Appendix B

Generalized Linear Model Theory

We describe the generalized linear model as formulated by Nelder and Wedderburn (1972), and discuss estimation of the parameters and tests of hypotheses.

B.1 The Model

Let $y_1, \ldots, y_n$ denote n independent observations on a response. We treat y_i as a realization of a random variable Y_i. In the general linear model we assume that Y_i has a normal distribution with mean μ_i and variance σ^2

$$Y_i \sim \mathrm{N}(\mu_i, \sigma^2),$$

and we further assume that the expected value μ_i is a linear function of p predictors that take values $\mathbf{x}_i' = (x_{i1}, \ldots, x_{ip})$ for the i-th case, so that

$$\mu_i = \mathbf{x}_i'\boldsymbol{\beta},$$

where $\boldsymbol{\beta}$ is a vector of unknown parameters.

We will generalize this in two steps, dealing with the stochastic and systematic components of the model.

B.1.1 The Exponential Family

We will assume that the observations come from a distribution in the exponential family with probability density function

$$f(y_i) = \exp\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\}. \tag{B.1}$$

Here θ_i and ϕ are parameters and $a_i(\phi)$, $b(\theta_i)$ and $c(y_i, \phi)$ are known functions. In all models considered in these notes the function $a_i(\phi)$ has the form

$$a_i(\phi) = \phi/p_i,$$

where p_i is a known *prior weight*, usually 1.

The parameters θ_i and ϕ are essentially location and scale parameters. It can be shown that if Y_i has a distribution in the exponential family then it has mean and variance

$$\begin{aligned} \mathrm{E}(Y_i) &= \mu_i = b'(\theta_i) && \text{(B.2)} \\ \mathrm{var}(Y_i) &= \sigma_i^2 = b''(\theta_i)a_i(\phi), && \text{(B.3)} \end{aligned}$$

where $b'(\theta_i)$ and $b''(\theta_i)$ are the first and second derivatives of $b(\theta_i)$. When $a_i(\phi) = \phi/p_i$ the variance has the simpler form

$$\mathrm{var}(Y_i) = \sigma_i^2 = \phi b''(\theta_i)/p_i.$$

The exponential family just defined includes as special cases the normal, binomial, Poisson, exponential, gamma and inverse Gaussian distributions.

Example: The normal distribution has density

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2}\frac{(y_i - \mu_i)^2}{\sigma^2}\}.$$

Expanding the square in the exponent we get $(y_i - \mu_i)^2 = y_i^2 + \mu_i^2 - 2y_i\mu_i$, so the coefficient of y_i is μ_i/σ^2. This result identifies θ_i as μ_i and ϕ as σ^2, with $a_i(\phi) = \phi$. Now write

$$f(y_i) = \exp\{\frac{y_i\mu_i - \frac{1}{2}\mu_i^2}{\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\}.$$

This shows that $b(\theta_i) = \frac{1}{2}\theta_i^2$ (recall that $\theta_i = \mu_i$). Let us check the mean and variance:

$$\begin{aligned} E(Y_i) &= b'(\theta_i) = \theta_i = \mu_i, \\ \mathrm{var}(Y_i) &= b''(\theta_i)a_i(\phi) = \sigma^2. \end{aligned}$$

Try to generalize this result to the case where Y_i has a normal distribution with mean μ_i and variance σ^2/n_i for known constants n_i, as would be the case if the Y_i represented sample means.□

Example: In Problem Set 1 you will show that the exponential distribution with density

$$f(y_i) = \lambda_i \exp\{-\lambda_i y_i\}$$

belongs to the exponential family.□

In Sections B.4 and B.5 we verify that the binomial and Poisson distributions also belong to this family.

B.1.2 The Link Function

The second element of the generalization is that instead of modeling the mean, as before, we will introduce a one-to-one continuous differentiable transformation $g(\mu_i)$ and focus on

$$\eta_i = g(\mu_i). \tag{B.4}$$

The function $g(\mu_i)$ will be called the *link* function. Examples of link functions include the identity, log, reciprocal, logit and probit.

We further assume that the transformed mean follows a linear model, so that

$$\eta_i = \mathbf{x}_i'\boldsymbol{\beta}. \tag{B.5}$$

The quantity η_i is called the *linear predictor*. Note that the model for η_i is pleasantly simple. Since the link function is one-to-one we can invert it to obtain

$$\mu_i = g^{-1}(\mathbf{x}_i'\boldsymbol{\beta}).$$

The model for μ_i is usually more complicated than the model for η_i.

