國 立 交 通 大 學

統計學研究所

碩 士 論 文

稀少和非稀少的潛在類別迴歸模型之適合度檢定

Goodness-of-fit Test for Sparse and Unsparse
Latent Class Regression Models

研 究 生 ：鄭俊凱
指導教授 ：黃冠華 博士

中 華 民 國 九 十 五 年 六 月

稀少和非稀少的潛在類別迴歸模型之適合度檢定

Goodness-of-fit Test for Sparse and Unsparse Latent Class Regression Models

研 究 生：鄭俊凱　　Student ： Chun-Kai Cheng
指導教授：黃冠華　　Advisor ： Dr. Guan-Hua Huang

國 立 交 通 大 學
統計學研究所
碩 士 論 文

A Thesis
Submitted to Institute of Statistics
College of Science
National Chiao Tung University
in Partial Fulfillment of the Requirements
for the Degree of
Master
in
Statistics
June 2006

Hsinchu, Taiwan, Republic of China
中華民國九十五年六月

稀少和非稀少的潛在類別迴歸模型之適合度檢定

研究生：鄭俊凱　　　　指導教授：黃冠華　博士

國立交通大學統計學研究所

摘要

潛在類別迴歸(latent class regression) 模型被廣泛利用在先前的許多文獻裡，這種模型能將多重指標的共同特徵整合成基本的類別變數。這篇論文中我們將提出一個潛在類別迴歸模型的適合度檢定，此檢定的基礎是由所有可能回答的選項以及相伴變數分群所組成的列聯表，這個概念是由 Hosmer 與 Lemeshow 在邏輯斯迴歸中所提出來的。而當列聯表有稀少情形發生時，我們將用一階和二階邊際來取代並且修正檢定統計量。我們在不同的條件下作模擬，來測試所提出的適合度檢定表現。

Goodness-of-fit Test for Sparse and Unsparse Latent Class Regression Models

Student：Chun-Kai Cheng   Advisor：Dr. Guan-Hua Huang

Institute of Statistic
National Chiao Tung University

ABSTRACT

Latent class regression (LCR) models have been utilized previously in many literatures. Such models can summarize shared features of the multiple indicators as an underlying categorical variable. In this paper, we propose a goodness-of-fit for the LCR model. The basis of the proposed test is a contingency table, which groups the population through all possible response patterns and concomitant covariates. The idea is from Hosmer-Lemeshow statistic for the multiple logistic regression model. When the contingency table is sparse, we replace it with the first- and second-order marginals and modify the test statistic. A simulation study is carried out to examine the behavior of the proposed goodness-of-fit test under different situations.

誌謝

在交大統計研究所學習的這兩年，要謝謝所上的各位教授細心教導，也要感謝自己的指導教授黃冠華老師，不厭其煩的指導來協助我解決遭遇到的許多難題，讓論文能夠順利的完成。另外也要感謝黃文瀚老師、徐南蓉老師以及陳鄰安老師，在口試時給予我建議，並指正我論文裡需要修改的地方。

我也要感謝研究所同學明曄、泓毅、志強、秀慧…等，以及我的女朋友和我的家人，在各方面給我的幫助和鼓勵，讓我能快樂、順利完成這兩年學業。最後將此論文獻給我的家人、朋友、統計所同學及學弟妹。

鄭俊凱　　謹誌于

國立交通大學統計學研究所

中華民國九十五年六月二十八號

Contents

List of Tables

1 Introduction

In recent years, questions of psychosocial and medical research investigate the relationship between multiple categorical outcome variables and continuous predictor variables. These relationships may be unobserved, hence valid surrogates are necessary. Latent class regression (LCR) models (Huang and Bandeen-Roche 2004) are useful tools for assessing association of measured indicators. The LCR model allow both the distribution of the underlying class variable and the within-class distributions of measured indicators to be functionally related to individual-level independent variables. Hence, LCR model may mitigate errors in measurement and can give well-summarized inference between multiple indicators and covariates of interest. However, we do not observe the true class membership of individuals. So we should carefully do model checking.

When no covariates, the population can be grouped by all possible response patterns. Pearson χ^2 test and log likelihood ratio test statistic (LRT) (Doodman 1974, Bartholomew 1987, Formann 1992) can be applied for evaluating overall model fit. However, when there are continuous covariates, Pearson χ^2 test is invalid because the degree of freedom increases when sample size increases.

In this paper, we apply the idea of the Hosmer-Lemeshow statistic (Hosmer and Lemeshow 1980) to our LCR model. We extend the outcome variable into not only binary but category and each individual has multiple outcome variables. Therefore, an adequate chi-square test statistic can be used to assess to our LCR model. Sometimes, when response patterns are large and sample size is moderate or small, some cells of the contingency table formed by the all response patterns will be sparse. In this situation, the chi-square test is also not valid. When sparseness occurs, informal remedies such as combining cells are often recommended. Here, we substitute the first-order and second-order marginal frequencies (Reiser and Lin 1999) for the original contingency table, and then we modify the chi-square test statistic which is mentioned above.

In section 2, we review four parts: 1.The LCR model and some assumptions which complete the model; 2.The goodness-of-fit of the multiple logistic regression model; 3.Theorem5.1 in Moore and Spruill (1975) and its required regularity conditions, which is applied to prove the asymptotic distribution of the proposed goodness-of-fit test; 4.The approach of second-order marginal frequencies. In section 3, we propose the goodness-of-fit of our LCR model and propose another test statistic when sparseness occurs. Section 4 presents the results of a simulation study and power of the test statistic. Some discussions and recommendation are presented in section 5.

2 Literature Review

2.1 Latent class regression model

To describe the latent class regression (LCR) model (Hung and Bandeen-Roche 2004), let $Y_i = (Y_{i1},\ldots,Y_{iM})^T$ represent the $M\times 1$ response vector for the ith individual in a study sample of N persons. Y_{im} can take value $\{1,\ldots,K_m\}$, where $K_m \geq 2$, $m = 1,\ldots,M$. And let $(\mathbf{x}_i, \mathbf{z}_i)$ be the concomitant covariates of the ith person, where $\mathbf{x}_i = (1, x_{i1},\ldots,x_{ip})^T$ are primary predictors for latent class membership S_i, S_i can take values $\{1,\ldots,J\}$, and $\mathbf{z}_i = (z_{i1},\ldots,z_{iM})$ with $\mathbf{z}_{im} = (z_{im1},\ldots,z_{imL})^T$, $m = 1,\ldots,M$, are secondary covariates used for $Pr(Y_{im} = k|S_i = j)$. These covariates may include any combination of continuous and discrete measures, and they may be mutually exclusive or overlapped. Then the LCR model can be represented as

$$Pr(Y_{i1} = y_1,\ldots,Y_{iM} = y_m|\mathbf{x}_i,\mathbf{z}_i) = \sum_{j=1}^{J}\{\eta_j(\mathbf{x}_i)\prod_{m=1}^{M}\prod_{k=1}^{K_m}[p_{mkj}(\mathbf{z}_{im})]^{y_{mk}}\}. \quad (1)$$

with $\eta_j(\mathbf{x}_i)$ and $p_{mkj}(\mathbf{z}_{im})$ as in the generalized linear framework (McCullagh and Nelder 1989). Here, $y_{mk} = I(y_m = k) = 1$ if $y_m = k$; otherwise. Various link functions could be chosen like probit, ordinal, or etc. We specifically propose to use the generalized logit link function (Agresti 1984) :

$$log\left[\frac{\eta_j(\mathbf{x}_i)}{\eta_J(\mathbf{x}_i)}\right] = \beta_{0j} + \beta_{1j}x_{i1} + \ldots + \beta_{pj}x_{ip} = \mathbf{x}_i^T\boldsymbol{\beta}_j, \quad (2)$$

and

$$log\left[\frac{p_{mkj'}(\mathbf{z}_{im})}{p_{mK_mj'}(\mathbf{z}_{im})}\right] = \gamma_{mkj'} + \alpha_{1mk}z_{im1} + \ldots + \alpha_{Lmk}z_{imL} = \gamma_{mkj'} + \mathbf{z}_{im}^T\boldsymbol{\alpha}_{mk}, \quad (3)$$

for $i = 1,\ldots,N; m = 1,\ldots,M; k = 1,\ldots,K_m - 1; j = 1,\ldots,J-1; j' = 1,\ldots,J$.

Parameters, γ_{mkj}, α_{mk} and β_j can be estimated by Expectation-Maximization (EM) algorithm (Dempster, Laird and Rubin 1977). EM algorithm is an iterative approach which is usually for computing maximum likelihood when model includes missing data.

Adding following three assumptions can complete the model (1) :

1. Latent class membership probabilities are associated with only :

$$Pr(S_i = j|\mathbf{x}_i, \mathbf{z}_i) = Pr(S_i = j|\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta}_j)}{1 + \sum_{l=1}^{J-1} \exp(\mathbf{x}_i^T \boldsymbol{\beta}_l)}, j = 1, \ldots, J-1$$

2. Conditioning on class membership, measured responses are associated with $\mathbf{z}_i$:

$$Pr(Y_{i1} = y_1, \ldots, Y_{iM} = y_m|S_i, \mathbf{x}_i, \mathbf{z}_i,) = Pr(Y_{i1} = y_1, \ldots, Y_{iM} = y_m|S_i, \mathbf{z}_i)$$

$$with \quad Pr(Y_{iM} = k|S_i = j', \mathbf{z}_i) = \frac{\exp(\gamma_{mkj'} + \mathbf{z}_{im}^T \boldsymbol{\alpha}_{mk})}{1 + \sum_{s=1}^{K_m-1} \exp(\gamma_{msj'} + \mathbf{z}_{im}^T \boldsymbol{\alpha}_{ms})},$$

$$for \quad m = 1, \ldots, M; k = 1, \ldots, K_m - 1; j' = 1, \ldots, J.$$

3. The multiple measurements are conditionally independent given class membership and $\mathbf{z}_i$:

$$Pr(Y_{i1} = y_1, \ldots, Y_{iM} = y_m|S_i, \mathbf{z}_i) = \prod_{m=1}^{M} Pr(Y_{im} = y_m|S_i, \mathbf{z}_i)$$

2.2 Goodness-of-fit test for logistic regression

Hosmer and Lemeshow (1980) proposed a goodness-of-fit test, which determines the adequacy of the fitted multiple logistic regression model. The logistic regression model will be stated as follows :

Let Y_i= 0 or 1 be outcome variables and $\mathbf{x}_i^T = (x_{i1},\ldots,x_{ip})$ be the independent variables. Let $\pi(x_i) = Pr(Y_i = 1|x_i) = \exp(\beta_0 + \boldsymbol{\beta}^T x_i)/(1 + \exp(\beta_0 + \boldsymbol{\beta}^T x_i))$ where $\boldsymbol{\beta}^T = (\beta_1, \ldots, \beta_p)$. Under these assumptions, the likelihood function is :

$$L(\mathbf{y};\mathbf{x},\beta_0,\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i} \ , \ where\ \pi_i = \pi(\mathbf{x}_i),\ for\ i = 1,\ldots,n.$$

From log of $L(\mathbf{y};\mathbf{x},\beta_0,\boldsymbol{\beta})$, the maximum likelihood estimates $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}$ by solving $(p+1)$ likelihood equations. The basis of Hosmer-Lemeshow statistic builds on a 2×g contingency table. To obtain the table, let $\hat{\pi}_i = \pi(x_i)|_{(\beta_0,\boldsymbol{\beta})=(\hat{\beta}_0,\hat{\boldsymbol{\beta}})}$ and define a random variable W where $w_i = j$ if $c_{j-1} \leq \hat{\pi}_i < c_j$, for $j = 1,\ldots,g$; $i = 1,\ldots,n$. The $c_j{}'s$ are known constants such that $0 = c_0 < c_1 < \ldots < c_{g-1} < c_g = 1$ .Denote the counts in the cell of table as n_{kj} where n_{kj} is the frequency of occurrence of the pair $(y_i = k, w_i = j)$ in the sample, $k = 0, 1$ and $j = 1,\ldots,g$. Notationally the observed frequencies may tabulated as Table 1.

