MATH3012 – Statistical Methods II
S-Plus Worksheet 8 – Modelling counts

Analysis of Road traffic accident data

1. The `accident` dataset concerns the number of road accidents and the volume of traffic observed on Mill Road and Trumpington Road in Cambridge during morning, midday and afternoon. By analysing this we should be able to answer questions like: (i) Is Mill Road more dangerous than Trumpington Road? (ii) How does time of day affect the rate of road accident? The data have been obtained from Prof Pat Altham (Cambridge University).

Accidents 1978–81, for traffic *into* Cambridge			
	Time of day	Number of accidents	Estimated traffic volume
Mill Road	(07.00–09.30)	4	1399
	(09.30–15.00)	20	2276
	(15.00–18.30)	4	1417
Trumpington Road	(07.00–09.30)	11	2206
	(09.30–15.00)	9	3276
	(15.00–18.30)	4	1999

2. We assume
$$Y_{ij} \sim \text{Poisson}(\mu_{ij}),\ i = 1, 2 \text{ for Road}, j = 1, 2, 3 \text{ for time.}$$

3. We might reasonably expect the number of accidents to depend on traffic `volume`, v_{ij}. It is better to work with log volume rather than volume itself since those are very high.

4. Use the Poisson GLM with canonical log link.
$$\log(\mu_{ij}) = \mu + \alpha_i + \beta_j + \gamma \log v_{ij}.$$
This is equivalent to assuming:
$$\mu_{ij} = \text{constant} \times i\text{th road effect} \times j\text{th time effect} \times \text{volume}^{\gamma}$$

5. By setting the following treatment contrasts we set $\alpha_1 = 0$. Thus $\alpha_2 = 0$ if the roads are equally risky. By the same command we also set $\beta_1 = 0$. Then β_2 represents difference between time 2 and 1, and β_3 represents difference between time 3 and 1.
```
options(contrasts=c("contr.treatment","contr.poly"))
```

6. Issue the command
```
acc.glm <- glm(nacc ~ road + time + log(volume), data=accident, family=poisson)
```

7. `summary(acc.glm)`. The output seems to say Mill road is more dangerous than Trumpington road. The mornings and afternoons are about as dangerous as each other and each is quite a lot more dangerous than the midday.

8. `anova(acc.glm, test="Chisq")`. The model seems to fit well, deviance 1.88 is non-significant when referred to χ^2 with 1 degree of freedom. The accident rate has a strong dependence on the traffic volume.

Analysis of Hodgkins data

1. Consider the `hodgkins` dataset where 538 patients with Hodgkin's disease have been cross-classified according to two factors, H, the histological type of their disease (4 levels) and R, their response to treatment (3 levels).

	Response to treatment		
Histological type	Positive	Partial	None
Lymphocyte predominance	74	18	12
Nodular sclerosis	68	16	12
Mixed cellularity	154	54	58
Lymphocyte depletion	18	10	44

2. Classification data are extremely common, and can be modelled very effectively using generalised linear models. The observations of the response variable are taken to be the counts (in this case the 12 patient totals) and a generalised linear model is used to determine how the expected counts depend on any explanatory variables (in this case, the factors H and R).

3. Counts are non-negative integers, so one approach is to treat them as observations of Poisson random variables. The canonical link function is then the log function, and Poisson generalised linear models with the log link are called *log-linear models*. A possible log-linear model for this data set is:

$$Y_i \sim \text{Poisson}(\mu_i) \qquad \log \mu_i = \alpha + \beta_H(h_i) + \beta_R(r_i) \qquad i = 1, \ldots, 12, \tag{1}$$

   where Y_i is the ith count (in the S-Plus spread-sheet), and h_i and r_i are the corresponding levels of H and R.

4. *Note:* This can be written with the i and j notations like the previous example, e.g. we can write: Y_{ij} = number of patients with the ith histological type and jth response to treatment, $i = 1, 2, 3, 4$ and $j = 1, 2, 3$.

5. We use the commands:

   (a) `hod.glm <- glm(y~h+r, data=hodgkins, family=poisson)`

   (b) `summary(hod.glm)`

   (c) `anova(hod.glm, test="Chisq")`

   (d) `plot(hod.glm)`

   (e) `u <- resid(hod.glm, type="pearson") # The Pearson residuals are in u`

   (f) `v <- resid(hod.glm, type="deviance") # The deviance residuals are in v`

(g) `sum(u^2) # The result is the Pearson X^2 statistic`

(h) `sum(v^2) # The result is the scaled deviance`

6. By comparing the $H + R$ model with the model $H + R + HR$ $(= H * R)$ we are determining whether H and R are independent or whether there is significant evidence of association. As every combination of H and R appears exactly once, the model $H * R$ is the saturated model, so the test we require compares model (1) with the saturated model (a goodness of fit test).

7. The independence model $[H + R;\ (1)]$ fails to fit. Its scaled deviance is 68.3 on 6 degrees of freedom, which is far to large to have reasonably come from a χ^2_6 distribution (p-value $< 10^{-12}$). The residual plot indicates why the model fails to fit. In particular, the residual for observation 12 is very high, and observation 10 is rather low. Therefore, the main reason that we are unable to draw the conclusion that response to treatment is independent of histological type is that the prognosis for individuals with Lymphocyte depletion is significantly worse than for other patients. Conversely, patients with Lymphocyte predominance and Nodular sclerosis fare rather better. Our conclusion is that response to treatment is associated with histological type.

8. The residual plot displays both the deviance (circle) and Pearson (square) residuals. They are generally quite close, as might be expected with relatively large counts. The Pearson X^2 statistic is 75.8, also on 6 degrees of freedom, providing even stronger evidence against the independence model.

Exercise: Analyse the `job` satisfaction data.