Poisson Regression

For response variables that have counts or frequencies as outcomes it is often reasonable to assume an underlying **Poisson** distribution and describe the impact of **explanatory variables** on their means by some **regression** function. Poisson regression models, as a widely applicable class of models particularly useful in biostatistics, emerged in the late 1970s; see, for example, [6, 11, 21–25, 28, 29, 31], and [32].

As an example consider the data given in Table 1, taken from [27]. Randomly chosen household members from a **probability sample** of Oakland, CA, were asked to note which stressful events had occurred within the last 18 months and to report the month of occurrence of these events. A scattergram of the data indicates a decline of recalls as events lie farther in the past, possibly due to the fallibility of human memory (see Figure 1). To define a Poisson regression model, assume that (i) the number of recalls is a random variable Y distributed as Poisson with mean μ, and (ii) μ is some function of X, the number of months before interview. Plotting logarithms of frequencies against months suggests a linear relationship

$$\log \mu = \alpha + \beta x.$$

For this **loglinear model**, the mean satisfies the exponential relationship,

$$\mu = \exp(\alpha + \beta x) = e^{\alpha}(e^{\beta})^{x}.$$

A one-unit increase in X has a multiplicative effect of e^{β} on μ, i.e. the mean of Y at $x + 1$ equals the mean of Y at x multiplied by e^{β}.

Most of the widely available software packages are capable of fitting **generalized linear models**, and can be used to obtain **maximum likelihood** estimates for the parameters of Poisson regression models as well. For these data one finds $\hat{\alpha} = 2.803$ and $\hat{\beta} = -0.0838$; hence

$$\hat{\mu} = 16.5 \times 0.920^{x},$$

indicating a negative trend in time.

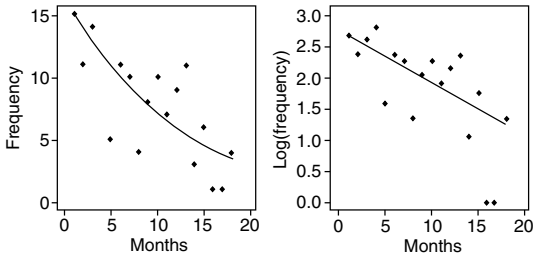

Figure 1 Scattergram of observed frequencies and their logarithms against months before interview. Solid lines represent fitted means and respective values for the linear predictor for the Poisson regression model mentioned in the text

Using the relationship between the **multinomial** and conditional Poisson distributions, this is shown to be equivalent to an exponential decay model for the probability of remembering an event. For a more detailed discussion, see [27] or [33].

Definition

To define the basic version of a Poisson regression model, suppose that we have observations $y_1, \ldots, y_n$ for the response variable $Y_1, \ldots, Y_n$, assumed to be independently distributed Poisson variates with means $\mu_1, \ldots, \mu_n$, i.e.

$$f(y_i|\mu_i) = \frac{\mu_i^{y_i}}{y_i!}\exp(-\mu_i). \qquad (1)$$

The systematic component of the model is specified by some regression function η, depending on regression parameters $\beta_1, \ldots, \beta_k$, with each component relating values $x_{i1}, \ldots, x_{ik}$ of explanatory variables to respective means, i.e.

$$\mu_i = \eta_i(\beta) = \eta_i(x_{i1}, \ldots, x_{ik}; \beta_1, \ldots, \beta_k). \qquad (2)$$

Table 1 Distribution by months prior to interview of stressful events reported from subjects: 147 subjects reporting exactly one stressful event in the period from 1 to 18 months prior to interview. Reprinted from [27, p. 3] by permission of Academic Press, Inc.

Months	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Number	15	11	14	17	5	11	10	4	8	10	7	9	11	3	6	1	1	4

Often, this relationship is such that some monotone transformation g of the means is connected to a *linear predictor* of explanatory variables,

$$g(\mu_i) = \sum_{j=1}^{k} x_{ij}\beta_j.$$

In this situation g is called the *link function* and the model defined in this manner is an instance of a **generalized linear model** (see [35] and [36] or, for an introductory text, [18]). For $\eta_i(\beta) = \exp\left(\sum_{j=1}^{k} x_{ij}\beta_j\right)$ we have the familiar loglinear model,

$$\log \mu_i = \sum_{j=1}^{k} x_{ij}\beta_j.$$

For the model specified by the stochastic component (1) and regression function (2), the log **likelihood** function is written as

$$\ell_y(\beta) = \sum_{i=1}^{n} \{y_i \log[\eta_i(\beta)] - \eta_i(\beta) - \log(y_i!)\}. \quad (3)$$

