Chapter 5. Analysis of Contingency Tables by Log Linear Models

MAST90139 Statistical Modelling for Data Science Slides

Guoqi Qian

SCHOOL OF MATHEMATICS AND STATISTICS

THE UNIVERSITY OF MELBOURNE

Contents

§5.1 Two-way contingency tables

- It is not possible to use (standard) logistic regression models for categorical response variables/factors with more than two categories.
- One method for dealing with this type of data is based on the Poisson distribution with models referred to as **log-linear models**.
- Log-linear models can be used to analyse all types of categorical data — you don't even need a response variable/factor!
- The data for a log-linear model analysis often can be set out as a contingency table, where the table can be of two or more dimensions.

Example: Income and job satisfaction

	Job satisfaction				
Income ($)	Very Dis.	Mod. Dis.	Mod. Sat.	Very Sat.	Total
< 6000	20	24	80	82	206
6000 — 15000	22	38	104	125	289
15000 — 25000	13	28	81	113	235
> 25000	7	18	54	92	171
Total	62	108	319	412	901

The classic approach to testing for association between income and job satisfaction is to calculate expected frequencies, assuming no association, and then to compare the observed and expected frequencies using Pearson's X^2 statistic.

Expected frequencies

Income ($)	Job satisfaction Very Dis.	Mod. Dis.	Mod. Sat.	Very Sat.	Total
< 6000	14.2	24.7	72.9	94.2	206
6000 — 15000	19.9	34.6	102.3	132.2	289
15000 — 25000	16.2	28.2	83.2	107.5	235
> 25000	11.8	20.5	60.5	78.2	171
Total	62	108	319	412	901

- Here $X^2 = \sum \frac{(O-E)^2}{E} = 11.98$ which is less than $\chi^2_{0.95}(9)$, hence we do not reject the hypothesis that there is no association between income and job satisfaction.
- Similar result can be obtained by using a *log linear model*.

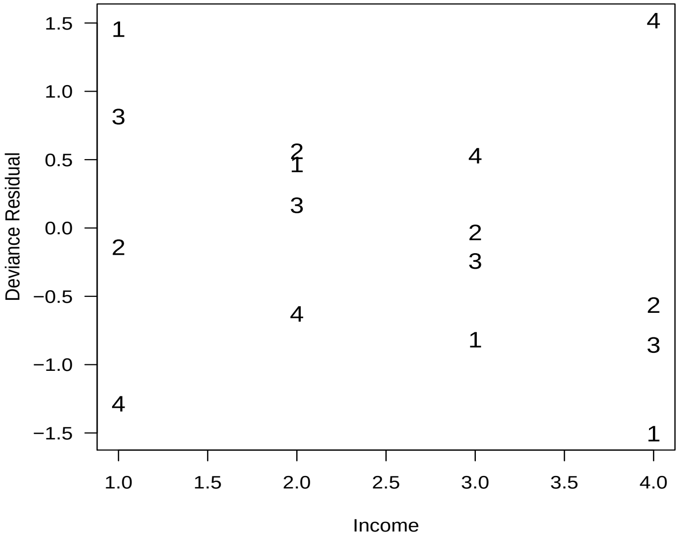
R analysis based on log linear model

```
> y <- c(20, 22, 13, 7, 24, 38, 28, 18, 80, 104, 81, 54, 82, 125,
+     113, 92)
> inc <- rep(1:4, 4)
> sat <- rep(1:4, c(4, 4, 4, 4))
> print(cbind(y, inc, sat)[1:12, ])

        y inc sat
 [1,]  20   1   1
 [2,]  22   2   1
 [3,]  13   3   1
 [4,]   7   4   1
 [5,]  24   1   2
 [6,]  38   2   2
 [7,]  28   3   2
 [8,]  18   4   2
 [9,]  80   1   3
[10,] 104   2   3
[11,]  81   3   3
[12,]  54   4   3

> jobsat.1 <- glm(y ˜ factor(inc) + factor(sat), family = poisson)
```

R analysis

```
> anova(jobsat.1)

Analysis of Deviance Table

Model: poisson, link: log

Response: y

Terms added sequentially (first to last)


             Df Deviance Resid. Df Resid. Dev
NULL                            15     445.76
factor(inc)   3    32.92        12     412.84
factor(sat)   3   400.81         9      12.04

> 1 - pchisq(12.0369, 9)

[1] 0.2112370
```

The plot of deviance residuals against income follows ...


Deviance Residual
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Income
1
3
2
4
2
1
3
4
4
2
3
1
4
2
3
1

```
> anova(glm(y ~ factor(inc) * factor(sat), family = poisson), test = "Chi")
Analysis of Deviance Table

Model: poisson, link: log

Response: y

Terms added sequentially (first to last)


                        Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                                       15     445.76
factor(inc)              3    32.92        12     412.84 3.349e-07
factor(sat)              3   400.81         9      12.04 1.481e-86
factor(inc):factor(sat)  9    12.04         0  2.842e-14      0.21
```

The model is **saturated** since its residual deviance equals 0.

Log-linear model analysis (1)

Let $Y_{ij} \overset{d}{=}$ independent Poisson random variables with mean λ_{ij}, then

$$\mathbb{E}(Y_{ij}) = \mathrm{Var}(Y_{ij}) = \lambda_{ij}$$

Log-linear model for λ_{ij}, i being the row index and j column index:

$$\log(\lambda_{ij}) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}; \quad i = 1, \cdots, I;\ j = 1, \cdots, J;$$

where α_i, β_j and $(\alpha\beta)_{ij}$ represent row effect, column effect, and row-column interaction effect, respectively. These effects must satisfy the following $I + J + 1$ constraints or equivalent ones in order to be estimable:

$$\sum_i \alpha_i = \sum_j \beta_j = \sum_{i,j} (\alpha\beta)_{ij} = 0$$

$$\sum_j (\alpha\beta)_{1j} = \cdots = \sum_j (\alpha\beta)_{Ij}, \quad \sum_i (\alpha\beta)_{i1} = \cdots = \sum_i (\alpha\beta)_{iJ}$$

Log-linear model analysis (2)

An equivalent expression of the above log-linear model is

$$\log(\lambda) = \gamma_0 + \gamma_2 R_2 + \cdots + \gamma_I R_I + \delta_2 C_2 + \cdots + \delta_J C_J \\ + \eta_{22} R_2 C_2 + \eta_{23} R_2 C_3 + \cdots + \eta_{IJ} R_I C_J$$

where R_i's are dummy variables for the row factor, such as $R_i = 1$ for row i and 0 for other rows; C_j's are dummy variables for the column factor. It is easy to determine the relationships among parameters in these two expressions, e.g. $\lambda_{11} = \mu + \alpha_1 + \beta_1 + (\alpha\beta)_{11} = \gamma_0$.

Independence between row and column factors

$$\Longrightarrow \quad (\alpha\beta)_{ij} = 0 \quad \text{or} \quad \eta_{ij} = 0 \quad \text{for all } i, j.$$

(Pearson) residuals: $r_{ij}^{(P)} = \frac{y_{ij} - \hat{\lambda}_{ij}}{\sqrt{\hat{\lambda}_{ij}}}$

Deviance residuals: $r_{ij}^{(D)} = \text{sign}(y_{ij} - \hat{y}_{ij}) \sqrt{2[y_{ij} \log(\frac{y_{ij}}{\hat{y}_{ij}}) - (y_{ij} - \hat{y}_{ij}]}$

§5.2 Independence in 2-way contingency tables

2×2 tables: with factors A and B.

