A Modified Model-Selection Criteria in a Generalised Estimating Equation for Latent Class Regression Models

Jerry Dwi Trijoyo Purnomo∗**, Chih-Rung Chen and Guan-Hua Huang**

Department of Statistics, Faculty of Mathematics, Computing, and Data Science
Institut Teknologi Sepuluh Nopember, 60111 Surabaya, Indonesia
∗Corresponding author: jerry@statistika.its.ac.id

Abstract In recent years, generalised estimating equations (GEEs) have played an important role in many fields of research, such as biomedicine. In this paper, we use GEEs for latent class regression (LCR) with covariate effects on underlying and measured variables. However, there are only a few model-selection criteria in GEEs. The widely known Akaike information criterion (AIC) cannot be used directly, since AIC is a full likelihood-based model, whereas GEEs are nonlikelihood based. Hence, we propose a modification to AIC in GEEs for LCR models, where the likelihood is replaced by quasi-likelihood, and a proper adjustment is made by giving a penalty term. The data of the modified hospital elder life program (mHELP) project are used to illustrate our method.

Keywords Generalised Estimating Equation; Latent Class Regression; Quasi-likelihood; Akaike information criterion; biomedicine

Mathematics Subject Classification 62H20, 62J12, 62P10

1 Introduction

In the latent variable framework, there are two types of variables to be considered. These are the variables that can be measured directly, also known as manifest variables, and the variables that are inferred from measured variables and cannot be measured directly, known as latent variables. The measured and latent variables can be discrete or continuous. In this paper, we concentrate on measured variables and latent variables that are categorical, called a latent class analysis (LCA). The LCA model was initially proposed by Green [1]. Goodman [9] provided a beneficial theory for LCA and developed a maximum likelihood (ML) approach to estimate model parameters as well. In recent years, LCA is admitting a lot of attention in many fields of research. For instance, LCA has been used in psychosocial (Pickles *et al.* [10]; Hudziak *et al.* [11]; Garrett and Zeger [12]; Neuman *et al.* [13]; Nylund *et al.* [14]; Li and Lee [15]; and Larance *et al.* [16]) and in medical research (Moustaki [17]; Sullivan *et al.* [18]; Bandeen-Roche *et al.* [19]; Klonsky and Olino [20]; and Neumann *et al.* [21]). When

a set of covariates is available, the LCA model can be extended to incorporate the covariate associated with the class membership, as well as to build a direct relationship with measured responses. This is a latent class regression (LCR) model as developed by Clogg and Goodman [22] and [23], Forman [24], Dayton and Macready [25], Forman [26], Hagenaars [27], Melton, Liang, and Pulver [28], Van Der Heijden, Dessens, and Bockenholt [29], Bandeen-Roche *et al.* [19], Leisch [30], Huang and Bandeen-Rochee [31], and Huang [32]. It is necessary to build LCR models under the conditional independence assumption, where the responses are independent given class membership. However, this assumption is not necessarily true and can be empirically checked through analyses stratifying on inferred class memberships. This is problematic since these inferred latent classes can be wrong if the model assumptions are violated. The use of generalised estimating equations (GEEs) is fast becoming the solution for such a problem, since this approach is built by constructing a working covariance structure to relax the dependency among variables. Unfortunately, there are a few model-selection criteria in GEE. The likelihood-based criteria, such as AIC, cannot be used directly in the GEEs approach since they are nonlikelihood based. Hence, we propose modified model-selection criteria for GEEs, particularly for LCR models. These criteria use quasi-likelihood instead of full likelihood to run the procedure.

2   Method

In this section, we briefly review the literature and illustrate the proposed method used in this paper.

