Generalized Linear Model under the Extended Negative Multinomial Model and Cancer Incidence

S. Lahiri & Sunil K. Dhar *

Department of Mathematical Sciences, CAMS
New Jersey Institute of Technology, Newark, NJ-07111, USA

E-mail: dhar@njit.edu

SUMMARY

The generalized linear model for a multi-way contingency table for several independent populations that follow the extended negative multinomial distributions is introduced. This model represents an extension of negative multinomial log-linear model. The parameters of the new model are estimated by the quasi-likelihood method and the corresponding score function

*Corresponding author
which gives a close from estimate of the regression parameters. The goodness-of-fit test for the model is also discussed. An application of the log-linear model under the generalized inverse sampling scheme representing cancer incidence data is given as an example of this model.

**Keywords:** generalized inverse sampling, extended negative multinomial distribution, quasi-likelihood, asymptotic distribution.

1 INTRODUCTION

A generalized linear model (GLIM) for functions of the extended negative multinomial frequency counts for $s$ independent sub populations is defined along the lines of Grizzle, Starmer and Koch (1969), Bonett (1985a) and Bonett(1985b). The extended negative multinomial model (ENMn) is defined in Dhar (1995) and the extended negative multinomial log-linear model is defined in Lahiri and Dhar (2008). Let $(f_{-_r1}, \cdots, f_{-_2i}, f_{-_1i}, f_{1i}, f_{2i}, \cdots, f_{ni})'$ be $s$ sets of count observations from $s$ subpopulations labeled $i = 1, 2, \cdots, s$, and let $j = -r, \cdots, -1, 1, \cdots, n$ be the set of response categories as summarized in Table 1 below. Here, $'$ denotes transpose of a matrix. To define the model the following notations are needed. Denote the vector $f_{-_i} = (f_{-_ri}, \cdots, f_{-_2i}, f_{-_1i})'$, $f_i = (f_{1i}, f_{2i}, \cdots, f_{ni})'$ and their corresponding expected valued vectors $\mu_{-_i} = (\mu_{-_ri}, \cdots, \mu_{-_1i})'$, $\mu_i = (\mu_{1i}, \cdots, \mu_{ni})'$, respectively. The vector $f^{(i)} = (f^{(i)}_{-_i}, f^{(i)}_i)'$ is assumed to follow an extended negative multinomial distribution with parameters, $k_i = \sum_{j=1}^r f_{-_ji}$ and $\mu^{(i)} = (\mu_{-_i}, \mu_i)' = k_i(\sum_{j=1}^r p_{-_ji})^{-1} p_i$, where $p_i = (p_{-_ri}, \cdots, p_{-_1i}, p_{1i}, \cdots, p_{ni})'$ is from Lahiri and Dhar (2008, Section 2). Let $f$ be the augmented vector de-
The generalized linear model for count data is defined by

$$y^* = X\beta + \delta$$

(1.1)

where $$y^* = (y_1^*, y_2^*, \ldots, y_s^*)$$ and $$y_i^* = \langle g(f_{-ri}), \ldots, g(f_{-1i}), g(f_{1i}), g(f_{2i}), \ldots, g(f_{ni}) \rangle$$, $$i = 1, 2, \ldots, s$$, and $$g$$ is function from $$\mathbb{R}$$ to $$\mathbb{R}$$, which has an inverse. Let $$X$$ be a known full rank $$q$$ and of dimensions $$(r + n)s \times q$$ ($$q \leq [r + n]s$$) design matrix with intercept, main effects, and interaction effects, $$\beta$$ is a $$q \times 1$$ vector of unknown non-random parameters, and $$\delta$$ is a $$(r + n)s \times 1$$ unobservable error vector such that $$y'_0 = \langle g(k_1), g(k_2), \ldots, g(k_s) \rangle^T$$, $$y'_0 - A\beta = 0$$ and $$y^* - X\beta$$ is asymptotically zero, as $$k \to \infty$$. Here, $$k = \sum_{i=1}^s k_i$$, the fractions of the total sample $$k$$, namely, $$\frac{k_i}{k} = w_i$$, $$f_0 = (k_1, k_2, \ldots, k_s)$$, and $$A = (a_1, a_2, \ldots, a_s)^T$$, $$s \times q$$, are known constants, with linearly independent constraints $$y'_0 - A\beta = 0$$. The data consists of observing the vector $$f$$ along with design matrix $$X$$ as shown in the $$r \times c$$ contingency Table 1. This model is known to be the generalized linear model under the extended inverse sampling scheme or the extended negative multinomial distribution, with $$g$$ in (1.1) as the link function. This model is specifically known to be linear or the log linear model according to $$g$$ being linear or the log function.

