J.Natn.Sci.Foundation Sri Lanka 2008 **36** (2):125-135

RESEARCH ARTICLE

A novel method for testing goodness of fit of a proportional odds model : an application to an AIDS study

W. W. M. Abeysekera and Roshini Sooriyarachchi*
Department of Statistics, Faculty of Science, University of Colombo, Colombo 03.

Revised: 29 June 2007 ; Accepted: 04 September 2007

Abstract: Ordinal categorical responses occur commonly in real world situations and many authors discuss the advantages of this type of response. Generalized logit models are popular for analyzing ordinal categorical responses. Of these models, the proportional odds model is the simplest to interpret. However, Lipsitz *et al*. illustrate that the goodness of fit statistics provided by standard statistical packages for this model may not be reliable in justifying the fit of the model. There is no freely available software for computing and analyzing residuals or expected counts for these models.

In their paper, Lipsitz *et al*. propose several goodness of fit statistics and residual analysis that are suitable for ordinal response regression models. However, the new methods are applied to a small artificial set of data. In this paper, the methods of Lipsitz *et al*. are examined, programmes developed in SAS and S-plus softwares and the methods applied to a large scale real-life data set on HIV/AIDS/STD. A proportional odds model was fitted to this data and goodness of fit and residual analysis were carried out using the methods of Lipsitz *et al*.

The methods examined suggest that the goodness of the fitted model is satisfactory. According to the methodology, the expected counts, residuals and approximated (standardized) residuals were calculated and the overall goodness of fit of our model and the reliability of the chi-square approximation of the goodness of fit statistics were confirmed.

Keywords: AIDS study, goodness of fit, ordinal categorical data, proportional odds model, residual analysis

INTRODUCTION

Standard methods of goodness of fit statistics such as likelihood ratio deviance and Pearson chi-square statistics, which compare the fit of the model with the saturated model, are available for most of the ordinal regression models for categorical responses. These measures can be directly obtained from most of the established statistical packages. However under the null hypothesis of a well fitted model, the distribution of the Goodness of fit statistics are even approximately chi-square distributed only if most expected counts formed by the cross classification of the response levels and all covariates are greater than 5. If many of these expected counts (more than 20%) are less than 5, then the usual Likelihood ratio deviance or Pearson chi-square statistics may not be appropriate for testing the goodness of fit of the model[1].

For a continuation ratio model[2], which is a kind of model that can be fitted to an ordinal categorical response with more than two categories, the goodness of fit statistics and residual analysis available for binary logistic regression such as the Hosmer-Lemeshow statistic[3] can be applied since the model actually is a combination of two or more independent binary logistic models.

The proportional odds model[4], is an alternative for modeling an ordinal response with more than two categories. It differs from the continuation ratio model as the proportional odds model has the additional feature of proportionality of odds. This feature while simplifying interpretation leads to the combination of binary logit models which are not independent to each other. This results in complications in obtaining residuals from available statistical packages.

Lipsitz *et al.* [5] suggest a novel approach for testing the goodness of fit of the proportional odds models and for computing its residuals and expected counts. This alternative approach is basically an extension of the method proposed by Hosmer and Lemeshow[3] in testing goodness of fit of logistic models for ordinal

*Corresponding author

binary responses. This method is based on the notion of partitioning the subjects into groups or regions. The goodness of fit statistic is calculated as a quadratic form in the observed minus expected responses in these regions or partitions. This concept of partitioning the covariate space into regions was actually initiated by Tsiatis [6], but he was unable to provide a particular method for how the partitioning should be done. Later, Hosmer and Lemeshow[3] proposed the partitioning of subjects into regions on the basis of the percentiles of the predicted probabilities from the fitted ordinary (binary) logistic regression model. However, Hosmer and Lemeshow did not extend their methodology to suit ordinal categorical responses with more than 2 categories (polytomous). Lipsitz *et al.*[5] extended the method of Hosmer and Lemeshow for models with polytomous categorical responses which are of ordinal scale, in a way that the covariate space can be partitioned into a suitable number of regions on the basis of the percentiles of the predicted mean scores (a formulation of predicted probabilities) of the fitted model. They have also showed that the residuals (observed minus expected counts) can be computed for the cross classification of the response levels and these regions.

Lipsitz *et al.*[5] have illustrated their novel methodologies of goodness of fit and residual analysis for ordinal categorical responses (polytomous) by applying the methods to a small data set of 25 observations and 4 covariates. Other than that, an application or a practice of these novel methods could not be found in the literature of analyzing ordinal categorical data. However, in practice usually data sets are large with many covariates, sometime resulting in a large percentage of cells with expectations less than 5. Thus, it is essential to test this method on a real life data set.

This paper is written with the objective of exemplifying the methodologies of Lipsitz *et al.* [5] using a large data set. This large data set is from the HIV/AIDS/STD Control Programme[7] in Sri Lanka which is a baseline survey to assess the knowledge of HIV/AIDS of plantation workers in the up-country estates of Sri Lanka. The response variable–'Knowledge of HIV/AIDS' – has a natural ordering with 3 levels, 1-Good, 2-Fair and 3-Poor. A proportional odds model was fitted to this ordinal categorical response along with the explanatory variables which represent the demographical characteristics and sexual behaviors of the respondents, and goodness of fit and residual analysis were carried out using the method of Lipsitz *et al.* [5].