2.1   Latent Class Regression (LCR)

To postulate the LCR model, the covariates of interest are incorporated into the LCA model. Let $(\mathbf{x}_i, \mathbf{z}_i)$ be the concomitant covariates of the ith individual for $i = 1, \ldots, n$, where $\mathbf{x}_i = (x_{i0}, x_{i1}, \ldots, x_{ip})^T, x_{i0} = 1$, are covariates assumed to be associated with the ith latent class membership, S_i, and $\mathbf{z}_i = (\mathbf{z}_{i1}, \ldots, \mathbf{z}_{iM})^T$ with $z_{im} = (z_{im1}, \ldots, z_{imL_m})^T$, and $m = 1, \ldots, M$ are covariates used to determine the direct effects on the response given the latent class membership. The covariates may include any combination of continuous and/or discrete measures and could be mutually exclusive or overlap. The regression extension of latent class analysis, called the LCR model, is defined as in equation (1):

$$P\left(\{Y_{i1} = y_1, \ldots, Y_{iM} = y_M\} \mid \mathbf{x}_i, \mathbf{z}_i\right)$$
$$= \sum_{j=1}^{J}\left\{\prod_{m=1}^{M}\prod_{k=1}^{K_m}\left[p_{mkj}\left(\boldsymbol{\gamma}_{mj} + \mathbf{z}_{im}^T\boldsymbol{\alpha}_m\right)\right]^{1_{\{y_m=k\}}}\right\}\eta_j\left(\mathbf{x}_i^T\boldsymbol{\beta}\right), \quad (1)$$

with $\eta_j\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)$ and $p_{mkj}\left(\boldsymbol{\gamma}_{mj} + \mathbf{z}_{im}^T\boldsymbol{\alpha}_m\right)$ being defined as in the generalised linear model (GLM) framework proposed by McCullagh and Nelder [2]. However, Agresti [3] stated that the logit link function was commonly adopted in GLMs, hence, we specify $\eta_j\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)$ and $p_{mkj}\left(\boldsymbol{\gamma}_{mj} + \mathbf{z}_{im}^T\boldsymbol{\alpha}_m\right)$ as in equation (2) and (3), respectively:

$$p_{imkj} = p_{mkj}\left(\boldsymbol{\gamma}_{mj} + \mathbf{z}_{im}^T\boldsymbol{\alpha}_m\right) = \frac{\exp\left(\gamma_{mkj} + \alpha_{1mk}z_{im1} + \cdots + \alpha_{L_mmk}z_{imL_m}\right)}{1 + \sum\limits_{k'=1}^{K_m-1}\left(\gamma_{mk'j} + \alpha_{1mk'}z_{im1} + \cdots + \alpha_{L_mmk'}z_{imL_m}\right)} \quad (2)$$

for $i = 1, \ldots, n$; $m = 1, \ldots, M$; $k = 1, \ldots, (K_m - 1)$; $j = 1, \ldots, J$, and

$$\eta_{ij} = \eta_j \left(\mathbf{x}_i^T \boldsymbol{\beta} \right) = \frac{\exp \left(\beta_{0j} + \beta_{1j} x_{i1} + \cdots + \beta_{Pj} x_{iP} \right)}{1 + \sum_{j'=1}^{J-1} \exp \left(\beta_{0j'} + \beta_{1j'} x_{i1} + \cdots + \beta_{Pj'} x_{iP} \right)} \tag{3}$$

for $i = 1, \ldots, n$; $j = 1, \ldots, J - 1$.

2.2 Generalised Estimating Equation (GEE) for LCR Models

The GEE was proposed by Liang and Zeger [4] as an extension of the generalised linear model (GLM) framework of Nelder and Weddenburn [5] to deal with correlated data. The general form of a GEE can be defined as in equation (4):

$$G(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\partial E_{\boldsymbol{\theta}} \left(\mathbf{Y}_i^T \right)}{\partial \boldsymbol{\theta}} \mathbf{V}_i^{-1} (\boldsymbol{\theta}) \left[\mathbf{Y}_i - E_{\boldsymbol{\theta}} \left(\mathbf{Y}_i \right) \right] = \mathbf{0} \tag{4}$$