It may be worthwhile noting that this reduces to a k-parameter **exponential family** log likelihood,

$$\ell_y(\beta) = \sum_{j=1}^{k} \left(\sum_{i=1}^{n} x_{ij} y_i\right) \beta_j - \sum_{i=1}^{n} \exp\left(\sum_{j=1}^{k} x_{ij}\beta_j\right) - \sum_{i=1}^{n} \log(y_i!), \quad (4)$$

with jointly **sufficient statistics** $\sum_{i=1}^{n} x_{ij} y_i$, $j = 1, \ldots, k$, if the model is log linear.

Some Special Cases

Loglinear Models for Contingency Tables

Suppose, in obvious notation, y_{ij} with indices $i = 1, \ldots, I$ and $j = 1, \ldots, J$ form a two-dimensional **contingency table**, according to some classifying factors A and B having I and J categories, respectively. A common method for analyzing data of this kind is to assume that cell frequencies Y_{ij} are independently distributed as Poisson and to use loglinear models, where in an **analysis-of-variance**-like fashion logarithms of expected cell frequencies μ_{ij} are assumed to be sums of several effects, e.g. for the **multiplicative model**,

$$\log(\mu_{ij}) = \beta_o + \beta_i^A + \beta_j^B, \quad (5)$$

subject to some constraints on the βs. Sums of independent Poisson variates are again distributed as Poisson with means equal to the sum of respective means. Row totals Y_{i+}, column totals Y_{+j}, and grand total Y_{++} are, therefore, Poisson variates with means $\mu_{i+} = \mu_{i1} + \cdots + \mu_{iJ}$, $\mu_{+j} = \mu_{1j} + \cdots + \mu_{Ij}$, and $\mu_{++} = \sum_{i,j} \mu_{ij}$, respectively. Under the assumption of the multiplicative model these quantities are related by

$$\mu_{ij} = \frac{\mu_{i+}\mu_{+j}}{\mu_{++}},$$

showing that the joint distribution of the contingency table is, in a multiplicative manner, completely determined by the marginal distributions.

The Poisson model assumption implies that marginals are random. If, instead, the total is fixed by the sampling design, it may be more appropriate to assume a multinomial distribution for the table. Formally, the multinomial model can be inferred from the Poisson model by conditioning on the total y_{++} (*see* **Conditional Probability**). For the probability π_{ij} of an observation falling into row i and column j, we then have $\pi_{ij} = \mu_{ij}/\mu_{++}$, and from assuming the multiplicative model (5), it follows that

$$\pi_{ij} = \pi_{i+}\pi_{+j}, \quad (6)$$

where π_{i+} and π_{+j} are the marginal probabilities of an observation falling into row i and column j, respectively. Hence, row and column variables A and B are independently distributed.

Likewise, if row totals are fixed, then each row may be assumed to be multinomially distributed. Again, this can be inferred from the Poisson model by conditioning on the row totals, and the multiplicative model (5) implies identical distributions for the rows – a condition usually called *homogeneity*. It may be worthwhile noting that maximum-likelihood estimates for the parameters in the Poisson models are identical to those obtained for some other sampling designs, such as the multinomial designs just mentioned, making this class of model particularly interesting and useful.

Loglinear models for two- and higher-dimensional contingency tables, used to describe the association and interaction structure connecting the variables, are

Table 2 Number of recurrences of superficial bladder cancer for 31 male patients with grade 2, stage T_1, solitary primary tumors and respective times under observation (in months) by size of primary tumor. Subset of data analyzed in [38]

Size	Recurrences	Time under observation
$\leq$3 cm	1	2, 3, 6, 8, 9, 10, 11, 13, 14, 16, 21, 22, 24, 26, 27
	2	7, 13, 15, 18, 23
	3	20
	4	24
>3 cm	1	1, 5, 17, 18, 25
	2	18, 25
	3	4
	4	19

discussed in the article on **Loglinear Model**. Usually, the goal is to find a **parsimonious** model that fits the data well and allows meaningful substantive interpretation. Most commonly, this search is restricted to **hierarchical models**.

Multiplicative Rate Models

If occurrences of some kind of event are counted over time, then often interest lies in the rate at which events occur. The rate describes the instantaneous risk for an event to happen at a given point in time. To be more specific, the probability of observing exactly one event in the interval ranging from t to $t+h$, divided by its length h, is assumed to tend to some value $\lambda(t)$, as h tends to 0. $\lambda(t)$, as a function of time t, is called the *rate* or *intensity function*.

