STATISTICAL MODELLING Part IIC

Practical 8: Contingency tables and gamma regression IAC/Lent 2011

Comments and corrections to ioana@statslab.cam.ac.uk

The first data set we will look at is `CancerData`. For 400 patients with a form of skin cancer called a malignant melanoma, the site and histological type of the tumour were recorded. Download `cancer.txt` from www.statslab.cam.ac.uk/∼ioana/statsmod.html, save it in your `Rwork` directory, and open it in R.

```
> Cancer <- read.table("cancer.txt", header = TRUE)
```

It's probably easiest to have the data in the form of a vector.

```
> y <- as.vector(as.matrix(Cancer))
```

We now need to form factors for the site and type. The first can be done with

```
> Site <- gl(3, 4, 12, names(Cancer))
> Site
```

```
 [1] Head   Head   Head   Head   Trunk  Trunk  Trunk  Trunk
 [9] Extrem Extrem Extrem Extrem
Levels: Head Trunk Extrem
```

Exercise: Create a similar factor for type, ensuring that the levels of the factor match up correctly with the data. We can fit a multinomial model for independence of type and site by means of a surrogate Poisson model:

```
> IndepMod <- glm(y ~ Type + Site, family = poisson)
> summary(IndepMod)
```

```
Call:
glm(formula = y ~ 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)   1.7544     0.2040   8.600  < 2e-16 ***
TypeSuper     1.6940     0.1866   9.079  < 2e-16 ***
TypeNodular   1.3020     0.1934   6.731 1.68e-11 ***
TypeIndet     0.4990     0.2174   2.295  0.02173 *
SiteTrunk     0.4439     0.1554   2.857  0.00427 **
SiteExtrem    1.2010     0.1383   8.683  < 2e-16 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 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
```

Exercise: Write down both the multinomial model and the surrogate Poisson model here.

Exercise: The residual deviance, 51.795, is certainly large by comparison with χ^2_6, and by comparing the data with the fitted values, confirm that this is largely because Hutchinson's melanotic freckle appears on the head more often, and less often on the neck and extremities, than would be expected if site and type were independent. (Moreover, the superficial spreading melanoma appears less often on the head than would be expected if site and type were independent.)

This can also be seen with an interaction plot. See Figure 1. What should the plot look like if site and type are independent?

One way to fit a single interaction term (perhaps not the best) is with

```
> Interaction <- c(1, rep(0, 11))
> NewMod <- glm(y ~ Type + Site + Interaction, family = poisson)
> summary(NewMod)


Call:
glm(formula = y ~ Type + Site + Interaction, family = poisson)
```

```
> par(mfrow = c(2, 1))
> interaction.plot(Type, Site, y, main = "Interaction: Type v. Site")
> interaction.plot(Site, Type, y, main = "Interaction: Site v. Type")
```