where $E_{\boldsymbol{\theta}} \left(\mathbf{Y}_i^T \right)$ is the mean response of $\mathbf{Y}_i$, $\mathbf{V}_i^{-1} (\boldsymbol{\theta})$ is the working covariance structure, and $\left[\mathbf{Y}_i - E_{\boldsymbol{\theta}} \left(\mathbf{Y}_i \right) \right]$ is the vector of centralised responses. To simplify the notation in equation (4), we set $\mathbf{D}_i^T = \partial E_{\boldsymbol{\theta}} \left(\mathbf{Y}_i^T \right) / \partial \boldsymbol{\theta}$. The generalised estimating equation approach relaxes the use of 'true' covariance structure by introducing working covariance structure, $\mathbf{V}_i^{-1} (\boldsymbol{\theta})$. This method was developed by Zeger and Liang [33], Zeger, Liang, and Albert [34], Lipsitz, Laird, and Harrington [35], Liang and Zeger [36], Preisser and Qaqish [37], Burton, Gurrin, and Sly [38], Pan [6], Diggle *et al.* [39], Hanley *et al.* [40], Hardin and Hilbe [41], Ballinger [42], Chaganty [43], Twisk [44], Balan and Schiopu-Kratina [45], Ogungbenro and Aarons [46], Natarajan *et al.* [47], Goetgeluk and Vansteelandt [48], Koper and Manseau [49], Shults *et al.* [50], Chen, Yi, and Cook [51], Warton [52], Shen and Chen [53], Wang, Zhou, and Qu [54], Shen and Chen [55], Touloumis, Agresti, and Kateri [56], Stoklosa, Gibb, and Warton [57], Yang, Chen, and Lung [58], Chen, Liang, and Wang [59], Kalema and Molenberghs [60], Jaman *et al.* [61]; Nikoloulopoulos [62], and Wang *et al.* [63]). To use a GEE for an LCR model, we extended model (4) to incorporate covariate effects as in equation (5):

$$G(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\partial E_{\boldsymbol{\theta}} \left(\mathbf{Y}_i^T \,\middle|\, \mathbf{x}_i \right)}{\partial \boldsymbol{\theta}} \mathbf{V}_i^{-1} \left(\boldsymbol{\theta}; \, \mathbf{x}_i, \, \mathbf{z}_i \right) \left[\mathbf{Y}_i - E_{\boldsymbol{\theta}} \left(\mathbf{Y}_i \,\middle|\, \mathbf{x}_i, \, \mathbf{z}_i \right) \right] = \mathbf{0}_{W \times 1} \tag{5}$$

Here, $\mathbf{x}_i$ and $\mathbf{z}_i$ are the covariates that associate with latent class membership and the covariates that are related to measured responses, respectively. The estimates of parameters are obtained using the Newton-Raphson procedure. The selection procedure to select the best working covariance structure, however, is beyond the scope of this paper.

2.3 Model Selection Criteria in GEEs for LCR Models

Traditional model-selection criteria such as AIC and BIC cannot be used directly in GEEs since they are likelihood-based, and full multivariate likelihoods are not expressed in GEE estimations. Instead, the estimation is based (in marginal) on quasi-likelihood. A modification of AIC, called the quasi-information criterion (QIC), was introduced by Pan [6]. The QIC

was constructed by replacing the likelihood in the Kullback-Leibler information with the quasi-likelihood under the working independence assumption. The QIC is defined as in equation (6):

$$QIC\,(R) = -2Q\left(\hat{\beta}\,(\mathbf{R})\,;\,I,\,D_{\text{True}}\right) + 2tr\left(\Omega\hat{V}_{\text{san}}\right) \tag{6}$$

where $\Omega = \sum_{i=1}^{n} \mathbf{D}_i^T \mathbf{V}_i \mathbf{D}$ is the model-based variance estimator under the independence working correlation structure, and $\hat{V}_{\text{san}}$ is the sandwich variance estimator under the working correlation structure $\mathbf{R}$ described by Liang and Zeger [4]. Since the QIC is built under an independence working correlation structure, we face the problem of misspecification if we apply another working correlation structure. Hence, Hin and Wang [7] proposed another criterion used for a more general working correlation structure by eliminating the first term of the right-hand side of equation 6. This criterion is called the correlation information criterion (CIC), and is defined as in equation (7):

$$CIC\,(R) = 2tr\left(\Omega\hat{V}_{\text{san}}\right) \tag{7}$$

It improves QIC performance. In this paper, we modify the QIC of Pan [6] and CIC of Hin and Wang [7] for LCR models with a covariate effect in both measured and underlying variables. These criteria assume the independence of m observed polytomous responses within the same individual. The independence working covariance for each individual $\mathbf{V}_i$ can be derived as a block-diagonal matrix as in equation (8):