A	B 1	 2	total
1	p_{11}	p_{12}	$p_{1.}$
2	p_{21}	p_{22}	$p_{2.}$
total	$p_{.1}$	$p_{.2}$	1

$$\begin{aligned} \text{Independence} \quad &\Leftrightarrow \quad p_{ij} = p_{i.}p_{.j} \\ &\Leftrightarrow \quad \frac{p_{11}p_{22}}{p_{12}p_{21}} = \frac{p_{1.}p_{.1}p_{2.}p_{.2}}{p_{1.}p_{.2}p_{2.}p_{.1}} = 1. \end{aligned}$$

Log-linear model $\log(\lambda_{ij}) = \log(Np_{ij}) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$

$$\begin{aligned} \text{Hence, } \log\frac{p_{11}p_{22}}{p_{12}p_{21}} \quad &= \quad (\alpha\beta)_{11} + (\alpha\beta)_{22} - (\alpha\beta)_{12} - (\alpha\beta)_{21} = 0 \\ \Longleftrightarrow \quad & \quad \text{all } (\alpha\beta)_{ij} = 0 \quad \Longleftrightarrow \quad A \text{ and } B \text{ are independent.} \end{aligned}$$

$r \times c$ tables

			B			
A	1	$\cdots$	j	$\cdots$	c	total
1	p_{11}	$\cdots$	p_{1j}	$\cdots$	p_{1c}	$p_{1.}$
$\vdots$	$\vdots$	$\cdots$	$\cdots$	$\cdots$	$\vdots$	$\vdots$
i	p_{i1}	$\cdots$	p_{ij}	$\cdots$	p_{ic}	$p_{i.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
r	p_{r1}	$\cdots$	p_{rj}	$\cdots$	p_{rc}	$p_{r.}$
total	$p_{.1}$	$\cdots$	$p_{.j}$	$\cdots$	$p_{.c}$	1

$$\text{Independence} \quad \Leftrightarrow \quad p_{ij} = p_{i.}p_{.j} \quad \Leftrightarrow \quad \frac{p_{ij}p_{rc}}{p_{rj}p_{ic}} = 1 \quad \text{for all } i, j.$$

Log-linear model $\quad \log(\lambda_{ij}) = \log(Np_{ij}) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij},$

$$\begin{aligned} \log \frac{p_{ij}p_{rc}}{p_{rj}p_{ic}} &= (\alpha\beta)_{ij} + (\alpha\beta)_{rc} - (\alpha\beta)_{rj} - (\alpha\beta)_{ic} = 0 \\ \Longleftrightarrow \quad & (\alpha\beta)_{ij} = 0 \quad \text{for all } i, j \quad \Longleftrightarrow \quad A \text{ and } B \text{ are independent.} \end{aligned}$$

Models based on frequencies

Let y_{ij} denote the observed frequency in cell ij. Then

$$\mathbb{E}(Y_{ij}) = \lambda_{ij} = Np_{ij}$$

where N denotes the total sample size.
Hence

$$\begin{aligned}\log(\lambda_{ij}) &= \log(Np_{ij}) \\ &= \log(N) + \log(p_{ij}) \\ &= \log(N) + \log(p_0) + \alpha_i + \beta_j + (\alpha\beta)_{ij} \\ &= \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} \qquad \text{(implying } \mu = \log(Np_0).)\end{aligned}$$

Again, independence implies no interaction, i.e. $(\alpha\beta)_{ij} = 0 \quad$ for all i, j.

Cross product ratio for 2 × 2 tables

The **cross product ratio**

$$\gamma = \frac{p_{11}p_{22}}{p_{12}p_{21}} = \frac{\lambda_{11}\lambda_{22}}{\lambda_{12}\lambda_{21}}$$

is a basic measure of association for 2 × 2 tables.

Properties:

1. 'Invariant' under interchange of rows, or columns.
2. Invariant under row and/or column multiplications, by a constant.
   E.g. multiply row 1 by 39.
3. 'Clear' interpretation. γ is the relative odds (or **odds ratio**) of being in the first column (say).
4. Can be used in $r \times c$ tables, either through a series of 2×2 partitionings or by looking at 2×2 subtables.
5. $0 \leq \gamma \leq \infty$. Independence $\iff \gamma = 1$.
6. $\hat{\gamma} = \frac{y_{11}y_{22}}{y_{12}y_{21}}$ can be used to test for independence (test of $\gamma = 1$).

§5.3 Study design

- Most studies that give rise to categorical (response) data are observational studies, from which it is very difficult to deduce a **causal relationship**.
- Often the best that can be done is to conclude that there is some form of relationship between the factors (or variables).
- Even among observational studies there are different types of study, the nature of which affects the type of conclusions that can be reached.
- We illustrate here with three (hypothetical) studies for investigating the association between smoking and lung cancer.

Case–control (retrospective) study

Cause of death

	Lung Cancer	Other	Total
Smoker	225	100	
Non-smoker	75	200	
Total	300	300	600

$$\text{odds ratio} = \frac{225 \times 200}{100 \times 75} = 6.0$$

$$\hat{\beta} = \ln(6.0) = 1.792; \quad se(\hat{\beta}) = 0.181$$

Case-control study tells us something about the **association between smoking and lung cancer** and *something about the proportion of people who die of lung cancer/other causes who are smokers*, but **nothing about** the prevalence of lung cancer deaths among smokers/non-smokers, nor the prevalence of smoking.

Cohort (prospective) study

	Cause of death		
	Lung Cancer	Other	Total
Smoker	180	120	300
Non-smoker	60	240	300
Total			600

$$\text{odds ratio} = \frac{180 \times 240}{120 \times 60} = 6.0$$

$$\hat{\beta} = \ln(6.0) = 1.792; \quad se(\hat{\beta}) = 0.186$$

Cohort study tells us something about the **association between smoking and lung cancer** and *something about the proportions of smokers and of non-smokers who die from lung cancer*, but **nothing about** the prevalence of lung cancer deaths, nor of smoking among lung cancer/other death patients.

Cross–sectional study

	Lung Cancer	Other	Total
Smoker	216	144	
Non-smoker	48	192	
Total			600

$$\text{odds ratio} = \frac{216 \times 192}{144 \times 48} = 6.0$$

$$\hat{\beta} = \ln(6.0) = 1.792; \quad se(\hat{\beta}) = 0.194$$

Cross-sectional study tells us something about **the association between smoking and lung cancer** and something about **the prevalence of both lung cancer and of smoking**.

§5.4 Test for trend in 2-way tables

One of the advantages of log-linear models is that they enable us to test for relationships between factors other than just independence.

Example: Survival versus severity of disease

	Severity			
	1	2	3	total
Survivors	19	15	6	40
Deaths	4	10	6	20
total	23	25	12	60

R analysis

```
> trend.y <- c(19, 4, 15, 10, 6, 6)
> survival <- c(1, 2, 1, 2, 1, 2)
> severity <- c(1, 1, 2, 2, 3, 3)
> trend.1 <- glm(trend.y ~ factor(survival) + factor(severity),
+     family = poisson)
> anova(trend.1)

Analysis of Deviance Table

Model: poisson, link: log

Response: trend.y

Terms added sequentially (first to last)


                 Df Deviance Resid. Df Resid. Dev
NULL                                 5    16.9643
factor(survival)  1   6.7960         4    10.1683
factor(severity)  2   5.3264         2     4.8419
```

R analysis

```
> 1 - pchisq(4.84189, 2)

[1] 0.08883763

> print(matrix(resid(trend.1, type = "deviance"), nrow = 2, ncol = 3,
+      byrow = F))

           [,1]       [,2]       [,3]
[1,]  0.9023535 -0.4153527 -0.7401453
[2,] -1.4589835  0.5595514  0.9303662
```

The (deviance) residuals show a tendency for deaths to increase as severity increases, though the ”independence” model does provide an adequate fit to the data.

R analysis: allow for a trend

```
> trend.2 <- glm(trend.y ~ factor(survival) + factor(severity) +
+      factor(survival):severity, family = poisson)
> anova(trend.2, test = "Chi")

Analysis of Deviance Table

Model: poisson, link: log

Response: trend.y

Terms added sequentially (first to last)


                          Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                                          5    16.9643
factor(survival)           1   6.7960         4    10.1683    0.0091
factor(severity)           2   5.3264         2     4.8419    0.0697
factor(survival):severity  1   4.4131         1     0.4288    0.0357
```

Adding the trend term decreases the deviance by 4.413, which is significant at the 5% level.

R analysis

There is statistically significant evidence that the chance of survival decreases as severity increases.

```
> print(matrix(resid(trend.2, type = "deviance"), nrow = 2, ncol = 3,
+      byrow = F))

           [,1]       [,2]       [,3]
[1,]  0.1328823 -0.2888514  0.2419701
[2,] -0.2740200  0.3774868 -0.2270031

> summary(trend.2)$coef

                            Estimate Std. Error    z value     Pr(>|z|)
(Intercept)                3.7064248  0.5103761  7.2621443 3.810013e-13
factor(survival)2         -2.1861789  0.8085440 -2.7038464 6.854197e-03
factor(severity)2          0.6605444  0.4249859  1.5542738 1.201191e-01
factor(severity)3          0.3627776  0.6061402  0.5985045 5.495034e-01
factor(survival)1:severity -0.7926268  0.3893116 -2.0359704 4.175333e-02
```

§5.5 Relationship between logistic and log-linear models

Consider the following $2 \times c$ contingency table where the rows correspond to the (two) levels of a response variable and where there is a value of a covariate, x, associated with each column.

