# Part III

# Count Data

# Count Data

Count data, in which there is no upper limit to the number of counts, usually fall into two types

- **rates** counts per unit of time/area/distance, etc
- **contingency tables** counts cross-classified by categorical variables

We will see that both of these types of count data can be modelled using Poisson glms with a log link.

# Poisson Processes

Often counts are based on events that may be assumed to arise from a Poisson process, where

- counts are observed over fixed time interval
- probability of the event approximately proportional to length of time for small intervals of time
- for small intervals of time probability of $> 1$ event is neglibile compared to probability of one event
- numbers of events in non-overlapping time intervals are independent

Examples include

- number of household burglaries in a city in a given year
- number of customers served by a saleperson in a given month
- number of train accidents in a given year

In such situations, the counts can be assumed to follow a Poisson distribution, say

$$Y_i \sim Poisson(\lambda_i)$$

# Rate Data

In many cases we are making comparisons across observation units $i = 1, \ldots, n$ with *different levels of exposure to the event* and hence the measure of interest is the **rate of occurrence**, e.g.

- number of household burglaries per 10,000 households in city $i$ in a given year
- number of customers served per hour by salesperson $i$ in a given month
- number of train accidents per billion train-kilometers in year $i$

Since the counts are Poisson distributed, we would like to use a glm to model the expected rate, $\lambda_i/t_i$, where $t_i$ is the exposure for unit $i$.

Typically explanatory variables have a multiplicative effect rather than an additive effect on the expected rate, therefore a suitable model is

$$\log(\lambda_i/t_i) = \beta_0 + \sum_{r=1}^{p} x_{ir}\beta_r$$

$$\Rightarrow \log(\lambda_i) = \log(t_i) + \beta_0 + \sum_{r=1}^{p} x_{ir}\beta_r$$

i.e. Poisson glm with the canonical log link.

This is known as a **log-linear** model.

# Offsets

The standardizing term $\log(t_i)$ is an example of an *offset*: a term with a fixed coefficient of 1.

Offsets are easily specified to `glm`, either using the `offset` argument or using the `offset` function in the formula, e.g. `offset(time)`.

If all the observations have the same exposure, the model does not need an offset term and we can model $\log(\lambda_i)$ directly.

# Ship Damage Data

The `ships` data from the `MASS` package concern a type of damage caused by waves to the forward section of cargo-carrying vessels. The variables are

- `incidents` number of damage incidents
- `service` aggregate months of service
- `period` period of operation : 1960-74, 75-79
- `year` year of construction: 1960-64, 65-69, 70-74, 75-79
- `type` type: ""A"' to ""E"'

Here it makes sense to model the expected number of incidents per aggregate months of service.

Let us consider a log-linear model including all the variables. We first exclude ships with 0 months of service and convert the `period` and `year` variables to factors:

```
library(MASS)
data(ships)

ships2 <- subset(ships, service > 0)
ships2$year <- as.factor(ships2$year)
ships2$period <- as.factor(ships2$period)

glm1 <- glm(formula = incidents ~ type + year + period,
    family = poisson(link = "log"), data = ships2,
    offset = log(service))
```

We notice that the deviance is somewhat larger than the degrees of freedom.

# Overdispersion

Lack of fit may be due to inadequate specification of the model, but another possibility when modelling discrete data is **overdispersion**.

Under the Poisson or Binomial model, we have a fixed mean-variance relationship:

$$\text{var}(Y_i) = V(\mu_i)$$

Overdispersion occurs when

$$\text{var}(Y_i) > V(\mu_i)$$

This may occur due to correlated responses or variability between observational units.

We can adjust for over-dispersion by estimating a dispersion parameter

$$\text{var}(Y_i) = \phi V(\mu_i)$$

This changes the assumed distribution of our response, to a distribution for which we do not have the full likelihood.

However the score equations in the IWLS

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{a_i(y_i - \mu_i)}{V(\mu_i)} \times \frac{x_{ij}}{g'(\mu_i)} = 0$$

only require the variance function, so we can still obtain estimates for the parameters. Note the score equations do not depend on $\phi$, so we will obtain the same estimates as if $\phi = 1$.

This approach is known as **quasi-likelihood estimation**. Whilst estimating $\phi$ does not affect the parameter estimates, it will change inference based on the model.