$$\underset{\left(\sum_{m=1}^{M} K_m - M\right) \times \left(\sum_{m=1}^{M} K_m - M\right)}{\mathbf{V}_i} = \begin{bmatrix} \mathbf{V}_{i1} & 0 & \cdots & 0 \\ 0 & \mathbf{V}_{i2} & 0 & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & \mathbf{V}_{iM} \end{bmatrix} \tag{8}$$

where the elements of $\mathbf{V}_i$, $\mathbf{V}_{im} \equiv \left\{V_{i,\,(m,k),\,(m,k')}\,(\theta;\,\mathbf{x}_i,\,\mathbf{z}_i)\right\} = \left\{Cov\,(Y_{imk},\,Y_{imk'}|\,\mathbf{x}_i,\,\mathbf{z}_i)\right\}$, is a $(K_m - 1) \times (K_m - 1)$ block matrix of the working covariance for the mth responses.

3   Data Example

In this section, we use data from the mHELP project to illustrate our approach. The mHELP project investigates the indicators that jointly affect frailty among patients after undergoing abdominal surgery. These indicators are based on Fried's criteria: shrinking (weight loss), weakness, exhaustion, low activity, and a slow walking speed. This project recruited elderly Taiwanese adults after undergoing abdominal surgical procedures. mHELP was hypothesised to improve the frailty of these patients at discharge. The sample included 377 elderly patients. However, for the purpose of our study, we only used the samples with complete observations. Hence, we used 320 out of the 377 patients, 149 of whom received usual care and were treated as the control group, and the remaining patients who received mHELP care served as the intervention group. Of the 320 patients, 184 were male and the rest were female. We utilised covariates suggested by Chen *et.al.* [8]. The covariates included the patients' demographics (age (years), gender (1 = male; 2 = female)) and medical statistics (duration of surgery (minutes), length of hospital stay (days), comorbidity, baseline frailty status at admission (using Fried's criteria; 1 criterion present = not frail, 2 criteria present = pre-frail, 3 or more criteria present

= frail), cancer stage (1 = stage 1, 2 = stage 2, 3 = stage 3, 4 = stage 4)). The data on demographic and medical characteristics were obtained from the medical report of each patient. Comorbidities were measured using the Charlson Comorbidity Index, where higher scores indicate a higher mortality risk. We also set the care group indicator (1 = usual care, 2 = mHELP care) as our primary covariate to be related to the latent frailty.

4   Model Fitting and Data Analysis

We fit LCR models using the five dichotomised frailty criteria: shrinking, weakness, exhaustion, low activity, and a slow walking speed as the observed responses. These five criteria were used to jointly reflect the underlying frailty, which was modelled as a categorical latent variable (i.e. the latent class). These measurements need to be combined appropriately to derive summaries of frailty. Seven covariates which were probably related to observed responses (also known as secondary covariates) were chosen as potential candidates to be incorporated into the LCR, including age, gender, duration of surgery, length of hospital stay (LOS), comorbidity, baseline frailty, and cancer stage. We set "treatment", the care group indicator (1 = usual care, 2 = mHELP care), as our primary covariate to be related to the latent class, which can be used to evaluate the effect of mHELP on frailty.

The GEE approach that we propose allows LCR modelling to be performed by regressing different covariates for different frailty indicators. For the purpose of the present study, we performed a pre-analysis to select significant covariates for each response (frailty indicator), which will be incorporate into LCR

Table 1: The Selected Covariate(s) for Each Response

Responses	Covariates
Y_{i1} = shrink	z_{i11} = gender z_{i12} = Length of hospital stay (day)
Y_{i2} = weakness	z_{i21} = age (year) z_{i22} = Length of hospital stay (LOS) (day)
Y_{i3} = exhaustion	z_{i31} = age (year)
Y_{i4} = low activity	z_{i41} = Baseline frailty criteria
Y_{i5} = slow walking speed	z_{i51} = age (year) z_{i52} = Length of hospital stay (LOS) (day) z_{i53} = comorbidity

as the secondary covariates $\mathbf{z}_{im}$. For each binary response (frailty indicator), we fit a logistic regression to all seven potential candidate covariates (introduced by Chen *et. al.*[8]) and used the stepwise procedure to obtain the important secondary covariate(s) $\mathbf{z}_{im}$. Table 1 presents the selected secondary covariate(s) for each response.