A choice of forming $c_0,\ldots,c_g$ in the 2×g contingency table is to make the distribution of W to be uniform. That is, the cut points $c_0,\ldots,c_g$ depend on the data and hence are no longer fixed constants. So there will be n/g value of in each interval. Let's define $\hat{\pi}_{(1)} \leq \hat{\pi}_{(2)} \leq \ldots \leq \hat{\pi}_{(n)}$ as the ordered values of $\hat{\pi}$ and let $\hat{c}_j = \hat{\pi}_{[jn/g]}$, where $\left[\frac{jn}{g}\right]$ represents the largest integer less than or equal to $\frac{jn}{g}$, $j = 0,\ldots,g$. Let $\hat{w}_i = j$ if $\hat{c}_{j-1} \leq \hat{\pi}_i < \hat{c}_j$. Define $\hat{n}_{kj}$ as the observed frequency of the pair $(y_i = k, \hat{w}_i = j)$ in the sample. If $\hat{J}_j = \{i : \hat{c}_{j-1} \leq \hat{\pi}_i < \hat{c}_j\}$ then the test statistic is

$$C_g = \sum_{j=1}^{g}\{\frac{(\hat{n}_{1j} - \sum_{r\in\hat{J}_j}\hat{\pi}_r)^2}{\sum_{r\in\hat{J}_j}\hat{\pi}_r} + \frac{\left[\hat{n}_{0j} - \sum_{r\in\hat{J}_j}(1-\hat{\pi}_r)\right]^2}{\sum_{r\in\hat{J}_j}(1-\hat{\pi}_r)}\} \qquad (4)$$

There are two problems for the application of the usual theory used for chi-square goodness-of-fit test to the distribution of C_g.

1. Parameter estimates are determined using likelihood functions for "un-grouped" data.

2. The frequencies, $\hat{n}_{kj}$ in the $2 \times g$ contingency table depend on the estimated parameters, namely the cells are random not fixed.

Chernoff and Lehmann (1954) first mention a chi-square test under problem 1 and then Watson (1959). Moore(1971) and Moore and Spruill (1975) considered the distribution of the chi-square goodness of fit statistic under both problems 1 and 2. They extended Watson's results to the case of random rectangular cells. Drust (1979) generalized these results to include random cells other than rectangles. By results of Moor and Spruill (1975) and Drust (1979), the asymptotic distribution of C_g can be obtained as follows.

Theorem Under distributional assumptions, the distribution of C_g will be asymptotically ($N \to \infty$)

$$\chi^2(2g - g - (p+1)) + \sum_{i=1}^{p+1} \lambda_i \chi_i^2(1)$$

where $0 < \lambda_i \leq 1$, $i = 1, \ldots, (p+1)$, and $\lambda_i's$ are eigenvalues of some matrix. The detailed statement of the matrix can see Theorem 1 in 3.1.

2.3 General chi-squared statistic for individual likelihood and random cells

The proof of the above theorem follows from verifying that the regularity conditions necessary for the proof of theorem 4.2, lemma 5.1 and theorem

5.1 in Moore and Spruill (1975) are satisfied.

Before describing these results, the notations are defined as follows. Let $F(\mathbf{y}|\theta,\eta)$ be the cdf of $\{\mathbf{Y}_1,\ldots,\mathbf{Y}_n\}$. The parameter θ ranges over an open set Ω_1 in R^m, while η ranges over a neighborhood of a point η_0 in R^p. The cells for the following χ^2 tests are rectangles in R^k. They are functions of a variable φ defined on Ω_2 in R^r. The resulting cells are denoted by $I_\sigma(\varphi)$. Here, the null hypothesis (H_0) is that $\mathbf{Y}_i$ have a cdf $F(\mathbf{y}|\theta,\eta_0)$. We will explore the large-sample behavior of tests for the null hypothesis under the sequences of parameter values (θ_0,η_n) where $\theta_0 \in \Omega_1$ and $\eta_n = \eta_0 + n^{-1/2}\gamma$ for fixed γ in R^p. H_0 is the special case $\gamma = 0$. θ is estimated by $\theta_n = \theta_n(\mathbf{Y}_1,\ldots,\mathbf{Y}_n)$. The cells are chosen by $\varphi_n = \varphi_n(\mathbf{Y}_1,\ldots,\mathbf{Y}_n)$. We will assume that under (θ_0,η_n), $\varphi_n - \varphi_0 = o_p(1)$ for some φ_0 and $\theta_n - \theta_0 = o_p(1)$. We will suppress arguments θ, φ, η whenever they take the values θ_0, φ_0, η_0 respectively.

The number of $\mathbf{Y}_1,\ldots,\mathbf{Y}_N$ falling in the cell $I_\sigma(\varphi)$ will be denoted by $N_{n\sigma}(\varphi)$. The cell probabilities are denoted by $p_\sigma(\theta,\eta,\varphi)$ where $\sigma = 1,2,\ldots,p\times g$ and $p_\sigma(\theta,\eta,\varphi) = \int_{I_\sigma(\varphi)} dF(\mathbf{x}|\theta,\eta)$.

The regularity conditions for the following theorem are as :

A1. Under (θ_0,η_n), $\theta_n - \theta_0 = O_p(n^{-1/2})$ and $\varphi_n - \varphi_0 = o_p(1)$. Every vertex $x(\varphi)$ of every cell $I_\sigma(\varphi)$ is a continuous R^k-valued function of φ in a neighborhood of φ_0.

A2. For each σ, $p_\sigma(\theta,\eta,\varphi)$ is continuous in (θ,η,φ) and continuously differentiable in (θ,η) in a neighborhood of $(\theta_0,\eta_0,\varphi_0)$. Moreover, $\sum_1^M p_\sigma = 1$ and $p_\sigma > 0$ for each σ.

A3. $F(x) = F(x|\theta_0, \eta_0)$ is continuous at every vertex $x(\varphi_0)$ of every cell $I_\sigma(\varphi_0)$. As $n \to \infty$, $sup_x|F(x|\eta_n) - F(x)| \to 0$.

A4. $K(\theta, \varphi) = S(\theta, \varphi)S(\theta, \varphi)^T$ for an $M \times M$ matrix $S(\theta, \varphi)$ with entries continuous in (θ, φ) at (θ_0, φ_0).

A5. Under (θ_0, η_N)

$$n^{1/2}(\theta_n - \theta_0) = n^{-1/2} \sum_{i=1}^{N} h(\mathbf{Y}_i, \eta_n) + A_\gamma + o_p(1)$$

for some $m \times p$ matrix A and measurable function $h(x, \eta)$ from $R^k \times R^p$ to R^m satisfying

$$E\left[h(\mathbf{Y}, \eta_n)|(\theta_0, \eta_n)\right] = 0$$

$$E\left[h(\mathbf{Y}, \eta_n)h(\mathbf{Y}, \eta_n)^T|(\theta_0, \eta_n)\right] = L(\eta_n)$$

where $L(\eta_n)$ is a $m \times m$ matrix converging to the finite and matrix $L = E\left[h(\mathbf{Y})h(\mathbf{Y})^T\right]$ as $n \to \infty$

Theorem 4.2 in Moore and Spruill (1975)

Let $V_n(\theta_n, \varphi_n)$ be a $M \times 1$ vector with σth component

$$v_{n\sigma}(\theta_n, \varphi_n) = \frac{N_{n\sigma}(\varphi_n) - np_\sigma(\theta_n, \eta_0, \varphi_n)}{\left[np_\sigma(\theta_n, \eta_0, \varphi_n)\right]^{1/2}}$$

Define also,

$q^T = (p_1^{\frac{1}{2}}, \ldots, p_M^{\frac{1}{2}})$

B is a $M \times m$ matrix and has (i, j)th entry $p_i^{-1/2} \frac{\partial p_i}{\partial \theta_j}$.

$J = E\left[(\frac{\partial \log f}{\partial \theta})(\frac{\partial \log f}{\partial \theta})^T\right]$

$\sum = I_M - qq^T + BLB^T - BE\left[h(Y)W(Y)^T\right] - E\left[W(Y)h(Y)^T\right]B^T$

$\sum_0 = S^T \sum S$

If A1,…, A5 hold, $V_n^T(\theta_n, \varphi_n)k(\theta_n, \varphi_n)V_n(\theta_n, \varphi_n)$ has limiting distribution

$$\sum_{j=1}^{M} \lambda_j \chi_{1j}^2 \qquad under \quad (\theta_0, \eta_0)$$

Where $\lambda_j' s$ are eigenvalues of $\sum_0$

One more regularity condition is needed for the following lemma :

C1. $m \leq M$ and the matrix with entries $\partial p_i / \partial \theta_j$ has rank m.

Lemma 5.1 in Moore and Spruill (1975)

When C1 regularity condition holds

1. $\left[P^T q q^T P e\right]_j = 0 \quad j = 1, \ldots, M-1$
   $\left[P^T q q^T P e\right]_j = 1 \quad j = M$

2. $\left[P^T C P e\right]_j = 0 \quad j = 1, \ldots, M-m-1, M$
   $\left[P^T C P e\right]_j = 1 \quad j = M-m, \ldots, M-1$

3. $\left[P^T B J^{-1} B^T P e\right]_j = 0 \quad j = 1, \ldots, M-m-1, M$
   $\left[P^T B J^{-1} B^T P e\right]_j = 1 - \lambda_j \quad j = M-m, \ldots, M-1$

where P is an orthogonal matrix which simultaneously diagonalizes qq^T, C and $BJ^{-1}B^T$. $C = B(B^T B)^{-1} B^T$.

