Chapter 5
Generalized Linear Models and Extensions

Abstract The generalized linear model (GLM) is reviewed and the log-linear models are integrated in this family. For GLMs, maximum likelihood estimation, model fit, and model selection are discussed. In the GLM framework the analysis of incomplete tables is more straightforward. The quasi-independence model is defined and illustrated in R. Furthermore, the family of generalized log-linear models (GLLMs) is briefly presented and a GLLM is illustrated with a representative example in R.

Keywords Generalized linear models • Exponential family • Maximum likelihood estimation • Model selection and fit • Log-linear models • Quasi independence • Multinomial Poisson homogeneous model

5.1 The Generalized Linear Model (GLM) in Keywords

Log-linear models for contingency tables are members of the family of *generalized linear models* (GLMs). The GLM is a broad class of statistical models, introduced by Nelder and Wedderburn (1972), that allows for unified consideration and treatment of many models of different types of response variables and error structures. Characteristic special cases of the GLM are the models of regression, logistic regression, Poisson regression, and the log-linear models. The GLM is an extension of the classical regression model that relates a *response variable* Y to a set of q *explanatory variables* X_j, $j = 1,\ldots,q$, by equating a function of the expected response $\mathrm{E}(Y)$ to a linear predictor based on $\mathbf{X} = (X_1,\ldots,X_q)$.

Under the classical linear regression model, if $\mathbf{y} = (y_1,\ldots,y_{n_y})'$ is a sample of size n_y of the response variable Y and $\mathbf{x} = (x_{ij})_{n_y \times q}$ is the $n_y \times q$ matrix with the corresponding sample values on the explanatory variables X_j, $j = 1,\ldots,q$, then in matrix notation we have

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\ ,$$

M. Kateri, *Contingency Table Analysis: Methods and Implementation Using R*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-4811-4_5,

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)'$ is the parameter vector and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_{n_y})'$, the vector of errors. The distributional assumptions are that (i) Y_i are independent normal distributed with $\mathrm{E}(Y_i) = \propto_i$ $(i = 1, \ldots, n_y)$ and common variance $\mathrm{Var}(Y_i) = \sigma^2$ and (ii) the errors are also independent normal distributed with zero mean and common variance σ_ε^2. In summary, the regression model has a *random component*, the response variable Y, and a *systematic component*, the linear combination of the explanatory variables $\mathbf{X}\boldsymbol{\beta}$, that links to the vector of the expected response values, i.e.,

$$\boldsymbol{\propto} = \mathrm{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} \ , \tag{5.1}$$

with $\boldsymbol{\propto} = (\propto_1, \ldots, \propto_{n_y})'$ and $\mathbf{Y} = (Y_1, \ldots, Y_{n_y})'$.

The GLM extends the regression models by relaxing the assumption about normal distributed response variable Y and by linking the systematic component not directly to $\boldsymbol{\propto}$ but to a function of it $g(\boldsymbol{\propto})$. Thus, the systematic component of the GLM is

$$\boldsymbol{\eta} = g(\boldsymbol{\propto}) = \mathbf{X}\boldsymbol{\beta} \ , \tag{5.2}$$

with $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_{n_y})'$. Function g is called the *link function*. The linear model (5.1) is a special case of (5.2) for the *identity link*, i.e. for $\boldsymbol{\eta} = g(\boldsymbol{\propto}) = \boldsymbol{\propto}$.

Under GLM, the distribution of the response Y may be any member of the *exponential family*. For univariate responses, as considered in this book, the corresponding density function is

$$f(y_i;\ \theta_i, \psi, \omega_i) = \exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{\psi}\omega_i + c(y_i, \psi, \omega_i) \right\} \ , \tag{5.3}$$

where ω_i is a weight with

$$\omega_i = \begin{cases} 1, & \text{ungrouped data } (i = 1, \ldots, n_y) \\ n_i^c, & \text{grouped data} \quad (i = 1, \ldots, g) \end{cases} \ ,$$

and $c = 1$ or -1, according to whether as group response is considered the average or the sum of the individuals' responses in a group, respectively. Parameter θ is called *natural parameter*, because it determines the mean, since

$$\propto = \mathrm{E}(Y) = b'(\theta) \ . \tag{5.4}$$

Parameter ψ controls the variance

$$\sigma^2 = \mathrm{Var}(Y) = \frac{\psi}{\omega_i} b''(\theta) \tag{5.5}$$

and is therefore called the *dispersion parameter*. $b(\cdot)$ and $c(\cdot)$ are specific functions determined by the type of the exponential family.

Many commonly used distributions are members of the exponential family, like the normal, the gamma, the binomial, the multinomial, and the Poisson. For one-parameter families the dispersion parameter ψ is fixed. For example, the Poisson $\mathscr{P}(\theta)$ and the binomial $\mathscr{B}(n,\theta)$, for fixed n, have $\psi = 1$. These distributions are in the simpler *natural exponential family*. Furthermore, for the Poisson $\omega = 1$ while for the binomial $\omega = n$ or n^{-1}, according to whether as response y is considered the success proportion or the number of successes.

The link function $\eta_i = g(\propto_i)$ can theoretically be any monotonic and differentiable function. However, the link options are practically limited, since the link is chosen so that the inverse $\propto_i = g^{-1}(\eta_i)$ leads to admissible values for $\propto_i$ and simple functions of θ_i. Characteristic example is the case of a binomial response $\mathscr{B}(n,\pi_i)$. Then $\propto_i = \pi_i$ and it must be in $(0,1)$. The three links that are more often used for binomial data are the *logit*, the *probit*, the *complementary log–log* , and the *complementary log*. In Chap. 8, we will apply the logit link $g(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$ and refer briefly to the other options. The link function specifies the nature of the distribution considered for the error ε_i. A convenient link with nice properties is the *canonical link* that expresses $\propto_i$ in terms of the parameter θ_i, i.e., the canonical link is $g(\propto_i) = B^{-1}(\theta_i)$, where $B = b'$. Under the canonical link, $\mathbf{X}'\mathbf{Y}$ is a *sufficient* statistic for $\boldsymbol{\beta}$.

In summary, GLM is a framework that unifies a wide range of models, flexible through the choices for the distribution of its random component, for the link and eventually the error distribution. Beyond the powerful theoretical setup, it is practically attractive because it allows to draw inference for all possible GLM models by the same algorithm, simplifying thus their implementation in statistical software.

5.2 Log-Linear Model: Member of the GLM Family

Classical log-linear models, presented in Chap. 4, can be viewed in the framework of GLM for specific selection of the link function and the error distribution, as will be stated next. Doing so has specific advantages. Beyond convenience in model selection and inference by adopting the procedures developed for the GLM family, it allows for easy handling of the structural zeros in log-linear modeling (see Sect. 5.5) and it provides a platform for extending the log-linear model to model the marginals as well (see Sect. 5.6).

In order to adjust to GLM's notation, contingency tables are expanded to vectors. Thus, the $I \times J$ table $\mathbf{n} = (n_{ij})$ is expanded (by rows) to the $n_y \times 1$ vector $\mathbf{y}$ as

$$\mathbf{y} = (y_1, y_2, \ldots, y_{n_y})' = (n_{11}, n_{12}, \ldots, n_{1J}, n_{21}, \ldots, n_{IJ})' ,$$

with $n_y = IJ$. Additionally, this vector approach ensures unified treatment for tables of any dimension. Throughout this book whenever tables are expanded in vectors, expansion is considered by rows, followed by columns, layers, etc.

Under the GLM setup, the log-linear models for contingency tables are easier derived considering the Poisson distribution for the random component, i.e., $Y_i \sim \mathscr{P}(\theta_i)$ and for link the $g(\propto_i) = \log \propto_i$, $i = 1, \ldots, n_y$. The *log link* is the canonical link for the Poisson distribution. They are referred as *Poisson log-linear models*. Considering Poisson sampling is not restrictive due to the equivalence of the three possible sampling schemes (see Sect. 2.2.1). Recall that also in the classical log-linear framework, estimation was based on the Poisson likelihood (2.33).

Thus, the log-linear models for $I \times J$ tables discussed in this section can be expressed in matrix notation, as follows:

$$\log(\boldsymbol{\propto}) = \mathbf{X}\boldsymbol{\beta} \,, \tag{5.6}$$

where $\boldsymbol{\propto}$ is the $IJ \times 1$ vector of expected cell frequencies under the model, $\boldsymbol{\beta}$ is the $q \times 1$ vector of parameters, and $\mathbf{X}$ is the $IJ \times q$ associated design matrix. The table of expected cell frequencies $\mathbf{m}_{I\times J}$ is expanded the same way as the table of observed frequencies.

