A ROBUST PROCEDURE FOR GEE MODEL

By Netti Herawati

WORD COUNT 2667
TIME SUBMITTED 17-SEP-2019 04:17PM
PAPER ID 50419675

Far East Journal of Mathematical Sciences (FJMS)

http://www.pphmj.com
http://dx.doi.org/10.17654/MS102030645
Volume 102, Number 3, 2017, Pages 645-654 ISSN: 0972-0871

A ROBUST PROCEDURE FOR GEE MODEL

Netti Herawati and Khoirin Nisa

Department of Mathematics
University of Lampung
Indonesia

Abstract

In longitudinal studies, multiple measurements are taken on the same subject at different points in time. Thus, observations for the same subject are correlated. This paper proposes a robust procedure for estimating parameters of regression model when generalized estimating equation (GEE) applied to longitudinal data that contains outliers. The procedure is a combination of the iteratively reweighted least square (IRLS) and least trimmed square (LTS) methods and is called iteratively reweighted least trimmed square (IRLTS). We conducted a simulation study for gamma model and Poisson model using the proposed method, the result shows that our approach can provide a better result than the classical GEE.

1. Introduction

In statistics, generalized estimating equation (GEE) [5] is used to estimate the parameters of a generalized linear model (GLM) [6] with a possible unknown correlation between outcomes. It is a general statistical approach to fit a marginal model for longitudinal data analysis, and it has been popularly applied into clinical trials and biomedical studies. GEEs belong to a class of regression techniques that are referred to as

Received: March 24, 2017; Accepted: June 5, 2017
2010 Mathematics Subject Classification: 62J12.
Keywords and phrases: longitudinal data, outlier, regression model.

semiparametric because they rely on specification of only the first two moments. Under correct model specification and mild regularity conditions, parameter estimates from GEEs are consistent. The generalized estimating equation approach requires correct specification of the first two moments of a model. However, these moment assumptions can be distorted by contaminated or irregular measurements namely outliers. As a result, the generalized estimating equation method fails to give consistent estimators, and more seriously this will lead to incorrect conclusions [1, 8]. In this situation, we need a robust method that can minimize the effect of outliers.

In recent years a few studies have considered robust methods for longitudinal data analysis, see e.g. [1, 2, 4, 8, 11]. In this paper, we combine the IRLS and LTS for obtaining a robust estimation of GEE when data contain outliers. We have shown the effectiveness of this procedure for normal model [7]. In this paper we apply the proposed procedure to gamma and Poisson models.

2. Generalized Estimating Equation

Let the vector of measurements on the ith subject be $\mathbf{Y}_i = [Y_{i1}, \ldots, Y_{ini}]^T$ with corresponding vector of means $\boldsymbol{\mu}_i = [\mu_{i1}, \ldots, \mu_{ini}]^T$ and $\mathbf{X}_i = [X_{i1}, \ldots, X_{ini}]^T$ be the $n_i \times p$ matrix of covariates. In general, the components of $\mathbf{Y}_i$ are correlated but $\mathbf{Y}_i$ and $\mathbf{Y}_k$ are independent for any $i \neq k$. To model the relation between the response and covariates, we can use a regression model similar to the generalized linear models:

$$g(\boldsymbol{\mu}_i) = \boldsymbol{\eta}_i = \mathbf{X}_i\boldsymbol{\beta},$$

where $\boldsymbol{\mu}_i = \mathrm{E}(\mathbf{Y}_i | \mathbf{X}_i)$, g is a specified link function, and $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]^T$ is a vector of unknown regression coefficients to be estimated. The GEE for estimating the $p \times 1$ vector of regression parameter $\boldsymbol{\beta}$ is given by:

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \mathbf{V}_i^{-1} [\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})] = 0, \qquad (1)$$

where $\mathbf{V}_i$ be the covariance matrix of $\mathbf{Y}_i$ modeled as $\mathbf{V}_i = \lambda \mathbf{A}_i^{1/2} \mathbf{R}(\boldsymbol{\alpha}) \mathbf{A}_i^{1/2}$, $\mathbf{A}_i$ is a diagonal matrix of variance functions $v(\mu_{ij})$, and $\mathbf{R}(\alpha)$ is the working correlation matrix of $\mathbf{Y}_i$ indexed by a vector of parameters $\boldsymbol{\alpha}$. Solution to equation (1) are obtained by alternating between estimation of λ and $\boldsymbol{\alpha}$. There are several specific choices of the form of working correlation matrix $\mathbf{R}_i(\boldsymbol{\alpha})$ commonly used to model the correlation matrix of $\mathbf{Y}_i$, among them are *exchangeable* and *autoregressive* correlation matrices.