	$1(x_1)$	$2(x_2)$	$\ldots$	$c(x_c)$
1	$y_{11}(p_{11})$	$y_{12}(p_{12})$	$\ldots$	$y_{1c}(p_{1c})$
2	$y_{21}(p_{21})$	$y_{22}(p_{22})$	$\ldots$	$y_{2c}(p_{2c})$
	$y_{.1}(p_{.1})$	$y_{.2}(p_{.2})$	$\ldots$	$y_{.c}(p_{.c})$

$$y_{..} = N \quad p_{..} = 1$$

Logistic regression models for a 2×c table

Logistic modesl may be considered for this $2 \times c$ table:

$$Y_{1j} \overset{d}{=} \text{Bin}(y_{.j}, \theta_j) \quad \text{where} \quad \theta_j = \frac{p_{1j}}{p_{.j}} \quad \text{cond. prob. of 'success' in col. } j$$

1. **Model for independence:**

$$\text{logit}(\theta_j) = \theta^*, \quad \text{a constant for all } j$$

2. **Model for straight line regression on x (trend):**

$$\text{logit}(\theta_j) = \alpha^* + \beta^* x_j$$

Log-linear models for a $2 \times c$ table

Log-linear models may also be considered for this $2 \times c$ table:

$$\mathbb{E}(Y_{ij}) = \lambda_{ij} = Np_{ij}$$

$$\text{hence } \lambda_{1j} = N\theta_j p_{.j} \quad \text{and} \quad \lambda_{2j} = N(1-\theta_j)p_{.j}.$$

1. **Model for independence:**

$$\begin{array}{rrcl} & \log(\lambda_{ij}) & = & \mu + \alpha_i + \beta_j \\ \Rightarrow & \log(\lambda_{1j}) - \log(\lambda_{2j}) & = & \text{logit}(\theta_j) = \alpha_1 - \alpha_2 \\ & (& = & \theta^*) \quad \text{for all } j \end{array}$$

2. **Model for trend:**

$$\begin{array}{rrcl} & \log(\lambda_{ij}) & = & \mu + \alpha_i + \beta_j + \gamma_i x_j \\ \Rightarrow & \log(\lambda_{1j}) - \log(\lambda_{2j}) & = & \text{logit}(\theta_j) \\ & & = & (\alpha_1 - \alpha_2) + (\gamma_1 - \gamma_2)x_j \\ & (& = & \alpha^* + \beta^* x_j) \end{array}$$

Example: Survival versus severity of disease (ctd)

Logistic regression analysis

```
> trend.ly <- c(19, 15, 6)
> trend.ln <- c(23, 25, 12)
> trend.ls <- c(1, 2, 3)
> trend.3 <- glm(trend.ly/trend.ln ~ trend.ls, family = binomial,
+     weight = trend.ln)
```

Note the equivalent null and deviance results and the parameter estimates between `trend.3` and `trend.2` on Slides 23& 24.

```
> summary(trend.3)

Call:
glm(formula = trend.ly/trend.ln ~ trend.ls, family = binomial,
    weights = trend.ln)

Deviance Residuals:
      1        2        3
 0.3045  -0.4753   0.3318

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.1862     0.8085   2.704  0.00685 **
trend.ls     -0.7926     0.3893  -2.036  0.04175 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4.84189  on 2  degrees of freedom
Residual deviance: 0.42876  on 1  degrees of freedom
AIC: 14.134

Number of Fisher Scoring iterations: 3
```

§5.6 Higher order contingency tables

- All (most) of the important aspects can be illustrated with 3-way tables
- The main issue is with the concept of independence

§5.7 Types of independence

Possible associations between the 3 factors A, B and C in a 3-way table:

1. A, B and C mutually independent
2. A independent of B and C, together
3. A independent of B for each level of C (i.e. conditional independence between A and B, given the level of C)
4. A independent of B, marginally over C (i.e. ignoring C)
5. no three-factor interaction (i.e. all two-factor associations are homogeneous)

The roles of A, B and C can be interchanged in each of (2), (3) and (4).

Log-linear models for three-way tables

Let y_{ijk} denote the observed frequencies, with marginal totals $y_{ij.}$, $y_{i..}$ etc. Let λ_{ijk}, $\lambda_{ij.}$, $\lambda_{i..}$ etc, denote the corresponding expected frequencies.

The (saturated) log-linear model is:

$$\log(\lambda_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk}$$

A saturated model is one that has as many (independent) parameters as there are observations $\quad \Rightarrow \ \hat{\lambda}_{ijk} = y_{ijk}$

- Suppose factors A, B, C have I, J, K levels, respectively. Then there will be $1 + I + J + K + IJ + IK + JK$ linear constraints among all parameters in the saturated log-linear model.

§5.9 Hierarchical log-linear models and their notations

Note: Inclusion of a term implies inclusion of all lower-order ‘relatives’.

Notation for hierarchical models

$[ABC]$	$\log(\lambda_{ijk}) = \mu+\alpha_i+\beta_j+\gamma_k+(\alpha\beta)_{ij}+(\alpha\gamma)_{ik}+(\beta\gamma)_{jk}+(\alpha\beta\gamma)_{ijk}$
$[AB][AC][BC]$	$\log(\lambda_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk}$
$[AB][AC]$	$\log(\lambda_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik}$
$[AB][C]$	$\log(\lambda_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij}$
$[A][B][C]$	$\log(\lambda_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k$

§5.10 Interpreting log-linear models w.r.t. independence

In most cases it is possible to express the various forms of independence in terms of a log-linear model.

(a)	A, B and C mutually independent:	$[A][B][C]$
(b)	A independent of B and C together:	$[A][BC]$
(c)	A independent of B for each level of C:	$[AC][BC]$
(d)	A independent of B, marginally over C:	$[A][B]$ model
(e)	no three-factor interaction: (Also called **homogeneous association model**)	$[AB][AC][BC]$

Interpretation of the three-factor interaction

- No three-factor interaction implies that the association between any two factors, say A and B, is the same for each level of the third factor, say C. Hence the resultant model is often called the **homogeneous association model**.
- For example, for a table with factors `gender`, `smoking` and `lung-cancer`, no three-way interaction would imply that the odds ratio in favour of a smoker dying of lung cancer is the same for males and females, even though the incidence of lung cancer might be substantially different for male and female smokers.

Example: Car preferences (as a 3-way contingency table)

```
> carpref.dat[, c(1:4)]

  gender residence     pref count
1   male      city    local   168
2   male   country    local    32
3 female      city    local    84
4 female   country    local   164
5   male      city imported    68
6   male   country imported    12
7 female      city imported    16
8 female   country imported    24
```

```
> deviance(glm(count ~ gender + residence + pref, family = poisson,
+      data = carpref.dat))   #model  [G][R][P]

[1] 172.0275

> deviance(glm(count ~ gender + residence * pref, family = poisson,
+      data = carpref.dat))  #model  [G][RP]

[1] 164.4114

> deviance(glm(count ~ gender * residence + pref, family = poisson,
+      data = carpref.dat))  #model  [GR][P]

[1] 19.23630

> deviance(glm(count ~ gender * pref + residence, family = poisson,
+      data = carpref.dat))  #model  [GP][R]

[1] 153.3955

> deviance(glm(count ~ gender * pref + pref * residence,
+      family = poisson, data = carpref.dat))  #model  [GP][PR]

[1] 145.7794
```

```
> deviance(glm(count ~ gender * residence + pref * residence,
+     family = poisson, data = carpref.dat))  #model [GR][PR]

[1] 11.62017

> deviance(glm(count ~ gender * residence + pref * gender,
+     family = poisson, data = carpref.dat))  #model [GR][PG]

[1] 0.6043126

> deviance(glm(count ~ gender * residence + pref * gender +
+     residence * pref, family = poisson, data = carpref.dat))
                 #model [GR][PG][RP]

[1] 0.1351234
```

G = gender; R = residence and P = preference.

Analysis of deviance table

Model	deviance	df
1. [G][R][P]	172.0	4
2. [G][RP]	164.4	3
3. [GR][P]	19.2	3
4. [GP][R]	153.4	3
5. [GP][PR]	145.8	2
6. [GR][PR]	11.6	2
7. [GR][PG]	0.60	2
8. [GR][PG][RP]	0.14	1

- Based on the chi-square goodness of fit test, the only models that provide an adequate fit to the data are $[GR][PG]$ and $[GR][PG][RP]$.
- Further, $[GR][PG][RP]$ is not significantly better than $[GR][PG]$ hence we conclude that $[GR][PG]$ is the most appropriate model.
- The interpretation of this model $[GR][PG]$ is that preference depends on gender, **but**, given gender, preference is independent of residence.

