MAS3906 - Generalized Linear Models

Project - Statistical Modelling

a project authored by:

Stephen Henry Cole and Calum Doran

Supervised by David Walshaw

Academic year 2019/20

Modelling hypertension (Question 1)

This model is based on a historical study which had the innovative idea of investigating whether snoring had an effect on the presence of hypertension (high blood pressure). The data we have is downloaded from blackboard and saved as **snoring.txt** which can then be directed into R for statistical modelling.

We have four explanatory variables, 'x', which are sex, smoking, obesity and snoring which explain the response variable hypertension, 'y'.

All variables included

Fitting this into a generalised linear model using the **glm** function, we get the following summary table:

```
Call:
glm(formula = cbind(hyper, total - hyper) ~ factor(sex) + smoking +
    obesity + snoring, family = binomial, data = snoring)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4569  -0.4381  -0.0414   0.3405   2.9874

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.34774    0.25053  -9.371  < 2e-16 ***
factor(sex)2  0.09573    0.19450   0.492  0.62260
smoking       0.01459    0.21854   0.067  0.94679
obesity       0.82539    0.22033   3.746  0.00018 ***
snoring       0.76947    0.24217   3.177  0.00149 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 44.123  on 14  degrees of freedom
Residual deviance: 14.987  on 10  degrees of freedom
AIC: 74.258

Number of Fisher Scoring iterations: 4
```

Figure 1: Summary table for all variables.

We can see that obesity and snoring are the only significant variables. Notice that the variables that influences hypertension the most are our variables obesity and snoring. We shall place it in order of significance and evaluate its ANOVA table to see the significance of each variable. We observe the following:

```
Terms added sequentially (first to last)


             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                           14     44.123
obesity       1  17.7480        13     26.375 2.522e-05 ***
snoring       1  11.1458        12     15.230 0.0008422 ***
factor(sex)   1   0.2380        11     14.992 0.6256598
smoking       1   0.0044        10     14.987 0.9468257
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Figure 2: ANOVA table for all ordered variables.

We can confirm that there is no evidence to suggest that sex and smoking has any effect on hypertension. But, there is overwhelming evidence that suggests that obesity and snoring does

have an effect on hypertension as they are both significant at the 0.1% level. We can check to see if there are any interaction terms which may help to explain the response variable a bit better and determine if sex and smoking are not needed in the model.

Interaction terms

When checking the interaction terms individually, we only observe one significant interaction term and that is **snoring**×**smoking**.

```
Call:
glm(formula = cbind(hyper, total - hyper) ~ obesity + snoring +
    factor(sex) + smoking + snoring * smoking, family = binomial,
    data = snoring)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.38515  -0.47508   0.08357   0.34796   1.29740

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -2.6225     0.2851  -9.200  < 2e-16 ***
obesity          0.8461     0.2217   3.817 0.000135 ***
snoring          1.0959     0.2818   3.889 0.000101 ***
factor(sex)2     0.1253     0.1947   0.643 0.519972
smoking          1.2911     0.4762   2.712 0.006695 **
snoring:smoking -1.5320     0.5329  -2.875 0.004039 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 44.1233  on 14  degrees of freedom
Residual deviance:  7.4751  on  9  degrees of freedom
AIC: 68.746

Number of Fisher Scoring iterations: 4
```

Figure 3: Summary table for reduced number of variables.

We can see that the Residual deviance has decreased significantly to 7.4751 and the Df value has decreased by 1. That is a reduction by 51.123% in residual deviance. It is also clear to see that the coefficients for each variable has changed. In comparison to the previous summary table, the variables that make the most impact on hypertension are **smoking** and its interaction term with **snoring**.

Again we shall look at its ANOVA table to see the significance of each variable:

```
               Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                              14     44.123
obesity         1  17.7480        13     26.375 2.522e-05 ***
snoring         1  11.1458        12     15.230 0.0008422 ***
factor(sex)     1   0.2380        11     14.992 0.6256598
smoking         1   0.0044        10     14.987 0.9468257
snoring:smoking 1   7.5120         9      7.475 0.0061290 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Figure 4: ANOVA table for reduced number of variables.