The asymptotic theory for maximum likelihood also applies to quasi-likelihood, in particular $\boldsymbol{\beta}$ is approximately distributed as

$$N(\boldsymbol{\beta}, \phi(X^T \hat{W} X)^{-1})$$

so compared to the case with $\phi = 1$, the standard errors of the parameters are multiplied by $\sqrt{(}\phi)$.

Since $\phi$ is estimated, Wald tests based on the Normal assumption are $t$ rather than $Z$ tests.

The deviance based on the likelihood of the exponential family distribution with the same variance function may be used as a *quasi-deviance*. Since $\phi$ is estimated rather than fixed at 1, nested models are compared by referring

$$\{D_{big} - D_{small}\}/\{\hat{\phi}(p_{big} - p_{small})\}$$

to the $F$ distribution with $p_{big} - p_{small}, n - p_{big}$ degrees of freedom.

The AIC is undefined for quasi-likelihood models.

In the Ships Damage data, it is likely that there is inter-ship variability in accident-proneness. Therefore we might expect some over-dispersion.

We can switch to a quasi-likelihood estimation using the corresponding `quasi-` family in R:

```
glm2 <- update(glm1, family = quasipoisson(link = "log"))
```

The dispersion parameter is estimated as 1.69, much larger than the value of 1 assumed under the Poisson model.

We can now check the significance of the predictors adjusting for the over dispersion:

```
anova(glm2, test = "F")
```

All the variables are significant. Adding second order interactions does not improve the model.

# Ship Damage Model

```
Call:  glm(formula = incidents ~ type + year + period,
family = quasipoisson(link = "log"), data = ships2,
offset = log(service))

Coefficients:
(Intercept)        typeB        typeC        typeD
   -6.40590     -0.54334     -0.68740     -0.07596
      typeE       year65       year70       year75
    0.32558      0.69714      0.81843      0.45343
   period75
    0.38447

Degrees of Freedom: 33 Total (i.e. Null);  25 Residual
Null Deviance:      146.3
Residual Deviance: 38.7  AIC: NA
```

# Intepretation of Ship Damage Model

We have the model

$$\log(\lambda_{typ}) = \log(s_{typ}) + \beta_0 + \beta_{1t} + \beta_{2y} + \beta_{3p}$$

Consider ships of type C and E. We have

$$\log(\lambda_{Eyp}) - \log(\lambda_{Cyp}) = \log(s_{Eyp}) - \log(s_{Cyp}) + \beta_{1E} - \beta_{1C}$$

Since $\beta_{1A} = 0$, we have

$$\beta_{1E} - \beta_{1C} = \log\left(\frac{\lambda_{Eyp}}{s_{Eyp}}\right) - \log\left(\frac{\lambda_{Cyp}}{s_{Cyp}}\right) = \log\left(\frac{r_{Eyp}}{r_{Cyp}}\right)$$

So $\exp(\beta_{1E} - \beta_{1C})$ is the ratio of the rates (expected number of damages per month in service)

We can conclude the following

- Types B and C have the lowest risk, E the highest. The rate for E is $\exp(0.33 - (-0.69)) = 2.75$ times that for C.
- The incident rate increased by a factor of $\exp(0.38) = 1.47$ after 1974
- The ships built between 1960 and 1964 seem to be the safest, with ships built between 1965 and 1974 having the highest risk

Also we have found evidence of inter-ship variability. When estimated, the coefficient of `service` is 0.90 (s.e. 0.13), confirming that `damage` is roughly proportional to `service`.

# Contingency Tables

The counts in contingency tables could arise from different sampling schemes.

It may be that the cell counts are realizations of independent Poisson processes, e.g. different groups of patients attending a health clinic during a fixed period of time. Thus we have counts $n_c, c = 1, \ldots, C$ distributed as Poisson($\mu_c$).

More commonly, the cell counts may an observation of a multinomial response, e.g a fixed sample of patients is taken and cross-classified by cholesterol level and whether or not they had heart disease. Thus we have a set of counts $n_1, \ldots n_C$ distributed as Multinomial($p_1, \ldots, p_c, n$).

It can be shown that if the cell counts are realizations of independent Poisson processes but the total count is fixed *a priori*, then the cell counts are Multinomial$(\mu_1/n, \ldots, \mu_c/n, n)$.

