GENLOG
Poisson Loglinear Model

This chapter describes the algorithm to calculate maximum-likelihood estimates for the Poisson loglinear model. This algorithm is applicable only to aggregated data. See “Aggregated Data (Poisson)” on p. 16 for producing aggregated data.

Notation

The following notation is used throughout this chapter unless otherwise stated:

B	Generic categorical dependent (response) variable. Its categories are indexed by an array of integers.
r	Number of categories of B.
p	Number of nonredundant (nonaliased) parameters.
i	Generic index for the category of B.
k	Generic index for the parameter.
n_i	Observed count in the ith response of B.
N	Total observed count, equal to $\sum_{i=1}^{r} n_i$.
m_i	Expected count.
z_i	Cell structure value.
β_k	The kth nonredundant parameter.
β	Vector of $\left(\beta_0, \beta_1, \ldots, \beta_p\right)'$.
x_{ik}	An element in the ith row and the kth column of the design matrix.

- Because of the Poisson distribution assumptions, the logit model is not applicable for a Poisson distribution.
- The Poisson distribution is available in GENLOG only.

Components of the Model

There are two components in a loglinear model: the random component and the systematic component.

Random Component

The random component describes the joint distribution of the counts.

- The count $\{n_i\}$ has a Poisson distribution with parameter m_i.
- The counts n_i and n_i' are independent if $i \neq i'$.
- The joint probability distribution of $\{n_i\}$ is the product of these r independent Poisson distributions. The probability density function is

$$\prod_{i=1}^{r} \frac{m_i^{n_i} e^{-m_i}}{n_i!}$$

- The expected count is $\mathrm{E}(n_i) = m_i$.
- The covariance is

$$\mathrm{cov}(n_i, n_i') = \begin{cases} m_i & \text{if } i = i' \\ 0 & \text{if } i \neq i' \end{cases}$$

Systematic Component

The systematic component describes the linkage function between the expected counts and the parameters. The expected counts are themselves functions of parameters. For $i = 1, \ldots, r$,

$$m_i = \begin{cases} z_i e^{\beta_0 + v_i} & \text{if } z_i > 0 \\ 0 & \text{if } z_i \leq 0 \end{cases}$$

where

$$v_i = \sum_{k=1}^{p} x_{ik}\beta_k \qquad \textbf{(1)}$$

Since there are no constraints on the observed counts, β_0 is a free parameter in a Poisson loglinear model.

Cell Structure Values

Cell structure values play two roles in SPSS loglinear procedures, depending on their signs. If $z_i > 0$, it is a usual weight for the corresponding cell and $\log(z_i)$ is sometimes called the **offset**. If $z_i \leq 0$, a **structural zero** is imposed on the cell ($B = i$). Contingency tables containing at least one structural zero are called **incomplete tables**. If $n_i = 0$ but $z_i > 0$, the cell ($B = i$) contains a **sampling zero**. Although SPSS still considers a structural zero part of the contingency table, it is not used in fitting the model. Cellwise statistics are not computed for structural zeros.

Maximum-Likelihood Estimation

The multinomial log-likelihood is

$$L(\beta) = L(\beta_0, \ldots, \beta_p) = \text{constant} + \sum_{i=1}^{r} \left(n_i \log(m_i) - m_i\right) \qquad \textbf{(2)}$$

Likelihood Equations

It can be shown that

$$\frac{\partial L}{\partial \beta_0} = \sum_{i=1}^{r} (n_i - m_i)$$

$$\frac{\partial L}{\partial \beta_k} = \sum_{i=1}^{r} (n_i - m_i) x_{ik} \qquad k = 1, \ldots, p$$

Let $\mathbf{g}(\beta) = \left(g_0(\beta),\ldots,g_p(\beta)\right)'$ be the $(p+1)\times 1$ gradient vector with

$$g_k(\beta) = \frac{\partial L}{\partial \beta_k}$$

The maximum-likelihood estimates $\hat{\beta} = \left(\hat{\beta}_0,\ldots,\hat{\beta}_p\right)'$ are regarded as a solution to the vector of likelihood equations:

$$\mathbf{g}(\beta) = 0 \qquad \textbf{(3)}$$

Hessian Matrix

The likelihood equations are nonlinear functions of β. Solving them for $\hat{\beta}$ requires an iterative method. The Newton-Raphson method is used. It can be shown that

$$\frac{\partial^2 L}{\partial^2 \beta_0} = -\sum_{i=1}^{r} m_i$$

$$\frac{\partial^2 L}{\partial \beta_0 \partial \beta_1} = -\sum_{i=1}^{r} m_i x_{il}$$

$$\frac{\partial^2 L}{\partial \beta_k \partial \beta_0} = -\sum_{i=1}^{r} m_i x_{ik}$$

$$\frac{\partial^2 L}{\partial \beta_k \partial \beta_l} = -\sum_{i=1}^{r} m_i x_{ik} x_{il}$$