Further, as Lipsitz *et al.*[5] have not provided any software programmes for the analysis, this paper sets out to develop software programmes in SAS, and S-Plus as a secondary objective.

Under the methods and materials, the underlying theories and methodologies of the novel methods of goodness of fit tests and residual analysis proposed by Lipsitz *et al.*[5] are briefly explained. An example, which is a complete illustration of the methods proposed, is given in the results. All the results and conclusions drawn within the entire structure of this paper are comprehensively presented under the discussion and conclusion. Finally, the Annex provides some important SAS and S-plus codes that were developed for the statistical analysis of this study.

METHODS AND MATERIALS

Proportional odds model[4]: Models for categorical variables can easily be extended from logistic models which handle only two outcomes[1,2] to cope with polytomous response variables.

Suppose there is a R-category ordinal response variable with 2 explanatory variables A and B having I and J categories respectively.

Then the proportional odds model is given by,

$$\log\left(\frac{Q_{ijr}}{1-Q_{ijr}}\right) = \alpha_r + \beta_i^A + \beta_j^B \ ;$$

$i = 1,....,I \quad j = 1,....., j \quad r = 1,....,(R-1)$ ..... (1)

This model assumes that the effect of the explanatory variables A and B on the odds of response below category *r* (i.e. L.H.S. model) is the same for all *r*.

where,

P_{ijr} = Pr [response r for an observation with explanatory variables (i, j)]

$$Q_{ijr} = \sum_{l=1}^{r} P_{lij}$$

The log odds of response below category *r* for a person with levels *(i,j)* for A and B respectively is given by equation (1). The log odds of response below category *r* for a person with levels *(i',j')* for A and B respectively is

$$\log\left(\frac{Q_{i'j'r}}{1-Q_{i'j'r}}\right) = \alpha_r + \beta_{i'}^A + \beta_{j'}^B \quad .....(2)$$

The log odds ratio of response below category *r* for a person with levels (i, j) to a person with levels (i', j') for A and B respectively is

(1) – (2) ;

$$\log\left(\frac{Q_{ijr}\left(1-Q_{i'j'r}\right)}{Q_{i'j'r}\left(1-Q_{ijr}\right)}\right)=\left(\beta_i^{\,A}-\beta_{i'}^{\,A}\right)+\left(\beta_j^{\,B}-\beta_{j'}^{\,B}\right) \quad \text{..... (3)}$$

This shows that the log of the odds ratio (i.e. the L.H.S.) is constant over all values of r =1,…,(R-1). This property is called the proportional odds and thus this model is called the proportional odds model.

Testing proportionality of odds – score test [8]: The score test is used to test the validity of the proportional odds assumption where it tests the null hypothesis "H_0: the effect of the explanatory variables A and B on the odds of response below category r is the same for all r".

Suppose the parameter vector (ϕ) of the proportional odds model (1) is,

$$\phi=\left[\alpha_1,\alpha_2,...,\alpha_{R-1},\beta_1',\beta_2',,...\beta_{R-1}',\right]'$$

where $\beta_r'=\left[\beta_{ir}^{A},\beta_{jr}^{B}\right]$

Then, the multivariate analogue of the quasi-score function[9] for model (1) is,

$$U=\sum_{t=1}^{n}\left(\frac{\partial e_t}{\partial\phi}\right)'V_t^{-1}(Y_t-e_t)$$

where,

t = 1,2,….,n (number of individuals in the sample), Y_t is the ordinal response (with R levels) for the t th individual, which is a (R-1)x1 vector such that $Y_t=\left[Y_{t1},Y_{t2},...,Y_{t(R-1)}\right]'$, where Y_{tr} = *1* when the response of the t th individual is r, and 0 otherwise,

e_t is the mean of Y_t { i.e. $E(Y_t)=e_t=\left[e_{t1},e_{t2},...,e_{t(R-1)}\right]'$ }

and, V_t is referred to as the working covariance matrix of Y_t vector, which contains (R-1) multinomial variables.

U is asymptotically multivariate normally distributed, and hence a score-like statistic (S) to test the null hypothesis of proportional odds model (that is, H_0 : $\beta_1=\beta_2=...=\beta_{R-1}$) can be derived as follows,

$$S=U'(\hat{\phi}_0)W^{-1}(\hat{\phi}_0)U(\hat{\phi}_0)$$

where $\hat{\phi}_0$ notation indicates that U and W are evaluated under the null hypothesis, that is, at $\beta_r=\hat{\beta}$ and $\alpha=\hat{\alpha}_r$ for $r=1,...,(R-1)$.

The asymptotic distribution of S is chi-square with (R-2)P degrees of freedom where P is the number of parameters of the explanatory variables in the model. If the p-value corresponding to this test statistic is greater than 0.05 then the fitted model satisfies the proportional odds assumption. If it is less than 0.05, then the fitted model does not satisfy the proportional odds assumption.

