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

Sunil Kumar Dhar

Center for Applied Mathematics and Statistics, Department of Mathematical Sciences, New Jersey Institute of Technology

Soumi Lahiri

Center for Applied Mathematics and Statistics, Department of Mathematical Sciences, New Jersey Institute of Technology

Abstract

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, which gives a close form 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 to demonstrate the effectiveness of the model.

Key Words : *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 (ENMn) frequency counts for s independent sub populations is defined along the lines of Grizzle et al. (1969), Bonett (1985a)

Received : December 2013

and Bonett(1985b). The ENMn model is defined in Dhar (1995) and the extended negative multinomial log-linear model is defined in Lahiri and Dhar (2008). Let $(f_{-(r-1)i}, \cdots, f_{-1i},\ f_{1i}, f_{2i}, \cdots, f_{ni})'$, be s sets of count observations from s subpopulations labeled $i = 1, 2, \cdots, s$, and let $j = -r, \cdots, -1, 1, \cdots, n$ represent the set of response categories with counts f_{ji} corresponding to events $A_j, j \in \{-r, \cdots, -1, 1, \cdots, n\}$, where the counts corresponding to $A_j, j \in \{-r, \cdots, -1\}$ add up to k_i known from data. This data for the model is summarized in Table 1 below. Here, $'$ denotes transpose of a matrix. To define the model the following notations are needed. Denote the vector $\mathbf{f}_{-i} = (f_{-(r-1)i}, \cdots, f_{-2i}, f_{-1i})'$, $\mathbf{f}_i = (f_{1i}, f_{2i}, \cdots, f_{ni})'$ and their corresponding expected valued vectors $\boldsymbol{\mu}_{-i} = (\mu_{-(r-1)i}, \cdots, \mu_{-1i})'$, $\boldsymbol{\mu}_i = (\mu_{1i}, \cdots, \mu_{ni})'$, respectively. Let $\mathbf{f}^{(i)} = (\mathbf{f}_{-i},\ \mathbf{f}_i)'$ and $\boldsymbol{\mu}^{(i)} = (\boldsymbol{\mu}_{-i}, \boldsymbol{\mu}_i)'$. The vector $(f_{-ri}, \cdots, f_{-2i}, f_{-1i}, f_{1i}, f_{2i}, \cdots, f_{ni})'$ is assumed to follow an extended negative multinomial distribution with parameters, $k_i = \sum_{j=1}^{r} f_{-ji}$ and $E\{(f_{-ri}, \cdots, f_{-2i}, f_{-1i}, f_{1i}, f_{2i}, \cdots, f_{ni})'\} = k_i(\sum_{j=1}^{r} p_{-ji})^{-1}\mathbf{p}_i$, where $\mathbf{p}_i = (p_{-ri}, \cdots,\ p_{-1i}, p_{1i}, \cdots, p_{ni})'$ is from Lahiri and Dhar (2008, Section 2). Let $\mathbf{f}$ be the augmented vector defined by $\mathbf{f} = (\mathbf{f}^{(1)}, \mathbf{f}^{(2)}, \cdots, \mathbf{f}^{(s)})'$ with expected value $\boldsymbol{\mu} = (\boldsymbol{\mu}^{(\mathbf{1})}, \boldsymbol{\mu}^{(\mathbf{2})}, \cdots, \boldsymbol{\mu}^{(\boldsymbol{s})})'$. The generalized linear model for count data is defined by

$$E(\mathbf{y}^*) = \mathbf{X}\boldsymbol{\beta}, \tag{1}$$

where $\mathbf{y}^{*'} = (\mathbf{y}_1^{*'},\ \mathbf{y}_2^{*'},\ \cdots,\ \mathbf{y}_s^{*'})$ and $\mathbf{y}_i^{*'} = \langle g(f_{-(r-1)i}), \cdots, g(f_{-2i}), g(f_{-1i}), g(f_{1i}), g(f_{2i}), \cdots, g(f_{ni})\rangle$, $i = 1, 2, \cdots, s$ and g is function from $\Re$ to $\Re$, which has an inverse. Let $\mathbf{X}$ be a known full rank q and of dimensions $(r+n-1)s \times q$ $(q \leq [r+n-1]s)$ design matrix with intercept, main effects, and interaction effects, $\boldsymbol{\beta}$ is a $q \times 1$ vector of unknown non-random parameters. The data consists of observing the vector $\mathbf{f}$ along with design matrix $\mathbf{X}$ as shown in the $r \times c$ contingency Table 1, where r is the number of rows and c is the number of columns. This model is known to be the generalized linear model under the extended inverse sampling scheme or the ENMn 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 ENMn 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, also known as the ENMn model, and are described below for the sake of completeness. The resulting distinct observed frequency counts can then be summarized as a $s \times (r + n - 1)$ contingency table.