Let $\mathbf{H}(\beta)$ be the $(p+1)\times(p+1)$ information matrix, where $-\mathbf{H}(\beta)$ is the Hessian matrix of (2). The elements of $\mathbf{H}(\beta)$ are

$$h_{kl}(\beta) = \frac{\partial^2 L}{\partial \beta_k \partial \beta_l} \qquad k = 0,\ldots,p \text{ and } l = 1,\ldots,p \qquad \textbf{(4)}$$

Note: $\mathbf{H}(\beta)$ is a symmetric positive definite matrix. The asymptotic covariance matrix of $\hat{\beta}$ is estimated by $\mathbf{H}^{-1}(\beta)$.

Newton-Raphson Method

Let $\beta^{(s)}$ denote the *s*th approximation for the solution to (3). By the Newton-Raphson method,

$$\beta^{(s+1)} = \beta^{(s)} + \mathbf{H}^{-1}\left(\beta^{(s)}\right)\mathbf{g}\left(\beta^{(s)}\right)$$

Define $\mathbf{q}(\beta) = \mathbf{H}(\beta)\beta + \mathbf{g}(\beta)$. The *k*th element of $\mathbf{q}(\beta)$ is

$$q_k(\beta) = \sum_{i=1}^{r} \eta_i x_{ik} \qquad \textbf{(5)}$$

where

$$\eta_i = \begin{cases} m_i v_i + (n_i - m_i) & \text{if } z_i > 0 \text{ and } \mathrm{m_i} > 0 \\ 0 & \text{otherwise} \end{cases}$$

Then

$$\mathbf{H}\left(\beta^{(s)}\right)\beta^{(s+1)} = \mathbf{q}\left(\beta^{(s)}\right) \qquad \textbf{(6)}$$

Thus, given $\beta^{(s)}$, the $(s+1)$th approximation $\beta^{(s+1)}$ is found by solving the system of equations in (6).

Initial Values

SPSS uses the $\beta^{(0)}$, which corresponds to a saturated model as the initial value for β. Then the initial estimates for the expected cell counts are

$$m_i^{(0)} = \begin{cases} n_i + \Delta & \text{if } z_i > 0 \\ 0 & \text{if } z_i \le 0 \end{cases} \qquad \textbf{(7)}$$

where $\Delta \ge 0$ is a constant.

Note: For saturated models, SPSS adds Δ to n_i if $z_i > 0$. This is done to avoid numerical problems in the case that some observed counts are 0. We advise users to set Δ to 0 whenever all observed counts (other than structural zeros) are positive.

The initial values for η_i are

$$\eta_i^{(0)} = \begin{cases} m_i \log\left(m_i^{(0)} / z_i\right) + \left(n_i - m_i^{(0)}\right) & \text{if } z_i > 0 \text{ and } m_i^{(0)} > 0 \\ 0 & \text{otherwise} \end{cases} \quad \textbf{(8)}$$

Stopping Criteria

SPSS checks the following conditions for convergence:

1. $\max_i\left(\left|m_i^{(s+1)} - m_i^{(s)}\right| / m_i^{(s)}\right) < \varepsilon$ provided that $m_i^{(s)} > 0$
2. $\max_i\left(\left|m_i^{(s+1)} - m_i^{(s)}\right|\right) < \varepsilon$
3. $\sqrt{\left(\sum_{k=0}^{p} g_k^2\left(\hat{\beta}\right)\right) / (p+1)} < \varepsilon$

The iteration is said to be **converged** if either conditions 1 and 3 or conditions 2 and 3 are satisfied. The iteration is said to be **not converged** if neither pair of conditions is satisfied within the maximum number of iterations.