Ordinary methods in testing Goodness of fit[1,2]: For ordinal regression models with categorical predictors, the usual Likelihood ratio deviance and Pearson chi-square statistics which measure the fit of the given model versus the saturated model can be used. If the P-value given by the test is greater than 0.05, then the model fits the data at 5% significance level.

However, as mentioned in the introduction, for the chi-square distribution to be a good approximation to the distribution of the Goodness of fit test statistics, most expected counts (formed by the cross classification of the response levels and all covariates) should be greater than 5. If many of those expected counts are less than 5, then the score test statistic used for testing the proportional odds assumption is suggested as an appropriate test statistic for testing the Goodness of fit of the proportional odds models[4].

Alternative approach in testing Goodness of fit [5]: When most expected counts are not greater than 5, there are some alternative approaches other than the score test statistic that can be used to assess the Goodness of fit. The Hosmer-Lemeshow test statistic[3] is one of the alternative approaches for testing the fit of ordinal logistic regression models with binary responses. As an extension of the case of binary, an alternative approach on testing ordinal polytomous response variables have been proposed by Lipsitz *et al* [5].

To form the goodness of fit statistic used in this alternative approach, firstly a score s_r is assigned to response category r. The assigned scores may in some instances be the actual numerical response or the midpoint of the interval when the response is a crude grouping of an underlying continuous variable. When the response has no underlying numerical scale, such as a response with 3 levels – poor, moderate, good – often an integer score is used such as, 1=poor, 2=moderate, 3=good.

Then a *fitted score* or a *predicted mean score* can be defined as,

$$\hat{\mu}_t=\sum_{l=1}^{r}s_l\hat{p}_{tl}\ ;\ t=1,2,\ldots..,n \quad \text{..... (4)}$$

where n is the number of subjects and $\hat{p}_{t1}, \hat{p}_{t2}, ..., \hat{p}_{tr}$ are the predicted probabilities for the t^{th} subject for the r response levels.

Then to form the Goodness of fit statistic, the subjects should be partitioned or grouped into G regions based on the percentiles of the predicted mean scores $\hat{\mu}_t$. As a general rule, the value of G should be decided such that $6 \leq G < n/5r$. In practice, any G that satisfies the inequality can be used; for the Hosmer-Lemeshow statistic[3] for binary responses, $G = 10$ has become popular. The G regions can be partitioned such that the first group contains subjects with smallest predicted mean scores and the last group contains subjects with largest predicted mean scores.

Given the partition of the data, the goodness of fit statistic is formulated by defining G–1 group indicators,

$$I_{tg} \begin{cases} 1 & ; \text{if } \hat{\mu}_t \text{ is in region g} \\ 0 & ; \text{if otherwise} \end{cases}$$

where g = 1, 2, …,G – 1.

Suppose the fitted proportional odds model is,

$$\log it(Q_{tr}) = \alpha_r + X_t'\beta_t \quad \text{..... (5)}$$

where X_t is the design matrix and β_t is the vector of parameters (coefficients) for the t^{th} subject.

Then to assess the Goodness of fit of model (5) the following alternative model can be constructed.

$$\log it(Q_{tr}) = \alpha_r + X_t'\beta_t + \sum_{g=1}^{G-1} I_{tg}\gamma_g \quad \text{..... (6)}$$

where γ_g is the coefficient corresponding to the indicator variable I_g.

If model (5) is correctly specified, $\gamma_1 = \gamma_2 = ... = \gamma_{G-1} = 0$ in equation (6), regardless of how the regions are chosen. (Regardless of what scores are used). To assess the goodness of fit of model (5), the Likelihood ratio, Wald or the Score test statistic can be used in testing $H_0 : \gamma_1 = \gamma_2 = ... = \gamma_{G-1} = 0$. If model (5) has been correctly specified, each of these statistics has an approximate chi-square distribution with *G-1* degrees of freedom when the sample size (n) is large.

Expected counts and residuals [5]: It has already been mentioned in this paper that the Goodness of fit test statistics follows the assumed chi-square distribution only if most of the expected counts are greater than 5 (at least 80%). This indicates the necessity of studying the behaviour of expected counts in order to assess the appropriateness of Goodness of fit statistics of a fitted model. The residuals at the same time should be studied carefully so that if any inadequacy is found, it is easy to locate where the error is present.

The conventional method of analyzing the residuals of ordinal categorical data is computing residuals for each level of response in each region specified in the previous section.

- The residual for the r^{th} level of response in g^{th} region is,

$$O_{gr} - E_{gr}$$

where O_{gr} and E_{gr} are observed and expected counts for the r^{th} level in g^{th} region. The estimated expected counts (number of subjects) with response r in region g is given by,

$$E_{gr} = \sum_{t=1}^{n} I_{tg}\hat{p}_{tr}$$

where $\hat{p}_{tr}$ is the predicted probability for the t^{th} subject with the r^{th} response level.

