J.Natn.Sci.Foundation Sri Lanka 2012 ***40****(2):137-148*

RESEARCH ARTICLE

Generalized Cochran Mantel Haenszel test for multilevel correlated categorical data: an algorithm and R function

D.B.U.S. De Silva and M.R. Sooriyarachchi*
Department of Statistics, Faculty of Science, University of Colombo, Colombo 03.

Revised: 12 December 2011 ; Accepted: 20 January 2012

Abstract: Multilevel data are a commonly encountered phenomenon in many data structures. Modelling such data requires careful consideration of the association between underlying variables at each level of the data structure. This requires the use of effective univariate techniques prior to modelling. However, currently no univariate tests are used to handle this situation. This paper presents the modification and novel application of a test developed by Zhang and Boos for testing the association between categorical variables measured on clusters of observations, for examining initial association in a multilevel framework. Zhang and Boos have used a SAS/IML programme (unpublished) for performing their test. This paper presents an R function for the application of the test, which will be freely available to users, since R is an open source software. The function is tested on a dataset from the medical field pertaining to respiratory disease severity of patients, attending several different clinics. The explanatory variables pertaining to this study are Age, Gender, Duration and Symptom, while the response variable indicating the severity of the diagnosis made is termed Diagnosis. The results indicate that when the experimental units show low levels of correlation within clusters with respect to a particular explanatory variable, the test performs similarly to the Standard Cochran Mantel Haenszel (CMH) test. When the corresponding correlation is high, the Generalized CMH (GCMH) test results in a smaller p-value than the Standard CMH test. Of the four variables, only Symptom and Duration are significant with respect to association with Diagnosis.

Keywords: Algorithm, clustered data, generalized Cochran Mantel Haenszel (GCMH) test, multilevel correlated categorical data, R functions.

INTRODUCTION

Ordered Categorical Responses are often encountered especially in educational, social and medical data (Kuruppumullage & Sooriyarachchi, 2007). Common examples are attitude measurements, disease severity and examination grades. The analysis of multilevel ordered categorical responses is another interesting branch of study, which requires special statistical methods (Hedeker & Gibbons, 1994; Fielding *et al.*, 2003; Goldstein, 2003; Hedeker *et al.*, 2006). The standard procedure for dealing with the ordered category scale in multilevel modelling is to assign a score to these categories and to treat it as a continuous variable (Goldstein, 1991). However, there are many questionable aspects to this method. A more suitable approach is the Generalized Linear Multilevel Model (GLMM), which preserves the grouping of the response variable (Fielding & Yang, 2005). Models are highly technical and difficult to fit, particularly in the presence of a large number of variables. In addition to this, non-statisticians find it difficult to understand these models. Univariate tests on the other hand, are much simpler and can easily be understood by anyone with even a basic statistical knowledge. Univariate tests have many uses. While the most general use would be to identify relationships between variables, univariate tests also provide a prelude to modelling by helping to select important variables for modelling when there are several variables. Most of the commonly used univariate tests such as the Pearson's Chi-Squared test are only applicable to non-clustered and uncorrelated data. However, recent years have brought about the development of many advanced univariate tests, capable of handling different data structures.

The focus of this paper is on the univariate techniques available for multilevel data, specifically with respect to ordered categorical responses. Though multilevel modelling techniques have been in development since the late 1980s (Aitkin *et al.*, 1981; Aitkin & Longford, 1986; Longford, 1987), multilevel modelling for ordinal

* Corresponding author (roshini@mail.cmb.ac.lk)

categorical responses is somewhat of a modern approach (Rashbash *et al.*, 2004). The developers of multilevel methodology, Goldstein (2003), Hedeker and Gibbons (1994), Rashbash *et al.*(2004) and many other authors have used examples in the development of multilevel methodology where every available variable was used in the model. Thus, it is clear that application of specific univariate techniques for multilevel data structures have not yet been explored especially in the case of ordered categorical responses. Even though many univariate techniques that deal with unordered/ordered categorical data such as the Pearson's Chi-Squared test are in existence, certain inherent characteristics of multilevel data structures render these usual univariate techniques inapplicable. Two such characteristics are the stratified nature of the data and the correlation between individual responses of units clustered within the higher levels. The first characteristic causes the usual Pearson Chi-Squared test inapplicable. A popular solution to this problem is the use of the GCMH test (Landis *et al*., 1978). However, the presence of intra-cluster correlations poses a significant problem with this test. Zhang and Boos (1997) explains the implications of the intra-cluster correlation when carrying out univariate tests and proposes three new statistics that may be used in the presence of correlated categorical data. The three new statistics proposed by Zhang and Boos (1997) takes the same form as the GCMH statistic proposed by Landis *et al.* (1978), but uses different covariance matrix estimators to avoid failure in the presence of correlation. The paper comprehensively discusses the formulation of these statistics under three different types of alternative hypotheses.

There are two main objectives to this paper. Firstly to present the modification and novel application of a test proposed by Zhang and Boos (1997) for correlated categorical data, to an ordered categorical multilevel data structure, and secondly to develop a user-friendly R function that can be freely used for the computation of this particular statistic for 2-level non-repeated multilevel data. The R function written is developed for the most basic case, but is easily adaptable for higher order multilevel data after certain modifications. Thus this paper attempts to make new contributions to research in the fields of multilevel modelling and computational statistics.

At this point, it is important to briefly examine the nature of the data used in this study. The data for this study was provided by the Primary Care Respiratory Group Sri Lanka, a member of the prestigious International Primary Care Respiratory Group (IPCRG) based in the United Kingdom (IPCRG, 2007). Fourteen different family physicians participated in this study, where each physician collected the relevant data during a stipulated time period. The data set consists of 7 variables spread across two main levels. The second level unit of the dataset can be identified as the physician, while the 1st level unit comprises individual patients. The level one variables, which are of relevance to this paper, comprise the patient age (in years), gender, most prevalent symptom, duration of the symptom (in days) and the severity of the disease diagnosed by the physician. The age of patients, most prevalent symptom and the duration were categorized into an appropriate number of categories for modelling purposes. The categorization was done on a logical basis and on medical grounds.

