A composite logistic regression approach for ordinal panel data regression

Ronghua Luo and Hansheng Wang*

Guanghua School of Management
Peking University
Beijing 100871, PR China
E-mail: luoronghua@gsm.pku.edu.cn
E-mail: hansheng@gsm.pku.edu.cn
*Corresponding author

Abstract: We propose in this article a Composite Logistic Regression (CLR) approach for ordinal panel data regression. The new method transforms the original ordinal regression problem into a number of binary ones. Thereafter, the method of conditional logistic regression (Chamberlain, 1984; Wooldridge, 2001; Hsiao, 2003) can be directly applied. As a result, the new method allows the unobserved subject effects to be correlated with the observed predictors in an arbitrary manner. Computationally, the new method is able to profile out unobserved subject effects in a very neat manner. This not only makes computational implementation very easy but also makes theoretical treatment straightforward. In particular, we show theoretically that the resulting estimator is $\sqrt{n}$-consistent and asymptotically normal. Both simulations and a real example are reported to demonstrate the usefulness of the new method.

Keywords: Composite Logistic Regression; CLR; conditional logistic regression; ordinal response; panel data; unobserved subject effect.

Reference to this paper should be made as follows: Luo, R. and Wang, H. (2008) ‘A composite logistic regression approach for ordinal panel data regression’, *Int. J. Data Analysis Techniques and Strategies*, Vol. 1, No. 1, pp.29–43.

Biographical notes: Mr. Ronghua Luo is currently a PhD candidate at the Department of Business Statistics & Econometrics, Guanghua School of Management, Peking University, Beijing, PR China. Mr. Luo obtained his BS in Statistics from the School of Mathematical Science, Nankai University, Tianjin, PR China.

Dr. Hansheng Wang is currently an Associate Professor of Statistics at the Department of Business Statistics & Econometrics, Guanghua School of Management, Peking University, Beijing, PR China. Dr. Wang obtained his BS in Statistics from the School of Mathematical Science, Peking University, Beijing, PR China, and his PhD in Statistics from the Department of Statistics, University of Wisconsin–Madison, Madison, Wisconsin, USA.

1 Introduction

Panel (or longitudinal) data refer to the data that follow a given sample of subjects over time (Diggle *et al.*, 1994; Wooldridge, 2001; Hsiao, 2003). Compared with typical cross-sectional data (providing one observation per individual), panel data collect multiple or repeated observations on every sampled subject. This immediately raises one question: what are the advantages of a panel data set as compared with its cross-sectional counterpart?

With the same number of subjects, a panel data set provides more observations than its cross-sectional counterpart, which naturally leads to estimators with better efficiency. Then, one may wonder whether such a merit is the only advantage of panel data. To fix the idea, consider the following standard panel linear regression model with an unobserved subject effect:

$$Y_{it} = \alpha_i + \mathbf{X}_{it}\beta + \varepsilon_{it},$$

where:

Y_{it} = the t-th repeated observation ($1 \le t \le T$) collected from the i-th subject ($1 \le i \le n$)

$\mathbf{X}_{it} \in \mathbb{R}^p$ = the associated p-dimensional predictor

ε_{it} = an independent noise

α_i = the unobserved random or fixed effect specific to the i-th subject.

Under a typical mixed-effects model set-up, one might assume that the subject effect α_i is a random effect, which is independent of the predictor $\mathbf{X}_{it}$ (Pinheiro and Bates, 2000). With such an assumption, one can easily show that the typical Ordinary Least Squares (OLS) estimator based on cross-sectional data (*i.e.*, $\{Y_{it}, \mathbf{X}_{it}\}_{i=1}^n$ with a fixed t) is still consistent for the unknown regression coefficient β (Wooldridge, 2001). Thus, the only merit of a panel data set under this set-up is the capability to provide more observations, and hence yield more accurate estimates.

Nevertheless, for many economics and/or medical research problems, the study is observational in nature. Therefore, the independence assumption between the unobserved subject effect α_i and the observed predictor $\mathbf{X}_{it}$ might be very questionable (Wooldridge, 2001; Hsiao, 2003). If α_i is indeed correlated with $\mathbf{X}_{it}$, the typical OLS estimator produced by the cross-sectional data $\{Y_{it}, \mathbf{X}_{it}\}_{i=1}^n$ (with an arbitrarily fixed t) could be biased and such a bias cannot be eliminated by purely increasing sample size n. However, such a problem can be fixed easily by panel data. For example, for two arbitrary distinct time points t_1 and t_2, one can easily verify that $\left(Y_{it_1} - Y_{it_2}\right) = \left(\mathrm{X}_{it_1} - \mathbf{X}_{it_2}\right)\beta + \left(\varepsilon_{it_1} - \varepsilon_{it_2}\right)$, which nicely profiles out the unobserved subject effect α_i and implies that a standard OLS estimator with $\left(Y_{it_1} - Y_{it_2}\right)$ as the response and $\left(\mathbf{X}_{it_1} - \mathbf{X}_{it_2}\right)$ as the predictor might lead to a consistent estimator for β (regardless of whether α_i is correlated with $\mathbf{X}_{it}$ or not). Although such a naive estimator might not be fully efficient, it clearly demonstrates that the advantage of panel data is much more than simply producing more observations. As a result, by controlling for the unobserved subject effect, panel data might help researchers to address many important questions, which can hardly be answered by its cross-sectional counterpart (Hsiao, 2003). Thus, we argue that the capability of controlling for unobserved subject effects is indeed the primary advantage of panel data. (For a more detailed discussion, see Wooldridge (2001) and Hsiao (2003).)