An important special case, termed the **Poisson process**, assumes that waiting times between successive events are independent and **exponentially distributed** with common mean $1/\lambda$. Here, the rate function is constant over time, $\lambda(t) \equiv \lambda$. Furthermore, the number $Y(t)$ of events that occur up to time t is distributed as Poisson with mean $\mu = \lambda t$. Note that the mean of $Y(t)/t$ equals the rate λ. This suggests a Poisson regression approach

$$\log \lambda = \log\left(\frac{\mu}{t}\right) = \alpha + \beta x$$

for modeling the dependence of the rate function on an explanatory variable X. This can be rewritten as

$$\log \mu = \alpha + \beta x + \log t,$$

with $\log(t)$ as an *offset*, i.e. a variable in the linear predictor, the corresponding regression parameter of which is set equal to 1. Observe that this defines a multiplicative model for the rate function,

$$\lambda = \mathrm{e}^{\alpha}(\mathrm{e}^{\beta})^{x}, \tag{7}$$

with a *baseline rate* $\lambda_0 = \exp(\alpha)$ and proportionality factor $\exp(\beta x)$.

For illustrative purposes a subset of the data analyzed in [38] is reprinted in Table 2. For 31 male patients, who have been treated for superficial bladder cancer, the number of recurrent tumors has been recorded for some time after removal of the primary tumor. Defining X to be 1 for larger primary tumors (>3 cm) and 0 otherwise, and assuming a Poisson process with rate (7), yields parameter estimates $\hat{\alpha} = -1.95$ and $\hat{\beta} = 0.385$. The (baseline) rate for smaller tumors is (estimated as) 0.142, the rate for larger tumors being 1.47 times larger. In terms of waiting times between recurrences, means are estimated as 7.06 and 4.80 months, respectively.

Now suppose that we have recorded, for n individuals, time under observation, t_i, and the number y_i of events occurred. Observation times are assumed to be nonrandom and counts to be mutually independent. We also have a set of explanatory variables $x_{i1}, \ldots, x_{ik}$ available for each subject. Under the assumption of *proportional rates*, $\lambda_i = \lambda_0 \exp\left(\sum_{j=1}^{k} x_{ij}\beta_j\right)$, we have

$$\mu_i = \lambda_i \times t_i = \lambda_0 \exp\left[\log(t_i) + \sum_{j=1}^{k} x_{ij}\beta_j\right]$$
$$= \lambda_0 t_i \prod_{j=1}^{k} \exp(x_{ij}\beta_j), \tag{8}$$

i.e. a loglinear model for the mean of the Poisson process, involving the logarithm of observation times as an "explanatory" variable, with the associated regression parameter fixed at a value of 1.

If the process is such that it can be characterized by a time-varying rate function $\lambda(t)$, it is called a *nonhomogeneous Poisson process*. Writing

$$\Lambda(t) = \int_0^t \lambda(u)\,\mathrm{d}u$$

for the *integrated rate* or *intensity function*, the number of occurrences of the event in period until time point t is again distributed as Poisson, but with mean equal to $\Lambda(t)$. Note that events in nonoverlapping time intervals are independent, but waiting times between successive events are, contrary to the homogeneous process with constant rate, neither identically distributed nor independent. In this situation model (8) can be modified, using a *baseline rate* function $\lambda_0(t|\alpha)$, possibly depending on some additional parameter α, to give

$$\mu_i = \exp\left\{\log[\Lambda_0(t_i|\alpha)] + \sum_{j=1}^{k} x_{ij}\beta_j\right\}.$$

Choosing $\Lambda_0(t|\alpha)$ to be t, t^α, or $\exp(\alpha t)$ corresponds to an exponential, a **Weibull**, or an **extreme value** intensity function, respectively, and results in a loglinear model for the Y_is. Disregarding constant terms, the likelihood function for this model is

$$L(\alpha, \beta) = \prod_{i=1}^{n}\left[\Lambda_0(t_i|\alpha)\exp\left(\sum_{j=1}^{k} x_{ij}\beta_j\right)\right]^{y_i} \times \exp\left[-\Lambda_0(t_i|\alpha)\exp\left(\sum_{j=1}^{k} x_{ij}\beta_j\right)\right]. \quad (9)$$

If times for occurrences of each event were known, a multiplicative term, depending on the parameter α, would be added to (9); see [32].