Example: Occupation, education and aptitude

- Occupation (1 = self-employed, business; 2 = teacher; 3 = self-employed, professional; 4 = salary-employed);
- Education (4 levels);
- Aptitude (5 levels).

		E			
O	A	1	2	3	4
1	1	42	55	22	3
	2	72	82	60	12
	3	90	106	85	25
	4	27	48	47	8
	5	8	18	19	5
2	1	0	0	1	19
	2	0	3	3	60
	3	1	4	5	86
	4	0	0	2	36
	5	0	0	1	14
3	1	1	2	8	19
	2	1	2	15	33
	3	2	5	25	83
	4	2	2	10	45
	5	0	0	12	19
4	1	172	151	107	42
	2	208	198	206	92
	3	279	271	331	191
	4	99	126	179	97
	5	36	35	99	79

```
> deviance(glm(oea.y ~ o.f + a.f + e.f, family = poisson,
+     data = oea.dat))  #model [O][A][E]

[1] 1356.970

> deviance(glm(oea.y ~ o.f + a.f * e.f, family = poisson,
+     data = oea.dat))  #model [O][AE]

[1] 1179.640

> deviance(glm(oea.y ~ o.f * a.f + e.f, family = poisson,
+     data = oea.dat))  #model [OA][E]

[1] 1319.561

> deviance(glm(oea.y ~ o.f * e.f + a.f, family = poisson,
+     data = oea.dat))   #model [OE][A]

[1] 228.2215
```

```
> deviance(glm(oea.y ~ o.f * e.f + o.f * a.f, family = poisson,
+     data = oea.dat))  #model [OE][OA]

[1] 190.8123

> deviance(glm(oea.y ~ o.f * e.f + e.f * a.f, family = poisson,
+     data = oea.dat)) #model [OE][EA]

[1] 50.89152

> deviance(glm(oea.y ~ o.f * a.f + e.f * a.f, family = poisson,
+     data = oea.dat))  #model [OA][EA]

[1] 1142.231

> deviance(glm(oea.y ~ (o.f + a.f + e.f)^2, family = poisson,
+     data = oea.dat))  #model [OA][EA][OE]

[1] 25.10476
```

Analysis of Deviance results

Model	deviance	df
1. [O][A][E]	1357	69
2. [O][AE]	1180	57
3. [OA][E]	1320	57
4. [OE][A]	228	60
5. [OE][OA]	191	48
6. [OE][EA]	51	48
7. [OA][EA]	1142	45
8. [OA][EA][OE]	25	36

- The model $[OE][EA]$ (in which given education, occupation is independent of aptitude) provides a good fit to the data (with p-value of 0.36),
- but the model $[OE][OA][EA]$ provides a significantly better fit ($\Delta D = 51 - 25$, $\Delta df = 48 - 36$, p-value $=$0.0107).

Try looking at the residuals for the model $[OE][EA]$.

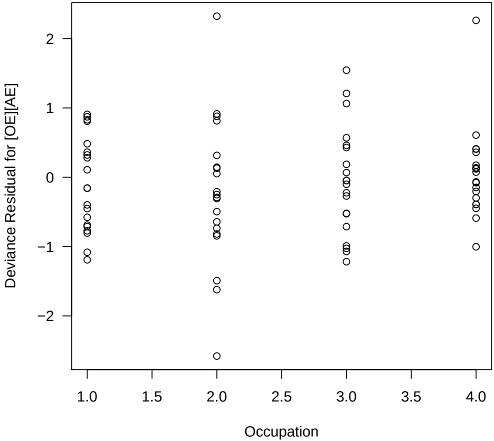


Figure 5.1: Deviance residuals against occupation.

Most of the large residuals come from teachers (code 2). What happens when teachers are omitted?

```
> no.teachers <- glm(oea.y ~ o.f * e.f + a.f * e.f + o.f *
+     a.f, family = poisson, data = oea.dat, subset = (o.f !="2"))
```

```
> anova(no.teachers, test = "Chisq")

Analysis of Deviance Table

Model: poisson, link: log

Response: oea.y

Terms added sequentially (first to last)

        Df Deviance Resid. Df Resid. Dev  P(>|Chi|)
NULL                       59     4773.1
o.f      2   2955.7        57     1817.4        0.0
e.f      3    121.9        54     1695.5  3.013e-26
a.f      4    925.1        50      770.4 5.981e-199
o.f:e.f  6    560.8        44      209.5 6.490e-118
e.f:a.f 12    184.4        32       25.2  5.438e-33
o.f:a.f  8      8.6        24       16.6        0.4
```

Occupation depends on education but, given education level, is independent of aptitude, especially for occupations other than teaching.

§5.11 Collapsibility

- It is often desirable to collapse higher order tables (> 2 factors) over one or more factors
  — to simplify the analysis, and subsequent interpretation.
- In some cases collapsing is a reasonable thing to do; in others it is totally inappropriate.
- The results provided in this section which focus upon collapsing over one factor in a three-way table, can readily be extended to cover tables of higher dimension.

A three-way table with factors A, B and C

- If we collapse over factor C (say), to produce a two-way table with factors A and B, then the only association we can investigate is the association between factors A and B.

- It can be shown that the association between factors A and B in the collapsed table will be the same as in the three-way table if, and only if, in the three-way table:

  (i) there is no three-factor interaction, and

  (ii) at least one of the two-factor interactions involving C is zero.

- In practice we interpret no interaction as meaning negligible interaction.

Example: Car preferences (continued)

E.g. $152.79 = D([PG][R]) - D([GR][PG]) = 153.396 - 0.604 = D([GR])$.

	Regression parameter estimates (standard errors) deviance change by dropping the term		
Model	$G:R$	$P:G$	$R:P$
$[GR][PG][RP]$	–2.29 (0.21) 145.64	–0.82 (0.25) 11.48	0.18 (0.26) 0.47
$[GR][PG]$	–2.31 (0.21) 152.79	–0.91 (0.22) 18.63	0.00 (0.00) 0.00
Collapsed tables			
over gender $[RP]$			0.60 (0.22) 7.62
over residence $[PG]$		–0.91 (0.22) 18.63	
over preference $[GR]$	–2.31 (0.21) 152.79		

- For the model $[GR][PG][RP]$ there is significant association between `gender` and `residence` and between `gender` and `preference`, but not between `residence` and `preference`.
- Omitting the $R : P$ interaction we find that the estimates of the other interaction parameters are different, but only slightly.
- Looking at the estimates of the interaction parameters obtained from the collapsed tables we find that we get exactly the same values for the estimates of $G : R$ and $P : G$ as we did for $[GR][PG]$, but that we get a substantially different value for $R : P$.
- This implies that it is OK to collapse over residence (or preference), but that it is not OK to collapse over gender.
- The parameter estimates, their standard errors, and the deviances that are obtained from the collapsed 2-way tables, can be obtained by fitting appropriate models to the (full) 3-way table.

§5.12 Model selection (1)

- A trade-off between adequacy of fit and simplicity.
- Usually want the simplest model(s) which provide a good fit to the data.
- The number of possible models is often tens to hundreds of thousands.
- There will often be a number of models that are not possible to choose between on statistical testing.
- Only pairs of models for which one is nested within the other can be formally compared.
- There is no all-purpose, best method of model selection.
- Stepwise procedures are available to implement a model selection criterion.

§5.12 Model selection (2)

- Often appropriate to consider only a restricted range of models — depending on the study design. E.g. include single response factor and several predictor factors only. In this case, it is not much meaningful to study whether the predictors are independent of each other conditional on the response factor.
- Sometimes the design of the study implies that certain terms should be in any model considered — the **minimal** model.

Approaches to model selection (1)

1. With one response factor (R) and explanatory factors (E1, E2, . . ., Ek), start with [E1 E2 . . . Ek][R] and add interactions between R and the Ej's until you find an adequate fit (or fits).
   Approach extends in an "obvious" way if there is more than one response factor.

Approaches to model selection (2)

2 Fit the sequence of models:

   1. main effects only
   2. all two-factor interactions
   3. all three-factor interactions
   4. etc

   - Determine the simplest such model that provides a good fit to the data.
   - Drop interaction terms from that model until the smallest (hierarchical) model that provides an adequate fit is obtained.

Approaches to model selection (3)

3. Use the "*Akaike Information Criterion*" (AIC):

   $$\text{AIC} = \text{(residual) deviance} + 2k$$

   where k is the number of (linearly independent) parameters for the model.
   (The residual deviance has $n - k$ degrees of freedom).
   The model with the smallest AIC is preferred.
   This is what R uses with the **step** command.