Thus under either sampling scheme, the cell counts can be modelled using a Poisson glm. In the multinomial case we condition on the total count by including an intercept in the model.

We will consider models for contingency tables from the viewpoint of multinomial sampling.

# Independence Model

If the two cross-classifying variables are independent, the joint probabilities for the cells in that table are simply determined by the marginal probabilities:

$$P(X = i \text{ and } Y = j) = P(X = i)P(Y = j)$$

or $p_{ij} = p_i p_j$

In terms of log expected frequencies we have

$$\begin{aligned}\log(\mu_{ij}) = \log(np_{ij}) &= \log n + \log(p_i p_j) \\ &= \log n + \log p_i + \log p_j\end{aligned}$$

i.e. a Poisson log-linear model. We represent this **independence** model as

$$\log(\mu_{ij}) = \lambda_0 + \lambda_i^X + \lambda_j^Y$$

# Diagnosis of Respiratory Tract Infections

Hueston and Stott (2000) report a study of clinicians' diagnoses of respiratory tract infections over a 14-month period. The aim was to determine whether a reduction in prescription of antibiotics to acute bronchitis patients was due to clinicians assigning an alternative diagnosis.

| | Time period | | | | |
|---|---|---|---|---|---|
| Diagnosis | 1-3/96 | 4-6/96 | 7-9/96 | 10-12/96 | 1-2/97 |
| Acute bronchitis | 113 | 58 | 40 | 108 | 100 |
| Acute sinusitis | 99 | 37 | 23 | 50 | 32 |
| URI | 410 | 228 | 125 | 366 | 304 |
| Pneumonia | 60 | 43 | 30 | 56 | 45 |
| Total | 682 | 366 | 218 | 580 | 481 |

# Exploring the Data

We can explore the pattern of the contingency table using a **mosaic plot**:

```
diag <- rep(c("bron", "sinus", "URI", "pneu"), 5)
time <- rep(c("win96", "spr96", "sum96", "aut96", "spr97"),
            rep(4, 5))
rt <- data.frame(diag = factor(diag, unique(diag)),
                 time = factor(time, unique(time)),
                 count = c(113, 99, 410, 60, 58, 37, 228,
                 43, 40, 23, 125, 30, 108, 50, 366, 56,
                 100, 32, 304, 45))
plot(xtabs(count ~ time + diag, rt))
```

The pattern is similar to what would be given by an independence model.

We can see this by fitting this model and plotting a mosaic plot of the fitted counts

```
ind <- glm(count ~ diag + time, poisson, rt)
plot(xtabs(fitted(ind) ~ time + diag, rt))
```

However the deviance shows that there is significant lack of fit (D $= 29.59$, d.f. $= 12$). Rejecting this model is equivalent to rejecting a null hypothesis of independence using a Pearson $\chi^2$ test.

# Residual Analysis

For small contingency tables, it can often be useful to tabulate the residuals to check for residual patterns in the data:

```
round(t(xtabs(residuals(ind)~ time + diag, rt)), 1)
```

Acute sinusitis diagnoses are decreasing over time and there is a corresponding increase in acute bronchitis diagnosis.

# Other Models for Two-way Tables

If we add and interaction term to the independence model

$$\log(\mu_{ij}) = \lambda_0 + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$$

the model is saturated - the observed data are fitted exactly.

More interesting intermediate models have been proposed for tables with more structure, e.g. ordered categories or square tables, but we shall not consider these here.

# Mutual Independence Model

If each pair of variables are independent, then

$$p_{ijk} = p_i p_j p_k$$

which is represented by the **mutual independence** model

$$\log(\mu_{ijk}) = \lambda_0 + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$$

This model is rarely interesting - we are more interested in associations between the variables.