Algorithm

The iteration process uses the following steps:

1. Calculate $m_i^{(0)}$ using (7) and $n_i^{(0)}$ using (8).
2. Set $s = 0$.
3. Calculate $\mathbf{H}\left(\beta^{(s)}\right)$ using (4) evaluated at $m_i^{(s)}$, and $\mathbf{q}\left(\beta^{(s)}\right)$ using (5) evaluated at $\eta_i = \eta_i^{(s)}$.
4. Solve for $\beta^{(s+1)}$ using (6).
5. Calculate $v_i^{(s+1)} = \sum_{k=1}^{p} x_{ik}\beta_k^{(s+1)}$ and

$$m_i^{(s+1)} = \begin{cases} z_i e^{\beta_0^{(s+1)} + v_i^{(s+1)}} & \text{if } z_i > 0 \\ 0 & \text{if } z_i \le 0 \end{cases}$$

6. Check whether the stopping criteria are satisfied. If yes, iteration stops and the process is declared converged. Otherwise continue.
7. Increase *s* by 1 and check whether the maximum iteration has been reached. If yes, iteration stops and the process is declared not converged. Otherwise, repeat steps 3-7.

Estimated Cell Counts

The estimated expected count is

$$\hat{m}_i = \begin{cases} z_i e^{\hat{\beta}_0 + \hat{v}_i} & \text{if } z_i > 0 \\ 0 & \text{if } z_i \le 0 \end{cases}$$

where

$$\hat{v}_i = \sum_{k=1}^{p} x_{ik} \hat{\beta}_k$$

Goodness-of-Fit Statistics

The Pearson chi-square statistic is

$$X^2 = \sum_{i=1}^{r} X_i^2$$

where

$$X_i^2 = \begin{cases} \left(n_i - \hat{m}_i\right)^2 / \hat{m}_i & \text{if } z_i > 0,\ n_i > 0,\ \text{and } \hat{m}_i > 0 \\ \text{SYSMIS} & \text{if } z_i > 0,\ n_i > 0,\ \text{and } \hat{m}_i = 0 \\ 0 & \text{if } z_i \le 0 \text{ or } n_i = \hat{m}_i \end{cases}$$

If any X_i^2 is system missing, then X^2 is also system missing.

The likelihood-ratio chi-square statistic is

$$G^2 = 2\sum_{i=1}^{r} G_i^2$$

where

$$G_i^2 = \begin{cases} n_i\left(\log\left(n_i / \hat{m}_i\right)\right) - \left(n_i - \hat{m}_i\right) & \text{if } z_i > 0,\ n_i > 0,\ \text{and } \hat{m}_i > 0 \\ \text{SYSMIS} & \text{if } z_i > 0,\ n_i > 0,\ \text{and } \hat{m}_i = 0 \\ \hat{m}_i & \text{if } z_i > 0,\ n_i = 0,\ \text{and } \hat{m}_i > 0 \\ 0 & \text{if } z_i \le 0 \text{ or } n_i = \hat{m}_i \end{cases}$$

If any G_i^2 is system missing, then G^2 is also system missing.

Degrees of Freedom

The degrees of freedom for each statistic is defined as $a = r - 1 - p - E$, where E is the number of cells with $z_i \le 0$ or $\hat{m}_i = 0$.

Significance Level

The significance level (or the p value) for the Pearson chi-square statistic is $\text{Prob}\left(\chi_a^2 > X^2\right)$ and that for the likelihood-ratio chi-square statistic is $\text{Prob}\left(\chi_a^2 > G^2\right)$. In both cases, χ_a^2 is the central chi-square distribution with a degrees of freedom.

Residuals

Goodness-of-fit statistics provide only broad summaries of how models fit data. The pattern of lack of fit is revealed in cell-by-cell comparisons of observed and fitted cell counts.

Simple Residuals

The **simple residual** of the *i*th cell is

$$r_i = \begin{cases} n_i - \hat{m}_i & \text{if } z_i > 0 \\ \text{SYSMIS} & \text{if } z_i \le 0 \end{cases}$$

Standardized Residuals

The **standardized residual** for the *i*th cell is

$$r_i^S = \begin{cases} \left(n_i - \hat{m}_i\right) / \sqrt{\hat{m}_i} & \text{if } z_i > 0 \text{ and } 0 < \hat{m}_i \\ 0 & \text{if } z_i > 0 \text{ and } n_i = \hat{m}_i \\ \text{SYSMIS} & \text{otherwise} \end{cases}$$

Notice that $\sum_{i=1}^{r} \left(r_i^S\right)^2 = X^2$ when all $z_i > 0$. Hence, standardized residuals are also known as Pearson residuals. Although the standardized residuals are asymptotically normal, their asymptotic variances are less than 1.