A closed-form estimator of the model parameters, estimate of the covariance matrix, and the general Wald test are derived under the assumption of extended negative multinomial sampling. The commonly used GLIM, does not accommodate contingency tables of inverse sampling schemes as has also been observed by Bonett (1985a).
Observations from a subpopulation are sampled following generalized inverse sampling and are described below for the sake of completeness. The resulting observed frequency counts can be summarized in the following $s \times (r + n)$ contingency table.

1.1 Definition of ENMn

Consider a sequence of independent trials, where one of the events $A_i$ occurs with probability $p_i$, $i \in \{-r, \cdots, -1, 1, \cdots, n\}$, $\sum_{i=-r, i\neq 0}^n p_i = 1$. Suppose that $A_{-r}, A_{-(r-1)}, \cdots, A_{-1}$ are the rare events. Let $f_i$ represent the frequency with which $A_i$, $i \in \{-r, \cdots, -1\}$ occurs. Here, $f_i$’s are counted until we get a total of $k$ (predetermined value) observations of at least one of the $A_i$’s , $i \in \{-r, \cdots, -1\}$. Then the distribution of $f = (f_{-r}, \cdots, f_{-1}, f_1, \cdots, f_n)'$ is said to follow an extended negative multinomial distribution with parameters $k$ and $p = (p_{-r}, \cdots, p_{-1}, p_1, \cdots, p_n)'$ with the joint probability density function given as

$$\frac{(\sum_{i=1}^n f_i + k - 1)!}{\prod_{i=1}^n f_i! (k - 1)!} (\sum_{i=1}^r p_{-i})^k \prod_{i=1}^n p_i^{f_i} \prod_{i=1}^r (p_i^*)^{f_{-i}} \prod_{i=1}^n f_i!$$

$f_{-i}$ and $f_i \geq 0$, $i \in \{-r, \cdots, -1, 1, \cdots, n\}$, where $p_i^* = \frac{p_{-i}}{\sum_{i=1}^r p_{-i}}$, $i = 1, \cdots, r$, $\sum_{i=-r, i\neq 0}^n p_i = 1$ and $k = \sum_{i=1}^r f_{-i}$. An observation that fall in cell $(j,i)$ has the probability $p_{ji}$.
Table 1: Frequency Distribution

<table>
<thead>
<tr>
<th>Categories of responses</th>
<th>Population</th>
<th>$-r$</th>
<th>$-2$</th>
<th>$-1$</th>
<th>$1$</th>
<th>$2$</th>
<th>...</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1$</td>
<td>$f_{-r1}$</td>
<td>$f_{-21}$</td>
<td>$f_{-11}$</td>
<td>$f_{11}$</td>
<td>$f_{21}$</td>
<td>...</td>
<td>$f_{n1}$</td>
<td></td>
</tr>
<tr>
<td>$2$</td>
<td>$f_{-r2}$</td>
<td>$f_{-22}$</td>
<td>$f_{-12}$</td>
<td>$f_{12}$</td>
<td>$f_{22}$</td>
<td>...</td>
<td>$f_{n2}$</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$s$</td>
<td>$f_{-rs}$</td>
<td>$f_{-2s}$</td>
<td>$f_{-1s}$</td>
<td>$f_{1s}$</td>
<td>$f_{2s}$</td>
<td>...</td>
<td>$f_{ns}$</td>
<td></td>
</tr>
</tbody>
</table>

