: hght
           Df  Sum Sq Mean Sq F value    Pr(>F)
ftmp        1 0.58742 0.58742 34.9526 8.259e-07 ***
htme        1 0.37277 0.37277 22.1805 3.447e-05 ***
qot         1 0.80860 0.80860 48.1135 3.446e-08 ***
hdtme       1 0.12917 0.12917  7.6858  0.008663 **
ftmp:htme   1 0.00350 0.00350  0.2084  0.650708
ftmp:qot    1 0.08585 0.08585  5.1084  0.029785 *
ftmp:hdtme  1 0.01505 0.01505  0.8956  0.350096
htme:qot    1 0.32835 0.32835 19.5376 8.318e-05 ***
htme:hdtme  1 0.00460 0.00460  0.2738  0.603892
qot:hdtme   1 0.00880 0.00880  0.5237  0.473801
Residuals  37 0.62183 0.01681
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So, the predictor `ttme` is not significant, but there are two significant two-way interactions. Note that due to the nature of the design of this experiment (i.e., which settings were chosen) the predictors are "orthogonal", so that rearranging the order of the predictors in the model does not change the partial sums-of-squares, and thus there is no need to reorder the second ANOVA table for testing specific hypothesis (and indeed the reordering of the first ANOVA table was unnecessary as well, but it serves to demonstrate the validity of the above statements). Now, the fitted values at each of the settings (excluding `ttme` since it is insignificant) are:

```
> spg.aov3 <- aov(hght~ftmp+htme+qot+hdtme+ftmp*qot+htme*qot)
> ftmpset <- c(rep(0, 8),rep(1, 8))
> htmeset <- rep(c(rep(0, 4),rep(1, 4)), 2)
> qotset <- rep(c(rep(0, 2),rep(1, 2)), 4)
```

```
> hdtmeset <- rep(c(0, 1), 8)
> int1set <- ftmpset*qotset
> int2set <- htmeset*qotset
> settings <- cbind(1, ftmpset, htmeset, qotset, hdtmeset, int1set, int2set)
> inches <- settings%*%spg.aov3$coef
> bestset <- cbind(settings[, -1], inches)
> bestset <- bestset[rev(order(inches)), ]
> bestset[1:3, ]

     ftmpset htmeset qotset hdtmeset int1set int2set
[1,]       1       0      0        1       0       0 8.056875
[2,]       1       0      0        0       0       0 7.953125
[3,]       0       0      0        1       0       0 7.920208
```

So, the best production settings are when the furnace temperature is high, the heating time is short and the quench oil temperature is low. The other factors do not seem to be overly important. Note that the hold-down time was statistically significant, but it does not appear to be practically significant, an important distinction.

(b) The preceding analysis is only valid under the assumption of homoscedasticity. Examine this assumption by comparing the "within group" variances determined by the model you chose in part (a). In other words, calculate the variance of each group of data points for which the predictor variables included in your part (a) model are the same.

Solution: The necessary *R* commands are:

```
> gpdat <- matrix(hght, ncol=3, byrow=TRUE)
> s2 <- apply(gpdat,1,var)
> s2

 [1] 0.000300000 0.027300000 0.001200000 0.010433333 0.003600000 0.049633333
 [7] 0.008400000 0.015633333 0.037300000 0.064533333 0.001200000 0.009233333
[13] 0.004800000 0.004233333 0.001633333 0.025433333
```

So, it does appear that there is a difference in the variances, but is this spread consistent with chance variation?

(c) More formally, we can test our homoscedasticity assumption by fitting a GLM to the "within group" variances and determining if any of the predictors are significant. Generally, we would use a gamma GLM with logarithmic link and the same model structure as was chosen in part (a). Is there evidence of heteroscedasticity based on this approach?