Table 2: Information Criteria for LCR Models with Two and Three Latent Classes

		LCR	
Working Covariance	Information criterion	class 2	class 3
Independence (GEE-I)	QIC	**1601.32**	2424.46
	CIC	**21.53**	66.08

In this paper, we proposed a GEE approach for LCR models (henceforth, the GEE for the LCR approach) by performing independence working covariance structures (GEE-I). Then, we selected the number of latent classes for our LCR model. Table 2 shows the summary of QIC and CIC for LCR's with two and three latent classes under independence working covariance structures. The results showed that the LCR with two latent classes provided the lowest values of both QIC and CIC. This indicates that this model gave the best result. Hence, in this paper, we chcose to concentrate on LCRs with two latent classes (not frail versus frail after surgery).

The direct relationship between frailty indicators and their confounding variables can be shown using the odds ratio (OR). The odds ratios are obtained by taking the exponential transformation of the estimates of regression coefficients in equation (2) [i.e., $\exp(\alpha_{imk})$'s], which is shown in Table 3. To summarise the results, we found that: (a) females were more at risk of shrinking than males; (b) patients who stayed in hospital longer were more likely to experience shrinkage, weakness, and a slow walking speed;

Table 3: Conditional Probability Regression from LCR (1) for the Direct Relationship between Frailty Indicators and Confounding Variable: mHELP Project

Frailty Indicators	Comparison	Confounding Variables	OR*	95% CI for OR
Shrink	no vs. yes	Female	**4.18**	(1.95, 8.97)
		LOS	**0.95**	(0.91, 0.99)
Weakness	no vs. yes	Age	**0.88**	(0.83, 0.93)
		LOS	**0.93**	(0.89, 0.97)
Exhaustion	no vs. yes	Age	**0.96**	(0.92, 0.99)
Low activity	no vs. yes	Frailty at baseline	**0.19**	(0.06, 0.56)
Slow walking speed	no vs. yes	Age	**1.14**	(1.13, 1.15)
		LOS	**0.98**	(0.97, 0.99)
		Comorbidity	**2.23**	(2.22, 2.24)

*Value in bold are significantly different from 1 at the 0.05 level

(c) older patients were more likely to experience weakness and exhaustion, although they did not develop a relatively slow walking speed; (d) patients with a high level of comorbidity were less likely to have a slow walking speed; and (e) patients who had baseline frailty were more likely to show low activity. In addition, the association between latent class membership and the risk factor was obtained by taking the exponential

Table 4: Latent Prevalence Regression from LCR (1) for the Relationship between Underlying Frailty Status and the Risk Factor: mHELP Project

Risk Factor	Comparison	OR	95% CI for OR
Treatment (1 = mHELP)	no frail vs. frail	**2.28**	(1.34, 3.87)

Value in bold are significantly different from 1 at the 0.05 level

transformation of the estimate of the regression coefficient in equation (3) [i.e., $\exp(\beta_{pj})$'s]. This association is shown in Table 4. Thus, mHELP reduces the risk of patients becoming frail after surgery by approximately 2.28 times compared to usual care.

5   Discussion

The generalised estimating equation (GEEs) approach relaxes the role of "true" covariance structure by introducing a working covariance; as a result, there is no need to establish the joint distribution of the responses. The GEE is an extension of the GLM framework to deal with correlated data, hence, it is an appropriate approach for the case of repeated measures, such as LCR models. Therefore, in this paper, we emphasised the use of the GEE approach for LCR models. The main issue coming from the GEE approach is that there are only a few model-selection criteria in GEEs. We cannot use either the AIC or BIC criteria directly, since GEEs are not full-likelihood based models. One way to solve this issue is to use QIC and CIC as a modification of AIC. This paper presented the modification of QIC and CIC for the LCR model under an independence working covariance structure. We used our proposed criteria to determine how many classes should be involved in the LCR model. The lower the QIC and/or CIC, the more appropriate the number of classes in the LCR model. Instead of using these criteria, we could utilise another method to determine the number of classes in the LCR model. For instance, we can utilise the bayesian method to decide the number of classes by using prior and posterior properties. Furthermore, to increase the efficiency of GEE approach, probably we could apply another working covariance structure, such as conditional pairwise independence working covariance.