In order to define the covariance matrix of $f$, the following notations are introduced. Let $p_{ji}^* = \frac{p_{ji}}{\sum_{j=1}^{r} p_{ji}}$ and $D_{\mu_i}$ be a diagonal matrix with the elements of the vector $\mu_i$, $n \times 1$, along the main diagonal. Then the covariance matrix of $f$ is given by the $(r+n)s \times (r+n)s$ block diagonal matrix $\Sigma_f$ of rank $(r+n)s$ with

$$
\Sigma_f^{(i)}(k_i) = \begin{pmatrix}
\Sigma_1^{(i)}(k_i) & 0 \\
0' & \Sigma_2^{(i)}(k_i)
\end{pmatrix}, \quad i = 1, 2, \ldots, s
$$

(1.2)

as the blocks, where $0$ is $r \times n$ zero matrix and where

$$
\Sigma_1^{(i)}(k_i) = \begin{pmatrix}
k_i p_{11}^*(1 - p_{11}^*) & -k_i p_{11}^* p_{21}^* & \cdots & -k_i p_{11}^* p_{ri}^* \\
-k_i p_{11}^* p_{21}^* & k_i p_{21}^*(1 - p_{21}^*) & \cdots & -k_i p_{21}^* p_{ri}^* \\
\vdots & \vdots & \ddots & \vdots \\
-k_i p_{11}^* p_{ri}^* & -k_i p_{21}^* p_{ri}^* & \cdots & k_i p_{ri}^*(1 - p_{ri}^*)
\end{pmatrix}
$$

(1.3)
and \( \Sigma_2^{(i)}(k_i) = \frac{(\mu_i \mu'_i)}{k_i} + D_{\mu_i}. \) \( (1.4) \)

## 2 Estimation

The quasi-likelihood estimator of \( \beta \) (Myers et al. 2002, Section 5.4) with constraints has been given in this section. An efficient estimator of \( \beta \), under the constraint \( A\beta = y_0 \), minimizes

\[
X^2 = (y^* - X\beta)'\hat{\Omega}(y^* - X\beta) - \lambda'(A\beta - y_0), \tag{2.5}
\]

where \( \hat{\Omega}^{-1} = [\partial y^*/\partial f] \hat{\Sigma}_f [\partial y^*/\partial f]' \), \( \hat{\Sigma}_f \) is the estimate of the block diagonal matrix \( \Sigma_f \) with blocks given by the matrix in (1.2), with \( k_i = 1 \), \( i = 1, 2, ..., s \) times \( w_i \), where \( w_i = \frac{k_i}{k} \) known, \( \lambda \) is an \( s \times 1 \) vector of Lagrange multipliers, \( y_0 = \langle g(k_1), g(k_2), \cdots, g(k_s) \rangle' \). The maximum likelihood estimator of the covariance matrix, \( \hat{\Sigma}_f \), is obtained by replacing the elements of \( p \) and \( \mu \) (which is a function of \( p \)), by their corresponding sample proportions based on \( f \) (Dhar 1995). Please see the expression for the true parameter \( \Omega^{-1} \) in equation (3.7) in the following section. Minimizing (2.5) with respect to \( \beta \) and using multivariate calculus, the quasi-likelihood estimator of \( \beta \) with constraints is computed and has the following closed form:

\[
\hat{\beta}^* = \hat{\beta} - (X'\hat{\Omega}X)^{-1}A'[A(X'\hat{\Omega}X)^{-1}A']^{-1}(A\hat{\beta} - y_0), \tag{2.6}
\]

where \( \hat{\beta} = (X'\hat{\Omega}X)^{-1}X'\hat{\Omega}y^* \), which is shown more generally by Ferguson (1958). The estimated covariance matrix of \( \hat{\beta}^* \) is

\[
\hat{\Sigma}_{\hat{\beta}^*} = (X'\hat{\Omega}X)^{-1} - (X'\hat{\Omega}X)^{-1}A'[A(X'\hat{\Omega}X)^{-1}A']^{-1}A(X'\hat{\Omega}X)^{-1}.
\]

The cell frequencies that follow the extended negative multinomial constraints are predicted by \( \hat{y}^* = X\hat{\beta}^* \).
3 Hypothesis Testing

In this section, the test of the general linear hypothesis \( H_0 : H\beta = h \) versus its negation is derived, where \( H \) is a \((n + r) \times q\) known matrix of rank \((n + r)\) and \( h \) is a known \((n + r) \times 1\) vector. The Wald’s statistics

\[
W = (H\hat{\beta} - h)'(H(X'\hat{\Omega}X)^{-1})H' - 1 (H\hat{\beta} - h)/k,
\]

where \( k = \sum_{i=1}^{s} k_i \) is used to test these hypotheses. The asymptotic distribution of \( W \) is now derived.