Solution: The necessary R commands are:

```
> preds <- cbind(ftmp, htme, hdtme, qot, ftmp*qot, htme*qot)
> preds <- preds[seq(1, 46, 3), ]
> s2.glm <- glm(s2 ~ preds ,family = Gamma(link = log), maxit = 50)
> anova(s2.glm, test="Chisq")

Analysis of Deviance Table

Model: Gamma, link: log

Response: s2

Terms added sequentially (first to last)


      Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                     15     26.566
preds  6    13.35         9     13.217  0.04007 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> summary(s2.glm)$dispersion

[1] 1.011844
```

So, there is some evidence of heteroscedasticity.

(d) In fact, it can be shown that sample variances tend to have gamma distributions with $\alpha = 0.5$. Does the analysis here seem consistent with this fact?

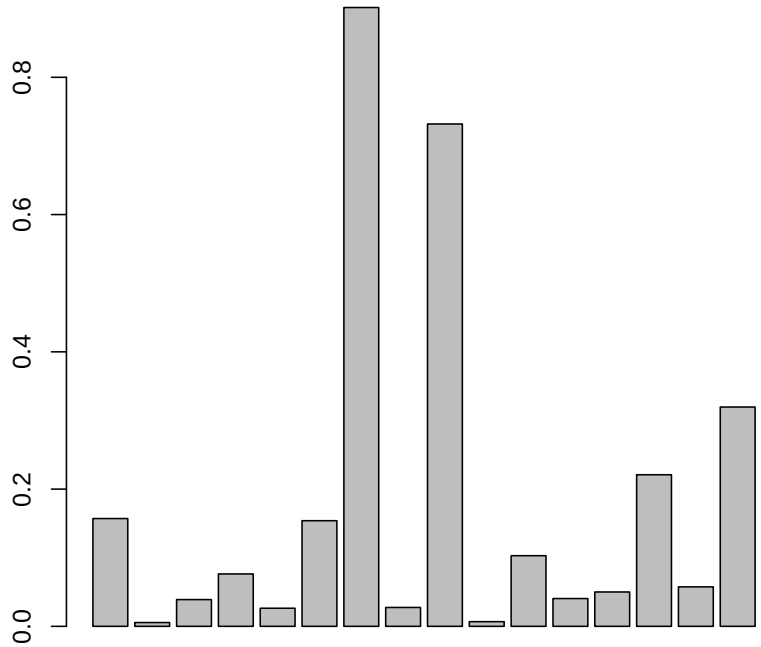
Solution: From above, we see that the dispersion estimate is 1.011844, which is not overly consistent with a true value of $1/\alpha = 1/0.5 = 2$.

(e) For the GLM fit in part (c), calculate the Cook's distances. Do any data points seem problematic? Rerun the analysis of part (c) without these data points. Does this new analysis seem consistent with the fact that sample variances have gamma distributions with $\alpha = 0.5$?

Solution: The required R commands are:

```
> CooksD <- 0
> for(i in 1:16) {
+ tmp <- glm(s2[-i]~preds[-i, ], family=Gamma(link=log), maxit=60)
+ CooksD[i] <- t(tmp$coef-s2.glm$coef)%*%
+ solve(summary(s2.glm)$cov.unscaled)%*%(tmp$coef-s2.glm$coef)/
+   (7*summary(s2.glm)$dispersion)}
```

```
> barplot(CooksD)
```



So, it appears that the seventh and ninth variances are somewhat influential. Refitting the model without this data point yields:

```
> s2.glm1 <- glm(s2[-c(7, 9)]~preds[-c(7, 9), ],family=Gamma(link=log))
> anova(s2.glm1, test="Chisq")

Analysis of Deviance Table

Model: Gamma, link: log

Response: s2[-c(7, 9)]

Terms added sequentially (first to last)


                 Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                13    25.2700
preds[-c(7, 9), ]  6   20.598         7     4.6721 3.249e-07 ***
```

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

Thus, with the removal of these outliers, there now appears to be strong evidence of heteroscedasticity. The dispersion parameter is still not in agreement with the theoretical value of 2, however.

(f) Use the predicted values for the "within group" variances from the model you fit in part (e) to perform an appropriate weighted ANOVA for the spring heights, thereby accounting for any potential heteroscedasticity. Does this change your results for the appropriate settings you selected in part (a)?

