MATH 523 - ASSIGNMENT 2: SOLUTIONS

1 (a) The identifiability constraints are identical to those applied in any two factor main effects model, that is, we need to set one parameter for each factor to zero, or to have a sum-to-zero constraint. The most straightforward constraint sets $\alpha_1 = \beta_1 = 0$; on the μ_{jk} scale, we could also require that

$$\prod_{j=1}^{J} \lambda_j^{(1)} = \prod_{k=1}^{K} \lambda_k^{(2)} = 1$$

which is the equivalent to a pair of sum-to-zero constraints.

2 MARKS

(b) Following the above notation, we may write for model 1

$$\mu_{jkl} = \lambda_{jk}^{(12)} \lambda_l^{(3)}$$

which implies that factors 1 and 2 **jointly** affect the response, **independently** of factor 3. For model 2, factors 2 and 3 jointly affect the response, independently of factor 1. For the numbers of parameters

$$\begin{aligned} \text{Model 1} &: 1 + (J-1) + (K-1) + (L-1) + (J-1)(K-1) \\ \text{Model 2} &: 1 + (J-1) + (K-1) + (L-1) + (K-1)(L-1) \end{aligned}$$

6 MARKS

(c) Using a backward selection strategy, starting from the saturated model, the first test assesses whether the three way interaction `X1:X2:X3` can be omitted: for this test we have

$$\begin{aligned} \triangle \text{ Deviance} &: 0.61327 \\ \triangle \text{ Rdf} &: 1 \end{aligned}$$

so the comparison of the change in deviance against a χ_1^2 distribution leads to the conclusion of a non-significant result (p-value is 0.4336). Hence the three-way interaction can be omitted.

Next, we attempt to drop one of the two-way interactions; using the `drop1` command, we obtain the following results

```
Single term deletions

Model:
Y ~ X1 + X2 + X3 + X1:X2 + X1:X3 + X2:X3
       Df Deviance    AIC    LRT Pr(Chi)
<none>      0.6133 68.062
X1:X2   1   4.5627 70.011 3.9494 0.04689 *
X1:X3   1   2.9799 68.428 2.3666 0.12395
X2:X3   1   2.0626 67.511 1.4493 0.22864
```

From this, we conclude that perhaps the terms `X1:X3` and `X2:X3` can be omitted from the model with all three two-way interactions: for this test we use `anova` to obtain

```
Analysis of Deviance Table

Model 1: Y ~ X1 + X2 + X3 + X1:X2
Model 2: Y ~ X1 + X2 + X3 + X1:X2 + X1:X3 + X2:X3
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         3     4.6014
2         1     0.6133  2   3.9881    0.1361
```

so the comparison of the change in deviance 3.9881 against a χ^2_2 distribution leads to the conclusion of a non-significant result (p-value is 0.1361).

After dropping these term, we now attempt to drop one of the remaining terms:

```
Single term deletions

Model:
Y ~ X1 + X2 + X3 + X1:X2
       Df Deviance     AIC     LRT Pr(Chi)
<none>        4.60   68.05
X3      1  1651.83 1713.28 1647.23 < 2e-16 ***
X1:X2   1     8.72   70.17    4.12 0.04234 *
```

It appears that neither term can be removed, so we are left with the final model

$$1 + X_1 + X_2 + X_3 + X_1.X_2$$

The final output for this model is as follows:

```
Call:
glm(formula = Y ~ X1 + X2 + X3 + X1:X2, family=poisson, data=Stress)

Deviance Residuals:
      1       2       3       4       5       6       7       8
-0.2481 -0.2686  0.4786 -0.2241  0.7714  0.8223 -1.5631 0.6885

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.04978    0.09497  32.112   <2e-16 ***
X1B          -0.10920    0.09549  -1.144   0.2528
X2Normal      1.23160    0.07462  16.505   <2e-16 ***
X3Normal      2.30155    0.07550  30.485   <2e-16 ***
X1B:X2Normal  0.21757    0.10726   2.028   0.0425 *
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2432.6119  on 7  degrees of freedom
Residual deviance:    4.6014  on 3  degrees of freedom
```

To assess model adequacy, we inspect the dispersion estimate

$$\widehat{\phi} = \frac{\text{Residual Deviance}}{\text{Residual df}} = \frac{4.6014}{3} = 1.5338$$

which is perhaps a little large, but represents no major cause for concern; the Poisson model seems adequate.

It seems that (X_1, X_2) act jointly, and independently from X_3, to influence response.

6 MARKS

2 Let `Thrombo`, `Smoke` and `Med` denote the three factors. Starting from the saturated model, the first test assesses whether the three way interaction `Thrombo:Smoke:Med` can be omitted: for this test we have

$$\begin{aligned} \triangle \text{ Deviance} &: 2.3533 \\ \triangle \text{ Rdf} &: 1 \end{aligned}$$

so the comparison of the change in deviance against a χ^2_1 distribution leads to the conclusion of a non-significant result (p-value is 0.1250). Hence the three-way interaction can be omitted.