Given the importance of the panel data, much research effort has been devoted to this subject during the past decade. Particularly, the models with continuous response have been extensively studied (Diggle *et al*., 1994; Wooldridge, 2001; Hsiao, 2003; Fan and Li, 2004; Fan *et al*., 2007). Nevertheless, how to extend those useful results from continuous response to discrete response (*e.g.*, binary and ordinal) is not straightforward. Taking binary data as one example, one can follow the idea of linear mixed-effects models (Pinheiro and Bates, 2000) and propose a generalised mixed-effects model for binary panel data (Chamberlain, 1980; Gibbons and Hedeker, 1997; Liu and Hedeker, 2006). Unfortunately, the resulting likelihood function might not have a close form, which makes computational implementation very difficult (Butler and Moffitt, 1982; Breslow and Clayton, 1993). A similar problem also exists for the latent variable approaches (Qu *et al*., 1992; 1995). Another way to model discrete panel data is by completely ignoring the subject effect α_i and then specifying the regression relationship between Y_{it} and $\mathbf{X}_{it}$ marginally. For example, the method of the generalised estimation equation (Liang and Zeger, 1986, Generalised Estimation Equation (GEE)) and many others (Molenberghs and Lesaffre, 1994; Heagerty and Zeger, 1996) belong to this category. However, those marginal methods might not be able to provide consistent estimators if the ignored subject effect is indeed correlated with the predictor $\mathbf{X}_{it}$.

As a simple summary, compared with those extensively studied panel data models for continuous response, much less has been done for binary panel data, and even less has been done for ordinal panel data, which are typically encountered in real practice (Agresti, 1990). Thus, real applications call for effective methods for ordinal panel data modelling. As one can see from our previous discussion, a useful model must carry at least the following two important characteristics. Firstly, it should allow for unobserved subject effects and such subject effects should be allowed to correlate with the observed predictor vector in an arbitrary manner. Secondly, such a model should be computationally tractable without incurring any computationally forbidden task (*e.g.*, a high dimensional integration).

With these two objectives in mind, we propose in this article a Composite Logistic Regression (CLR) approach for ordinal panel data regression. The new method transforms the original ordinal regression problem into a number of binary ones, for which the method of conditional logistic regression (Chamberlain, 1984; Wooldridge, 2001; Hsiao, 2003) can be directly applied. As a result, the new method allows unobserved subject effects to be correlated with the observed predictors in an arbitrary manner. Computationally, the new method is able to profile out unobserved subject effects in a very neat manner. This not only makes computational implementation very easy but also makes theoretical treatment straightforward. In particular, we show theoretically that the resulting estimator is $\sqrt{n}$-consistent and asymptotically normal. Numerical studies are reported to illustrate the usefulness of the new method.

The rest of the article is organised as follows: Section 2 introduces the model and notations. The proposed new method and its main theoretical properties are also presented in this section. Numerical studies based on both simulated and real data sets are reported in Section 3. The article is concluded with a brief discussion in Section 4. All technical details are left to the Appendix.

2 Composite logistic regression

2.1 Model and notations

Let (Y_{it} $1 \le i \le n$ and $1 \le t \le m_i$) be the t-th ordinal response (taking values in $\{1, \cdots, K+1\}$) collected from the i-th subject, and let $\mathbf{X}_{it} = \left(X_{it1}, \cdots, X_{itp}\right)^{\top} \in \mathbb{R}^p$ be the associated p-dimensional predictor. Furthermore, we assume that the response Y is determined by the latent variable Y_{it}^* by the relationship $Y_{it} = k$ if $\gamma_{k-1} \le Y_{it}^* \le \gamma_k$, where γ_k are threshold values satisfying $\gamma_0 = -\infty < \gamma_1 < \cdots < \gamma_{K+1} = \infty$ and Y_{it}^* is related to X_{it} as:

$$Y_{it}^* = \alpha_i + X_{it}^{\top}\beta + \varepsilon_{it}.$$

Moreover, we assume that the error term ε_{it} follows the logistic distribution. As a result, we obtain the following cumulative logistic regression model:

$$P(Y_{it} \le k \mid \mathbf{X}_{it}) = \frac{\exp\left\{\alpha_{ik} + \mathbf{X}_{it}^{\top}\beta\right\}}{1 + \exp\left\{\alpha_{ik} + \mathbf{X}_{it}^{\top}\beta\right\}} \tag{2.1}$$

with $1 \le k \le K$, where $\beta = (\beta_1, \cdots, \beta_p)^{\top} \in \mathbb{R}^p$ is the interested regression coefficient with the true value given by β_0, and α_{ik} is the unobserved subject effect or incidental parameter (Wooldridge, 2001; Hsiao, 2003). Note that α_{ik} is allowed to be correlated with $\mathbf{X}_{it}$ in an arbitrary yet unknown manner. Lastly, we assume that conditional on $\left\{\alpha_{ik}\right\}_{k=1}^{K}$ and $\left\{\mathbf{X}_{it}\right\}_{t=1}^{m_i}$, the responses $\left\{Y_{it}\right\}_{t=1}^{m_i}$ are mutually independent (Wooldridge, 2001; Hsiao, 2003). By treating the subject effect α_{ik} as unknown parameters (*i.e.*, fixed effects), one can easily get the likelihood function of $\{Y_{it}, \mathbf{X}_{it}\}$. Thereafter, the Maximum Likelihood Estimator (MLE) can be obtained by optimising the joint likelihood with respect to both β and $\{\alpha_{ik}\}$ ($1 \le i \le n$ and $1 \le k \le K$). One can note immediately that the number of unknown parameters incurred by α_{ik} diverges to infinity as the sample size n increases. Such a problem is referred to as the 'incidental parameter problem' in econometrics literature (Wooldridge, 2001; Hsiao, 2003). As pointed out by Wooldridge (2001) and Hsiao (2003), the incidental parameter problem not only makes computation very difficult but also makes the resulting MLE inconsistent. To remedy such an issue, the method of conditional logistic regression has been developed for binary panel data (Chamberlain, 1980; Wooldridge, 2001; Hsiao, 2003).