Solution: The necessary *R* commands are:

```
> lgfts <- cbind(1, ftmp, htme, hdtme, qot, ftmp*qot, htme*qot)%*%s2.glm1$coef
> wgts <- as.vector(1/exp(lgfts))
> spg.aov4 <- lm(hght~(ftmp+htme)*qot+hdtme, weights=wgts)
> inches <- settings%*%spg.aov4$coef
> bestset <- cbind(settings[, -1], inches)
> bestset <- bestset[rev(order(inches)), ]
> bestset[1:3, ]

     ftmpset htmeset qotset hdtmeset int1set int2set
[1,]       1       0      0        1       0       0 8.017709
[2,]       1       0      0        0       0       0 7.933179
[3,]       0       0      0        1       0       0 7.894122
```

So, the best settings are indeed the same, except that we have a clear best setting which is closer to 8 inches.

2. The data file `Eyes.txt` is located on Wattle and contains a contingency table tabulating the eye test information on 7477 female employees aged between 30 and 39 years at the Royal Ordinance Factories. The data tabulates the visual acuity of each of the eyes of the subjects, rated on a four point scale: High, Moderate, Low and Poor. The columns are associated with the visual acuity of the left eye while rows are associated with the right eye.

(a) Test whether there is independence/homogeneity in this contingency table. Briefly comment on the result of your test and the structure of the observed counts.

Solution: The necessary *R* commands are:

```
> eyes <- read.table("Eyes.txt",header=T)
> attach(eyes)
> eyes

     high  mod  low poor
high 1520  266  124   66
```

```
mod   234 1512  432   78
low   117  362 1772  205
poor   36   82  179  492
```

```
> rtot <- apply(eyes, 1, sum)
> ctot <- apply(eyes, 2, sum)
> eij <- rtot%*%t(ctot)/sum(rtot)
> pres <- (eyes-eij)/sqrt(eij)
> c(sum(pres^2), 1-pchisq(sum(pres^2), (4-1)*(4-1)))
```

```
[1] 8096.877    0.000
```

So, the visual acuities are clearly not independent. Indeed, a quick glance at the table would seem to indicate that most women's eyes both have the same visual acuity level.

(b) Another hypothesis of interest might be whether the table is "symmetric". In other words, if a woman has eyes which are of two different visual acuities, is the better one equally likely to be the right as the left? To answer this question, we can use a Pearson chi-squared test. The only difficulty is that we need to prescribe the appropriate E_{ij} values under the hypothesis of symmetry and determine the appropriate number of degrees of freedom. Now, the actual hypothesis of symmetry is $H_0 : \pi_{ij} = \pi_{ji}$. For this hypothesis, it can be shown that:

$$E_{ij} = \frac{Y_{ij} + Y_{ji}}{2}.$$

In addition, the appropriate degrees of freedom is, as usual, the number of parameters which have been removed from the saturated model in the designation of the model under study. Using these facts, do you think that the table is symmetric? If not, where is the asymmetry primarily located in the table?

Solution: The necessary R commands are:

```
> eij1 <- (eyes+t(eyes))/2
> pres1 <- (eyes-eij1)/sqrt(eij1)
> df <- (4*4)-(4+3+2+1)
> c(sum(pres1^2), 1-pchisq(sum(pres1^2), df))
```

```
[1] 19.10655022  0.00398742
```

```
> pres1
```

```
           high        mod        low       poor
high  0.0000000  1.0119289  0.3188413  2.1004201
mod  -1.0119289  0.0000000  1.7565996 -0.2236068
low  -0.3188413 -1.7565996  0.0000000  0.9381942
poor -2.1004201  0.2236068 -0.9381942  0.0000000
```

So, it appears that the assumption is not justified. From the table it appears the left eye has poorer sight that the right eye. Note the symmetric model has 10 effective parameters.

