國 立 交 通 大 學

統 計 學 研 究 所

碩 士 論 文

潛在類別迴歸模型之適合度檢定

Goodness-of-fit Test for Latent Class Regression Model

研究生：高徽宜

指導教授：黃冠華 博士

中華民國 九十四 年 六 月

潛在類別迴歸模型之適合度檢定

Goodness-of-fit Test for Latent Class Regression Model

研究生：高徽宜　　Student：Hui-Yi Kao

指導教授：黃冠華 博士　　Advisor：Dr. Guan-Hua Huang

國立交通大學理學院

統計研究所

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 2005

Hsinchu Taiwan Republic of China

中華民國 九十四 年 六 月

潛在類別迴歸模型之適合度檢定

研究生：高徽宜　　　　指導教授：黃冠華　博士

國立交通大學統計學研究所

摘要

生物醫學以及社會心理方面的研究近來越來越常使用潛在類別迴歸(latent class regression)模型來分析多重類別資料與有興趣的共變數之間的關係。在潛在類別迴歸模型中多重類別資料會被整合摘要，而與風險因子之間的關係也會藉由模型中的線性迴歸方法整理出來。這些模型較於精簡並且能夠將多重類別資料的一些分析方法的理論基礎整合起來，然而這些優點卻是伴隨著一些很強的模型假設而來，這些假設有可能會對分析結果造成嚴重的影響，因此評估這些模型是否很適當的被使用是必須的。這篇論文中我們將簡介應用在logistic regression中Hosmer與Lemeshow提出的統計量並且將之延伸到潛在類別迴歸模型之中來做適合度檢定。

關鍵字：

多重類別資料；適合度檢定；潛在類別;迴歸；卡方分配

GOODNESS-OF-FIT TEST FOR LATENT CLASS REGRESSION MODEL

Student: Hui-Yi Kao   Advisor: Dr. Guan-Hua Huang

Institute of statistics

National Chiao Tung Unerversity

Abstract

Biomedical and psychosocial researchers increasingly utilize latent class regression (LCR) models to analyze relationships between measured multiple categorical outcomes and covariates of interest. In LCR, the multiple outcomes are summarized and their associations with risk factors are determined in a single modeling step. These models are parsimonious and can incorporate theory underlying the multiple response choices. However, these advantages come at the price of strong modeling assumptions which may critically influence analytic findings. Careful evaluation of model appropriateness is necessary. In this thesis, we first introduced Hosmer-Lemeshow statistic for multiple logistic regression model and then extended the method to LCR model to assess overall fit of the LCR model. An analysis of how measured health impairments affect older persons' functioning is used for illustration.

KEY WORDS: categorical data; goodness-of-fit test; latent class regression; chi-square distribution.

誌謝

在統計所的這兩年來，我要感謝我的指導教授—黃冠華老師，指引我研究的方向，遇到困難時老師在忙碌之餘也總是會抽空為我解惑並且耐心的引導我，讓我能將研究做的更完整，進而完成這篇論文。另外也要感謝我的口試委員—徐南蓉老師，盧鴻興老師以及楊志堅老師的悉心指教，讓我將論文修正的更清楚。

在生活上我要感謝我的媽媽含辛茹苦栽培我、支持我，此外也要感謝視我如己出的姨媽、李老師與叔叔，他們在我迷惘時總能適時的指引我方向，給我關懷與溫暖。接著要感謝的是朋友們的陪伴，讓我無論是快樂或悲傷總是有人與我分享。最後謹將我這小小的成就獻給我最敬愛的母親、李老師與叔叔、姨媽。

Contents

1 INTRODUCTION

Many concepts in medical research are unobservable, hence valid surrogates must be measured in place of these concepts. Models that permit exploration of relationships between unobservable variables and their surrogates are referred to as latent variable models. When measured surrogates are discrete, the latent class analysis (LCA) model (Green [1], Lazarsfeld and Henry [2], Goodman [3], Haberman [4,5])is the most commonly applied latent variable approach. LCA is distinguished by its treatment of the unobservable variable as categorical (i.e., as defining latent "classes"). The model assumes that the unobservable (latent) variable fully explains the associations between observed indicators, thus measured indicators are independent of one another within any class of the latent variable. Recently, several authors extended the LCA model to incorporate covariate effects on estimating the underlying mechanism (Dayton and Macready [6], Formann [7], Bandeen-Roche et al. [8], Muthén and Shedden [9]), or on estimating measured indicator distributions within latent classes (Formann [7], Melton et al. [10], Muthén and Shedden [9]). This thesis studies a LCA model that uses covariates on describing distributions of both the underlying latent class and the measured indicators themselves henceforth, latent class regression-LCR (Huang and Bandeen-Roche [11]).

The LCR model is parsimonious, explicitly recognize and hence may mitigate errors in measurement, and can give well-summarized inferences on the theory underlying the choice of multiple indicators and their relationships with covariates of interest. However, these come at the price of assuming conditional independence between measured items within each latent class and parametric models of incorporated covariates with the latent class and measured indicators. Therefore, careful evaluation of these model assumptions is necessary and important to prevent that scientific findings are to be driven by the statistical assumptions rather than by the data.

In application of LCA, overall population can be grouped by possible response patterns, and therefore Pearson χ^2 and likelihood ratio goodness-of-fit can be applied for evaluating overall model fit (Gooman [3], Bartholomew [12], Formann [7]). However, when the model includes continuous covariates, every cell contains only one observation, and the saturated model would contain as many parameters as there are observed. Therefore, the χ^2 statistic of the model does not work directly, because the degree of freedom increases with the sample size. This is not acceptable as the χ^2 sampling distribution hold only when the sample size is large relative to the degree of freedom.

The most widely used goodness-of-fit for situations with a continuous covariate is the Hosmer-Lemeshow statistic (Hosmer and Lemeshow [19,20]) for multiple logistic regression models. In this thesis, we applied the idea of doing the Hosmer-Lemeshow statistic to LCR model where the outcome variable is not only binary but category, and each individual has not only one outcome variable but multiple outcome variables. Therefore, a χ^2 test can be applied to assess the goodness of fit of the LCR model.

To summarize organization of this thesis, section 2 provides brief description of the LCR model that this thesis studies and its associated model assumptions. In section 3, we first describe the motivated method in logistic regression. Then the idea of doing goodness-of-fit test of LCR model is developed. After the test statistic is proposed, we begin to simulate it's distribution in section 4, and the power of the proposed statistic is also examined for the alternative models. Visual functioning data are used to illustrate the proposed goodness-of-fit method in section 5. Discussion is provided in section 6.

2 LATENT CLASS REGRESSION

To specifically describe the model, let $(Y_{i1}, \ldots, Y_{iM})^T$ represent the $M \times 1$ response vector and S_i be the unobservable latent class, for the ith individual in a study sample of N persons. Y_{im} can take values $\{1, \ldots, K_m\}$, where $K_m \geq 2$, $m = 1, \ldots, M$ and S_i can take values $\{1, \ldots, J\}$. The basic structure of latent class analysis model for the ith individual can be represented as

$$Pr(Y_{i1} = y_1, \ldots, Y_{iM} = y_m) = \sum_{j=1}^{J} \{\eta_j \prod_{m=1}^{M} \prod_{k=1}^{K_m} p_{mkj}^{y_{mk}}\}. \tag{1}$$

Here, $y_{mk} = I(y_m = k) = 1$ if $y_m = k$; 0 otherwise, $\eta_j = Pr(S_i = j)$ are the "latent class probabilities" of each underlying variable category, and $p_{mkj} = Pr(Y_{im} = k|S_i = j)$ are the "conditional probabilities" of the measured responses given the underlying variable category. The model of LCA is based on the concept of conditional independence- i.e., the observed variables are assumed to be statistically independent within latent classes.

To incorporate covariate effects into LCA, let $(\mathbf{x}_i, \mathbf{z}_i)$ be the associated covariate vector for the ith person, where $\mathbf{x}_i = [1, x_{i1}, \ldots, x_{iP}]^{\mathrm{T}}$ are predictors for estimating $Pr(S_i = j)$, and $\mathbf{z}_i = [\mathbf{z}_{i1}, \ldots, \mathbf{z}_{iM}]^{\mathrm{T}}$ with $\mathbf{z}_{im} = [z_{im1}, \ldots, z_{imL}]^{\mathrm{T}}$, $m = 1, \ldots, M$ are covariates used for $Pr(Y_{im} = k|\mathrm{S}_i = j)$. The two sets of covariates may include any combination of continuous and discrete measures, and they may be mutually exclusive or overlapped. The latent class regression (LCR) model is then stated as

$$Pr(Y_{i1} = y_1, \ldots, \mathrm{Y}_{i\mathrm{M}} = y_m|\mathbf{x}_i, \mathbf{z}_i) = \sum_{j=1}^{J} \{\eta_j(\mathbf{x}_i) \prod_{m=1}^{M} \prod_{k=1}^{K_m} [p_{mkj}(\mathbf{z}_{im})]^{y_{mk}}\}, \tag{2}$$

with $\eta_j(\mathbf{x}_i)$ and $p_{mkj}(\mathbf{z}_{im})$ as in the generalized linear framework (McCullagh and Nelder [14]). Often, (2) is implemented assuming generalized logit (Agresti [15]) link functions:

$$\log[\frac{\eta_j(\mathbf{x}_i)}{\eta_J(\mathbf{x}_i)}] = \beta_{0j} + \beta_{1j}x_{i1} + \ldots + \beta_{Pj}x_{i\mathrm{P}} = \mathbf{x}_i^T\boldsymbol{\beta}_j, \tag{3}$$

and

$$\log[\frac{p_{mkj'}(\mathbf{z}_{im})}{p_{mK_mj'}(\mathbf{z}_{im})}] = \gamma_{mkj'} + \alpha_{1mk}z_{im1} + \ldots + \alpha_{Lmk}z_{imL} = \gamma_{mkj'} + \mathbf{z}_{im}^T\boldsymbol{\alpha}_{mk}, \tag{4}$$

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

Through (3), we can summarize the effects of risk factors on the underlying mechanism. (4) aims to isolate the classification of subject's measured indicators to the underlying outcome apart from variables that confound measurements, hence hopefully improve the accuracy of classifying of individuals. For example, in evaluating functional disability, some data have suggested that women rate tasks as "difficult" more readily than men (Bandeen-Roche et al [16]). Without adjusting for a gender effect, the model might well classify some men and women with identical underlying functioning differently (men as "able", women as "disabled"). Parameters in (3) and (4) can be estimated through the EM algorithm (Dempster, Laird and Rubin [17]), which is a broadly applicable approach to the interactive computation of maximum likelihood estimates while the model can be viewed as an "incomplete-data" problem. Following three assumptions are necessary for obtaining the LCR model (2), (3) and (4):

(C1) Latent class membership is associated with $\mathbf{x}_i$ only, and their relationship can be stated as (3):

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

(**C2**) Conditioning on class membership, measured responses are only associated with $\mathbf{z}_i$ and their marginal mean associations with $\mathbf{z}_i$ can be stated as (4):

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

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

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

(**C3**) Multiple measurements are conditionally independent given class membership and $\mathbf{z}_i$:

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

More detailed model characteristics, parameter estimations and theoretical properties of the proposed LCR can be found in Hung and Bandeen-Roche [11].