- The approximation to the estimated variance of residual for the r^{th} level of response in g^{th} region is,

$$\hat{Var}(O_{gr} - E_{gr}) \approx n_g \bar{p}_{gr}(1 - \bar{p}_{gr})$$

Where n_g is the number of subjects in region g and $\bar{p}_{gr}$ is the mean predicted probability of r^{th} level of response in g^{th} region.

$$\text{i.e. } \bar{p}_{gr} = \frac{\sum_{t=1}^{n} I_{tg}\hat{p}_{tr}}{n_g} = \frac{E_{gr}}{n_g} = \bar{E}_{gr}$$

- The approximated (standardized) residual for the r^{th} level in g^{th} region is,

$$R_{gr} = \frac{O_{gr} - E_{gr}}{\sqrt{n_g \bar{E}_{gr}(1 - \bar{E}_{gr})}} \quad \text{..... (7)}$$

The expression (7) for computing approximated residuals is occupied only when the predicted probabilities $\hat{p}_{tr}$ are fairly similar in each group and hence $p_{tr} \approx \bar{p}_{gr}$. Under the null hypothesis, when n_g is large, R_{gr} is approximately normal with mean 0 and variance slightly less than 1. That is $n_g \bar{E}_{gr}(1 - \bar{E}_{gr})$ tends to over estimate the variance of $O_{gr} - E_{gr}$. Thus, the expression (7) can be adjusted as follows to formulate residuals which are more closely approximates the standard normal distribution under the null hypothesis.

$$R^*_{gr} = \frac{O_{gr} - E_{gr}}{\hat{\sigma}\sqrt{n_g \bar{E}_{gr}(1-\bar{E}_{gr})}} = R_{gr}/\hat{\sigma} \quad \text{..... (8)}$$

where $\hat{\sigma} = \sqrt{\frac{\sum_{g=1}^{G}\sum_{r=1}^{R} R_{gr}^2}{GR}}$, which can be considered as an estimated common standard deviation of GxR number of R_{gr}'s. As a rough rule of thumb, if more than 5% of the R_{gr} s (or R^*_{gr}s) is not within the band -2 to +2, then special attention should be paid on the profile of the covariates, response and predicted response for each of the subject within those regions.

RESULTS

The HIV/AIDS/STD related baseline survey among plantation workers in Sri Lanka-2005[7] is one of the recent surveys conducted by the *HIV/AIDS/STD* Control Programme. In this study, the collected data of the said survey was taken into consideration as an example based on a large sample (with 594 plantation workers responding to the questions based on their knowledge of HIV/AIDS and STDs, social and demographical background and sexual behaviours), to illustrate how the new methods that were comprehensively presented in this paper can be applied to a large scale problem.

The study of this data was mainly to determine the factors that have significant influence on the knowledge of HIV/AIDS of plantation workers. In accomplishing this requirement, ordinal categorical response - knowledge of HIV/AIDS - which consists of three levels – 'poor', 'fair' and 'good' – and important explanatory variables (filtered from a large set of variables, by using univariate analysis) – gender (*sex*), age group (*age*), education level (*edu*), ethnic group (*ethn*), frequency of condom use (*frqc*), level of knowledge of STDs (*kstd*), level of exposure to mass media (*comm*) and level of sexual practices (*prac*) of the respondent – are incorporated into the modeling procedure.

According to the main aspect of this study, a proportional odds model for our response, which is a polytomous categorical variable in ordinal nature, is fitted in order to carry out the Goodness of fit and residual analysis. The forward selection procedure[1] is used in the SAS statistical software to carry out the model selection.

Finally, the Goodness of fit and residuals of the selected model is evaluated by incorporating the predicted values of the model using several programmes written in both SAS and S-plus softwares (Annex). The software programmes were written to incorporate the methodologies described in this section which were actually proposed by Lipsitz *et al.*[5].

The proportional odds model selected by the forward selection procedure[1] is,

$$\log it(Q_{rijklmnpq}) = \alpha_r + \beta_i^{sex} + \beta_j^{age} + \beta_k^{edu} + \beta_l^{ethn} + \beta_m^{frqc} + \beta_n^{kstd} + \beta_p^{comm} + \beta_q^{prac} \quad \text{... (9)}$$
; r =1,2

where i = 1 (male), 2 (female)
j = 1 (18 to 29), 2 (30 to 39), 3 (40 to 49)
k = 1 (no education), 2 (primary), 3 (secondary or higher)
l = 1 (tamil), 2 (other)
m = 1 (no), 2 (low), 3 (moderate), 4 (high)
n = 1 (poor), 2 (fair), 3 (good)
p = 1 (low), 2 (medium), 3 (high)
q = 1 (low), 2 (medium), 3 (high)

The variables in the selected model have been defined earlier in this section.

Testing the proportionality of odds

The model (9) should satisfy the assumption of proportional odds to be fully accepted as a *'proportional odds'* model. If this model fails the assumption, then a continuation ratio model instead of a proportional odds model should be fitted.