We can see that smoking is still not significant. However, the interaction term is significant at the 1% level and so we have strong evidence to suggest that **snoring**×**smoking** is a significant term in explaining the hypertension. Sex is still not significant so we can remove this variable from the model.

Final model

As previously mentioned, we remove sex from the model as it does not have any significance on its own, or with any interactions. We therefore have a model which includes the other explanatory variables and the interactive term. Observe the following summary table:

```
Call:
glm(formula = cbind(hyper, total - hyper) ~ obesity + snoring +
    smoking + snoring * smoking, family = binomial, data = snoring)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.41221  -0.42654   0.02932   0.36142   1.41725

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)     -2.5338     0.2483 -10.206  < 2e-16 ***
obesity          0.8373     0.2210   3.788 0.000152 ***
snoring          1.0614     0.2765   3.838 0.000124 ***
smoking          1.2601     0.4738   2.660 0.007819 **
snoring:smoking -1.5109     0.5318  -2.841 0.004498 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 44.1233  on 14  degrees of freedom
Residual deviance:  7.8884  on 10  degrees of freedom
AIC: 67.16

Number of Fisher Scoring iterations: 4
```

Figure 5: Summary table for our final model.

Clearly, all terms are now significant and that most of the effect on hypertension is caused by **snoring, smoking** and their interaction term. Our Residual deviance has slightly increased, but not by any noticeable difference, so it is not an issue.

```
                Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                14     44.123
obesity          1  17.7480        13     26.375 2.522e-05 ***
snoring          1  11.1458        12     15.230 0.0008422 ***
smoking          1   0.0005        11     15.229 0.9827446
snoring:smoking  1   7.3406        10      7.888 0.0067413 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Figure 6: ANOVA table for our final model.

We can see that smoking is not significant on its own, but it is significant when interacting with snoring. Therefore, we need to follow the hierarchical structure and keep it within our model.

Our fitted model is therefore:

$$\begin{aligned} hypertension = &-2.5338 + 0.8373 * obesity + 1.0614 * snoring + 1.2601 * smoking \\ &- 1.5109 * snoring * smoking \end{aligned}$$

If the individual has **obesity** or **smokes** or **snores**, they will have increased levels of hypertension. However, smoking and snoring will also decrease the level of hypertension due to the interaction term being negatively significant.

Modelling Ulcers (Question 2)

Here we have a retrospective study of peptic ulcers and blood groups. The particular data set from the study in which we are analysing has three factors, *blood group, place* and *case/control*. We have two blood groups we are considering from Manchester and Newcastle. For each cell we have both number of cases and a control as well.

		Blood Group		
		A	*B*	*Total*
Manchester	*Cases*	246	361	607
	Control	3775	4532	8307
Newcastle	*Cases*	291	396	687
	Control	5261	6598	11859

We would like to investigate whether there is an association between ulcers and blood group.

The Model

To find out if there is any association between ulcers and blood group, we will first create the minimal model for the study. This model contains only marginal effects and any interactions built in to the study design. The response group for the study is blood group and therefore we will add the totals for the cross-classification of case/control and place.

Let y_{jkl} denote the count for cell (j, k, l) where $j = 1, 2$ indexes case/control, $k = 1, 2$ indexes blood group, and $l = 1, 2$ indexes place.

A log linear model for this three-way contingency table is defined as,

$$Y_{jkl} \sim Po(\mu_{jkl})$$
$$log(\mu_{jkl}) = \mu + \alpha_j + \beta_k + \gamma_l + (\alpha\beta)_{jk} + (\alpha\gamma)_{jl} + (\beta\gamma)_{kl} + (\alpha\beta\gamma)_{jkl}$$

where $(\alpha\beta)_{jk}$ stands for the interaction between case/control and blood group, $(\alpha\gamma)_{jl}$ stands for the interaction between case/control and place and $(\beta\gamma)_{kl}$ stands for the interaction between blood group and place. With $(\alpha\beta\gamma)_{jkl}$ standing for the interaction between all factors.
This is the maximal model, containing all the marginal and all the interaction terms.
As mentioned before the minimal model only contains the marginal effects and the interaction terms built into the study. This is therefore,

$$log(\mu_{jkl}) = \mu + \alpha_j + \beta_k + \gamma_l + (\alpha\gamma)_{jl}$$

The constraints for these models are,

$$\alpha_j = 0, \beta_k = 0, \gamma_l = 0, (\alpha\beta)_{jk} = 0, (\alpha)_{jl} = 0,$$
$$(\beta\gamma)_{kl} = 0, (\alpha\beta\gamma)_{jkl} = 0 \text{ whenever } j = 1, k = 1 \text{ or } l = 1$$

To test for the interaction between case and blood group we will consider the model,

$$log(\mu_{jkl}) = \mu + \alpha_j + \beta_k + \gamma_l + (\alpha\gamma)_{jl} + (\alpha\beta)_{jk}$$

Checking for Interaction

Here we have the output from the R code to check for significant interaction between case and blood group using analysis of deviance.