Table 1: Cross Classification Table of ENMn's Frequencies

	Categories of responses							
Population	$\vert -(r-1)$	$\dots$	-2	-1	1	2	$\dots$	n
1	$\vert f_{-(r-1)1}$	$\dots$	f_{-21}	f_{-11}	f_{11}	f_{21}	$\dots$	f_{n1}
2	$\vert f_{-(r-1)2}$	$\dots$	f_{-22}	f_{-12}	f_{12}	f_{22}	$\dots$	f_{n2}
$\vdots$	$\vdots$	$\dots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$
s	$\vert f_{-(r-1)s}$	$\dots$	f_{-2s}	f_{-1s}	f_{1s}	f_{2s}	$\dots$	f_{ns}

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 events of interest. Let f_i represent the frequency with which A_i, $i \in \{-r, \cdots, -1, 1, \cdots, n\}$, occurs. Here, f_i's represent count of the event A_i, $i \in \{-r, \cdots, -1, 1, \cdots, n\}$, respectively, until we get $k > 0$ (predetermined value) total count of either of the A_i's , $i \in \{-r, \cdots, -1\}$. Then the distribution of $(f_{-r}, \cdots, f_{-1}, f_1, \cdots, f_n)'$ is said to follow an ENMn distribution with parameters k and $\mathbf{p} = (p_{-r}, \cdots, p_{-1}, p_1, \cdots, p_n)'$ with the joint probability density function given as

$$\frac{(\sum_{i=1}^{n} f_i + k - 1)!}{\prod_{i=1}^{n} f_i!\,(k-1)!} \left(\sum_{i=1}^{r} p_{-i}\right)^k p_1^{f_1} \cdots p_n^{f_n} \frac{k!}{\prod_{i=1}^{r} f_{-i}!} (p_1^*)^{f_{-1}} \cdots (p_r^*)^{f_{-r}},$$

f_{-i} and $f_i \geq 0$, $i \in \{-r, \cdots, -1, 1, \cdots, n\}$, where $p_i^* = \dfrac{p_{-i}}{\sum_{i=1}^{r} p_{-i}}$, $i = 1, ...r$,

$\sum_{i=-r, i\neq 0}^{n} p_i = 1$ and $k = \sum_{i=1}^{r} f_{-i}$.

The count data taken from s populations can now be described as non negative integers that fall in each of the cells (j,i) with probability $p_{ji}, i \in \{-(r-1), \cdots, -1, 1, \cdots, n\}$ and $j \in \{, 1, \cdots, s\}$, with $k_i, i \in \{1, \cdots, s\}$ known, summarized in Table 1.

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

$$\boldsymbol{\Sigma}_f^{(i)}(k_i) \;=\; \begin{pmatrix} \boldsymbol{\Sigma}_1^{(i)}(k_i) & \mathbf{0} \\ \mathbf{0}' & \boldsymbol{\Sigma}_2^{(i)}(k_i) \end{pmatrix}, i = 1, 2, \cdots, s, \tag{2}$$

as the blocks, where $\mathbf{0}$ is $(r-1) \times n$ zero matrix, $\boldsymbol{\Sigma}_1^{(i)}(k_i) =$

$$\begin{pmatrix} k_i p^*_{1i}(1-p^*_{1i}) & -k_i p^*_{1i} p^*_{2i} & \cdots & -k_i p^*_{1i} p^*_{(r-1)i} \\ -k_i p^*_{1i} p^*_{2i} & k_i p^*_{2i}(1-p^*_{2i}) & \cdots & -k_i p^*_{2i} p^*_{(r-1)i} \\ \vdots & \vdots & \ddots & \vdots \\ -k_i p^*_{1i} p^*_{(r-1)i} & -k_i p^*_{2i} p^*_{(r-1)i} & \cdots & k_i p^*_{(r-1)i}(1-p^*_{(r-1)i}) \end{pmatrix}, \tag{3}$$

$$\text{and} \quad \boldsymbol{\Sigma}_2^{(i)}(k_i) \;=\; \frac{(\boldsymbol{\mu}_i \boldsymbol{\mu}_i')}{k_i} + \mathbf{D}_{\mu_i}. \tag{4}$$

2 Estimation

The quasi-likelihood estimator of $\boldsymbol{\beta}$ (Myers et al. 2002, Section 5.4) has been given in this section. An efficient estimator of $\boldsymbol{\beta}$ minimizes

$$X^2 \;=\; (\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta})' \widehat{\boldsymbol{\Omega}} (\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta}), \tag{5}$$

where $\widehat{\boldsymbol{\Omega}}^{-1} = [\partial \mathbf{y}^* / \partial \mathbf{f}]\, \widehat{\Sigma_f^*}\, [\partial \mathbf{y}^* / \partial \mathbf{f}]'$, $\widehat{\Sigma}_f^*$ is the estimate of the block diagonal matrix $\boldsymbol{\Sigma}_f^*$ 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. The definition of y^* below (1.1) reveals that $\partial \mathbf{y}^* / \partial \mathbf{f}$ is the diagonal matrix with diagonal elements $g'(f_{ji})$, $j = -(r-1), \cdots, -1, 1, \cdots, n$, $i = 1, \cdots, s$. The maximum likelihood estimator of the covariance matrix, $\widehat{\Sigma_f^*}$, is obtained by replacing the elements of $\mathbf{p}$ and $\boldsymbol{\mu}$ (which is a function of $\mathbf{p}$), by their corresponding sample proportions based on $\mathbf{f}$ (Dhar 1995). Please see the expression for the true parameter $\boldsymbol{\Omega}^{-1}$ in the asymptotic normal r.v.'s covariance matrix in equation (3.7) in the following section. Minimizing (2.5) with respect to $\boldsymbol{\beta}$ and using vector calculus, the quasi-likelihood estimator of $\boldsymbol{\beta}$ is computed and has the following closed form:

$$\widehat{\boldsymbol{\beta}} \;=\; (\mathbf{X}' \widehat{\boldsymbol{\Omega}} \mathbf{X})^{-1} \mathbf{X}' \widehat{\boldsymbol{\Omega}} \mathbf{y}^*, \tag{6}$$

which is shown more generally by Ferguson (1958). The estimated covariance matrix of $\widehat{\boldsymbol{\beta}}$ is $\widehat{\boldsymbol{\Sigma}}_{\widehat{\beta}} = (\mathbf{X}'\widehat{\boldsymbol{\Omega}}\mathbf{X})^{-1}$. The cell frequencies that follow the extended negative multinomial constraints are predicted by $\widehat{\boldsymbol{y}^*} = \mathbf{X}\widehat{\boldsymbol{\beta}}$.