The score test statistic for proportional odds assumption is the option in achieving the objective mentioned above. The test is designed to test the hypothesis,

H_0 : The proportional odds assumption is valid

vs.

H_1 : The proportional odds assumption is not valid

The results of the score test given by the "SAS" proc logistic procedure is given in Table 1.

Since the p-value of the score test is 0. 127 (>> 0.05), H_0 cannot be rejected at 5% significant level. Thus, the model (9) satisfies the proportional odds assumption and hence it is not necessary to go for a continuation ratio model.

Testing the Goodness of fit

The usual Likelihood ratio deviance and Pearson chi-square statistics which measure the fit of the given model

versus the saturated model were studied before moving to alternative techniques for assessing the Goodness of fit. The Hypothesis is;

H_0 : The model fits well to the data
vs.
H_1 : The model does not fit well to the data

The package - *MINITAB* provided the Pearson and Deviance test results for model (9). These are given in Table 2.

The p-values of the two tests are greater than 0.05 indicating that the model (9) satisfies the Goodness of fit at 5% significant level.

However, as it is mentioned in the previous sections, for the chi-square distribution to be a good approximation to the distribution of the Goodness of fit test statistics, most expected counts (formed by the cross classification of the response levels and all covariates) should be greater than 5. From the expected counts given by model (9), about 99.7% percent was less than 5 (out of the 11,664 total cells formed by the cross classification of the response levels and all covariates, 11,633 cells contains expected counts less than 5). Thus, the score test statistic for the proportional odds assumption is suggested in place of Pearson and Deviance statistics for testing the Goodness of fit of the model. According to the score test results given in the previous section, it can be concluded that the model (9) fits well to the data.

The alternative approach [5] that can be used to asses the Goodness of fit is also adopted in parallel to the above mentioned Pearson and Deviance tests and is briefly described in the following section.

According to the method, the first step is to partition the 594 subjects (sample of this study) into 10 regions according to the predicted mean score such that each group consists of approximately 59 subjects. The number of regions 10 is not fixed, but is the most popular one which is proposed by Hosmer and Lemeshow[3].

The predicted mean score in this case is,

$$\hat{\mu}_t = \hat{p}_{t1} + 2\hat{p}_{t2} + 3\hat{p}_{t3}$$

where $\hat{p}_{t1}, \hat{p}_{t2}$ and $\hat{p}_{t3}$ are the predicted probabilities for the three response levels estimated by the model (9) for each subject, and t = 1,2,…..,594.

The 10 regions are such that the first group contains subjects with smallest predicted mean scores and the last group contains subjects with largest predicted mean scores.

Given the partition of the data, the goodness of fit statistic is formulated by defining 9 group indicators,

Table 1: Results of the Score test for Model (9)

Goodness of fit test	Chi-square value	Degrees of freedom	p value
Score	21.3316	15	0.1266

Table 2: Results of Pearson and Likelihood ratio deviance tests for Model (9) (Goodness of fit tests)

Goodness of fit test	Chi-square value	Degrees of freedom	p value
Pearson	538.615	509	0.176
Deviance	474.313	509	0.863

Table 3: Results of testing significance of the Alternative Model (10) against Model (9)

Test	Models	Test statistic	Difference in statistic	Difference in d.f.	p value
Likelihood ratio	(9)	114.7653			
	(10)	105.1835	9.5818	9	0.385388
Wald	(9)	106.6943			
	(10)	95.9209	10.7734	9	0.291562
Score	(9)	96.0341			
	(10)	89.9757	6.0584	9	0.734059

$$I_{tg} \begin{cases} 1 & ; \text{if } \hat{\mu}_t \text{ is in region g} \\ 0 & ; \text{if otherwise} \end{cases}$$

where $g = 1, 2, \ldots, 9$.

Then to assess the goodness of fit of model (9) the following alternative model is constructed.

$$\log it(Q_{rijklmnpq}) = \alpha_r + \beta_i^{sex} + \beta_j^{age} + \beta_k^{edu} + \beta_l^{ethn} + \beta_m^{frqc} + \beta_n^{kstd} + \beta_p^{comm} + \beta_q^{prac} + \sum_{g=1}^{9} I_g \gamma_g \quad \ldots\ldots\ldots (10)$$

If model (9) is correctly specified, then $H_0 : \gamma_1 = \gamma_2 = \ldots\ldots = \gamma_9 = 0$ is not rejected.

The model (10) was fitted using "SAS" proc logistic procedure and the testing of H_0 is carried out using the likelihood ratio, Wald and the score test statistic on models (9) and (10).

By looking at the p-values of the three test results illustrated in Table 3, it can be concluded that model (9) is preferable to the alternative model (10). The hypothesis $H_0 : \gamma_1 = \gamma_2 = \ldots\ldots = \gamma_9 = 0$ is not rejected at 5% significant level and hence, it can be concluded that model (9) fits well.

Expected counts and residuals

Based on the methods explained under the methods and materials, the expected counts and residuals of the fitted model that were calculated and analyzed are presented below. The observed counts, estimated expected counts and the residuals (including the approximated residuals) for the 10 regions are displayed in Table 4.