2.2 Composite logistic regression

To fix the idea about ordinal panel data, we first consider an arbitrary but fixed $1 \le k \le K$ value. We can then define a binary response vector $\mathbf{Z}_{ik} = \left(Z_{i1k}, \cdots, Z_{im_ik}\right)^{\top} \in \mathbb{R}^{m_i}$ with $Z_{itk} = I(Y_{it} \le k)$, where $I(\cdot)$ is an indicator function (*i.e.*, $I = 1$ if $Y_{it} \le k$ but $I = 0$ if $Y_{it} > k$).

Furthermore, for one particular realisation of $\mathbf{Z}_{ik}$, we define $c_{ik} = \sum_{t=1}^{m_i} Z_{itk}$. Then, for such a fixed k, one can verify that (Wooldridge, 2001):

$$P\left(\mathbf{Z}_{ik}\left|\sum_t Z_{itk}=c_{ik}\right.\right)=\frac{\exp\left\{\mathbf{Z}_{ik}^{\top}\mathbb{X}_i\beta\right\}}{\sum_{\tilde{\mathbf{Z}}_{ik}\in S_{ik}}\exp\left\{\tilde{\mathbf{Z}}_{ik}^{\top}\mathbb{X}_i\beta\right\}} \tag{2.2}$$

where $\mathbb{X}_i=\left(\mathbf{X}_{i1},\cdots,\mathbf{X}_{im_i}\right)^{\top}$, $\tilde{\mathbf{Z}}_{ik}=\left(\tilde{Z}_{i1k},\cdots,\tilde{Z}_{im_ik}\right)^{\top}\in\mathbb{R}^{m_i}$ and $S_{ik}=\left\{\tilde{\mathbf{Z}}_{ik}:\tilde{Z}_{i1k}+\cdots.+\tilde{Z}_{im_ik}=c_{ik}\right\}$. As one can see, the conditional probability (Equation 2.2) is free of the subject effect α_{ik}. Consequently, instead of optimising a typical likelihood function, which includes the incidental parameter α_{ik}, we can optimise the following conditional log likelihood function:

$$\mathcal{L}_k(\beta)=\sum_{i=1}^{n}\log P\left(\mathbf{Z}_{ik}\left|\sum_t Z_{itk}=c_{ik}\right.\right) \tag{2.3}$$

where the probability $P\left(\mathbf{Z}_{ik}\left|\sum_t Z_{itk}=c_{ik}\right.\right)$ is defined according to Equation (2.2). By maximising the conditional log likelihood function (Equation 2.3), a conditional MLE for β can be obtained, which is $\sqrt{n}$-consistent and also asymptotically normal (Chamberlain, 1980; Wooldridge, 2001; Hsiao, 2003). Note that, by focusing on one fixed k value ($1\le k\le K$), the estimator produced by Equation (2.3) might not be very efficient. To combine the information provided by different $\mathcal{L}_k(\beta)$ functions, we can further construct the following composite conditional log likelihood function:

$$\mathcal{L}(\beta)=\sum_{k=1}^{K}\mathcal{L}_k(\beta). \tag{2.4}$$

Then, the final estimator can be obtained by maximising the composite log likelihood function (Equation 2.4). We refer to such an approach as 'Composite Logistic Regression' (CLR) and the resulting estimator $\hat{\beta}$ as the CLR estimator.

2.3 Theoretical properties

To establish a rigorous asymptotic theory for the CLR estimator $\hat{\beta}$, the following technical conditions are needed:

(C.1) Assume that $E\left\{\sum_{i=1}^{m_i}\|X_{it}\|^2\right\}<\infty$

(C.2) Assume that $E\left\{\operatorname{var}\left(\mathbb{X}_i^{\top}\mathbf{Z}_{ik}\left|\sum_t Z_{itk}=c_{ik}\right.\right)\right\}$ is positive definite.

We remark that both assumptions are very reasonable. Specifically, the condition (C.1) is trivially satisfied if $E\|\boldsymbol{X}_{it}\|^2<\infty$ and $m_i\le m_{\max}$ for some $m_{\max}>0$. Simply put, as long as the number of repeated observations is bounded and the marginal distribution of the predictor has a finite second-order moment, condition (C.1) is well satisfied. The second condition (C.2) is also very reasonable. This condition has a similar flavour to requiring $\operatorname{var}(X_{it})$ to be positive definite in a typical OLS regression setting. Then, we have the following two theorems:

Theorem 1. (Consistency) *Assuming conditions (C.1) and (C.2), we then must have* $\hat{\beta} - \beta_0 \to_p 0$, *where "$\to_p$" stands for convergence in probability.*

Theorem 2. (Asymptotic normality) *Assuming conditions (C.1) and (C.2), we then must have* $\sqrt{n}(\hat{\beta} - \beta_0) \to_d N(0, \Sigma)$, *where "$\to_d$" stands for convergence in distribution:*

$$\Sigma = \left(\sum_{k=1}^{K} var\left\{ \mathbb{X}_i^{\mathsf{T}} \mathbf{Z}_{ik} - \eta_{ik} \right\} \right)^{-1} var\left(\sum_{k=1}^{K} \left\{ \mathbb{X}_i^{\mathsf{T}} \mathbf{Z}_{ik} - \eta_{ik} \right\} \right) \left(\sum_{k=1}^{K} var\left\{ \mathbb{X}_i^{\mathsf{T}} \mathbf{Z}_{ik} - \eta_{ik} \right\} \right)^{-1}$$

$$\eta_{ik} = \left(\sum_{\tilde{\mathbf{Z}}_{ik} \in S_{ik}} \left\{ \mathbb{X}_i^{\mathsf{T}} \mathbf{Z}_{ik} \right\} \exp\left\{ \tilde{\mathbf{Z}}_{ik}^{\mathsf{T}} \mathbb{X}_i \beta \right\} \right) \left(\sum_{\tilde{\mathbf{Z}}_{ik} \in S_{ik}} \exp\left\{ \tilde{\mathbf{Z}}_{ik}^{\mathsf{T}} - \mathbb{X}_i \beta \right\} \right)^{-1}.$$

By Theorem 1, we know that the proposed CLR estimator is indeed consistent as long as both conditions (C.1) and (C.2) are satisfied. Furthermore, by Theorem 2, we know that such an estimator is also asymptotically normal with an asymptotic covariance matrix Σ. By substituting the unknown quantity β by its CLR estimator $\hat{\beta}$, and then replacing $E(\cdot)$ and var(·) by their sample versions, a consistent estimator for Σ can be easily obtained, which is denoted by $\hat{\Sigma}$.