(c) A slightly more flexible model to fit to these data is:

$$H_0 : \pi_{ij} = \theta\pi_{ji} \quad i < j,$$

for some unknown parameter θ. This model implies that if a woman's eyes are of different visual acuities then the odds that it is the right eye which is better is equal to θ. Under this model, it can be shown that appropriate fitted values are:

$$E_{ij} = \begin{cases} \frac{Y_{ij}+Y_{ji}}{1+\widehat{\theta}} & i > j; \\ Y_{ii} & i = j; \\ \frac{\widehat{\theta}(Y_{ij}+Y_{ji})}{1+\widehat{\theta}} & i < j. \end{cases}$$

where $\widehat{\theta}$ is the MLE of θ. Recall the definition of what θ actually measures and construct a common sense estimate $\widehat{\theta}$. Use your estimated value of θ to test whether the hypothesis H_0 fits the data adequately. Suppose we were asked to construct a 95% confidence interval for θ. Would you expect the value $\theta = 1$ to be in the interval?

Solution: The necessary *R* commands are:

Since θ measures the odds of the right eye being the better one among those women who have eyes of different visual acuities, the most obvious estimate of θ is just to take the ratio of those women whose right eye is better to those women whose left eye is better (and indeed, this turns out to be the MLE):

$$\hat{\theta} = \frac{\sum_{i<j} Y_{ij}}{\sum_{i>j} Y_{ij}} = \frac{266 + 124 + 66 + 432 + 78 + 205}{234 + 117 + 362 + 36 + 82 + 179} = 1.159406$$

So, to test the hypothesis H_0:

```
> thtahat <- 1.159406
> eij2 <- (eyes+t(eyes))/2
> rowi <- matrix(rep(1:4,4), ncol=4)
> colj <- matrix(rep(1:4,4), ncol=4, byrow=TRUE)
> eij2 <- (rowi!=colj)*1*2*eij2/(1+thtahat)+(rowi==colj)*1*eij2
> eij2 <- (rowi<colj)*1*thtahat*eij2+(rowi>=colj)*1*eij2
> pres2 <- (eyes-eij2)/sqrt(eij2)
> c(sum(pres2^2), 1-pchisq(sum(pres2^2), df-1))

[1] 7.2612035 0.2019277
```

So, it appears that the model described by H_0 is a reasonable description of the observed data. Note that the appropriate degrees of freedom here is just one less than the degrees of freedom in part b, since the model under study here has exactly one additional parameter. Since the value $\theta = 1$ implies exact symmetry, and we rejected this hypothesis in part b, it stands to reason that any reasonable confidence interval for θ should not be expected to contain the value 1.

3. The following table contains information regarding the one-year survival rates of heart transplant patients. Thirty-nine patients who received heart transplants during the period 1968 to 1971 were tracked to determine whether they survived for at least one year after their operation:

	Year of transplant:			
Survived 1 year	1968	1969	1970	1971
Yes	2	6	2	6
No	7	5	4	7

Use a contingency table analysis (treating both categorical variables as nominal) to test whether the two variables (i.e. one-year survival and transplant year) are independent of one another. Do you think such an analysis is very reliable in this instance?

Solution: The necessary *R* commands are:

```
> hrtran <- read.table("hrtran.txt", header=TRUE)
> attach(hrtran)
> hrtran

    X1968 X1969 X1970 X1971
yes     2     6     2     6
no      7     5     4     7

> rtot <- apply(hrtran, 1, sum)
> ctot <- apply(hrtran, 2, sum)
> eij <- rtot%*%t(ctot)/sum(rtot)
> pres <- (hrtran-eij)/sqrt(eij)
> c(sum(pres^2), 1-pchisq(sum(pres^2), 3))

[1] 2.4342885 0.4872837
```

The p-value is greater than 0.05 so we fail to reject the null hypothesis and conclude that the two variables are independent of one another. To see if the analysis is reliable, we must look at the E_{ij}'s.

```
> eij
        X1968    X1969    X1970    X1971
[1,] 3.692308 4.512821 2.461538 5.333333
[2,] 5.307692 6.487179 3.538462 7.666667
```

We note that half of them are below 5. This means our analysis is not that reliable as the distribution of the Pearson chi-squared statistic is only approximately chi-squared. The majority of the E_{ij} values (about 80%) must be greater than 5 to make the approximation truly reliable.