It is seen in Table 4 that almost all the expected cell counts are significantly greater than 1 and there are six expected cell counts that are lower than 5 (out of 30) indicating that there is 20% of expected cell counts which are less than 5. Thus, using the principles that all estimated expected cell counts should be greater than 1, and at least 80% should be greater than 5, it can be concluded that our chi-square approximations used in the Goodness of fit tests are not misleading. The Goodness of fit tests based on the methods of Lipsitz *et al* [5] practiced in this paper can be accepted and it can be concluded that our model fits well to the data.

To study residuals more clearly, these were plotted as shown in Figure 1. It is clearly demonstrated in Figure 1 that all the residuals are within the band of -2 and +2. At the same time, the plots did not show any systematic pattern or any outliers. Therefore, the residuals also indicated that the fitted proportional odds model is satisfactory.

Finally, based on the results of this Goodness of fit tests and residual analysis evaluated in this example, it can be concluded that the fitted proportional odds

Table 4: Results of Residual Analysis for Model (9)

Region		Residuals for the response levels		
		1	2	3
1	Observed	12	37	10
	Expected	15.285	36.004	7.711
	Residual	-3.285	0.996	2.289
	Approximated residual	-0.976	0.266	0.884
2	Observed	6	35	18
	Expected	8.915	37.136	12.949
	Residual	-2.915	-2.136	5.051
	Approximated residual	-1.060	-0.576	1.589
3	Observed	8	38	14
	Expected	7.086	36.694	16.219
	Residual	0.914	1.306	-2.219
	Approximated residual	0.365	0.346	-0.645
4	Observed	10	35	14
	Expected	5.585	34.451	18.965
	Residual	4.415	0.549	-4.965
	Approximated residual	1.964	0.145	-1.384
5	Observed	2	31	27
	Expected	4.394 +	32.483	23.124
	Residual	-2.394	-1.483	3.876
	Approximated residual	-1.186	-0.384	1.028
6	Observed	4	30	25
	Expected	3.648 +	30.042	25.310
	Residual	0.352	-0.042	-0.310
	Approximated residual	0.190	-0.011	-0.081
7	Observed	4	27	28
	Expected	2.686 +	26.232	30.082
	Residual	1.314	0.768	-2.082
	Approximated residual	0.820	0.201	-0.542
8	Observed	4	21	35
	Expected	2.173 +	23.664	34.163
	Residual	1.827	-2.664	0.837
	Approximated residual	1.262	-0.704	0.218
9	Observed	2	22	35
	Expected	1.563 +	19.318	38.119
	Residual	0.437	2.682	-3.119
	Approximated residual	0.354	0.744	-0.849
10	Observed	2	10	48
	Expected	0.811 +	11.955	47.234
	Residual	1.189	-1.955	0.766
	Approximated residual	1.328	-0.632	0.242

+ Expected count less than 5

model is adequate and hence further interpretation of the model was carried out. Table 5 illustrates the odds ratios computed for the fitted model.

The odds ratio analysis of the fitted model shows that males are more knowledgeable than females whereas younger age groups show higher knowledge compared to the elders. Simultaneously, the ethnic group –Tamils have a poor knowledge campared to other ethnic groups. The plantation workers with higher sexual practices showed no higher knowledge about HIV/AIDS which actually exposes the vulnerability of the plantation workers. Education and exposure to mass media show positive influence in increasing the knowledge of HIV/AIDS.

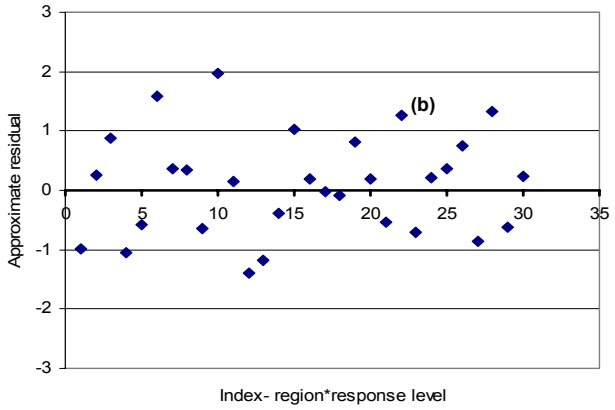

Figure 1: Plot of Standardized Residuals of the Model (10)

DISCUSSION AND CONCLUSION

Reliable techniques in testing Goodness of fit are essential in testing the Goodness of any type of model. The lack of having reliable Goodness of fit tests and residual analysis techniques for categorical models with ordinal categorical predictors (especially for proportional odds models) is discussed throughout this paper. Identifying this problem, Lipsitz *et al.*[5] have suggested several new methods which are more reliable in testing Goodness of fit, and also techniques in analyzing residuals of proportional odds models.

This paper was written with the objective of discussing the new methods proposed by Lipsitz *et al.*[5] in the light of a large scale example illustrating the application of those proposed methodologies. This example is based on a survey and data was collected from 594 plantation workers in the upcountry estates of Sri Lanka to collect information of their level of knowledge on HIV/AIDS along with their social and sexual behaviours. A secondary objective was to develop statistical software programmes for this new methodology.