3 Numerical studies

3.1 A simulation study

To evaluate the finite sample performance of the proposed CLR estimator, extensive Monte Carlo experiments are conducted. Because the results are qualitatively similar, we only report one example here. More specifically, the data are generated according to model (2.1) with $K = 2$ (*i.e.*, $Y_{it} \in \{1, 2, 3\}$), $\beta_0 = (\beta_{01}, \cdots, \beta_{05})^{\mathsf{T}} = (0.75, 0.50, 0.25, 0, 0)^{\mathsf{T}} \in \mathbb{R}^5$, $\alpha_{ik} = u_i + 4k$ $(1 \le k \le 2)$, where u_i is simulated from a standard normal distribution. With a given u_i, $\{X_{it}\}_{t=1}^{m_i}$ is then simulated independently from a normal distribution with mean $(u_i, \cdots, u_i)^{\mathsf{T}} \in \mathbb{R}^5$ and the $\mathrm{cov}(\mathbf{X}_{it}|u_i) = I_5$, where I_p stands for a $p \times p$ identity matrix. Thus, the unobserved subject effect α_{ik} is indeed correlated with the observed predictor $\{\mathbf{X}_{it}\}$. Lastly, the number of repeated observations m_i is simulated from a uniform distribution on $\{2, \cdots, m_{\max}\}$. Various sample sizes, from very small *(n* = 50) to relatively large *(n* = 400), are considered.

For each parameter set-up, a total of 1000 simulation iterations are conducted. Denote by $\hat{\beta}^{(s)} = \left(\hat{\beta}_1^{(s)}, \cdots, \hat{\beta}_p^{(s)} \right)^{\mathsf{T}} \in \mathbb{R}^p$ the CLR estimator obtained in the *s*th simulation iteration $(1 \le s \le 1000)$. Then, we estimate the true Root Mean Squared Error (RMSE) of $\hat{\beta}_j$ by:

$$\mathrm{RMSE} = \left\{ \frac{1}{1000} \sum_{s=1}^{1000} \left(\hat{\beta}_j^{(s)} - \beta_{0j} \right)^2 \right\}^{1/2}$$

which is also a consistent estimator for the asymptotic standard deviation of $\hat{\beta}_j$. Within each simulation iteration, we also recorded the estimated asymptotic standard deviation of $\hat{\beta}_j$ (denoted by $\hat{\sigma}_j^{(s)}$), whose sample average should be a consistent estimate for $E\left(\hat{\sigma}_j^{(s)}\right)$. We denote such a sample average by $\widehat{\text{RMSE}}$. For easy interpretation, we also compute the Standard Deviation (SD) of $\left\{\hat{\sigma}_j^{(s)}\right\}_{s=1}^{1000}$ and the Relative Bias (RB) as $\text{RB} = 100 \times (\widehat{\text{RMSE}} / \text{RMSE} - 1)$. Lastly, to test the hypothesis $H_0 : \beta_j = 0$ versus $H1 : \beta_j \neq 0$, a *Z*-type test statistic can be constructed as $t_j^{(s)} = \hat{\beta}_j^{(s)} / \hat{\sigma}_j^{(s)}$, which is asymptotically distributed as a standard normal distribution if β_j is indeed 0. We then compute the p-value as $p_j^{(s)} = 2\left\{1 - \Phi\left(\left|t_j^{(s)}\right|\right)\right\}$, where $\Phi(\cdot)$ stands for the cumulative distribution function of a standard normal distribution. With a prespecified significance level α (*e.g.*, 0.10, 0.05), we can reject the null hypothesis whenever we find $p_j^{(s)} < \alpha$. We then summarise in the last two columns of Table 1 the percentage of the experiments with the null hypothesis rejected.

From the third column of Table 1, we find that the value of RMSE steadily decreases as sample size *n* increases. This confirms that the proposed CLR estimator is indeed consistent. Next, from the fourth column, we find that the value of $\widehat{\text{RMSE}}$ is very close to that of the RMSE with a very small RB, which is reported in the sixth column. This means that the proposed covariance estimator is asymptotically unbiased. Furthermore, from the fifth column of Table 1, we find that the standard deviation of $\hat{\sigma}_j^{(s)}$ steadily decreases as the sample size increases. This further confirms that the proposed covariance estimator is also consistent. Lastly, the results reported in the sixth and seventh columns indicate that if $\beta_j \neq 0$ (*i.e.*, β_1, β_2, and β_3), the aforementioned standard *Z*-test does have nontrivial power. As the sample size increases, the power approaches 100% very quickly. The results also indicate that if $\beta_j = 0$ (*i.e.*, β_4 and β_5), the empirical sizes estimated by our Monte Carlo experiments are very close to their nominal level. This confirms the asymptotic normality of our proposed CLR estimator. Lastly, we note that a larger number of repeated observations (*e.g.*, $m_{\max} = 10$ versus $m_{\max} = 5$) typically lead to better estimates with smaller RMSE values.

3.2 A real example

To demonstrate the usefulness of the proposed CLR method in real applications, we consider here a data set about the Chinese stock market. The data set is derived from the CCER Chinese Stock Database,[1] which is partially developed by the Chinese Center for Economic Research (CCER) at Peking University. It is one of the most authoritative and widely used databases about the Chinese stock market (Wang *et al.*, 2007a). After eliminating those incomplete and/or extreme observations, we obtain a total of 4917 observations, which are generated by a total of 1192 different firms in the period between year 2000 and year 2004 (*i.e.*, a total of five years). Those firms account for about 90% of the entire Chinese stock market during that period.