3 Hypothesis Testing

In this section, the test of the general linear hypothesis $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{h}$ versus its negation is derived, where $\mathbf{H}$ is a $(n + r - 1) \times q$ known matrix of rank $(n + r - 1)$ and $\mathbf{h}$ is a known $(n + r - 1) \times 1$ vector. Here, $\mathbf{H}$ can represent the design matrix corresponding to a single population ($s = 1$). The Wald's statistics is used to test this null hypothesis and is given by $\mathbf{W} = (\mathbf{H}\widehat{\boldsymbol{\beta}} - \mathbf{h})'(\mathbf{H}(\mathbf{X}'\widehat{\boldsymbol{\Omega}}\mathbf{X})^{-1})\mathbf{H}')^{-1}(\mathbf{H}\widehat{\boldsymbol{\beta}} - \mathbf{h})/k$, where $k = \sum_{i=1}^{s} k_i$ and $\frac{k_i}{k} = w_i$ are known constants. The asymptotic distribution of $\mathbf{W}$ is derived next.

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

Proof: Define,

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

for $m = 1, \cdots, k_i; \quad j = 1, \cdots, (r-1); \quad i = 1, \cdots, s,$
and

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

for $m = 1, \cdots, k_i; \quad j = 1, \cdots, n; \quad i = 1, \cdots, s.$
Then

$$\begin{aligned} \mathbf{f}^{(i)} &= (\mathbf{f}_{-i},\ \mathbf{f}_i)' \\ &= (f_{-1i}, f_{-2i}, \cdots, f_{-(r-1)i}, f_{1i}, f_{2i}, \cdots, f_{ni})' \\ &= (\sum_{m=1}^{k_i} B_{-1i}^{(m)}, \sum_{m=1}^{k_i} B_{-2i}^{(m)}, \cdots, \sum_{m=1}^{k_i} B_{-(r-1)i}^{(m)}, \sum_{m=1}^{k_i} D_{1i}^{(m)}, \cdots, \sum_{m=1}^{k_i} D_{ni}^{(m)})' \\ &= \sum_{m=1}^{k_i} (B_{-1i}^{(m)}, B_{-2i}^{(m)}, \cdots, B_{-(r-1)i}^{(m)}, D_{1i}^{(m)}, \cdots, D_{ni}^{(m)})' \end{aligned}$$

$$= \sum_{m=1}^{k_i} (\mathbf{C}_i^{(m)})',$$

$i = 1, \cdots, s$.
Therefore, $\mathbf{C}_i^{(m)}$, $j = -(r-1), \cdots, -1, 1, \cdots, n$, follows an extended negative multinomial distribution with parameters $(1, \boldsymbol{\mu^{(i)}}/k_i)$, where $\boldsymbol{\mu^{(i)}}/k_i$ is as described in the introduction section. Now, $\mathbf{C}_i^{(m)}$ are iid random vectors with mean $\boldsymbol{\mu^{(i)}}/k_i$ and covariance matrix $\boldsymbol{\Sigma}_f^{(i)}(1)$, for m = $1, \cdots, k_i$. The Central Limit Theorem (CLT) yields $\sqrt{k_i}\ (\frac{1}{k_i}\sum_{m=1}^{k_i}(\mathbf{C}_i^{(m)})' - \boldsymbol{\mu}^{(i)}/k_i) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_f^{(i)}(1))$ as $k_i \to \infty$, $i = 1, 2, \cdots, s$. This implies that $\frac{1}{\sqrt{k_i}}(\mathbf{f}^{(i)} - \boldsymbol{\mu}^{(i)}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_f^{(i)}(1))$, as $k_i \to \infty$. Hence, $\frac{1}{\sqrt{k}}(\mathbf{f} - \boldsymbol{\mu}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_f^*)$, as $k \to \infty$, where $\frac{k_i}{k} = w_i$ are known constants and $\boldsymbol{\Sigma}_f^*$ is the block diagonal matrix with blocks $w_i\boldsymbol{\Sigma}_f^{(i)}(1)$, $i = 1, 2, \cdots, s$, along the diagonal of rank $s(n+r-1)$. Therefore,

$$\frac{1}{\sqrt{k}}(\mathbf{y}^* - \boldsymbol{\mu}^*) \xrightarrow{d} N(\mathbf{0}, \mathbf{D}_{\boldsymbol{\mu}}(\boldsymbol{\mu}^*)\boldsymbol{\Sigma}_f^*\mathbf{D}_{\boldsymbol{\mu}}\ (\boldsymbol{\mu}^*)') \tag{7}$$