For example, the model of independence (4.1) subject to last category zero constraints is equivalently expressed by (5.6), where the $IJ \times 1$ vector of expected frequencies is $\propto = (m_{11}, m_{12}, \ldots, m_{1J}, m_{21}, \ldots, m_{IJ})'$, the $(I + J - 1) \times 1$ vector of parameters is $\beta = (\lambda, \lambda_1^X, \ldots, \lambda_{I-1}^X, \lambda_1^Y, \ldots, \lambda_{J-1}^Y)'$, and

$$\mathbf{X} = \begin{pmatrix} \mathbf{1} & \mathbf{1}^{(1)} & \mathbf{I}^* \\ \mathbf{1} & \mathbf{1}^{(2)} & \mathbf{I}^* \\ \vdots & \vdots & \vdots \\ \mathbf{1} & \mathbf{1}^{(I-1)} & \mathbf{I}^* \\ \mathbf{1} & \mathbf{0}_{J\times(I-1)} & \mathbf{I}^* \end{pmatrix}$$

is the $IJ \times (I + J - 1)$ design matrix, with $\mathbf{1}$ the $J \times 1$ matrix of 1's, $\mathbf{1}^{(i)}$ the $J \times (I - 1)$ matrix with 1's at the ith column and 0's in all other entries, $\mathbf{0}_{s\times t}$ the $s \times t$ matrix of 0's, and

$$\mathbf{I}^* = \begin{pmatrix} \mathbf{I}_{J-1} \\ \mathbf{0}_{1\times(J-1)} \end{pmatrix} \,,$$

where $\mathbf{I}_s$ is the $s \times s$ identity matrix.

The application of the independence model through local odds ratios (2.52), though simpler in expression, is more advanced and computationally involved, because it is not in the GLM family. It does not apply to the expected cell frequencies directly but to a function of them. For this, a generalization of the GLM is needed, briefly discussed in Sect. 5.6.

5.3 Inference for GLMs

5.3.1 ML Estimation for GLMs

For the maximum likelihood estimation of $\boldsymbol{\beta}$ for model (5.2), the log-likelihood of a given sample needs to be maximized with respect to $\boldsymbol{\beta}$. Thus, for a random sample $\mathbf{y}$ of size n_y, from a population distributed by (5.3), the log-likelihood is

$$\ell = \sum_{i=1}^{n_y} \log f(y_i;\ \theta_i, \psi, \omega_i) = \sum_{i=1}^{n_y} \frac{y_i\theta_i - b(\theta_i)}{\psi}\omega_i + \sum_{i=1}^{n_y} c(y_i, \psi, \omega_i) \tag{5.7}$$

and is a function of $\boldsymbol{\beta}$ due to (5.2) and (5.4).

The first derivative of the log-likelihood function is the *Fisher's score function*

$$s(\boldsymbol{\beta}) = \frac{\partial \ell}{\partial \boldsymbol{\beta}} = \left(\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_1}, \ldots \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_q} \right)' .$$

Equating the score function's components to zero, the corresponding likelihood equations are obtained

$$s(\beta_j) = \frac{\partial \ell}{\partial \beta_j} = \frac{\partial}{\partial \beta_j} \left(\sum_{i=1}^{n_y} \log f(y_i;\ \theta_i, \psi, \omega_i) \right) = 0\ , \quad j = 1, \ldots, q,$$

and are finally equal to

$$\sum_{i=1}^{n_y} \left(\frac{y_i - \mathrm{E}(Y_i)}{\mathrm{Var}(Y_i)} \cdot \frac{\partial g^{-1}(\eta_i)}{\partial \eta_i} \cdot x_{ij} \right) = 0\ , \quad j = 1, \ldots, q\ , \tag{5.8}$$

where $\eta_i = \sum_{j=1}^{q} \beta_j x_{ij}$. The likelihood equations (5.8) are derived applying the chain rule, since $\theta_i = (b')^{-1}(\propto_i)$, $\propto_i = g^{-1}(\eta_i)$, and using (5.4) and (5.5).

For certain distributional assumption for Y_i and particular link function g, the likelihood equations (5.8) take their explicit form and specify the MLE $\hat{\boldsymbol{\beta}}$. For the canonical link, $\eta_i = \theta_i$ and $g^{-1} = b'$, leading to $\frac{\partial g^{-1}(\eta_i)}{\partial \eta_i} = b''(\theta_i)$. Thus, by (5.5), (5.8) are simplified to

$$\sum_{i=1}^{n_y} [y_i - \mathrm{E}(Y_i)]\, x_{ij} = 0\ , \quad j = 1, \ldots, q\ , \tag{5.9}$$

stating that the likelihood equations for the canonical link equate the β_j's sufficient statistic $\sum_{i=1}^{n_y} y_i x_{ij}$ to its expected value, for $j = 1, \ldots, q$.

The asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ is derived from the second derivative of the log-likelihood, since it is equal to

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \mathscr{I}_F^{-1}\ ,$$

where $\mathscr{I}_F = \mathrm{Cov}(s(\boldsymbol{\beta}))$ is the *expected Fisher information matrix*. In our case

$$\mathscr{I}_F = \mathrm{Cov}(s(\boldsymbol{\beta})) = \mathrm{E}\left(\frac{\partial \ell}{\partial \boldsymbol{\beta}} \frac{\partial \ell}{\partial \boldsymbol{\beta}'}\right) = \mathrm{E}\left(-\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right) = \mathbf{X}'\mathbf{W}\mathbf{X} \,,$$

where $\mathbf{W}$ is a diagonal matrix with diagonal entries

$$w_i = (\partial \circ_i / \partial \eta_i)^2 [\mathrm{Var}(Y_i)]^{-1} \,. \tag{5.10}$$

For large n_y,

$$\hat{\boldsymbol{\beta}} \sim \mathscr{N}_q(\boldsymbol{\beta},\ \mathscr{I}_F^{-1}) \,.$$

The matrix of the negative second derivatives of the score function is the *observed information matrix*

$$\mathscr{I}_F^{obs} = -\mathbf{H} = -\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \,,$$

where the matrix of second derivatives $\mathbf{H}$ is usually referred as the *Hessian* matrix. It holds that

$$\mathscr{I}_F = \mathrm{E}\left(\mathscr{I}_F^{obs}\right) = \mathrm{E}\left(-\mathbf{H}\right) \,. \tag{5.11}$$

For GLMs with canonical link functions, $\eta_i = \theta_i$ implies $\frac{\partial \circ_i}{\partial \eta_i} = \frac{\partial \circ_i}{\partial \theta_i}$ and the Hessian matrix becomes

$$\mathbf{H} = -\mathbf{X}'\mathbf{W}\mathbf{X} \,, \tag{5.12}$$

with $\mathbf{W}$ a diagonal matrix with entries $w_i = \omega_i \left[g^{-1}(\theta_i)\right]' / \psi$, $i = 1, \ldots, n_y$, independent of $\mathbf{y}$. Hence

$$\mathscr{I}_F = \mathrm{E}\left(-\mathbf{H}\right) = -\mathbf{H} = \mathscr{I}_F^{obs} \,,$$

i.e., for canonical link functions, the expected and observed information matrices are identical.

The likelihood equations (5.8) or (5.9) do not usually lead to closed form expressions for the $\hat{\boldsymbol{\beta}}$ and have to be solved iteratively. The two algorithms usually applied for solving the likelihood equations are the *Newton–Raphson* and the *Fisher scoring*.

If $\boldsymbol{\beta}^{(t)}$ is the value assigned to $\hat{\boldsymbol{\beta}}$ at stage t of the iterative procedure ($t = 0, 1, 2, \ldots$), then the updating equations of the Newton–Raphson algorithm at stage $t+1$ are

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^t - \left(\mathbf{H}^{(t)}\right)^{-1} s(\boldsymbol{\beta}^{(t)}) \,, \tag{5.13}$$

where $s(\boldsymbol{\beta}^{(t)})$ and $\mathbf{H}^{(t)}$ are the score function $s(\boldsymbol{\beta})$ and the Hessian matrix $\mathbf{H}$ evaluated at $\boldsymbol{\beta}^{(t)}$. For matrix inversion to be possible, $\mathbf{H}^{(t)}$ has to be non-singular.

The algorithm converges and stops when a termination criterion is met, say after t_c iterations, leading to $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(t)}$. A termination criterion checks whether $\boldsymbol{\beta}^{(t)}$ and $\boldsymbol{\beta}^{(t+1)}$ are sufficiently close, for example, whether

$$|\ell(\boldsymbol{\beta}^{(t_c+1)}) - \ell(\boldsymbol{\beta}^{(t_c)})| \leq \varepsilon \quad \text{or} \quad \frac{\left\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\right\|}{\left\|\boldsymbol{\beta}^{(t)}\right\|} \leq \varepsilon \,,$$

for a pre-chosen small positive ε.

The Fisher's scoring algorithm is similar to the Newton–Raphson algorithm with the only difference being that it is based on the expected information matrix, instead of the observed information matrix. In particular, the updating equations for the Fisher scoring algorithm are

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^{t} + \left(\mathscr{I}_F^{(t)}\right)^{-1} s(\boldsymbol{\beta}^{(t)}) \,, \tag{5.14}$$

where $\mathscr{I}_F^{(t)}$ is $\mathscr{I}_F$ evaluated at $\boldsymbol{\beta}^{(t)}$.

The asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ is estimated for the Fisher's scoring algorithm by $\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}) = \hat{\mathscr{I}}_F^{-1}$ and for the Newton–Raphson algorithm by $\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}) = (-\hat{\mathbf{H}})^{-1}$, where $\hat{\mathscr{I}}_F$ and $\hat{\mathbf{H}}$ are $\mathscr{I}_F$ and $\mathbf{H}$, respectively, evaluated at $\hat{\boldsymbol{\beta}}$.

Due to (5.11), the Newton–Raphson and the Fisher scoring algorithm coincide for GLMs of canonical link function. For noncanonical link functions, the choice between the algorithms relates to issues of ease of application, algorithm's convergence, and efficiency of implementation. It is a choice between observed and expected information matrix. For a related discussion, we refer to the classical discussion paper by Efron and Hinkley (1978) and Palmgren (1981). Alternatively, other methods have been proposed like the *Quasi-Newton* (or *Newton's unidimensional*) method that is easier to apply since it does not require matrix inversion but does not provide estimate of the asymptotic covariance matrix. We will illustrate the Newton's unidimensional method for association models in Sect. 6.2.