REVIEW OF LITERATURE

Description

Prior to fitting any statistical model, it is always important to test the nature and the strength of the relationship between the explanatory variables and the response. Univariate tests provide the means of assessing these relationships and hence are a precursor for selecting variables to the initial stage of the model. Univariate methods vary according to the nature of the variables in question. Additionally, most techniques also depend on various other conditions that need to be met, prior to using these. The variables involved in this study were all nominal/ordinal categorical variables, while the response, namely 'severity of disease' was an ordinal categorical variable. In addition to this, the data involved in the study had a multilevel structure. As indicated in the introduction, this factor renders usual univariate techniques such as the Pearson Chi-squared test and the GCMH test (Landis *et al*., 1978) to be inapplicable. Thus it was necessary to identify a suitable technique for assessing the nature of the relationships among variables, taking into account the structure of the data.

A brief review of past developments

The theory governing the Cochran Mantel Haenszel (CMH) statistic first surfaced in 1954. The theory was presented by Cochran (1954), as an alternative method to overcome problems of the Chi-Squared test in the presence of stratified data. This statistic was further developed by Mantel and Haenszel (1959). This initial statistic was developed to compare two binary variables, adjusting for control variables. Though the standard Cochran Mantel Haenszel test was able to overcome the problem with respect to stratified variables, one major drawback to the test was its limitation to binary variables. As a solution to this problem Landis *et al*. (1978)

proposed the Generalized Cochran Mantel Haenszel (GCMH) statistic, which was a multivariate extension of the CMH test, and thus was able to handle variables with two levels or more. This statistic was shown to have an approximate Chi-Squared distribution under the assumption of independence between observations. However, in the presence of within strata/cluster correlation, their variance-covariance matrix was shown to be invalid. Thus the need for a test statistic capable of handling clustered correlated categorical data arose.

Liang (1985) proposed a GCMH score test for correlated binary responses, which was also able to handle sparse data. This statistic had two major limitations. Firstly it was only applicable to binary data. Thus, this test could not be applied to variables with more than two levels. Secondly, simulation studies showed that this statistic was somewhat conservative in the presence of a smaller number of strata.

As indicated in the introduction, Zhang and Boos(1997) proposed three new statistics, based on the GCMH statistic (Landis *et al*., 1978) that was both able to handle correlated data and variables with two or more levels.

Summarized review of Zhang and Boos (1997) test procedure

Description: Zhang and Boos (1997) proposed a new testing approach for use in the presence of correlated categorical data. The GCMH test for correlated categorical data provide three different forms of test statistics, namely, T_{EL}, T_P and T_U, which could be used in place of more traditional tests for testing associations between stratified categorical variables. Zhang and Boos (1997) presented a detailed discussion about the suitability of each of the statistics for various situations involving stratified data, sparse data and missing data. The statistic T_{EL} was a direct generalization of the statistic proposed by Liang (1985), for categorical data and was specifically designed to handle the many-strata situation. Thus T_{EL} proved to be more liberal in the presence of a smaller number of strata.

The other two statistics, T_P and T_U were extensions of the statistics previously proposed by Zhang and Boos for binary data. Both T_P and T_U are asymptotically valid when there are many strata with the number of observations in each strata being relatively small, as well as in the presence of a fewer number of large strata. As indicated by Zhang and Boos (1997), one major drawback of T_{EL} is its use of strata as the primary sampling unit in its variance estimator. This affects the efficiency of the statistic. However, both T_P and T_U use individual subjects as the primary sampling unit. Hence, both these statistics are more efficient than T_{EL} especially in the case of a smaller number of strata. Thus for a multilevel correlated categorical data structure, the above reasons justify the use of the statistics T_P and T_U proposed by Zhang and Boos (1997) over all the other statistics presented above.

This paper presents the modification and application of the statistic T_P to the 2-level non-repeated measures data structure. The main reason governing the selection of the statistic T_P over T_U and T_{EL} for application to the multilevel data structure is the simulation study done by Zhang and Boos (1997), which showed that T_P maintained error values even for a small number of strata. In addition to this, T_P was also shown to be a better choice over T_U due to its usage of pooled estimators for estimating the variance. Thus, considering the relatively small number of strata in the considered dataset (14 practices/physician) and taking into account the simplicity of calculation of the statistic, T_P was selected as the most suitable statistic for this study.

Theory: The following sections present the theory of the GCMH test (Landis *et al*., 1978) and the T_P test procedure proposed by Zhang and Boos (1997).

Consider a data structure with q strata where in each stratum subjects are exposed to a treatment scheme (explanatory variable) with R being the number of levels of treatment, and a response structure consisting of C number of response categories. Let x_{hijk} denote the number of times the k^{th} individual in the i^{th} treatment level of the h^{th} stratum, receives a response of level j. Then n_{hik} refers to the number of repeated measurements on the k^{th} individual, and n_{hi} refers to the number of subjects in the i^{th} treatment level of the h^{th} stratum. Table 1 gives the data structure for the h^{th} stratum.

Table 1: Data Structure in Stratum h

Explanatory variable levels (i)	Response Variable Categories (j) 1	2	.	j	.	C	Total
1	$x_{h11.}$	$x_{h12.}$	.	$x_{h1j.}$	.	$x_{h1c.}$	$n_{h1.}$
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
R	$x_{hR1.}$	$x_{hR2.}$	.	$x_{hRj.}$	.	$x_{hRC.}$	$n_{hR.}$
Total	t_{h1}	t_{h2}	.	t_{hj}	.	t_{hC}	N_h

Let $\pi_{hi*} = (\pi_{hi1}, \pi_{hi2},..., \pi_{hiC})'$ where π_{hij} is the probability that a single multinomial response is in the j^{th} category for the i^{th} treatment level and the h^{th} stratum.