More regularity conditions are needed for the following theorem :

C2. $\log f(x|\theta, \eta)$ is differentiable with respect to (θ, η) at (θ_0, η_0). The matrix J is pd and J_{12} is finite. $(\partial/\partial\theta)F(x|\theta)$ may be evaluated by differentiating $f(x|\theta)$ under the integral sign for all x and $\theta = \theta_0$.

C3. $n^{1/2}(\hat{\theta}_n - \theta_0) = n^{-1/2}\sum_{i=1}^{n} J^{-1}\frac{\partial \log f(Y_i|\eta_n)}{\partial\theta} + J^{-1}J_{12}\gamma + o_p(1)$. Here J is the information matrix for $F(x|\theta)$ at θ_0.

$$J = E\left[(\frac{\partial \log f}{\partial\theta})(\frac{\partial \log f}{\partial\theta})^T\right],$$

J_{12} is the $m \times p$ matrix

$$J_{12} = E\left[(\frac{\partial \log f}{\partial\theta})(\frac{\partial \log f}{\partial\eta})^T\right].$$

C4. $J - B^T B$ is pd.

Theorem 5.1 in Moore and Spruill (1975)

When A1,. . ., A5 and C1,. . ., C4 regularity conditions hold, $\| V_n(\hat{\theta}_n, \varphi_n) \|^2$ has limiting distribution

$$\chi^2_{M-m-1} + \sum_{j=M-m}^{M-1} \lambda_j \chi^2_{1j} \quad under \quad (\theta_0, \eta_0)$$

and $\lambda_{M-m}, \ldots, \lambda_{M-1}$ are the m roots of the determinantal equation

$$|B^T B - (1-\lambda)J| = 0$$

2.4 First and second-order marginals

In practice, when response patterns are large and the sample size, n, is moderate or small, some response patterns of $Y_i's$ are usually less than 5 even to 0. This kind of contingency table is said to be *sparse* (Agresti and Yang 1987). However, the chi-square approximation for the test distribution may not be valid. So when sparseness occurs, informal remedies such as combining cells or adding a small constant like 0.5 to each cell are sometimes recommended. One kind of combining method is first order and second-order marginals (Reiser and Lin 1999). The advantage of it is that the frequencies are almost always substantially larger than zero, even with small samples. The combing technique states as follows :

To make the presentation clear, we assume dichotomous response cases. Let Y_i= 0 or 1 be outcome variables, for $i = 1,\dots,k$. The response patter is a k-dimensional vector of zeros and 1's. A set of T response patterns can be generated by varying the index of the kth variable most rapidly, the $k-1$th variable next, etc. Let $\pi_s(\boldsymbol{\beta})$ represent probability of response pattern s and w_{is} represent element i of response pattern s. Under the model, the first order and second-order marginal proportion for variable Y_i and Y_j can be defined as

$$P_i(1|\boldsymbol{\beta}) = P(Y_i = 1|\boldsymbol{\beta}) = \sum_s w_{is}\pi_s(\boldsymbol{\beta}) ,$$

$$P_{ij}(1,1|\boldsymbol{\beta}) = P(Y_i = 1, Y_j = 1|\boldsymbol{\beta}) = \sum_s w_{is}w_{js}\pi_s(\boldsymbol{\beta}) ,$$

The summation across the frequencies associated with the response patterns to obtain the marginal proportions represents a transformation of the frequencies in the multinomial vector $\boldsymbol{\pi}^T = (\pi_1, \pi_2,\dots,\pi_T)$, which can be implemented via multiplication by matrix $\mathbf{H}$ where for $j = 1,\dots,k$; $i = j, j+1,\dots,k$; $s = 1,\dots,T$; and $l = (j-1)k - 0.5j(j-1) + i$, element (l, s) of $\mathbf{H}$ is given by

$$h_{ls} = \begin{cases} 1 & \text{if} \quad w_{is}{=}w_{js}{=}1 \\ 0 & \text{otherwise} \end{cases}$$

Using matrix $\mathbf{H}$

$$P_{ij}(1,1|\boldsymbol{\beta}) = P(Y_i = 1, Y_j = 1|\boldsymbol{\beta}) = \mathbf{h}_l^T\boldsymbol{\pi}(\boldsymbol{\beta}),$$

where $\mathbf{h}_l^T$ is row l of matrix $\mathbf{H}$.

3 Methodologies

3.1 Goodness-of-fit test of LCR model

We imitate the Hosmer-Lemeshow goodness-of-fit to create the test statistic for LCR model. Let the joint probability of the ith individual be

$$Pr(\mathbf{Y}_i = \mathbf{y}_h; \phi) = Pr\{(Y_{i1}, \ldots, Y_{iM}) = (y_{h1}, \ldots, y_{hM}); \phi\} = \pi_{ih}(\phi) \qquad (5)$$

Where $i = 1, \ldots, N; h = 1, \ldots, K^*; K^* = \prod_{m=1}^{M} K_m$ and $\phi = (\gamma_{mj}, \alpha_m, \beta)$ is the vector of parameters. Here $\{\mathbf{y}_1, \ldots, \mathbf{y}_{K^*}\}$ represent the all possible multiple outcomes. The basis of the goodness-of-fit test statistic of our LCR model is a $K^* \times g$ contingency table as shown in Table 2.

In Kuo (2004), she defined a random variable W to form the contingency table, where $W_i = j$ if $c_{j-1} < \hat{\pi}_{i1} < c_j$, for $j = 1, \ldots, g$; $i = 1, \ldots, n$. The $c_j's$ are known constants such that $0 = c_0 < c_1 < \ldots < c_{g-1} < c_g = 1$, and $\hat{\pi}_{i1}$ is the estimated probability of the ith individual at the first response pattern.

We choose another different method to group the population. Here, We apply two partition methods in the R package cluster, clara and fanny, to group the population into g groups depending on the covariates associated with conditional probabilities and latent prevalence. We explain the main difference between the clara method and the fanny method first. In clara method, if we assume that one person belongs to group 3, then the probability of his falling into group 3 would be one. While the probability of his fallying into other groups would be zero. In the same case, if we apply the fanny method, the probabilities of his falling into other groups would be all larger than zero, but smaller than the probability of falling into group 3. The number of groups, g, is constant and it is determined by the highest average silhouette width which calls the silhouette coefficient (SC). SC is defined as the average of the $s(i)$. The detailed statements of $s(i)$ can see Appendix A. Experience has led to the subjective interpretation of the (SC) as listed in Table 3. The $K^* \times g$ contingency table is obtained by defining a random variable W, where $W_i = j$ if ith person fall in the jth group, $j = 1,\ldots,g$. Under the hypothesis of LCR model holds, the goodness-of-fit test statistic will be obtained by comparing "observed" frequencies $O'_{hj}s$ to versus "expected" frequencies $E'_{hj}s$. Hence, we will discuss under three situations of O_{hj} and E_{hj}.

Situation 1 : O_{hj} and E_{hj} from clara method are denoted as follows :
Denote O_{hj} is the observed frequency of occurrence of the pair $(\mathbf{Y} = \mathbf{y}_h, W = j)$ in the sample, where $h = 1,\ldots,K^*$; $K^* = \prod_{m=1}^{M} K_m$; $j = 1,\ldots,g$. The total observed frequencies may show as Table 2. Denote the expected frequency E_{hj} in the hth response pattern and jth group. The expression is obtained as $E_{hj} = \sum_{r \in I_j} \pi_{rh}(\hat{\boldsymbol{\phi}})$, where $I_j = \{i : W_i = j\}$, $j = 1,\ldots,g$, and $\pi_{rh}(\hat{\boldsymbol{\phi}}) = \pi_{rh}(\boldsymbol{\phi})|_{\boldsymbol{\phi}=\hat{\boldsymbol{\phi}}}$.

Situation 2 : O_{hj} and E_{hj} from fanny method are denoted as follows :
The denotation of O_{hj} is the same as situation 1. Denote the expected frequency E_{hj} as the hth group response pattern and jth group. The expression is obtained as $E_{hj} = \sum_{i=1}^{n} \pi_{ih}(\hat{\boldsymbol{\phi}}) \times \rho_{ij}$, where ρ_{ij} is the estimated probability of the ith individual falling into the jth group. $i = 1,\ldots,n$, $j = 1,\ldots,g$.

Situation 3 : Another O_{hj} and E_{hj} obtaining from fanny method are denoted as follows :

Denote $O_{hj} = \sum_{i=1}^{n} I(Y_i = y_h) \times \rho_{ij}$. And the denotation of E_{hj} is the same in situation 2. Notationally set-up of the frequencies in LCR model may tabulated as Table 2.

Then, the statistic is

$$T_1 = \sum_{h=1}^{K^*} \sum_{j=1}^{g} \frac{(O_{hj} - E_{hj})^2}{E_{hj}} \tag{6}$$

The large sample distribution of T_1 is following the following theorem.

Theorem 1 Under LCR assumptions (1), (2), and (3), the distribution of T_1 will be asymptotically ($N \to \infty$)

$$\sum_{i=1}^{K^* g} \lambda_i \chi_{1i}^2$$

where $\lambda_i' s$ are the eigenvalues of the matrix $\sum(T_1) = I - qq^T - BJ^{-1}B^T$; $i = 1, \ldots, K^* \times g$. I is a $K^*g\times$ (number of parameter) identity matrix. q is a $K^*g \times 1$ vector with elements $\sqrt{P_{hj}}$, $h = 1, \ldots, K^*$, $j = 1, \ldots, g$, where $P_{hj} = Pr(\mathbf{Y} = \mathbf{y}_h, W = j) = \frac{1}{N} E_{hj}$. **B** is a $K^*g \times K^*g$ matrix and has a general element given by $\frac{1}{\sqrt{P_{hj}}} \frac{\partial P_{hj}}{\partial \phi_l}$, $\boldsymbol{\phi} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha})$. J^{-1} is the asymptotic variance covariance matrix of estimates $\hat{\boldsymbol{\phi}}$.

Proof The proof of this theorem follows from verifying that the regularity conditions necessary for the proof of theorem 4.2 in Moore and Spruill (1975) are satisfied. For details, see Appendix B.

In order to estimate the asymptotic large sample distribution of T_1 , we must calculate the eigenvalues of the matrix $\sum(T_1)$. Here, we substitute

$\widehat{\sum(T_1)} = I - \hat{q}\hat{q}^T - \hat{B}\hat{J}^{-1}\hat{B}^T$ for $\sum(T_1)$ and calculate the eigenvalues of the matrix $\widehat{\sum(T_1)}$, where $\hat{q}$, $\hat{B}$ and $\hat{J}^{-1}$ are estimators of q, B and J^{-1}. Then, the nominal asymptotic distribution will be $\sum_{i=1}^{K^*g} \hat{\lambda}_i \chi^2_{1i}$ by substituting $\hat{\boldsymbol{\phi}}$ for $\boldsymbol{\phi}$.