References

[1] Green, B.F. A general solution for the latent class model of latent structure analysis. *Psychometrika.* (1951). 16: 151-166.

[2] McCullagh, P., and Nelder, J.A. *Generalized Linear Models Second Edition.* London: Chapman and Hall. 1989.

[3] Agresti, A. *Categorical data analysis, 3rd Edition.* New Jersey: John Wiley & Sons. 2013.

[4] Liang, K.Y., and Zeger, S.L. Longitudinal data analysis using generalized linear model. *Biometrika.* 1986. 73(1): 13-22.

[5] Nelder, J.A., and Weddenburn, R.W.M. Generalized linear models. *Journal of the Royal Statistical Society Series A.* 1972. 135(3): 370-384.

[6] Pan, W. Akaike's information criterion in generalized estimating equations. *Biometrics.* 2001. 57(1): 120-125.

[7] Hin, L.Y., and Wang, Y.G. Working-correlation-structure identification in generalized estimating equations. *Statist. Med.* 2009. 28: 642-658.

[8] Chen, C.C.H., Chen, C.N., Lai, I.R., Huang. G.H., Saczynski, J.S., and Inouye, S.K. Effects of a modified hospital elder life program on frailty in individuals undergoing major elective abdominal surgery. *The American Geriatrics Society.* 2014. 62: 261-268.

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

[10] Pickles, A., Bolton, P., Macdonald, H., Bailey, A., Le Couteur, A., Sim, C.H., and Rutter, M. Latent class analysis of recurrence risks for complex phenotypes with selection and measurement error: a twin and family history study of autism. *Am. J. Hum. Genet.* 1995. 57: 717-726.

[11] Hudziak, J.J., Heath, A.C., Madden, P.F., Reich, W., Bucholz, K.K., Slutske, W., Bierut, L.J., Neuman, R.J., and Todd, R.D. Latent class and factor analysis of DSM-IV ADHD: A twin study of female adolescents. *Journal of the American Academy of Child and Adolescent Psychiatry.* 1998. 37: 848–857.

[12] Garrett, E.S., and Zeger, S.L. Latent class model diagnosis. *Biometrics.* 2000. 56: 1055–1067.

[13] Neuman, R.J., Heath, A., Reich, W., Bucholz, K.K., Madden, P.A.F., Sun, L., Todd, R.D., and Hudziak, J.J. Latent class analysis of ADHD and comorbid symptoms in a population sample of adolescent female twins. *Journal of Child Psychology and Psychiatry and Allied Disciplines.* 2001. 42: 933–942.

[14] Nylund, K., Nishina, A., Bellmore, A., and Graham, S. Subtypes, severity, and structural stability of peer victimization: what does latent class analysis say?. *Child Development.* 2007. 78: 1706-1722.

[15] Li, J.J., and Lee, S.S. Latent class analysis of antisocial behavior: interaction of serotonin transporter genotype and maltreatment. *J Abnorm Child Psychol.* 2010. 38: 789–801.

[16] Larance, B., Carragher, N., Mattick, R.P., Lintzeris, N., Ali, R., and Dagenhardt, L. A latent class analysis of self-reported clinical indicators of psychosocial stability and adherence among opioid substitution therapy patients: do stable patients receive more unsupervised doses?. *Drug and Alcohol Dependence.* 2014. 142: 45-55.

[17] Moustaki, I. A latent trait and a latent class model for mixed observed variables. *British Journal of Mathematical and Statistical Psychology.* 1996. 49: 313–334.

[18] Sullivan, P.F., Kessler, R.C., and Kendler, K.S. Latent class analysis of lifetime depressive symptoms in national comorbidity survey. *American Journal of Psychiatry.* 1998. 155: 1398–1406.