Table 5: Odds Ratios computed for Model (9)

Effect	Description	Point estimate $\hat{\varphi}$	95% Wald confidence interval		$H_0 : \varphi = 1$ $H_1 : \varphi \neq 1$
			Lower	Upper	
SEX	male vs female	1.76	1.23	2.50	reject H0
AGE	18-29 vs 40-49	1.65	1.01	2.68	reject H0
	30-39 vs 40-49	0.95	0.58	1.55	do not reject H0
EDU	no edu vs secondary or higher	0.50	0.28	0.89	reject H0
	primary vs secondary or higher	0.69	0.46	1.03	do not reject H0
ETHN	tamil vs other	0.56	0.38	0.82	reject H0
FRQC	no vs high	2.02	0.91	4.50	do not reject H0
	low vs high	1.03	0.39	2.68	do not reject H0
	moderate vs high	2.61	1.93	7.38	reject H0
KSTD	poor vs good	0.10	0.03	0.32	reject H0
	fair vs good	0.28	0.09	0.84	reject H0
COMM	low vs high	0.21	0.07	0.64	reject H0
	medium vs high	0.94	0.63	1.40	do not reject H0
PRAC	low vs high	3.15	0.98	10.09	do not reject H0
	medium vs high	4.23	1.29	13.87	reject H0

Note: φ corresponding to the odds ratio and $\hat{\varphi}$ to its maximum likelihood estimate. A 5% level of significance is used in testing $H_0 : \varphi = 1$ versus $H_1 : \varphi \neq 1$.

According to the main aspect of this study, firstly, a proportional odds model was fitted to the collected data taking the ordinal categorical variable – knowledge of HIV/AIDS – which is of 3 categories (1-Good, 2-Fair and 3-Poor) as the response. The model selection was done by the usual forward selection method and at the end of the selection procedure, an appropriate main effects model was chosen as the best.

The next step was to test the Goodness of fit of the model and simultaneously carry out the residual analysis on the fitted model. First of all, the assumption of proportional odds was tested using the score test statistic which is provided by the SAS proc logistic procedure, and concluded that the assumption was not violated and hence the model fitted well.

Prior to examining new methods, the classical Goodness of fit measures - Likelihood ratio deviance and Pearson chi-square statistics – were measured for the fitted model and concluded that the model fitted well. But, this conclusion was made under the caution that the conclusion is accepted only if the chi-square approximation for these test statistics is valid.

Then, the new methods of Goodness of fit statistics which are actually more reliable were applied to the fitted model. As it was proposed [5], the score test for proportional odds model is one alternative test that can replace the Likelihood ratio deviance and Pearson chi-square statistics. As the score test did not reject the proportional odds assumption, this is one indication that the model fits well to the data. According to the alternative approach suggested by Lipsitz *et al.*[5], the Likelihood ratio, Wald and Score test statistics indicated that the Goodness of fit of the model is satisfactory.

Finally the residuals and expected counts were explored. In this study, the programmes were developed in SAS and S-plus softwares to compute the expected counts and residuals (also approximated residuals) for the cross classification of the three response levels and the ten regions. Among the calculated thirty (3 response levels x 10 regions) expected counts, it was found that almost all the expected cell count were greater than 1 and only 20% of the expected cell counts were significantly lower than 5, implying that the test statistics can be reliably assumed to follow the chi-square distribution, in deciding the Goodness of fit of the model.

The calculated approximate residuals which approximately follow the Standard Normal distribution [*N(0,1)]* were plotted in order to observe these clearly. Since the distribution of the approximated residuals is *N(0,1)*, the 95% confidence band of these residuals is approximately -2 to +2. From the plot, it was seen that all the residuals fell within this band and hence supported the conclusion that the overall Goodness of fit of the chosen model is satisfactory.

The selected model indicated that the knowledge on HIV/AIDS among females, elderly people, Tamils and those with high sexual practices in the plantation sector were poor, and that these groups should be targeted for interventions. The study identified exposure to mass media and technical education as successful interventions.

Finally, with the support of this example, it was possible to discuss comprehensively, the novel methods proposed by Lipsitz *et al.*[5] in model validation which consists of testing Goodness of fit and residual analysis. Mode of illustration and interpretation of these novel methods were also clearly stated in this paper.

References

1. Agresti A. (1984). *Analysis of Ordinal Categorical Data*. Wiley, New York.
2. Fienberg S.E. (1980). *The Analysis of Cross-Classified Categorical data*, second edition. MIT Press, Cambridge.
3. Hosmer D.W. & Lemeshow S. (1980). *Applied Logistic Regression.* Wiley, New York.
4. McCullagh P. (1980). Regression models for ordinal data. *Journal of the Royal Statistics Society*, Series B. **42**:109-142.
5. Lipsitz S.R., Fitzmaurice G.M. & Molenberghs G. (1996). Goodness-of-Fit tests for ordinal response regression models. *Applied Statistics* **45**(2):175-190.
6. Tsiatis A.A. (1980). A note on the goodness of fit for the logistic regression model. *Biometrika* **67**:250-251.
7. National STD/AIDS Control programme in Sri Lanka (2005). *HIV/AIDS/STD related baseline survey among plantation workers in Sri Lanka.*
8. Thomas R.S., Huiman X.B. & John M.W. (1999). Testing proportionality in the proportional odds model fitted with GEE. *Statistics in Medicine* **18**:1419-1433.
9. Wedderburn R.W.M. (1974). Quasi-likelihood functions, generalized linear models and the Gaussian method. *Biometrika* **61**: 439-444.

