A Study of Bias for Non-Iterative Estimates of the Linear-by-Linear Association Parameter from the Ordinal Log-Linear Model

Sidra Zafar, Salman A. Cheema, Eric J. Beh, Irene L. Hudson

Abstract Ordinal log-linear models (OLLM) are commonly used to analyse the association in a contingency table with ordered categorical variables. Recently, a non-iterative procedure was introduced to estimate the linear-by-linear association parameter of the OLLM. This paper will consider two of these non-iterative estimates and show that they are unbiased.

1 Introduction

Log-linear models (LLM) are one of the most commonly used techniques to analyse the association between categorical variables in a contingency table. One may refer to, for example, Christensen (1997) and Agresti (2002) for a detailed discussion of many of the issues. Much of this work has focused on the study of variables consisting of nominally structured categories. However, there are many practical situations where variables consist of ordered categories and it is important to reflect the structure of these variables. Hence, LLM’s for these cases are therefore commonly referred to as ordinal log-linear models, or OLLMs. An important aspect of LLMs, and OLLMs, is the estimation of the parameters from these models.

*Sidra Zafar, School of Mathematical and Physical Sciences, University of Newcastle, Callaghan, 2308, NSW, Australia; email: sidra.zafar@uon.edu.au*
*Salman A. Cheema, School of Mathematical and Physical Sciences, University of Newcastle, Callaghan, 2308, NSW, Australia; email: salman.cheema@uon.edu.au*
*Eric J. Beh, School of Mathematical and Physical Sciences, University of Newcastle, Callaghan, 2308, NSW, Australia; email: eric.beh@newcastle.edu.au*
*Irene L. Hudson, School of Mathematical and Physical Sciences, University of Newcastle, Callaghan, 2308, NSW, Australia; email: irenelena.hudson@gmail.com*
Traditionally, the estimation of parameters from LLM’s and OLLM’s have involved the use of iterative procedures. The most commonly used are Newton’s algorithm and iterative proportional fitting; see, for example, Haberman (1974), Agresti (2010) and Ishii-Kuntz (1994). For OLLM’s, the parameter of interest reflects the linear-by-linear association between the ordered categorical variables. Rather than considering an iterative algorithm for estimating this parameter, one may consider a more direct, non-iterative, approach originally proposed by Beh and Davy (2004) and further studied by Beh and Farver (2009, 2012a, 2012b, 2012c). The two key estimates we shall consider here are the LogNI estimate (first proposed in Beh and Farver, 2009) and the BDNI estimate (Beh and Davy, 2004).

The underlying mathematical links between these two non-iterative estimates of the linear-by-linear association parameter is well established. Their reliability and stability have been verified through simulation. This paper demonstrates that the LogNI and BDNI estimates of the linear-by-linear association parameter are unbiased. A formal proof of this result for the LogNI estimate is given, and a simulation study is undertaken to confirm the unbiasedness of both non-iterative estimates.

2 Non-Iterative Estimation of the Linear-by-Linear Association Parameter

For a doubly ordered \( I \times J \) contingency table, \( N \), denote the proportion of individuals/units in the \((i, j)\)th cell as \( p_{ij} = n_{ij}/n \) where \( n_{ij} \) is the \((i, j)\)th cell value of \( N \), for \( i = 1,2,\ldots,I \) and \( j = 1,2,\ldots,J \). Hence \( \sum_{i=1}^{I} \sum_{j=1}^{J} p_{ij} = 1 \). Let the \( i \)th row and \( j \)th column marginal frequencies be denoted by \( n_i = \sum_{j=1}^{J} n_{ij} \) and \( n_j = \sum_{i=1}^{I} n_{ij} \) respectively so that \( \sum_{i=1}^{I} n_i = \sum_{j=1}^{J} n_j = n \). Let \( p_i = n_i/n \) and \( p_j = n_j/n \) be the \( i \)th row marginal and \( j \)th column marginal proportions respectively. Moreover, let \( m_{ij} \) be the expected cell frequency of the \((i, j)\)th cell under some model. The OLLM for the doubly ordered two way table that is considered by Beh and Farver (2009) and Agresti (2010), is

\[
\ln m_{ij} = \mu + \alpha_i + \beta_j + \phi (u_i - \bar{u})(v_j - \bar{v})
\]