```
           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                         7    20865.3
place       1    617.7         6    20247.6 < 2.2e-16 ***
case        1  19973.2         5      274.4 < 2.2e-16 ***
blood       1    250.0         4       24.4 < 2.2e-16 ***
place:case  1     16.2         3        8.2 5.698e-05 ***
case:blood  1      5.4         2        2.8   0.01997 *
```

We can see that the p-value is 0.01997 for testing the hypothesis,

$$H_0 : (\alpha\beta)_{jk} = 0 \text{ versus } H_1 : (\alpha\beta)_{jk} \neq 0 \text{ for at least one pair } (j, k)$$

Although not extremely significant, we can reject H_0 at the 5% level and conclude that the probability of having an ulcer is associated with blood group.

Checking for Further Interactions

We now want to check if this association varies from city to city. Here we have the output when checking the maximal model using analysis of deviance and therefore the three-way interaction.

```
                Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                           7    20865.3
place            1    617.7         6    20247.6 < 2.2e-16 ***
case             1  19973.2         5      274.4 < 2.2e-16 ***
blood            1    250.0         4       24.4 < 2.2e-16 ***
place:case       1     16.2         3        8.2 5.698e-05 ***
case:blood       1      5.4         2        2.8   0.01997 *
place:blood      1      1.7         1        1.0   0.19111
place:case:blood 1      1.0         0        0.0   0.30725
```

From the R code we see that the p-value for the hypothesis

$$H_0 : (\alpha\beta\gamma)_{jkl} = 0 \text{ versus } H_1 : (\alpha\beta\gamma)_{jkl} \neq 0 \text{ for at least one combination of } (j, k, l)$$

The p-value for the hypothesis is 0.30725 and therefore the three-way interaction is not significant and we can conclude that the association does not vary from city to city.

Nature of the Significant

We will now use the summary function in R on the maximal model to check the co-efficients.

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        5.50533    0.06376  86.348  < 2e-16 ***
place2             0.16799    0.08661   1.940   0.0524 .
case2              2.73082    0.06580  41.500  < 2e-16 ***
blood2             0.38355    0.08267   4.639  3.5e-06 ***
place2:case2       0.16393    0.08920   1.838   0.0661 .
case2:blood2      -0.20078    0.08556  -2.347   0.0189 *
place2:blood2     -0.07546    0.11312  -0.667   0.5048
place2:case2:blood2  0.11914    0.11672   1.021   0.3074
```

From the table we can see that $(\alpha\beta)_{22}$ is negatively significant at the 5% level. This tell us that there is fewer controls of blood group B than is expected. So blood group B has significantly more values than blood group A.