Here, we propose another two test statistics.

$$T_2 = V_n^T(I_{K^*g} - qq^T - BJ^{-1}B^T)^{-1}V_n \quad and \quad T_3 = V_n^T(I_{K^*g} - BJ^{-1}B^T)^{-1}V_n.$$

Where $\mathbf{V}_n$ is a $K^*g \times 1$ vector with elements $V_{hj} = \frac{O_{hj}-E_{hj}}{\sqrt{E_{hj}}}$, for $h = 1,\ldots,k^*, j = 1,\ldots,g$.

It is easy to show that the asymptotic distribution of T_2 is χ_{k^*g}. Because qq^T is usually very small, we can ignore it. (Lemma 5.1 (1) of Moore and Spruill(1975)). The asymptotic distribution of T_3 is also χ_{k^*g}.

3.2 First- and second-order marginals of LCR model

when sparseness occurs, we substitute the second-order marginal frequencies for original contingency table. Then, if the LCR model is rejected based on the use of the first- and second-order marginals, it could be concluded that the model does not hold in the joint frequencies either. Notationally set-up of the frequencies of the first- and second-order marginals may tabulated as Table 4. The rows of Table 4 are constituted by the following first- and second-order marginals :

$$\begin{cases} Pr(Y_{ij} = s\,; \boldsymbol{\phi}) & \text{for} \quad j = k \\ Pr(Y_{ij} = s, Y_{ik} = t\,; \boldsymbol{\phi}) & \text{for} \quad j \neq k \end{cases}$$

where $k = 1,\ldots,m$; $j = k,\ldots,m$; $s = 1,\ldots,K_k - 1$; $t = 1,\ldots,K_j - 1$; $\boldsymbol{\phi} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha})$.

The summation across the frequencies associated with the response patterns to obtain the marginal proportions represents a transformation of the frequencies in the multinomial vector $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_{k^*})$, which can be implemented via multiplication by a matrix $\mathbf{H}$.

The new O_{hj} and E_{hj} of three situations are as follows :
By matrix H, We can transform the original observed frequency table into new observed frequency table for each situation. Then, O^*_{hj} is hth row and jth column of new observed frequency table. In the same way, We can transform the original expected frequency table into new expected frequency table for each situation. And E^*_{hj} is hth row and jth column of new expected frequency table.
Hence, the new test statistic is

$$T_1^* = \sum_{h=1}^{K^{***}} \sum_{j=1}^{g} \frac{(O^*_{hj} - E^*_{hj})^2}{E^*_{hj}} \tag{7}$$

where $K^{***} = \sum_{m=1}^{M}(K_m - 1) + \sum_{m=1}^{M-1}\Big[(K_m - 1) \times \sum_{m'=m+1}^{M}(K_{m'} - 1)\Big]$ is the total number of response pattern of the first- and second-order marginals.

Similar to the Theorem 1, we rewrite the theorem as follows :

Theorem 2 Under LCR assumptions (1), (2), and (3), the distribution of T_1^* will be asymptotically $(N \to \infty)$

$$\sum_{i=1}^{K^{***}g} \lambda_i^* \chi^2_{1i}$$

where $\lambda_i^{*'}s$ are the eigenvalues of the matrix $\sum(T_1^*) = Z_N^T H S_N^T \sum(T_1) S_N H^T Z_N^T$; $i = 1, \ldots, K^* \times g$. Here, $\sum(T_1)$ is mentioned in section 3.1. Z_N is a $K^{***}g \times K^{***}g$ diagonal matrix with elements $\frac{1}{\sqrt{E^*_{hj}}}$, for $h = 1, \ldots, k^{***}$, $j = 1, \ldots, g$. S_N is a $K^*g \times K^*g$ diagonal matrix with elements $\sqrt{E_{hj}}$, for

$h = 1, \ldots, k^*$, $j = 1, \ldots, g$. The detailed proof of Theorem 2 can be found in Appendix C.

In order to estimate the large sample distribution of T_1^* , we must calculate the eigenvalues of the matrix $\sum(T_1^*)$. Here, we substitute $\widehat{\sum(T_1^*)} = \hat{Z}_N^T H \hat{S}_N^T \widehat{\sum(T_1)} \hat{S}_N H^T \hat{Z}_N^T$ for $\sum(T_1^*)$ and calculate the eigenvalues of the matrix $\widehat{\sum(T_1^*)}$. Then, the nominal distribution will be $\sum_{i=1}^{K^{***}g} \hat{\lambda}_i^* \chi_{1i}^2$, where $\hat{\lambda}_i^{*\prime} s$ are eigenvalues of $\widehat{\sum(T_1^*)}$.

Under sparse situation, we rewrite test statistic T_2 and T_3 as follow:

$$T_2^* = W_n^* \left[Z_n^T H S_n^T (I_{K^*g} - qq^T - BJ^{-1}B^T)^{-1} S_n H^T Z_n \right]^{-1} W_n^*$$

and

$$T_3^* = W_n^* \left[Z_n^T H S_n^T (I_{K^*g} - BJ^{-1}B^T)^{-1} S_n H^T Z_n \right]^{-1} W_n^*$$

Where $\mathbf{W}_n$ is a $K^{***}g \times 1$ vector with elements $W_{hj} = \frac{O_{hj}^* - E_{hj}^*}{\sqrt{E_{hj}^*}}$, for $h = 1, \ldots, k^{***}$, $j = 1, \ldots, g$. The asymptotic distributions to T_2^* and T_3^* are $\chi_{k^{***}g}$.

4 Simulation Studies

4.1 Generated data from the LCR model

Here, we are going to simulate two major situations to discuss. One is "balance" and the other is "unbalance". "Balance" means the contingency table is not sparse and "unbalance" means contingency table is sparse.

In balanced case, we simulate three-class LCR with five-two level measured indicator, two covariates associated with conditional probabilities, two covariates associated with latent prevalence and sample size is 2500 ($i.e.$, $J = 3, M = 5, K_1 =\ldots= K_5 = 2, P = L = 2, N = 2500$). Then, β_{pj} , which are the model parameters, can be determined randomly by setting $\beta_{pj} = k_1 U_j$, $U_j \sim U(0, 1)$, for each $p \in \{0, 1,\ldots, P\}$; $j = 1,\ldots,(J-1)$. k_1 is constant such that $\sum_{p=1}^{P}\sum_{j=1}^{J-1}\beta_{pj}$ equal the preselected total. Similarly, we can use the same way to determine $\{\gamma_{jmk}, j = 1,\ldots,(J-1)\}$ for all m, k and $\{\alpha_{qmk}, m = 1,\ldots,M;\ k = 1,\ldots,(K_m-1)\}$ for all q. Here, we set the parametric values of $\sum_{l=1}^{L}\sum_{m=1}^{M}\alpha_{lm}$ and $\sum_{i=1}^{m}\sum_{j=1}^{J}\alpha_0$ as 1 and of $\sum_{p=1}^{P}\sum_{j=1}^{J-1}\beta_{pj}$ and $\sum_{j=1}^{J-1}\beta_0$ as 0.6. And observable $Y_i's$ are generated with 100 replications. Table 5 shows the values of α_0 and α_{lm}. Table 6 shows the values of β_0 and β_{pj}.

The covariates associated with conditional probabilities (z_{im1}, z_{im2}), $m = 1,\ldots,5$ and latent prevalences (x_{i1}, x_{i2}) are generated as follows:

For each m

$z_{im1} \sim Bernoulli(0.4), z_{im2} \sim Normal(50, 5)\ i = 1 \sim 500$

$z_{im1} \sim Poisson(20), z_{im2} \sim Gamma(4, 3)\ i = 501 \sim 1000$

$z_{im1} \sim Binomial(14, 0.6), z_{im2} \sim Uniform(1, 10)\ i = 1001 \sim 1500$

$z_{im1} \sim Binomial(6, 0.4), z_{im2} \sim Exponential(6)\ i = 1501 \sim 2000$

$z_{im1} \sim Poisson(3), z_{im2} \sim Unifotm(20, 30)\ i = 2001 \sim 2500$

and covariates associated with latent prevalences are generated as

$x_{i1} \sim Bernoulli(0.6), x_{i2} \sim Normal(0, 1)\ i = 1, \ldots, 2500$

In unbalanced case, we simulate five-class LCR with six-two level measured indicator, two covariates associated with conditional probabilities, two covariates associated with latent prevalence and sample size is 2500 (*i.e.*, $J = 5, M = 6, K_1 = \ldots = K_6 = 2, P = L = 2, N = 2500, g = 5$). Here, we set the parametric values of $\sum_{l=1}^{L}\sum_{m=1}^{M}\alpha_{lm}$ and $\sum_{i=1}^{m}\sum_{j=1}^{J}\alpha_0$ as 1.5 and of $\sum_{p=1}^{P}\sum_{j=1}^{J-1}\beta_{pj}$ and $\sum_{j=1}^{J-1}\beta_0$ as 0.8. Table 7 shows the values of α_0 and α_{lm}. Table 8 shows the values of β_0 and β_{pj}. Then, the covariates associated with conditional probabilities $(z_{im1}, z_{im2}), m = 1, \ldots, 5$ and latent prevalences (x_{i1}, x_{i2}) are generated by the same ways in balanced case. Table 9 is the averaged O's over 100 simulations in the contingency table forming by all response patterns in balanced case and Table 10 is the averaged O's over 100 simulations in unbalanced case. Table 11 is table 10 after combining as first- and second order marginals.

The simulation results are represented from Table 12 to Table 17. According to the results of balanced case, test statistics of fanny are well approximated to nominal distribution. Nevertheless, behaviors of three test statistics of clara are not as good as behaviors of fanny, because the values of clara are obviously lower than nominal distribution.

On the other hand, according to the results of unbalanced case, the values of test statistics of fanny are higher than nominal distribution. While the values of test statistics of clara are lower than nominal distribution.

4.2 Assess power of the proposed test statistics

The simulations considered thus far have demonstrated that the test statistic have well defined distributions under the null hypotheses that the LCR model holds. To examine the power of the proposed test statistics, data were generated the same as section 4.1. Then, we use a simpler model to fit the data which were generated from a complicated model.

The selected sample size is 2500 and $Y_i's$ are generated with 100 replication. In balanced case, we use two-class LCR with five-two level measured indicator, one covariate associated with conditional probabilities, one covariates associated with latent prevalence (*i.e.*, $J = 2, M = 5, K_1 = \ldots = K_5 = 2, P = L = 1$) and divide the population into three groups to fit alternative model. The covariates associated with conditional probabilities z_{im1} ,$m = 1,\ldots,5$ and latent prevalences x_{i1} are generated as follows :

For each m

$z_{im2} \sim Normal(20, 5)$ $i = 1 \sim 800$

$z_{im2} \sim Gamma(4, 2)$ $i = 801 \sim 1600$

$z_{im2} \sim Poisson(15)$ $i = 1601 \sim 2500$

and covariates associated with latent prevalences are generated as

$x_{i2} \sim Normal(0, 1)$ $i = 1,\ldots,2500$

In unbalanced case, we use three-class LCR with six-two level measured indicator, one covariate associated with conditional probabilities, one covariates associated with latent prevalence($i.e.$, $J = 3, M = 6, K_1 = \ldots = K_6 = 2, P = L = 1$) and divide the population into three groups to fit alternative model. The covariates associated with conditional probabilities z_{im1} ,$m = 1, \ldots, 6$ and latent prevalences x_{i1} are generated as the same in balanced case.

