MSc: Statistical Analysis (Spring Term)

Generalized Linear Models: Exercises 3 Solutions

As usual, appropriate R code for fitting the relevant models is shown below. This has been slightly edited.

1. (a) The test for independence, effectively tests that the model without the interaction term between SITE and TYPE provides an appropriate fit to the data.

```
> count <- c(22, 2, 10, 16, 54, 115, 19, 33, 73, 11, 17, 28)
> sum(count)
[1] 400
> t <- c(rep("F", 3), rep("S", 3), rep("N", 3), rep("I", 3))
> s <- rep(c("H", "T", "E"), 4)
> type <- factor(t)
> site <- factor(s)

> skin <- data.frame(type, site, count)
> skin
   type site count
1     F    H    22
2     F    T     2
3     F    E    10
4     S    H    16
5     S    T    54
6     S    E   115
7     N    H    19
8     N    T    33
9     N    E    73
10    I    H    11
11    I    T    17
12    I    E    28

> skin.glm <- glm(count ~ type + site, family = poisson)
> summary(skin.glm)

Call:
glm(formula = count ~ type + site, family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0453  -1.0741   0.1297   0.5857   5.1354

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.9554     0.1770  16.696  < 2e-16 ***
typeI         0.4990     0.2174   2.295   0.0217 *
typeN         1.3020     0.1934   6.731 1.68e-11 ***
typeS         1.6940     0.1866   9.079  < 2e-16 ***
siteH        -1.2010     0.1383  -8.683  < 2e-16 ***
siteT        -0.7571     0.1177  -6.431 1.27e-10 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 295.203  on 11  degrees of freedom
Residual deviance:  51.795  on  6  degrees of freedom
AIC: 122.91
```

```
Number of Fisher Scoring iterations: 5
```

The residual deviance of 51.795 on 6 d.f. is clearly significant, so that we reject the hypothesis of independence between tumour type and site. The model parameters, and fitted values can be calculated as follows (but these are not of any real interest since the model does not fit).

```
> dummy.coef(skin.glm)
Full coefficients are

(Intercept):      2.955431
type:                    F          I          N         S
                 0.0000000  0.4989912  1.3019532 1.6939953
site:                    E          H          T
                 0.0000000 -1.2010273 -0.7570959

> fitted(skin.glm)
      1       2       3       4       5       6       7       8       9      10
  5.780   9.010  19.210  31.450  49.025 104.525  21.250  33.125  70.625   9.520
     11      12
 14.840  31.640

 > sum(fitted(skin.glm))
[1] 400
```

(b) The fitted values for the two-way table (under independence) are

$$\frac{\text{(row total)} * \text{(column total)}}{\text{overall total}}$$

which match the expected values (E_i) used in the chi-squared test.

The statistic $\sum(O_i - E_i)^2/E_i$ can be constructed 'by hand'...

```
> Obs <- count
> Exp <- fitted(skin.glm)
> X2 <- sum((Obs - Exp)^2/Exp)
> X2
65.81293
```

This should be compared with a Chi-squared distribution on (J-1)(K-1)=6 d.f., clearly resulting in the same conclusion.

Note: It can be shown that, provided we use the log link and include a constant term in the model, the residual deviance from the *Poisson* model is $2\sum O_i \log(O_i/E_i)$, and that the two statistics are asymptotically equal.

The chi-squared test can be carried out in R using `chisquare.test`, in the following way

```
> mat <- tapply(count, list(type, site), mean)
> mat
    E  H  T
F  10 22  2
I  28 11 17
N  73 19 33
S 115 16 54

> sum(mat)
[1] 400

> chisq.test(mat)
```

```
        Pearson's Chi-squared test

data:  mat
X-squared = 65.813, df = 6, p-value = 2.943e-12
```

(Note that R has permuted the factor levels in the table of counts to be in alphabetical order).

To try and understand the relationship between site and tumour type it is often useful to take the independence model (which does not fit) and examine the deviance residuals. Consider the following output.

```
> restabskin <- tapply(resid(skin.glm), list(type, site), mean)
> print(restabskin, digits = 2)

      E     H      T
F -2.32  5.14 -2.828
I -0.66  0.47  0.548
N  0.28 -0.50 -0.022
S  1.01 -3.05  0.699
```

From this we observe that for tumour types 'Superficial spreading melanoma' and 'Nodular', the hypothesis of independence gives a good fit. There are 3 large residuals. For 'Hutchinson's melantonic freckle' there are obviously far more tumours sited on the head and neck than would be expected using the independence model, with correspondingly far fewer on the trunk and extremities. So this type of tumour occurs significantly more on the head and neck. For 'Intermediate' skin cancer there are obviously far fewer tumours occurring on the head and neck than would be expected using the independence model.

2. (a) Since the margin corresponding to TREATMENT is fixed by the design, we treat TREATMENT as an explanatory variable, and can only test for *homogeneity* of the RESPONSE level between the placebo and vaccine groups.

(b)
```
> count <- c(25, 8, 5, 6, 18, 11)
> r <- rep(c("S", "M", "L"), 2)
> t <- c(rep("P", 3), rep("V", 3))
> sum(count)
[1] 73
> resp <- factor(r)
> treat <- factor(t)
> trial <- data.frame(resp, treat, count)
> trial
  resp treat count
1    S     P    25
2    M     P     8
3    L     P     5
4    S     V     6
5    M     V    18
6    L     V    11

> rm(count, r, t, resp, treat)
> trial.glm <- glm(count ~ resp + treat, family = poisson, data = trial)
> summary(trial.glm)

Call:
glm(formula = count ~ resp + treat, family = poisson, data = trial)
```

```
Deviance Residuals:
     1       2       3       4       5       6
 2.040  -1.630  -1.247  -2.615   1.469   1.128

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.11972    0.27408   7.734 1.04e-14 ***
respM        0.48551    0.31774   1.528   0.1265
respS        0.66140    0.30783   2.149   0.0317 *
treatV      -0.08224    0.23428  -0.351   0.7256
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 23.807  on 5  degrees of freedom
Residual deviance: 18.643  on 2  degrees of freedom
AIC: 51.771

Number of Fisher Scoring iterations: 5
```

The residual deviance of 18.643 is very significant on the χ^2_2 distribution. Thus we reject the hypothesis of homogeneity between the vaccine and placebo groups.