**Theorem 1:** Under the true model as described by (1.1), the assumptions \( g \) is a differentiable function from \( \mathbb{R} \) to \( \mathbb{R} \), and \( \sup_k E\|y^* - X\beta\|^{1+\eta} < \infty \) for some \( \eta > 0 \), where \( \|\cdot\| \) represents the Euclidean norm, then \( W \xrightarrow{d} \chi^2_{(n+r)} \) as \( k \to \infty \).

**Proof:** Define,

\[
B^{(m)}_{-ji} = \begin{cases} 
1, & \text{if } m^{th} \text{ unit drawn at random from } i^{th} \text{ population belongs} \\
& \text{to } -j^{th} \text{ category,} \\
0, & \text{otherwise,}
\end{cases}
\]

for \( m = 1, \cdots, k_i; \ j = 1, \cdots, r; \ i = 1, \cdots, s, \)

and

\[
D^{(m)}_{ji} = \begin{cases} 
1, & \text{if } m^{th} \text{ unit drawn at random from } i^{th} \text{ population belongs} \\
& \text{to } j^{th} \text{ category,} \\
0, & \text{otherwise,}
\end{cases}
\]

for \( m = 1, \cdots, k_i; \ j = 1, \cdots, n; \ i = 1, \cdots, s. \)
Then

\[ f^{(i)} = (f_{-i}, f_i)' \]

\[ = (f_{-1}, f_{-2}, \ldots, f_{-r_i}, f_{-r_i}, f_{-2}, \ldots, f_{-1}, f_{1}, f_{2}, \ldots, f_{m})' \]

\[ = \left( \sum_{m=1}^{k_i} B_{-i1}^{(m)}, \sum_{m=1}^{k_i} B_{-i2}^{(m)}, \ldots, \sum_{m=1}^{k_i} B_{-ri}^{(m)}, \sum_{m=1}^{k_i} D_{1i}^{(m)}, \ldots, \sum_{m=1}^{k_i} D_{ni}^{(m)} \right)' \]

\[ = \sum_{m=1}^{k_i} (B_{-i1}^{(m)}, B_{-i2}^{(m)}, \ldots, B_{-ri}^{(m)}, D_{1i}^{(m)}, \ldots, D_{ni}^{(m)})' \]

\[ = \sum_{m=1}^{k_i} (C_i^{(m)})', \]

\( i = 1, \ldots, s. \)

Therefore, \( C_i^{(m)}, j = -r, \ldots, -1, 1, \ldots, n, \) follows an extended negative multinomial distribution with parameters \((1, \mu_i^{(i)}/k_i),\) where \( \mu^{(i)}/k_i = (\sum_{j=1}^{r} p_{-ji})^{-1} p_i. \) Now, \( C_i^{(m)} \) are iid random vectors with mean \( \mu_i^{(i)}/k_i \) and covariance matrix \( \Sigma_i^{(i)}(1) \), for \( m = 1, \ldots, k_i. \) The central limit theorem (CLT) yields, \( \sqrt{k_i} \left( \frac{1}{k_i} \sum_{m=1}^{k_i} (C_i^{(m)}) - \mu_i^{(i)}/k_i \right) \xrightarrow{d} N(0, \Sigma_i^{(i)}(1)) \) as \( k_i \to \infty, \ i = 1, 2, \ldots, s. \) This implies that \( \frac{1}{\sqrt{k_i}} (f^{(i)} - \mu^{(i)}) \xrightarrow{d} N(0, \Sigma_i^{(i)}(1)) \), as \( k_i \to \infty, \) hence \( \frac{1}{\sqrt{k}} (f - \mu) \xrightarrow{d} N(0, \Sigma^*) \), as \( k \to \infty, \) where \( \frac{k}{k_i} = w_i \) is a known constant and \( \Sigma^*_i \) is the block diagonal matrix with blocks \( w_i \Sigma_i^{(i)}(1), \ i = 1, 2, \ldots, s, \) along the diagonal of rank \( s(n + r). \) Therefore,