Next, we attempt to drop one of the two-way interactions; using the `drop1` command, we obtain the following results

```
Single term deletions
Model:
Y ~ Thromb + Smoke + Med + Thromb:Smoke + Thromb:Med + Smoke:Med
             Df Deviance    AIC     LRT   Pr(Chi)
<none>            2.3533 51.809
Thromb:Smoke  1   3.5820 51.038  1.2287   0.26766
Thromb:Med    1  28.3732 75.829 26.0199 3.379e-07 ***
Smoke:Med     1   6.4265 53.882  4.0732   0.04357 *
```

From this, we conclude that the term `Thrombo:Smoke` can be omitted from the model with all three two-way interactions. For the other two two-way interactions, we have

```
Single term deletions
Model:
Y ~ Thromb + Smoke + Med + Thromb:Med + Smoke:Med
          Df Deviance    AIC     LRT  Pr(Chi)
<none>           3.582 51.038
Thromb:Med  1   33.082 78.538 29.5005 5.59e-08 ***
Smoke:Med   1   11.136 56.591  7.5538 0.005988 **
```

that is, two strongly significant results. Hence the final model is

```
1 + Thromb + Smoke + Med + Thromb.Med + Smoke.Med
```

for which the model summary is

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)       4.4275     0.1066  41.537  < 2e-16 ***
ThrombYes        -1.1977     0.2017  -5.938 2.89e-09 ***
SmokeYes         -1.3241     0.2089  -6.337 2.34e-10 ***
MedYes           -2.7127     0.3655  -7.422 1.15e-13 ***
ThrombYes:MedYes  2.1532     0.4232   5.087 3.63e-07 ***
SmokeYes:MedYes   1.1009     0.3952   2.786  0.00534 **
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 166.439  on 7  degrees of freedom
Residual deviance:   3.582  on 2  degrees of freedom
```

which, despite some evidence of over-dispersion, seems an adequate fit, as $\widehat{\phi} = 3.582/2 = 1.791$. In the analysis listing, the baseline group is the control patients who do not smoke and do not take medication.

- The results indicate that all three factors influence the response; for the `Thromb` factor, this is somewhat expected, due to the case-control design.
- Also, in this model, consider the expected responses for the two levels of factor `Med`: at each level, the other two factors combine additively to influence response. This implies a conditional independence structure; conditional on `Med`, `Thromb` and `Smoke` act independently.
- For each of the main effects, the effect of moving from `No` to `Yes` **decreases** the rate compared to baseline; however, the two two-way interactions both act to **increase** the response significantly. That is, for example, the rate of thromboembolisms **increases significantly** in those taking the medication.
- Define the notation μ_{jkl} for $j, k, l = 0, 1$, where j, k or l equal to 0 corresponds to `No` etc. In the original table, this yields the mean structure There are six coefficients that are estimated;

		Smoker		Non-smoker	
		Medication		Medication	
		Y	N	Y	N
Thromboembolism	Y	μ_{111}	μ_{110}	μ_{101}	μ_{100}
	N	μ_{011}	μ_{010}	μ_{001}	μ_{000}

they combine as follows to explain $\log \mu_{jkl}$:

$\log \mu_{000} = \beta_1$ =4.4275

$\log \mu_{100} = \beta_1 + \beta_2^{(T)}$ =4.4275+(-1.1977)

$\log \mu_{010} = \beta_1 + \beta_2^{(S)}$ =4.4275+(-1.3241)

$\log \mu_{110} = \beta_1 + \beta_2^{(T)} + \beta_2^{(S)}$ =4.4275+(-1.1977)+(-1.3241)

$\log \mu_{001} = \beta_1 + \beta_2^{(M)}$ =4.4275+(-2.7127)

$\log \mu_{101} = \beta_1 + \beta_2^{(T)} + \beta_2^{(M)} + \gamma_{22}^{(TM)}$ =4.4275+(-1.1977)+(-2.7127)+2.1532

$\log \mu_{011} = \beta_1 + \beta_2^{(S)} + \beta_2^{(M)} + \gamma_{22}^{(SM)}$ =4.4275+(-1.3241)+(-2.7127)+1.1009

$\log \mu_{111} = \beta_1 + \beta_2^{(T)} + \beta_2^{(S)} + \beta_2^{(M)} + \gamma_{22}^{(TM)} + \gamma_{22}^{(SM)}$ =4.4275+(-1.1977)+(-1.3241)+(-2.7127)+2.1532+1.1009

- A comparison of interest is the relative rate of thromboembolisms in cases and controls in the four groups defined by the other two factors: Then, from the tables,

$$\log\left(\frac{\mu_{111}}{\mu_{011}}\right) = \beta_2^{(T)} + \gamma_{22}^{(TM)} \qquad \log\left(\frac{\mu_{100}}{\mu_{000}}\right) = \beta_2^{(T)}$$

and hence

$$\log\left(\frac{\mu_{111}}{\mu_{011}}\right) - \log\left(\frac{\mu_{100}}{\mu_{000}}\right) = \gamma_{22}^{(TM)}.$$

Similarly

$$\log\left(\frac{\mu_{101}}{\mu_{001}}\right) - \log\left(\frac{\mu_{100}}{\mu_{000}}\right) = \gamma_{22}^{(TM)}.$$

Hence

$$\left(\frac{\mu_{111}}{\mu_{011}}\right) = \left(\frac{\mu_{101}}{\mu_{001}}\right)$$

Similar calculations follow for $\gamma_{22}^{(SM)}$.

6 MARKS