Table 1 The detailed simulation results

n	Regression coefficient	$m_{max} = 5$ RMSE	$\widehat{RMSE}$	SD of $\hat{\sigma}_j^{(k)}$	RB(%)	Nominal level $\alpha = 10\%$	$\alpha = 5\%$	$m_{max} = 10$ RMSE	$\widehat{RMSE}$	SD of $\hat{\sigma}_j^{(k)}$	RB(%)	Nominal level $\alpha = 10\%$	$\alpha = 5\%$
50	β_1	0.3871	0.3138	0.0983	−18.9392	0.9130	0.8140	0.1973	0.1815	0.0323	−8.0066	0.9990	0.9930
	β_2	0.3702	0.2911	0.0864	−21.3795	0.6620	0.5340	0.1884	0.1719	0.0281	−8.7677	0.9170	0.8550
	β_3	0.3214	0.2774	0.0654	−13.6950	0.2900	0.2080	0.1761	0.1658	0.0264	−5.8240	0.4590	0.3320
	β_4	0.3113	0.2717	0.0619	−12.7299	0.1350	0.0750	0.1727	0.1626	0.0266	−5.8650	0.1250	0.0780
	β_5	0.3208	0.2761	0.0765	−13.9314	0.1290	0.0660	0.1773	0.1617	0.0251	−8.7691	0.1290	0.0730
100	β_1	0.2251	0.2065	0.0368	−8.2627	0.9970	0.9890	0.1281	0.1264	0.0154	−1.3148	1.000	1.0000
	β_2	0.2102	0.1934	0.0307	−7.9744	0.8890	0.8060	0.1237	0.1188	0.0133	−3.9290	0.9960	0.9910
	β_3	0.1951	0.1854	0.0278	−4.9817	0.4380	0.3200	0.1196	0.1155	0.0137	−3.4182	0.7280	0.6080
	β_4	0.1918	0.1829	0.0267	−4.6787	0.1110	0.0600	0.1117	0.1138	0.0129	1.8531	0.0910	0.0430
	β_5	0.1990	0.1829	0.0283	−8.0801	0.1430	0.0650	0.1137	0.1130	0.0122	−0.6330	0.1170	0.0530
200	β_1	0.1447	0.1417	0.0169	−2.0758	1.0000	1.0000	0.0887	0.0883	0.0075	−0.4901	1.0000	1.0000
	β_2	0.1323	0.1329	0.0144	0.4480	0.9920	0.9820	0.0836	0.0837	0.0065	0.0746	1.0000	1.0000
	β_3	0.1382	0.1279	0.0129	−7.4964	0.6280	0.5240	0.0850	0.0811	0.0063	−4.5478	0.9260	0.8780
	β_4	0.1325	0.1263	0.0128	−4.6925	0.1100	0.0510	0.0806	0.0799	0.0060	−0.9105	0.0970	0.0560
	β_5	0.1298	0.1261	0.0127	−2.8767	0.1040	0.0570	0.0801	0.0801	0.0065	−0.0671	0.1010	0.0470
400	β_1	0.1004	0.0990	0.0080	−1.3099	1.0000	1.0000	0.0644	0.0622	0.0035	−3.4422	1.0000	1.0000
	β_2	0.0927	0.0933	0.0071	0.6261	1.0000	1.0000	0.0559	0.0592	0.0033	5.9733	1.0000	1.0000
	β_3	0.0937	0.0896	0.0064	−4.4180	0.8930	0.8180	0.0595	0.0570	0.0032	−4.1616	0.9940	0.9880
	β_4	0.0871	0.0886	0.0064	1.6929	0.0880	0.0460	0.0584	0.0566	0.0031	−3.2251	0.1110	0.0590
	β_5	0.0891	0.0881	0.0062	−1.1458	0.1030	0.0530	0.0569	0.0563	0.0031	−0.9184	0.1050	0.0520

For each firm, the following predictors are collected in the current year (recorded as the t-th year): Return on Equity (ROE) of the current year (ROE_t), asset turnover ratio (ATO), Profit Margin (PM), debt to asset ratio (LEV), sales growth rate (GROWTH) and the logarithm of total asset (ASSET). Specifically, ATO measures how effectively a company can make use of its assets. PM captures the company's profitability. LEV is a standard measure for the liability level and GROWTH reflects how fast the company is growing. Lastly, ASSET represents the firm's size. (For a more detailed discussion about the economics meanings of those variables, we refer to Harrison and Horngren (2001).) Next, we present summary statistics of these explanatory variables in Table 2, from which we find that the kurtosis of every variable (except ASSET) is very large. This suggests that the first- and second-order moments of many predictors could be infinite. Thus, the statistical validity of many standard regression methods (*e.g.*, logistic regression, CLR) becomes questionable. To fix the problem, we replace the value of ATO, PM, LEV and GROWTH by their normal scores. For consistency, the same operation is also conducted for ASSET.

Table 2 The descriptive analysis of the real example

Variable	*Mean*	*Median*	*SD*	*Kurtosis*	*Skewness*
ROE	0.0352	0.0767	0.4376	137.6314	–8.8023
ROE_t	0.0500	0.0840	0.3428	142.2203	–8.1871
ATO	0.5568	0.4485	0.4177	9.3458	2.0752
PM	–0.0329	0.0635	1.2859	681.4239	–23.3777
LEV	0.4664	0.4505	0.2457	70.2585	5.3035
GROWTH	0.2397	0.1276	0.7921	127.9780	8.8230
ASSET	21.0011	20.9375	0.8573	3.2889	0.3363

The response of interest (denoted by Y) is the discretised next year's (*i.e.*, the $(t + 1)$th year) earning. More specifically, $Y = 0$ if the firm's next year ROE is negative. Thus, $Y = 0$ implies that the firm is suffering loss. By the requirement of the Chinese Security Regulation Commission (CSRC, the government body overseeing the stock market), the firms reporting too many negative ROE values (*i.e.*, $Y = 0$) might be Specially Treated (ST), which is a severe punishment received from the CSRC. (For more detailed background information about the ST policy, we refer to Wang (2007).) Thus, the definition of the $Y = 0$ is very relevant. Next, we define $Y = 1$ as those firms with their ROE values falling into [0,0.10]. One might wonder why 0.10 is a critical number. Also by the requirement of the CSRC, if a firm wishes to do additional share offering, its ROE value must be maintained above the level of 0.10 for a number of years. Thus, 0.10 becomes another critical value. Lastly, we define $Y = 2$ if the reported ROE value is larger than 0.10. As a result, the percentage of the observations with $Y = 0$, 1 and 2 are given by 12.87%, 51.17% and 35.96%, respectively. Thus, the effect of extreme responses as evidenced by Wang *et al.* (2007b) is well controlled. A similar discretisation technique was also used by Zou and Yuan (2008). Per one careful anonymous referee's advice, the same operation is also conducted for ROE_t for consistency.