3 THE TEST STATISTIC

3.1 Hosmer-Lemeshow goodness-of-fit test (1980)

In multiple logistic regression, Hosmer and Lemeshow ([19,20]) proposed the following test statistic for evaluation goodness-of-fit:

Let Y_i= 0 or 1 be the outcome variable, and $\mathbf{x}_i^T$ =$(x_{i1}, \ldots, x_{ip})$ be the independent variables. Let $\pi(\mathbf{x}_i) = \Pr(Y_i = 1|\mathbf{x}_i)$=$\exp(\beta_0 + \boldsymbol{\beta}^T\mathbf{x}_i)/(1$+$\exp(\beta_0 + \boldsymbol{\beta}^T\mathbf{x}_i))$ where $\boldsymbol{\beta}^T = (\beta_1, \ldots, \beta_p)$. The likelihood function is $L(\mathbf{y}; \mathbf{x}, \beta_\mathbf{0}, \boldsymbol{\beta}) = \prod_{i=1}^n \pi_i^{y_i}(1-\pi_i)^{1-y_i}$, where $\pi_i = \pi(\mathbf{x}_i)$, i=$1, \ldots, n$. So $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}$ can be obtained as the maximum likelihood estimators, and hence $\hat{\pi}_i$ can be estimated. The basis of

Hosmer-Lemeshow statistic is a $2 \times g$ contingency table which was obtained by defining a random variable W, where $w_i = j$ if $c_{j-1} \leq \hat{\pi}_i < c_j$, j=1, ..., g; i=1, ..., 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 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, ..., g. Notationally the "observed" frequencies may tabulated as Table 1.

One way of selecting the cut points $c_0, \ldots, c_g$ is by defining $\hat{\pi}_{(1)} \leq \hat{\pi}_{(2)} \leq \ldots \leq \hat{\pi}_{(n)}$ as the ordered values of $\hat{\pi}$ and let $\hat{c}_j = \hat{\pi}_{([jn/g])}$, where $[\frac{jn}{g}]$ represents the largest integer less than or equal to $\frac{jn}{g}$, j=0, 1, ..., g. Let $\hat{w}_i = j$ if $\hat{c}_{j-1} \leq \hat{\pi} < \hat{c}_j$. 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

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

and the simulation result indicated that a good approximation to the distribution of C_g is $\chi^2(g-2)$ distributed.

3.2 Proposed goodness-of-fit test for LCR

Similar to the Hosmer-Lemeshow goodness-of-fit, we can extend the method to our LCR model and get a test statistic by grouping our outcome variables as follows. Let the joint probability

$$Pr(\mathbf{Y}_i = \mathbf{y}_h; \boldsymbol{\phi}) = Pr\{(Y_{i1}, \ldots, Y_{iM}) = (y_{h1}, \ldots, y_{hM}); \boldsymbol{\phi}\} = \pi_{ih}(\boldsymbol{\phi}), \tag{2}$$

where $i = 1, \ldots, N$; $h = 1, \ldots, K^*$; $K^* = \prod_{m=1}^{M} K_m$; and $\boldsymbol{\phi}$ is the vector of parameters. Here, the observation $\mathbf{Y}_i$ for each i may take values $\{\mathbf{y}_1, \ldots, \mathbf{y}_{K^*}\}$ where $\mathbf{y}_h$ could be one of all possible multiple outcome for $\mathbf{Y}_i$, h=1, ..., K^*. The basis of the proposed goodness-of-fit statistic is a $K^* \times g$ contingency table in our LCR model. This table is obtained by defining a random variable W, where $W_i = j$ if $c_{j-1} \le \pi_{i1}(\hat{\boldsymbol{\phi}}) < c_j$; $j = 1, \ldots, g; i = 1, \ldots, N$; The c_j's are known constants such that $0 = c_0 < c_1 < \ldots < c_{g-1} < c_g = 1$, and $\pi_{i1}(\hat{\boldsymbol{\phi}})$ is the estimate of $\pi_{i1}(\boldsymbol{\phi})$ evaluated at the MLE of $\boldsymbol{\phi}$. Denote the counts in jth group as n_j, that is, n_j is the number of persons whose $W_i = j$. And denote O_{hj} is the observed frequency of occurrence of the pair $(\mathbf{Y}_i = \mathbf{y}_h,\, W_i = j)$ in the sample, where $h = 1, \ldots, K^*;\ K^* = \prod_{m=1}^{M} K_m;\ j = 1, \ldots, g$. So the total observed frequencies may be tabulated as Table2.

The goodness-of-fit statistic is obtained by comparing the "observed" frequencies to ones which are "expected" if the hypothesis of a LCR model holds. The expected frequency for the hth combination and the jth group is obtained as $E_{hj} = \sum_{r \in I_j} \pi_{rh}(\hat{\boldsymbol{\phi}})$, where $I_j = \{i : c_{j-1} \le \pi_{i1}(\hat{\boldsymbol{\phi}}) < c_j\}$, j=1,2, …, g. Hence, the test statistic is

$$T = \sum_{h=1}^{K^*} \sum_{j=1}^{g} \frac{(O_{hj} - E_{hj})^2}{E_{hj}}. \tag{3}$$

There are many methods to group the observations (that is, to define the cut points, $c_0,\ c_1,\ \ldots,\ c_g$). In this thesis, we adopt the following strategy:
Define $\pi_{(1)1}(\hat{\boldsymbol{\phi}}) \le \pi_{(2)1}(\hat{\boldsymbol{\phi}}) \le \ldots \le \pi_{(N)1}(\hat{\boldsymbol{\phi}})$as ordered values of $\pi_{i1}(\hat{\boldsymbol{\phi}})$ for all i. In other words, the cut points depend on the data and are determined so that n/g persons fall in each interval. Let $c_j = \pi_{([\frac{jn}{g}])1}(\hat{\boldsymbol{\phi}})$ where $[\frac{jn}{g}]$ represents the largest integer less than or equal to $\frac{jn}{g}$.

3.3 Large sample of T

The distribution of T cannot be obtained from a straightforward application of usual theory used for χ^2 goodness-of-fit test because:

(a). Parameter estimates are determined using likelihood functions for "ungrouped" data.

(b). The frequency, O_{hj} in the $K^* \times g$ table depend on the estimated parameters, namely the cells are random not fixed.

A χ^2 test under (a) were first addressed by Chernoff and Lehmann (1954) and then Watson (1959). Moore (1971) and Moore and Spruill [22] considered the distribution of the χ^2 goodness of fit statistics under both (a) and (b). Their work extended the results of Watson to the case of random rectangular cells. Drust (1979) generalized these results to include random cells other than rectangles. The application of the results of Moore and Spruill [22] and Drust [21] to the problem is contained in the following theorem.

Theorem 1 *Let* $\lambda_1, \ldots, \lambda_{K^{**}}$ *are the non-zero or 1 eigenvalues of the matrix* $\Sigma(T) = I - \mathbf{q}\mathbf{q}^T - BJ^{-1}B^T$. *Here,* I *is a* $K^*g \times K^*g$ *identity matrix and* $\mathbf{q}$ *is a* $K^* \times g$ *vector with elements* $\sqrt{P_{hj}}$, $h = 1, \ldots, K^*$, *where* P_{hj}=$Pr(\mathbf{Y} = \mathbf{y}_h, W_i = j)$. $\mathbf{B}$ *is a* $(K^* \times g)K^{**}$ *matrix and has a general element given by* $\frac{1}{\sqrt{P_{hj}}}\frac{\partial P_{hj}}{\partial \boldsymbol{\phi}_l}$, $\boldsymbol{\phi} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha})$. J^{-1} *is the asymptotic variance covariance matrix of the discriminant function estimates* $\boldsymbol{\phi}$. *Then under LCR assumptions (2), (3), and(4), the distribution of* T *will be asymptotically* $(N \to \infty)$

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

where $0 < \lambda_i < 1$; $i = 1, \ldots, K^{**}$; $K^{**} = (P+1)(J-1) + (J+L)\sum_{m=1}^{M}(K_m - 1)$ *is the total number of the parameters in the LCR model.*

Proof:

The proof of the theorem follows from verifying that the regularity conditions necessary for the proof of *theorem* 5.1 in Moore and Spruill [22] are satisfied, see appendix.