4. Or use "*Schwarz Information criterion*" (SIC or BIC):

   $$\text{BIC} = \text{(residual) deviance} + k \log n$$

   where n is the number of sample units in the data.

Ex: Compliance with a screening test for bowel cancer

		Education Level								
		Primary			Secondary			Tertiary		
		C	N	R	C	N	R	C	N	R
Smokers	O	3	11	4	4	19	7	1	2	2
	E	1.4	12.7	3.9	5.0	17.2	7.8	1.6	2.1	1.4
	%	8	71	21	17	57	26	32	41	27
Non-	O	13	26	4	34	23	5	16	3	1
smokers	E	14.1	25.2	3.7	34.0	23.0	5.0	14.9	3.8	1.2
	%	33	58	9	55	37	8	75	19	6

Legend:

C = compliance; N = non-compliance; R = refuse to participate

O = observed frequency; E = expected frequency;

% = expected frequency as a percentage for that group of subjects (e.g. primary educated smokers)

Factors: S = smoking; EL = education level; RES = response (C, N or R)

Variable: $x = EL$

Model	Scaled Deviance	df	AIC
$S + EL + RES + S : EL$	40.36	10	56.36
$S + EL + RES + S : EL + EL : RES$	25.76	6	49.76
$S + EL + RES + S : EL + S : RES$	16.71	8	36.71
$S + EL + RES + S : EL + S : RES + EL : RES$	2.61	4	30.61
$S + EL + RES + S : EL + S : RES + RES : x^*$	3.04	6	27.04

* model used to calculate expected frequencies

§5.13 Uniform association

Consider a two-way table with factors R and C, both of which are **ordinal**. Let $x = r \times c$, where r and c denote levels of R and C, then the (log-linear) model $R + C + x$ is the model for uniform association.

Interpretation:
If β denotes the coefficient of x in the model then, for all adjacent 2×2 sub-tables, the odds ratio $= \exp(\beta)$, i.e. for all adjacent 2×2 sub-tables, the odds ratio is constant (uniform) throughout the table.
(It is assumed here that r and c are unit spaced integers.)

Expressions of βx in log-linear model $R + C + x$

	C				
R	1	2	3	4	5
1	β	2β	3β	4β	5β
2	2β	4β	6β	8β	10β
3	3β	6β	9β	12β	15β
4	4β	8β	12β	16β	20β
5	5β	10β	15β	20β	25β

Example: Income and job satisfaction revisited (1)

	Job satisfaction				
Income ($)	Very Dis.	Mod. Dis.	Mod. Sat.	Very Sat.	Total
< 6000	20	24	80	82	206
6000 — 15000	22	38	104	125	289
15000 — 25000	13	28	81	113	235
> 25000	7	18	54	92	171
Total	62	108	319	412	901

Example: Income and job satisfaction revisited (2)

```
> y <- c(20, 22, 13, 7, 24, 38, 28, 18, 80, 104, 81, 54,
+     82, 125, 113, 92)
> inc <- rep(1:4, 4)
> sat <- rep(1:4, c(4, 4, 4, 4))
> jobsat.1 <- glm(y ~ factor(inc) + factor(sat), family = poisson)
```

Example: Income and job satisfaction revisited (3)

```
> anova(jobsat.1)

Analysis of Deviance Table

Model: poisson, link: log

Response: y

Terms added sequentially (first to last)


            Df Deviance Resid. Df Resid. Dev
NULL                           15     445.76
factor(inc)  3    32.92        12     412.84
factor(sat)  3   400.81         9      12.04

> 1 - pchisq(12.04, 9)

[1] 0.2110637
```

Try the uniform association model

```
> jobsat.2 <- glm(y ~ factor(inc) + factor(sat) + I(inc *
+     sat), family = poisson)
> anova(jobsat.2, test = "Chi")
```

```
Analysis of Deviance Table

Model: poisson, link: log

Response: y

Terms added sequentially (first to last)


             Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                            15     445.76
factor(inc)   3    32.92        12     412.84 3.349e-07
factor(sat)   3   400.81         9      12.04 1.481e-86
I(inc * sat)  1     9.65         8       2.39 1.893e-03
```

```
> summary(jobsat.2)$coef

                   Estimate Std. Error     z value     Pr(>|z|)
(Intercept)      2.85061005 0.14825615 19.22760010 2.174611e-82
factor(inc)2 -0.01053008 0.14343926 -0.07341142 9.414787e-01
factor(inc)3 -0.57692154 0.24745025 -2.33146475 1.972887e-02
factor(inc)4 -1.26388882 0.36668853 -3.44676397 5.673440e-04
factor(sat)2  0.30752526 0.17547363  1.75254402 7.968031e-02
factor(sat)3  1.13006167 0.20826080  5.42618528 5.757120e-08
factor(sat)4  1.11197428 0.28091454  3.95840767 7.545110e-05
I(inc * sat)  0.11193941 0.03640759  3.07461716 2.107729e-03
```

- Uniform association odds ratio $= e^{0.1119} = 1.118$.
- **Interpretation:** If income increases by one level then the odds of being one level more satisfied increase by a factor of 1.118 (i.e. by 11.8%).
- The independence (or no association) model `jobsat.1`, even though having no evidence against its goodness of fit, is not powerful enough to reveal the >1 odds ratio. The odds ratio equals 1 based on the independence model.

§5.14 Case study: Occupational mobility

Example: Occupational status of fathers and sons
Occupations of a sample of fathers and sons (UK data), with occupation categorised according to status (1 = high; 5 = low).

	Sons				
Fathers	1	2	3	4	5
1	50	45	8	18	8
2	28	174	84	154	55
3	11	78	110	223	96
4	14	150	185	714	447
5	3	42	72	320	411

R analysis

```
> Omobility.dat <- data.frame(son = rep(1:5, 5), father = rep(1:5,
+     rep(5, 5)), freq = c(50, 45, 8, 18, 8, 28, 174, 84,
+     154, 55, 11, 78, 110, 223, 96, 14, 150, 185, 714,
+     447, 3, 42, 72, 320, 411))
> mobility.1 <- glm(freq ~ factor(father) + factor(son),
+     family = poisson, data = Omobility.dat)
```

ANOVA

```
> anova(mobility.1)

Analysis of Deviance Table

Model: poisson, link: log

Response: freq

Terms added sequentially (first to last)


               Df Deviance Resid. Df Resid. Dev
NULL                              24     4008.8
factor(father)  4   1555.7        20     2453.1
factor(son)     4   1660.9        16      792.2
```

Residual matrix

```
> matrix(resid(mobility.1, type = "deviance"), nrow = 5,
+     ncol = 5, byrow = T)
          [,1]       [,2]       [,3]       [,4]       [,5]
[1,] 12.757064  5.3288659 -2.4191759 -5.5395146 -5.8528616
[2,]  2.994580 10.5546430  2.2648569 -3.5331563 -8.4806046
[3,] -1.251392  0.6532514  4.6795401  0.7843029 -4.7629329
[4,] -5.506220 -4.4288802 -0.9360735  3.8290953  0.3920227
[5,] -5.699249 -8.1143568 -3.9767817 -1.4278538  9.5582160
```

Extract 1,1

```
> f1s1 <- ifelse(Omobility.dat$father == 1 & Omobility.dat$son ==
+      1, 1, 0)
> matrix(f1s1, nrow = 5, ncol = 5, byrow = T)

     [,1] [,2] [,3] [,4] [,5]
[1,]    1    0    0    0    0
[2,]    0    0    0    0    0
[3,]    0    0    0    0    0
[4,]    0    0    0    0    0
[5,]    0    0    0    0    0

> mobility.2 <- glm(freq ~ factor(father) + factor(son) +
+      f1s1, family = poisson, data = Omobility.dat)
```

ANOVA

```
> anova(mobility.2)

Analysis of Deviance Table

Model: poisson, link: log

Response: freq

Terms added sequentially (first to last)


               Df Deviance Resid. Df Resid. Dev
NULL                              24     4008.8
factor(father)  4   1555.7        20     2453.1
factor(son)     4   1660.9        16      792.2
f1s1            1    207.9        15      584.3
```

Residual matrix

```
> matrix(resid(mobility.2, type = "deviance"), nrow = 5,
+     ncol = 5, byrow = T)
              [,1]       [,2]       [,3]       [,4]        [,5]
[1,] -4.712161e-08  7.5153033 -0.8596712 -2.9014410 -3.73984873
[2,]  5.390798e+00 10.4152837  2.1466084 -3.7242494 -8.62766823
[3,]  7.823366e-01  0.5323405  4.5533949  0.5783768 -4.92360055
[4,] -2.416535e+00 -4.6221096 -1.1313759  3.4717920  0.09712822
[5,] -3.590891e+00 -8.2466844 -4.1151804 -1.6856465  9.31658204
```