For comparison, we firstly ignore the panel data structure and treat it as typical cross-sectional data. We then apply a standard cumulative logistic regression model to the data, with detailed results reported in Table 3. As one can see, all estimated regression coefficients are statistically significant at the 5% level of significance. By Table 3, we know that, if we compare two different firms (*i.e.*, interfirm comparison), the firm with larger values in ATO, PM, LEV, GROWTH and ASSET tends to have better earnings than the other one. Those results are reasonable and might be useful for interfirm comparison. Nevertheless, they could be rather misleading if intrafirm comparison is our focus. In other words, one might also wish to know what would happen to a firm's earnings (compared with its own past history) if its size (*i.e.*, ASSET value for example) is getting larger and larger. To address this question, one has to control the firm's specific effect, *i.e.*, subject effect α_{ik} in (model 2.1). Otherwise, those statistically significant results due to interfirm differences cannot be differentiated from those due to intrafirm differences. Thus, our proposed CLR method becomes useful.

Table 3 The cross-sectional analysis results of the real example

Variable	*Estimate*	*SE*	*t-statistics*	*p-value*
ROE_t:1	0.4895	0.1276	–3.8366	0.0001
ROE_t:2	–1.2086	0.1566	7.7170	0.0000
ATO	–0.4794	0.0400	11.9806	0.0000
PM	–0.6327	0.0514	12.3122	0.0000
LEV	–0.1323	0.0361	3.6645	0.0002
GROWTH	–0.1130	0.0345	3.2732	0.0011
ASSET	–0.1014	0.0314	3.2236	0.0013

By recognising the panel data structure and applying the CLR method to the data, we obtain another set of regression coefficient estimates, which are very different from that of Table 3 and are reported in Table 4. Firstly, we note that all estimated coefficients are statistically significant at the 5% level of significance. Comparing the results of Table 4 with Table 3, we find that they all agree that increasing a firm's ATO, PM, LEV and GROWTH values can lead to better earnings in the future. Nevertheless, Table 4 presents a totally different sign for ASSET. According to Table 4, we find that if a firm's size (*i.e.*, ASSET) increases too much (compared with itself), its future earnings capability can be affected adversely (as compared with its past earning capability). We find this result very interesting but also reasonable. A firm's size and its earning capability are two components that have to be balanced for the firm's long-run health. If a firm's size keeps on increasing abnormally, then the firm's future earnings capability might be sacrificed to some extent, which is then reflected by the positive sign of ASSET in Table 4. In addition to that, from Table 4, we find that the firms with negative current year ROE are most likely to report better earnings next year. Such a result is very interesting but also reasonable. As we mentioned before, due to the ST policy enforced by the CSRC, those firms that currently suffer a negative year earning tend to have the strongest motivation to report positive earnings the following year, very often by means of earnings management (see Jiang and Wang, 2008). We remark that such a result is not available by a standard cross-sectional data analysis, as demonstrated in Table 3.

Table 4 The panel analysis results of the real example

Variable	*Estimate*	*SE*	*t-statistics*	*p-value*
ROE_t:1	1.4616	0.1481	9.8706	0.0000
ROE_t:2	1.3413	0.1885	7.1170	0.0000
ATO	–0.2322	0.1134	–2.0467	0.0407
PM	–0.7524	0.0833	–9.0368	0.0000
LEV	–0.4815	0.1049	–4.5889	0.0000
GROWTH	–0.1934	0.0514	–3.7601	0.0002
ASSET	1.2532	0.1783	7.0267	0.0000

4 Concluding remarks

We proposed in this article a CLR method for a simple yet effective analysis of ordinal panel data. We show both theoretically and numerically that the proposed method is rather useful. Its extension to semiparametric and/or nonparametric models is under investigation (Xia and Li, 1999; Fan and Li, 2004; Xia, 2006; Fan *et al.*, 2007).

Acknowledgements

The authors are grateful to the Editor and the referees for their helpful comments, which led to a substantially improved manuscript. This research is supported in part by NSFC (Grant No. 10771006).

References

Agresti, A. (1990) *Categorical Data Analysis*, New York, NY: John Wiley.

Breslow, N.E. and Clayton, D.G. (1993) 'Approximate inference in generalized linear mixed models', *Journal of the American Statistical Association*, Vol. 88, pp.9–25.

Butler, J.S. and Moffitt, R.A. (1982) 'A computationally efficient quadrature procedure for the one-factor multinomial probit model', *Econometrica*, Vol. 50, pp.761–764.

Chamberlain, G. (1980) 'Analysis of covariance with qualitative data', *Review of Economic Studies*, Vol. 47, pp.225–238.

Chamberlain, G. (1984) 'Panel data', in Z. Griliches and M.D. Intriligator (Eds.) *Handbook of Econometrics*, Amsterdam: North Holland, Vol. 2, pp.1247–1318.

Diggle, P.J., Liang, K.Y. and Zeger, S.C. (1994) *Analysis of Longitudinal Data*, Oxford: Oxford University Press.

Fan, J., Huang, T. and Li, R. (2007) 'Analysis of longitudinal data with semiparametric estimation of covariance function', *Journal of the American Statistical Association*, Vol. 102, pp.632–641.