In practice, the expected frequencies of some possible response patterns of $\mathbf{Y}_i$ usually less than 5 even to 0. However, the χ^2 approximation for the test distribution loses validity when a large number of response patterns have low expected frequencies. So we should add those response patterns which their expected frequency less than 5 to the next ones until no one take value less than 5. Therefore, we can apply our proposed goodness-of-fit method on the new response patterns and theorem1 still holds when K^* become the new number of response patterns after the combination.

4 SIMULATION STUDY

4.1 Data generation under the LCR model

Here, we simulated three-class LCR with five two-level measured indicators, two covariates associated with conditional probabilities, two covariates associated with latent prevalences, two, five, and ten groups (i.e., $J = 3,\ M = 5,\ K_1 = \ldots = K_5 = 2,\ P = L = 2,\ g = 2, 5, 10$). The model parameters β_{pj} can be determined through the method. For each $p \in \{0, 1, \ldots, P\}$

- randomly selected: $\beta_{pj} = k_1 U_j,\ U_j \sim U(0,1),\ j = 1, \ldots, (J-1)$;

where k_1 was constants such that $\sum_{j=1}^{J-1} \beta_{pj}$ equaled the preselected total. The method was also applied to create $\{\gamma_{jmk},\ j = 1, \ldots, (J-1)\}$ for all m, k, and $\{\alpha_{qmk},\ m = 1, \ldots, M;\ k = 1, \ldots, (K_m - 1)\}$ for all q. All $(\beta_{pj}, \gamma_{jmk}, \alpha_{qmk})$ pairs were generated by the same method.

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

$z_{im1} \sim Bernoulli(0.4),\ z_{im2} \sim Normal(0,1),\ i = 1, \ldots, N\ for\ each\ m,$

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

all z_{imq} and x_{ip} are mutually independent.

The selected sample size was 2400, 2400, and 4800 which gave roughly 15 individuals per cell of the contingency tables for the goodness-of-fit tests with two , five and ten group. The observable $\mathbf{Y}_i$ were then generated with 100 replications. Actually, by the common collected data the calculated expected frequencies table usually not equally distributed. For example, in evaluation functional disability, most people would task as "not difficult" more than " difficult", so the number of person who tasks "difficult" item would be very sparse. Therefore, in this thesis we simulate two situations to discuss their large sample behavior: one is equally distributed data and another is not equally distributed data.

The simulation results were represented in Table3. Here, "balance" represented the equally distributed data and $\chi^2(142)$ indicated the real χ^2 distribution; "unbalance" represented the not equally distributed data. From the results shown in Table 3, the equally distributed data was well approximated to a χ^2 distribution with degree of freedom 142 and the unequally distributed data was bad approximated.

4.2 Data generation under alternative models

This simulations considered thus far have demonstrated that the test statistic have well defined distributions under the null hypotheses that the LCR (2) model holds.

To examine the power of the proposed test statistic, data were generate from the alternative models which their covariates were generated from the distributions presented in Table 4 and we considered the situations when we use a simpler model to fit the data generated from a complicated model.

For situations 1-17 we use three different link functions for η to simulate. The selected sample size was 2400 and $\mathbf{Y}_i's$ were generated with 100 replication for each situation. All $(\beta_{pj}, \gamma_{jmk}, \alpha_{qmk})$ pairs were generated by the same method as we mentioned in section 4.1. The generated data were then fitted by the LCR model stated in section 4.1. The fitted model was three-class LCR with five-two level measured indicators, two covariates associated with conditional probability, two covariates associated with latent prevalence and five groups(i.e., J=3, M=5, $K_1 = \ldots = K_5 = 2$, P=L=2, g=5).

Situations 1-5 used the original link function (3) we represented in section 2 and 6-10 used probit link as follows:

$$\begin{aligned}
\eta_1 &= \Phi(\beta_{01} + \beta_{11}x_1 + \beta_{21}x_2); \\
\eta_2 &= \Phi(\beta_{02} + \beta_{12}x_1 + \beta_{22}x_2) \times (1 - \eta_1); \\
\eta_3 &= 1 - \eta_1 - \eta_2.
\end{aligned}$$

Situation 11-15 used proportional odds model:

$$\begin{aligned}
\eta_1 &= \frac{\exp(\beta_{01} + \beta_{11}x_1 + \beta_{21}x_2)}{1 + \exp(\beta_{01} + \beta_{11}x_1 + \beta_{21}x_2)}; \\
\eta_2 &= \frac{\exp(\beta_{02} + \beta_{11}x_1 + \beta_{21}x_2)}{1 + \exp(\beta_{02} + \beta_{11}x_1 + \beta_{21}x_2)} - \eta_1, \;\; \beta_{02} > \beta_{01}; \\
\eta_3 &= 1 - \eta_1 - \eta_2.
\end{aligned}$$

Situation 16, 17 used probit and proportional odds link with the original covariates.

For situation 18, we generated three-class LCR model with five-two level stated in section 4.1 and then fitted by the two-class LCR model with five-two levels. The test results of the above alternative models were illustrated in Table 5.

The simulation results indicated that the LCR (2) did not fit the probit data particularly well, that is, the test statistic had higher power as with the probit alternative model than the other two alternative models. The test statistic did not appear to be particularly powerful in detecting the difference between the proportional odds alternative and LCR (2) models and was not powerful to detect the LCR (2) models with different covariates. For the differences between three-class and two-class LCR models, the statistic was also not sensitive. These not powerful simulation results may be due to the limiting of the replicated times, there were some difficulties to increase the replication times of our proposed models, and the way the grouping was defined or the selected alternative models were not far from the LCR model.

5 THE SALISBURY EYE EVALUATION PROJECT

5.1 Background

To illustrate the proposed diagnostic methods, we use data from Salisbuty Eye Evaluation(SEE) project. The SEE project is describe in detail in West et al. [24]. Briefly, SEE is a population-based, prospective study of risk factors for ocular pathology and of how vision affects functioning in older persons. An age- and race-stratified random sample of Salisbury, Maryland residents between the ages of 65 and 84 years was drawn from the Health Care Financing Administration (HCFA) Medicare Database. To be eligible for the study, participants had to be able to communicate in English, travel to the clinic for vision tests, and score greater than 17 on the Mini-Mental

State Examination (MMSE: Folstein et al. [29]). The response rate to both the home interview and clinic examination was 65%, excluding the ineligible. Twenty-five hundred and twenty persons agreed to participate in both activities. Table 6 shows the demographic characteristics of SEE participants.

The analysis we report in this thesis aims to describe the association between functioning in activities that require seeing at a distance (far vision functioning) and psychophysical measures of visual impairment, adjusting for potential confounding variables. In the SEE project, visual functioning was determined using the Activities of Daily Vision Scale (ADVS) questionnaire (Mangione et al. [25], Valbuena et al. [26]). The analysis we report used selfreported difficulty doing five ADVS activities as responses: *reading street signs in daylight; reading street signs at night; walking down steps during daylight; walking down steps in dim light;* and *watching TV.* Here, we measure difficulty as a binary indicator (1=having difficulty; 2=no difficulty) for each activity, except for *reading street signs at night* which is measured as a three-level categorical indicator(1=extreme or moderate difficulty; 2=a little difficulty; 3=no difficulty). The hope was that, together, these five questions characterized the underlying far visual functioning. The frequency distributions of far vision subscale items are shown in Table 6. The distributions are severely skewed with most participants reporting no difficulty at all in all items.

In the SEE project, visual impairment was determined using multiple psychophysical vision tests (Rubin et al. [27]). Our analysis include five test: visual acuity of both eyes at regular luminance, contrast sensitivity of the better eye, glare sensitivity of the worse eye, stereoacuity of both eyes and central visual field of both eyes. For all the measures except contrast sensitivity, a higher score indicates worse vision.

5.2 Assess the goodness of fit and analysis result

A latent class regression model (2) for self-reported visual disability was fitted as a function of visual impairment variables, the number of reported comorbid diseases, and the following personal demographic characteristics: age at clinic exam, MMSE score, years of education, indicator of being female, indicator of being African-American, and General Health Questionnaire depression subscale score (GHQ score: Goldberg [28]). The vision and disease variables were treated as primary predictors of latent class membership ($\mathbf{x}_i$), and the personal characteristics were modeled as having direct effects on measured indicators themselves within classes ($\mathbf{z}_i$). The analysis was applied to the subsample of participants who rated each far vision item and also had no missing covariates(N=1641). Table 6 presented the characteristics and frequency distribution of far vision difficulty items in the SEE project, and from Table 6 we can find that the proportion of choosing no difficulty is much larger than difficulty.

We started with a three-, four-, and five-class LCR model, and the hypothesis is:

H_0: The fitted model explains the data well. vs. H_1: Not H_0.

The goodness of fit for the models began by the grouping method we proposed before. Table 7 and Table 8 displays the original contingency table for the expected and observed values when we fix a five-class model where $\mathbf{y}_1$ represented the response persons who self-reported as: *signs-day:have difficult; signs-night: extreme difficult; steps-day: extreme difficult; steps-dim: have difficult; watch TV: have difficult;* $\mathbf{y}_2$ represented the response persons who self-reported as: *signs-day:no difficult; signs-night: extreme difficult; steps-day: extreme difficult; steps-dim: have difficult; watch TV: have difficult,* and so on. We can find that some cells of the expected frequencies table are very sparse, and there are many cells take values less than 5. However, the χ^2 approximation for the test distribution loses validity when a large number of response patterns have low expected frequencies. So we added one row of the expected frequencies table to the next ones until no element of the row take value less than 5, and then we got a new contingency table as Table 12. Here, the first row combined the first 21 rows of the original observed frequencies table; the $2nd$ row combined $22th - 25th$ rows ,the $3rd$ row combined $26th - 30th$ rows, the $4th$ row combined $31th - 34th$ rows, the $5th$ row combined $35th - 38th$ rows ; the $6th$ row combined $39th - 44th$ rows, the $7th$ row combined $45th - 46th$ rows, the $8th$ row combined $47th - 48th$ rows. Table 9 and 10 are the expected and observed contingency tables after the combination. We can apply the χ^2 test after the combination and the test statistic is 38.53494. According to Hosmer and Lemeshow, the contribution of $\sum_{i=1}^{K^{**}} \lambda_i \chi_i^2(1)$ in theorem we proposed in Section 3 is approximately that of $\chi^2(K^{**} - 2)$. So the distribution of our statistic is approximately to $\chi^2(K^* g - g - 2)$ where K^* is number of the response patterns after the combination, and hence in five-class LCR model of SEE project is $\chi^2(33)$. Because $38.53494 < \chi^2_{0.95}(33) = 47.39988$, so we can conclude that the five-class LCR model explains the data well.