as $k \to \infty$, where $\boldsymbol{\mu}^*$ is equal to $g(\boldsymbol{\mu})$, i.e., g applied to each component of $\boldsymbol{\mu}$ and $\mathbf{D}_{\boldsymbol{\mu}}(\boldsymbol{\mu}^*)$ is the differential of $\boldsymbol{\mu}^*$ with respect to $\boldsymbol{\mu}$ (Theorem A, Serfling, 2002, p. 122). Here, $\boldsymbol{\Omega}^{-1} = \mathbf{D}_{\boldsymbol{\mu}}(\boldsymbol{\mu}^*)\boldsymbol{\Sigma}_f^*\mathbf{D}_{\boldsymbol{\mu}}\ (\boldsymbol{\mu}^*)'$. Now consider $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\widehat{\boldsymbol{\Omega}}\mathbf{X})^{-1}\mathbf{X}'\widehat{\boldsymbol{\Omega}}\mathbf{y}$, as described in Section 2, which converges as follows: $\frac{1}{\sqrt{k}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, (\mathbf{X}'\boldsymbol{\Omega}\mathbf{X})^{-1})$ as $k \to \infty$ (Theorem A, Serfling, 2002, p.122). Similarly, as $k \to \infty$, $\frac{1}{\sqrt{k}}(\mathbf{H}\widehat{\boldsymbol{\beta}} - \mathbf{H}\boldsymbol{\beta}) \xrightarrow{d} N(0, \mathbf{H}(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X})^{-1}\mathbf{H}')$, where $\mathbf{H}$ is a $(n+r-1) \times q$ known matrix with rank $(n+r-1)$. Therefore, from the fact that $\mathbf{x}'\mathbf{H}(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X})^{-1}\mathbf{H}')\mathbf{x}$ is a continuous function of $\mathbf{x}$ gives
$\mathbf{W} = \frac{(\mathbf{H}\widehat{\boldsymbol{\beta}} - \mathbf{h})'}{\sqrt{k}}(\mathbf{H}(\mathbf{X}'\boldsymbol{\Omega}\mathbf{X})^{-1})\mathbf{H}')^{-1}\frac{(\mathbf{H}\widehat{\boldsymbol{\beta}} - \mathbf{h})}{\sqrt{k}} \xrightarrow{d} \chi^2_{(n+r-1)}$ (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 $(\mathbf{y}^* - \mathbf{X}\widehat{\boldsymbol{\beta}})'\widehat{\boldsymbol{\Omega}}(\mathbf{y}^* - \mathbf{X}\widehat{\boldsymbol{\beta}})/k$. 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 $s(n+r-1)-q$ 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 by an example the usefulness of GLIM under an ENMn model, which is a given data set from $s = 3$ subpopulations. The application involves modeling of the incidence of several related diseases (d) in different cities. Here, k_i are taken to be the sum of the counts for which the correlations between cities are negative to reflect the multinomial part of the data. More general models could be developed treating k_i as the shape parameters and then estimating them. There are no maximum likelihood estimates for the shape parameters in general. An estimates based on quantiles of Pearson's chi-squared statistics can be applied to estimate the shape parameters based on 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 numbers have been changed to reflect both positive and negative correlation among diseases over cities as is the case with the ENMn 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

	Disease Label Number								
City	1	2	3	4	5	6	7	8	9
City 1	35	41	31	440	488	159	523	169	268
City 2	40	28	27	270	337	133	378	107	160
City 3	31	25	28	190	212	91	254	77	137

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 ENMn model is expected to be the most appropriate one to fit the data due to its covariance structure (equations 1.2 to 1.4). As an example, treating disease homogeneously over cities one can see that the correlation between the count of eye tumor is negatively correlated with that of the gallbladder tumor, while the remaining correlations are positive. Thus, we can take k_1, k_2, and k_3 to be 66, 67, and 59, respectively. Consider the log-linear model of means with no interaction between city and disease type as

$$\ln \mu_{ji} \quad = \quad \mu + {\alpha_i}^{city} + {\beta_j}^{disease},$$

where ${\alpha_i}^{city}$, $i = 1, 2, 3$, 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 see effectiveness of the GLIM under the ENMn distribution, we consider the above model with identity link, i.e., the ln is replaced by the identity function and compared with the count estimates obtained by running a 3×9 Completely Randomized Design (CRD) using *proc glm* with 'disease' and 'city' considered as categorical variables. Interestingly, the linear model with the normality assumption produced negative count estimates for disease d1, d2, and d3 corresponding to City 3 as -21.852, -25.852, and -28.519, respectively. In contrast, the GLIM in this paper gave the following count estimates using the estimation method of Section 2, showing its superiority over common practices of model fitting. However, the bigger count numbers

Table 3: Cancer Deaths in the Three Different Cities Estimate Using GLIM under ENMn Model and Linear Link

	Disease Label Number								
City	1	2	3	4	5	6	7	8	9
City 1	35.93	28.03	30.07	237.07	273.43	108.18	313.11	95.25	154.81
City 2	40.44	24.52	26.56	233.55	269.92	104.67	309.60	91.74	151.30
City 3	29.76	29.24	27.20	236.24	272.61	107.36	312.28	94.42	153.98

have a much larger residual in magnitude than the smaller count numbers. This suggests a log link function for a more reasonable model. The design matrix used to achieve the estimated counts in Table 3 is given by $\mathbf{X}$. Let $\mathbf{0}$ be the 7×1 zero column vector, $\mathbf{1}$ represent the 7×1 vector with all one's, and $\mathbf{I}$ represent the 7×7 identity matrix. Then, $\mathbf{X}$ is given by

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

Note that the regression coefficient vector $\boldsymbol{\beta}$ is given by $(\mu, \alpha_1, \alpha_2, \beta_1, \cdots, \beta_7)'$ because the alpha and beta effects add up to zero. Also note that since k_1, k_2 and k_3 are known, the data corresponding to the disease eye is dropped and hence, the design matrix has dimension 24×10. Using Section 2, (2.6) and the covariance matrix the GLIM log-link model can be fitted and using the diagnostic method based on Pearson's residuals, GLIM demonstrates its superiority over Least Squares Method with log transformation, where the latter is implemented through SAS. The regression parameter estimates along with the Wald tests are given in the following table.