# Joint Independence Model

If the first and second variables are dependent, but jointly independent of the third, then

$$p_{ijk} = p_{ij}p_k$$

which in terms of log expected frequencies is

$$\log(\mu_{ij}) = \log n + \log p_{ij} + \log p_k$$

We include all main effects to give the **joint independence** model

$$log(\mu_{ijk}) = \lambda_0 + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY}$$

The two-way interaction shows which two variables are dependent

# Conditional Independence Model

Now suppose that the first two variables are independent given the value of the third variable, then

$$p_{ij|k} = p_{i|k} p_{j|k}$$

and so

$$p_{ijk} = p_{ik} p_{jk} / p_k$$

which gives

$$\log(\mu_{ij}) = \log n + \log p_{ik} + \log p_{jk} - \log p_k$$

Again, we include all the main effects to give the **conditional independence** model

$$\log(\mu_{ijk}) = \lambda_0 + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XZ} + \lambda_{ij}^{YZ}$$

# Further Models

Including all two-way interactions results in the **uniform association model**, considered later.

All the models described so far are nested within the saturated model, which includes the three-way interaction and all lower-order terms.

A simple approach to identify the appropriate association model is to start with the saturated model and determine how the model can be simplified.

# Example: Drug Use

The following 2 x 2 x 2 table cross-classifies students according to their alcohol, cigarette and drug use

| | | Marijuana Use | |
|---|---|---|---|
| Alcohol Use | Cigarette Use | Yes | No |
| Yes | Yes | 911 | 538 |
| | No | 44 | 456 |
| No | Yes | 3 | 43 |
| | No | 2 | 279 |

# Modelling the Drugs Data

We don't need to fit the saturated model since we know it has a deviance of zero on zero d.f. so we start with the uniform association model:

```
lab <- c("Y", "N")
drugs <- data.frame(alcohol = gl(2, 4, 8, labels = lab),
                    cigarette = gl(2, 2, 8, labels = lab),
                    marijuana = lab,
                    count = c(911, 538, 44, 456, 3, 43, 2, 279))

unif <- glm(count ~ . - alcohol:cigarette:marijuana,
            poisson, data = drugs)
summary(unif)
```

The likelihood ratio test statistic to compare the uniform association model to the saturated model is simply the deviance of the uniform association model. So we can see that adding the three-way interaction does not significantly improve the model.

It is clear that dropping any further terms from the model will significantly increase the deviance.

# Uniform Association Model

The uniform association model is so called because the odds ratios between two variables are the same for any level of the third variable. E.g. for any level of marijuana use $i$

$$\frac{\text{odds of alcohol use}|\text{cigarette use}}{\text{odds of alcohol use}|\text{no cigarette use}} = \frac{\mu_{YYi}/\mu_{NYi}}{\mu_{YNi}/\mu_{NNi}}$$

$$= \exp(\lambda_{YY}^{alc,cig}) = \exp(2.05) = 7.8$$

i.e. students who have smoked cigarettes have estimated odds of alcohol use that are 7.8 times the estimated odds for students who have not smoked cigarettes.

# Higher Dimensional Tables

The same ideas extend to higher dimensional tables, although model-building and interpretation can be quite complex.

In the drug use example, students were also classified by sex and race.

We use `ftable` to view the full data:

```
drugs2 <- read.table("drugs.txt", header = TRUE)
ftable(xtabs(count ~ sex + marijuana + alcohol +
       cigarette + race, drugs2))
```

# Response and Explanatory Factors

Here sex and race are **explanatory** factors. We treat the marginal totals of these factors as being fixed.

The **minimal model** must contain the interaction of all the explantory factors.

Interactions between the **response** factors – here alcohol, cigarette and marijuana use – and the explantory factors indicate interesting structure.

# Model Building

We consider blocks of terms to determine the order of the model:

```
ind <- glm(count ~ . + race:sex, poisson, data = drugs2)
homog <- glm(count ~ (.)^2, poisson, data = drugs2)
ord3 <- glm(count ~ (.)^3, poisson, data = drugs2)
anova(ind, homog, ord3, test = "Chisq")
```

It seems we do not need to consider terms of higher order than 2.

Now we try to simplify the homogeneous association model.

We can consider the effect of single deletions using `drop1`:

```
drop1(homog)
homog <- update(homog, . ~ . - sex:cigarette)
```

Dropping the `race:cigarette` interaction leads to the smallest increase in deviance, so we drop this term.

We continue dropping terms until no terms can be dropped without significantly increasing the deviance:

```
drop1(homog)
homog <- update(homog, . ~ . - race:sex)
homog <- update(homog, . ~ . - race:cigarette)
homog <- update(homog, . ~ . - race:marijuana)
```

The final model has seven two-way interactions, with a deviance of 20.54 on 20 d.f.