Adjusted Residuals

The **adjusted residual** is the simple residual divided by its estimated standard error. This statistic for the *i*th cell is

$$r_i^A = \begin{cases} \left(n_i - \hat{m}_i\right) / \sqrt{\hat{m}_i\left(1 - a_{ii}\right)} & \text{if } z_i > 0,\ n_i \neq \hat{m}_i,\ \text{and } \hat{m}_i > 0 \\ 0 & \text{if } z_i > 0 \text{ and } n_i = \hat{m}_i \\ \text{SYSMIS} & \text{otherwise} \end{cases}$$

where

$$a_{ii} = \hat{m}_i \left(h^{00} + 2\sum_{k=1}^{p} x_{ik} h^{k0} + \sum_{k=1}^{p}\sum_{l=1}^{p} x_{ik} x_{il} h^{kl} \right)$$

h^{kl} is the *(k,l)*th element of $\mathbf{H}^{-1}\left(\hat{\beta}\right)$. The adjusted residuals are asymptotically standard normal.

Deviance Residuals

Pierce and Schafer (1986) and McCullagh and Nelder (1989) define the signed square root of the individual contribution to the G^2 statistic as the **deviance residual**. This statistic for the *i*th cell is

$$r_i^D = \text{sign}\left(n_i - \hat{m}_i\right)\sqrt{d_i}$$

where

$$d_i = \begin{cases} 2\left(n_i\left(\log\left(n_i / \hat{m}_i\right)\right) - \left(n_i - \hat{m}_i\right)\right) & \text{if } z_i > 0,\ \hat{m}_i > 0,\ \text{and } n_i > 0 \\ 2\hat{m}_i & \text{if } z_i > 0,\ \hat{m}_i \geq 0,\ \text{and } n_i = 0 \\ 0 & \text{if } z_i > 0 \text{ and } n_i = \hat{m}_i \\ \text{SYSMIS} & \text{otherwise} \end{cases}$$

When all $z_i > 0$, $\sum_{i=1}^{r}\left(r_i^D\right)^2 = G^2$.

Generalized Residual

Consider a linear combination of the cell counts $\sum_{i=1}^{r} d_i n_i$, where d_i are real numbers.

The estimated expected value is

$$\sum_{i=1}^{r} d_i \hat{m}_i$$

The simple residual for this linear combination is

$$\sum_{i=1}^{r} d_i (n_i - \hat{m}_i)$$

The standardized residual for this linear combination is

$$\frac{\sum_{i=1}^{r} d_i (n_i - \hat{m}_i)}{\sqrt{\sum_{i=1}^{r} d_i^2 \hat{m}_i}}$$

Using the results in Christensen (1990, p. 227), the adjusted residual for this linear combination is

$$\frac{\sum_{i=1}^{r} d_i (n_i - \hat{m}_i)}{\sqrt{V}}$$

where

$$V = \sum_{i=1}^{r}\sum_{j=1}^{r} d_i d_j \hat{m}_i \left(\delta_{ij} - a_{ij}\right)$$
$$= \sum_{i=1}^{r} d_i^2 \hat{m}_i - \sum_{i=1}^{r}\sum_{j=1}^{r} d_i d_j a_{ij} \hat{m}_i$$

where

$$a_{ij} = \hat{m}_i \left(h^{00} + \sum_{k=1}^{p} \left(x_{ik} + x_{jk}\right) h^{k0} + \sum_{k=1}^{p}\sum_{l=1}^{p} x_{ik} x_{jl} h^{kl} \right)$$

h^{kl} is the *(k,l)*th element of $\mathbf{H}^{-1}(\beta)$.

Generalized Log-Odds Ratio

Consider a linear combination of the natural logarithm of cell counts

$$\sum_{i=1}^{r} d_i \log(m_i) \qquad \textbf{(9)}$$

where d_i are real numbers with the restriction

$$\sum_{i=1}^{r} d_i = 0$$

The quantity (9) is estimated by

$$\sum_{i=1}^{r} d_i \log(\hat{m}_i) = \sum_{i=1}^{r} d_j \log(z_i) + \sum_{i=1}^{r}\sum_{k=1}^{p} d_i x_{ik} \hat{\beta}_k$$

The variance is

$$\text{var}\left(\sum_{i=1}^{r} d_i \log(\hat{m}_i)\right) = \sum_{k=1}^{p}\sum_{l=1}^{p} w_k w_l h^{kl} \qquad \textbf{(10)}$$

where

$$w_k = \sum_{i=1}^{r} d_i x_{ik} \qquad k = 1,\ldots,p$$

Wald Statistic

The null hypothesis is

$$H_0: \sum_{i=1}^{r} d_i \log(m_i) = 0$$

The Wald statistic is

$$W = \frac{\left(\sum_{i=1}^{r} d_i \log(\hat{m}_i)\right)^2}{\sum_{k=1}^{p}\sum_{l=1}^{p} w_k w_l h^{kl}}$$

Under H_0, W asymptotically distributes as a chi-square distribution with 1 degree of freedom. The significance level is $\text{Prob}\left(\chi_1^2 \geq W\right)$. *Note*: W will be system missing if (10) is 0.