Mess with Weights Instead - the same

```
> mobility.2 <- glm(freq ~ factor(father) + factor(son),
+     weight = (1 - f1s1), family = poisson, data = Omobility.dat)
> matrix(resid(mobility.2, type = "deviance"), nrow = 5,
+     ncol = 5, byrow = T)

           [,1]       [,2]       [,3]       [,4]        [,5]
[1,]  0.0000000  7.5153033 -0.8596712 -2.9014410 -3.73984873
[2,]  5.3907982 10.4152837  2.1466084 -3.7242494 -8.62766823
[3,]  0.7823366  0.5323405  4.5533949  0.5783768 -4.92360055
[4,] -2.4165350 -4.6221096 -1.1313759  3.4717920  0.09712822
[5,] -3.5908908 -8.2466844 -4.1151804 -1.6856465  9.31658204
```

Try a Diagonal Term

```
> diag <- ifelse(Omobility.dat$father == Omobility.dat$son,
+     1, 0)
> matrix(diag, nrow = 5, ncol = 5, byrow = T)

     [,1] [,2] [,3] [,4] [,5]
[1,]    1    0    0    0    0
[2,]    0    1    0    0    0
[3,]    0    0    1    0    0
[4,]    0    0    0    1    0
[5,]    0    0    0    0    1

> mobility.3 <- glm(freq ~ factor(father) + factor(son) +
+     diag, family = poisson, data = Omobility.dat)
```

ANOVA

```
> anova(mobility.3, test = "Chisq")

Analysis of Deviance Table

Model: poisson, link: log

Response: freq

Terms added sequentially (first to last)


               Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                              24     4008.8
factor(father)  4   1555.7        20     2453.1       0.0
factor(son)     4   1660.9        16      792.2       0.0
diag            1    316.2        15      476.0 9.66e-71
```

Curses! Where the heck is that LoF??

```
> matrix(resid(mobility.3, type = "deviance"), nrow = 5,
+     ncol = 5, byrow = T)

          [,1]       [,2]      [,3]       [,4]      [,5]
[1,]  9.503123  5.0468280 -2.598090 -4.7414410 -5.798369
[2,]  2.778147  3.7357323  2.718397 -0.4827347 -7.319801
[3,] -1.471663  0.9817873 -1.558603  3.9596568 -3.588880
[4,] -4.841539 -1.5822965  2.038829 -3.4133001  6.027807
[5,] -5.444961 -6.6397925 -2.363667  4.5998085  1.334610
```

Everywhere! Get 1,1 again as well

```
> mobility.4 <- glm(freq ~ factor(father) + factor(son) +
+     diag + f1s1, family = poisson, data = Omobility.dat)
> deviance(mobility.4)

[1] 343.8607

> matrix(resid(mobility.4, type = "deviance"), nrow = 5,
+     ncol = 5, byrow = T)

              [,1]      [,2]      [,3]      [,4]      [,5]
[1,] -2.107342e-08  7.124104 -1.114342 -2.484878 -3.853167
[2,]  5.058285e+00  4.378936  2.487446 -1.063596 -7.654731
[3,]  4.807584e-01  0.765404 -0.965491  3.347611 -3.942329
[4,] -2.138734e+00 -2.140135  1.459872 -2.905815  5.063660
[5,] -3.556414e+00 -6.998082 -2.749780  3.617850  2.007667
```

It's not enough! We need more power!

```
> dist <- c(0, 1, 2, 3, 4, 1, 0, 1, 2, 3, 2, 1, 0, 1, 2,
+     3, 2, 1, 0, 1, 4, 3, 2, 1, 0)
> matrix(dist, nrow = 5, ncol = 5, byrow = T)

     [,1] [,2] [,3] [,4] [,5]
[1,]    0    1    2    3    4
[2,]    1    0    1    2    3
[3,]    2    1    0    1    2
[4,]    3    2    1    0    1
[5,]    4    3    2    1    0

> mobility.5 <- glm(freq ~ factor(father) + factor(son) +
+     dist + factor(dist), family = poisson, data = Omobility.dat)
```

A small advance

```
> anova(mobility.5)

Analysis of Deviance Table

Model: poisson, link: log

Response: freq

Terms added sequentially (first to last)


               Df Deviance Resid. Df Resid. Dev
NULL                              24     4008.8
factor(father)  4   1555.7        20     2453.1
factor(son)     4   1660.9        16      792.2
dist            1    699.2        15       93.0
factor(dist)    3     39.0        12       54.0
```

Curses! Where the heck is that LoF??

```
> matrix(resid(mobility.5, type = "deviance"), nrow = 5,
+     ncol = 5, byrow = T)

           [,1]       [,2]       [,3]       [,4]       [,5]
[1,]  4.3563240 -0.7327473 -3.4390485 -1.4076868  0.3409451
[2,] -1.9638355  0.1164528 -0.4321956  0.7950587  0.6313393
[3,] -1.9104307 -1.6871941  0.5309160  1.6608222 -0.5896614
[4,] -1.1119299  1.0717728  0.5553674 -0.8859217  0.3959667
[5,] -0.4877915  1.0915308  0.6167783 -0.1650866 -0.3872946

> mobility.6 <- glm(freq ~ factor(father) + factor(son) +
+     f1s1 + factor(dist), family = poisson, data = Omobility.dat)
```

Better

```
> anova(mobility.6)
Analysis of Deviance Table

Model: poisson, link: log

Response: freq

Terms added sequentially (first to last)


               Df Deviance Resid. Df Resid. Dev
NULL                              24     4008.8
factor(father)  4   1555.7        20     2453.1
factor(son)     4   1660.9        16      792.2
f1s1            1    207.9        15      584.3
factor(dist)    4    568.2        11       16.1
```

Hmmm

```
> summary(mobility.6)$coef

                  Estimate Std. Error    z value     Pr(>|z|)
(Intercept)      2.4214033 0.19454423  12.446544 1.460329e-35
factor(father)2  1.2049374 0.12693965   9.492206 2.261956e-21
factor(father)3  0.9777009 0.12805064   7.635268 2.253519e-14
factor(father)4  1.9020692 0.12439394  15.290690 8.821099e-53
factor(father)5  1.4615041 0.12654356  11.549415 7.432426e-31
factor(son)2     1.5262047 0.14499938  10.525594 6.584451e-26
factor(son)3     1.2301726 0.14596178   8.428046 3.514856e-17
factor(son)4     2.2661544 0.14168867  15.993899 1.409268e-57
factor(son)5     2.1238884 0.14217149  14.938920 1.839330e-50
f1s1             1.4906197 0.24051498   6.197617 5.732449e-10
factor(dist)1   -0.3504798 0.03919791  -8.941288 3.846612e-19
factor(dist)2   -0.8890118 0.05497953 -16.169868 8.227381e-59
factor(dist)3   -1.7218437 0.09771157 -17.621697 1.678927e-69
factor(dist)4   -2.5632214 0.31893052  -8.036927 9.211937e-16
```

Try squaring the Distances

```
> mobility.7 <- glm(freq ~ factor(father) + factor(son) +
+     f1s1 + dist + I(dist^2), family = poisson, data = Omobility.dat)
```

MUCH better

```
> anova(mobility.7)

Analysis of Deviance Table

Model: poisson, link: log

Response: freq

Terms added sequentially (first to last)


               Df Deviance Resid. Df Resid. Dev
NULL                              24     4008.8
factor(father)  4   1555.7        20     2453.1
factor(son)     4   1660.9        16      792.2
f1s1            1    207.9        15      584.3
dist            1    538.3        14       46.1
I(dist^2)       1     29.5        13       16.6
```

Try the diagnoal as well

```
> mobility.8 <- glm(freq ~ factor(father) + factor(son) +
+     f1s1 + dist + I(dist^2) + diag, family = poisson,
+     data = Omobility.dat)
```

It's not needed

```
> anova(mobility.8)

Analysis of Deviance Table

Model: poisson, link: log

Response: freq

Terms added sequentially (first to last)


               Df Deviance Resid. Df Resid. Dev
NULL                              24     4008.8
factor(father)  4   1555.7        20     2453.1
factor(son)     4   1660.9        16      792.2
f1s1            1    207.9        15      584.3
dist            1    538.3        14       46.1
I(dist^2)       1     29.5        13       16.6
diag            1      0.1        12       16.5
```