Note that we do not transform the response y_i, but rather its expected value μ_i. A model where $\log y_i$ is linear on x_i, for example, is not the same as a generalized linear model where $\log \mu_i$ is linear on x_i.

Example: The standard linear model we have studied so far can be described as a generalized linear model with normal errors and identity link, so that

$$\eta_i = \mu_i.$$

It also happens that μ_i, and therefore η_i, is the same as θ_i, the parameter in the exponential family density.□

When the link function makes the linear predictor η_i the same as the canonical parameter θ_i, we say that we have a *canonical link*. The identity is the canonical link for the normal distribution. In later sections we will see that the logit is the canonical link for the binomial distribution and the log is the canonical link for the Poisson distribution. This leads to some natural pairings:

Error	Link
Normal	Identity
Binomial	Logit
Poisson	Log

However, other combinations are also possible. An advantage of canonical links is that a minimal sufficient statistic for $\boldsymbol{\beta}$ exists, i.e. all the information about $\boldsymbol{\beta}$ is contained in a function of the data of the same dimensionality as $\boldsymbol{\beta}$.

B.2 Maximum Likelihood Estimation

An important practical feature of generalized linear models is that they can all be fit to data using the same algorithm, a form of *iteratively re-weighted least squares*. In this section we describe the algorithm.

Given a trial estimate of the parameters $\hat{\boldsymbol{\beta}}$, we calculate the estimated linear predictor $\hat{\eta}_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}}$ and use that to obtain the fitted values $\hat{\mu_i} = g^{-1}(\hat{\eta}_i)$. Using these quantities, we calculate the working dependent variable

$$z_i = \hat{\eta}_i + (y_i - \hat{\mu_i})\frac{\mathrm{d}\eta_i}{\mathrm{d}\mu_i}, \tag{B.6}$$

where the rightmost term is the derivative of the link function evaluated at the trial estimate.

Next we calculate the iterative weights

$$w_i = p_i/[b''(\theta_i)(\frac{\mathrm{d}\eta_i}{\mathrm{d}\mu_i})^2], \tag{B.7}$$

where $b''(\theta_i)$ is the second derivative of $b(\theta_i)$ evaluated at the trial estimate and we have assumed that $a_i(\phi)$ has the usual form ϕ/p_i. This weight is inversely proportional to the variance of the working dependent variable z_i given the current estimates of the parameters, with proportionality factor ϕ.

Finally, we obtain an improved estimate of $\boldsymbol{\beta}$ regressing the working dependent variable z_i on the predictors $\mathbf{x}_i$ using the weights w_i, i.e. we calculate the weighted least-squares estimate

$$\hat{\beta} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{z}, \tag{B.8}$$

where $\mathbf{X}$ is the model matrix, $\mathbf{W}$ is a diagonal matrix of weights with entries w_i given by (B.7) and $\mathbf{z}$ is a response vector with entries z_i given by (B.6).

The procedure is repeated until successive estimates change by less than a specified small amount. McCullagh and Nelder (1989) prove that this algorithm is equivalent to Fisher scoring and leads to maximum likelihood estimates. These authors consider the case of general $a_i(\phi)$ and include ϕ in their expression for the iterative weight. In other words, they use $w_i^* = \phi w_i$, where w_i is the weight used here. The proportionality factor ϕ cancels out when you calculate the weighted least-squares estimates using (B.8), so the estimator is exactly the same. I prefer to show ϕ explicitly rather than include it in $\mathbf{W}$.

Example: For normal data with identity link $\eta_i = \mu_i$, so the derivative is $d\eta_i/d\mu_i = 1$ and the working dependent variable is y_i itself. Since in addition $b''(\theta_i) = 1$ and $p_i = 1$, the weights are constant and no iteration is required.□

In Sections B.4 and B.5 we derive the working dependent variable and the iterative weights required for binomial data with link logit and for Poisson data with link log. In both cases iteration will usually be necessary.

Starting values may be obtained by applying the link to the data, i.e. we take $\hat{\mu}_i = y_i$ and $\hat{\eta}_i = g(\hat{\mu}_i)$. Sometimes this requires a few adjustments, for example to avoid taking the log of zero, and we will discuss these at the appropriate time.