Similarly, Table 11 and Table 12 represents the original contingency tables of expected and observed frequencies for the four-class LCR model, and Table 15 and Table 16 represents those for three-class LCR model. For the three- and four-class LCR model, the tables also need to merge some patterns to others to let all the elements of expected table equal or larger than 5. Table 13, 14 and Table 17, 18 represents the new tables after combination for four- and three-class LCR model. The statistic for the new table of four-class LCR model was $35.92977 < \chi^2_{0.95}(38) = 53.38354$, so the four-class LCR model also explains the data well. The statistic for the new table of three-class LCR model was $83.83406 > \chi^2_{0.95}(33) = 47.39988$. Hence, the data is not well explained by the three-class LCR model. From the above tests, we can conclude that the four-class LCR model is the simplest model which can well explain the data .

6 DISCUSSION

In this thesis, we implement a latent class regression model that allows two types of covariates effects: relationships between primary predictors and responses that are mediated through the underlying variable, and direct effects of secondary covariates on the measured indicators themselves. This model is very useful in addressing scientific questions, however, the model is so complex that scientific findings are likely to be driven by the statistical assumptions rather than by the data. We develop goodness-of-fit test for assessing overall model fit. As long as operating with careful evaluation of model appropriateness, a great deal can be learned from the LCR model.

The number of groups was determined by cases when we forming the contingency table. In this thesis, we selected the one such that the contingency table would not be too large or too small, and then the replicated times could be reduced in reasonable ranges.

There is a lack of simulation-based investigation into the success in detecting the targeted model violations. Additional works on how various model violations appear on the proposed method is needed to identify strengths and weakness.

APPENDIX

Let θ be the unknown parameter in forming χ^2 statistics and is estimated by $\theta_N = \theta_N(\mathbf{Y}_1, \ldots, \mathbf{Y}_N)$. The parameter θ ranges over an open set Ω_1 in R^g. The cells are chosen by φ_N=$\varphi_N(\mathbf{Y}_1, \ldots, \mathbf{Y}_N)$. Let $F(\mathbf{y}|\theta, \boldsymbol{\phi})$ be the cdf of $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_N\}$. The null hypothesis is that $\mathbf{Y}_i$ have a cdf $F(\mathbf{y}|\theta)$. We will explore the large-sample behavior of tests for the null hypothesis under the sequences of parameter values $(\theta_0, \boldsymbol{\phi}_N)$ where $\theta_0 \in \Omega_1$ and $\boldsymbol{\phi}_N = \boldsymbol{\phi}_0 + N^{-1/2}\gamma$ for fixed γ in $R^{K^{**}}$. H_0 is the special case $\gamma = 0$. We will assume that under $(\theta_0, \boldsymbol{\phi}_N)$, $\varphi_N - \varphi_0 = o_{K^{**}}(1)$ for some φ_0 and $\theta_N = \theta_0 = o_{k^{**}}(1)$. We will suppress arguments θ, φ, $\boldsymbol{\phi}$ whenever they take the values θ_0, φ_0, $\boldsymbol{\phi}_0$ respectively. The resulting cells are denoted by $I_\sigma(\varphi)$, 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 $P_\sigma(\theta, \boldsymbol{\phi}, \varphi)$ where σ=1, 2, ..., $K^{**}g$. Then regular conditions of *theorem* 5.1 in Moore and Spruill [22] are satisfied as follows:

(**A1**). Under $(\theta_0, \boldsymbol{\phi}_N)$, $\theta_N - \theta_0 = O_{K^{**}}(N^{-1/2})$ and $\varphi_N - \varphi_0 = o_{K^{**}}(1)$. Every vertex $\mathbf{y}(\varphi)$ of every cell $I_\sigma(\varphi)$ is a continuous R^M-valued function of φ in a neighborhood of φ_0.

(**A2**). For each σ, $P_\sigma(\theta, \boldsymbol{\phi}, \varphi)$ is continuous in (θ, $\boldsymbol{\phi}$, φ) and continuously differentiable in a neiborhood of (θ_0, $\boldsymbol{\phi}_0$, φ_0). Moreover, $\sum_{\sigma=1}^{K^{**}g} P_\sigma = 1$ and $P_\sigma > 0$ for each σ.

(**A3**). $F(\mathbf{y})$=$F(\mathbf{y}|\theta_0, \boldsymbol{\phi}_0)$ is continuous at every vertex $\mathbf{y}(\varphi_0)$ of every cell $I_\sigma(\varphi_0)$. As $N \to \infty$, $sup_{\mathbf{y}}|F(\mathbf{y}|\boldsymbol{\phi}_N) - F(\mathbf{y})| \to 0$.

(**A4**). Under $(\theta_0, \boldsymbol{\phi}_N)$

$N^{1/2}(\theta_N - \theta_0) = N^{-1/2}\sum_{i=1}^{N} h(\mathbf{Y}_i, \boldsymbol{\phi}_N) + A\gamma + o_{K^{**}}(1)$

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

$$E[h(\mathbf{Y}, \boldsymbol{\phi}_N)|(\theta_0, \boldsymbol{\phi}_N)] = 0$$
$$E[h(\mathbf{Y}, \boldsymbol{\phi}_N)h(\mathbf{Y}, \boldsymbol{\phi}_N)'|(\theta_0, \boldsymbol{\phi}_N)] = L(\boldsymbol{\phi}_N)$$

where $L(\boldsymbol{\phi}_N)$ is a $g\times g$ matrix converging to the finite nnd matrix L=$E[h(\mathbf{Y})h(\mathbf{Y})']$ as $N \to \infty$.

(**A5**). The df's $F(\mathbf{y}|\boldsymbol{\phi})$ possess pdf's $f(\mathbf{y}|\boldsymbol{\phi})$ with respect to a σ-finite dominating measure ν. As $N \to \infty$, $f(\mathbf{y}|\boldsymbol{\phi}_N) \to f(\mathbf{y}|\boldsymbol{\phi}_0)$ and $h(\mathbf{y}, \boldsymbol{\phi}_N) \to h(\mathbf{y})$ a.e. (ν).

(**A6**).

$$N^{1/2}(\hat{\theta}_N - \theta_0) = N^{-1/2}\sum_{i=1}^{N} J^{-1}\frac{\partial \log f(\mathbf{Y}_i|\boldsymbol{\phi}_N)}{\partial\theta} + J^{-1}J_{12}\gamma + o_{K^{**}}(1).$$

Here $\hat{\theta}_N$ maximizes $\sum_{i=1}^{N} \log f(\mathbf{Y}_i|\theta)$, J is the information matrix for $F(\mathbf{y}|\theta)$ at θ_0,

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

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

$$J_{12} = E\left[\left(\frac{\partial \log f}{\partial\theta}\right)\left(\frac{\partial \log f}{\partial\boldsymbol{\phi}}\right)'\right].$$

(**A7**). Let $\mathbf{V}_N(\theta, \boldsymbol{\phi}, \varphi)$ be an M-vector and it's σth component is

$$\frac{n_{N\sigma}(\varphi) - NP_\sigma(\theta, \boldsymbol{\phi}, \varphi)}{[NP_\sigma(\theta, , \boldsymbol{\phi}, \varphi)]^{1/2}}.$$

Then

$$N^{1/2}(\bar{\theta} - \theta_0) = (B'B)^{-1}B'\mathbf{V}_N(\boldsymbol{\phi}_N) + (B'B)^{-1}B'B_{12}\gamma + o_{K^{**}}(1),$$

where $\bar{\theta}$ maximizes $\sum_{\sigma=1}^{K^{**}} n_{N\sigma}(\varphi) \log P_\sigma(\theta, \varphi_N)$ and the $K^* \times g$ matrix B_{12} has $(i,\, j)th$ entry

$$P_i^{-1/2}\frac{\partial P_i}{\partial \boldsymbol{\phi}_j}.$$

(**A8**). $g \leq K^*g$ and the matrix with entries $\partial P_i/\partial \theta_j$ has rank g.

(**A9**). **A7** holds, so that $\bar{\theta}_N$ satisfies **A4** with $A = (B'B)^{-1}B'B_{12}$ and $h(\mathbf{y}) = (B'B)^{-1}B'W(\mathbf{y})$, where $\chi_\sigma(\mathbf{y})$ indicator function of $I_\sigma(\varphi_0)$ and $W(\mathbf{y})$ the K^*g vector with σth component $[\chi_\sigma(\mathbf{y}) - P_\sigma]/P_\sigma^{1/2}$.

(**A10**). $\log f(\mathbf{y}|\theta, \boldsymbol{\phi})$ is differentiable with respect to $(\theta, \boldsymbol{\phi})$ at $(\theta_0, \boldsymbol{\phi}_0)$. The matrix J is pd and J_{12} is finite. $(\partial/\partial\theta)F(\mathbf{y}|\theta)$ may be evaluated by differentiating $f(\mathbf{y}|\theta)$ under the integral sign for all $\mathbf{y}$ and $\theta = \theta_0$.