Annex

SAS CODES:

SAS codes used in fitting the model and the diagnostics analysis

```
/* Fitting the best proportional odds model [Model-(M)] selected from the Forward selection procedure */

proc logistic data=rowdata;

class SEX AGE EDU EMPL ETHN MARR PRAC FRQC KSTD COMM K;
model K=SEX AGE EDU ETHN FRQC KSTD COMM PRAC /scale=none;

output out= pred p=phat lower=lcl upper=ucl
            predprob=(individual crossvalidate)
            resdev=r2 H=h xbeta=lp;
run;

/* Formulating predicted mean scores and Indicator variables to separate the subjects into 10 regions */

data goodnes;
/*import data*/
set rowdata;
/*import the predicted probabilities of the fitted model*/
set pred;
/*defining the 9 indicator variables*/
I1=0; I2=0; I3=0; I4=0; I5=0; I6=0; I7=0; I8=0; I9=0;

g=594/10;

/*calculation of predicted mean score (m) from predicted probabilities*/
m=(1*ip_1)+(2*ip_2)+(3*ip_3);

proc sort data=goodnes;
    by m;
run;

data good;
set goodnes;
i=_n_;
/*Grouping the 594 subjects into 10 regions by using the 9 indicator variables*/
if i<=g then I1=1;
else if i<=2*g then I2=1;
else if i<=3*g then I3=1;
else if i<=4*g then I4=1;
else if i<=5*g then I5=1;
else if i<=6*g then I6=1;
else if i<=7*g then I7=1;
else if i<=8*g then I8=1;
else if i<=9*g then I9=1;

/*Fitting the alternative model [Model-(M*)]*/
proc logistic data=good;
    class SEX AGE EDU EMPL ETHN MARR PRAC FRQC KSTD COMM K I1-I9;
    model K=SEX AGE EDU ETHN FRQC KSTD COMM PRAC I1-I9;
run;

/*defining a single indicator variable that separates the subjects into 10 regions*/
if i<=g then I=1;
else if i<=2*g then I=2;
else if i<=3*g then I=3;
else if i<=4*g then I=4;
else if i<=5*g then I=5;
else if i<=6*g then I=6;
else if i<=7*g then I=7;
else if i<=8*g then I=8;
else if i<=9*g then I=9;
else if i<=10*g then I=10;

proc sort data=good;
    by I K;

run;

/*printing the response level (K), predicted probabilities, indicator that define 10 regions (I) and predicted mean score (m)*/proc print data = good;

    var K ip_1 ip_2 ip_3 I m;/*to go for EXCEL*/

run;
```

S-PLUS CODES:

The above printed output were imported to the S-plus programme given below,

```
#Read the saved SAS output saved as a text file

data<-read.table("C:\\paperHIV\\diagnostics\\main3pred.txt", header=T)
attach(data)

i[595]_11
r_1
c_1
x_1
y_1
exp_0
obs_0
grpn_0
```

```
while(r<11){
      p1_0
      p2_0
      p3_0
      gn1_0
      gn2_0
      gn3_0
      while[i(x)==r]{
            p1_p1+ip1(x)
            p2_p2+ip2(x)
            p3_p3+ip3(x)
            if [K(x)==1] gn1_gn1+1
                  else if [K(x)==2] gn2_gn2+1
                      else if [K(x)==3] gn3_gn3+1
            x_x+1
      }
      exp(c)_p1
      obs(c)_gn1
      grpn(c)_gn1+gn2+gn3
      c_c+1
      exp(c)_p2
      obs(c)_gn2
      grpn(c)_gn1+gn2+gn3
      c_c+1
      exp(c)_p3
      obs(c)_gn3
      grpn(c)_gn1+gn2+gn3
      c_c+1
r_r+1
}
```

```
res_0
ebar_0
vari_0
apres_0
j_1
while(j<31){

      res(j)_obs(j)-exp(j)
      ebar(j)_exp(j)/grpn(j)
      vari(j)_{grpn(j)*ebar(j)*[1-ebar(j)]}
      apres(j)_res(j)/sqrt[vari(j)]
      j_j+1
}

# Print the 'Observed counts', 'Expected counts', 'residuals'
and 'approximated residuals'
obs
exp
res
apres

#Plotting the approximated residuals
g_c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,
23,24,25,26,27,28,29,30)

xyplot(apres~g, xlab="10 (groups) * 3 (response levels)=30
groups", ylab="approximated residual")
```