Since the first column representing cancer with respect to the organ eye was dropped due to treating $k_i's$ to be known, the labeling of the disease effects only are as follows: 1 = gallbladder 2 = oral cavity, 3 = lung, 4 = breast, 5 = genitals, 6 = urinary organs, 7 = leukemia and 8 = lymphatic tissues and eye. Note that from Table 4 only the City 1 effect is significant, which also means that the sum of City 2 and 3 effects combined is significant. However, City 2 and City 3 individually are not significant. Similarly, all the individual disease effects except breast cancer and the combined effect of lymphatic tissue and eye only are significant, which in turn implies that the combined effect of all the diseases except breast cancer, is not significant. Further, since the eye data column was removed because $k_i's$ are treated to be known and the disease effects add up to zero, the last effect labeled 8 gives the estimate of the combined effects lymphatic tissues and eye diseases. To get all the individual disease effects, one could run the GLIM model again, this time removing column 3 from Table 2, keeping $k_i's$ to be the same as before and getting the individual disease effects for 1 = eye, 2 = oral cavity, 3 = lung, 4 = breast, 5 = genitals, 6 = urinary organs, 7 = leukemia and 8 = lymphatic tissues and gall bladder. Now using only the individual eye disease effect or the gallbladder disease effect and subtracting either one of them from the disease effect corresponding to label 8 from either of the GLIM models, respectively, would give all the individual effects. However, for the time being let us shift our attention on the first GLIM model to show its effectiveness when compared to CRD under normal errors.

Table 4: Individual Parameter Significance Using GLIM under ENMn Model and Linear Link

Estimator	Estimate	Variance Estimator	Wald Statistic	p-value
$\widehat{\mu}$	4.862052	0.0044311	5334.916	0
$\widehat{\alpha_1}$	0.234421	0.00567	9.690035	0.001853
$\widehat{\alpha_2}$	-0.03289	0.006023	0.179593	0.671724
$\widehat{\alpha_3}$	-0.20153	0.017374	2.337682	0.126277
$\widehat{\beta_1}$	-1.51183	0.009238	247.4098	9.5×10^{-56}
$\widehat{\beta_2}$	-1.45094	0.0084861	248.0798	6.8×10^{-56}
$\widehat{\beta_3}$	0.8041441	0.0013417	481.96149	8×10^{-107}
$\widehat{\beta_4}$	0.9454499	0.0012315	725.84289	7×10^{-160}
$\widehat{\beta_5}$	-0.04681	0.0024662	0.8884827	0.3458888
$\widehat{\beta_6}$	1.0526489	0.0011565	958.12340	2×10^{-210}
$\widehat{\beta_7}$	-0.13260	0.0026323	6.6794171	0.0097533
$\widehat{\beta_8}$	0.3399331	0.0018357	62.948473	5.4×10^{-08}

Let f_{ji} denote the number of cancer deaths of site j in city i and apply the ENMn distribution, since counts have both positive and negative covariances. The Pearson χ^2 test statistic will have the following form:

$$\chi^2(s(n+r-1)-q) \quad = \quad \sum_{i=1}^{s(n+r-1)} \frac{(y_i^* - \widehat{y}_i^*)^2}{Var(y_i^*)}, \tag{8}$$

where $s(n+r-1)-q$ is its degrees of freedom. The above test statistic can be used to measure the goodness-of-fit of the model and the Pearson residuals are used to do the diagnostics of the model, which itself has the following form:

$$\frac{y_i^* - \widehat{y}_i^*}{SD(y_i^*)}.$$

Here, the standard deviation (SD) of y_i^* can be estimated from the diagonal elements of $\widehat{\mathbf{\Omega}}^{-1}$ by taking their square root.

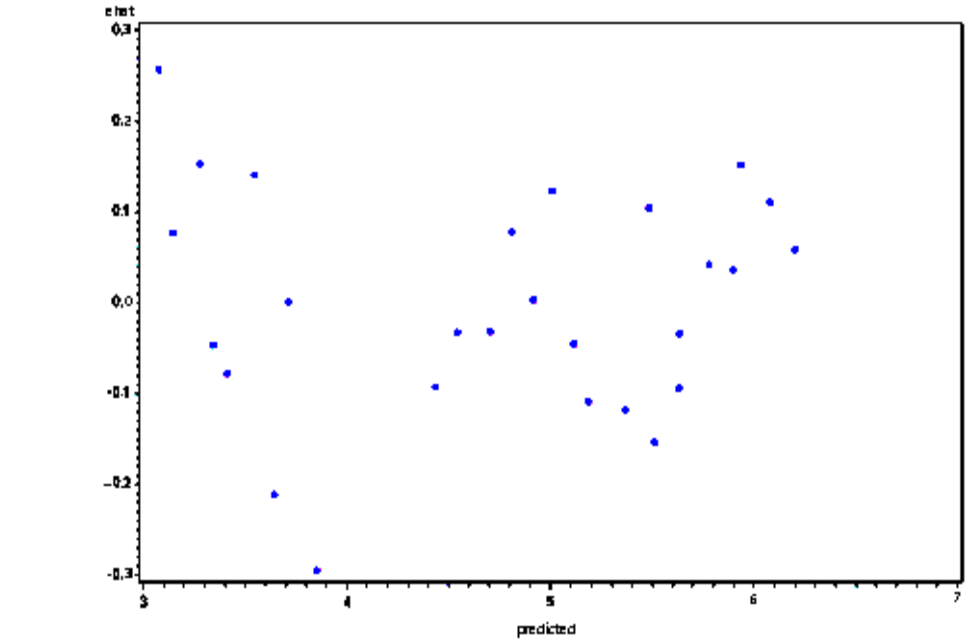
In fact, the Pearson Chi-square value computes to 19.43996578, with p-value = 0.148809569 and the Pearson Chi-square/df = 1.388568984. The Pearson Chi-square test value divided by the degrees of freedom is close to