The solutions of the likelihood equations correspond actually to local maxima and not to the global maximum of the log-likelihood function ℓ, as is expected for the MLE $\hat{\boldsymbol{\beta}}$. Whenever ℓ is concave, the local and global maxima are identical. For non-concave ℓ, the choice of the initial estimate $\boldsymbol{\beta}^{(0)}$ is important, to ensure that it is in the region of the global maxima.

5.3.2 *Evaluating Model Fit for GLMs*

Given a sample $\mathbf{y}$ of n_y observations, let $\hat{\boldsymbol{\alpha}}$ denote the corresponding ML estimate of $\boldsymbol{\alpha} = \text{E}(\mathbf{Y})$ under a model $\mathscr{M}$ of q parameters. The quality of the model fit is assessed by comparing the maximum log-likelihood for the model $\ell(\hat{\boldsymbol{\alpha}};\mathbf{y})$ to the maximum log-likelihood for the model that describes the data perfectly, i.e., the

saturated model. A saturated model has as many parameters as the observations in the sample. We have seen so far saturated models in the context of log-linear models. For the saturated GLM, the number of parameters is n_y, $\hat{\boldsymbol{\alpha}} = \mathbf{y}$ and the corresponding log-likelihood is $\ell(\mathbf{y};\mathbf{y})$. It is obvious that always $\ell(\hat{\boldsymbol{\alpha}};\mathbf{y}) < \ell(\mathbf{y};\mathbf{y})$ with model $\mathscr{M}$ fitting as better as its log-likelihood approaches the saturated log-likelihood. Hence, the goodness of fit of a model is expressed in terms of their difference by the test statistic

$$-2\left[\ell(\hat{\boldsymbol{\alpha}};\mathbf{y}) - \ell(\mathbf{y};\mathbf{y})\right] ,$$

which for the exponential family (5.3) becomes

$$\frac{D(\mathbf{y};\hat{\boldsymbol{\alpha}})}{\psi} = \frac{2}{\psi}\sum_{i=1}^{n_y} \omega_i \left(y_i(\tilde{\theta}_i - \hat{\theta}_i) - [b(\tilde{\theta}_i) - b(\hat{\theta}_i)]\right) , \tag{5.15}$$

where $\hat{\theta}_i$ is the ML estimate of parameter θ_i under the model $\mathscr{M}$ and $\tilde{\theta}_i$ is the estimate under the saturated model. The statistic $D(\mathbf{y};\hat{\boldsymbol{\alpha}})$ is known as *deviance*. Analogously, the Pearson's X^2 statistic can be used for testing the adequacy of model $\mathscr{M}$. In this context

$$X^2(\mathscr{M}) = \sum_{i=1}^{n_y} \frac{(y_i - \hat{\alpha}_i)^2}{\hat{\alpha}_i} . \tag{5.16}$$

For Poisson and binomial GLMs, the deviance (5.15) turns out to equal the LR statistic for testing the null hypothesis that model $\mathscr{M}$ holds against the saturated model

$$G^2(\mathscr{M}) = 2\sum_{i=1}^{n_y} y_i \log(\frac{y_i}{\hat{\alpha}_i}) . \tag{5.17}$$

The statistics above can be used for testing goodness of fit of $\mathscr{M}$, if their asymptotic distribution can be specified. For this to be possible, the data have to be grouped (each y_i occurs n_i times) with the number of observations in each group n_i being sufficiently large. In this case, the distribution for the statistics (5.15)–(5.17) is approximately $\mathscr{X}^2_{df}$, with $df = n_y - q$, the difference between the number of parameters for the saturated model (n_y) and the model under testing (q). For more on the test statistics refer to McCullagh and Nelder (1989).

These goodness-of-fit tests do not account for model complexity while they are increasing in sample size n_y, giving thus significant values even for good models if the sample size is large. Alternatively, the fit of a model $\mathscr{M}$ can be evaluated by *Akaike's information criterion* (Akaike 1974)

$$AIC = -2\ell(\hat{\boldsymbol{\alpha}};\mathbf{y}) + 2q . \tag{5.18}$$

It is based on the maximum likelihood under $\mathscr{M}$ but penalizes its value for model complexity. Furthermore, the *Bayesian information criterion* (Schwarz 1978)

$$BIC = -2\ell(\hat{\boldsymbol{\alpha}};\mathbf{y}) + (\log n)q \tag{5.19}$$

is another maximum likelihood-based measure, incorporating Bayesian thinking, that beyond complexity takes into account also the sample size n. The *AIC* and *BIC* are used for comparing models, with smaller values indicating better models. They can be used to compare also non-nested models. They will be illustrated in the log-linear model context in Sect. 5.4.1.

5.3.3 Residuals

Residuals are critical for diagnosing lack of model fit and identifying possible underlying patterns. The types of residuals used in GLM analysis are the same as those discussed in the context of independence testing for two-way tables (see Sect. 2.2.4). In the GLM setup, the raw residuals $e_i = y_i - \hat{\alpha}_i$ $(i = 1,\ldots,n_y)$ are transformed to the Pearsonian residuals

$$e_i^P = \frac{y_i - \hat{\alpha}_i}{\sqrt{\widehat{\mathrm{Var}}(y_i)}}\,, \quad i = 1,\ldots,n_y\,. \tag{5.20}$$

For the Poisson GLM, $\widehat{\mathrm{Var}}(y_i) = \hat{\alpha}_i$ in (5.20) above, while for testing independence in two-way tables, (5.20) is (2.40), expressed in vector form. Pearson's residuals are asymptotic normal distributed but not standard normal, as explained in Sect. 2.2.4. Thus, dividing the raw residuals by their asymptotic standard errors, the standardized residuals are derived

$$e_i^s = \frac{e_i^P}{\sqrt{1-\hat{h}_i}} = \frac{e_i}{\sqrt{\widehat{\mathrm{Var}}(y_i)(1-\hat{h}_i)}}\,, \quad i = 1,\ldots,I,\;\; j = 1,\ldots,J\,, \tag{5.21}$$

where $\hat{h}_i$ is the estimate of the diagonal element h_i, $i = 1,\ldots,n_y$ of the $n_y \times n_y$ matrix

$$\mathbf{Hat} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{W}^{1/2}\,,$$

known as *hat matrix*, with $\mathbf{W}$ the diagonal matrix with entries (5.10).

The *deviance residuals* decompose the deviance to the individual contributions of each observation i. Hence, for the exponential family (5.3), they are equal to

$$e_i^d = \mathrm{sign}(y_i - \hat{\alpha}_i)\cdot\left[2\omega_i\left(y_i(\tilde{\theta}_i - \hat{\theta}_i) - [b(\tilde{\theta}_i) - b(\hat{\theta}_i)]\right)\right]^{1/2}\,, \quad i = 1,\ldots,n_y\,, \tag{5.22}$$

satisfying $D(\mathbf{y};\hat{\boldsymbol{\alpha}}) = \sum_i^{n_y}\left(e_i^d\right)^2$. For testing independence in two-way tables, (5.22) simplify to (2.43).

5.3.4 Model Selection in GLMs

Deviance plays a predominant role in comparing GLMs, via the likelihood ratio criterion, for responses y_i, $i = 1, \ldots, n_y$, in the exponential family with $\psi = 1$. In this case, by (5.15), the deviance of a model is equal to the corresponding LR statistic (4.33) for testing its fit.

Let $\mathscr{M}_1$ be a GLM of q_1 parameters. Let also $\mathscr{M}_0$ be a simpler GLM, produced from $\mathscr{M}_1$ by eliminating r of its q_1 parameters. Then, $\mathscr{M}_0$ is said to be *nested* in $\mathscr{M}_1$ and denoted by $\mathscr{M}_0 \subset \mathscr{M}_1$. Model $\mathscr{M}_0$ has $q_0 = q_1 - r$ parameters and is more parsimonious than $\mathscr{M}_1$.

If $\hat{\boldsymbol{\alpha}}_0$ and $\hat{\boldsymbol{\alpha}}_1$ are the ML estimates of $\boldsymbol{\alpha}$ under $\mathscr{M}_0$ and $\mathscr{M}_1$, respectively, then, for $\psi = 1$, the deviances of models $\mathscr{M}_0$ and $\mathscr{M}_1$ are

$$\begin{aligned} D(\mathbf{y}; \hat{\boldsymbol{\alpha}}_0) &= -2\left[\ell(\widehat{\boldsymbol{\alpha}_0}; \mathbf{y}) - \ell(\mathbf{y}; \mathbf{y})\right] \\ D(\mathbf{y}; \hat{\boldsymbol{\alpha}}_1) &= -2\left[\ell(\widehat{\boldsymbol{\alpha}_1}; \mathbf{y}) - \ell(\mathbf{y}; \mathbf{y})\right] . \end{aligned}$$

Since reducing the number of model's parameters implies increase of model's distance from the perfect fit of the saturated model, it will always be $D(\mathbf{y}; \hat{\boldsymbol{\alpha}}_0) > D(\mathbf{y}; \hat{\boldsymbol{\alpha}}_1)$.