B.3 Tests of Hypotheses

We consider Wald tests and likelihood ratio tests, introducing the *deviance* statistic.

B.3.1 Wald Tests

The Wald test follows immediately from the fact that the information matrix for generalized linear models is given by

$$\mathbf{I}(\boldsymbol{\beta}) = \mathbf{X}'\mathbf{W}\mathbf{X}/\phi, \tag{B.9}$$

so the large sample distribution of the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ is multivariate normal

$$\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\phi). \tag{B.10}$$

with mean $\boldsymbol{\beta}$ and variance-covariance matrix $(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\phi$.

Tests for subsets of $\boldsymbol{\beta}$ are based on the corresponding marginal normal distributions.

Example: In the case of normal errors with identity link we have $\mathbf{W} = \mathbf{I}$ (where $\mathbf{I}$ denotes the identity matrix), $\phi = \sigma^2$, and the *exact* distribution of $\hat{\boldsymbol{\beta}}$ is multivariate normal with mean $\boldsymbol{\beta}$ and variance-covariance matrix $(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$.

B.3.2 Likelihood Ratio Tests and The Deviance

We will show how the likelihood ratio criterion for comparing any two nested models, say $\omega_1 \subset \omega_2$, can be constructed in terms of a statistic called the *deviance* and an unknown scale parameter ϕ.

Consider first comparing a model of interest ω with a *saturated* model Ω that provides a separate parameter for each observation.

Let $\hat{\mu}_i$ denote the fitted values under ω and let $\hat{\theta}_i$ denote the corresponding estimates of the canonical parameters. Similarly, let $\tilde{\mu}_O = y_i$ and $\tilde{\theta}_i$ denote the corresponding estimates under Ω.

The likelihood ratio criterion to compare these two models in the exponential family has the form

$$-2\log\lambda = 2\sum_{i=1}^{n}\frac{y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)}{a_i(\phi)}.$$

Assume as usual that $a_i(\phi) = \phi/p_i$ for known prior weights p_i. Then we can write the likelihood-ratio criterion as follows:

$$-2\log\lambda = \frac{D(\mathbf{y}, \hat{\boldsymbol{\mu}})}{\phi}. \tag{B.11}$$

The numerator of this expression does not depend on unknown parameters and is called the *deviance*:

$$D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{n} p_i[y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)]. \tag{B.12}$$

The likelihood ratio criterion $-2\log L$ is the deviance divided by the scale parameter ϕ, and is called the *scaled deviance*.

Example: Recall that for the normal distribution we had $\theta_i = \mu_i$, $b(\theta_i) = \frac{1}{2}\theta_i^2$, and $a_i(\phi) = \sigma^2$, so the prior weights are $p_i = 1$. Thus, the deviance is

$$\begin{aligned} D(\mathbf{y}, \hat{\boldsymbol{\mu}}) &= 2\sum\{y_i(y_i - \hat{\mu}_i) - \frac{1}{2}y_i^2 + \frac{1}{2}\hat{\mu}_i{}^2\} \\ &= 2\sum\{\frac{1}{2}y_i^2 - y_i\hat{\mu}_i{}^2 + \frac{1}{2}\hat{\mu}_i{}^2\} \\ &= \sum(y_i - \hat{\mu}_i)^2 \end{aligned}$$

our good old friend, the residual sum of squares.□

Let us now return to the comparison of two nested models ω_1, with p_1 parameters, and ω_2, with p_2 parameters, such that $\omega_1 \in \omega_2$ and $p_2 > p1$.

The log of the ratio of maximized likelihoods under the two models can be written as a difference of deviances, since the maximized log-likelihood under the saturated model cancels out. Thus, we have

$$-2\log\lambda = \frac{D(\omega_1) - D(\omega_2)}{\phi} \tag{B.13}$$

The scale parameter ϕ is either known or estimated using the larger model ω_2.

Large sample theory tells us that the asymptotic distribution of this criterion under the usual regularity conditions is χ^2_ν with $\nu = p_2 - p_1$ degrees of freedom.