1 and also the large p-value are indicative that the fit of the GLIM is good, Meyers et al. (2002). Also, the diagnostic plot comparison from Appendix Section 5 shows that there is very little difference between the CRD model with normal errors and log transformation through SAS and GLIM, with minor differences in favor of the latter. The residual plots in Figure 1, Appendix, for CRD shows a slight backward fan shaped pattern indicating that the variances may not be equal, whereas the Pearson's residuals are quite well spread out in a circular cloud pattern. Again, the diagonal plot between ln count predicted versus ln count observed in Figure 2, Appendix, is closer to the $y = x$ line for the GLIM than the diagonal plot for the CRD. This can be seen by comparing the first cluster to the left in the two graphs. The CRD's cluster's linear trend points towards the 60 degree slope line whereas, the first cluster in the second diagonal plot corresponding to the GLIM points towards the 45 degree line. Figure 3, Appendix, shows that the normal probability of the variable ehat, for the CRD, spreads between -0.3 and 0.3. This plot has a pair of consecutive arch-shaped points with one point each in the two extremes. This indicates the presence of some distribution that differs from the normal distribution. However, this plot is roughly indicative of normal because the expected counts for the corresponding cross classification table are all higher than 5. In the case of GLIM, since Person's residual is standardized, it is well spread from -2 to 2 without any distinct pattern in the normal probability plot.

In conclusion, CRD under least squares estimation through SAS, and GLIM, both with log transformation, could be used to obtain final results in data analysis. The advantage of GLIM is that it is based on a cost saving sampling scheme, where the sampling is stopped when a predetermined number of $k_i's$ for each population i has been observed. Moreover under such a sampling design, implementing a GLIM to the count data that represents the ENMn distribution is more appropriate as has been observed from the diagnostics plots.

5 Appendix

Figure 1. Residual Plots Comparison between CRD with Normal Error and GLIM with ENMn Distribution.



Pearson residual

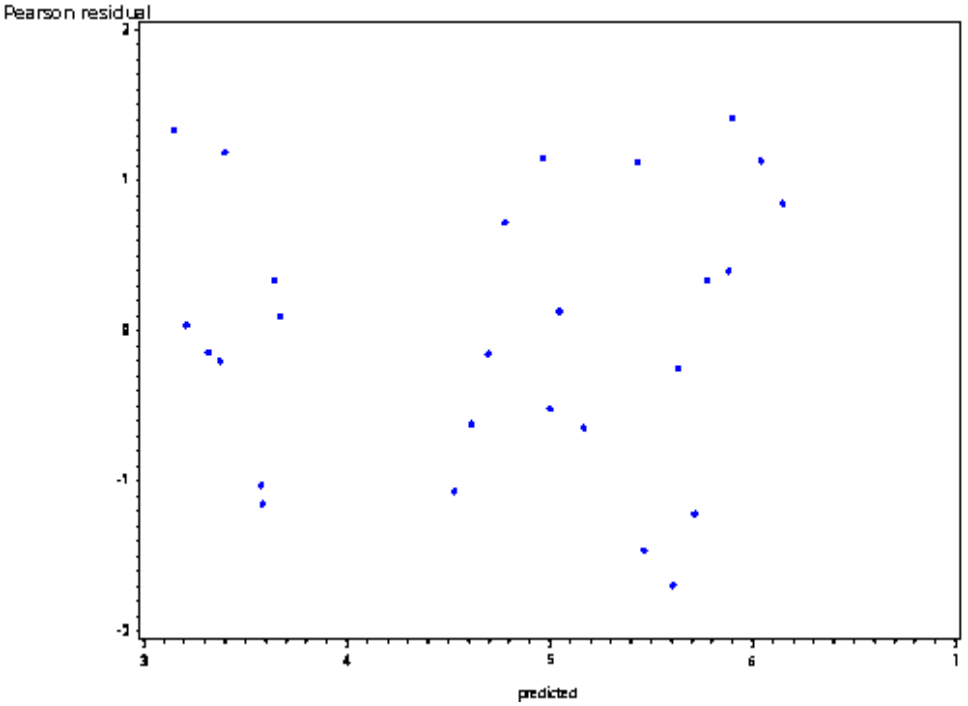
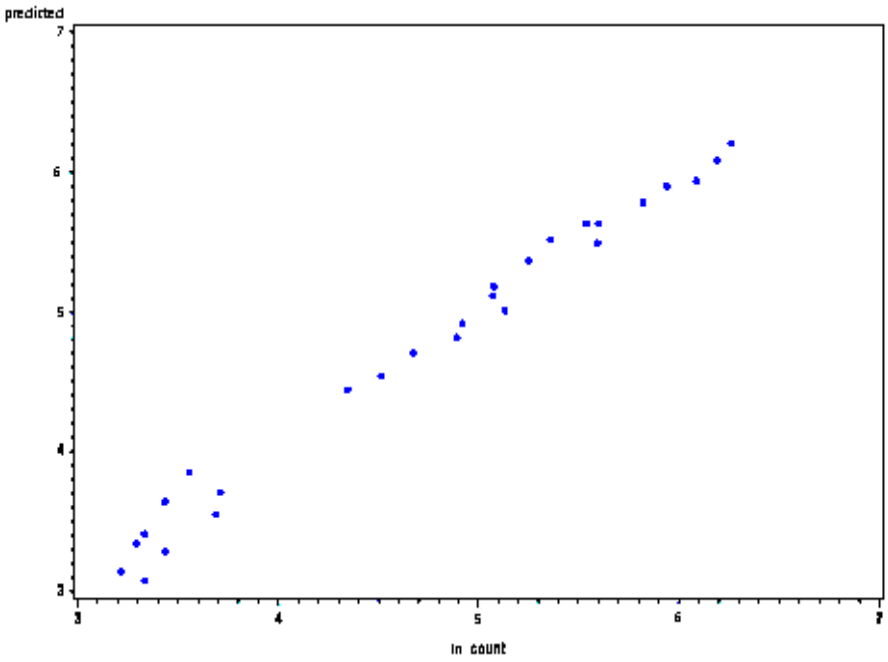