\[ \frac{1}{\sqrt{k}} (y^* - \mu^*) \xrightarrow{d} N(0, D\mu(\mu^*) \Sigma^*_i D\mu(\mu^*)) \]  

(3.7)

as \( k \to \infty, \) where \( \mu^* \) is equal to \( g(\mu), \) i.e., \( g \) applied to each component of \( \mu \) and \( D\mu(\mu^*) \) is the differential of \( \mu^* \) with respect to \( \mu \) (Theorem A, Serfling, 2002, p. 122). Here, \( \Omega^{-1} = D\mu(\mu^*) \Sigma^*_i D\mu(\mu^*) \). Now consider \( \hat{\beta} = (X'\hat{\Omega}X)^{-1}X'\hat{\Omega}Y, \) as described in Section 2, which converges as follows:

\[ \frac{1}{\sqrt{k}} (\hat{\beta} - \beta) \xrightarrow{d} N(0, (X'\Omega X)^{-1}) \]  

as \( k \to \infty \) (Theorem A, Serfling, 2002, p.122).
Similarly, as $k \to \infty$, \( \frac{1}{\sqrt{k}}(H\widehat{\beta} - H\beta) \xrightarrow{d} N(0, H(X'\Omega X)^{-1}H') \), where $H$ is a $(n + r) \times q$ known matrix with rank $(n + r)$. Therefore, from the fact that $x'H(X'\Omega X)^{-1}H'x$ is a continuous function of $x$ gives

\[
W = \left(\frac{H\widehat{\beta} - h}{\sqrt{k}}\right)'(H(X'\Omega X)^{-1})^{-1}\left(\frac{H\widehat{\beta} - h}{\sqrt{k}}\right) \xrightarrow{d} \chi^2_{(n+r)} \text{ (Rao, 1973, p. 188).}
\]

Hence the proof.

To evaluate the goodness-of-fit of this model with linearly independent constraints, one can consider the statistics $\left(\mathbf{y}^* - X\widehat{\mathbf{\beta}}^*\right)'\widehat{\Omega}\left(\mathbf{y}^* - X\widehat{\mathbf{\beta}}^*\right)/k$. Similar to Bonnett (1989 and 1985a) and Haber et al. (1986) one can show that this statistic converges in distribution to the chi-square distribution with $s(n + r) - q$ degrees of freedom. This statement holds true under conditions of the aforementioned Theorem 1 and is similar to its proof as well as Bonnett (1985a, Section 4, page 128) or Bonnett (1989), which uses Haber et al. (1986).

### 4 An Example

This section illustrates an example of an extended negative multinomial model to a given data set from $s = 3$ subpopulations. The application involves modeling of the incidence of several related diseases in different cities. There is no maximum likelihood estimate for the shape parameter in general. An estimate based on quantiles of Pearson’s chi-squared statistics is applied to model the cancer incidence for three cities in Ohio. This approach is similar to those proposed by Williams (1982), Breslow (1984), and more recently, in Waller and Zelterman (1997) for estimating over-dispersion parameters.
4.1 Model of Cancer Incidence for Three Cities

Suppose that the following table gives the cancer incidence for the three largest cities according to the site of primary cancer during a particular year. It is assumed that the overall disease incidence may be higher (lower) in one location than the other, but this increase (decrease) is not disease-specific — that is, the relative frequencies of disease do not change across cities. Waller and Zelterman (1997) used the log-linear model with the negative multinomial distribution to fit the cancer deaths, during 1989, in the three largest Ohio cities. The structure of the data used in this example is similar to the one described in Waller and Zelterman (1997), but hypothetical numbers have been used to illustrate the proposed model. The sites of primary tumors are as follows: 1 = eye, 2 = oral cavity, 3 = gallbladder, 4 = lung, 5 = breast, 6 = genitals, 7 = urinary organs, 8 = leukemia and 9 = lymphatic tissues.