(1)

where \( \mu \) is the grand mean of the expected cell frequencies while \( \alpha_i \) and \( \beta_j \) are the main effects of the \( i \)th row and \( j \)th column respectively. These parameters can be treated as nuisance parameters and have closed form estimates; refer to Beh and Davy (2004). To reflect the structure of ordinal variable, scores are assigned to each of the categories. Here \( u_i \) and \( v_j \) are \textit{a priori} chosen scores and are associated with the \( i \)th row and \( j \)th column respectively. The choice, and impact, of \textit{a priori} chosen scores has received considerable attention in the categorical data analysis literature – see, for example Ishii-Kuntz (1994, pg 30 – 31), Beh (1998), Rayner and Best (2001) and Agresti (2010, pg 8). In this paper, the parameter of interest in (1) is \( \phi \) and is the measure of linear-by-linear association between the variables. When adjacent integer value scores are chosen, so that \( u_i = i \) and \( v_j = j \), \( \phi \) is the common log-odds ratio of the contingency
Beh and Farver (2009) considered two non-iterative estimation methods for the parameter $\phi$. The first estimate is

$$\hat{\phi}_{\text{LogNI}} = \frac{1}{\sigma_1 \sigma_2} \sum_{i=1}^{l} \sum_{j=1}^{f} p_i p_j (u_i - \bar{u})(v_j - \bar{v}) \ln \left( \frac{p_{ij}}{p_i p_j} \right)$$

and is termed the LogNI estimate of the parameter $\phi$. The second non-iterative estimate of $\phi$, and originally proposed by Beh and Davy (2004), is

$$\hat{\phi}_{\text{BDNI}} = \frac{1}{\sigma_1 \sigma_2} \sum_{i=1}^{l} \sum_{j=1}^{f} p_i p_j (u_i - \bar{u})(v_j - \bar{v})$$

and is termed the BDNI estimate of the parameter. For (2) and (3), $\bar{u} = \sum_{i=1}^{l} p_i u_i$, $\bar{v} = \sum_{j=1}^{f} p_j v_j$, $\sigma_1^2 = \sum_{i=1}^{l} p_i u_i^2 - \bar{u}^2$ and $\sigma_2^2 = \sum_{j=1}^{f} p_j v_j^2 - \bar{v}^2$. Since the logarithm function in (2) leads to non-existent estimates of $\phi$ for zero cell’s, strategies need to be considered for dealing with the presence of random zero cell frequencies. Here, we shall consider replacing a zero cell frequency with 0.05. Note that $\hat{\phi}_{\text{BDNI}}$ does not suffer from this zero cell frequency problem. For both (2) and (3), when there is complete independence between the ordered row and ordered column categories, so that $p_{ij} = p_i p_j$ for all $i = 1, 2, ..., l$ and $j = 1, 2, ..., f$, the non-iterative estimate of $\phi$ is zero. In the following section, we shall demonstrate that $\hat{\phi}_{\text{LogNI}}$ is an unbiased estimate of the linear-by-linear parameter $\phi$. It can be similarly shown that the BDNI estimate is also unbiased. A formal mathematical treatment of this issue will be left for future discussion, but section 4 provides a simulation study verifying this property for both non-iterative estimates.

### 3 Unbiasedness of $\hat{\phi}_{\text{LogNI}}$ and $\hat{\phi}_{\text{BDNI}}$

In this section, the mean of the sampling distributions of $\hat{\phi}_{\text{LogNI}}$ is determined. Our discussion of the asymptotic distributional issues of the estimates is made using the delta method (Casella and Berger, 2002, pp.240; Agresti, 2002, pp.73). We shall also be making use of the following restrictions on the a priori scores $u_i$ and $v_j$:

$$\sum_{i=1}^{l} p_i (u_i - \bar{u})(u'_i - \bar{u}) = \begin{cases} \sigma_1^2 & \text{if } u_i = u'_i \\ 0 & \text{if } u_i \neq u'_i \end{cases}$$

$$\sum_{j=1}^{f} p_j (v_j - \bar{v})(v'_j - \bar{v}) = \begin{cases} \sigma_2^2 & \text{if } v_j = v'_j \\ 0 & \text{if } v_j \neq v'_j \end{cases}$$

in the following derivation.

Suppose that the cell frequencies of $N$, represented as elements of the vector $\mathbb{n} = \{ n_{ij}; \forall i, \forall j \}$, are independent Poisson random variables with parameter $\lambda$. By
substituting \( n_{ij} = np_{ij} \) into (2) and taking the expectation of both sides of the resulting expression, we get

\[
E[\hat{\phi}_{LogNI}] = \frac{1}{\sigma_i} \sum_{i=1}^{l} \frac{1}{\sigma_j} \sum_{j=1}^{l} p_i p_j (u_i - \bar{u})(v_j - \bar{v}) E[\ln(p_{ij})] - \frac{1}{\sigma_i} \sum_{i=1}^{l} \frac{1}{\sigma_j} \sum_{j=1}^{l} p_i p_j (u_i - \bar{u})(v_j - \bar{v}) E[\ln(p_{i,j})]
\] (5)