Figure 2. Diagonal Line Plots Log Count Predicted versus Observed Log Count for CRD and GLIM.



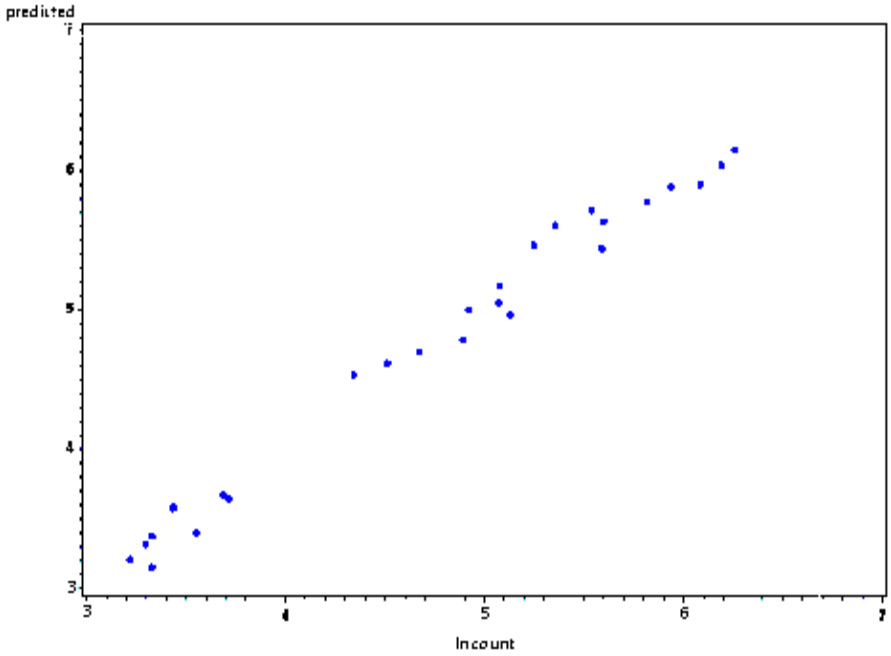
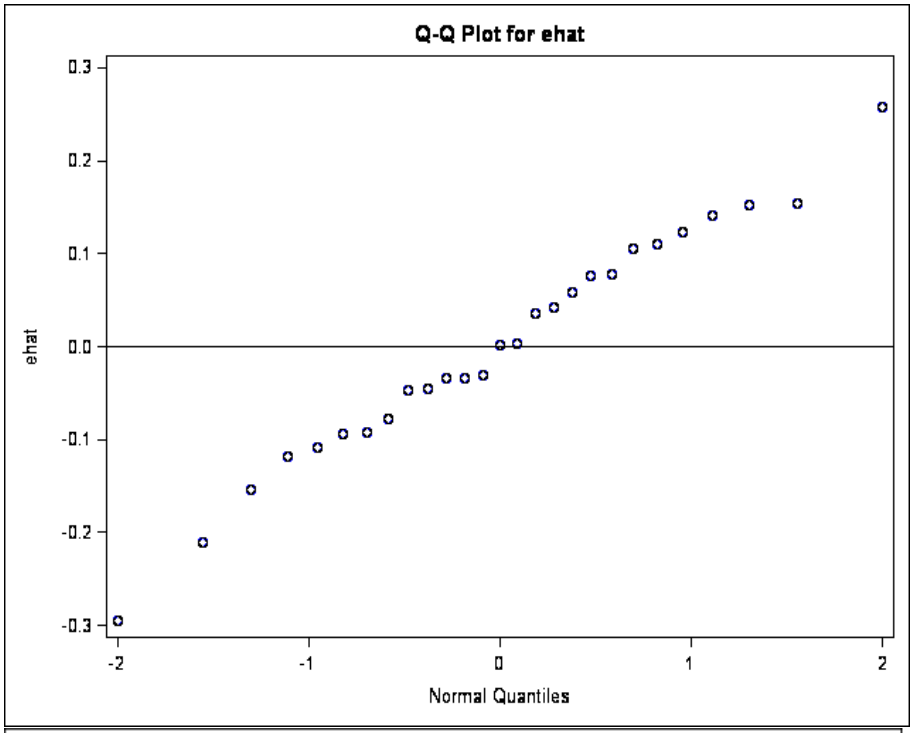
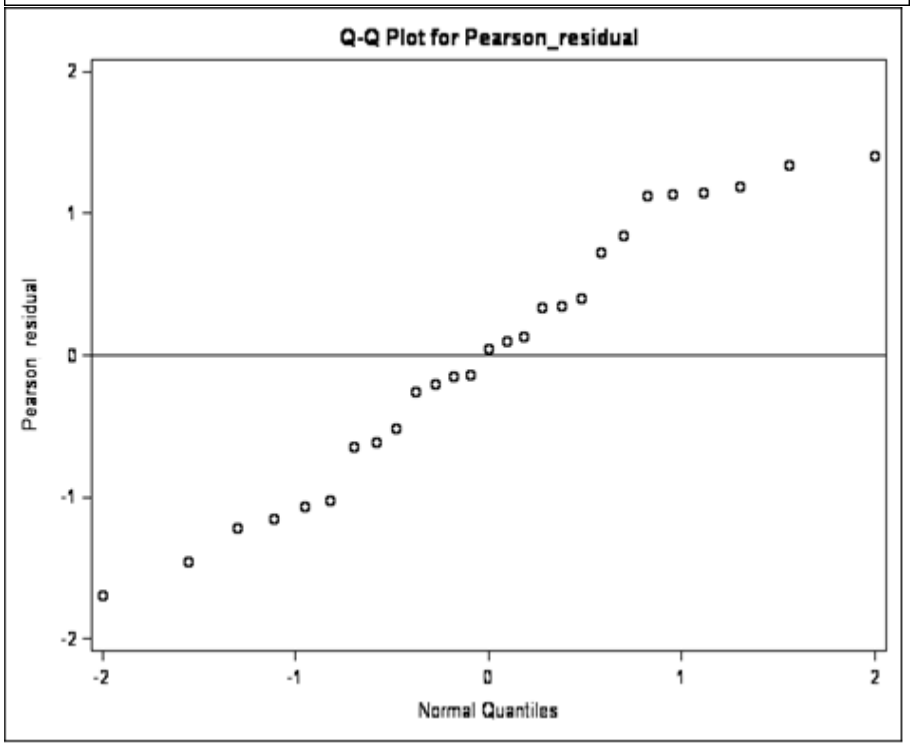