(**A11**). **A7** holds, so that $\hat{\theta}_N$ satisfies **A4** with $A = J^{-1}J_{12}$ and $h(\mathbf{y}) = J^{-1}(\partial \log f(\mathbf{y}|\theta, \boldsymbol{\phi})/\partial\theta)|_{\theta_0, \boldsymbol{\phi}_0}$.

(**A12**). $J - B'B$ is pd.

REFERENCES

1. Green BF. A general solution of the latent class model of latent structure analysis and latent profile analysis. *Psychometrika* 1951; **16**: 151-166.

2. Lazarsfeld PF, Henry NW. *Latent Structure Analysis*. New York: Houghton-Mifflin, 1968.

3. Goodman LA. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* 1974; **61**: 215-231.

4. Haberman SJ. Log-linear models for frequency tables derived by indirect observation: maximum likelihood equations. *AnnualsofStatistics* 1974; **2**: 911-924.

5. Haberman SJ. *Analysis of Qualitative Data. Vol. 2 : New Developments.* New York: Academic Press, 1979.

6. Dayton CM, Macready GB. Concomitant-variable latent-class models. *Journal of the American Statistical Association.* 1988; **83**: 173-178.

7. Formann AK. Linear logistic latent class analysis for polytomous data. *Journal of the American Statistical Association.* 1992; **87**: 476-486.

8. Bandeen-Roche K, Miglioretti DL, Zeger SL, Rathouz PJ. Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association.* 1997; **92**: 1375-1386.

9. Muthén B, Shendden K. Finite mixture modeling with mixture outcomes using EM algorithm. *Biometrics* 1999, **55**: 463-469.

10. Melton B, Liang K-Y, Pulver AE. Extended latent class approach to the study of familial/sporadic forms of a disease: its application to the study of the heterogeneity of schizophrenia. *Genetic Epidemiology* 1994; **11**: 311-327.

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

12. Batholomew DJ. *Latent Variable Models and Factor Analysis.* London: Charles Griffin & Co. Ltd, 1987.

13. Titterington DM, Smith AFM, Makov UE. *Statistical Analysis of Finite Mixture Distributions.* Chichester, U.K.: Wiley, 1985.

14. McCullagh P, Nelder JA. *Generalized Linear Models, 2nd edition.* London: Chapman and Hall, 1989.

15. Agresti A. *Analysis of Categorical Data.* New York: J. Wiley and Sons, 1984.

16. Bandeen-Roche K, Huang GH, Munoz B, Rnbin GS. Determination of risk factor associations with questionnaire outcomes: a methods case study. *American Journal of Epidemiology* 1999; **150**: 1165-1178.

17. Demster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* 1977; **39**: 1-38.

18. Hosmer, D. W., Lemeshow, S. Goodness-of-fit Tests for the Multiple Logistic Regression Model. *Communications in Statistics,* 1980; **A10**: 1043-1069.

19. Lemeshow S., Hosmer D.W.. The Use of Goodness-of-fit Statistics in the Development of Logistic Regression Models. *American Journal of Epidemiology,* 1982; **115**: 92-106.

20. Hosmer DW, Lemeshow S. *Applied Logistic Regression.* New York: John Wiley & Sons.

21. Drust MC. Donsker. Vapnik-Chervonenkis classes and chi-square tests of fit with random cells, 1980; Unpublished doctoral dissertation, Department of Mathematics, M.I.T., Cambridge, MA.

22. Moore D.S., Spruill M.C. Unified Large-sample Theory of General Chi-squared Statistics for Tests of Fit. *Annals of Statistics,* 1975; **3**: 599-616.

23. Huang GH. Selecting the number of classes under latent class regression models: a factor analysis analogous approach. Submitted for publication.

24. West SK, Munoz B, Rubin GS, Schein OD, Bandeen-Roche K, Zeger SL, German PS, Fried LP. Function and visual impairment in a population-based study of older adults: SEE project. *Investigative Ophthalmology and Visual Science* 1997; **38**: 72-82.

25. Mangione CM, Phillips RS, Seddon JM, Lawrence MG, Cook EF, Dailey R, Goldman L. Development of the "activities of daily vision" scale: a measurement of visual functional status. *Medical Care* 1992; **30**: 1111-1126.

26. Valbuena M, Bandeen-Roche K, Rubin GS, Munoz B, West SK, SEE project team. Self-reported assessment of visual functioning in a population based setting. *Investigative Ophthalmology and Visual Science* 1999; **40**: 280-288.

27. Rubin GS, West SK, Munoz B, Bandeen-Roche K, Zeger SL, Scgein O, Fried LP. A comprehensive assessment of visual impairment in an older American population: SEE study. *Investigative Ophthalmology and Visual Science* 1997; **38**: 557-568.

28. Goldberg D. *GHQ The Selection of Psychiatric Illness by Questionnaire.* London: Oxford University Press, 1972.

29. Folstein MF, Folstein SE, McHgh PR. Mini-mental state: a practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*1975; **12**: 189.

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

	1	2	...	g	Total
$y = 0$	n_{01}	n_{02}	...	n_{0g}	n_0
$y = 1$	n_{11}	n_{12}	...	n_{1g}	n_1
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	...	$n_{\bullet g}$	n

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

	1	2	...	g
$\mathbf{y}_1$	O_{11}	O_{12}	...	O_{1g}
$\mathbf{y}_2$	O_{21}	O_{22}	...	O_{2g}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
$\mathbf{y}_{K^*}$	O_{K^*1}	O_{K^*2}	...	O_{K^*g}
	n_1	n_2	...	n_g

Table 3: Simulation results for latent class data were generated equally and unequally

		Mean	Variance	% above 90th %-ile	% above 95th %-ile	% above 99th %-ile
Balance						
g = 2		46.81888	72.56164	58.06453	63.56119	67.00705
g = 5		141.8764	250.6917	160.9970	169.2281	184.6254
g = 10		295.7811	508.613	326.8426	338.3959	347.1322
Unbalance						
g = 2		51.3062	113.7762	62.91475	66.27286	94.70239
g = 5		137.9329	339.7012	158.4159	161.6915	192.0311
g = 10		282.3167	1057.653	316.5099	336.3160	364.9463
Nominal distribution						
g = 2	$\chi^2(47)$	47	94	59.77429	64.00111	72.44331
g = 5	$\chi^2(142)$	142	284	163.9799	170.8092	184.1176
g = 10	$\chi^2(296)$	296	592	327.5783	337.1254	355.5251

Table 4: Generated covariates for alternative models

Situation	z_{im1}	x_{i1}	z_{im2}	x_{i2}
1, 6,11	Ber(0.9)	Ber(0.1)	N(0,15)	N(0,15)
2, 7,12	Ber(0.9)	Ber(0.1)	N(2,15)	N(2,15)
3, 8,13	Ber(0.9)	Ber(0.1)	exp(30)	exp(30)
4, 9,14	Poisson(0.9)	Poisson(0.1)	exp(30)	exp(30)
5,10,15	Poisson(15)	Poisson(15)	exp(30)	exp(30)

Table 5: Simulation results for situations 1-18

			α=0.05
Situation	mean	variance	power
1	140.1411	329.5329	0.07
2	140.9580	311.0997	0.04
3	141.0234	304.7142	0.09
4	139.8990	290.2688	0.07
5	139.4221	258.986	0.06
6	140.6912	304.635	0.06
7	141.3010	254.3497	0.03
8	142.0696	288.4761	0.07
9	140.6636	291.4900	0.08
10	141.1024	294.3013	0.1
11	140.8594	265.5376	0.02
12	140.3471	270.9357	0.04
13	144.2373	274.0058	0.09
14	140.1092	267.8568	0.06
15	140.0726	250.0764	0.03
16	140.7425	241.9495	0.07
17	141.3161	247.0320	0.06
18	144.4393	269.5813	0.09

Table 6: Demographic characteristics and frequency distribution of far vision difficulty items: SEE project (N=2520)

Characteristics	%
Age(year)	
65-69	31.0
70-74	33.1
75-79	22.0
≥ 80	13.9
Gender	
Male	42.1
Female	57.9
Race	
White	73.6
African American	26.4
Education (year)	
< 7	8.2
7-11	43.3
12	20.4
> 12	28.1
MMSEscore	
< 24	16.2
24-29	65.4
30	18.4
GHQ depression score	
0	90.5
1-2	6.8
≥ 3	2.7
Number of comorbid diseases	
≤ 1	31.4
2-3	47.0
4-5	17.5
≥ 6	4.1

Frequency distribution	Degree of difficulty (%)			
Activities	extreme diff.	a little diff.	having diff.	no diff.
signs-day	–	–	18.3	81.7
signs-night	16.2	26.4	–	57.4
step-day	–	–	11.5	88.5
steps-dim	–	–	18.6	81.4
watch TV	–	–	10.4	89.6

Table 7: Contingency table of the expected frequencies for the five-class LCR model of SEE project