Models $\mathscr{M}_0$ and $\mathscr{M}_1$ apply both on the same $\mathbf{y}$, thus their difference is

$$D(\mathbf{y}; \hat{\boldsymbol{\alpha}}_0) - D(\mathbf{y}; \hat{\boldsymbol{\alpha}}_1) = -2\left[\ell(\widehat{\boldsymbol{\alpha}_0}; \mathbf{y}) - \ell(\widehat{\boldsymbol{\alpha}_1}; \mathbf{y})\right] = \mathrm{LRS}(\mathscr{M}_0, \mathscr{M}_1) ,$$

where $\mathrm{LRS}(\mathscr{M}_0, \mathscr{M}_1)$ is the LR statistic for testing the null hypothesis that $\mathscr{M}_0$ holds against the alternative that $\mathscr{M}_1$ holds. In particular, by (5.15), the difference in deviances equals

$$D(\hat{\boldsymbol{\alpha}}_0; \hat{\boldsymbol{\alpha}}_1) = D(\mathbf{y}; \hat{\boldsymbol{\alpha}}_0) - D(\mathbf{y}; \hat{\boldsymbol{\alpha}}_1) = 2\sum_{i=1}^{n_y} \omega_i \left(y_i(\hat{\theta}_{i1} - \hat{\theta}_{i0}) - [b(\hat{\theta}_{i1}) - b(\hat{\theta}_{i0})]\right) . \tag{5.23}$$

Under $\mathscr{M}_0$, (5.23) is approximately $\mathscr{X}^2_r$ distributed, where $r = q_1 - q_0$ is the difference between the number of parameters of the two compared models. This asymptotic result is the key for models' comparison.

For Poisson log-linear models, (5.23) simplifies to (4.34), i.e.,

$$G^2(\mathscr{M}_0 | \mathscr{M}_1) = 2\sum_{i=1}^{n_y} \hat{\alpha}_{i1} \log\left(\frac{\hat{\alpha}_{i1}}{\hat{\alpha}_{i0}}\right) = G^2(\mathscr{M}_0) - G^2(\mathscr{M}_1) ,$$

where $G^2(\mathscr{M}_0)$ and $G^2(\mathscr{M}_1)$ are as in (5.17).

Upon considering a sequence of nested models from a very simple $\mathscr{M}_0$ up to the saturated $\mathscr{M}_{\mathrm{sat}}$,

$$\mathscr{M}_0 \subset \mathscr{M}_1 \subset \mathscr{M}_2 \subset \ldots \subset \mathscr{M}_{\mathrm{sat}} ,$$

the importance of the parameters added gradually can be evaluated by successive comparisons of neighbor models. Thus, the appropriate model can be built by selecting this model $\mathcal{M}_s$ for which $D(\hat{\boldsymbol{\alpha}}_s; \hat{\boldsymbol{\alpha}}_{s+1})$ is nonsignificant and $D(\hat{\boldsymbol{\alpha}}_{s-1}; \hat{\boldsymbol{\alpha}}_s)$ is significant. This means that adding more parameters would complicate the model without improving its fit significantly, while removing any parameters further would lead to a model of significantly poorer fit. Hence, comparisons of nested models serve for developing procedures of "best model" selection. Furthermore, once the "best model" is selected, model comparison can serve as a tool for evaluating the individual importance of each parameter or group of parameters. Model selection can also be based on *AIC* and *BIC*. For a comparative study of *AIC* and *BIC* and a corrected for finite samples version of *AIC* with emphasis on their role in model selection, we refer to Burnham and Anderson (2004). These criteria will be illustrated in the context of log-linear models for multi-way tables next (see Sect. 5.4.1).

5.4 Software for GLMs

All general-purpose statistical packages (like SAS, SPSS, Stata, and SYSTAT) have procedures for GLM analysis. For example, GLMs are fitted in SAS by the procedure GENMOD. The corresponding `R` function is `glm`, which is based on the S-function "glm" (Hastie and Pregibon 1992). The basic form for calling the `glm` function is

```
> Mfit <- glm(formula, family=..., data=...)
```

where `formula` defines the model to be fitted, `family` determines the error distribution and link function of the model, and `data` specifies the data frame on which the model will be applied. `Mfit` is the object where output of `glm` is saved. `formula` is provided in a form of the type `Y~X1+X2+X3+X1:X2`, where `Y` is the dependent variable, `X1, X2, X3` the independent, and `X1:X2` denotes the interaction between `X1` and `X2`. The expression above is equivalent to `Y~X3+X1*X2`, where `X1*X2` stands for the generating term of a hierarchical model, i.e., it is equivalent to `Y~X1+X2+X1:X2`. For log-linear models the choice for `family` is `family=poisson (link = "log")`. The specification of data frame is optional. If it is omitted, the variables are taken from the environment from which `glm` is called.

The minimum output is printed on screen by simply typing `Mfit` while more detailed output is provided by `summary(Mfit)`. The content of object `Mfit` can be viewed by `names(Mfit)`. An item, say `A`, of `Mfit` is located in `Mfit$A` and can be saved in a variable for further use (e.g., `V1 <- Mfit$A`). Due to the predominant role deviance plays in GLM's analysis, the residuals saved in `Mfit`, the output object of `glm`, are the deviance residuals. For results not provided in `Mfit`, a variety of special functions is available that apply on the `glm` output. Function `step()` for model selection between nested models and `anova()` for analysis of variance can be activated also in `glm` framework, as will be illustrated in the examples that follow.

Table 5.1 Summary output of the independence model applied on Table 2.3, fitted by `glm`

```
Call:
glm(formula = freq ~ WELFARE + DEGREE, family = poisson,data = nt.frame)

Deviance Residuals:
Min                  1Q  Median            3Q  Max
-1.3419         -0.5377  -0.1352       0.3366  1.6724

Coefficients:
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  3.54654    0.10253    34.590    < 2e-16 ***
WELFARE2     0.32962    0.08276     3.983    6.81e-05 ***
WELFARE3     0.34666    0.08247     4.204    2.63e-05 ***
DEGREE2      1.26567    0.09855    12.843    < 2e-16 ***
DEGREE3     -0.42845    0.13858    -3.092     0.00199 **
DEGREE4      0.30458    0.11473     2.655     0.00793 **
DEGREE5     -0.38299    0.13670    -2.802     0.00508 **
--
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

( Dispersion parameter for poisson family taken to be 1 )

   Null deviance: 478.046 on 14 degrees of freedom

Residual deviance: 10.363 on 8 degrees of freedom

AIC: 110.74

Number of Fisher Scoring iterations: 4
```

For historical reasons, let us note that GLIM (generalized linear interactive modeling) was the first package with the ability of fitting a variety of GLMs in a unified manner. It was developed by the GLIM working party of the Royal Statistical Society in 1974. GLIM4, the latest release (1993), had many links as standard options and was convenient for GLM fit and model selection. A rich macro library was available while users could write their own macros in GLIM language. The associated journal *GLIM Newsletter*, issued from 1979 to 1998, was publishing GLIM macros.

5.4.1 Example 2.4 by `glm`

The log-linear model of independence (4.1) will be fitted on Table 2.3, by `glm` of `R`. The variables are required in vector form; thus we apply `glm` on the data frame `nt.frame`, constructed in Sect.4.2.1. Model (4.1) is then fitted by

```
> I.glm <- glm(freq ~ WELFARE+DEGREE,family=poisson,data=nt.frame)
```

and the extended output (provided in Table 5.1) is obtained by

```
> summary(I.glm)
```

The value of the G^2 statistic is reported under "Residual Deviance" and is saved in `I.glm$deviance`, as can be verified by typing `names(I.glm)`. Its asymptotic p-value is not provided but can easily be calculated by

```
> p.value <- 1-pchisq(I.glm$deviance, I.glm$df.residual)
```

We find p-value = 0.240; thus the independence model describes adequately this data set. Furthermore the value of the AIC is given ($AIC = 110.74$) while the BIC, defined by (5.19), can be computed as

```
> n <- sum(ntfare$freq);  q <- I.glm$df.null-I.glm$df.residual
> BIC <- I.glm$aic-(2-log(n))*q
```

giving $BIC = 139.91$. The level of the AIC and BIC values can be judged in comparison to alternative models. In this case, for the saturated model

```
> sat <- glm(freq ~ WELFARE*DEGREE,family=poisson,data=nt.frame)
```

$AIC = 116.4$, while for the models of only one main effect

```
> welfr <- glm(freq ~ WELFARE,family=poisson,data=nt.frame)
```

and

```
> degr <- glm(freq ~ DEGREE,family=poisson,data=nt.frame)
```

we get $AIC = 548.1$ and $AIC = 129$, respectively. Hence, the choice of the independence model is justified.

Function `glm` produces parameter estimates subject to the first category zero constraints. Recall that only the effect differences between different categories are of interest and these remain invariant under different types of constraints. Observe that $\hat{\lambda}_3^X - \hat{\lambda}_1^X = 0.347 - 0$, equal to the corresponding value derived in Sect. 4.2.1 subject to the sum to zero constraints.

The residuals saved in object `I.glm` are the working residuals. The Pearsonian residuals are calculated by `residuals(I.glm, type = c("pearson"))` and the deviance by changing the type option to `"deviance"`. Standardized residuals are obtained by `rstandard(I.glm)`.