Table 18 presents the results of clara method in balanced case. Three test statistics virtually have no power. This method seems to cluster the population unsuitably under the balanced situation. Table 19 and Table 20 present the results of fanny method in balanced case. In Table 19, T_1 and T_3 have higher power in detecting the difference between fitted model and alternative model. While T_2 have comparably lower power. In Table 20, power of T_1 is lower than powers of T_2 and T_3. Table 21, Table 22 and Table 23 present the results of clara and fanny method in unbalanced case. The conclusions are similar to balanced case.

5 Discussion

In this paper, we use the latent class regression model to fit the relationship between a latent class outcome and latent factor predictors. We propose the goodness-of-fit test statistic to assess the adequacy of the model. The number of the group is determined before forming the contingency table. Then, we use two clustering methods, clara and fanny, to cluster the population.

The fanny method is a good approach for our grouping the population of the LCR model. Under fanny method, situation 2 is well than situation 3. So we suggest using method of situation 2. But fanny method is sensitive to covariates which are selected to do the clustering. There is a serious influence on the results of the cluster. Therefore, when we select covariates to do the clustering, we should select carefully to avoid the inappropriate results.

Appendix A: Silhouette coefficient

For each object i, we denote A the cluster to which it belongs, and compute

$$a(i) := \frac{1}{|A|-1} \sum_{j \in A, j \neq i} d(i,j)$$

It is the average dissimilarity of i to all other objects of A.

Here, $d(i,j)$ is defined as

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}} \quad \in [0,1]$$

where

$d_{ij}^{(f)}$ = contribution of variable f to $d(i,j)$, which depends on its type :

1. f binary or nominal : $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$, and $d_{ij}^{(f)} = 1$ otherwise,
2. f interval-scaled : $d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{\max_h x_{hf} - \min_h x_{hf}}$,
3. f ordinal or ratio-scaled : compute ranks r_{if} and $z_{if} = \frac{r_{if}-1}{\max_h r_{hf}-1}$ and treat these z_{if} as interval-scaled,

and

$\delta_{ij}^{(f)}$ = weight of variable f :

1. $\delta_{ij}^{(f)} = 0$ if x_{if} or x_{jf} is missing,
2. $\delta_{ij}^{(f)} = 0$ if $x_{if} = x_{jf} = 0$ and variable f is asymmetric binary,
3. $\delta_{ij}^{(f)} = 1$ otherwise.

and p is number of variables.

Now consider any cluster C different from A and put

$$d(i,C) := \frac{1}{|C|} \sum_{j \in C} d(i,j)$$

It is the average dissimilarity of i to all other objects of C.

After computing $d(i, C)$ for all clusters $C \neq A$ we take the smallest of those:

$$b(i) := \min_{C \neq A} d(i, C).$$

The cluster B which attains this minimum [that is, $d(i, B) = b(i)$] is called the *neighbor* of object i. This is the second-best cluster for object i.

The *silhouette value* $s(i)$ of the object i is defined as

$$s(i) := \frac{b(i) - a(i)}{max\{a(i), b(i)\}}.$$

clearly $s(i)$ always lies between -1 and 1.

Appendix B: Proof of theorem 1

Then regular conditions of *theorem* 4.2 in Moor and Spruill are satisfied as follows :

1. Under (ϕ_N, φ), $\phi_N - \phi_0 = o_K**(1)$ and $\varphi_n = \varphi(\mathbf{x}, \mathbf{z})$. Every vertex $\mathbf{y}(\phi)$ of every cell $I_\sigma(\phi)$ is a continuous R^M-valued function of ϕ in a neighborhood of ϕ_0.

2. For each σ, $P_\sigma(\phi, \varphi)$ is continuous in (ϕ, φ) and continuously differentiable in (ϕ) in a neighborhood of (ϕ_0, φ_0). Moreover, $\sum_{\sigma=1}^{K^{**}g} P_\sigma = 1$ and $P_\sigma > 0$ for each σ.

3. $F(y) = F(\mathbf{y}|\phi_0)$ is continuous at every vertex $\mathbf{y}(\phi_0)$ of every cell $I_\sigma(\phi_0)$. As $N \to \infty$, $sup_y|F(\mathbf{y}|\phi_N) - F(\mathbf{y})| \to 0$.

4. $K(\phi) = S(\phi)S(\phi)^T$ for an $K^{**}g \times K^{**}g$ matrix $S(\phi)$ with entries continuous in ϕ at ϕ_0.

5. Under ϕ_N

$$N^{1/2}(\phi_N - \phi_0) = N^{-1/2} \sum_{i=1}^{N} h(\mathbf{Y}_i, \phi_N) + A_\gamma + o_{K^{**}}(1)$$

for some $g \times K^{**}$ matrix A and measurable function $h(\mathbf{y}, \phi)$ from $R^M \times R^{K^{**}}$ to R^g satisfying

$$E\left[h(\mathbf{Y}, \phi_N)|\phi_N\right] = 0$$

$$E\left[h(\mathbf{Y}, \phi_N)h(\mathbf{Y}, \phi_N)^T|\phi_N\right] = L(\phi_N)$$

where $L(\phi_N)$ is a $g \times g$ matrix converging to the finite and matrix $L = E\left[h(\mathbf{Y})h(\mathbf{Y})^T\right]$ as $N \to \infty$

6. $g \leq K^* g$ and the matrix with entries $\partial p_i / \partial \phi_j$ has rank g.

7. $\log f(\mathbf{y}|\phi)$ is differentiable with respect to ϕ at ϕ_0. The matrix J is pd and J_{12} is finite. $(\partial/\partial\phi)F(\mathbf{y}|\phi)$ may be evaluated by differentiating $f(\mathbf{y}|\phi)$ under the integral sign for all $\mathbf{y}$ and $\phi = \phi_0$.

8. $n^{1/2}(\hat{\phi}_n - \phi_0) = n^{-1/2} \sum_{i=1}^{n} J^{-1} \frac{\partial \log f(Y_i|\eta_n)}{\partial \phi} + J^{-1} J_{12} \gamma + o_p(1)$. Here J is the information matrix for $F(\mathbf{y}|\phi)$ at ϕ_0.

$$J = E\left[(\frac{\partial \log f}{\partial \phi})(\frac{\partial \log f}{\partial \phi})^T \right],$$

J_{12} is the $m \times p$ matrix

$$J_{12} = E\left[(\frac{\partial \log f}{\partial \phi})(\frac{\partial \log f}{\partial \eta})^T \right].$$

9. $J - B^T B$ is pd, where matrix B has (i, j)th entry $p_i^{-1/2} \frac{\partial p_i}{\partial \phi_j}$.

Appendix C: Distributions of test statistic T_1^*, T_2^* and T_3^*

N = total number of individuals

$h = 1, 2, \ldots, K^*$, where $K^* = \prod_{m=1}^{M} k_m$

$j = 1, 2, \ldots, g$, where g = number of groups

$\mathbf{V_N} = K^* g \times 1$ vector

$$\mathbf{V_N} = \begin{pmatrix} v_{11} \\ \vdots \\ v_{1g} \\ \vdots \\ \vdots \\ v_{K^*1} \\ \vdots \\ v_{K^*g} \end{pmatrix}, \quad where \quad v_{hj} = \frac{O_{hj} - E_{hj}}{\sqrt{E_{hj}}}$$

E_{hj} = expected number of observation in (h, j)

$T_N = V_N^T K_N V_N = \| S_N^T V_N \|^2 = \sum_{h=1}^{K^*} \sum_{j=1}^{g} (O_{hj} - E_{hj})^2$

$$where \quad K_N = \begin{pmatrix} E_{11} & & & & & & & 0 \\ & \ddots & & & & & & \\ & & E_{1g} & & & & & \\ & & & \ddots & & & & \\ & & & & \ddots & & & \\ & & & & & E_{K^*1} & & \\ & & 0 & & & & \ddots & \\ & & & & & & & E_{K^*g} \end{pmatrix}$$

and $K_N = S_N S_N^T$

$$So \quad S_N = \begin{pmatrix} \sqrt{E_{11}} & & & & & & \\ & \ddots & & & 0 & & \\ & & \sqrt{E_{1g}} & & & & \\ & & & \ddots & & & \\ & & & & \ddots & & \\ & & & & & \sqrt{E_{K^*1}} & & \\ & & 0 & & & & \ddots & \\ & & & & & & & \sqrt{E_{K^*g}} \end{pmatrix}$$

By theorem 4.2 in Moore & Spruill

$$S_N^T V_N \xrightarrow{d} N(\mu, \Sigma_0) \quad where \quad \Sigma_0 = S_N^T (I_{K^*g} - qq^T - BJ^{-1}B^T) S_N$$

$$\left(\begin{array}{c} \therefore \parallel S_N^T V_N \parallel^2 = V_N^T S_N V_N \xrightarrow{d} \Sigma_{i=1}^{K^*g} \lambda_i \chi_{1i}^2 \\ \text{where } \lambda_i's \text{ are eigenvalues of } \Sigma_0 \end{array} \right)$$

Let $K^{***} = \sum_{m=1}^{M} (K_m - 1) + \sum_{i \neq j = 1}^{M} (K_i - 1)(K_j - 1)$

$\mathbf{W_N} = K^{***} \times 1$ vector

$$\mathbf{W_N} = \begin{pmatrix} w_{11} \\ \vdots \\ w_{1g} \\ \vdots \\ \vdots \\ w_{K^{***}1} \\ \vdots \\ w_{K^{***}g} \end{pmatrix}, \quad where \quad w_{sj} = \frac{O_{sj}^* - E_{sj}^*}{\sqrt{E_{sj}^*}}$$

O^*_{sj} = number of observation in (s,j) after combining.

E^*_{sj} = expected number of observation in (s,j) after combining.

Let $\mathbf{W_N} = H^* S_N^T V_N$

$$\mathbf{H}^*_{K^{***}g \times K^{***}g} = \begin{pmatrix} \mathbf{H} & & 0 \\ & \ddots & \\ 0 & & \mathbf{H} \end{pmatrix},$$

where $\mathbf{H}$ is matrix which is mentioned in second-order.