Solving for $\boldsymbol{\beta}$ is done with iteratively reweighted least squares (IRLS). The following is an algorithm for fitting the specified model using GEEs as described in [3] and [8]:

1. Compute an initial estimate of $\hat{\boldsymbol{\beta}}_{GEE}$, for example with an ordinary generalized linear model assuming independence.
2. A current estimate $\hat{\boldsymbol{\beta}}_{GEE}$ is updated by regressing the working response vector $\mathbf{Z}^* = \mathbf{X}\hat{\boldsymbol{\beta}} + \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}}(\mathbf{y} - \hat{\boldsymbol{\mu}})$ on $\mathbf{X}$. A new estimate $\hat{\boldsymbol{\beta}}_{new}$ is obtained by:

$$\hat{\boldsymbol{\beta}}_{new} = (\mathbf{X}^{\mathrm{T}} \mathbf{W}^* \mathbf{X})^{-1} \mathbf{X}^{\mathrm{T}} \mathbf{W}^* \mathbf{Z}^*,$$

   where $\mathbf{W}^*$ is a block diagonal weight matrix whose ith block is the $n_i \times n_i$ matrix $\mathbf{W}_i^* = \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)^{-1} \mathbf{A}_i^{-1} \mathbf{R}_i(\hat{\boldsymbol{\alpha}}) \mathbf{A}_i^{-1} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)^{-1}$.
3. Use $\hat{\boldsymbol{\beta}}_{new}$ to update $\hat{\boldsymbol{\eta}} = \mathbf{X}\hat{\boldsymbol{\beta}}_{new} = \mathbf{H}\mathbf{Z}^*$, where

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}} \mathbf{W}^* \mathbf{X})^{-1} \mathbf{X}^{\mathrm{T}} \mathbf{W}^*.$$

4. Iterate until convergence.

3. Iterated Reweighted Least Trimmed Square

Let us briefly recall that the robust estimation of regression parameters

using LTS [9] method is given by:

$$\hat{\boldsymbol{\beta}}_{LTS} = \arg\min \sum_{i=1}^{h} e_i^2,$$

where $e_1^2 \le e_2^2 \le \cdots \le e_h^2 \le \cdots \le e_n^2$ are the ordered squared residuals, from smallest to largest. LTS is calculated by minimizing the h ordered squares residuals, where h can be chosen between the range $\frac{n}{2}+1 \le h \le \frac{3n}{4}+\frac{n+1}{4}$, with n being sample size and number of parameters, respectively. One can refer to e.g. [9, 10] for some details on LTS method.

The IRLTS procedure is stated in the following short algorithm. To motivate this method, it is convenient to write the algorithm with involving the residuals.

1. Compute an initial estimate of $\hat{\boldsymbol{\beta}}_{GEE}$ using IRLS, use the estimate to calculate fitted value: $\hat{\boldsymbol{\mu}}_i = g^{-1}(\mathbf{X}_i\hat{\boldsymbol{\beta}})$.
2. Calculate residuals: $e_{ij} = Y_{ij} - \hat{\mu}_{ij}$. Sort $|e_{ij}|$ for $i = 1, 2, \ldots, t$ and $j = 1, 2, \ldots, n$ in ascending order: $|e_{11}| \le |e_{12}| \le \cdots \le |e_m|$.
3. Choose h observations which have the lowest h-residuals, we denote as subset H.
4. Improve estimates of β by solving $\hat{\boldsymbol{\beta}}_{new}$ based on subset H using IRLS.
5. Iterate until convergence.