By using the delta method, such that

\[
[\ln(n_{ij}) - \ln(m_{ij})] \rightarrow N \left( 0, \frac{1}{nm_{ij}} \right)
\]

the expectation of \( \hat{\phi}_{LogNI} \), given by (5), can be expressed as

\[
E[\hat{\phi}_{LogNI}] = \frac{1}{\sigma_i} \sum_{i=1}^{l} \frac{1}{\sigma_j} \sum_{j=1}^{l} p_i p_j (u_i - \bar{u})(v_j - \bar{v}) E[\ln(m_{ij})] - \frac{1}{\sigma_i} \sum_{i=1}^{l} \frac{1}{\sigma_j} \sum_{j=1}^{l} p_i p_j (u_i - \bar{u})(v_j - \bar{v}) E[\ln(np_{i,j})]
\] (6)

By considering an unsaturated (linear-by-linear) association model of the type described in Beh and Farver (2009, eqn 24), \( m_{ij} \) is the expected cell frequency of the \((i, j)\)th cell where

\[
m_{ij} = np_{ij} \left( 1 + \phi \frac{(u_i - \bar{u})(v_j - \bar{v})}{\sigma_i \sigma_j} \right)
\]

Therefore, by substituting this unsaturated association model into (6), and by using (4), yields, and after some simplification,

\[
E[\hat{\phi}_{LogNI}] = \phi
\] (7)

Equation (7) shows that \( \hat{\phi}_{LogNI} \) is an unbiased estimator of the linear-by-linear association parameter \( \phi \) in the OLLM of (1). We can similarly show that \( \hat{\phi}_{BDNI} \) is also an unbiased estimate of \( \phi \). The delta method also demonstrates that both non-iterative estimates are asymptotically normally distributed with a calculable variance. This issue will be a topic of further consideration in the future.

4 Computational Study

To study the unbiasedness of the estimates of \( \phi \), \( \hat{\phi}_{LogNI} \) and \( \hat{\phi}_{BDNI} \), 10,000 contingency tables of varying size (2×2 to 5×5) are randomly generated. To reflect the varying magnitude of the cell frequencies these estimates are randomly generated using \( \lambda = 5 \) and 50 reflecting small and large randomly generated cell frequencies.
respectively. For each combination of contingency table size and $\lambda$, the non-iterative estimates of $\phi$ are calculated yielding $\hat{\phi}_{\text{LogNI}}$ and $\hat{\phi}_{\text{BDNI}}$. Therefore a population of 10,000 $\hat{\phi}_{\text{LogNI}}$ values and a population of 10,000 $\hat{\phi}_{\text{BDNI}}$ values are obtained. The means of these estimates are denoted by $\overline{\phi}_{\text{LogNI}}^p$ and $\overline{\phi}_{\text{BDNI}}^p$ respectively. From each population, samples of $\hat{\phi}_{\text{LogNI}}$ and $\hat{\phi}_{\text{BDNI}}$ are randomly selected; small samples of $\hat{\phi}_{\text{LogNI}}$ and $\hat{\phi}_{\text{BDNI}}$ are deemed to be those representing 0.1%, 1% and 5% of the population. Similarly, moderate sized samples (15%) and large samples (30% and 50%) are also considered. For each sample size, the mean of the sampling distribution of $\hat{\phi}_{\text{LogNI}}$ and $\hat{\phi}_{\text{BDNI}}$ values obtained are denoted by $\hat{\phi}_{\text{LogNI}}^n$ and $\hat{\phi}_{\text{BDNI}}^n$ respectively.

Table 1: Study of bias for $\hat{\phi}_{\text{LogNI}}$ for varying contingency table size, sample size and $\lambda$ of simulated contingency tables

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>n</th>
<th>2×2</th>
<th>3×4</th>
<th>4×5</th>
<th>5×5</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.1%</td>
<td>-0.002</td>
<td>0.0004</td>
<td>-0.0029</td>
<td>0.0007</td>
</tr>
<tr>
<td></td>
<td>1%</td>
<td>-0.0007</td>
<td>-0.0004</td>
<td>0.0006</td>
<td>0.0017</td>
</tr>
<tr>
<td></td>
<td>5%</td>
<td>0.0002</td>
<td>0.0003</td>
<td>0.0022</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>15%</td>
<td>-0.0002</td>
<td>-0.0002</td>
<td>-0.0007</td>
<td>-0.0013</td>
</tr>
<tr>
<td></td>
<td>30%</td>
<td>-0.0002</td>
<td>-0.0002</td>
<td>-0.0017</td>
<td>0.0001</td>
</tr>
<tr>
<td></td>
<td>50%</td>
<td>-0.0026</td>
<td>-0.0026</td>
<td>0.0013</td>
<td>0.0027</td>
</tr>
<tr>
<td>50</td>
<td>0.1%</td>
<td>0.0003</td>
<td>0.0018</td>
<td>-0.0004</td>
<td>-0.0004</td>
</tr>
<tr>
<td></td>
<td>1%</td>
<td>-0.0007</td>
<td>-0.0003</td>
<td>-0.0001</td>
<td>-0.0002</td>
</tr>
<tr>
<td></td>
<td>5%</td>
<td>0.0005</td>
<td>0.0004</td>
<td>0.0006</td>
<td>-0.0001</td>
</tr>
<tr>
<td></td>
<td>15%</td>
<td>0.0004</td>
<td>0.0004</td>
<td>-0.0003</td>
<td>-0.0003</td>
</tr>
<tr>
<td></td>
<td>30%</td>
<td>0.0009</td>
<td>0.0009</td>
<td>0.0000</td>
<td>0.0004</td>
</tr>
<tr>
<td></td>
<td>50%</td>
<td>0.0016</td>
<td>0.0016</td>
<td>0.0003</td>
<td>-0.0004</td>
</tr>
</tbody>
</table>