There is a close connection to **relative risk models**, which are very frequently used in epidemiology. This class of models assumes that risk factors interact in a multiplicative way. See [9] and [12], and, for a critical review, [26].

Proportional Hazard Models for Censored Survival Times

Now suppose that individuals are under observation until either a single event of interest occurs or the period of observations ends for some other reason. For each subject the data are of the form (y_i, c_i), where y_i is the time under observation, and c_i is an indicator variable for **censoring**, taking the value 1 if the event has occurred at time y_i, and the value 0 if the event has not occurred until time y_i. This is a similar situation to the one in the previous example, but with one terminal event that stops the process; interest, however, lies in the analysis of the *survival times* y_i (*see* **Survival Analysis, Overview**).

The distribution of the survival time can be uniquely described by the rate function, in the context of survival analysis usually called **hazard rate** or *force of mortality*. As before, a common approach assumes **proportional hazard** rates,

$$\lambda(y_i|\alpha, \beta) = \lambda_0(y_i|\alpha)\exp\left(\sum_{j=1}^{k} x_{ij}\beta_j\right),$$

with a *baseline hazard* $\lambda_0(y_i|\alpha)$.

Assuming a noninformative censoring mechanism (and continuous survival times), the kernel of the likelihood function is $\prod_{i=1}^{n} f(y_i)^{c_i} \times S(y_i)^{1-c_i}$, where $f(y_i)$ denotes the density for the ith survival time, and $S(y_i) = 1 - F(y_i)$, the *survival function*, i.e. the probability for the ith survival time to exceed y_i. The ratio $f(y_i)/S(y_i)$ is identical to the hazard function. For proportional hazard rates, the likelihood function can therefore be expressed as

$$L_{y,c}(\alpha, \beta) = \prod_{i=1}^{n}\left[\Lambda_0(y_i|\alpha)\exp\left(\sum_{j=1}^{k} x_{ij}\beta_j\right)\right]^{c_i} \times \exp\left[-\Lambda_0(y_i|\alpha)\exp\left(\sum_{j=1}^{k} x_{ij}\beta_j\right)\right],$$

where $\Lambda_0(y_i|\alpha) = \int_0^{y_i} \lambda_0(u|\alpha)\,\mathrm{d}u$ denotes the cumulative baseline hazard rate. Writing, as we did before, $\mu_i = \exp\left\{\log[\Lambda_0(y_i|\alpha)] + \sum_{j=1}^{k} x_{ij}\beta_j\right\}$, $L(\alpha, \beta)$ is the likelihood function for n independent Poisson variates C_i with means μ_i. Aitkin & Clayton [2] used

this fact to bring survival analysis into the framework of generalized linear models (see also [7, 28], and [31]).

If no assumptions on the functional form of the baseline hazard function are made, then this is Cox's proportional hazards model [13, 14] that can be fitted by maximizing a "**partial likelihood**" (*see* **Cox Regression Model**). Another **semiparametric model**, due to Breslow [7], assumes a piecewise exponential distribution for the survival times, the baseline hazard function in this case is constant over prespecified intervals of time (*see* **Grouped Survival Times**). To be more specific, suppose that the time axis is split into intervals $(a_{p-1}, a_p]$, $p = 1, \ldots, P$, with $0 = a_0 < a_1 < \cdots < a_p < a_{p+1} = \infty$. The baseline hazard can now be written as

$$\lambda_0(y|\alpha) = \exp(\alpha_p), \quad \text{if } a_{p-1} < y \leq a_p.$$

To simplify notation, for individual i and interval $(a_{p-1}, a_p]$ the proportional hazard assumption can be expressed in terms of a constant λ_{ip}, where

$$\lambda_{ip} = \exp\left(\alpha_p + \sum_{j=1}^{k} x_{ij}\beta_j\right). \tag{10}$$

Define P_i to be such that y_i is contained in interval $(a_{P_{i-1}}, a_{P_i}]$ and e_{ip} to be the exposure time of individual i in the pth interval, i.e.

$$e_{ip} = \begin{cases} a_p - a_{p-1}, & \text{if } p = 1, \ldots, P_i - 1, \\ y_i - a_{P_{i-1}}, & \text{if } p = P_i \end{cases}.$$

Also, introduce an extended censoring indicator variable to be

$$c_{ip} = \begin{cases} 1, & \text{if } p = 1, \ldots, P_i - 1, \\ c_i, & \text{if } p = P_i. \end{cases}$$