The items of the output object are all in vector form but can easily be transformed to the more friendly table form by `xtabs()`. For example, the ML estimates of the expected cell frequencies under independence and the standardized residuals are derived in table form by

```
> MLEs <- xtabs(I.glm$fitted.values ~ WELFARE+DEGREE,data=natfare)
> stdres <- xtabs(rstandard(I.glm) ~ WELFARE+DEGREE,data=natfare)
```

Thus, the standardized residuals are

```
> stdres
```

	DEGREE				
WELFARE	LT HS	HS	JColg	BA	Grad
too little	2.0983151	-1.039894	-0.9517240	0.1790943	-0.1654438
about right	-1.6533505	-0.543633	0.3659428	0.4422752	1.7955727
too much	-0.4040979	1.462390	0.4702723	-0.6127615	-1.7788921

The only standardized residual that exceeds in absolute value 1.96 corresponds to cell $(1,1)$. That is, responders with educational level lower than high school tend to believe that welfare spending is too little with higher probability than expected under the independence model.

The sequence of commands followed above is unified in function `fit.I()` of the web appendix (see Sect. A.3.4), which additionally provides the values for Pearson's X^2 along with its p-value, the dissimilarity index (4.18) and the *BIC*. The function requires the vector of frequencies (by rows) and the number of rows and columns of the table. For this example, it is called as `fit.I(freq,3,5)`.

The standardized residuals can be displayed on the mosaic plot as shown below. We apply

```
> mosaic(natfare, gp=shading_Friendly, residuals=stdres,
+   residuals_type="Std\nresiduals",labeling = labeling_residuals)
```

where `stdres` is the table of standardized residuals derived above. The mosaic plot derived is given in Fig. 5.1 (right). The figure on the left is the mosaic plot for standardized residuals for Example 2.2 and is derived analogously.

The residuals illustrated in the mosaic plots so far were all for the independence model (default). To refer to residuals of a different model, the output object of the assumed model has to take the position of the data matrix as input in `mosaic()`. Thus,

```
> mosaic(natfare, gp=shading_hcl, residuals_type="deviance")
```

is equivalent to

```
> mosaic(I.glm, gp=shading_hcl, residuals_type="deviance")
```

To incorporate the residuals of the model with only the row (opinion) main effect

```
> X.glm <- glm(freq ~ WELFARE+DEGREE,family=poisson,data=nf.frame)
```

the mosaic plot function should be

```
> mosaic(X.glm, gp=shading_hcl, residuals_type="deviance")
```

From the ML estimates it can be verified that the estimated under independence $\hat{\theta}_{ij}$ ($i = 1,2$, $j = 1,\ldots,4I$) are, as expected, all equal to 1. The same holds also for the global and cumulative odds ratios. The ML estimates of any set of generalized odds ratios expected under the assumed model can be calculated in `R`, using the corresponding functions of the web appendix (see Sect. A.3.2). The procedure is that described for the sample generalized odds ratios at the end of Sect. 2.2.5 and illustrated in the example of Sect. 2.2.6. Only the vector of observed frequencies has to be replaced with the vector of ML estimates of the expected cell frequencies under the assumed model. The equivalent independence model (2.52) in terms of the local odds ratios will be illustrated for this example in Sect. 5.6.

5.4.2 *Example 3.1 (Revisited)*

For the example of Table 3.1, we have seen in Sect. 3.3, applying the Breslow–Day test (or the Woolf test), that the association between smoking and depression is homogeneous for males and females. At this point, we shall select the appropriate log-linear model for describing the underlying association structure of Table 3.1. The data are available in `R` in matrix `depsmok3`. In order to fit the models in the GLM setup applying `glm`, the data have to be expanded from a matrix to a vector and the factors corresponding to the classification variables have to be defined. This is carried out easily as follows:

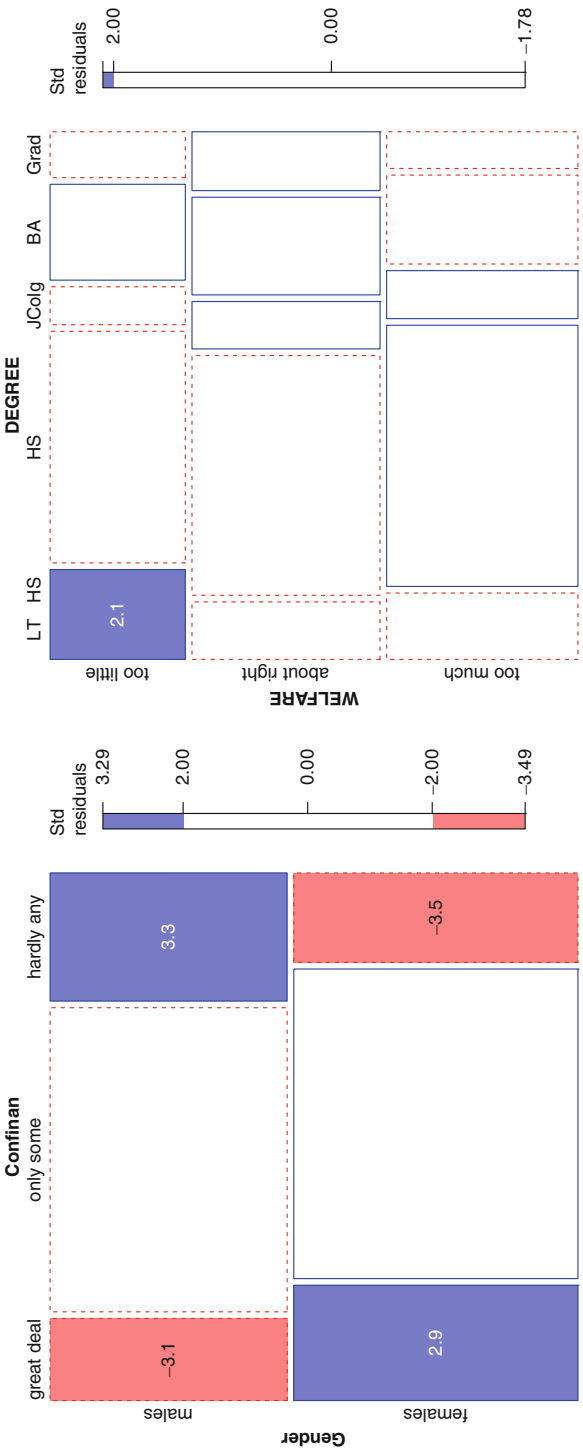

Fig. 5.1 Mosaic plots of standardized residuals for the independence model applied on Table 2.2 (*left*) and Table 2.3 (*right*)

```
> obs <- as.vector(depsmok3)
> row <- rep(1:2, 4);  col <- rep(1:2, each=2,2)
> lay <- rep(1:2, each=4);  row.lb <- c("yes","no")
> col.lb <- c("yes","no");  lay.lb <- c("male", "female")
> S <- factor(row,labels=row.lb);  D <- factor(col,labels=col.lb)
> G <- factor(lay, labels=lay.lb)
> depres.fr <- data.frame(obs,S,D,G)
```

The appropriate log-linear model is selected via the backward stepwise procedure based on *AIC*. Thus, we first save the saturated model under object `saturated` and then proceed with the backward model selection procedure as follows:

```
> saturated <- glm(freq S*D*G, poisson, data = depres.fr)
> step(saturated, direction="backward")
```

The stepwise procedure concludes to the model of no three-factor interaction (*SD*, *DG*, *SG*), giving the following output:

```
Start: AIC=71.38
freq  S * D * G:
                Df          Deviance   AIC
- S:D:G          1           0.77135   70.155
<none>                       0.00000   71.384
Step: AIC=70.16
freq ~ S + D + G + S:D + S:G + D:G
                Df          Deviance     AIC
<none>                         0.771     70.155
- S:D            1            33.024     100.408
- D:G            1            34.386     101.769
- S:G            1           112.298     179.682
Call: glm(formula = freq~S+D+G+S:D+S:G+D:G, family=poisson,
data=depres.fr)

Coefficients:

Intercept       Sno               Dno   Gfemale
3.7393        -1.6684          3.0485   0.8850
Sno:Dno     Sno:Gfemale   Dno:Gfemale
0.9187        0.7834         -0.9369

Degrees of Freedom: 7 Total (i.e. Null); 1 Residual
Null Deviance:  3315
Residual Deviance: 0.7713   AIC: 70.16
```

The (*SD*, *DG*, *SG*) is also the model of *homogeneous association* since under this model the association in all two-way partial tables is homogeneous across the levels of the remaining third classification variable, as explained in Sect. 4.3. This model is fitted in R, as shown below, giving the output provided in Table 5.2.

```
> hom.assoc <- glm(freq~S*D+S*G+D*G, poisson,data=depres.fr);
summary(hom.assoc)
```

The p-value of testing the model fit based on G^2 statistic is 0.380, which is close to the corresponding p-values of the Woolf's or the Breslow–Day test (Sect. 3.3.3).