Table 2: Cancer Deaths in the Three Different Cities

<table>
<thead>
<tr>
<th>City</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>City 1</td>
<td>35</td>
<td>41</td>
<td>31</td>
<td>440</td>
<td>488</td>
<td>159</td>
<td>523</td>
<td>169</td>
<td>268</td>
</tr>
<tr>
<td>City 2</td>
<td>40</td>
<td>28</td>
<td>27</td>
<td>270</td>
<td>337</td>
<td>133</td>
<td>378</td>
<td>107</td>
<td>160</td>
</tr>
<tr>
<td>City 3</td>
<td>31</td>
<td>25</td>
<td>28</td>
<td>190</td>
<td>212</td>
<td>91</td>
<td>254</td>
<td>77</td>
<td>137</td>
</tr>
</tbody>
</table>

Our objective is to fit the data with an appropriate model. Poisson models are appropriate when the sample mean and the sample variance are equal. Multinomial models can be used when the cell counts are negatively correlated and the negative multinomial models are used when the cell counts are
positively correlated. It can be observed from the above data that some of the frequency counts are positively correlated, while others are negatively correlated. In this situation, the extended negative multinomial model is expected to be the most appropriate one to fit the data due to its covariance structure (equations 1.2 to 1.4).

Consider the log-linear model of means with no interaction between city and disease type as

$$\ln \mu_{js} = \mu + \alpha_s^{city} + \beta_j^{disease},$$

where $\alpha_s^{city}, s = 1, 2, 3,$ is the effect due to city and $\beta_j^{disease}, j = 1, 2, \cdots, 9,$ is the effect due to disease. Note that each of these effects add up to zero. No interaction between city and disease type implies that an increase (decrease) in incidence of one disease is accompanied by a similar increase (decrease) in incidence of the same disease across all the other cities. It has been assumed that the state has a large amount of manufacturing and industry. So, if there is an environmental cause for high incidence of one type of cancer, this may translate into high incidence of another type of cancer. To describe the design matrix $X$ for the above model let $\mathbf{0}$ be the $8 \times 1$ zero column vector, $\mathbf{1}$ represent the $8 \times 1$ vector with all one’s, and $\mathbf{I}$ represent the $8 \times 8$ identity
matrix. Then, $X$ is given by

$$
\begin{pmatrix}
1 & 1 & 0 & I \\
1 & 1 & 0 & 0' \\
1 & 0 & 1 & I \\
1 & 0 & 1 & 0' \\
1 & 0 & 0 & I \\
1 & 0 & 0 & 0' \\
\end{pmatrix}.
$$

Let $f_{js}$ denote the number of cancer deaths of site $s$ in city $j$ and use the extended negative multinomial distribution, since many of the counts have both positive and negative covariances. The test statistic $\chi^2_s(k_s)$ will have the following form:

$$
\chi^2_s(k_s) = \sum_j \frac{(f_{js} - \hat{\mu}_{js})^2}{\hat{\mu}_{js}}, \quad (4.8)
$$

where $k_s$ is the shape parameter of the extended negative multinomial distribution. The estimated means $\hat{\mu}_{js}$ are the same as described in Section 2. The above test statistic not only measures fit, but also suggests a method of estimating $k_s$. In each cities case, solve for $k_s$, such that the median of the limiting chi-squared distribution in (4.8) with 8 d.f. is 7.34. This method is similar to the one proposed by Waller and Zelterman, which is used to find values of shape parameter. Using these values of $k_s$, the model parameters are computed by maximizing the log-likelihood with penalized constraints using Newton-Raphson algorithm iteratively and by R program. By improving the estimate of $k_s$, the model will be improved.
Acknowledgments

Professor Hira Lal Koul is my "statistics guru." I wish to thank him for all the knowledge he has given me.

Bibliography