Example: In the linear model with normal errors we estimate the unknown scale parameter ϕ using the residual sum of squares of the larger model, so the criterion becomes

$$-2\log\lambda = \frac{\mathrm{RSS}(\omega_1) - \mathrm{RSS}(\omega_2)}{\mathrm{RSS}(\omega_2)/(n - p_2)}.$$

In large samples the approximate distribution of this criterion is χ^2_ν with $\nu = p_2 - p_1$ degrees of freedom. Under normality, however, we have an exact result: dividing the criterion by $p_2 - p_1$ we obtain an F with $p_2 - p_1$ and $n - p_2$ degrees of freedom. Note that as $n \to \infty$ the degrees of freedom in the denominator approach ∞ and the F converges to $(p_2 - p_1)\chi^2$, so the asymptotic and exact criteria become equivalent.□

In Sections B.4 and B.5 we will construct likelihood ratio tests for binomial and Poisson data. In those cases $\phi = 1$ (unless one allows overdispersion and estimates ϕ, but that's another story) and the deviance is the same as the scaled deviance. All our tests will be based on asymptotic χ^2 statistics.

B.4 Binomial Errors and Link Logit

We apply the theory of generalized linear models to the case of binary data, and in particular to logistic regression models.

B.4.1 The Binomial Distribution

First we verify that the binomial distribution $B(n_i, \pi_i)$ belongs to the exponential family of Nelder and Wedderburn (1972). The binomial probability distribution function (p.d.f.) is

$$f_i(y_i) = \left(\begin{array}{c} n_i \\ y_i \end{array} \right) \pi_i^{y_i} (1-\pi_i)^{n_i - y_i}. \tag{B.14}$$

Taking logs we find that

$$\log f_i(y_i) = y_i \log(\pi_i) + (n_i - y_i) \log(1-\pi_i) + \log \left(\begin{array}{c} n_i \\ y_i \end{array} \right).$$

Collecting terms on y_i we can write

$$\log f_i(y_i) = y_i \log(\frac{\pi_i}{1-\pi_i}) + n_i \log(1-\pi_i) + \log \left(\begin{array}{c} n_i \\ y_i \end{array} \right).$$

This expression has the general exponential form

$$\log f_i(y_i) = \frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)$$

with the following equivalences: Looking first at the coefficient of y_i we note that the canonical parameter is the logit of π_i

$$\theta_i = \log(\frac{\pi_i}{1-\pi_i}). \tag{B.15}$$

Solving for π_i we see that

$$\pi_i = \frac{e^{\theta_i}}{1+e^{\theta_i}}, \quad \text{so} \qquad 1-\pi_i = \frac{1}{1+e^{\theta_i}}.$$

If we rewrite the second term in the p.d.f. as a function of θ_i, so $\log(1-\pi_i) = -\log(1+e^{\theta_i})$, we can identify the cumulant function $b(\theta_i)$ as

$$b(\theta_i) = n_i \log(1+e^{\theta_i}).$$

The remaining term in the p.d.f. is a function of y_i but not π_i, leading to

$$c(y_i, \phi) = \log \left(\begin{array}{c} n_i \\ y_i \end{array} \right).$$

Note finally that we may set $a_i(\phi) = \phi$ and $\phi = 1$.

Let us verify the mean and variance. Differentiating $b(\theta_i)$ with respect to θ_i we find that

$$\mu_i = b'(\theta_i) = n_i \frac{e^{\theta_i}}{1 + e^{\theta_i}} = n_i \pi_i,$$

in agreement with what we knew from elementary statistics. Differentiating again using the quotient rule, we find that

$$v_i = a_i(\phi) b''(\theta_i) = n_i \frac{e^{\theta_i}}{(1 + e^{\theta_i})^2} = n_i \pi_i (1 - \pi_i),$$

again in agreement with what we knew before.

In this development I have worked with the binomial count y_i, which takes values $0(1)n_i$. McCullagh and Nelder (1989) work with the proportion $p_i = y_i/n_i$, which takes values $0(1/n_i)1$. This explains the differences between my results and their Table 2.1.

B.4.2 Fisher Scoring in Logistic Regression

Let us now find the working dependent variable and the iterative weight used in the Fisher scoring algorithm for estimating the parameters in logistic regression, where we model

$$\eta_i = \text{logit}(\pi_i). \tag{B.16}$$

It will be convenient to write the link function in terms of the mean μ_i, as:

$$\eta_i = \log(\frac{\pi_i}{1 - \pi_i}) = \log(\frac{\mu_i}{n_i - \mu_i}),$$

which can also be written as $\eta_i = \log(\mu_i) - \log(n_i - \mu_i)$.