Response	Group				
pattern	1	2	3	4	5
$\mathbf{y}_1$	0.42049628	0.99968124	1.81263384	3.69370896	29.72270173
$\mathbf{y}_2$	0.18643264	0.27539497	0.39021325	0.57620746	0.99525688
$\mathbf{y}_3$	0.11119960	0.33222896	0.51137721	0.76755720	1.48092621
$\mathbf{y}_4$	0.29903223	0.41844374	0.56736963	0.81923390	1.45762043
$\mathbf{y}_5$	0.06300168	0.10078633	0.14188887	0.21270749	0.50788740
$\mathbf{y}_6$	0.21401378	0.21398787	0.25865891	0.36033524	0.53801246
$\mathbf{y}_7$	0.26104599	0.67933664	1.06554182	1.68858513	6.17137860
$\mathbf{y}_8$	0.17457210	0.23507525	0.31356743	0.44322213	0.76267136
$\mathbf{y}_9$	0.16607883	0.44464038	0.67558871	1.01762602	1.88077908
$\mathbf{y}_{10}$	0.38645248	0.54911854	0.70523430	0.94257781	1.37228041
$\mathbf{y}_{11}$	0.04200933	0.05644380	0.07398084	0.10387555	0.26671268
$\mathbf{y}_{12}$	0.27667556	0.29077849	0.32698076	0.41438756	0.58527210
$\mathbf{y}_{13}$	0.09097031	0.15926094	0.24235955	0.41409407	2.46833773
$\mathbf{y}_{14}$	0.05874608	0.06395631	0.07767421	0.10047405	0.13134480
$\mathbf{y}_{15}$	0.02980672	0.05806012	0.07983519	0.10911793	0.15647378
$\mathbf{y}_{16}$	0.11552310	0.14904297	0.18270672	0.22112555	0.26985210
$\mathbf{y}_{17}$	0.01862574	0.01941415	0.02351588	0.03066136	0.07072115
$\mathbf{y}_{18}$	0.08751291	0.08388996	0.09421222	0.11162741	0.13523778
$\mathbf{y}_{19}$	0.63768755	1.14340365	1.58956562	2.15151985	3.84862639
$\mathbf{y}_{20}$	0.80648470	0.80947657	0.91017970	1.12810305	1.68052656
$\mathbf{y}_{21}$	0.75838308	1.10924521	1.43204584	1.90607478	2.82379096
$\mathbf{y}_{22}$	1.85887290	1.85283452	1.98053637	2.29910468	2.97559958
$\mathbf{y}_{23}$	0.16384062	0.16270179	0.18390915	0.22323372	0.37089297
$\mathbf{y}_{24}$	2.44569602	2.10736084	1.98915184	1.98861364	1.83753862

Response pattern	Group 1	2	3	4	5
$\mathbf{y}_{25}$	0.54042227	1.73156563	2.81586975	4.15098379	10.27999890
$\mathbf{y}_{26}$	1.09828562	1.56198456	2.15043059	3.01754843	4.44128717
$\mathbf{y}_{27}$	0.60144036	1.60463972	2.40415869	3.31413678	4.71608401
$\mathbf{y}_{28}$	1.79694626	2.93859969	4.04802492	5.24541618	7.40559661
$\mathbf{y}_{29}$	0.32890596	0.47399871	0.63245982	0.85320599	1.13644414
$\mathbf{y}_{30}$	1.26454454	1.44735271	1.82453947	2.38028614	3.02789299
$\mathbf{y}_{31}$	0.85062691	2.65441513	4.26555237	6.09161241	10.728768035
$\mathbf{y}_{32}$	1.15130224	1.69965419	2.26762283	2.99815672	4.51704113
$\mathbf{y}_{33}$	0.91258949	2.22682930	3.28979151	4.52216227	7.06590127
$\mathbf{y}_{34}$	3.94518861	6.69430173	8.72773722	10.55980526	12.50051423
$\mathbf{y}_{35}$	0.22046238	0.27909237	0.35011903	0.45957534	0.94005296
$\mathbf{y}_{36}$	4.37773774	5.29408268	5.97981629	6.82891761	7.61818812
$\mathbf{y}_{37}$	0.13023080	0.30000679	0.43644729	0.59433628	1.12270149
$\mathbf{y}_{38}$	0.34588037	0.42898907	0.53640686	0.66122105	0.77008515
$\mathbf{y}_{39}$	0.16235453	0.28784858	0.38270096	0.49530050	0.62961065
$\mathbf{y}_{40}$	0.74436805	1.48056945	1.95359229	2.17974196	2.25137670
$\mathbf{y}_{41}$	0.09688252	0.09316940	0.10517277	0.12963315	0.15425575
$\mathbf{y}_{42}$	0.53769469	0.72674833	0.89919330	1.02940472	1.09668809
$\mathbf{y}_{43}$	3.45079913	5.78162334	7.70581607	9.90671502	13.73782580
$\mathbf{y}_{44}$	7.54361584	10.63312563	11.80221091	12.53484838	13.67550209
$\mathbf{y}_{45}$	4.26883963	5.88123132	7.22737492	9.06883811	12.62221788
$\mathbf{y}_{46}$	40.06027971	54.67193447	56.43316835	54.42342720	43.95310627
$\mathbf{y}_{47}$	1.12298026	1.40463204	1.51398265	1.64979597	2.36773250
$\mathbf{y}_{48}$	242.77446186	205.38904195	184.61908324	163.18115618	99.73068633

Table 8: Contingency table of the observed frequencies for the five-class LCR model of SEE project

Response pattern	Group 1	2	3	4	5
$\mathbf{y}_1$	1	2	1	2	30
$\mathbf{y}_2$	0	1	0	2	2
$\mathbf{y}_3$	0	1	0	1	1
$\mathbf{y}_4$	1	0	0	1	3
$\mathbf{y}_5$	1	0	0	0	0
$\mathbf{y}_6$	0	1	0	0	0
$\mathbf{y}_7$	0	1	1	1	8
$\mathbf{y}_8$	0	0	0	1	1
$\mathbf{y}_9$	0	0	1	2	0
$\mathbf{y}_{10}$	0	0	0	2	2
$\mathbf{y}_{11}$	0	0	0	0	0
$\mathbf{y}_{12}$	0	0	0	0	1
$\mathbf{y}_{13}$	0	0	1	0	1
$\mathbf{y}_{14}$	0	0	0	0	0
$\mathbf{y}_{15}$	0	0	0	1	0
$\mathbf{y}_{16}$	0	0	0	1	0
$\mathbf{y}_{17}$	0	0	0	1	0
$\mathbf{y}_{18}$	0	0	0	0	0
$\mathbf{y}_{19}$	0	0	0	4	9
$\mathbf{y}_{20}$	0	0	1	0	1
$\mathbf{y}_{21}$	0	0	1	4	2
$\mathbf{y}_{22}$	3	3	4	1	3
$\mathbf{y}_{23}$	1	0	0	0	0
$\mathbf{y}_{24}$	4	2	0	2	1

Response pattern	Group 1	2	3	4	5
$\mathbf{y}_{25}$	1	4	3	3	14
$\mathbf{y}_{26}$	0	0	2	2	4
$\mathbf{y}_{27}$	0	2	4	1	4
$\mathbf{y}_{28}$	2	3	4	7	8
$\mathbf{y}_{29}$	0	0	0	0	0
$\mathbf{y}_{30}$	2	0	3	5	7
$\mathbf{y}_{31}$	1	5	1	5	15
$\mathbf{y}_{32}$	2	1	3	0	6
$\mathbf{y}_{33}$	0	4	5	6	5
$\mathbf{y}_{34}$	6	3	7	18	6
$\mathbf{y}_{35}$	0	0	1	0	1
$\mathbf{y}_{36}$	5	8	3	3	7
$\mathbf{y}_{37}$	0	1	1	0	2
$\mathbf{y}_{38}$	0	0	0	0	0
$\mathbf{y}_{39}$	0	0	1	0	1
$\mathbf{y}_{40}$	1	1	0	0	4
$\mathbf{y}_{41}$	0	0	0	1	0
$\mathbf{y}_{42}$	0	2	0	0	1
$\mathbf{y}_{43}$	5	9	2	11	12
$\mathbf{y}_{44}$	7	13	9	13	15
$\mathbf{y}_{45}$	4	7	6	6	13
$\mathbf{y}_{46}$	38	55	62	47	47
$\mathbf{y}_{47}$	1	2	1	1	3
$\mathbf{y}_{48}$	242	197	200	173	89

Table 9: Contingency table of the expected frequencies for the five-class LCR model of SEE project after combining the rows which were less than 5

Response pattern	Group 1	2	3	4	5
$\mathbf{y}'_1$	5.20475070	8.19166609	11.47513052	17.21282251	57.32641060
$\mathbf{y}'_2$	5.00883181	5.85446277	6.96946711	8.66193583	15.46403006
$\mathbf{y}'_3$	5.09012273	8.02657539	11.05961349	14.81059352	20.72730491
$\mathbf{y}'_4$	6.85970724	13.27520034	18.55070393	24.17173667	34.81222466
$\mathbf{y}'_5$	5.07431128	6.30217091	7.30278947	8.54405028	10.45102772
$\mathbf{y}'_6$	12.53571477	19.00308472	22.84868631	26.27564373	31.54525907
$\mathbf{y}'_7$	44.32911934	60.55316579	63.66054327	63.49226532	56.57532415
$\mathbf{y}'_8$	243.89744211	206.79367399	186.13306590	164.83095215	102.09841883

Table 10: Contingency table of the observed frequencies for the five-class LCR model of SEE project after combining the rows of expected frequencies table which were less than 5

Response pattern	Group 1	2	3	4	5
$\mathbf{y}'_1$	2	7	6	22	62
$\mathbf{y}'_2$	9	8	8	5	19
$\mathbf{y}'_3$	3	6	15	15	21
$\mathbf{y}'_4$	9	13	13	31	33
$\mathbf{y}'_5$	5	8	6	4	9
$\mathbf{y}'_6$	13	26	10	26	33
$\mathbf{y}'_7$	42	64	66	54	59
$\mathbf{y}'_8$	245	196	204	171	93

Table 11: Contingency table of the expected frequencies for the four-class LCR model of SEE project