Then it can be stated that $x_{hi*k} = (x_{hi1k}, x_{hi2k},..., x_{hiCk})'$ has a correlated multinomial distribution with parameters π_{hi*}, n_{hik} and covariance matrix $\sum_{hi}$. Also note that $x_{hi*\bullet} = (x_{hi1\bullet}, x_{hi2\bullet},..., x_{hiC\bullet})'$ denotes the sum of x_{hi*k} over k.

Let us define $x_h = (x_{h1*\bullet}', x_{h2*\bullet}',..., x_{hR*\bullet}')'$, and $m_h = N_h(P_{h\bullet*} \otimes P_{h*\bullet})$ where $P_{h*\bullet} = (P_{h1\bullet}, P_{h2\bullet},..., P_{hR\bullet})'$ and $P_{h\bullet*} = (P_{h\bullet1}, P_{h\bullet2},..., P_{h\bullet C})'$. Here, $P_{hi\bullet} = n_{hi\bullet}/N_h$, $P_{h\bullet j} = t_{hj}/N_h$ and N_h refers to the h^{th} stratum total, $n_{hi\bullet}$ refers to the i^{th} treatment total in the h^{th} stratum and t_{hj} refers to the j^{th} response category total in the h^{th} stratum. $\otimes$ denotes the Kronecker product multiplication.

In addition to the above definitions, Zhang and Boos also make the assumptions that $\{x_{hi*k}\}$ are independent from each other within and across the strata, and the expectation of x_{hi*k} is given by $n_{hik}\pi_{hi*}$. Zhang and Boos (1997) presented three different alternative hypotheses that may be tested and clearly present the different adjustments that need to be made. However, our focus is mainly on the alternative hypothesis of general association given as follows. The reason for this selection is that with respect to categorical variables in a multilevel data structure, the most applicable hypothesis to be tested, in order to select variables for the initial stage of modelling, is the hypothesis of general association (Zhang & Boos, 1997). Accordingly the overall null hypothesis of no treatment effect is given as,

$H_0 : \pi_{h1*} = \pi_{h2*} = ... = \pi_{hR*}$, for $h = 1,2,...q$

The GCMH statistic proposed by Landis *et al.* (1978) is as follow.

$$T_{CMH} = G'V_{CMH}^{-1}G \quad ...(1)$$

where $$G = \sum_{h=1}^{q} G_h = \sum_{h=1}^{q} B_h(x_h - m_h) \quad ...(2)$$

and $$B_h = C_h \otimes R_h \quad ...(3)$$

For the alternative hypothesis of general association $R_h = (I_{R-1}, -J_{R-1})$ and $C_h = (I_{C-1}, -J_{C-1})$ where I_{R-1} and I_{C-1} are identity matrices of rank R-1 and C-1 respectively. J_{R-1} and J_{C-1} are each a $(R\text{-}1)\times 1$ and $(C\text{-}1)\times 1$ vector of 1s.

$$V_{CMH} = \sum_{h=1}^{q} B_h V_{Gh} B_h' \quad ...(4)$$

$$V_{Gh} = \frac{N_h^2}{(N_h - 1)\{(D_{Ph\bullet*} - P_{h\bullet*}P'_{h\bullet*}) \otimes (D_{Ph*\bullet} - P_{h*\bullet}P'_{h*\bullet})\}} \quad ...(5)$$

The matrix D_a represents a diagonal matrix with the elements of a along its main diagonal.

The T_P statistic proposed by Zhang and Boos (1997) is as follows.

$$T_P = G'V_P^{-1}G \quad ...(6)$$

where G takes the same form as explained previously.

$$V_p = \sum_{h=1}^{q} [B_h V_{Ph} B_h'] \quad ...(7)$$

$$V_{Ph} = \sum_{i=1}^{R} \{A_{hi} \{ \sum_{k=1}^{n_{hi}} \left[\frac{(x_{hi*k} - n_{hik}\hat{\pi}_h)(x_{hi*k} - n_{hik}\hat{\pi}_h)'}{1 - \frac{n_{hik}}{N_h}} \right] \} A_{hi}' \} \quad ...(8)$$

In the above expression $\hat{\pi}_h = \left(\frac{t_{h1}}{N_h}, \frac{t_{h2}}{N_h}, \ldots, \frac{t_{hC}}{N_h}\right)'$

and $A_{hi} = I_C \otimes \Lambda_{hi}$ where

$\Lambda_{hi} = (\lambda_{h1}, \lambda_{h2}, \ldots, \lambda_{h(i-1)}, \lambda_{hi}^*, \lambda_{h(i+1)}, \ldots, \lambda_{hR})'$

and $\lambda_{hi} = -\frac{n_{hi\bullet}}{N_h}$ and $\lambda_{hi}^* = 1 - n_{hi\bullet}/N_h$

Then following the theorems in Zhang and Boos (1997), $T_P = G'V_P^{-1}G \sim x_{df}^2$ where df is the rank of B_h. Accordingly if, $T_P > x_{df,\alpha\%}^2$, we reject H_0 in favour of H_1, at the $\alpha\%$ level of significance.

The derivation of T_{EL} and T_U are also similar and are clearly presented in Zhang and Boos (1997). Their derivation will not be presented here as this paper concentrates only on the statistic T_P. Interested individuals should refer Zhang and Boos (1997) for the formulation of these two statistics.

MODIFIED TEST FOR TWO DIMENSIONAL MULTILEVEL DATA WITHOUT REPEATED MEASURES

The general theory and algorithm proposed by Zhang and Boos (1997) for correlated categorical data were

introduced in the previous section. As explained above, the original test proposed by Zhang and Boos was for correlated categorical data with repeated measures. This section describes the modifications made to the algorithm in order to apply the test to a two dimensional multilevel dataset without repeated measures. The algorithm explained below describes step by step all matrices and their operations that will be applied in the R function developed in a systematic manner.