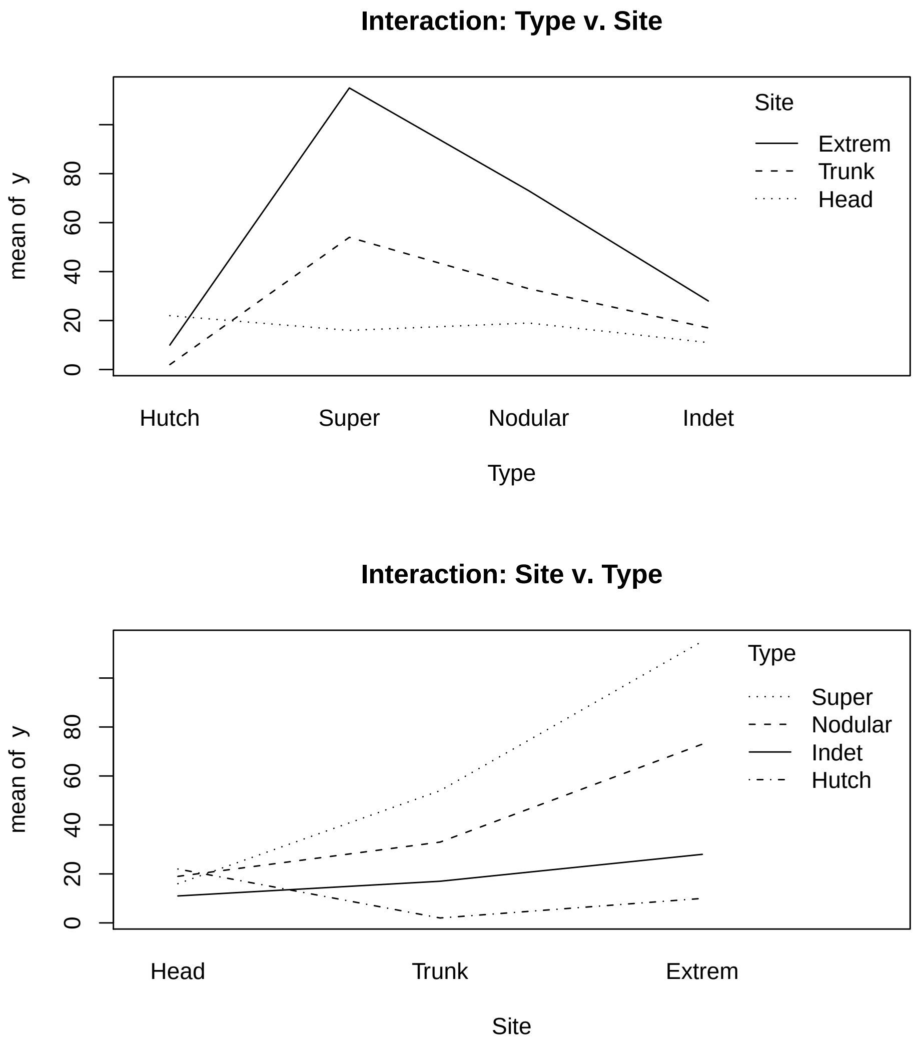

Figure 1: Interaction plots for cancer type v. site and vice–versa

```
Deviance Residuals:
         1           2           3           4           5
-4.712e-08  -1.594e+00   8.033e-01   1.379e+00  -1.031e+00
         6           7           8           9          10
```

```
 3.256e-01  -3.235e-01   3.410e-01   6.188e-01   4.630e-01
       11          12
-1.624e-01  -9.495e-01

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.5452     0.3289   1.658   0.0974 .
TypeSuper     2.6011     0.2985   8.713  < 2e-16 ***
TypeNodular   2.2091     0.3029   7.294 3.01e-13 ***
TypeIndet     1.4061     0.3187   4.412 1.03e-05 ***
SiteTrunk     0.7980     0.1769   4.511 6.44e-06 ***
SiteExtrem    1.5551     0.1621   9.593  < 2e-16 ***
Interaction   2.5458     0.3920   6.495 8.32e-11 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 295.2030  on 11  degrees of freedom
Residual deviance:   8.0021  on  5  degrees of freedom
AIC: 81.113

Number of Fisher Scoring iterations: 4
```

How is this new model coded in R? The baseline levels are `Hutch` and `Head` for factor variables `Type` and `Site`, respectively. Let $\mathbb{I}_{\texttt{Type[i]=Super}}$ be an indicator variable that equals 1 if the ith observation has `Type`[i] = `Super` and 0 otherwise. NewMod is defined as

$$\begin{aligned}\log(\mu_i) = \beta_0 &+ \beta_1\mathbb{I}_{\texttt{Type[i]=Super}} + \beta_2\mathbb{I}_{\texttt{Type[i]=Nodular}} + \beta_3\mathbb{I}_{\texttt{Type[i]=Indet}} \\ &+ \beta_4\mathbb{I}_{\texttt{Site[i]=Trunk}} + \beta_5\mathbb{I}_{\texttt{Site[i]=Extrem}} + \beta_6\mathbb{I}_{\texttt{Type[i]=Hutch\&Site[i]=Head}},\end{aligned}$$

where $Y_i \sim \text{Poisson}(\mu_i)$. Compare the data with the fitted values from the new model. Test for the significance of the interaction term. What do you conclude? What can you say about the independence and interaction models from Figure 2?