Response pattern	Group 1	2	3	4	5
$\mathbf{y}_1$	0.33253387	0.88876030	1.72617907	3.70358088	29.82976505
$\mathbf{y}_2$	0.12301973	0.29909110	0.45631283	0.63958536	0.93435164
$\mathbf{y}_3$	0.09338816	0.33887229	0.50027383	0.74302915	1.34521607
$\mathbf{y}_4$	0.29088420	0.58132585	0.79495813	1.17422478	1.49356677
$\mathbf{y}_5$	0.04996797	0.10351446	0.14507650	0.20892670	0.31921990
$\mathbf{y}_6$	0.21421744	0.27089226	0.34000770	0.47852932	0.63182957
$\mathbf{y}_7$	0.32521086	0.64098222	1.00166670	1.63079331	7.57719802
$\mathbf{y}_8$	0.23988203	0.33552952	0.45431556	0.65294843	1.12506714
$\mathbf{y}_9$	0.18425023	0.40953031	0.58268994	0.92654563	1.83566779
$\mathbf{y}_{10}$	0.54745725	0.60195163	0.69168003	0.95225818	1.44328960
$\mathbf{y}_{11}$	0.07806263	0.08085038	0.09054321	0.12065982	0.18556317
$\mathbf{y}_{12}$	0.40141747	0.32906583	0.31712040	0.37208971	0.51899129
$\mathbf{y}_{13}$	0.11497551	0.19533397	0.31431747	0.54472756	3.38344896
$\mathbf{y}_{14}$	0.04865135	0.06449317	0.07778130	0.09975095	0.12446216
$\mathbf{y}_{15}$	0.03083251	0.05871804	0.07084292	0.09545449	0.11064886
$\mathbf{y}_{16}$	0.13405810	0.15558773	0.16793868	0.22338508	0.26777677
$\mathbf{y}_{17}$	0.02121085	0.02436612	0.02713968	0.03467698	0.04114157
$\mathbf{y}_{18}$	0.11628616	0.10593940	0.10703523	0.13178468	0.17296853
$\mathbf{y}_{19}$	0.51431176	0.87882308	1.20544306	1.64677892	3.44862179
$\mathbf{y}_{20}$	0.81644864	0.87137810	1.06146977	1.32381667	2.00307882
$\mathbf{y}_{21}$	0.66508082	0.95165435	1.18439411	1.62976061	2.34866568
$\mathbf{y}_{22}$	1.78011284	1.77966199	1.98442276	2.28541383	3.19828556
$\mathbf{y}_{23}$	0.16019129	0.14357463	0.15698200	0.18779582	0.27001319
$\mathbf{y}_{24}$	2.13143014	1.82902966	1.76025016	1.76125173	1.57661959

Response pattern	Group 1	2	3	4	5
$\mathbf{y}_{25}$	0.35008546	1.35659933	2.20562021	3.20297809	9.50318998
$\mathbf{y}_{26}$	0.74971719	1.70878508	2.52018322	3.28079288	4.22723731
$\mathbf{y}_{27}$	0.54646183	1.89329414	2.68646415	3.50791726	4.35527884
$\mathbf{y}_{28}$	1.73592421	3.20052301	4.26688494	5.78982934	6.94303779
$\mathbf{y}_{29}$	0.30625298	0.58797149	0.78197613	1.00356313	1.20707350
$\mathbf{y}_{30}$	1.31455773	1.52752317	1.89741854	2.48843086	3.22194006
$\mathbf{y}_{31}$	0.82709725	2.42881286	3.78911593	5.58826854	11.03174882
$\mathbf{y}_{32}$	1.63680029	2.31386936	3.11301295	4.16099641	6.48235890
$\mathbf{y}_{33}$	1.23046411	2.67246263	3.68526124	5.21172681	7.75399850
$\mathbf{y}_{34}$	3.89644176	5.06612099	5.97139292	7.33903102	9.86007980
$\mathbf{y}_{35}$	0.50905302	0.51441872	0.56835118	0.69374280	0.97892556
$\mathbf{y}_{36}$	4.89962474	6.35564147	6.95464367	7.35305454	7.59107044
$\mathbf{y}_{37}$	0.10765676	0.22050313	0.29985853	0.39987288	1.09156276
$\mathbf{y}_{38}$	0.30021062	0.38517418	0.46821588	0.57362767	0.75073846
$\mathbf{y}_{39}$	0.18316997	0.32686049	0.38314876	0.47461805	0.51148020
$\mathbf{y}_{40}$	0.82563371	1.01004948	1.15725911	1.44404737	1.77213408
$\mathbf{y}_{41}$	0.13074221	0.14082310	0.15350743	0.18209205	0.21774938
$\mathbf{y}_{42}$	0.85496731	1.04591266	1.22574937	1.45074646	1.76654086
$\mathbf{y}_{43}$	3.09297696	5.32888010	6.96298252	8.91537464	11.97127840
$\mathbf{y}_{44}$	7.69814344	9.91887907	11.55472314	12.48776853	15.04856313
$\mathbf{y}_{45}$	4.80917652	6.59156023	7.89861846	9.89917022	12.64121238
$\mathbf{y}_{46}$	40.52719260	57.36985264	58.14678568	56.11497384	44.71831148
$\mathbf{y}_{47}$	1.44144168	1.50706755	1.56976274	1.64623074	1.95424084
$\mathbf{y}_{48}$	240.6123	202.5895	184.5202	163.2234	99.21479

Table 12: Contingency table of the observed frequencies for the four-class LCR model of SEE project

Response	Group				
pattern	1	2	3	4	5
$\mathbf{y}_1$	3	0	1	3	29
$\mathbf{y}_2$	0	1	0	1	3
$\mathbf{y}_3$	0	0	1	1	1
$\mathbf{y}_4$	0	1	0	2	2
$\mathbf{y}_5$	0	1	0	0	0
$\mathbf{y}_6$	0	1	0	0	0
$\mathbf{y}_7$	0	1	0	2	8
$\mathbf{y}_8$	0	0	1	0	1
$\mathbf{y}_9$	0	0	1	2	0
$\mathbf{y}_{10}$	0	0	0	2	2
$\mathbf{y}_{11}$	0	0	0	0	0
$\mathbf{y}_{12}$	0	0	0	0	1
$\mathbf{y}_{13}$	0	1	0	0	1
$\mathbf{y}_{14}$	0	0	0	0	0
$\mathbf{y}_{15}$	0	0	0	1	0
$\mathbf{y}_{16}$	0	0	0	1	0
$\mathbf{y}_{17}$	0	0	0	1	0
$\mathbf{y}_{18}$	0	0	0	0	0
$\mathbf{y}_{19}$	0	0	0	4	9
$\mathbf{y}_{20}$	0	0	1	0	1
$\mathbf{y}_{21}$	0	1	0	3	3
$\mathbf{y}_{22}$	3	3	2	3	3
$\mathbf{y}_{23}$	1	0	0	0	0
$\mathbf{y}_{24}$	5	1	0	1	2

Response	Group				
pattern	1	2	3	4	5
$\mathbf{y}_{25}$	1	4	2	5	13
$\mathbf{y}_{26}$	0	0	3	1	4
$\mathbf{y}_{27}$	0	2	3	2	4
$\mathbf{y}_{28}$	2	3	4	8	7
$\mathbf{y}_{29}$	0	0	0	0	0
$\mathbf{y}_{30}$	1	1	5	5	5
$\mathbf{y}_{31}$	0	4	3	6	14
$\mathbf{y}_{32}$	1	2	3	0	6
$\mathbf{y}_{33}$	0	4	4	6	6
$\mathbf{y}_{34}$	5	5	8	15	7
$\mathbf{y}_{35}$	0	0	1	0	1
$\mathbf{y}_{36}$	7	6	3	4	6
$\mathbf{y}_{37}$	0	1	1	0	2
$\mathbf{y}_{38}$	0	0	0	0	0
$\mathbf{y}_{39}$	0	1	0	1	0
$\mathbf{y}_{40}$	1	1	0	0	4
$\mathbf{y}_{41}$	0	0	0	1	0
$\mathbf{y}_{42}$	0	1	1	0	1
$\mathbf{y}_{43}$	4	9	3	10	13
$\mathbf{y}_{44}$	8	10	12	14	13
$\mathbf{y}_{45}$	6	5	6	7	12
$\mathbf{y}_{46}$	41	52	59	49	48
$\mathbf{y}_{47}$	1	2	1	1	3
$\mathbf{y}_{48}$	238	204	199	166	94

Table 13: Contingency table of the expected frequencies for the four-class LCR model of SEE project after combining the rows which were less than 5

Response	Group				
pattern	1	2	3	4	5
$\mathbf{y}_1'$	5.342148	8.186660	11.317186	17.333307	59.140539
$\mathbf{y}_2'$	5.171537	6.817651	8.627458	10.718232	18.775346
$\mathbf{y}_3'$	6.367094	11.951994	16.534873	22.539006	33.241438
$\mathbf{y}_4'$	5.126906	7.738584	9.656654	12.550758	17.614078
$\mathbf{y}_5'$	5.408678	6.870060	7.522995	8.046797	8.569996
$\mathbf{y}_6'$	5.495358	8.458203	10.650722	13.440379	18.081484
$\mathbf{y}_7'$	7.698143	9.918879	11.554723	12.487769	15.048563
$\mathbf{y}_8'$	45.33637	63.96141	66.04540	66.01414	57.35952
$\mathbf{y}_9'$	242.0538	204.0966	186.0900	164.8696	101.1690

Table 14: Contingency table of the observed frequencies for the four-class LCR model of SEE project after combining the rows of expected frequencies table which were less than 5

Response	Group				
pattern	1	2	3	4	5
$\mathbf{y}_1'$	3	7	5	23	61
$\mathbf{y}_2'$	10	8	7	10	22
$\mathbf{y}_3'$	4	12	18	21	36
$\mathbf{y}_4'$	5	9	12	21	13
$\mathbf{y}_5'$	7	6	4	4	7
$\mathbf{y}_6'$	5	13	5	12	20
$\mathbf{y}_7'$	8	10	12	14	13
$\mathbf{y}_8'$	47	57	65	56	60
$\mathbf{y}_9'$	239	206	200	167	97

Table 15: Contingency table of the expected frequencies for the three-class LCR model of SEE project