Ok time for a glass of wine

```
> summary(mobility.7)$coef

                  Estimate Std. Error   z value     Pr(>|z|)
(Intercept)      2.4201373 0.18898988 12.805645 1.524537e-37
factor(father)2  1.2039282 0.12347851  9.750104 1.842801e-22
factor(father)3  0.9784189 0.12593904  7.768988 7.911588e-15
factor(father)4  1.9013648 0.12213723 15.567447 1.211464e-54
factor(father)5  1.4598416 0.12436349 11.738506 8.090336e-32
factor(son)2     1.5280595 0.14308988 10.679019 1.276133e-26
factor(son)3     1.2319248 0.14502883  8.494344 1.990532e-17
factor(son)4     2.2673732 0.14050013 16.137873 1.382231e-58
factor(son)5     2.1239566 0.14129854 15.031695 4.552091e-51
f1s1             1.4918857 0.23604486  6.320348 2.609745e-10
dist            -0.2376327 0.05196032 -4.573351 4.799860e-06
I(dist^2)       -0.1076130 0.02034376 -5.289729 1.224976e-07
```

Summary

Model	Residual deviance	df
F + S	792.2	16
F + S + f1s1	584.3	15
F + S + diag	476.0	15
F + S + diag + f1s1	343.9	14
F + S + dist	93.0	15
F + S + dist + f1s1	46.1	14
F + S + dist.f	54.0	12
F + S + dist.f + f1s1	16.1	11
F + S + dist + dist^2 + f1s1	16.6	13
F + S + dist + dist^2 + f1s1 + diag	16.5	12

Standardisation to fixed totals

- A useful way to compare two or more tables. The cell frequencies are manipulated so as to retain the association structure in the original tables while producing a desired set of marginal totals.
- **Detail:** Entries for the standardised table are obtained by expressing the cell values as percentages of the row totals, then expressing these percentages as the percentages of the resultant column totals, then expressing these values as percentages of the resultant row totals, and so on until the process converges.

Example: Occupational mobility in the UK and Denmark

Standardisation to fixed totals

Original data:

			Sons			
Fathers	1	2	3	4	5	Total
1	50 18	45 17	8 16	18 4	8 2	129 57
2	28 24	174 105	84 109	154 59	55 21	495 318
3	11 23	78 84	110 289	223 217	96 95	518 708
4	14 8	150 49	185 175	714 348	447 198	1510 778
5	3 6	42 8	72 69	320 201	411 246	848 530
Totals	106 79	489 263	459 658	1429 829	1017 562	

Upper number in cell — UK data.
Lower number in cell — Danish data.

Standardised data:

Fathers	Sons					Total
	1	2	3	4	5	
1	68.5	20.9	4.6	3.7	2.3	100
	58.6	25.0	13.0	2.6	1.8	100
2	17.8	37.5	22.5	14.7	7.5	100
	21.1	41.6	21.9	10.3	5.1	100
3	8.0	19.2	33.7	24.3	14.9	100
	11.7	19.3	33.7	21.9	13.5	100
4	4.1	14.7	22.6	31.1	27.6	100
	4.1	11.4	20.7	35.5	28.4	100
5	1.6	7.8	16.6	26.2	47.8	100
	4.5	2.7	11.8	29.8	51.2	100
Totals	100	100	100	100	100	
	100	100	100	100	100	

§5.15 Incomplete tables

- A contingency table may be incomplete in the sense that some cells in the table have zero value or missing value, with the missing value being coded as 0 sometimes.
- Therefore, it is important to distinguish between two types of zeros: **Sampling and structural zeros**
  - **sampling zeros** — due to sampling variability, thus are treated in the same way as the other values in the table.
  - **structural zeros** — cells known to have zero values, a priori, or having missing values. E.g. male patients in a maternity ward.
- In this section we provide two applications of log-linear model in cases involving structure zeros: **quasi independence** and **capture-recapture experiments**.

Quasi independence

- For a two-way table, independence $\Rightarrow \lambda_{ij} = a_i b_j$, for suitable a_i, b_j.
- If this equation holds only for S — a subset of cells of the table not containing any structural zeros, then row and column factors of S are said to be **quasi independent**.
- An independence log-linear model applied to S can assess the effect of quasi independence.
- One may also fit the quasi independence model to disjoint subsets of complete tables for which overall independence is inappropriate.

Example: Purum marriages

The table has structure zeros, thus can be analysed by a quasi independence model

	Sib of Husband				
Sib of Wife	Marrim	Makan	Parpa	Thao	Kheyang
Marrim	–	5	17	–	6
Makan	5	–	0	16	2
Parpa	–	2	–	10	11
Thao	10	–	–	–	9
Kheyang	6	20	8	0	1

R analysis

```
> purum <- read.csv("../data/purum.csv"); purum$wife <- factor(purum$wife)
> purum$husband <- factor(purum$husband); purum
   freq wife husband wt
1    99    1       1  0
2     5    2       1  1
3    99    3       1  0
4    10    4       1  1
5     6    5       1  1
6     5    1       2  1
7    99    2       2  0
8     2    3       2  1
9    99    4       2  0
10   20    5       2  1
11   17    1       3  1
12    0    2       3  1
13   99    3       3  0
14   99    4       3  0
15    8    5       3  1
16   99    1       4  0
17   16    2       4  1
18   10    3       4  1
19   99    4       4  0
20    0    5       4  1
21    6    1       5  1
```

A first try

```
> purum.1 <- glm(freq ~ wife + husband, family = poisson,
+      weight = wt, data = purum)
> anova(purum.1)

Analysis of Deviance Table

Model: poisson, link: log

Response: freq

Terms added sequentially (first to last)


        Df Deviance Resid. Df Resid. Dev
NULL                       16     86.479
wife     4    4.187        12     82.292
husband  4    6.041         8     76.251
```

A first try

```
> matrix(resid(purum.1, type = "deviance"), nrow = 5, ncol = 5,
+      byrow = F)

            [,1]      [,2]        [,3]       [,4]       [,5]
[1,]  0.00000000 -1.970226  1.78021524  0.0000000 -0.2141023
[2,]  0.06252835  0.000000 -3.62417090  2.6715392 -1.1182067
[3,]  0.00000000 -2.651545  0.00000000  0.1489020  2.2695727
[4,] -0.12381948  0.000000  0.00000000  0.0000000  0.1342204
[5,]  0.10694455  3.627028  0.08695497 -4.2226076 -2.0942630
```

- We conclude that, excluding those disallowed marriage types, some combinations of sibs are more favoured than others. The more favoured combinations are those with large, positive residuals; the less favoured combinations are those with large, negative residuals.
- The model need be further refined to obtain more insights, but not pursued here.

Capture-recapture experiments

Example: Comparison of two lists
The following data were obtained from two hospital lists. What is wanted is an estimate of the total number of patients or, equivalently, the number of patients absent from both lists.

	Interviewer's List	
Registrar's List	Present	Absent
Present	794	710
Absent	741	?

R analysis

```
> list.freq <- c(794, 741, 710, 0)
> reg.f <- factor(c(1, 2, 1, 2))
> int.f <- factor(c(1, 1, 2, 2))
> list.wt <- c(1, 1, 1, 0)
```

R analysis

```
> anova(list.1 <- glm(list.freq ~ reg.f + int.f, family = poisson,
+     weight = list.wt))

Analysis of Deviance Table

Model: poisson, link: log

Response: list.freq

Terms added sequentially (first to last)


      Df Deviance Resid. Df Resid. Dev
NULL                      2     4.8019
reg.f  1   0.1080         1     4.6939
int.f  1   4.6939         0 -1.616e-13
```

R analysis

```
> cbind(list.freq, fitted(list.1))

  list.freq
1       794 794.000
2       741 741.000
3       710 710.000
4         0 662.607
```

Example: Trapping of cottontail rabbits

- P: presence; A: absence; First half-year); Second half-year).
- Each cell value is the number of rabbits shot/not shot by hunters and trapped/not trapped by rangers in each half-year of a year.