So

$$\mathbf{W_N} = H^* S_N^T V_N \stackrel{d}{\rightarrow} N(H^*\mu, H^*\Sigma_0 H^{*^T})$$

and

$$H^*\Sigma_0 H^{*^T} = H^* S_N^T (I_{K^*g} - qq^T - BJ^{-1}B^T) S_N H^{*^T}$$

Let $W_N^* = Z_N^T W_N$, where Z_N is a $K^{***}g \times K^{***}g$ matrix.

$$Z_N = \begin{pmatrix} 1/\sqrt{E^*_{11}} & & & & & & & \\ & \ddots & & & & & 0 & \\ & & 1/\sqrt{E^*_{1g}} & & & & & \\ & & & \ddots & & & & \\ & & & & \ddots & & & \\ & & & & & 1/\sqrt{E^*_{K^{***}1}} & & \\ & & 0 & & & & \ddots & \\ & & & & & & & 1/\sqrt{E^*_{K^{***}g}} \end{pmatrix}$$

So $W_N^* = Z_N^T W_N \stackrel{d}{\rightarrow} N(Z_N^T H^*\mu, Z_N^T H^*\Sigma_0 H^{*^T} Z_N)$

$\Rightarrow W_N^{*^T} W_N^* = \sum_{s=1}^{K^{***}} \sum_{j=1}^{g} \frac{(O^*_{sj} - E^*_{sj})^2}{E^*_{sj}} \sim \sum_{i=1}^{K^{***}} \lambda_i^* \chi^2_{1i}$

where $\lambda_i^{*'} s$ is eigenvalues of $\Sigma^* = Z_N^T \Sigma Z_N$

References

[1] Agresti, A. (1984). *Analysis of Catagorical Data*. New York: J.Wiley and Sons.

[2] Agresti,A.,Yang,M.C. (1987).An Empirical Investigation of Some Effects of Sparseness in Contingency Tables. *Computational Statistics and Data Analysis*.5:9-21.

[3] Bandeen-Roche,K.,Miglioretti,D.L.,Zeger, S.L.,Rathouz,P.J. (1997).Latent Variable Regression For Multiple Discrete Outcome. *Journal of the American Statistical Association*. 92: 1375-1386.

[4] Batholomew, D.J. (1987).*Latent Variable Model and Factor Analysis*. London: Charles Griffin & Co. Ltd.

[5] Demster, A.P,Laird, N.M.,Rubin, D.B (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*;39:1-38.

[6] Drust,M.C.,Donsker.(1980).Vapnik-Chervonenkis classes and chi-square tests of fit with random cells; Unpublished doctoral dissertation, Department of Mathematics, M.I.T., Cambridge,MA.

[7] Formann, A.K. (1992).Linear logistic latent class analysis for polytomous data. *Journal of the American Statistical Association*. 87:476-486.

[8] Goodman, L.A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biomertrika.* 61:215-231

[9] Hosmer, D.W.,Lemeshow,s. *Applied Logistic Regression.* New York: John Wiley & Sons.

[10] Huang, G.H.,Bandeen-Roche, L. Latent variable regression with covariate effects on underlying and measured variables: an approach of analyzing multiple polytomous surrogates. Submitted for publication.

[11] Kuo, H.Y. (2004). Goodness-of-fit Test for Latent Class Regression Model. To be submitted.

[12] Lemeshow, S.,Hosmer, D.W. (1982). The Use of Goodness-of-fit Statistics in the Development of Logistic Regression Models.*American journal of Epidemiology.* 115:92-106

[13] McCullagh, P.,Nelder, J.A. (1989). *Generalized Linear Models, 2nd edition.* London: Chapman and Hall.

[14] Moore,D.S.,Spruill,M.C.(1975). Unified Large-sample Theory of General Chi-squared Statistic for Tests of Fit. *Annals of Statistics.* 3:599-616.

[15] Reiser,M.,Lin,Y.(1999). A Goodness-of-Fit Test for the Latent Class Model When Expected Frequencies are Small. *Sociological Methodology.*Vol. 29, pp.81-111

[16] Struyf,A.,Hubert,M.,Rousseeuw,P.J.(1996).Clustering in an Object-Oriented Environment. *Journal of Statistical Software*.l.

Table 1: Notational set-up of the frequencies in logistic regression model

	1	2	$\ldots$	g	Total
y=0	n_{01}	n_{12}	$\ldots$	n_{1g}	n_0
y=1	n_{11}	n_{12}	$\ldots$	n_{2g}	n_1
Total	$n_{.1}$	$n_{.2}$	$\ldots$	$n_{.g}$	n

Table 2: Notational set-up of the frequencies in LCR model

	1	2	$\ldots$	g
$(y_1 = 1, y_2 = 1, \ldots, y_m = 1)$	O_{11}	O_{12}	$\ldots$	O_{1g}
$(y_1 = 1, y_2 = 1, \ldots, y_m = 2)$	O_{21}	O_{22}	$\ldots$	O_{2g}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
$(y_1 = 1, y_2 = 1, \ldots, y_m = k_m)$	O_{m1}	O_{m2}	$\ldots$	O_{mg}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
$(y_1 = k_1, y_2 = k_2, \ldots, y_m = k_m)$	O_{k^*1}	O_{K^*2}	$\ldots$	O_{K^*g}
	n_1	n_2	$\ldots$	n_g

Table 3: Interpretation of the silhouette coefficient for partitioning method

SC	Proposed Interpretation
0.71-1.00	A strong structure has been found.
0.51-0.70	A reasonable structure has been found
0.26-0.50	The structure is weak and could be artificial, try additional method
≤ 0.25	No substantial structure has been found

Table 4: Notational set-up of the frequencies of first- and second-order marginals

	1	2	$\ldots$	g
$(y_1 = 1)$	O_{11}	O_{12}	$\ldots$	O_{1g}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
$(y_1 = k_1 - 1)$	$O_{h_1 1}$	$O_{h_1 2}$	$\ldots$	$O_{h_1 g}$
$(y_2 = 1)$	$O_{h_2 1}$	$O_{h_2 2}$	$\ldots$	$O_{h_2 g}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
$(y_2 = k_2 - 1)$	$O_{h_3 1}$	$O_{h_3 2}$	$\ldots$	$O_{h_3 g}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
$(y_M = 1)$	$O_{h_4 1}$	$O_{h_4 2}$	$\ldots$	$O_{h_4 g}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
$(y_M = k_M - 1)$	$O_{h_5 1}$	$O_{h_5 2}$	$\ldots$	$O_{h_5 g}$
$(y_1 = 1, y_2 = 1)$	$O_{h_6 1}$	$O_{h_6 2}$	$\ldots$	$O_{h_6 g}$
$(y_1 = 1, y_2 = 2)$	$O_{h_7 1}$	$O_{h_7 2}$	$\ldots$	$O_{h_7 g}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
$(y_1 = 1, y_2 = k_2 - 1)$	$O_{h_8 1}$	$O_{h_8 2}$	$\ldots$	$O_{h_8 g}$
$(y_1 = 2, y_2 = 1)$	$O_{h_9 1}$	$O_{h_9 2}$	$\ldots$	$O_{h_9 g}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
$(y_1 = k_1 - 1, y_2 = k_2 - 1)$	$O_{h_{10} 1}$	$O_{h_{10} 2}$	$\ldots$	$O_{h_{10} g}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
$(y_{M-1} = 1, y_M = 1)$	$O_{h_{11} 1}$	$O_{h_{11} 2}$	$\ldots$	$O_{h_{11} g}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
$(y_{M-1} = k_{M-1} - 1, y_M = k_M - 1)$	$O_{k^{***}1}$	$O_{K^{***}2}$	$\ldots$	$O_{K^{***}g}$
	n_1	n_2	$\ldots$	n_g

Note:

$$h_1 = k_1 - 1, \qquad h_2 = (k_1 - 1) + 1, \qquad h_3 = \sum_{i=1}^{2}(k_i - 1)$$

$$h_4 = \left[\sum_{i=1}^{M-1}(k_i - 1)\right] + 1, \qquad h_5 = \sum_{i=1}^{M}(k_i - 1), \qquad h_6 = \left[\sum_{i=1}^{M}(k_i - 1)\right] + 1$$

$$h_7 = \left[\sum_{i=1}^{M}(k_i - 1)\right] + 2, \quad h_8 = \left[\sum_{i=1}^{M}(k_i - 1)\right] + (k_2 - 1), \quad h_9 = \left[\sum_{i=1}^{M}(k_i - 1)\right] + k_2$$

$$h_{10} = \left[\sum_{i=1}^{M}(k_i - 1)\right] + (k_1 - 1)(k_2 - 1)$$

$$h_{11} = \left[\sum_{i=1}^{M}(k_i - 1)\right] + \left[\sum_{i \neq j, i<j}^{M-1}(k_i - 1)(k_j - 1)\right] + 1$$

Table 5: Values of α_0 and α_{Lm} in balanced case

	α_0				
	item 1	item 2	item 3	item 4	item 5
class 1	-0.6012	0.6358	0.2786	-0.3152	0.5294
class 2	0.1289	0.3371	0.1878	0.3102	0.3829
class 3	0.2698	0.0271	-0.5336	0.3746	-0.3508
	α_{lm}				
z_{1m}	-0.1741	-0.1904	0.1923	0.2254	0.2177
z_{2m}	0.1984	0.2835	0.2014	-0.2836	0.1674

Table 6: Values of β_0 and β_{Pj} in balanced case

	β_0	
	class 1 vs. class 3	class 2 vs. class 3
	0.2731	0.3269
	β_{pj}	
x_{i1}	-0.2170	0.3830
x_{i2}	0.4760	0.1240

Table 7: Values of α_0 and α_{Lm} in unbalanced case

	α_0					
	item 1	item 2	item 3	item 4	item 5	item 6
class 1	0.2797	0.4434	-0.4717	0.5080	0.5683	0.2855
class 2	0.2323	0.2686	0.3412	0.1323	0.4963	-0.2234
class 3	0.6281	-0.0856	0.0781	0.1472	0.6396	-0.4206
class 4	0.3330	0.4659	-0.2854	-0.1591	0.2062	0.1081
class 5	0.0268	0.2366	0.3235	0.5534	0.0693	0.4623
	α_{lm}					
z_{1m}	-0.2050	0.2243	0.2265	0.2655	0.2564	0.3224
z_{2m}	-0.1052	0.4443	0.7867	0.2103	0.3878	-0.0902

Table 8: Values of β_0 and β_{Pj} in unbalanced case

	β_0			
	class 1 vs. class 5	class 2 vs. class 5	class 3 vs. class 5	class 4 vs. class5
	0.2510	0.3041	0.0413	0.2035
	β_{pj}			
x_{i1}	0.1655	-0.2943	0.1719	0.3683
x_{i2}	0.0251	0.1850	0.4911	0.2988

Table 9: Observed contingency table of balanced case, averaging over 100 simulations