Response	Group				
pattern	1	2	3	4	5
$\mathbf{y}_1$	0.32405784	0.81798275	1.59660503	3.51864062	26.9483264
$\mathbf{y}_2$	0.06631802	0.16394991	0.25634506	0.46729546	1.6617218
$\mathbf{y}_3$	0.05371077	0.21023411	0.28972742	0.51072580	1.2754389
$\mathbf{y}_4$	0.08941671	0.20460213	0.29480615	0.50884533	0.9026324
$\mathbf{y}_5$	0.01755697	0.03783496	0.05666595	0.09876624	0.3019590
$\mathbf{y}_6$	0.03609294	0.04793101	0.06941731	0.10896863	0.1924848
$\mathbf{y}_7$	0.36756824	0.80212273	1.28023853	2.30544252	11.1359594
$\mathbf{y}_8$	0.26580013	0.43994338	0.60041050	0.90439845	1.8380178
$\mathbf{y}_9$	0.22004397	0.55432952	0.72276488	1.11827653	1.8701467
$\mathbf{y}_{10}$	0.48808858	0.68373037	0.91038068	1.30985962	1.9921334
$\mathbf{y}_{11}$	0.08047676	0.11233530	0.14895868	0.21518359	0.3922728
$\mathbf{y}_{12}$	0.23956990	0.24238449	0.31496340	0.40068843	0.6275601
$\mathbf{y}_{13}$	0.17243752	0.33990160	0.56193203	0.99269358	4.7531580
$\mathbf{y}_{14}$	0.10984894	0.16941069	0.23219291	0.34475372	0.7114983
$\mathbf{y}_{15}$	0.08858420	0.20995924	0.27181291	0.41062835	0.5975988
$\mathbf{y}_{16}$	0.21202841	0.27437334	0.36632562	0.52232204	0.7443654
$\mathbf{y}_{17}$	0.03408459	0.04419635	0.05886210	0.08294078	0.1481046
$\mathbf{y}_{18}$	0.10183665	0.09173190	0.12377457	0.15243051	0.2091379
$\mathbf{y}_{19}$	0.35816064	0.57949384	0.78205364	1.13307404	3.2120391
$\mathbf{y}_{20}$	0.64755337	0.69458258	0.83686299	1.01293982	1.4152484
$\mathbf{y}_{21}$	0.47504240	0.66757800	0.80626637	1.04835012	1.3248982
$\mathbf{y}_{22}$	1.51972616	1.57289672	1.85368123	2.12663334	2.5479860
$\mathbf{y}_{23}$	0.20877046	0.19071292	0.22859238	0.26357337	0.3393000
$\mathbf{y}_{24}$	2.16976587	1.91850717	2.02445231	2.02797815	1.8753881

Response pattern	Group				
	1	2	3	4	5
$\mathbf{y}_{25}$	0.34618763	1.24966160	1.87897444	3.09709248	8.8822582
$\mathbf{y}_{26}$	0.41998864	0.99242695	1.45233677	2.22456248	3.5973526
$\mathbf{y}_{27}$	0.41622423	1.44323607	2.03034090	3.06023537	4.9028305
$\mathbf{y}_{28}$	0.72897336	1.43692742	2.04567192	3.10515507	4.6623652
$\mathbf{y}_{29}$	0.12262888	0.24403947	0.33893170	0.49490892	0.7610699
$\mathbf{y}_{30}$	0.30753832	0.37095790	0.50106892	0.71078433	1.0286178
$\mathbf{y}_{31}$	1.18058453	3.07399241	4.12491926	5.88679677	9.7491025
$\mathbf{y}_{32}$	2.29924369	3.47367270	4.49002260	6.02125844	8.2445501
$\mathbf{y}_{33}$	1.82860813	4.00924885	5.20820143	7.10716757	9.4369980
$\mathbf{y}_{34}$	4.58041870	6.34962584	7.82626714	10.06252554	13.2375753
$\mathbf{y}_{35}$	0.69721477	0.88501791	1.09892286	1.39996599	1.8116868
$\mathbf{y}_{36}$	4.41785669	5.52047584	6.00410260	6.72210412	7.5024807
$\mathbf{y}_{37}$	0.44533935	1.08344795	1.43421384	1.98575668	3.2063911
$\mathbf{y}_{38}$	0.88417430	1.22496464	1.57919151	2.07845882	2.7209311
$\mathbf{y}_{39}$	0.71366107	1.46137652	1.88348772	2.52648921	3.2242573
$\mathbf{y}_{40}$	1.74918913	2.07598775	2.64761156	3.41781722	4.3899174
$\mathbf{y}_{41}$	0.27952619	0.32699495	0.40740409	0.50711334	0.6259024
$\mathbf{y}_{42}$	0.97502204	0.91612645	1.10491131	1.27840021	1.4815578
$\mathbf{y}_{43}$	2.26397726	3.44103394	4.06401867	4.96421676	5.9669476
$\mathbf{y}_{44}$	8.32469842	10.80300523	11.37814988	12.17386995	11.9393512
$\mathbf{y}_{45}$	4.12665914	5.31976653	6.16510334	7.26372174	8.1203980
$\mathbf{y}_{46}$	41.14785545	56.72699335	55.20611448	54.27339198	42.0665447
$\mathbf{y}_{47}$	2.29350821	2.33661203	2.47253783	2.57161914	2.5536178
$\mathbf{y}_{48}$	239.1044	202.1637	187.9694	163.4812	101.8699

Table 16: Contingency table of the observed frequencies for the three-class LCR model of SEE project

Response pattern	Group 1	2	3	4	5	Response pattern	Group 1	2	3	4	5
$\mathbf{y}_1$	2	1	0	4	29	$\mathbf{y}_{25}$	1	5	3	3	13
$\mathbf{y}_2$	0	1	0	1	3	$\mathbf{y}_{26}$	0	0	3	1	4
$\mathbf{y}_3$	0	0	1	1	1	$\mathbf{y}_{27}$	0	3	2	2	4
$\mathbf{y}_4$	0	1	0	1	3	$\mathbf{y}_{28}$	3	2	5	6	8
$\mathbf{y}_5$	0	1	0	0	0	$\mathbf{y}_{29}$	0	0	0	0	0
$\mathbf{y}_6$	0	1	0	0	0	$\mathbf{y}_{30}$	1	1	4	6	5
$\mathbf{y}_7$	0	1	1	2	7	$\mathbf{y}_{31}$	2	3	2	6	14
$\mathbf{y}_8$	0	0	0	1	1	$\mathbf{y}_{32}$	1	3	2	0	6
$\mathbf{y}_9$	0	0	1	2	0	$\mathbf{y}_{33}$	0	4	3	6	7
$\mathbf{y}_{10}$	0	0	0	2	2	$\mathbf{y}_{34}$	4	4	8	18	6
$\mathbf{y}_{11}$	0	0	0	0	0	$\mathbf{y}_{35}$	0	0	1	0	1
$\mathbf{y}_{12}$	0	0	0	0	1	$\mathbf{y}_{36}$	8	5	4	2	7
$\mathbf{y}_{13}$	0	0	1	0	1	$\mathbf{y}_{37}$	0	1	1	0	2
$\mathbf{y}_{14}$	0	0	0	0	0	$\mathbf{y}_{38}$	0	0	0	0	0
$\mathbf{y}_{15}$	0	0	0	1	0	$\mathbf{y}_{39}$	0	0	1	0	1
$\mathbf{y}_{16}$	0	0	0	0	1	$\mathbf{y}_{40}$	1	1	0	1	3
$\mathbf{y}_{17}$	0	0	0	1	0	$\mathbf{y}_{41}$	0	0	0	1	0
$\mathbf{y}_{18}$	0	0	0	0	0	$\mathbf{y}_{42}$	0	1	1	0	1
$\mathbf{y}_{19}$	0	0	0	4	9	$\mathbf{y}_{43}$	4	10	4	11	10
$\mathbf{y}_{20}$	0	0	1	0	1	$\mathbf{y}_{44}$	7	14	8	14	14
$\mathbf{y}_{21}$	0	1	0	3	3	$\mathbf{y}_{45}$	4	6	8	5	13
$\mathbf{y}_{22}$	1	5	2	2	4	$\mathbf{y}_{46}$	39	54	62	47	47
$\mathbf{y}_{23}$	1	0	0	0	0	$\mathbf{y}_{47}$	1	2	1	2	2
$\mathbf{y}_{24}$	6	0	0	1	2	$\mathbf{y}_{48}$	242	197	198	171	93

Table 17: Contingency table of the expected frequencies for the three-class LCR model of SEE project after combining the rows which were less than 5

Response	Group				
pattern	1	2	3	4	5
$\mathbf{y}_1'$	5.968004	8.961505	12.435048	19.293858	64.802688
$\mathbf{y}_2'$	5.900662	10.920462	14.625289	20.871087	35.798285
$\mathbf{y}_3'$	8.708271	13.832547	17.524491	23.190952	30.919123
$\mathbf{y}_4'$	5.115071	6.405494	7.103025	8.122070	9.314167
$\mathbf{y}_5'$	5.046912	7.088898	9.056820	11.794035	15.648957
$\mathbf{y}_6'$	10.58868	14.24404	15.44217	17.13809	17.90630
$\mathbf{y}_7'$	43.12666	59.31977	68.16510	54.26372	55.12040
$\mathbf{y}_8'$	241.3979	204.5003	190.4419	166.0528	104.4235

Table 18: Contingency table of the observed frequencies for the three-class LCR model of SEE project after combining the rows of expected frequencies table which were less than 5

Response	Group				
pattern	1	2	3	4	5
$\mathbf{y}_1'$	3	12	7	25	66
$\mathbf{y}_2'$	14	14	19	25	50
$\mathbf{y}_3'$	5	11	13	24	19
$\mathbf{y}_4'$	8	5	5	2	8
$\mathbf{y}_5'$	1	3	3	2	7
$\mathbf{y}_6'$	11	24	12	25	24
$\mathbf{y}_7'$	43	60	70	52	60
$\mathbf{y}_8'$	47	57	65	56	60
$\mathbf{y}_9'$	243	199	199	173	95