Table 5.2 Output for model (SD, DG, SG), fitted on data in Table 3.1

```
Call:
glm(formula = freq~S*D+S*G+D*G, family=poisson, data=depres.fr)

Deviance Residuals:
      1       2       3        4        5       6        7
-0.32157 0.70555 0.06943 -0.10112 0.20418 -0.32157 -0.07131
      8
0.07006
Coefficients:
             Estimate   Std. Error   z value     Pr(>|z|)
(Intercept)   3.73930      0.14417    25.936     < 2e-16 ***
SNo          -1.66844      0.17668    -9.443     < 2e-16 ***
DNo           3.04847      0.14705    20.731     < 2e-16 ***
Gfemale       0.88501      0.16620     5.325     1.01e-07 ***
SNo:DNo       0.91871      0.17059     5.385     7.23e-08 ***
SNo:Gfemale   0.78344      0.07529    10.405     < 2e-16 ***
DNo:Gfemale  -0.93691      0.17055    -5.493     3.94e-08 ***
--
Signif. codes:   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 3315.40325 on 7 degrees of freedom

Residual deviance: 0.77135 on 1 degrees of freedom

AIC: 70.155

Number of Fisher Scoring iterations: 4
```

Relation (4.27), adjusted in our setup, becomes

$$\log \theta_{(k)}^{SD} = \log\left(\frac{\pi_{11|k}\pi_{12|k}}{\pi_{21|k}\pi_{22|k}}\right) = \lambda_{22}^{SD} = \log \theta^{SD} \ , \ k = 1,2 \ ,$$

due to the identifiability constraints $\lambda_{11}^{SD} = \lambda_{12}^{SD} = \lambda_{21}^{SD} = 0$. Thus, the ML estimate of the common odds ratio θ^{SD} under the log-linear model of homogeneous association is

$$\hat{\theta}^{SD} = \exp\left(\hat{\lambda}_{22}^{SD}\right) = \exp(0.91871) = 2.506 \ ,$$

close in value to $\hat{\theta}_{MH}$ and $\hat{\theta}_W$, calculated in Sect. 3.3.3.

Furthermore, the asymptotic Wald $(1-\alpha)100\%$ CI for θ^{SD} is

$$\exp\left[\log \hat{\theta}^{SD} \pm z_{\alpha/2} s.e.\left(\log \hat{\theta}^{SD}\right)\right] \ ,$$

where $s.e.(\log \hat{\theta}^{SD})$ is the standard error of $\log \hat{\theta}^{SD}$ and is equal to $s.e.(\log \theta^{SD}) = s.e.(\lambda_{22}^{SD}) = 0.17059$.

This CI can easily be computed via the function

```
> CI <- function(t, SE, conf.level=0.95)
                  {exp(t+c(-1,1)*qnorm(0.5*(1+conf.level))*SE)}
```

with `t` and `SE` standing for $\log \hat{\theta}^{SD}$ and its standard error, respectively. Hence, the 95% CI for θ^{SD} in this case is computed as

```
> logSD <- 0.91871 ; SE.SD <- 0.17059
> CI(logSD, SE.SD)
[1] 1.793842  3.501041
```

The `xtabs()` function, used in the previous example (Sect. 5.4.1), is especially useful in multi-way tables, since it provides a straightforward way to extract marginal and partial tables of observed or expected cell frequencies. In this example for instance, the smoking-depression marginal table of the ML estimates of the expected cell frequencies under (SD, DG, SG) is

```
> MLE.SD <- xtabs(hom.assoc$fitted.values ~ S + D)
```

and, as expected, coincides with the corresponding marginal table of observed frequencies, which for arrays is obtained by

```
> margin.table(depsmok3, c(1,2))
```

or

```
> apply(depsmok3, c(1,2), sum)
```

However, were the data available only in the data frame format (`depres.fr`), with `obs` the vector of observed frequencies, then the smoking-depression observed marginal table would be

```
> MLE.SD <- xtabs(obs ~ S + D)
```

5.5 Independence for Incomplete Tables

In case of structural zeros existence (see also Sect. 4.9.1), the corresponding cells are of zero probability and must be excluded from the analysis . Thus, any model assumed will not apply on all cells of the contingency table under consideration but only on the subset of its nonstructural zero cells. Hence, structural zeros affect the assumed model in substance. A table with structural zeros is known as an *incomplete* or *truncated table*.

As an illustration, we will consider the independence model for an $I \times J$ table. Independence is considered not for all IJ cells but only for the subset of the nonstructural zero cells $S = \{(i, j) : \pi_{ij} > 0\}$. The model of independence applied on an incomplete table is known as the *quasi-independence* (QI) model, term introduced by Goodman (1968).

QI is defined naturally in the log-linear models framework, as the classical model of independence (4.1), applied on a subset S of the table

$$\log m_{ij} = \lambda + \lambda_i^X + \lambda_j^Y \ , \qquad (i, j) \in S \ . \tag{5.24}$$

The main effect parameters satisfy the identifiability constraints (4.4), and the associated df are $df = (I-1)(J-1) - s$, where $s = IJ - |S|$ is the number of structural zeros, i.e., the cardinality of the set of structural zeros S^c .

The restriction $(i,j) \in S$ can be incorporated in the model by introducing s additional parameters in (4.1), one for each structural zero. Hence, (5.24) is equivalent to

$$\log m_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + q_{ij} \mathrm{I}_{ij}^{S^c} \ , \ \ i = 1, \ldots, I, \ j = 1, \ldots, J \ , \tag{5.25}$$

where $\mathrm{I}_{ij}^{S^c}$ is the indicator function for structural zeros

$$\mathrm{I}_{ij}^{S^c} = \begin{cases} 1 \ , \ (i,j) \notin S \\ 0 \ , \ (i,j) \in S \end{cases} .$$

This way, the structural zero cells equal the observed counts ($n_{ij} = m_{ij} = 0$ for $(i,j) \notin S$), sacrificing thus $s \ df$. Structural zeros have no contribution to the value of the X^2 or G^2 test statistic.

QI is expressed directly on the cell probabilities, as

$$\pi_{ij} = \alpha_i \beta_j \ , \qquad (i,j) \in S \ ,$$

where the marginal parameters are no more the marginal probabilities.

Additionally, structural zeros serve as a powerful tool in contingency table analysis, since they can be activated by the needs of the analysis to exclude a specific cell or region of the table that is nonzero but exhibits "special behavior" and exacerbates the fit of the assumed model. This is often the case for mobility tables or panel studies, where the tables are square with augmented diagonal entries, corresponding to non-change. It is natural thus to exclude the diagonal from the analysis by considering $S = \{(i,j) : \ i \neq j\}$. Other incomplete square tables that received special attention are triangular tables. We will return to special QI models for square tables in Sect. 9.3. References on conditions for existence of ML estimates for truncated tables are provided in Sect. 5.7.1.

Structural zeros are incorporated easily in log-linear models analysis in the GLM framework. A cell (i,j) is excluded from the model, by the inclusion in the log-linear model (5.25) of the additional parameters q_{ij} that is responsible for fixing it to its observed frequency ($e_{ij} = 0$). In practice, this is achieved in standard software by adding in the log-linear model the index variable of (5.25) as an explanatory variable. In the presence of more structural zeros, additional index variables are added in the model, one for each structural zero. Alternatively, in the GLM context, all structural zeros can be indicated in one single variable that will be used to determine the subset of cells on which model (5.24) will be applied. SPSS handles structural zeros in the "general log-linear analysis" straightforward. An index variable has to be added in the data file, taking values 0 for structural zero cells and 1 otherwise. This index variable has to be declared in the "Cell Structure" field of the window:

Analyze > Loglinear > General...

QI will be illustrated in R, using Example 5.1 below.

When interaction is significant, model (4.5) is expressed for two-way incomplete tables as

$$\log m_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY} , \qquad (i,j) \in S . \tag{5.26}$$

The main effect parameters satisfy constraints (4.4) while the sum to zero constraints in (4.6) for the interaction parameters are corrected to

$$\sum_{i=1}^{I} \mathrm{I}_{ij}^{S^c} \lambda_{ij}^{XY} = \sum_{j=1}^{J} \mathrm{I}_{ij}^{S^c} \lambda_{ij}^{XY} = 0 .$$

Log-linear models for multi-way incomplete contingency tables can be defined and fitted in an analogous manner.

5.5.1 Example 5.1

A typical example of contingency table with structural zeros is a survey on teenagers' health concerns. Teenagers are cross-classified according to their health concerns (in four categories), gender, and age (in two categories: 12–15, 16–17) in a $4 \times 2 \times 2$ table. The table has two structural zeros, since the health concerns category "menstrual problems" cannot refer to boys. This example is analyzed by Grizzle and Williams (1972) and Fienberg (2007, pp. 148–150). Ignoring age, i.e., merging over the age, the data are provided in Table 5.3, and there exists 1 structural zero; thus, the test of QI will be based on $2\,df$. QI is rejected, since $G^2(\mathrm{QI}) = 12.60$ (p-value = 0.0018) and $X^2(\mathrm{QI}) = 12.39$ (p-value = 0.0020). The ML estimates of the expected under QI cell frequencies along with the standardized residuals are provided in Table 5.3 in parentheses. Observing them, we conclude that the greatest difference between genders lies on the category "how healthy I am," for which girls are significantly less concerned and boys more than under independence, followed by "sex, reproduction" for which boys are significantly less interested while girls more, though not as significant. Finally, boys are more health concerns-free than expected under independence and girls less, but these differences are at the limit of 5 % significance.