			Live Trappings			
		First half	P	P	A	A
Year	Hunters	Second half	P	A	P	A
1965	P		8	26	32	189
	A		8	41	41	–
1967	P		9	5	4	28
	A		30	17	22	–
1968	P		9	3	4	41
	A		31	41	29	–
1970	P		4	6	22	59
	A		15	17	32	–

R analysis

We aim to estimate the numbers missing from the table.

```
> rabbit.freq <- c(8, 26, 32, 189, 8, 41, 41, 0, 9, 5,
+     4, 28, 30, 17, 22, 0, 9, 3, 4, 41, 31, 41, 29, 0,
+     4, 6, 22, 59, 15, 17, 32, 0)
> year.f <- factor(rep(1:4, c(8, 8, 8, 8)))
> hunt.f <- factor(rep(rep(1:2, c(4, 4)), 4))
> first.f <- factor(rep(rep(1:2, c(2, 2)), 8))
> second.f <- factor(rep(1:2, 16))
> rabbit.wt <- rep(c(1, 1, 1, 1, 1, 1, 1, 0), 4)
> head(cbind(rabbit.freq, year.f, hunt.f, first.f, second.f,
+     rabbit.wt))
```

```
     rabbit.freq year.f hunt.f first.f second.f rabbit.wt
[1,]           8      1      1       1        1         1
[2,]          26      1      1       1        2         1
[3,]          32      1      1       2        1         1
[4,]         189      1      1       2        2         1
[5,]           8      1      2       1        1         1
[6,]          41      1      2       1        2         1
```

```
> rabbit.1 <- glm(rabbit.freq ~ year.f * hunt.f * first.f *
+     second.f - year.f:hunt.f:first.f:second.f, family = poisson,
+     weight = rabbit.wt)  #This model has no 4-factor interaction
```

```
> anova(rabbit.1, test = "Chi")

Analysis of Deviance Table

Model: poisson, link: log

Response: rabbit.freq

Terms added sequentially (first to last)


                       Df Deviance Resid. Df Resid. Dev  P(>|Chi|)
NULL                                     27     731.77
year.f                  3   148.50       24     583.27  5.546e-32
hunt.f                  1     0.28       23     582.99       0.60
first.f                 1   156.54       22     426.45  6.443e-36
second.f                1   199.07       21     227.37  3.329e-45
year.f:hunt.f           3    83.71       18     143.67  4.913e-18
year.f:first.f          3    28.22       15     115.45  3.267e-06
hunt.f:first.f          1    14.18       14     101.27  1.662e-04
year.f:second.f         3    39.95       11      61.32  1.093e-08
hunt.f:second.f         1    17.41       10      43.91  3.013e-05
first.f:second.f        1    18.22        9      25.69  1.971e-05
year.f:hunt.f:first.f   3     3.51        6      22.18       0.32
year.f:hunt.f:second.f  3     8.15        3      14.03       0.04
year.f:first.f:second.f 3    14.03        0 -2.043e-14  2.861e-03
hunt.f:first.f:second.f 0     0.00        0 -2.043e-14
```

```
> step(rabbit.1)

Start:  AIC= 187.55
 rabbit.freq ~ year.f * hunt.f * first.f * second.f - year.f:hunt.f:first.f:second.

Step:  AIC= 187.55
 rabbit.freq ~ year.f + hunt.f + first.f + second.f + year.f:hunt.f +
    year.f:first.f + hunt.f:first.f + year.f:second.f + hunt.f:second.f +
    first.f:second.f + year.f:hunt.f:first.f + year.f:hunt.f:second.f +
    year.f:first.f:second.f

                          Df   Deviance     AIC
- year.f:hunt.f:second.f   3      3.347 184.899
- year.f:hunt.f:first.f    3      4.571 186.123
<none>                       -2.043e-14 187.552
- year.f:first.f:second.f  3     14.032 195.584

Step:  AIC= 184.9
 rabbit.freq ~ year.f + hunt.f + first.f + second.f + year.f:hunt.f +
    year.f:first.f + hunt.f:first.f + year.f:second.f + hunt.f:second.f +
    first.f:second.f + year.f:hunt.f:first.f + year.f:first.f:second.f

                          Df Deviance     AIC
- year.f:hunt.f:first.f    3    6.715 182.267
- hunt.f:second.f          1    5.151 184.703
<none>                          3.347 184.899
- year.f:first.f:second.f  3   22.180 197.732
```

```
Step:  AIC= 182.27
 rabbit.freq ~ year.f + hunt.f + first.f + second.f + year.f:hunt.f +
    year.f:first.f + hunt.f:first.f + year.f:second.f + hunt.f:second.f +
    first.f:second.f + year.f:first.f:second.f

                          Df Deviance     AIC
- hunt.f:first.f           1    6.853 180.405
- hunt.f:second.f          1    8.645 182.197
<none>                          6.715 182.267
- year.f:first.f:second.f  3   25.693 195.245
- year.f:hunt.f            3   38.104 207.656

Step:  AIC= 180.4
 rabbit.freq ~ year.f + hunt.f + first.f + second.f + year.f:hunt.f +
    year.f:first.f + year.f:second.f + hunt.f:second.f + first.f:second.f +
    year.f:first.f:second.f

                          Df Deviance     AIC
<none>                          6.853 180.405
- hunt.f:second.f          1    9.150 180.702
- year.f:first.f:second.f  3   26.256 193.808
- year.f:hunt.f            3   39.761 207.313
```

```
Call:  glm(formula = rabbit.freq ~ year.f + hunt.f + first.f + second.f +
  year.f:hunt.f + year.f:first.f + year.f:second.f + hunt.f:second.f +
 first.f:second.f + year.f:first.f:second.f, family = poisson, weights = rabbit.wt)

Coefficients:
               (Intercept)                     year.f2
                   1.99551                     0.15716
                   year.f3                     year.f4
                  -0.19179                    -0.09681
                   hunt.f2                    first.f2
                   0.16136                     1.51787
                 second.f2             year.f2:hunt.f2
                   1.22666                     1.10016
           year.f3:hunt.f2             year.f4:hunt.f2
                   1.55916                     0.45139
          year.f2:first.f2            year.f3:first.f2
                  -1.92334                    -1.71024
          year.f4:first.f2           year.f2:second.f2
                  -0.47333                    -2.08359
         year.f3:second.f2           year.f4:second.f2
                  -1.43741                    -1.27761
         hunt.f2:second.f2          first.f2:second.f2
                   0.35217                     0.50171
year.f2:first.f2:second.f2  year.f3:first.f2:second.f2
                   1.94023                     1.81128
year.f4:first.f2:second.f2
                   0.68354
```

```
> rabbit.2 <- glm(rabbit.freq ~ year.f * first.f * second.f +
+     year.f * hunt.f + hunt.f * second.f, family = poisson,
+     weight = rabbit.wt)
> anova(rabbit.2, test = "Chi")
```

```
Analysis of Deviance Table

Model: poisson, link: log

Response: rabbit.freq

Terms added sequentially (first to last)

                        Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                                       27     731.77
year.f                   3   148.50        24     583.27 5.546e-32
first.f                  1   154.30        23     428.97 1.987e-35
second.f                 1   156.72        22     272.24 5.883e-36
hunt.f                   1    44.87        21     227.37 2.108e-11
year.f:first.f           3    61.06        18     166.32 3.493e-13
year.f:second.f          3    74.81        15      91.51 3.978e-16
first.f:second.f         1    44.80        14      46.70 2.181e-11
year.f:hunt.f            3    19.27        11      27.43 2.403e-04
second.f:hunt.f          1     1.18        10      26.26      0.28
year.f:first.f:second.f  3    19.40         7       6.85 2.256e-04
```

```
> rabbit.3 <- glm(rabbit.freq ~ year.f * first.f * second.f +
+     year.f * hunt.f, family = poisson, weight = rabbit.wt)
> anova(rabbit.3, test = "Chi")

Analysis of Deviance Table

Model: poisson, link: log

Response: rabbit.freq

Terms added sequentially (first to last)


                        Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                                       27     731.77
year.f                   3   148.50        24     583.27 5.546e-32
first.f                  1   154.30        23     428.97 1.987e-35
second.f                 1   156.72        22     272.24 5.883e-36
hunt.f                   1    44.87        21     227.37 2.108e-11
year.f:first.f           3    61.06        18     166.32 3.493e-13
year.f:second.f          3    74.81        15      91.51 3.978e-16
first.f:second.f         1    44.80        14      46.70 2.181e-11
year.f:hunt.f            3    19.27        11      27.43 2.403e-04
year.f:first.f:second.f  3    18.28         8       9.15 3.843e-04
```

```
> print(fitted(rabbit.3) * (1 - rabbit.wt))
```

```
       1        2        3        4        5        6        7        8 
  0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000 257.7273 
       9       10       11       12       13       14       15       16 
  0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000 107.3333 
      17       18       19       20       21       22       23       24 
  0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000 258.8125 
      25       26       27       28       29       30       31       32 
  0.0000   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000 118.0000 
```

The numbers of rabbits that were neither shot nor trapped are estimated to be 258, 107, 259 and 118 for the years 1965, 1967, 1968 and 1970 respectively.