Table 2: Study of bias for $\hat{\phi}_{\text{BDNI}}$ for varying contingency table size, sample size and $\lambda$ of simulated contingency tables

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>n</th>
<th>2×2</th>
<th>3×4</th>
<th>4×5</th>
<th>5×5</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.1%</td>
<td>0.0001</td>
<td>0.0011</td>
<td>-0.0020</td>
<td>-0.0006</td>
</tr>
<tr>
<td></td>
<td>1%</td>
<td>-0.0007</td>
<td>-0.0007</td>
<td>0.0009</td>
<td>0.0012</td>
</tr>
<tr>
<td></td>
<td>5%</td>
<td>0.0001</td>
<td>0.0002</td>
<td>0.0020</td>
<td>0.0002</td>
</tr>
<tr>
<td></td>
<td>15%</td>
<td>-0.0001</td>
<td>-0.0001</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>30%</td>
<td>-0.0002</td>
<td>-0.0002</td>
<td>-0.0008</td>
<td>-0.0008</td>
</tr>
<tr>
<td></td>
<td>50%</td>
<td>-0.0018</td>
<td>-0.0018</td>
<td>0.0007</td>
<td>-0.0007</td>
</tr>
<tr>
<td>50</td>
<td>0.1%</td>
<td>0.0005</td>
<td>-0.0012</td>
<td>-0.0004</td>
<td>-0.0004</td>
</tr>
<tr>
<td></td>
<td>1%</td>
<td>-0.0007</td>
<td>-0.0008</td>
<td>-0.0001</td>
<td>-0.0002</td>
</tr>
<tr>
<td></td>
<td>5%</td>
<td>0.0004</td>
<td>0.0004</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>15%</td>
<td>0.0004</td>
<td>0.0004</td>
<td>-0.0002</td>
<td>-0.0002</td>
</tr>
<tr>
<td></td>
<td>30%</td>
<td>0.0009</td>
<td>0.0009</td>
<td>-0.0001</td>
<td>-0.0001</td>
</tr>
<tr>
<td></td>
<td>50%</td>
<td>0.0016</td>
<td>0.0016</td>
<td>0.0003</td>
<td>0.0003</td>
</tr>
</tbody>
</table>

Table 1 summarises the computational study population mean and the mean of the sampling distribution of the LogNI and verifies the unbiasedness of the estimate - see
It can be seen that as the sample size increases, the population mean of the estimates obtained from (2), $\hat{\Phi}_{\text{LogNI}}^P$ and the mean of the sampling distribution, $\hat{\Phi}_{\text{LogNI}}^r$, become increasingly similar. It is also to be noted that for larger value of $\lambda$, the difference between $\hat{\Phi}_{\text{LogNI}}^P$ and $\hat{\Phi}_{\text{LogNI}}^r$ is zero for large samples of $\hat{\Phi}_{\text{LogNI}}$ and is minimal even for the small sample sizes. Similar conclusions can be made concerning the unbiasedness of the estimates, $\hat{\Phi}_{\text{BDNI}}$. Table 2 compares the population mean and the mean of the sampling distribution for these estimates. The unbiasedness of the BDNI estimate is apparent from observing the near zero difference between $\hat{\Phi}_{\text{LogNI}}^c$ and $\hat{\Phi}_{\text{BDNI}}^r$ for each contingency table size and Poisson parameter $\lambda$ chosen.

5 Discussion

This paper has theoretically, and through simulation, demonstrated that the LogNI estimate of the linear-by-linear association parameter is unbiased. This supports the findings of Beh and Farver (2009, 20012a, 2012b) who confirmed that this estimate is an excellent non-iterative alternative to the algorithms traditionally used. However, more research needs to be undertaken to study the distributional characteristics of $\hat{\Phi}_{\text{LogNI}}^P$ and $\hat{\Phi}_{\text{BDNI}}^r$, including the variance of the estimates. Such issues will be left for further consideration.

References