Disregarding constant terms, the likelihood function is then

$$L_{y,c}(\alpha, \beta) = \prod_{i=1}^{n} \prod_{p=1}^{P_i} (\lambda_{ip} e_{ip})^{c_{ip}} \exp(-\lambda_{ip} e_{ip}),$$

where λ_{ip} is defined by (10). Since this is a Poisson likelihood for the "counts" c_{ip}, the *piecewise exponential model* reduces to a loglinear model. If intervals are chosen such that their endpoints correspond to observed times of death, i.e. t_is with $c_i = 1$, then maximum likelihood estimates for the regression parameters β are found to be close to those obtained from the Cox model; see [3] and [39].

For an example consider the data printed in Table 3. For 33 patients treated for papillary thyroid carcinoma, survival time, censoring indicator, age, and gender are reported. This is a small subset of the data analyzed in [30]. For cutpoints $a_1 = 0.5$, $a_2 = 1$,

Table 3 Survival times: *time* in years, censoring indicator *cens* (= 0 for censored), *gender* (1 for male), and *age* for 33 patients treated for papillary thyroid carcinoma. Subset of data analyzed in [30]

Time	Cens	Gender	Age	Time	Cens	Gender	Age
27.42	0	1	21	2.33	0	1	76
8.50	1	2	31	1.33	0	1	46
0.13	1	1	62	0.08	1	2	84
0.83	1	1	53	2.83	0	2	69
5.92	0	2	52	2.25	1	2	90
1.92	0	2	67	0.25	1	2	52
0.92	1	1	73	3.42	0	2	71
11.67	0	2	56	1.92	1	2	75
0.17	1	1	57	3.00	1	1	69
5.00	1	2	71	1.00	1	1	75
0.08	1	1	53	8.50	1	2	73
0.08	1	2	53	4.17	0	2	36
0.92	1	2	48	3.50	1	1	38
5.08	1	2	65	1.25	1	2	69
5.42	1	2	49	0.33	1	2	77
0.25	0	1	61	0.67	1	1	87
0.17	1	1	71				

$a_3 = 2$, and $a_4 = 3$, the piecewise-constant baseline hazard function is, up to a constant, estimated as

$$\lambda_0(y) = \begin{cases} 0.45, & \text{if } y \leq 0.5, \\ 0.36, & \text{if } 0.5 < y \leq 1, \\ 0.11, & \text{if } 1 < y \leq 2, \\ 0.16, & \text{if } 2 < y \leq 3, \\ 0.20, & \text{if } 3 < y. \end{cases}$$

Regression parameters, estimated for gender and age, are -0.70 and 0.04, respectively.

Log-nonlinear Models

While loglinear models do have some desirable properties, it may not always be possible to find a parameterization such that the regression function is linear on the log scale. An example of this is given in [20], using a log-logistic regression function. The data come from a **radioimmunoassay**, a widely used technique to measure the quantity of a given biological substance by identifying the amount of a radioactive labeled antibody from a reagent by subsamples of increasing concentration. The response variable is the amount of radioactive material remaining measured in counts per minute. If these are very large, a normal distribution for the counts may be assumed, but if this is not the case, an underlying Poisson distribution seems to be more appropriate. For counts $y_1, \ldots, y_n$ and concentrations $x_1, \ldots, x_n$ a regression function of the form

$$\eta_i(\beta_1, \ldots, \beta_4) = \beta_1 + \frac{\beta_2}{1 + \exp\{-[\beta_3 + \beta_4 \log(x_i)]\}} \quad (11)$$

can be used to describe the relationship between mean counts and concentrations. Note that this model cannot be transformed into a loglinear one.

Other examples of log-nonlinear models arise frequently in the analysis of contingency tables, when specific structure in the data suggests inclusion of nonlinear **interaction** effects into the regression function. See, for instance, [1, pp. 287–293].

Likelihood Inference

When adopting a modeling approach it seems to be natural to estimate the parameters of a model by maximizing the likelihood function, or, equivalently, its logarithm. The likelihood function contains all the relevant information about the mechanism that generated the data as well as the data actually observed. The larger its value the stronger the support given, by the data, to the corresponding value of the parameters. When dealing with Poisson regression models, maximum-likelihood estimation is, by far, the most often used method to obtain estimates for the unknown parameters.