4. Simulation Study

We compare the performances of IRLTS and IRLS through simulation study. Two types of outcomes are considered, continuous and count responses. The models for data generation are as follows:

$$1/\mu_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij},$$

$$Log(\mu_{ij}) = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij},$$

where β_k's for $k = 0, 1, 2$ are randomly generated, $i = 1, 2, ..., 200$ and $j = 1, 2, ..., 5$. The covariates x_{1ij} are i.i.d. from a uniform distribution Unif$(1, 5)$ and x_2 is the measurement time variable, i.e., $x_{2i} = 1, 2, 3, 4, 5$. For each scenario, we generate the data based on the underlying true correlation structures as exchangeable and autoregressive with $\alpha = 0.5$. For the first model (inverse link) the gamma distributed model was used, and for the second model (log link) the Poisson distributed model was applied. In this simulation, 1000 Monte Carlo data sets were generated for each scenario. We considered contamination proportion in data $\varepsilon = 5\%, 10\%, 20\%$ and 30%. We evaluated the results using the mean square error (MSE) of the parameter estimates.

We provide the expected values and MSEs of parameter estimates resulted from our simulation on Table 1-Table 4. Table 1 and Table 2 show the expected values and MSEs of parameter estimates for the first model, while Table 3 and Table 4 for the second model.

Table 1. The expected values, standard errors and MSEs of $\hat{\beta}_i$ for gamma distributed model with exchangeable correlation matrix

Method	Classical GEE			IRLTS		
	$\hat{\beta}_0$					
ε	$E(\hat{\beta}_0)$	$SE(\hat{\beta}_0)$	MSE	$E(\hat{\beta}_0)$	$SE(\hat{\beta}_0)$	MSE
5%	1.833421	0.417356	0.419098	1.292539	0.280084	0.080563
10%	1.963008	0.551804	0.694453	1.343505	0.345998	0.119739
20%	2.105067	0.522999	0.861099	1.302799	0.245013	0.061308
30%	2.158723	0.712847	1.180859	1.432421	0.581205	0.346613

			$\hat{\beta}_1$			
ε	$E(\hat{\beta}_1)$	$SE(\hat{\beta}_1)$	MSE	$E(\hat{\beta}_1)$	$SE(\hat{\beta}_1)$	MSE
5%	0.664892	0.297197	0.205406	0.957103	0.034595	0.003693
10%	0.587415	0.348410	0.297492	0.968876	0.075239	0.007119
20%	0.314225	0.347369	0.600688	0.964959	0.039372	0.003323
30%	0.297561	0.338291	0.617832	0.937923	0.088894	0.012682
			$\hat{\beta}_2$			
ε	$E(\hat{\beta}_2)$	$SE(\hat{\beta}_2)$	MSE	$E(\hat{\beta}_2)$	$SE(\hat{\beta}_2)$	MSE
5%	0.743453	0.289166	0.174850	0.974127	0.039936	0.006689
10%	0.549449	0.423348	0.425292	0.958451	0.018566	0.007922
20%	0.355655	0.349199	0.597828	0.975630	0.032734	0.005954
30%	0.332191	0.336312	0.621917	0.969043	0.071512	0.010960

Table 2. The expected values, standard errors and MSEs of $\hat{\beta}_i$ for gamma distributed model with autoregressive correlation matrix

Method	Classical GEE			IRLTS		
			$\hat{\beta}_0$			
ε	$E(\hat{\beta}_0)$	$SE(\hat{\beta}_0)$	MSE	$E(\hat{\beta}_0)$	$SE(\hat{\beta}_0)$	MSE
5%	1.713422	0.174403	0.127509	1.169724	0.219998	0.102270
10%	2.051958	0.292341	0.476604	1.172668	0.169948	0.081395
20%	2.096749	0.279746	0.561178	1.182726	0.281815	0.127424
30%	2.208355	0.570003	0.975395	1.094088	0.105856	0.105908
			$\hat{\beta}_1$			
ε	$E(\hat{\beta}_1)$	$SE(\hat{\beta}_1)$	MSE	$E(\hat{\beta}_1)$	$SE(\hat{\beta}_1)$	MSE
5%	0.746269	0.169773	0.087549	0.977105	0.032371	0.001180
10%	0.493787	0.257591	0.311197	0.974237	0.026179	0.000892
20%	0.290721	0.257720	0.553459	0.982542	0.035247	0.001279
30%	0.228004	0.231963	0.595925	0.981851	0.023087	0.000579
			$\hat{\beta}_2$			
ε	$E(\hat{\beta}_2)$	$SE(\hat{\beta}_2)$	MSE	$E(\hat{\beta}_2)$	$SE(\hat{\beta}_2)$	MSE
5%	0.739142	0.224797	0.132593	0.981539	0.021532	0.002405
10%	0.511909	0.302973	0.355672	0.981291	0.028264	0.002762
20%	0.281145	0.270400	0.627330	0.983924	0.028529	0.002551
30%	0.225163	0.377600	0.646722	1.003760	0.035796	0.001758