Consider a data structure with q strata where in each stratum subjects are exposed to a treatment scheme (explanatory variable) with R being the number of levels of treatment, and a response structure consisting of C number of response categories. Also consider that each subject is exposed to the corresponding treatment scheme only once and hence only provides a single response. This implies that there are no repeated measures. Thus, if x_{hijk} denote the number of times the k^{th} individual in the i^{th} explanatory variable level of the h^{th} stratum receives a response of level j, then x_{hijk} takes the value of unity or zero. Also n_{hik} referring to the repeated measurements of the k^{th} individual takes the value of unity, and n_{hi} refers to the number of subjects in the i^{th} explanatory variable level of the h^{th} stratum.

Let $\pi_{hi*} = (\pi_{hi1}, \pi_{hi2}, ..., \pi_{hiC})'$ where π_{hij} is the probability that a single multinomial response is in the j^{th} category for the i^{th} explanatory variable level and the h^{th} stratum. The following steps describe in detail the algorithm used for the development of the R function. It should be noted that this algorithm is a slight modification of that proposed by Zhang and Boos (1997) in the sense that adjustments have been made to apply the algorithm to a two dimensional multilevel data structure without repeated measurements. Most of these adjustments are explained in Step I below.

Step I

For a two dimensional multilevel structure without repeated measures, the vector $x_{hi*k} = (x_{hi1k}, x_{hi2k}, ..., x_{hiCk})'$ can be denoted as $(0,0,...,1,...0)'$, if the k^{th} individual in the i^{th} explanatory variable level of the h^{th} stratum provides a response of level j. That is, x_{hi*k} denotes the response vector of the k^{th} individual in the i^{th} treatment group of the h^{th} stratum. Thus, x_{hi*k} has a correlated multinomial distribution with parameters $(\pi_{hi*}, 1)$ (since n_{hik} is unity) and covariance matrix $\sum_{hi}$. Also note that $x_{hi*\bullet} = (x_{hi1\bullet}, x_{hi2\bullet}, ..., x_{hiC\bullet})'$ denotes the sum of x_{hi*k} over k and since each individual in each treatment provides a single response, the vector $x_{hi*\bullet}$ simply denotes the number of subjects (individuals) in each response category in the i^{th} explanatory variable level of the h^{th} stratum.

Step II

The following definitions follow directly from Zhang and Boos (1997).

Let us define $x_h = (x_{h1*\bullet}', x_{h2*\bullet}', ..., x_{hR*\bullet}')'$, and $m_h = N_h(P_{h\bullet*} \otimes P_{h*\bullet})$ with $P_{h*\bullet} = (P_{h1\bullet}, P_{h2\bullet}, ..., P_{hR\bullet})'$ and $P_{h\bullet*} = (P_{h\bullet1}, P_{h\bullet2}, ..., P_{h\bullet C})'$ where $P_{hi\bullet} = n_{hi\bullet}/N_h$, $P_{h\bullet j} = t_{hj}/N_h$

The definitions of N_h, $n_{hi\bullet}$ and t_{hj} are the same as explained above. $\otimes$ denotes the Kronecker product multiplication.

Step III

The null hypothesis that is to be tested is that of no association between the explanatory variable and response. That is,

$H_0 : \pi_{h1*} = \pi_{h2*} = ... = \pi_{hR*}$, for $h = 1,2,...q$ (No association between the explanatory variable and the response)

The alternative hypothesis, if the null hypothesis is rejected, is that the distribution of the response variable differs significantly, in non specific patterns across levels of the row factor, adjusted for strata.

H_1 : General association between the response and levels of the explanatory variable

The following indicates the derivation of the test statistic T_P with necessary adjustments to fit a two dimensional multilevel data structure without repeated measures. As explained by Zhang and Boos (1997), the test statistic takes the form explained in equation (6). The adjustments made to their method are explained next.

For the alternative hypothesis of general association $R_h = (I_{R-1}, -J_{R-1})$ and $C_h = (I_{C-1}, -J_{C-1})$ where I_{R-1} and I_{C-1} are identity matrices of rank $R-1$ and $C-1$ respectively. J_{R-1} and J_{C-1} are each a $(R-1)\times 1$ and $(C-1)\times 1$ vector of 1s and V_P is as given in equation (7).

Then according to the modified form,

$$V_{Ph} = \sum_{i=1}^{R}\{A_{hi}\{\sum_{k=1}^{n_{hi}}\left[\frac{(x_{hi*k} - \hat{\pi}_h)(x_{hi*k} - \hat{\pi}_h)'}{1 - \frac{1}{N_h}}\right]\}A_{hi}'\} \quad ...(9)$$

In the above expression $\hat{\pi}_h = \left(\frac{t_{h1}}{N_h}, \frac{t_{h2}}{N_h}, ..., \frac{t_{hC}}{N_h}\right)'$ and $A_{hi} = I_C \otimes \Lambda_{hi}$ with

$$\Lambda_{h1} = (\lambda^*_{h1}, \lambda_{h2}, \dots, \lambda_{h(i-1)}, \lambda_{hi}, \lambda_{h(i+1)}, \dots, \lambda_{hR})'$$

$$\dots \quad = \quad \dots$$

$$\Lambda_{hi} = (\lambda_{h1}, \lambda_{h2}, \dots, \lambda_{h(i-1)}, \lambda^*_{hi}, \lambda_{h(i+1)}, \dots, \lambda_{hR})'$$

$$\dots \quad = \quad \dots$$

$$\Lambda_{hR} = (\lambda_{h1}, \lambda_{h2}, \dots, \lambda_{h(i-1)}, \lambda_{hi}, \lambda_{h(i+1)}, \dots, \lambda^*_{hR})'$$

Where $\lambda_{hi} = -n_{hi\bullet}/N_h$ and $\lambda^*_{hi} = 1 - n_{hi\bullet}/N_h$
Theories and assumptions related to the above test were presented previously. According to these theories, $T_P = G'V_P^{-1}G \sim x^2_{df}$, where *df* is the rank of B_h under H_0. That is, T_p has an approximate Chi-Squared distribution under H_0. Zhang and Boos (1997) presents two theorems, which include the conditions that need to be satisfied for the approximation to hold.