Fan, J. and Li, R. (2001) 'Variable selection via nonconcave penalized likelihood and its oracle properties', *Journal of the American Statistical Association*, Vol. 96, pp.1348–1360.

Fan, J. and Li, R. (2004) 'New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis', *Journal of the American Statistical Association*, Vol. 99, pp.710–723.

Gibbons, R.D. and Hedeker, D. (1997) 'Random effects probit and logistic regression models for three level data', *Biometrics*, Vol. 53, pp.1527–1537.

Harrison, W.T. and Horngren, C.T. (2001) *Financial Accounting*, New York, NY: Prentice Hall.

Heagerty, P.J. and Zeger, S.L. (1996) 'Marginal regression models for clustered ordinal measurements', *Journal of the American Statistical Association*, Vol. 91, pp.1024–1036.

Hsiao, C. (2003) *Analysis of Panel Data*, Cambridge: Cambridge University Press.

Jiang, G. and Wang, H. (2008) 'Should earnings thresholds be used as delisting criteria in stock market?', *Journal of Accounting and Public Policy* (to appear).

Liang, K.Y. and Zeger, S.L. (1986) 'Longitudinal data analysis using generalized linear models', *Biometrika*, Vol. 73, pp.13–22.

Liu, L.C. and Hedeker, D. (2006) 'A mixed-effects regression model for longitudinal multivariate ordinal data', *Biometrics*, Vol. 62, pp.261–268.

Molenberghs, G. and Lesaffre, E. (1994) 'Marginal modeling of correlated ordinal data using a multivariate plackett distribution', *Journal of the American Statistical Association*, Vol. 89, pp.633–644.

Pinheiro, J.C. and Bates, M.B. (2000) *Mixed Effects Models in S and S-Plus*, New York: Springer.

Qu, Y., Piedmonte, M.R. and Medendorp, S.V. (1995) 'Latent variable models for clustered ordinal data', *Biometrics*, Vol. 51, pp.268–275.

Qu, Y., Williams, G.W., Beck, G.J. and Medendorp, S.V. (1992) 'Latent variable models for clustered dichotomous data with multiple subclusters', *Biometrics*, Vol. 48, pp.1095–1102.

Wang, H. (2007) 'A note on iterative marginal optimization: a simple algorithm for maximum rank correlation estimation', *Computational Statistics & Data Analysis*, Vol. 51, pp.2803–2812.

Wang, H., Li, G. and Jiang, G. (2007a) 'Robust regression shrinkage and consistent variable selection via the LAD-LASSO', *Journal of Business and Economic Statistics*, Vol. 25, pp.347–355.

Wang, H., Li, G. and Tsai, C.L. (2007b) 'Regression coefficient and autoregressive order shrinkage and selection via lasso', *Journal of Royal Statistical Society, Series B*, Vol. 69, pp.63–78.

Wooldridge, J.M. (2001) *Econometric Analysis of Cross Section and Panel Data*, Cambridge, MA: The MIT Press.

Xia, Y. (2006) *Asymptotic Distributions of Two Estimators of the Single-Index Model, Econometric Theory*, Vol. 22, pp.1112–1137.

Xia, Y. and Li, W.K. (1999) 'On single-index coefficient regression models', *Journal of the American Statistical Association*, Vol. 94, pp.1275–1285.

Zou, H. and Yuan, M. (2008) 'Composite quantile regression and the oracle model selection theory', *The Annals of Statistics* (to appear).

Note

1 http://www.ccerdata.com/

Appendix A

Proof of Theorem 1

The basic idea of the proof follows that of Fan and Li (2001). More specifically, one can verify that objective function (2.4) is a strictly concave function in β. Thus, as long as we can show that there exists a $\sqrt{n}$-consistent local maximiser, it must be the $\sqrt{n}$-consistent global maximiser. As in Fan and Li (2001), the existence of a $\sqrt{n}$-consistent local maximiser is implied by the fact that for an arbitrarily small $\epsilon > 0$, there exists a sufficiently large constant C such that:

$$\lim_{n} \inf P\left\{ \sup_{\mathbf{u}\in\mathbb{R}^d:\|\mathbf{u}\|=C} \mathcal{L}\left(\beta_0 + n^{-1/2}u\right) < \mathcal{L}\left(\beta_0\right)\right\} > 1-\epsilon \tag{A.1}$$

where $\mathbf{u} = (u1, \cdots, up)^\top \in \mathbb{R}^p$ and $\|\cdot\|$ stands for the typical L_2 norm. By Taylor expansion, we have:

$$\mathcal{L}\left(\beta_0 + n^{-1/2}\mathbf{u}\right) - \mathcal{L}\left(\beta_0\right) = \left(\sum_{k=1}^{K} W_{nk}\right)^\top \mathbf{u} - \frac{1}{2}\mathbf{u}\left(\sum_{k=1}^{K}\Delta_{nk}\right)\mathbf{u} + o_p(1) \tag{A.2}$$

$$\leq \left\|\sum_{k=1}^{K} W_{nk}\right\| C - \frac{1}{2}\lambda_{\min}\left(\sum_{k=1}^{K}\Delta_{nk}\right)C^2 + o_p(1) \tag{A.3}$$

where $\lambda_{\min}(A)$ stands for the smallest eigenvalue of an arbitrary positive definite matrix $A, \mathbf{W}_{nk} = n^{-1/2}\sum_{i=1}^{n}\left(\mathbb{X}_i^\top \mathbf{Z}_{ik} - \eta_{ik}\right)$, and $\Delta_{nk} = n^{-1}\sum_{i=1}^{n}\delta_{ik}$ with:

$$\eta_{ik} = \frac{\sum_{\tilde{\mathbf{Z}}_{ik}\in S_{ik}}\left\{\mathbb{X}_i^\top \mathbf{Z}_{ik}\right\}\exp\left\{\tilde{\mathbf{Z}}_{ik}^\top \mathbb{X}_i\beta\right\}}{\sum_{\tilde{\mathbf{Z}}_{ik}\in S_{ik}}\exp\left\{\tilde{\mathbf{Z}}_{ik}^\top \mathbb{X}_i\beta\right\}}$$

$$\delta_{ik} = \frac{\sum_{\tilde{\mathbf{Z}}_{ik}\in S_{ik}}\left\{\mathbb{X}_i^\top \mathbf{Z}_{ik}\mathbf{Z}_{ik}^\top \mathbb{X}_i\right\}\exp\left\{\tilde{\mathbf{Z}}_{ik}^\top \mathbb{X}_i\beta\right\}}{\sum_{\tilde{\mathbf{Z}}_{ik}\in S_{ik}}\exp\left\{\tilde{\mathbf{Z}}_{ik}^\top \mathbb{X}_i\beta\right\}} - u_i u_i^\top.$$

Note that $\left\{\mathbb{X}_i^\top \mathbf{Z}_{ik}\right\}_{i=1}^{n}$ is a sequence of independent and identically distributed random vectors with a mean given by:

$$E\left\{\mathbb{X}_i^\top \mathbf{Z}_{ik} - \eta_{ik}\right\} = E\left\{E\left(\mathbb{X}_i^\top \mathbf{Z}_{ik} - \eta_{ik} \mid Z_{i1k}, \cdots, Z_{im_ik}\right)\right\} = 0. \tag{A.4}$$

Next, note that $E\left\{\mathbb{X}_i^\top \mathbf{Z}_{ik} - \eta_{ik}\right\}^2 \leq 2E\left\|\mathbb{X}_i^\top \mathbf{Z}_{ik}\right\|^2 + 2E\left\|\eta_{ik}\right\|^2$. Because both the quantities $\left\|\mathbb{X}_i^\top \mathbf{Z}_{ik}\right\|^2$ and $\left\|\eta_{ik}\right\|^2$ are bounded by $\sum_{t=1}^{m_i}\left\|X_{it}\right\|^2$, we know that $E\left\{\mathbb{X}_i^\top \mathbf{Z}_{ik} - \eta_{ik}\right\}^2 \leq 4E\left\{\sum_{t=1}^{m_i}\left\|X_{it}\right\|^2\right\} < \infty$ by condition (C.1). Combining this result with (A.4) and applying the Central Limit Theorem, we know that $W_{nk} = O_p(1)$. Under the same regularity conditions and by the Law of Large Numbers, we know that

$\Delta_{nk} \to_p \Delta_k = E(\delta_{ik}), \sum \Delta_{nk} \to_p \sum \Delta_k$, and $\lambda_{\min}\left(\sum \Delta_{nk}\right) \to_p \lambda_{\min}\left(\sum_k \delta_{ik}\right) > 0$ by (C.2). Thus, we know that the first term in (A.3) is linear in C with a $O_p(1)$ coefficient but the second term in (A.3) is quadratic in C. Thus, as long as C is sufficiently large, the second term in (A.3) will dominate its first term asymptotically. Thus, the conclusion (A.1) is correct. This proves the $\sqrt{n}$-consistency of $\hat{\beta}$ and completes the proof.

Appendix B

Proof of Theorem 2

Because $\hat{\beta}$ is the maximiser of objective function (2.4), it must satisfy the following normal equation:

$$\sum_{k=1}^{K}\sum_{i=1}^{n}\left(\mathbb{X}_i^{\top}\mathbf{Z}_{ik}-\hat{\eta}_{ik}\right)=0 \tag{A.5}$$

where:

$$\hat{\eta}_{ik}=\frac{\sum_{\tilde{\mathbf{Z}}_{ik}\in S_{ik}}\left\{\mathbb{X}_i^{\top}\mathbf{Z}_{ik}\right\}\exp\left\{\tilde{\mathbf{Z}}_{ik}^{\top}\mathbb{X}_i\hat{\beta}\right\}}{\sum_{\tilde{\mathbf{Z}}_{ik}\in S_{ik}}\exp\left\{\tilde{\mathbf{Z}}_{ik}^{\top}\mathbb{X}_i\hat{\beta}\right\}}.$$

By Theorem 1 we know that $\hat{\beta}-\beta_0=o_p(1),$ thus the difference between $\hat{\eta}_{ik}$ and η_{ik} can be approximated by the following Taylor's expansion:

$$\hat{\eta}_{ik}-\eta_{ik}=\Delta_{nk}(\hat{\beta}-\beta_0)\{1+o_p(1)\}. \tag{A.6}$$

Applying (A.6) back to (A.5), we get:

$$\sqrt{n}\left(\hat{\beta}_0-\beta_0\right)=\left\{\sum_{k=1}^{K}\Delta_{nk}\right\}^{-1}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{k=1}^{K}\left(\mathbb{X}_i^{\top}\mathbf{Z}_{ik}-\eta_{ik}\right)\right\}\{1+o_p(1)\}.$$

By the Law of Large Numbers, we know that $\Delta_{nk}\to_p \text{var}\left\{\mathbb{X}_i^{\top}\mathbf{Z}_{ik}-\eta_{ik}\right\}.$ Next, by Slusky's theory and the Central Limit Theorem, we have $\sqrt{n}\left(\hat{\beta}_0-\beta_0\right)\to_d N(0,\Sigma),$ where:

$$\Sigma=\sum_{k=1}^{K}\text{var}\left\{\mathbb{X}_i^{\top}\mathbf{Z}_{ik}-\eta_{ik}\right\}^{-1}\text{var}\left(\sum_{k=1}^{K}\left\{\mathbb{X}_i^{\top}\mathbf{Z}_{ik}-\eta_{ik}\right\}\right)\left(\sum_{k=1}^{K}\text{var}\left\{\mathbb{X}_i^{\top}\mathbf{Z}_{ik}-\eta_{ik}\right\}\right)^{-1}.$$

Thus the theorem conclusion is correct and completes the proof.