Response pattern	Group 1	2	3	4	5
$(y_1 = 1, y_2 = 1, y_3 = 1, y_4 = 1, y_5 = 1)$	18.22	19.03	27.87	27.66	29.21
$(y_1 = 1, y_2 = 1, y_3 = 1, y_4 = 1, y_5 = 2)$	9.35	10.27	16.27	16.99	18.88
$(y_1 = 1, y_2 = 1, y_3 = 1, y_4 = 2, y_5 = 1)$	21.72	22.51	17.38	18.77	20.99
$(y_1 = 1, y_2 = 1, y_3 = 1, y_4 = 2, y_5 = 2)$	12.33	11.51	10.89	11.65	14.09
$(y_1 = 1, y_2 = 1, y_3 = 2, y_4 = 1, y_5 = 1)$	12.43	12.59	20.24	19.35	23.73
$(y_1 = 1, y_2 = 1, y_3 = 2, y_4 = 1, y_5 = 2)$	6.82	6.63	12.42	12.89	17.13
$(y_1 = 1, y_2 = 1, y_3 = 2, y_4 = 2, y_5 = 1)$	14.53	15.16	13.32	14.13	16.77
$(y_1 = 1, y_2 = 1, y_3 = 2, y_4 = 2, y_5 = 2)$	7.63	7.63	7.69	8.64	11.51
$(y_1 = 1, y_2 = 2, y_3 = 1, y_4 = 1, y_5 = 1)$	10.82	11.68	22.88	22.18	23.63
$(y_1 = 1, y_2 = 2, y_3 = 1, y_4 = 1, y_5 = 2)$	5.68	5.81	13.36	14.39	16.16
$(y_1 = 1, y_2 = 2, y_3 = 1, y_4 = 2, y_5 = 1)$	12.61	13.13	15.24	15.73	15.62
$(y_1 = 1, y_2 = 2, y_3 = 1, y_4 = 2, y_5 = 2)$	6.81	7.29	8.99	8.61	10.46
$(y_1 = 1, y_2 = 2, y_3 = 2, y_4 = 1, y_5 = 1)$	7.15	8.01	16.34	16.21	19.19
$(y_1 = 1, y_2 = 2, y_3 = 2, y_4 = 1, y_5 = 2)$	4.03	3.95	9.34	10.07	14.02
$(y_1 = 1, y_2 = 2, y_3 = 2, y_4 = 2, y_5 = 1)$	8.93	8.82	10.72	11.22	13.26
$(y_1 = 1, y_2 = 2, y_3 = 2, y_4 = 2, y_5 = 2)$	4.72	4.23	6.78	6.74	9.61
$(y_1 = 2, y_2 = 1, y_3 = 1, y_4 = 1, y_5 = 1)$	36.65	38.49	27.67	28.68	32.02
$(y_1 = 2, y_2 = 1, y_3 = 1, y_4 = 1, y_5 = 2)$	18.89	20.14	16.32	16.41	21.21
$(y_1 = 2, y_2 = 1, y_3 = 1, y_4 = 2, y_5 = 1)$	43.55	45.03	18.61	19.09	24.91
$(y_1 = 2, y_2 = 1, y_3 = 1, y_4 = 2, y_5 = 2)$	23.61	23.09	11.00	11.46	14.87
$(y_1 = 2, y_2 = 1, y_3 = 2, y_4 = 1, y_5 = 1)$	11.28	12.27	13.77	13.80	16.16
$(y_1 = 2, y_2 = 1, y_3 = 2, y_4 = 1, y_5 = 2)$	13.02	13.49	11.78	12.71	16.73
$(y_1 = 2, y_2 = 1, y_3 = 2, y_4 = 2, y_5 = 1)$	29.85	30.07	12.98	14.88	18.42
$(y_1 = 2, y_2 = 1, y_3 = 2, y_4 = 2, y_5 = 2)$	15.47	16.00	8.31	8.86	12.37
$(y_1 = 2, y_2 = 2, y_3 = 1, y_4 = 1, y_5 = 1)$	22.33	23.32	22.28	21.95	25.73
$(y_1 = 2, y_2 = 2, y_3 = 1, y_4 = 1, y_5 = 2)$	11.28	12.27	13.77	13.80	16.16
$(y_1 = 2, y_2 = 2, y_3 = 1, y_4 = 2, y_5 = 1)$	27.09	26.44	14.28	15.32	19.05
$(y_1 = 2, y_2 = 2, y_3 = 1, y_4 = 2, y_5 = 2)$	13.78	14.14	8.48	8.84	11.61
$(y_1 = 2, y_2 = 2, y_3 = 2, y_4 = 1, y_5 = 1)$	14.13	15.52	16.34	16.42	20.01
$(y_1 = 2, y_2 = 2, y_3 = 2, y_4 = 1, y_5 = 2)$	8.09	7.96	9.52	11.07	14.34
$(y_1 = 2, y_2 = 2, y_3 = 2, y_4 = 2, y_5 = 1)$	18.15	17.43	10.72	11.24	14.16
$(y_1 = 2, y_2 = 2, y_3 = 2, y_4 = 2, y_5 = 2)$	8.77	9.78	6.46	7.16	9.49

Table 10: Observed contingency table of unbalanced case, averaging over 100 simulations

Response pattern	Group 1	2	3	4	5
$(y_1=1,y_2=1,y_3=1,y_4=1,y_5=1,y_6=1)$	15.44	15.48	13.09	13.01	18.91
$(y_1=1,y_2=1,y_3=1,y_4=1,y_5=1,y_6=2)$	18.72	18.92	9.83	12.26	14.98
$(y_1=1,y_2=1,y_3=1,y_4=1,y_5=2,y_6=1)$	7.05	6.93	9.13	8.32	11.27
$(y_1=1,y_2=1,y_3=1,y_4=1,y_5=2,y_6=2)$	9.17	8.90	6.50	7.33	9.01
$(y_1=1,y_2=1,y_3=1,y_4=2,y_5=1,y_6=1)$	11.45	11.80	12.84	11.83	13.85
$(y_1=1,y_2=1,y_3=1,y_4=2,y_5=1,y_6=2)$	14.09	13.64	9.74	10.15	11.39
$(y_1=1,y_2=1,y_3=1,y_4=2,y_5=2,y_6=1)$	5.49	5.16	9.12	7.81	8.40
$(y_1=1,y_2=1,y_3=1,y_4=2,y_5=2,y_6=2)$	6.34	6.83	6.60	6.72	6.93
$(y_1=1,y_2=1,y_3=2,y_4=1,y_5=1,y_6=1)$	12.15	12.29	14.02	14.31	18.95
$(y_1=1,y_2=1,y_3=2,y_4=1,y_5=1,y_6=2)$	15.09	14.68	10.81	12.77	14.73
$(y_1=1,y_2=1,y_3=2,y_4=1,y_5=2,y_6=1)$	5.46	5.95	9.26	9.00	11.29
$(y_1=1,y_2=1,y_3=2,y_4=1,y_5=2,y_6=2)$	7.10	7.01	6.82	7.65	9.16
$(y_1=1,y_2=1,y_3=2,y_4=2,y_5=1,y_6=1)$	8.57	9.32	13.77	12.33	14.05
$(y_1=1,y_2=1,y_3=2,y_4=2,y_5=1,y_6=2)$	10.64	11.50	10.74	10.57	11.25
$(y_1=1,y_2=1,y_3=2,y_4=2,y_5=2,y_6=1)$	3.73	4.69	9.93	7.87	8.91
$(y_1=1,y_2=1,y_3=2,y_4=2,y_5=2,y_6=2)$	5.05	5.04	7.57	7.24	6.60
$(y_1=1,y_2=2,y_3=1,y_4=1,y_5=1,y_6=1)$	14.83	15.09	7.69	9.73	12.55
$(y_1=1,y_2=2,y_3=1,y_4=1,y_5=1,y_6=2)$	18.15	19.07	5.92	9.23	11.63
$(y_1=1,y_2=2,y_3=1,y_4=1,y_5=2,y_6=1)$	6.36	7.51	5.30	5.46	7.69
$(y_1=1,y_2=2,y_3=1,y_4=1,y_5=2,y_6=2)$	8.33	8.70	3.95	5.44	7.44
$(y_1=1,y_2=2,y_3=1,y_4=2,y_5=1,y_6=1)$	10.79	11.38	7.05	7.90	9.67
$(y_1=1,y_2=2,y_3=1,y_4=2,y_5=1,y_6=2)$	13.85	14.08	5.36	7.63	8.37
$(y_1=1,y_2=2,y_3=1,y_4=2,y_5=2,y_6=1)$	4.60	5.35	4.75	5.05	5.68
$(y_1=1,y_2=2,y_3=1,y_4=2,y_5=2,y_6=2)$	6.29	6.66	3.70	4.27	4.75
$(y_1=1,y_2=2,y_3=2,y_4=1,y_5=1,y_6=1)$	12.11	11.92	8.03	9.41	12.20
$(y_1=1,y_2=2,y_3=2,y_4=1,y_5=1,y_6=2)$	15.47	14.55	6.27	9.77	10.46
$(y_1=1,y_2=2,y_3=2,y_4=1,y_5=2,y_6=1)$	5.28	5.97	5.37	6.06	7.89
$(y_1=1,y_2=2,y_3=2,y_4=1,y_5=2,y_6=2)$	6.98	7.01	4.33	4.86	5.82
$(y_1=1,y_2=2,y_3=2,y_4=2,y_5=1,y_6=1)$	8.44	8.36	7.82	8.87	9.03
$(y_1=1,y_2=2,y_3=2,y_4=2,y_5=1,y_6=2)$	11.22	10.49	5.85	7.67	7.72
$(y_1=1,y_2=2,y_3=2,y_4=2,y_5=2,y_6=1)$	4.29	4.33	5.59	5.33	4.94
$(y_1=1,y_2=2,y_3=2,y_4=2,y_5=2,y_6=2)$	5.08	5.22	4.33	4.58	4.49