# Association Graph

The final model can be represented by the association graph (shown on the board!)

Every path between `cigarettes` and {`sex`, `race` } involves a variable in {`alcohol`, `marijuana`}.

Thus given the outcome on alcohol and marijuana use, cigarette use is independent of race and gender.

Cigarette use can thus be safely summarised in a table collapsed over sex and race (proportion using cigarettes in each category):

| Alcohol Use | Marijuana Use | |
|---|---|---|
| | Yes | No |
| Yes | 95% | 54% |
| No | 60% | 13% |

Note here that the figure of 60% smoking in the 'marijuana but not alcohol' cell is based on only 5 cases and should therefore be treated with caution.

# Exercises

1. The following data are from a cross-sectional study of 400 patients with a form of skin cancer called malignant melanoma.

| | Site | | | |
|---|---|---|---|---|
| Tumour type | Head & neck | Trunk | Extremities | Total |
| Hutchinson's melanotic freckle | 22 | 2 | 10 | 34 |
| Superficial spreading melanoma | 16 | 54 | 115 | 185 |
| Nodular | 19 | 33 | 73 | 125 |
| Indeterminate | 11 | 17 | 28 | 56 |
| Total | 68 | 106 | 226 | 400 |

Create a data.frame in R from these data.

2. Use `xtabs` to reproduce the table shown in question 1. We would like to know if there is an association between Site and Tumour Type.

Use `margin.table` to find the row totals and save them in a vector. Now use `prop.table` to represent the table counts as percentages of the column totals and then the row totals as percentages of the grand total. If there is no association between Site and Tumour Type, the percentages in a given row should be approximately equal to the overall percentage for that row. Does this seem to be the case?

Repeat the above, this time finding percentages for the columns.

3. An alternative way to view the data is to use a mosaic plot as seen in the lectures. Create this plot. Do Site and Tumour Type seems to be independent?

4. We can test for independence using the conventional chi-squared test. Under independence, the expected frequencies for each cell can be calcuated from the marginal totals as follows:

$$e_{ij} = y_{i.}y_{.j}/n$$

These are compared to the observed values through the statistic

$$X^2 = \sum_i \sum_j \frac{(y_{ij} - e_{ij})^2}{e_{ij}}$$

which under independence follows a $\chi^2_{(i-1)(j-1)}$ distribution. Perform this test in R using `chisq.test`. What do you find?

5. Now fit the log-linear independence model to the data using `glm`. Testing for a significant interaction term in the model is equivalent to testing the hypothesis of independence. Is the interaction significant here?

Confirm that the sum of the squared pearson residuals from the independence model is equal to the chi-squared statistic found in question 4.

Note the statistics in both tests are assumed to be approximately $\chi^2_6$, but they are different! We should obtain similar conclusions if the $\chi^2$ approximation is valid.

6. Use `xtabs` to look at the residuals from the independence model. You should find that there is one particularly large residual.

7. In this case we can propose a simple alternative to the independe model, in which the count for cell with the large residual is modelled exactly. Create a factor in R which indicates this cell. Add this "cell effect" to the independence model. Is this model a significant improvement? Does the model adequately describe the data? Interpret your findings.

8. The data set "Long.txt" contains data on the productivity of biochemistry PhD students. The variables are as follows

- `art` Number of articles published by the student during last three years of PhD
- `fem` Gender: 1 if female, 0 if male
- `mar` Maritial status: 1 if married, 0 if not
- `kid5` Number of children five years old or younger
- `phd` Prestige rating of PhD department
- `ment` Number of articles published by mentor during last three years

Read the dataset into R and attach the data frame.

9. Since the exposure of all students is fixed at three years, we can model the students' article count directly, using a Poisson log-linear model with no offset. Investigate the bivariate relationships of `log(art)` with the other variables. Which variables does the article count appear to depend on?

10. Fit a Poisson log-linear model regressing `art` on the linear effect of the other variables. Notice that the deviance is much greater than the degrees of freedom. Could this be due to a need for second order terms?

11. Fit a quasi-Poisson model regressing `art` on the linear effect of the other variables. Do the data appear to be overdispersed?

12. Using a Poisson or quasi-Poisson model as you see fit, select an appropriate model for `art`. Interpret your final model.