Accordingly if $T_P > x^2_{df,\propto\%}$ we reject H_0 in favour of H_1, at $\alpha\%$ level of significance.

IMPLEMENTATION

This section presents the implementation of the algorithm described earlier. The R function explained in this section and presented in Appendix A, was designed specifically for testing the association between correlated categorical variables in a two dimensional multilevel data structure. Prior to explaining the function, a brief description of the design procedure will be explained. Three major factors were taken into account when designing the R function. These were, the original algorithm proposed by Zhang and Boos (1997), the SAS IML function developed by Zhang and Boos in order to calculate the test statistics proposed and the adjustments to the original algorithm presented above for applying the statistic T_p for correlated categorical variables in a two dimensional multilevel data structure without repeated measures.

Our function source code contains comments (followed by the # sign) at crucial points in order to explain the function of the codes. Prior to executing the function it is important to note the method of entering data into R. The function requires data to be entered in the form of a data frame.

The function is given the name Tp and requires four arguments to be passed to it, which are the data frame name and variables of the data frame that represent the explanatory variable, response variable and the stratification variable, respectively. At each point of the function relevant matrices have been specified following the same notations used in the algorithm presented in the above section as much as possible. It is noteworthy that the vectors $P_{h*\bullet}$, $P_{h\bullet *}$, x_h and m_h in the algorithm are derived from the columns of the matrices phrow, phcol, Xh and Mh, respectively (refer Appendix A). In addition to calculating the T_p statistic, the function is also designed to return the value of the GCMH (Landis *et al*., 1978) statistic along with its p value, for comparison purposes. The function was developed in R version 2.13.0 and does not require any special R packages for execution. However, since the most basic R functionalities are used in developing the function, it may be executed even in earlier versions of the software. The R function is presented in Appendix A. The function returns the value of T_p, the value of the Landis *et al*. (1978) GCMH statistic (T_{CMH}) and the corresponding p values along with the degrees of freedom. By executing the code as an R-script the relevant test can be carried out.

AN EXAMPLE SESSION

This section illustrates an example session where the GCMH test using test statistic T_p will be carried out on a two dimensional multilevel dataset, using the R function presented in the earlier section. Prior to carrying out the test it is important to present a comprehensive description of the dataset that will be used for the purpose.

Structure of the data

The data for this study was provided by the Primary Care Respiratory Group Sri Lanka, a not-for-profit organization established by ten family physicians who are interested in respiratory medicine. Fourteen different family physicians participated in this study, where each physician collected the relevant data during a stipulated time period (3 months).

The dataset consisted of 7 variables spread across two main levels. The 2nd level unit of the dataset was identified as the physician/practice, while the 1st level unit comprised individual patients. The level 2 variables comprised the qualification of the physician (qualification with regard to family medicine) and the number of years in service (not of relevance to this example). Level 1 variables comprised the patient age (in years), gender, most prevalent symptom, duration of the symptom (in days) and the severity of the disease diagnosed by the physician.

The response variable of interest is the severity of the respiratory infection diagnosed by the physician at the end of the examination. Though the initial data

classified symptoms and diagnosis according to the ICHPPC (International Classification of Health Problems in Primary Care) classification (Slocum, 1977), the diagnosed diseases were categorized as 'mild', 'moderate' or 'severe', according to the severity of the disease while the symptoms were categorized based on their frequency of occurrence as 'frequently encountered' and 'not frequently encountered', based on a medical basis to suit the statistical modelling procedures. The initial dataset contained 3814 patient's records. However, after carrying out the necessary data cleaning procedures, the final number of records used for the analysis was 2966. A limitation of the Zhang and Boos (1997) test is that no row/column totals could be zero. Thus, one of the practices was entirely removed from the study since it had no records in a particular age group, and hence the final analysis took into account only 13 practices.

As explained above, the dataset in this study took a hierarchical form with respect to patients being clustered within practices. In the point of view of the univariate analysis the 'Practice' was considered as the stratification factor according to which patients were clustered. The response variable was termed as 'Diagnosis', referring to the severity of the respiratory disease present in each patient. Of the four original explanatory variables at the patient level, two variables, namely age and the duration of symptom were continuous variables. Thus, they were categorized on medical grounds in order to maintain all four explanatory variables as categorical. The explanatory variables were denoted as 'Gender', 'Age', 'Symptom' and 'Duration'.

The variables used in the study along with their groupings are presented in Table 2.

Table 2: Variables and their groupings

Variable	Grouping
Practice (stratification variable)	1-13
Gender	Male Female
Age	Infants and pre-schoolers School going Adolescents and adults
Symptom	Frequently encountered Not frequently encountered
Duration	≤ 5 days > 5 days
Diagnosis (response variable)	Mild Moderate Severe

Calculations and interpretations

The portion of the univariate analysis presented in this paper is related to the patient level with the intension of identifying the effect of explanatory variables on the response. Since we expect intra-cluster correlation with respect to at least some explanatory variables, the GCMH test for correlated categorical data needs to be used in place of the traditional Chi-Squared techniques. The test statistic used in this study is the statistic termed T_P as indicated in Zhang and Boos (1997). It should be noted that the intension of this paper is to present modifications to the T_P test algorithm presented by Zhang and Boos (1997), so that the test will be applicable for two dimensional multilevel data without repeated measures and to present an R function for its computation. The reasons governing the choice of the statistic T_P over the other two were discussed earlier.