Again, we can look at the estimated parameters from the model and the fitted responses, noting that the fitted responses in each row of the table are constrained to match the row sums, as we require. Otherwise the model parameters are not particularly useful - since the model does not fit.

```
> dummy.coef(trial.glm)
Full coefficients are

(Intercept):      2.119715
resp:                    L          M         S
                 0.0000000  0.4855078 0.6613985
treat:                   P          V
                 0.0000000 -0.0822381

> fitted(trial.glm)
       1        2        3        4        5        6
16.136986 13.534247  8.328767 14.863014 12.465753  7.671233
> sum(fitted(trial.glm)[1:3])
[1] 38
> sum(fitted(trial.glm)[4:6])
[1] 35
```

Again we look at the residuals.

```
> attach(trial)
> restabtrial <- tapply(resid(trial.glm), list(treat, resp), mean)
> print(restabtrial, digits = 2)

     L    M    S
P -1.2 -1.6  2.0
V  1.1  1.5 -2.6
```

From this it is clear that, under the hypothesis of homogeneity, the vaccine sample contains significantly fewer individuals with a small antibody response than would be expected, and the placebo sample contains significantly more individuals with a small antibody response. This suggests that the vaccine is effective in that it affects the antibody response in a positive way.

3. This example was discussed in lectures.

   (a) The data can be input and the suggested model fitted using the following code.

```
> count <- c(176, 3, 293, 4, 197, 17, 23, 2)
> cl <- c(rep("A", 4), rep("B", 4))
> ca <- rep(c("L", "L", "H", "H"), 2)
> su <- rep(c("S", "D"), 4)
> clinic <- factor(cl)
> care <- factor(ca)
> surv <- factor(su)
> health <- data.frame(clinic, care, surv, count)
> health
  clinic care surv count
1      A    L    S   176
2      A    L    D     3
3      A    H    S   293
4      A    H    D     4
5      B    L    S   197
6      B    L    D    17
7      B    H    S    23
8      B    H    D     2

> health.glm <- glm(count ~ clinic + care + surv + clinic:care + clinic:surv,
    family = poisson, data = health)
> summary(health.glm)

Call:
glm(formula = count ~ clinic + care + surv + clinic:care + clinic:surv,
    family = poisson, data = health)

Deviance Residuals:
        1          2          3          4          5          6          7
-0.027693   0.221610   0.021487  -0.178476   0.000894  -0.003044  -0.002617
        8
 0.008894

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.47422    0.37964   3.883 0.000103 ***
clinicB       -0.78737    0.48226  -1.633 0.102535
careL         -0.50635    0.09462  -5.351 8.74e-08 ***
survS          4.20469    0.38077  11.042  < 2e-16 ***
clinicB:careL  2.65345    0.23157  11.458  < 2e-16 ***
clinicB:survS -1.75550    0.44963  -3.904 9.45e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1.0664e+03  on 7  degrees of freedom
Residual deviance: 8.2289e-02  on 2  degrees of freedom
AIC: 52.265

Number of Fisher Scoring iterations: 4
```

   The residual deviance is 0.0823, which is not significant when compared with a χ^2_2 distribution, so that the suggested model certainly fits the data. We should check that there is not a 'better' fitting (more parsimonious) model with fewer terms. Try dropping each of the remaining two factor interactions in turn.

```
> anova(health.glm, update(health.glm,  ~ . - clinic:care), test = "Chisq")
Analysis of Deviance Table

Model 1: count ~ clinic + care + surv + clinic:care + clinic:surv
Model 2: count ~ clinic + care + surv + clinic:surv
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1         2      0.082
2         3    193.736 -1  -193.65 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> anova(health.glm, update(health.glm,  ~ . - clinic:surv), test = "Chisq")
Analysis of Deviance Table

Model 1: count ~ clinic + care + surv + clinic:care + clinic:surv
Model 2: count ~ clinic + care + surv + clinic:care
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1         2     0.0823
2         3    17.8284 -1  -17.746 2.524e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The difference in deviance when CLINIC:CARE is removed from the fitted model is $193.74 - 0.0823 = 193.6677$, which is obviously significant against a chi-squared distribution on $3 - 2 = 1$ d.f. Similarly, there is strong evidence against removing the CLINIC:SURV interaction.

The estimated parameters and fitted responses from the chosen model are as follows.

```
> dummy.coef(health.glm)
Full coefficients are

(Intercept):      1.474224
clinic:                 A          B
                0.0000000 -0.7873732
care:                   H          L
                0.0000000 -0.5063463
surv:                   D          S
                 0.000000   4.204693
clinic:care:          A:H        B:H       A:L       B:L
                 0.000000   0.000000  0.000000  2.653447
clinic:surv:          A:D        B:D       A:S       B:S
                 0.000000   0.000000  0.000000 -1.755504

> fitted(health.glm)
         1          2          3          4          5          6          7
176.367647   2.632353 292.632353   4.367647 196.987448  17.012552  23.012552
         8
  1.987448
```