Asymptotic Confidence Interval

The asymptotic $(1-\alpha)\times 100\,\%$ confidence interval for (9) is

$$\sum_{i=1}^{r} d_i \log(\hat{m}_i) \pm z_{\alpha/2} \sqrt{\sum_{k=1}^{p} \sum_{l=1}^{p} w_k w_l h^{kl}}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ point of the standard normal distribution. The default value of α is 0.05.

Aggregated Data (Poisson)

This section shows how data are aggregated for a Poisson distribution. The following notation is used in this section:

v_i	The number of SPSS cases for $B = i \quad (i = 1, \ldots, r)$
n_{is}	The *s*th SPSS caseweight for $B = i \quad (s = 1, \ldots, v_i)$
x_{is}	Covariate
z_{is}	Cell weight
c_{is}	GRESID coefficient
e_{is}	GLOR coefficient
v_i^+	The number of positive z_{is} (cell weights) for $1 \le s \le v_i$

The cell count is

$$n_i = \begin{cases} \sum_{1 \le s \le v_i}^{*} n_{is}^+ & \text{if } v_i^+ > 0 \\ 0 & \text{if } v_i = 0 \text{ or } v_i^+ = 0 \end{cases}$$

where

$$n_{is}^+ = \begin{cases} n_{is} & \text{if } n_{is} > 0 \text{ and } z_{is} > 0 \\ 0 & \text{if } n_{is} \le 0 \text{ and } z_{is} > 0 \end{cases}$$

and $\sum_{1 \le s \le v_i}^{*}$ means summation over the range of s with the terms $z_{is} > 0$.

The cell weight value is

$$z_i = \begin{cases} \sum_{1 \le s \le v_i}^{*} n_{is}^{+} z_{is} / n_{ij} & \text{if } n_i > 0 \text{ and } v_i^{+} > 0 \\ \sum_{1 \le s \le v_i}^{*} z_{is} / v_i^{+} & \text{if } n_i = 0 \text{ and } v_i^{+} > 0 \\ 0 & \text{if } v_i^{+} = 0 \\ 1 & \text{if } v_i = 0 \end{cases}$$

If no variable is specified as the cell weight variable, then all cases have unit cell weights by default.

The cell covariate value is

$$x_{ij} = \begin{cases} \sum_{1 \le s \le v_i}^{*} n_{is}^{+} x_{is} / n_i & \text{if } n_i > 0 \text{ and } v_i > 0 \\ \sum_{1 \le s \le v_i}^{*} x_{is} / v_i^{+} & \text{if } n_i = 0 \text{ and } v_i^{+} > 0 \\ 0 & \text{if } v_i^{+} = 0 \text{ or } v_i = 0 \end{cases}$$

The cell GRESID coefficient is

$$c_i = \begin{cases} \sum_{1 \le s \le v_i}^{*} n_{is}^{+} c_{is} / n_i & \text{if } n_i > 0 \text{ and } v_i > 0 \\ \sum_{1 \le s \le v_i}^{*} c_{is} / v_i^{+} & \text{if } n_i = 0 \text{ and } v_i^{+} > 0 \\ 0 & \text{if } v_i^{+} = 0 \text{ or } v_i = 0 \end{cases}$$

There are no defaults for the GRESID coefficients.

The cell coefficient is

$$e_i = \begin{cases} \sum_{1 \le s \le v_i}^{*} n_{is}^{+} e_{is} / n_i & \text{if } n_i > 0 \text{ and } v_i > 0 \\ \sum_{1 \le s \le v_i}^{*} e_{is} / v_i^{+} & \text{if } n_i = 0 \text{ and } v_i^{+} > 0 \\ 0 & \text{if } v_i^{+} = 0 \text{ or } v_i = 0 \end{cases}$$

There are no defaults for the GLOR coefficients.

References

Agresti, A. 1990. *Categorical data analysis*. New York: John Wiley & Sons, Inc.

Christensen, R. 1990. *Log-linear models*. New York: Springer-Verlag.

Haberman, S. J. 1973. The analysis of residuals in cross-classified tables. *Biometrics*, 29: 205–220.

McCullagh, P., and Nelder, J. A. 1989. Generalized linear models. 2nd ed. London: Chapman and Hall.

Pierce, D. A., and Schafer, D. W. 1986. Residuals in generalized linear models. *Journal of the American Statistical Association*, 81: 977–986.