Poisson regression models as defined above are instances of curved exponential family models; even if the model is loglinear, it is an exponential family model. So a much more general theory applies to this class of models. Here, only the special case will be considered. Readers interested in a general and detailed treatment are referred to, for example, Barndorff-Nielsen & Cox [5].

To maximize the log likelihood function one usually calculates partial derivatives with respect to all the parameters, sets them equal to 0, and solves this system of equations for the unknowns. For a Poisson regression model with log likelihood (3), the *estimating equations*

$$u_j(\beta) = \sum_{i=1}^{n} \frac{\partial}{\partial \beta_j} \eta_i(\beta) \frac{1}{\eta_i(\beta)} [y_i - \eta_i(\beta)] = 0$$

need to be solved. If the model is loglinear, then this simplifies to

$$u_j(\beta) = \sum_{i=1}^{n} x_{ij} \left[y_i - \exp\left(\sum_{h=1}^{k} x_{ih} \beta_h \right) \right] = 0.$$

A generally applicable method for obtaining estimates numerically is provided by the *Fisher scoring algorithm* (*see* **Optimization and Nonlinear Equations**). In the present case, this is seen to be an *iteratively reweighted least squares procedure*, where, in each step of the iterative algorithm, a weighted **least squares** problem is to be solved. As a particular consequence to this fact, methods developed for diagnosing linear regression models can be modified for generalized linear models. To define *leverage* and *influence* one only needs to refer to respective quantities calculated from the last iteration step (*see* **Diagnostics**). Formulas needed to do so are lengthy to write down, but most of the widely used software packages provide, at least as an option, the figures. For more on diagnostics for generalized linear models

see [15]. Software packages found useful for fitting Poisson regression models include GLIM [20] and **S-PLUS** [10].

Not much is known about existence and uniqueness of maximum likelihood estimators in the general case. For loglinear models, however, if all observed sufficient statistics involved are larger than 0, then maximum likelihood estimates for the means, i.e. $\hat{\mu}_i = \eta_i(\hat{\beta})$, do exist and are unique, which is also true for $\hat{\beta}$, if the design matrix is of full rank. For a more detailed discussion, see [1] and the references therein.

A statistic capable of measuring the amount of support given by the data to a particular value of the parameter compared to its maximum likelihood estimate is the *deviance*, defined as minus two times the logarithm of the *normed likelihood*:

$$\begin{aligned} D_y(\beta) &= -2\log\left(\frac{L_y(\beta)}{L_y(\hat{\beta})}\right) \\ &= -2[\ell_y(\beta) - \ell_y(\hat{\beta})] \\ &= -2\sum_{i=1}^{n}\left\{y_i \log\left[\frac{\eta_i(\beta)}{\eta_i(\hat{\beta})}\right] \right. \\ &\qquad \left. - [\eta_i(\beta) - \eta_i(\hat{\beta})]\right\}. \end{aligned}$$

The deviance cannot be negative. It provides a measure of distance between the model described by β and the model characterized by the most likely parameter $\hat{\beta}$ and can, thus, be used to construct likelihood regions. Assuming β to be the "true" parameter, the deviance has an asymptotic χ^2 distribution with k degrees of freedom, where k is the dimension of the parameter β. This admits an interpretation of likelihood regions as confidence sets.

To obtain a measure of **goodness of fit** similar to the residual sum of squares in normal linear regression, the likelihood for the *maximal model* that perfectly fits the data can be compared to the likelihood of the model under consideration. This statistic is usually written as

$$\mathrm{dev}_y = 2\sum_{i=1}^{n}\left\{y_i \log\left[\frac{y_i}{\eta_i(\hat{\beta})}\right] - [y_i - \eta_i(\hat{\beta})]\right\}, \tag{12}$$

and termed *deviance* as well. Assuming the null model to be correct, the expected value for the latter statistic is approximately equal to the number of residual **degrees of freedom**, i.e. the number of observations minus the number of parameters in the model.

The deviance is a very important tool in searching for a "good", i.e. a parsimonious and well fitting, model, as it can be used to compare nested **hierarchical** models. Suppose we have a model with parameter β and a smaller one with a parameter γ, which can be obtained from β by setting r components to 0. Then, assuming the smaller model to be the correct one, the difference of deviances (12) is asymptotically χ^2 distributed with r degrees of freedom. Note that this is a **likelihood ratio test** for the smaller model with the null hypothesis against the larger model as the alternative.