Finally, we look at `DrinksData`. A soft drink bottler is analysing vending machine service routes in his distribution system, and is interested in predicting the amount of time required by the route driver to service the vending machines in an outlet. The industrial engineer responsible for the study has suggested that the two most important variables affecting the delivery time are the number of cases of product stocked and the distance walked by the route driver. The engineer has collected the 25 observations below on

```
> interaction.plot(Type, Site, fitted(IndepMod), main = "Interaction: Type v. Site",
+     ylim = c(min(y), max(y)), ylab = "fitted values",
+     col = c("blue", "green", "red"))
> lines(Type[Site == "Extrem"], fitted(NewMod)[Site ==
+     "Extrem"], pch = 1)
> lines(Type[Site == "Trunk"], fitted(NewMod)[Site ==
+     "Trunk"], pch = 1)
> lines(Type[Site == "Head"], fitted(NewMod)[Site ==
+     "Head"], pch = 1)
> points(Type[Site == "Extrem"], y[Site == "Extrem"],
+     col = "red", pch = 19)
> points(Type[Site == "Trunk"], y[Site == "Trunk"],
+     col = "green", pch = 19)
> points(Type[Site == "Head"], y[Site == "Head"], col = "blue",
+     pch = 19)
```

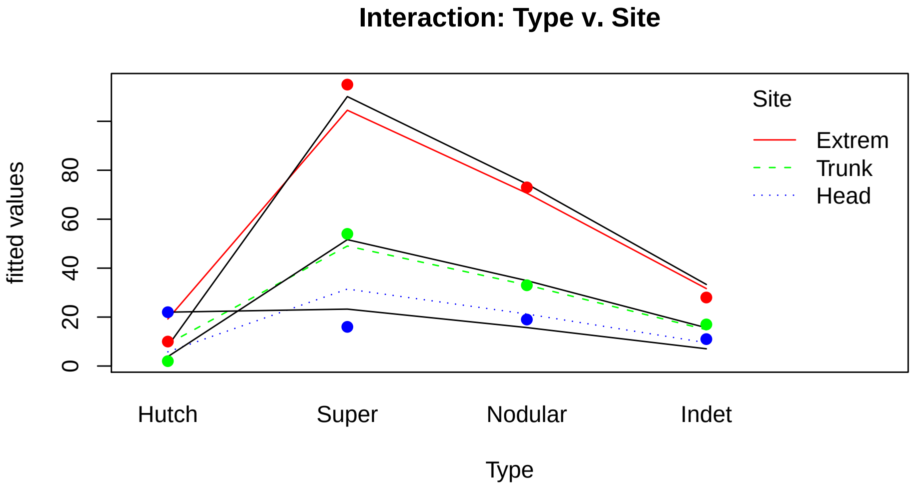

Figure 2: Interaction plot for cancer type v. site for independence and interaction models

delivery time (minutes), number of cases and distance walked (feet). Copy the `drinks.txt` file from the web page, save it in `Rwork` and read it into R.

```
> DrinksData <- read.table("drinks.txt", header = TRUE)
> attach(DrinksData, warn.conflicts = FALSE)
```

It could be argued that for this data, the standard deviation of the time should not be constant, but should be proportional to the number of cases and/or the distance walked.

Thus we have multiplicative, rather than additive, errors. One option is to transform the responses using a logarithmic transformation, much as we did with the `mammals` data. An alternative, which retains the original scale of measurement, is to observe that if

$$Y = \mu\epsilon,$$

where, without loss of generality, $\mathbb{E}(\epsilon) = 1$ and $\mathrm{Var}(\epsilon) = \sigma^2$, then $\mathrm{Var}(Y) = \sigma^2\mu^2$. This suggests using a gamma model for the data. Consider the Gamma(ν, θ) distribution with shape parameter $\nu > 0$ and scale parameter $\theta > 0$. The density function is given by

$$f(y; \nu, \theta) = y^{\nu-1}\frac{e^{-y/\theta}}{\Gamma(\nu)\theta^{\nu}}, y > 0.$$