Differentiating with respect to μ_i we find that

$$\frac{d\eta_i}{d\mu_i} = \frac{1}{\mu_i} + \frac{1}{n_i - \mu_i} = \frac{n_i}{\mu_i(n_i - \mu_i)} = \frac{1}{n_i \pi_i (1 - \pi_i)}.$$

The working dependent variable, which in general is

$$z_i = \eta_i + (y_i - \mu_i) \frac{d\eta_i}{d\mu_i},$$

turns out to be

$$z_i = \eta_i + \frac{y_i - n_i\pi_i}{n_i\pi_i(1-\pi_i)}. \tag{B.17}$$

The iterative weight turns out to be

$$\begin{aligned} w_i &= 1/\left[b''(\theta_i)(\frac{d\eta_i}{d\mu_i})^2\right], \\ &= \frac{1}{n_i\pi_i(1-\pi_i)}[n_i\pi_i(1-\pi_i)]^2, \end{aligned}$$

and simplifies to

$$w_i = n_i\pi_i(1-\pi_i). \tag{B.18}$$

Note that the weight is inversely proportional to the variance of the working dependent variable. The results here agree exactly with the results in Chapter 4 of McCullagh and Nelder (1989).

Exercise: Obtain analogous results for Probit analysis, where one models

$$\eta_i = \Phi^{-1}(\mu_i),$$

where $\Phi()$ is the standard normal cdf. *Hint:* To calculate the derivative of the link function find $d\mu_i/d\eta_i$ and take reciprocals.□

B.4.3 The Binomial Deviance

Finally, let us figure out the binomial deviance. Let $\hat{\mu_i}$ denote the m.l.e. of μ_i under the model of interest, and let $\tilde{\mu_i} = y_i$ denote the m.l.e. under the saturated model. From first principles,

$$\begin{aligned} D &= 2\sum[y_i\log(\frac{y_i}{n_i}) + (n_i - y_i)\log(\frac{n_i - y_i}{n_i}) \\ &\quad -y_i\log(\frac{\hat{\mu_i}}{n_i}) - (n_i - y_i)\log(\frac{n_i - \hat{\mu_i}}{n_i})]. \end{aligned}$$

Note that all terms involving $\log(n_i)$ cancel out. Collecting terms on y_i and on $n_i - y_i$ we find that

$$D = 2\sum[y_i\log(\frac{y_i}{\hat{\mu_i}}) + (n_i - y_i)\log(\frac{n_i - y_i}{n_i - \mu_i})]. \tag{B.19}$$

Alternatively, you may obtain this result from the general form of the deviance given in Section B.3.

Note that the binomial deviance has the form

$$D = 2\sum o_i \log(\frac{o_i}{e_i}),$$

where o_i denotes observed, e_i denotes expected (under the model of interest) and the sum is over both "successes" and "failures" for each i (i.e. we have a contribution from y_i and one from $n_i - y_i$).

For grouped data the deviance has an asymptotic chi-squared distribution as $n_i \to \infty$ for all i, and can be used as a goodness of fit test.

More generally, the difference in deviances between nested models (i.e. the log of the likelihood ratio test criterion) has an asymptotic chi-squared distribution as the number of groups $k \to \infty$ or the size of each group $n_i \to \infty$, provided the number of parameters stays fixed.

As a general rule of thumb due to Cochrane (1950), the asymptotic chi-squared distribution provides a reasonable approximation when all *expected* frequencies (both $\hat{\mu}_i$ and $n_i - \hat{\mu}_i$) under the *larger* model exceed one, and at least 80% exceed five.

B.5 Poisson Errors and Link Log

Let us now apply the general theory to the Poisson case, with emphasis on the log link function.