A detailed description of the computation of T_P for examining the association between the variable Symptom and the response Diagnosis is as follows.
According to the algorithm,

$$T_P = G' V_P^{-1} G$$

Prior to executing the R function, the dataset needs to be imported to R, using the following code. 'Data.csv' represents the data file saved in csv format. This file is first imported to a data frame in R.

```
#Loading csv datafile to dataframe A
A<-read.csv("Data.csv", header=TRUE)
```

Once the function T_P is entered in R, the corresponding T_P value, T_{CMH} value, p values for both statistics and degrees of freedom are obtained as follows. The corresponding arguments to the function T_P is passed using the following code (considering the relationship between the variable Symptom and response Diagnosis).

```
Tp (A,Symptom,Diagnosis,Practice)
```

The following values were obtained

$T_P = 83.206$ P-Value $[T_P] = 8.551e\text{-}19$

degrees of freedom = 2

$T_{CMH} = 79.023$ P-Value $[T_{CMH}] = < 2.2e\text{-}16$

The function presented in the implementation section computes the above information. In addition to the above values, any intermediate matrices/vectors can also be printed by adding print statements at required points in the function or by uncommenting (by removing the # sign) the print statements, which have already been included in the source code.

According to the above p value, it can be concluded that there is a highly significant relationship between Symptom and Diagnosis.

The T_P values and p values obtained for each of the four explanatory variables when tested with the response variable are indicated in Table 3. In addition the value of the T_{CMH} (GCMH of Landis *et al*.,1978) and the corresponding p values are also included for comparison in Table 3.

According to the p values associated with the T_P values in Table 3, it is evident that only the variables Symptom and Duration show significant associations with the response variable Diagnosis.

Comparison of T_P and T_{CMH} values

At this point it may also be of interest to observe the differences in values and the significance between the T_P statistic and the Landis *et al.*(1978) Cochran Mantel Haenszel statistic (T_{CMH}). As explained previously

Table 3: T_P *vs*. T_{CMH}

Variable	T_P Test		T_{CMH} Test		DF
	T_P	p-value	T_{CMH}	p-value	
Symptom	83.206	8.551e-19	79.023	< 2.2e-16	2
Duration	27.593	1.019e-06	26.347	1.900e-06	2
Age	7.441	0.114	6.668	0.155	4
Gender	0.290	0.865	0.291	0.865	2

categorical variables present in multilevel data structures render both the Pearson's Chi-Squared test and the T_{CMH} test to be inapplicable. While the stratified/clustered nature of the data (i.e. in multilevel data structures, lower level units are assumed to be clustered within upper level units) renders the Pearson's Chi-Squared test to be inapplicable, significant intra-cluster correlations among the units, if present, causes the T_{CMH} test to be inapplicable. Thus, one of the GCMH tests of Zhang and Boos (1997) were modified appropriately and applied to overcome these problems. However, it may be advantageous to observe the pattern of deviation between the statistic T_P and T_{CMH} for the example data. The R function presented in the implementation was used to compute the T_P and respective p values. Necessary codes have been provided within the R function [mantelhaen.test()] required to calculate the T_{CMH} statistic and its p value as well. Table 3 presents these results.

From the results indicated in Table 3, it is clearly seen that the variables showing significant relations to the response variable under the T_P test also show significant relations under the T_{CMH} test. However, the T_{CMH} value is less than the T_P value for the variable Symptom by a considerable margin, while a decrease by a smaller margin can be observed for the variable Duration. The variables Age and Gender show similar values for both statistics.

The data based results as well as the technical results obtained and interpreted above will be comprehensively discussed in the following section.

DISCUSSION AND CONCLUSION

In discussing the major findings and conclusions of the study, it is important to consider these in two angles. Firstly the technical conclusions and secondly the data based conclusions. When discussing the technical findings and conclusions, the most important aspect is the discussion of the R function developed. As explained earlier, no univariate techniques have thus far been applied to multilevel data structures. Hence, the development of this function is a contribution to this field. Another advantage of the function is it being developed in R, it can be used freely by all. The algorithm presented earlier that was used in the development of the function is a slightly adjusted version of the algorithm presented in Zhang and Boos (1997), to make it compatible for a two dimensional multilevel data structure without repeated measurements.

Zhang and Boos (1997) presented three test statistics based on the T_{CMH} statistic for dealing with correlated categorical data. These statistics were specifically designed and tested for the situation where repeated measurements were considered. In addition to this, as mentioned earlier, they also wrote a SAS IML programme capable of calculating the above statistics. However, SAS not being a free software and SAS IML being a separate module makes the programme difficult to be acquired and used freely. In designing the R function, the adjusted algorithm presented earlier as well as the SAS

IML programme of Zhang and Boos (1997) were taken into consideration, and the development was done in a systematic manner comparing the results given by the R-function to those yielded by the SAS IML programme at each stage.

Technical and data based findings and conclusion

The major technical findings of this study were the differences observed between the statistic T_P and the T_{CMH} statistic. It was observed that for the variable Symptom the value of T_P was higher than the value of T_{CMH}, while for the variable Duration a lesser increase was observed. The variable Age showed a very slight increase for T_P over T_{CMH}, while the variable Gender showed approximately equal values for both statistics.

As explained in a previous section, the statistic T_P and T_{CMH} varies in the presence of intra-cluster correlations (ICC) with respect to the considered explanatory variable. Thus, the results indicate that there is a significant correlation between patients within the same practice, with regard to the variable Symptom while a considerable correlation also seems to exist with respect to Duration. However, little correlation seems to exist with respect to Age. The patients within the same practice do not seem to be correlated with regard to the variable Gender. These results were also consistent with the values of the intra-cluster correlations calculated (Hu *et al.*, 1998) with respect to each of the four explanatory variables, using the results obtained in the univariate multilevel modelling carried out using the MLwiN package. The univariate multilevel models refer to the multilevel models fitted considering one explanatory variable at a time. According to Hu *et al*. (1998):

$$ICC = \frac{\sigma_v^2}{\sigma_v^2 + \pi^2/3} \quad ...(10)$$

The term σ_v^2 refers to the practice-level variance (obtained for each of the univariate multilevel models fitted) and the term $\pi^2/3$ corresponds to the variance of the standard logistic distribution, where $\pi = 22/7$. The σ_v^2 and ICC values calculated for each of the four explanatory variables are presented in Table 4. When modelling, it was observed that the value of σ_v^2 converged to negative values for two of the variables, namely Age and Gender. Accordingly, these values were set to zero by convention (Nadaraja & Sooriyarachchi, 2009). Table 4 also contains the absolute difference between the calculated T_P and T_{CMH} values for each variable and whether the T_P value was significant or not. According to Table 4, it is clear that the absolute difference between T_P and T_{CMH} was highest for variables Symptom followed by Duration.