[19] Bandeen-Roche, K., Miglioretti, D.L., Zeger, S.L., and Rathouz, P.J. Latent variable regression for multiple discrete responses. *Journal of the American Statistical Association.* 1997. 92: 1375–1386.

[20] Klonsky, E.D., and Olino, T.M. Identifying clinically distinct subgroups of self-injurers among young adults: a latent class analysis. *Journal of Consulting and Clinical Psychology.* 2008. 76(1): 22-27.

[21] Neuman, M., Wirtz, M., Ernstmann, N., Ommen, O., Langler, A., Edelhauser, F., Scheffer, C., Tauschel, D., and Pfaff, H. Identifying and predicting subgroups of information needs among cancer patients: an initial study using latent class analysis. *Support Care Cancer.* 2011. 19: 1197-1209.

[22] Clogg, C.C., and Goodman, L.A. Latent structure analysis of a set of multidimensional contingency tables. *Journal of the American Statistical Association.* 1984. 79: 762–771.

[23] Clogg, C.C., and Goodman, L.A. Simultaneous latent structure analysis in several groups. In *Sociological Methodology 1985*, Tuma N.B. (ed.). San Francisco: Jossey-Bass. 1985. 81–110.

[24] Formann, A.K. Constrained latent class models: Theory and applications. *British Journal of Mathematical and Statistical Psychology.* 1985. 38: 87–111.

[25] Dayton, C.M., and Macready, G.B. Concomitant-variable latent-class models. *Journal of the American Statistical Association.* 1988. 83: 173–178.

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

[27] Hagenaars, J.A. *Loglinear Models with Latent Variables.* Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-094. Newbury Park, CA: Sage. 1993.

[28] Melton, B., Liang, K.Y., and Pulver, A.E. 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.

[29] Van der Heijden, P.G.M., Dessens, J., and Bockenholt, U. Estimating the concomitant variable latent class model with the EM algorithm, *Journal of Educational and Behavioral Statistics.* 1996. 21: 215–229.

[30] Leisch, F.F. A general framework for finite mixture models and latent class regression in r. *Journal of Statistical Software.* 2004. 11: 1–18.

[31] Huang, G.H., and Bandeen-Roche, K. Building and identifiable latent class model with covariate effects on underlying and measured variables. *Psychometrika.* 2004. 69(1): 5-32.

[32] Huang, G.H. Selecting the number of classes under latent class regression: a factor analytic analogue. *Psychometrika.* 2005. 70(2): 325-345.

[33] Zeger, S.L., and Liang, K.Y. Longitudinal data analysis for discrete and continuous responses. *Biometrics.* 1986. 42(1): 121-130.

[34] Zeger, S.L., Liang, K.Y., and Albert, P.S. Models for longitudinal data: a generalized estimating equation approach. *Biometrics.* 1988. 44: 1049-1060.

[35] Lipsitz, S.R., Laird, N.N., and Harrington, D.P. Generalized estimating equations for correlated binary data using the odds ratio as a measure of association. *Biometrika.* 1991. 78(1): 153-160.

[36] Liang, K.Y., and Zeger, S.L. Regression analysis for correlated data. *Annu. Rev. Pub. Health.* 1993. 14: 43-68.

[37] Preisser, J.S., and Qaqish, B.F. Deletion diagnostic for generalized estimating equation. *Biometrika.* 1996. 83(3): 551-562.

[38] Burton, P., Gurrin, L., and Sly, P. Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modeling. *Statistics in Medicine.* 1998. 17: 1261-1291.

[39] Diggle, P.J., Heagerty, P., Liang, K.Y., and Zeger, S.L. *Analysis of Longitudinal Data second edition.* Oxford University Press. London. 2002.

[40] Hanley, J.A., Negassa, A., deB. Edwardes, M.D., and Forrester, J.E. Statistical analysis of correlated data using generalized estimating equations: an orientation. *American Journal of Epidemiology.* 2002. 157(4): 364-375.

[41] Hardin, J.W., and Hilbe, J.M. *Generalized Estimating Equations.* London: Chapman and Hall. 2003.

[42] Ballinger, G.A. Using generalized estimating equations for longitudinal data analysis. *Organizational Research Methods.* 2004. 7: 127-150.