The deviance is a useful measure of discrepancy, frequently supposed to have an approximate χ^2 distribution. However, this is to be taken with care, as χ^2 is not, in general, guaranteed to be a large sample distribution of (12). The deviance itself can be approximated by

$$X^2 = \sum_{i=1}^{n}\frac{[y_i - \eta_i(\hat{\beta})]^2}{\eta_i(\hat{\beta})}, \tag{13}$$

which is known as the *Pearson goodness-of-fit statistic* (*see* **Chi-square Tests**).

Another way of performing significance tests of hypotheses about single parameters is by applying a *Wald test* (*see* **Likelihood**). This uses the approximate normality of the maximum likelihood estimates and computes, as the test statistic, the ratio of the estimate of the parameter of interest and its asymptotic standard error. The formula is complex, but, again, many statistical packages provide the figures for the Wald test, sometimes under the heading *t-test*, as well as observed significance values. For more detailed accounts on likelihood inference for a generalized linear model with some emphasis on the Poisson regression model see [1, 19, 34] and [35] and the references therein.

An obvious way of defining **residual** quantities is to use square roots of contributions to the sums in (12) or (13), and attach the appropriate signs. Denoting raw residuals by $r_i = y_i - \eta_i(\hat{\beta})$, we have

$$\begin{aligned} r_i^{\mathrm{D}} = \mathrm{sgn}(r_i)(-2\{y_i \log\left[\frac{y_i}{\eta_i(\hat{\beta})}\right] \\ - [y_i - \eta_i(\hat{\beta})]\})^{1/2} \end{aligned} \tag{14}$$

for the *deviance residuals* and

$$r_i^{\mathrm{P}} = \frac{y_i - \eta_i(\hat{\beta})}{[\eta_i(\hat{\beta})]^{1/2}} \tag{15}$$

for the *Pearson residuals*. In any case, large residuals indicate large contributions to the respective goodness-of-fit statistics. Both deviance and Pearson residuals can (and should) be standardized. This requires computation of leverages for all observations. See [37] and [15] for more on residuals in generalized linear models.

For the time trend model fitted to the Stress Recall Data one calculates a deviance of 24.57 with 16 degrees of freedom. The deviance is 1.5 times larger than its approximate expected value, indicating a moderate amount of *overdispersion* (see [8, 16], and [17] for more on the phenomenon of overdispersion in Poisson regression models). Compared to a model with only the constant term included, we see a difference of deviances of 26.67. Referring to its approximate **chi-square distribution** (with 1 degree of freedom) clearly confirms the time trend. The regression parameter for the explanatory variable "months before interview" has been estimated as -0.0837, with an asymptotic standard error of 0.017, resulting in a t-value of -4.99, which is definitely large enough to reject the hypothesis of no time trend. The smallest deviance residual is -1.99, the largest 2.04, and there is no obvious pattern suggesting any specific inadequacies in the model.

References

[1] Agresti, A. (1990). *Categorical Data Analysis*. Wiley, New York.

[2] Aitkin, M. & Clayton, D. (1980). The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM, *Applied Statistics* **29**, 156–163.

[3] Aitkin, M., Laird, N. Francis, B. (1983). A reanalysis of the Stanford heart transplant data (with comments and rejoinder), *Journal of the American Statistical Association* **78**, 264–292.

[4] Aitkin, M., Anderson, D., Francis, B. & Hinde. J. (1989). *Statistical Modeling in GLIM*. Clarendon Press, Oxford.

[5] Barndorff-Nielsen, O.E. & Cox, D.R. (1994). *Inference and Asymptotics*. Chapman & Hall, London.

[6] Bishop, Y.M.M., Fienberg, S.E. & Holland, P. (1975). *Discrete Multivariate Analysis*. MIT Press, Cambridge, Mass.

[7] Breslow, N. (1974). Covariance analysis of censored survival data, *Biometrics* **30**, 89–100.

[8] Breslow, N. (1990). Tests of hypotheses in overdispersed Poisson regression and other quasi-likelihood models, *Journal of the American Statistical Association* **85**, 565–571.

[9] Breslow, N.E. & Day, N.E. (1980). *Statistical Methods in Cancer Research*, Vol. 1. *The Analysis of Case-Control Studies*. International Agency for Research on Cancer, Lyon.

[10] Chambers, J.M. & Hastie, T.J. (1992). *Statistical Models in S*. Wadsworth & Brooks, Pacific Gove.

[11] Charnes, A., Frome, E.L. & Yu, P.L. (1976). The equivalence of generalized least squares and maximum likelihood estimates in the exponential family, *Journal of the American Statistical Association* **71**, 169–172.