Equating the mean $\nu\theta$ and variance $\nu\theta^2$ to μ and $\sigma^2\mu^2$, respectively, results that we must use the Gamma$(1/\sigma^2, \mu\sigma^2)$ distribution. Express this distribution as a member of an exponential dispersion family; deduce that the variance function is $V(\mu) = \mu^2$ and the canonical link function is $g(\mu) = -1/\mu$. We can fit a gamma model with

```
> GammaMod <- glm(Time ~ Cases + Distance, family = Gamma)
> summary(GammaMod)


Call:
glm(formula = Time ~ Cases + Distance, family = Gamma)

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-0.563042  -0.186538  -0.008564   0.105451   0.497320

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.132e-02  4.154e-03  17.171 3.14e-14 ***
Cases       -1.728e-03  5.093e-04  -3.393  0.00261 **
Distance    -6.428e-06  1.039e-05  -0.618  0.54269
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

(Dispersion parameter for Gamma family taken to be 0.06801047)

    Null deviance: 7.7060  on 24  degrees of freedom
Residual deviance: 1.5431  on 22  degrees of freedom
AIC: 157.29

Number of Fisher Scoring iterations: 4
```

The style of most of the output should be familiar by now. Let Y_i denote the ith time, let x_i denote the ith number of cases, and let z_i denote the distance. The model is that $Y_i \sim \text{Gamma}(1/\sigma^2, \mu\sigma^2)$, $i = 1, \ldots, n$, are independent, where

$$\frac{1}{\mu_i} = \alpha + \beta x_i + \gamma z_i,$$

for $i = 1, \ldots, n$. Notice that for some reason, R uses $1/\mu$, rather than $-1/\mu$ as the link function (though of course the only effect is to multiply the parameter estimates by -1). One new piece of information in the summary is the estimate of the dispersion parameter, which you should check comes from the estimate

$$\tilde{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)^2}{a_i V(\hat{\mu}_i)},$$

as discussed in lectures (here we have $a_i = 1$ for all i). Check the link function formula above by comparing the fitted values from the model with the reciprocal of $\hat{\alpha} + \hat{\beta} x_i + \hat{\gamma} z_i$ for $i = 1, \ldots, n$. Notice that the residual deviance, 1.5431, is certainly small in comparison with χ^2_{22}.

Exercise: It appears that one of the explanatory variables could be removed from the model. Try fitting a new model with this term removed. Is the increase in deviance significant? To see this, apply the `anova` function with `test='F'`. Recall that when testing model $\mathcal{M}_1$ against $\mathcal{M}_2$, where $\mathcal{M}_1 \subset \mathcal{M}_2$ with parameters $q < p$, respectively, we employ the likelihood ratio statistic

$$\frac{D(y; \mathcal{M}_1) - D(y; \mathcal{M}_2)}{\sigma^2} \sim \chi^2_{p-q} \quad \text{approximately},$$

where $D(y; \mathcal{M}_1)$ and $D(y; \mathcal{M}_2)$ are the deviances of $\mathcal{M}_1$ and $\mathcal{M}_2$. If σ^2 is not known but it is estimated by $\tilde{\sigma}^2$ from $\mathcal{M}_2$, then the following approximate result is used

$$\frac{D(y; \mathcal{M}_1) - D(y; \mathcal{M}_2)}{\tilde{\sigma}^2 (p-q)} \sim \mathrm{F}_{p-q, n-p}\,.$$

Exercise: In addition to the canonical link, R also supports the logarithmic and identity links. Considering how the data came about, which function of the mean do you think can be described best as a linear combination of the explanatory variables? Fit the model again with all three built-in link functions and look at plots of standardised deviance residuals versus fitted values. Which link function gives the best fit? Recall that large values (compared to 1) of standardised deviance residuals are evidence for misfit, and an obvious sign of trend in the residuals is an indication of a problem with the link function.