This model was fitted in R by the function `fit.QI()`, provided in web appendix (see Sect. A.3.4). This function fits the QI model by (5.24), excluding the structural zero cells from the analysis. It needs to read the numbers of rows I and columns J of the table, the cell frequencies in a vector (by rows) of length IJ, where 0 are put in places of structural zeros, and an index vector of length IJ with entries the $\mathrm{I}_{ij}^{S^c}$ indices, given by rows. Thus for our example, the analysis is carried out by the commands

```
> freq<-c(6,16,0,12,49,29,77,102)
> zer<- c(0,0,1,0,0,0,0,0)
> fit.QI(freq,zer,4,2)
```

Table 5.3 Teenagers' cross-classification by gender and their health concerns (Brunswick 1971)

	Gender	
Health concerns	Male	Female
Sex, reproduction	6 (10.41, −2.13)	16 (11.59, 1.85)
Menstrual problems	–	12 (12.00, 0.00)
How healthy I am	49 (36.90, 3.08)	29 (41.10, −3.41)
Nothing	77 (84.69, −1.95)	102 (94.31, 1.90)

In parenthesis are provided the ML estimates under the QI model and the standardized residuals

The output of `fit.QI()`, beyond the results presented above, includes the overview of the fit provided by `summary()` and the estimates of the log-linear model parameters in vector forms for possible further use.

Alternatively, without restricting the cells on which the model applies, the QI model can be fitted by (5.25), including s extra parameters in the model, one for each structural zero. For this example, $s = 1$ and would have

```
> NI <- 4
> NJ <- 2
> row<-gl(NI,NJ,length=NI*NJ)
> col<-gl(NJ,1,length=NI*NJ)
> example <- data.frame(row, col, freq, zer)
> QI.model <- glm(freq ~ row+col+zer, poisson)
```

Under this approach, in the presence of $s > 1$ structural zeros, the index vector `zer` used in `glm()` above, needs to be replaced by a *factor* of $s + 1$ levels. Level 0 is assigned to the non-structural zero cells and a different level (from 1 to s) is assigned to every structural zero cell.

In case of existence of sampling zeros as well, they will not differ from the structural zeros in the frequency vector but in their index vector entry.

5.6 Models for Joint and Marginal Distributions

Model (5.6) applies directly on the cell entries of the table. In certain frameworks, it is of interest to model or test hypotheses about linear functions of the cell entries. For this, (5.6) is extended to

$$\log(\mathbf{Mm}) = \mathbf{X}\boldsymbol{\beta} \ , \tag{5.27}$$

with $\mathbf{M}$ a matrix suitably defined in order to form the desired functions of the expected cell entries when applied on $\mathbf{m}$.

The most famous models of this type are those modeling the marginals of a table, since some structures can easier be expressed in terms of marginal distributions, leading to the *marginal models*. Marginal models for contingency tables impose

structural restrictions on certain marginals of the classification variables and are usually of log-linear type. A characteristic example is the *marginal homogeneity* model for a square $I \times I$ table, presented in Sect. 9.2.2. For higher dimensional tables, modeling the marginal distributions is important for clustered and longitudinal categorical data (see Sects. 5.7.2 and 9.7.4).

However, if we would like to model the local odds ratios of an $I \times J$ table, model (5.27) is not appropriate; a further extension is needed. A brighter family of models is the generalized log-linear model (GLLM)

$$\mathbf{C}\log(\mathbf{Mm}) = \mathbf{X}\boldsymbol{\beta} \ . \tag{5.28}$$

Matrix $\mathbf{C}$ provides more flexibility and allows an even brighter variety of models to be included in this class. GLLM is introduced by Lang and Agresti (1994) and opened new origins in the analysis of multivariate categorical data, providing a powerful and flexible framework to model structures of associations. Model (5.28) is suitable for modeling, among others, the log of local or global odds ratios (see Sect. 2.2.5). Recall the matrix definition of the generalized odds ratios, given by (2.54) and (2.55), which correspond to the left-hand side of (5.28).

GLLM is itself a member of the broader *multinomial-Poisson homogeneous* (MPH) model, which is of the very general form

$$\mathrm{L}(\mathbf{m}) = \mathbf{X}\boldsymbol{\beta} \ , \tag{5.29}$$

where L is a link function. Details on inference for the MPH model are beyond the scope of this book and can be found in Lang (2004, 2005). Setting $\mathrm{L}(\mathbf{m}) = \mathbf{C}\log(\mathbf{Mm})$, (5.29) reduces to (5.28).

Another special case of the MPH model (5.29) is the

$$\mathrm{h}(\mathbf{m}) = \mathbf{0} \ , \tag{5.30}$$

where h() is a smooth constraint function with the constraints in (5.30) being nonredundant. With the adequate choice of the constraint function h(), model (5.30) reduces to the independence model (2.52), expressed in terms of the local odds ratios.

Though inference for the MPH model is not straightforward, it can be implemented in `R` by the `mph` function of Lang or the package `hmmm` of Colombi et al. (2013). We will illustrate `mph`, which is a powerful and flexible function that fits a big variety of general models via maximum likelihood. We limit its use only to GLLM models (5.28) and to model (5.30), both considered for the local odds ratios and the global odds ratios of a contingency table.

Function `mph` is available on request. The file "mph.Rcode.txt" is then sent and the routine `mph` is activated in `R` by

```
> source("c://...//mph.Rcode.txt")
```

The data are read in vector form that has to be defined as matrix. Thus, the $I \times J$ table of observed frequencies is expanded (by rows) in a $IJ \times 1$ vector `freq` and this vector finally forms the $IJ \times 1$ data matrix

```
> y <- matrix(freq)
```

The derived vector of expected cell frequencies $\mathbf{m}$ is also a matrix of size $IJ \times 1$.

The typical expression of the `mph` function for fitting (5.29) is

```
> mph.out <- mph.fit(y=y,L.fct=L.fct,X=X, strata=1)
```

where `L.fct` is the link function and `X` the design matrix of the MPH model (5.29) under consideration. The link for the GLLM model (5.28) is defined by

```
> L.fct <- function(m)C%*%log(M%*%m)
```

with `C` and `M` appropriate defined matrices. In the sequel, command

```
> mph.summary(mph.out,cell.stats=T,model.info=T)
```

produces summary output of the model, i.e., goodness-of-fit statistics, parameter estimates, expected cell frequency estimates under the assumed model, and information on the model applied and its convergence.

Model (5.30) is fitted by

```
mph.constr <- mph.fit(y, constraint=h.fct, strata=1)
```

where `h.fct` is the constraints function. For example, in order to fit the independence model (2.52), it should be

```
> h.fct <- function(m) {C%*%log(m)}
```

with `C` an appropriate $(I-1)(J-1) \times IJ$ matrix.

Examples of fitting the GLLM model through the `L.fct` option will be discussed in Sects. 6.6.4 and 7.1, for the local and the global odds ratios, respectively. The standard expression of `mph.fit()` assumes one single multinomial sample (`strata=1`). The extra option for defining more strata of data will be discussed in Sect. 5.6.2. At this point we will use `mph` to fit model (2.52) for our familiar Example 2.3, illustrating the use of `h.fct`.

5.6.1 *Example 2.4 by* `mph`

The function `local.odds.DM()` in the web appendix (see Sect. A.3.2) produces the matrix `C` needed to derive the logs of the local odds ratios when multiplied to `log(m)`, for tables of any size $I \times J$.

Hence, after actualizing `mph` in `R`, model (2.52) is fitted for our example by

```
> NI <- 3;  NJ <- 5
> freq <- c(45,116,19,48,23,40,167,33,68,41,47,185,34,63,26)
> C<-local.odds.DM(NI,NJ)
> h.fct <- function(m) {C%*%log(m)}
> ind.odds <- mph.fit(y, constraint=h.fct, strata=1)
```

The corresponding output is derived by

```
> mph.summary(ind.odds,cell.stats=T,model.info=T)
```

Part of this output is provided in Table 5.4.

Table 5.4 Output of the `mph` function, fitting the independence model on the local odds ratios of Example 2.4

```
MODEL GOODNESS OF FIT: Test of Ho: h(p)=0 vs. Ha: not Ho...

Likelihood Ratio Stat (df=8):  Gsq=10.36287 (pval=0.2405)
Pearson's  Score Stat (df=8):  Xsq=10.52048 (pval=0.2304)
Generalized Wald Stat (df=8):  Wsq=10.40275 (pval=0.2379)

Adj Resids: -1.709 -1.604 ...1.865 2.195,
Number |Adj Resid| > 2: 1

SAMPLING PLAN INFORMATION...
Number of strata: 1
Strata identifiers: 1
Strata with fixed sample sizes: all
Observed strata sample sizes: 955
CELL-SPECIFIC STATISTICS...
    strata OBS       FV StdErr.FV  PROB  StdErr.PROB ADJ.RESIDS
y1     1    45  34.6932   3.3753  0.0363    0.0035     2.1954
y2     1   116 123.0031   7.8052  0.1288    0.0082    -1.0299
y3     1    19  22.6031   2.6280  0.0237    0.0028    -0.9253
y4     1    48  47.0461   4.0679  0.0493    0.0043     0.1797
y5     1    23  23.6545   2.6971  0.0248    0.0028    -0.1647
y6     1    40  48.2387   4.4071  0.0505    0.0046    -1.6041
y7     1   167 171.0283   9.2226  0.1791    0.0097    -0.5415
y8     1    33  31.4283   3.4996  0.0329    0.0037     0.3690
y9     1    68  65.4147   5.2158  0.0685    0.0055     0.4452
y10    1    41  32.8901   3.5852  0.0344    0.0038     1.8653
y11    1    47  49.0681   4.4699  0.0514    0.0047    -0.4012
y12    1   185 173.9686   9.3027  0.1822    0.0097     1.4776
y13    1    34  31.9686   3.5528  0.0335    0.0037     0.4752
y14    1    63  66.5393   5.2853  0.0697    0.0055    -0.6073
y15    1    26  33.4555   3.6394  0.0350    0.0038    -1.7087
CONVERGENCE INFORMATION...
    Original counts used.
    iterations = 5 , time elapsed = 0.18
    norm.diff = 1.80924e-09 = dist between last and second
    last iterates.
           Norm diff convergence criterion [1e-06] was met.
    norm.score = 1.61128e-09 = norm of score at last iteration.
           Norm score convergence criterion [1e-06] was met.
```