Table 4: ICC values and (T_P-T_{CMH}) values

Variable	Whether T_P is significant	σ_v^2	ICC	($T_P - T_{CMH}$)
Symptom	Yes	0.760	0.183	4.183
Duration	Yes	0.707	0.170	1.246
Age	No	0	0	0.753
Gender	No	0	0	0

These two variables also show high ICC values and are also the two variables, which show significant T_P values. Thus, it can be concluded that the value of T_P tends to be significantly higher than the value of T_{CMH} in the presence of intra-cluster correlations (Zhang & Boos, 1997). The data based findings of the study revealed that the variables Symptom and Duration both show significant associations with the response variable Diagnosis, while the association between the response and the variables Age and Gender were each insignificant.

Further work

This research can be considered as the basis for future work, in several arenas. One significant area would be to update the R function presented in implementation for the calculation of the Generalized CMH statistic T_P into a fully programmed R package, which will then facilitate the use of this function as an inbuilt-function in R. In addition to this, since the programme developed is geared to handle two dimensional multilevel data without repeated measures, it may also be advantageous to update the programme to handle higher dimensional data with and without repeated measurements. Another area of development that can be considered is the development of R functions/ packages capable of computing the other two statistics T_U and T_{EL} as well (Zhang and Boos,1997).

Acknowledgement

The authors wish to thank Prof. Dennis Boos, for providing his publication and SAS programmes, Dr. A.L.P. De S. Senevirathna and the Primary Care Respiratory Group Sri Lanka for providing the much required data, and Mrs. A. A. Sunethra for her invaluable assistance regarding R programming.

REFERENCES

1. Aitkin M. & Longford N. (1986). Statistical modelling in school effectiveness studies (with discussion). *Journal of the Royal Statistical Society* **149** (A): 1– 43.

2. Aitkin M., Anderson D. & Hinde J. (1981). Statistical modelling of data on teaching styles (with discussion). *Journal of the Royal Statistical Society* **144** (A):148 – 61.
3. Cochran W. G. (1954). Some methods for strengthening the common Chi-Squared tests. *Biometrics* **10** (4) :417– 451.
4. Fielding A. & Yang M. (2005). Generalized linear mixed models for ordered responses in complex multilevel structures: effects beneath the school or college in education. *Journal of the Royal Statistical Society* **168** (1): 159 –183.
5. Fielding A., Yang M. & Goldstein H. (2003). Multilevel ordinal models for examination grades. *Statistical Modelling* **3** (2) : 127–153.
6. Goldstein H. (1991). Nonlinear multilevel models with an application to discrete response data. *Biometrika* **78** (1): 45 –51.
7. Goldstein H. (2003). *Multilevel Statistical Models*, 3rd edition. Edward Arnold Publishers Ltd., London, UK.
8. Hedeker D. & Gibbons R.D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics* **50** (4): 933 – 944.
9. Hedeker D., Berbaum M. & Mermelstein R. (2006). Location-scale models for multilevel ordinal data: between- and within- subjects variance modeling. *Journal of Probability and Statistical Science* **4** (1): 1–20.
10. Hu F.B., Goldberg J., Hedeker D. & Henderson W.G. (1998). Modelling ordinal responses from co-twin control studies. *Statistics in Medicine* **17** (9): 957–970.
11. International Primary Care Respiratory Group. (2007). Available at *http://www.theipcrg.org/members/srilanka.php*. Accessed on 30 April 2007.
12. Kuruppumullage P. & Sooriyarachchi M.R. (2007). Log linear models for ordinal multidimensional categorical data. *Journal of the National Science Foundation of Sri Lanka* **35** (1): 29– 40.
13. Landis J. R., Heyman E. H. & Koch G. G. (1978). Average partial association in three-way contingency tables: a review and discussion of alternative tests. *International Statistical Review* **46** (3): 237–254.
14. Liang K.Y. (1985). Odds ratio inference with dependent data. *Biometrika* **72** (3): 678 – 682.
15. Longford N.T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika* **74** (4): 817–827.
16. Mantel N. & Haenszel W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* **22**: 719 –748.
17. Nadaraja K. & Sooriyarachchi M.R. (2009). A Monte-Carlo simulation study of the properties of residual maximum likelihood (REML) estimators for the linear gausian mixed model. *Sri Lanka Journal of Applied Statistics* **10**: 119–136.
18. Rashbash J., Steele F., Browne W. J. & Goldstein H. (2004). *A User's Guide to MLwiN*. Centre for Multilevel Modelling, University of Bristol, Bristo, UK.
19. Slocum H. E. (1977). Personal medical reference files for family physicians. *The Journal of Familiy Practice* **5** (4): 593–595.
20. Zhang J. & Boos D.D. (1997). Generalized Cochran-Mantel-Haenszel Test Statistics for correlated categorical data. *Communications in Statistics - Theory and Methods* **26** (8):1813–1837.