[43] Chaganty, N.R., and Joe, H. Efficiency of generalized estimating equations for binary responses. *J.R. Statist. Soc. B.* 2004. 66(4): 851-860.

[44] Twisk, J.W.R. Longitudinal data analysis. A comparison between generalized estimating equations and random coefficient analysis. *European Journal of Epidemiology.* 2004. 19(8): 769-776.

[45] Balan, R.M., and Schiopu-Kratina, I. Asymptotic results with generalized estimating equations for longitudinal data. *The Annals of Statistics.* 2005. 33(2): 522-541.

[46] Ogungbenro, K., and Aarons, L. Sample size calculation based on generalized estimating equations for population pharmacokinetic experiments. *Journal of Biopharmaceutical Statistics.* 2006. 16: 135-150.

[47] Natarajan, S., Lipsitz, S., Parzen, M., and Lipshultz, S. A measure of partial association for generalized estimating equations. *Statistical Modelling.* 2007. 7(2): 175-190.

[48] Goetgeluk, S., and Vansteelandt, S. Conditional generalized estimating equations for the analysis of clustered and longitudinal data. *Biometrics.* 2008. 64: 772-780.

[49] Koper, N., and Manseau, M. Generalized estimating equations and generalized linear mixed-effects models for modeling resource selection. *Journal of Applied Ecology.* 2009. 46: 590-599.

[50] Shults, J., Sun, W., Tu, X., Kim, H., Amsterdam, J., Hilbe, J.M., and Ten-Have, T. A comparison of several approaches for choosing between working correlation structures in generalized estimating equation analysis of longitudinal binary data. *Statistics in Medicine.* 2009. 28: 2338-2355.

[51] Chen, B., Yi, G.Y., and Cook, R.J. Weighted estimating functions for longitudinal response and covariate data that are missing at random. *Journal of the American Statistical Association.* 2010. 105(489): 336-353.

[52] Warton, D.I. Regularized sandwich estimators for analysis of high-dimensional data using generalized estimating equations. *Biometrics.* 2011. 67: 116-123.

[53] Shen, C.W., and Chen, Y.H. Model selection for generalized estimating equations accommodating dropout missingness. *Biometrics.* 2012. 68: 1046-1054.

[54] Wang, L., Zhou, J., and Qu, A. Penalized generalized estimating equation for high-dimensional longitudinal data analysis. *Biometrics*. 2012. 68: 353-360.

[55] Shen, C.W., and Chen, Y.H. Model selection of generalized estimating equations with multiply imputed longitudinal data. *Biometrical Journal*. 2013. 6: 899-911.

[56] Touloumis, A., Agresti, A., and Kateri, M. GEE for multinomial responses using a local odds ratios parameterization. *Biometrics*. 2013. 69: 633-640.

[57] Stoklosa, J., Gibb, H., and Warton, D.I. Fast forward selection for generalized estimating equations with a large number of predictor variables. *Biometrics*. 2014. 70: 110-120.

[58] Yang, M.C., Chen, K.P., and Lung, F.W. Generalized estimating equation model and long-term exposure effect of antipsychotics on SH-SY5Y cells against oxidative stressors. *European Journal of Pharmacology*. 2014. 740: 697-702.

[59] Chen, J., Li, D., Liang, H., and Wang, S. Semiparametric GEE analysis in partially linear single-index models for longitudinal data. *The Annals of Statistics*. 2015. 43(4): 1682-1715.

[60] Kalema, G., and Molenberghs, G. Second-order generalized estimating equations for correlated count data. *Comput Stat*. 2016. 31: 749-770.

[61] Jaman, A., Latif, M.A.H.M., Bari, W., and Wahed, A.S. A determinant-based criterion for working correlation structure selection in generalized estimating equation. *Statistics in Medicine*. 2016. 35: 1819-1833.

[62] Nikoloulopoulos, A.K. Correlation structure and variable selection in generalized estimating equation via composite likelihood information criteria. *Statistics in Medicine*. 2016. 35: 2377-2390.

[63] Wang, M., Kong, L., Li, Z., and Zhang, L. Covariance estimators for generalized estimating equations (GEE) in longitudinal analysis with small sample. *Statistics in Medicine*. 2016. 35: 1706-1721.