If we wanted to express the independence model in terms of the global odds ratios, then $\mathrm{h}(\mathbf{m})$ in (5.30) equals $\mathrm{h}(\mathbf{m}) = \mathbf{C}\log(\mathbf{M}\mathbf{m})$, with matrices $\mathbf{C}$ and $\mathbf{M}$ appropriately defined. Function `global.odds.DM()` of the web appendix (see Sect. A.3.2) returns these two matrices for tables of size $I \times J$. The procedure above had to be adjusted as follows:

```
> C <- global.odds.DM(NI,NJ)$C;  M <- global.odds.DM(NI,NJ)$M
> h.fct <- function(m) {C%*%log(M%*%m)}
> ind.glob <- mph.fit(y, constraint=h.fct, strata=1)
```

5.6.2 *Example 3.3 by* `mph`

The hypothesis of homogeneous association (3.7) in $2 \times 2 \times K$ tables can be treated also in the GLLM framework, expressed by (5.28) with $\mathbf{m}$ the expected cell frequencies under the homogeneous association hypothesis expanded in a $4K \times 1$ matrix form, $\mathbf{X} = (1)_{K\times 1}$, and $\mathbf{C}$ the $K \times 4K$ block-diagonal matrix $\mathbf{C} = diag(\mathbf{C}_1, \ldots, \mathbf{C}_K)$ matrix with $\mathbf{C}_k = \mathbf{C}_0 = (1, -1, -1, 1)$, for $k = 1, \ldots, K$. $\mathbf{C}_0$ is the matrix for constructing the log odds ratios when applied on $\mathbf{m}$. It has this form, provided that the expected frequency table is expanded by columns. In this case the parameter is scalar and is equal to the assumed log odds ratio for all partial 2×2 tables under the homogeneous association hypothesis, i.e., $\beta = \log \theta$.

This approach is illustrated in `mph` for Example 3.3, as follows. Function `bdiag()` of library `Matrix` is applied to produce the block-diagonal matrix $\mathbf{C}$.

```
> source("c://Program Files//R//mph.Rcode.txt");
freq <- c(79,68,5,17,89,221,4,46,141,77,6,18,45,26,29,21,81,112,
          3,11,168,51,13,12);
y<- matrix(freq);  K <- 6;  X1 <- matrix(rep(1, K));
library(Matrix);  C0<-c(1, -1, -1, 1);
C <- t(bdiag(C0,C0,C0,C0,C0,C0)); # 6×6 block-diagonal matrix
L.fct <- function(m){C%*%log(m)};
mph.out <- mph.fit(y=y,strata=K,L.fct=L.fct,X=X1);
mph.summary(mph.out,cell.stats=T,model.info=T)
```

From the observed output we have that $G^2 = 7.950$ (p-value=0.159, df=5) and $X^2 = 7.896$ (p-value=0.162, df=5) while the ML estimate of the common under homogeneous association log odds ratio is $\hat{\beta} = 1.0759$, i.e., $\hat{\theta} = 2.9326$. This model is equivalent to the homogeneous association log-linear model applied on the cell frequencies (see Sect. 4.6.1.1). Recall from Sect. 3.3.4 that the Mantel–Haenszel estimate was $\hat{\theta}_{MH} = 2.96$.

5.7 Overview and Further Reading

The classical reference for GLMs is McCullagh and Nelder (1989). Additionally, a comprehensive reference is Fahrmeir and Tutz (2001). For application of GLMs in S-Plus and `R`, we refer to Venables and Ripley (2002, Chap. 7). Dobson and Barnett (2008) provide an easy to follow introduction to GLMs, with theoretical counterpart

but focusing on the analysis of particular types of data and their implementation in standard software, categorical data included. They consider also Bayesian analysis and Markov chain Monte Carlo (MCMC) methods to fit GLMs. A formulation and presentation of models for categorical data through the GLM family can be found in Agresti (2007, 2013).

GLMs have been extended in various directions, like for incorporating nonconstant variance, modeling dispersion, or generalizing the link function (McCullagh and Nelder 1989). In categorical data context, characteristic cases are, for example, the consideration of a negative binomial instead of a Poisson response or the introduction of dispersion effect in the cumulative link model (McCullagh 1980).

The Fisher information matrix plays an important role in statistics in many different aspects, the two most characteristic being in determining the variance of an estimator and the "noninformative" priors determination in the Bayesian setup. Spall (2005) reviews basic principles associated with the information matrix and presents a resampling-based method for computing the information matrix.

When the n_i's are small, the residuals are not approximately normal distributed. For such cases the transformed *Anscombe residuals* have been proposed (see McCullagh and Nelder 1989). For a survey on residuals for GLMs, we refer to Pierce and Schafer (1986). For goodness-of-fit testing of GLMs for sparse data, see Farrington (1996).

5.7.1 Incomplete Contingency Tables

Incomplete tables attracted researchers' attention very early. Stigler (1992), in an interesting and enlightening historical review, points out that in 1913, Karl Pearson was the first to consider the independence model for two-way incomplete tables. The historical fingerprint data set in Waite (1915) contains structural and sampling zeros while Harris and Treloar (1927) and Harris and Tu (1929) face for incomplete tables the problems occurring in the applicability of the contingency coefficient.

The existence of ML estimates for models considered on incomplete tables became a central issue in the late 1960s and 1970s. The most well-known model for incomplete tables is the QI model, presented in Sect. 5.5. Very popular, especially in the context of rater agreement and mobility tables, is the QI model for square tables having the main-diagonal entries missing or excluded. The key reference for the QI model is Goodman (1968), though the QI model for diagonal truncated square tables had been considered earlier by Savage and Deutsch (1960) and Goodman (1963a) in transaction flows analysis and White (1963) and Goodman (1965) in mobility table analysis. Fundamental papers in developing inference for QI in the log-linear model setup were Bishop and Fienberg (1969), Fienberg (1970a), and Haberman (1973a), with the last two providing conditions for existence of unique nonzero ML estimates. The QI model is discussed in detail in Bishop et al. (1975).

Interesting is the approach of Fienberg (1969) that locates the cells exhibiting interaction, when the number of such cells is relatively small compared with the total number of cells in the table, and applies then the QI model, excluding these cells. Mantel (1970) focused on determining the appropriate degrees of freedom and considered, beyond independence, also symmetry testing for incomplete square tables. Goodman (1971a) proposed a test procedure for testing the hypothesis of QI simultaneously for several different subsets of the cells of a table. Enke (1977) considered incomplete two-way tables of special structures that are decomposed to separable tables and lead to closed form MLEs. For the ML estimation of the diagonal truncated independence model, Morgan and Titterington (1977) compared the performance of the EM, Newton–Raphson, and iterative scaling algorithms, concluding empirically that the last is the least efficient method.

Another special type of incomplete square tables are the triangular tables. Such form of incomplete tables occurred already in Waite (1915), while is special referred in Goodman (1968) and Bishop and Fienberg (1969). Special on triangular QI are Goodman (1979a, 1994) and Altham (1975), who considered also the Bayesian analysis with conjugate prior. For ordinal triangular tables, Sarkar (1989) interpreted QI in terms of likelihood ratio dependence and Tsai and Sen (1995) provided an alternative test of QI. We considered in Sect. 5.5 the problem of incorporating structural zeros in the simple independence model for two-way tables. The diagonal and triangular truncated tables will be presented in Sect. 9.3.

Colombo and Ihm (1988) applied the QI model in an unusual context to estimate failure rates of components classified by two qualitative covariates. QI allows for different operating times in the various cells, zero operating time included.

Incomplete tables may occur in tables of higher dimension and of more complex association structures. Klimova et al. (2012) introduce a general family of models for contingency tables, the rational models, which provide a unified framework for analysis of complete and incomplete tables by log-linear models and others, like association models (Chap. 6) and rater agreement models (Sect. 9.5.2). They provide sufficient conditions for the existence of the ML estimates under this general model and prove the classical equivalence between the Poisson and multinomial likelihoods.

A nice review of the literature on the sensitivity analysis of overparameterized models for incomplete categorical data, Bayesian and frequentist, is provided by Poleto et al. (2011).

5.7.2 Marginal Distributions Modeling

Marginal models have been mainly developed by Lang and Agresti (1994), Lang (1996a), Lang et al. (1999), and Bergsma and Rudas (2002a,b). Their approach is based on earlier work by Haber (1985) and Haber and Brown (1986). Bartolucci et al. (2007) generalized the model of Bergsma and Rudas (2002a) to allow for global and continuation type logits, which may be more adequate for ordinal

data analysis. Rudas et al. (2010) formed conditional independence models in a marginal log-linear parameterization. Becker et al. (1998) explored similarities and differences between standard log-linear and marginal models with special emphasis on square tables and reference to multi-way tables as well in the social sciences framework. For a detailed presentation of marginal models and their features, we refer to the book by Bergsma et al. (2009).

Marginal models are applied for modeling repeated (or clustered) categorical data (see also Sect. 9.7.4).