Appendix A

```
#User Defined Funtion Tp-Initialisation
#The following variables should be passed from the
data frame to Data,Y,U,C
#Data-Dataframe name Y-Explanatory variable U-
Response variable C-Variable specifying #strata

Tp<-function(Data,Y,U,C)
{
attach(Data)
#Forming 3-way frequency table
B<-table(Y,U,C)
#Defining Xh vectors
xh<-matrix(nrow=dim(B)[1]*dim(B)[2],ncol=dim(B)
[3])
k<-1
while(k<=dim(B)[3])
{
xh[,k]<-t(B[1:dim(B)[1],,k])
k<-k+1
}
#Calculation of row totals
rowtot<-matrix(nrow=dim(B)[1],ncol=dim(B)[3])
k<-1
while(k<=dim(B)[3])
{
j<-1
while(j<=dim(B)[1])
{
rowtot[j,k]<-sum(B[j,1:dim(B)[2],k])
j<-j+1
}
k<-k+1
}
#Calculation of column totals
coltot<-matrix(nrow=dim(B)[2],ncol=dim(B)[3])
k<-1
while(k<=dim(B)[3])
{
j<-1
while(j<=dim(B)[2])
{
coltot[j,k]<-sum(B[1:dim(B)[1],j,k])
j<-j+1
}
k<-k+1
}
#Calculation of row proportions
phrow<-matrix(nrow=dim(rowtot)[1],ncol=dim(rowtot)
[2])
```

```
j<-1
while(j<=dim(rowtot)[2])
{
i<-1
while(i<=dim(rowtot)[1])
{
phrow[i,j]<-rowtot[i,j]/sum(t(t(rowtot[,j])))
i<-i+1
}
j<-j+1
}
#Calculation of column proportions
phcol<-matrix(nrow=dim(coltot)[1],ncol=dim(coltot)
[2])
j<-1
while(j<=dim(coltot)[2])
{
i<-1
while(i<=dim(coltot)[1])
{
phcol[i,j]<-coltot[i,j]/sum(t(t(coltot[,j])))
i<-i+1
}
j<-j+1
}
#Derivation of Mh vectors
m<-matrix(nrow=dim(phcol)[1]*dim(phrow)
[1],ncol=dim(phcol)[2])
j<-1
while(j<=dim(B)[3])
{
m[,j]<-
sum(t(t(rowtot[,j])))*kronecker(t(t(t(phrow[,j]))),t(t(t(ph
col[,j]))))
    j<-j+1
}
#Calculation of Xh-Mh
T<-xh-m
 #Derivation of Ic.Ir,Jc and Jr matrices
Ir_1<-matrix(0,nrow=dim(B)[1]-1,ncol=dim(B)[1]-1)
diag(Ir_1)=1
Ic_1<-matrix(0,nrow=dim(B)[2]-1,ncol=dim(B)[2]-1)
diag(Ic_1)=1
Jr_1<-matrix(-1, 1,dim(B)[1]-1)
Jc_1<-matrix(-1,1,dim(B)[2]-1)
#Manupulation of Rh and Ch matrices
Rh<-matrix(nrow= nrow(Ir_1)+nrow(Jr_1),ncol=
ncol(Ir_1))
Rh[1:ncol(Ir_1),]=Ir_1
Rh[ncol(Ir_1)+1,]=Jr_1
Ch<-matrix(nrow= nrow(Ic_1)+nrow(Jc_1), ncol=
ncol(Ic_1))
Ch[1:ncol(Ic_1),]=Ic_1
Ch[ncol(Ic_1)+1,]=Jc_1
#Manupulation of matrix Bh
Bh<-kronecker(t(Rh),t(Ch))
 #Derivation of Gh vectors
Gh<-matrix(0,nrow=dim(Bh)[1],ncol=dim(B)[3])
K<-matrix(0,nrow=dim(Bh)[1],ncol=1)
P<-matrix(0,nrow=dim(Bh)[1],ncol=dim(B)[3])
j<-1
while(j<=dim(Gh)[2])
{
P[,j]<-Bh%*%t(t(T[,j]))
K<-K+P[,j]
Gh[,j]<-K
j<-j+1
}
#Manupulation of vector G
G<-t(t(Gh[,dim(B)[3]]))
#uncomment above command in order to print vector G
#print(G)
#Derivation of Lambda vectors
L<-array(0,c(dim(phrow)[1],dim(phrow)[1],dim(phrow)
[2]))
for(i in 1:dim(phrow)[1])
{
for(j in 1:dim(phrow)[1])
{
for(k in 1:dim(phrow)[2])
{
if(i==j)
L[i,j,k]<-1-phrow[i,k]
else
L[i,j,k]<--phrow[i,k]
}
}
}
#Derivation of matrix Ic
Ic<-matrix(0,nrow=dim(B)[2],ncol=dim(B)[2])
diag(Ic)=1

#Derivation of matrix Vp
Vp<-matrix(0,nrow=(dim(B)[1]-1)*(dim(B)[2]-
1),ncol=(dim(B)[1]-1)*(dim(B)[2]-1))

for(k in 1:dim(B)[3])
{
Vph<-matrix(0,nrow=dim(B)[1]*dim(B)[2],
ncol=dim(B)[1]*dim(B)[2])
for(i in 1:dim(B)[1])
{
Vpi<-matrix(0,nrow=dim(B)[1]*dim(B)[2],
ncol=dim(B)[1]*dim(B)[2])
D1<-matrix(0,nrow=dim(B)[2],ncol=dim(B)[2])
for(j in 1:dim(B)[2])
{
D<-B[i,j,k]*((t(t(Ic[,j]))-phcol[,k])%*%t(t(t(Ic[,j]))-
```

```
phcol[,k]))/(1-1/sum(rowtot[,k]))
D1<-D1+D
}
A<-kronecker(t(t(L[,i,k])),Ic)
Vpi<-A%*%D1%*%t(A)
Vph<-Vph+Vpi
}
V<-Bh%*%Vph%*%t(Bh)
Vp<-Vp+V
}
# uncomment below command in order to print Vp
#print(Vp)
#uncomment below command in order toprint Vp
inverse
#print(solve(Vp))

#Calculation of test statistic Tp
Tp<-t(G)%*%solve(Vp)%*%G
#print value of Tp, degrees of freedom and p-value from
Chi-Squared distribution
#print(Tp)
#print(qr(Bh)$rank)
pp=cbind(Tp,qr(Bh)$rank)
colnames(pp)=c("Tp","DF")
rownames(pp)=c("")
print(pp)
p<-pchisq(Tp, qr(Bh)$rank, ncp=0, lower.tail = F, log.p
= FALSE)
colnames(p)=c("P Value")
rownames(p)=c("")
print(p)
#In addition to the above statistic, the standard CMH
statistic is also printed by the following #code
print(mantelhaen.test(Y,U,C))
detach(Data)}
```