Response pattern	Group 1	2	3	4	5
$(y_1=2,y_2=1,y_3=1,y_4=1,y_5=1,y_6=1)$	9.01	9.66	11.17	10.43	15.26
$(y_1=2,y_2=1,y_3=1,y_4=1,y_5=1,y_6=2)$	11.23	11.75	8.26	8.44	11.51
$(y_1=2,y_2=1,y_3=1,y_4=1,y_5=2,y_6=1)$	4.51	4.58	7.26	6.72	9.62
$(y_1=2,y_2=1,y_3=1,y_4=1,y_5=2,y_6=2)$	5.49	5.70	5.44	5.76	6.74
$(y_1=2,y_2=1,y_3=1,y_4=2,y_5=1,y_6=1)$	6.75	7.12	10.20	8.85	11.20
$(y_1=2,y_2=1,y_3=1,y_4=2,y_5=1,y_6=2)$	8.18	8.78	7.77	7.69	9.00
$(y_1=2,y_2=1,y_3=1,y_4=2,y_5=2,y_6=1)$	3.23	3.06	6.83	5.97	7.04
$(y_1=2,y_2=1,y_3=1,y_4=2,y_5=2,y_6=2)$	4.12	4.12	5.48	4.81	5.27
$(y_1=2,y_2=1,y_3=2,y_4=1,y_5=1,y_6=1)$	7.00	7.00	12.23	11.38	15.34
$(y_1=2,y_2=1,y_3=2,y_4=1,y_5=1,y_6=2)$	9.17	9.09	8.22	9.15	11.60
$(y_1=2,y_2=1,y_3=2,y_4=1,y_5=2,y_6=1)$	3.30	3.81	8.26	7.25	9.63
$(y_1=2,y_2=1,y_3=2,y_4=1,y_5=2,y_6=2)$	4.39	4.21	5.67	5.44	7.44
$(y_1=2,y_2=1,y_3=2,y_4=2,y_5=1,y_6=1)$	5.41	5.75	11.05	10.03	10.72
$(y_1=2,y_2=1,y_3=2,y_4=2,y_5=1,y_6=2)$	7.56	6.82	8.74	8.39	8.98
$(y_1=2,y_2=1,y_3=2,y_4=2,y_5=2,y_6=1)$	2.37	2.38	8.61	6.58	7.00
$(y_1=2,y_2=1,y_3=2,y_4=2,y_5=2,y_6=2)$	2.91	3.56	6.18	5.16	5.26
$(y_1=2,y_2=2,y_3=1,y_4=1,y_5=1,y_6=1)$	9.07	9.53	6.23	6.47	9.71
$(y_1=2,y_2=2,y_3=1,y_4=1,y_5=1,y_6=2)$	10.97	11.75	4.43	6.16	8.27
$(y_1=2,y_2=2,y_3=1,y_4=1,y_5=2,y_6=1)$	3.86	4.44	4.37	4.02	6.37
$(y_1=2,y_2=2,y_3=1,y_4=1,y_5=2,y_6=2)$	5.25	5.35	3.33	3.84	4.82
$(y_1=2,y_2=2,y_3=1,y_4=2,y_5=1,y_6=1)$	6.47	6.54	5.90	5.83	7.37
$(y_1=2,y_2=2,y_3=1,y_4=2,y_5=1,y_6=2)$	8.04	9.00	4.54	5.62	6.07
$(y_1=2,y_2=2,y_3=1,y_4=2,y_5=2,y_6=1)$	3.08	3.06	4.30	4.11	4.64
$(y_1=2,y_2=2,y_3=1,y_4=2,y_5=2,y_6=2)$	3.68	3.91	3.21	3.73	3.89
$(y_1=2,y_2=2,y_3=2,y_4=1,y_5=1,y_6=1)$	7.04	7.49	6.43	6.93	9.60
$(y_1=2,y_2=2,y_3=2,y_4=1,y_5=1,y_6=2)$	8.58	9.04	4.65	6.24	8.12
$(y_1=2,y_2=2,y_3=2,y_4=1,y_5=2,y_6=1)$	2.96	2.85	4.61	4.96	5.87
$(y_1=2,y_2=2,y_3=2,y_4=1,y_5=2,y_6=2)$	4.25	4.05	3.49	3.91	4.75
$(y_1=2,y_2=2,y_3=2,y_4=2,y_5=1,y_6=1)$	5.04	5.26	6.49	6.11	6.67
$(y_1=2,y_2=2,y_3=2,y_4=2,y_5=1,y_6=2)$	7.08	6.73	4.99	5.78	5.95
$(y_1=2,y_2=2,y_3=2,y_4=2,y_5=2,y_6=1)$	2.46	2.46	4.23	4.29	4.49
$(y_1=2,y_2=2,y_3=2,y_4=2,y_5=2,y_6=2)$	2.93	3.32	3.35	3.52	3.38

Table 11: Observed contingency table of first- and second-order marginals, averaging over 100 simulations

Response	Group				
pattern	1	2	3	4	5
$(y_1 = 1)$	307.61	313.83	251.08	270.43	318.42
$(y_2 = 1)$	250.17	255.53	291.14	281.22	341.29
$(y_3 = 1)$	273.89	283.85	219.29	230.59	287.71
$(y_4 = 1)$	283.77	290.28	230.37	251.71	327.04
$(y_5 = 1)$	337.61	343.88	270.13	290.94	355.11
$(y_6 = 1)$	217.60	226.52	260.93	252.22	315.81
$(y_1 = 1, y_2 = 1)$	155.54	158.14	159.77	159.17	189.68
$(y_1 = 1, y_3 = 1)$	170.95	175.50	120.57	132.14	160.93
$(y_1 = 1, y_4 = 1)$	177.69	179.98	126.32	144.61	182.39
$(y_1 = 1, y_5 = 1)$	211.01	212.57	148.83	167.44	199.74
$(y_1 = 1, y_6 = 1)$	136.04	141.53	142.76	142.29	175.28
$(y_2 = 1, y_3 = 1)$	140.27	142.43	139.26	136.10	170.38
$(y_2 = 1, y_4 = 1)$	144.28	145.96	145.97	149.22	195.44
$(y_2 = 1, y_5 = 1)$	170.46	173.60	172.48	171.59	211.72
$(y_2 = 1, y_6 = 1)$	110.92	114.98	166.77	151.69	191.44
$(y_3 = 1, y_4 = 1)$	157.44	163.36	111.90	122.62	164.19
$(y_3 = 1, y_5 = 1)$	187.04	193.59	130.02	141.23	179.74
$(y_3 = 1, y_6 = 1)$	121.99	126.69	125.23	121.51	159.23
$(y_4 = 1, y_5 = 1)$	194.03	197.31	137.28	155.69	203.82
$(y_4 = 1, y_6 = 1)$	125.43	130.50	132.45	133.46	182.15
$(y_5 = 1, y_6 = 1)$	149.57	153.99	154.01	153.42	195.08

Table 12: Simulation results of "**situation 1**" in balanced case

	no. group	mean	variance	% above 90th%-ile	% above 95th%-ile	% above 99th%-ile
T_1	5	140.419	381.386	165.298	168.832	184.317
T_2	5	140.419	381.387	165.298	168.833	184.318
T_3	5	140.420	381.387	165.298	168.833	184.318
Nominal asymptotic distribution						
T_1	5	159.229	345.390	186.458	191.088	199.370
T_2	5	159	318	182.234	189.424	203.399
T_3	5	159	318	182.234	189.424	203.399

Table 13: Simulation results of "**situation 2**" in balanced case

	no. group	mean	variance	% above 90th%-ile	% above 95th%-ile	% above 99th%-ile
T_1	5	162.180	339.864	188.266	191.263	202.460
T_2	5	159.509	338.743	183.911	188.962	199.383
T_3	5	162.184	339.870	188.272	191.266	202.465
Nominal asymptotic distribution						
T_1	5	159.230	345.391	186.460	190.090	199.370
T_2	5	159	318	182.234	189.424	203.399
T_3	5	159	318	182.234	189.424	203.399

Table 14: Simulation results of "**situation 3**" in balanced case

	no. group	mean	variance	% above 90th%-ile	% above 95th%-ile	% above 99th%-ile
T_1	5	166.843	569.964	202.338	206.251	212.555
T_2	5	159.508	338.742	183.912	188.962	199.972
T_3	5	162.184	339.870	188.272	191.266	202.464
Nominal asymptotic distribution						
T_1	5	159.230	345.391	186.460	190.090	199.370
T_2	5	159	318	182.234	189.424	203.399
T_3	5	159	318	182.234	189.424	203.399

Table 15: Simulation results of "**situation 1**" in unbalanced case

	no. group	mean	variance	% above 90th%-ile	% above 95th%-ile	% above 99th%-ile
T_1	5	31.753	74.620	41.902	45.463	54.409
T_2	5	65.096	97.559	77.240	83.331	85.622
T_3	5	65.506	98.267	77.480	83.333	85.622
Nominal asymptotic distribution						
T_1	5	93.825	796.756	134.624	150.099	155.742
T_2	5	105	210	123.947	129.918	141.620
T_3	5	105	210	123.947	129.918	141.620

Table 16: Simulation results of "**situation 2**" in unbalanced case

	no. group	mean	variance	% above 90th%-ile	% above 95th%-ile	% above 99th%-ile
T_1	5	112.617	463.602	140.582	150.719	166.195
T_2	5	82.007	119.623	97.725	102.131	109.548
T_3	5	85.545	114.977	100.786	106.047	111.311
Nominal asymptotic distribution						
T_1	5	93.827	796.468	134.776	150.026	156.049
T_2	5	105	210	123.947	129.918	141.620
T_3	5	105	210	123.947	129.918	141.620

Table 17: Simulation results of "**situation 3**" in unbalanced case

	no. group	mean	variance	% above 90th%-ile	% above 95th%-ile	% above 99th%-ile
T_1	5	129.527	806.741	173.690	181.156	185.074
T_2	5	82.010	119.623	97.730	102.130	109.550
T_3	5	85.544	114.980	100.790	106.046	111.310
Nominal asymptotic distribution						
T_1	5	93.827	796.468	134.776	150.026	156.049
T_2	5	105	210	123.947	129.918	141.620
T_3	5	105	210	123.947	129.918	141.620

Table 18: Power of "**situation 1**" in balanced case

			$\alpha = 0.05$	
test statistic	no.group	mean	variance	power
T_1	3	86.642	206.920	0.01
T_2	3	86.643	206.921	0
T_3	3	86.643	206.921	0

Table 19: Power of "**situation 2**" in balanced case

			$\alpha = 0.05$	
test statistic	no.group	mean	variance	power
T_1	3	131.240	224.264	0.79
T_2	3	115.122	219.229	0.46
T_3	3	131.447	224.234	0.76

Table 20: Power of "**situation 3**" in balanced case

			$\alpha = 0.05$	
test statistic	no.group	mean	variance	power
T_1	3	50.653	85.886	0
T_2	3	115.122	219.230	0.46
T_3	3	131.450	224.234	0.76

Table 21: Power of "**situation 1**" in unbalanced case

			$\alpha = 0.05$	
test statistic	no.group	mean	variance	power
T_1	3	28.918	158.416	0.01
T_2	3	46.668	257.740	0.01
T_3	3	44.542	102.021	0

Table 22: Power of "**situation 2**" in unbalanced case

			$\alpha = 0.05$	
test statistic	no.group	mean	variance	power
T_1	3	347.980	1094.687	1
T_2	3	143.270	726.348	0.98
T_3	3	86.203	122.648	0.65

Table 23: Power of "**situation 3**" in unbalanced case

			$\alpha = 0.05$	
test statistic	no.group	mean	variance	power
T_1	3	11.245	19.834	0
T_2	3	143.270	726.348	0.98
T_3	3	86.202	122.647	0.65