Figure3. Normal Probability Plots for Residuals from CRD and Pearson Residuals from GLIM





Acknowledgments

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

References

[1] Bonett, D. G. (1985a). A linear negative multinomial model. Statist.& Probability Letters, 3, 127-129.

[2] Bonett, D. G. (1985b). The negative multinomial logit model. Commun. Statist.-Theory Methods, 3, 127-129.

[3] Bonett, D. G. (1989). Pearson chi-square estimator and test for log-linear models with expected frequencies subject to linear constraints. Statist.& Probability Letters, 8, 175-177.

[4] Breslow, N. E. (1984). Extra-Poisson variation in log-linear models. Applied Statistics, 33, 38-44.

[5] Dhar, S. K. (1995). Extension of a negative multinomial model. Commun. Statist.-Theory Methods, 24(1), 39-57.

[6] Ferguson, T. S. (1958). A method of generating best asymptotically normal estimates with application to the estimating of bacterial densities. Ann. Math. Stat., 33, 38-44.

[7] Grizzle, J. E., Starmer, C. F. and Koch, G. G. (1969). Analysis of categorical variables by linear models. Biometrics, 25, 489-504.

[8] Haber, M. and Brown, M. B. (1986). Maximum likelihood methods for log-linear models when expected frequencies are subject to linear constraints. J. Amer. Statist. Assoc., 81, 477-482.

[9] Serfling, R. J. (2002). Approximation theorems of mathematical statistics. New York: Wiley.

[10] Lahiri, S. and Dhar, S. K. (2008). Log-linear Modeling under Generalized Inverse Sampling Scheme. Communications in Statistics - Theory and Methods, 37(8), 1237-1244.

[11] Myers, R. H., Montgomery, D. C., Vining, G. G. and Robinson, T. J. (2002). Generalized linear models with application in engineering and the sciences," second edition, New York: Wiley.

[12] Rao, C. R. (1973). Linear statistical inference and its applications. Second edition, New York: Wiley.

[13] Williams, D. A. (1982). Extra-binomial variation in logistic linear models. Applied Statistics, 31, 144-148.

[14] Waller, L. A. and Zelterman, D. (1997). Log-linear modeling with the negative multinomial distribution. Biometrics, 53, 971-982.

Sunil Kumar Dhar
Center for Applied Mathematics and Statistics, Department of Mathematical Sciences
New Jersey Institute of Technology, Newark, NJ-07102
E-mail: dhar@njit.edu

Soumi Lahiri
Center for Applied Mathematics and Statistics, Department of Mathematical Sciences
New Jersey Institute of Technology, Newark, NJ-07102
E-mail: soumi_lahiri@hotmail.com