[12] Clayton, D. & Hills, M. (1993). *Statistical Models in Epidemiology*. Oxford University Press, Oxford.

[13] Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

[14] Cox, D.R. (1975). Partial likelihood, *Biometrika* **62**, 269–276.

[15] Davison, A.C. & Snell, E.J. (1991). Residuals and diagnostics, in *Statistical Theory and Modelling. In Honour of Sir David Cox*, D.V. Hinkley, N. Reid & E.J. Snell, eds. Chapman & Hall, London, pp. 83–106.

[16] Dean, C.B. (1992). Testing for overdispersion in Poisson and binomial regression models, *Journal of the American Statistical Association* **87**, 451–457.

[17] Dean, C. & Lawless, J.F. (1989). Tests for detecting overdispersion in Poisson regression models, *Journal of the American Statistical Association* **84**, 467–472.

[18] Dobson, A. (1990). *An Introduction to Generalized Linear Models*, 2nd Ed. Chapman & Hall, London.

[19] Firth, D. (1991). Generalized linear models, in *Statistical Theory and Modelling. In Honour of Sir David Cox*, D.V. Hinkley, N. Reid & E.J. Snell, eds. Chapman & Hall, London, pp. 55–82.

[20] Francis, B., Green, M. & Payne, C., eds (1993). *The GLIM System. Release 4 Manual*. Clarendon Press, Oxford.

[21] Frome, E.L. (1981). Poisson regression analysis, *American Statistician* **35**, 262–263.

[22] Frome, E.L. (1983). The analysis of rates using Poisson regression models, *Biometrics* **39**, 665–674.

[23] Frome, E.L. & DuFrain, R.J. (1986). Maximum likelihood estimation for cytogenic dose-response curves, *Biometrics* **42**, 73–84.

[24] Frome, E.L. & Morris, M.D. (1989). Evaluating goodness of fit of Poisson regression models in cohort studies, *American Statistician* **43**, 144–147.

[25] Frome, E.L., Kutner, M.H. & Beauchamp, J.J. (1973). Regression analysis of Poisson-distributed data, *Journal of the American Statistical Association* **68**, 935–940.

[26] Greenland, S. & Maldonado, G. (1994). The interpretation of multiplicative-model parameters as standardized parameters, *Statistics in Medicine* **13**, 989–999.

[27] Haberman, S. (1978). *Analysis of Qualitative Data*. Vol. 1. *Introductory Topics*. Academic Press, New York.

[28] Holford, T.R. (1980). The analysis of rates and survivorship using log-linear models, *Biometrics* **36**, 299–306.

[29] Koch, G.G., Atkinson, S.S. & Stokes, M.E. (1986). Poisson regression, in *Encyclopedia of Statistics*, Vol. 7, S. Kotz & N.L. Johnson, eds. Wiley, New York, pp. 32–41.

[30] Ladurner, D. & Seeber, G. (1984). Das papilläre Schilddrüsenkarzinom-Prognose und prognostische Faktoren, *Langenbeck's Archiv für Chirurgie* **363**, 43–55.

[31] Laird, N. & Olivier, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques, *Journal of the American Statistical Association* **76**, 231–240.

[32] Lawless, J.F. (1987). Regression methods for Poisson process data, *Journal of the American Statistical Association* **82**, 808–815.

[33] Lindsey, J.K. (1995). *Modelling Frequency and Count Data*. Clarendon Press, Oxford.

[34] Lindsey, J.K. (1996). *Parametric Statistical Inference*. Clarendon Press, Oxford.

[35] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Ed. Chapman & Hall, London.

[36] Nelder, J.A. & Wedderburn, R.W. (1972). Generalized linear models, *Journal of the Royal Statistical Society, Series A* **135**, 370–384.

[37] Pierce, D.A. & Schafer, D.W. (1986). Residuals in generalized linear models, *Journal of the American Statistical Association* **81**, 977–986.

[38] Seeber, G.U.H. (1989). On the regression analysis of tumour recurrence rates, *Statistics in Medicine* **8**, 1363–1369.

[39] Selmer, R. (1990). A comparison of Poisson regression models fitted to multiway summary tables and Cox's survival model using data from a blood pressure screening in the city of Bergen, Norway, *Statistics in Medicine* **9**, 1157–1165.

(*See also* **Categorical Data Analysis**)

GILG U.H. SEEBER