B.5.1 The Poisson Distribution

A Poisson random variable has probability distribution function

$$f_i(y_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} \tag{B.20}$$

for $y_i = 0, 1, 2, \ldots$. The moments are

$$E(Y_i) = \text{var}(Y_i) = \mu_i.$$

Let us verify that this distribution belongs to the exponential family as defined by Nelder and Wedderburn (1972). Taking logs we find

$$\log f_i(y_i) = y_i \log(\mu_i) - \mu_i - \log(y_i!).$$

Looking at the coefficient of y_i we see immediately that the canonical parameter is

$$\theta_i = \log(\mu_i), \tag{B.21}$$

and therefore that the canonical link is the log. Solving for μ_i we obtain the inverse link

$$\mu_i = e^{\theta_i},$$

and we see that we can write the second term in the p.d.f. as

$$b(\theta_i) = e^{\theta_i}.$$

The last remaining term is a function of y_i only, so we identify

$$c(y_i, \phi) = -\log(y_i!).$$

Finally, note that we can take $a_i(\phi) = \phi$ and $\phi = 1$, just as we did in the binomial case.

Let us verify the mean and variance. Differentiating the cumulant function $b(\theta_i)$ we have

$$\mu_i = b'(\theta_i) = e^{\theta_i} = \mu_i,$$

and differentiating again we have

$$v_i = a_i(\phi)b''(\theta_i) = e^{\theta_i} = \mu_i.$$

Note that the mean equals the variance.

B.5.2 Fisher Scoring in Log-linear Models

We now consider the Fisher scoring algorithm for Poisson regression models with canonical link, where we model

$$\eta_i = \log(\mu_i). \tag{B.22}$$

The derivative of the link is easily seen to be

$$\frac{d\eta_i}{d\mu_i} = \frac{1}{\mu_i}.$$

Thus, the working dependent variable has the form

$$z_i = \eta_i + \frac{y_i - \mu_i}{\mu_i}. \tag{B.23}$$

The iterative weight is

$$\begin{aligned} w_i &= 1/\left[b''(\theta_i)(\frac{d\eta_i}{d\mu_i})^2\right] \\ &= 1/\left[\mu_i \frac{1}{\mu_i^2}\right], \end{aligned}$$

and simplifies to

$$w_i = \mu_i. \tag{B.24}$$

Note again that the weight is inversely proportional to the variance of the working dependent variable.

B.5.3 The Poisson Deviance

Let $\hat{\mu_i}$ denote the m.l.e. of μ_i under the model of interest and let $\tilde{\mu_i} = y_i$ denote the m.l.e. under the saturated model. From first principles, the deviance is

$$\begin{aligned} D &= 2\sum[y_i \log(y_i) - y_i - \log(y_i!) \\ &\quad -y_i \log(\hat{\mu_i}) + \hat{\mu_i} + \log(y_i!)]. \end{aligned}$$

Note that the terms on $y_i!$ cancel out. Collecting terms on y_i we have

$$D = 2\sum[y_i \log(\frac{y_i}{\hat{\mu_i}}) - (y_i - \hat{\mu_i})]. \tag{B.25}$$

The similarity of the Poisson and Binomial deviances should not go unnoticed. Note that the first term in the Poisson deviance has the form

$$D = 2\sum o_i \log(\frac{o_i}{e_i}),$$

which is identical to the Binomial deviance. The second term is usually zero. To see this point, note that for a canonical link the score is

$$\frac{\partial \log L}{\partial \boldsymbol{\beta}} = \mathbf{X}'(\mathbf{y} - \boldsymbol{\mu}),$$

and setting this to zero leads to the estimating equations

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'\hat{\boldsymbol{\mu}}.$$

In words, maximum likelihood estimation for Poisson log-linear models—and more generally for any generalized linear model with canonical link—requires equating certain functions of the m.l.e.'s (namely $\mathbf{X}'\hat{\boldsymbol{\mu}}$) to the same functions of the data (namely $\mathbf{X}'\mathbf{y}$). If the model has a constant, one column of $\mathbf{X}$ will consist of ones and therefore one of the estimating equations will be

$$\sum y_i = \sum \hat{\mu_i} \qquad \text{or} \qquad \sum(y_i - \hat{\mu_i}) = 0,$$

so the last term in the Poisson deviance is zero. This result is the basis of an alternative algorithm for computing the m.l.e.'s known as "iterative proportional fitting", see Bishop *et al.* (1975) for a description.

The Poisson deviance has an asymptotic chi-squared distribution as $n \to \infty$ with the number of parameters p remaining fixed, and can be used as a goodness of fit test. Differences between Poisson deviances for nested models (i.e. the log of the likelihood ratio test criterion) have asymptotic chi-squared distributions under the usual regularity conditions.