As shown in Table 1 and Table 2, our approach (IRLTS) performs better than the classical GEE. The MSEs of IRLTS are smaller than the MSEs of classical GEE, the outliers influence the estimation of $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$. The parameter estimates of classical GEE are much more influenced than the parameter estimates of IRLTS. The more outliers contained in the data the larger the deviation of classical GEE estimates from the parameter value. In Table 3 and Table 4, the behavior of MSEs of both methods is the same as the first case, here we can see that IRLTS performs better than the classical GEE because the MSEs of IRLTS are smaller than the MSEs of classical GEE.

Table 3. The expected values, standard errors and MSEs of $\hat{\beta}_i$ for Poisson distributed model with exchangeable correlation matrix

Method	Classical GEE			IRLTS		
			$\hat{\beta}_0$			
ε	$E(\hat{\beta}_0)$	$SE(\hat{\beta}_0)$	MSE	$E(\hat{\beta}_0)$	$SE(\hat{\beta}_0)$	MSE
5%	7.851928	0.578389	49.012922	2.242959	0.411831	1.041084
10%	9.299508	0.776268	71.575965	2.084935	0.454712	1.670854
20%	10.710302	0.425014	96.915019	2.351178	0.546663	1.478125
30%	11.413164	0.621123	111.439999	1.988002	0.486656	1.475743
			$\hat{\beta}_1$			
ε	$E(\hat{\beta}_1)$	$SE(\hat{\beta}_1)$	MSE	$E(\hat{\beta}_1)$	$SE(\hat{\beta}_1)$	MSE
5%	0.309635	0.108437	0.418891	1.052717	0.059291	0.014543
10%	0.194983	0.044670	0.568586	1.011790	0.074132	0.009602
20%	0.097441	0.077894	0.729016	0.918343	0.038557	0.002349
30%	0.092040	0.031461	0.733153	0.809874	0.101063	0.029211
			$\hat{\beta}_2$			
ε	$E(\hat{\beta}_2)$	$SE(\hat{\beta}_2)$	MSE	$E(\hat{\beta}_2)$	$SE(\hat{\beta}_2)$	MSE
5%	0.281547	0.064601	0.478304	0.947537	0.021808	0.000986
10%	0.159566	0.060507	0.660655	0.967633	0.065401	0.004284
20%	0.087528	0.031481	0.779957	0.900333	0.028748	0.005696
30%	0.067109	0.035320	0.816673	0.825752	0.057643	0.024164

The result for Poisson model shows similar behavior to the result for gamma model. For the result of Poisson model in Table 3 and Table 4, IRLTS also performs better than the classical GEE. The MSEs of IRLTS are smaller than the MSEs of classical GEE, the outliers influence the estimation of $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$. The parameter estimates of classical GEE are much more influenced than the parameter estimates of IRLTS. The more outliers contained in the data the larger the deviation of classical GEE estimates from the parameter value. In Table 3 and Table 4, the behavior of MSEs of both methods is the same as the first case, here we can see that IRLTS performs better than the classical GEE.

The estimation of IRLTS yields better results than classical GEE for both cases we considered here. The MSEs of IRLTS is smaller than classical GEE, this means that IRLTS can reduce the influence of the high leverage points better than the classical GEE.

Table 4. The expected values, standard errors and MSEs of $\hat{\beta}_i$ for Poisson distributed model with autoregressive correlation matrix

Method	Classical GEE			IRLTS		
			$\hat{\beta}_0$			
ε	$E(\hat{\beta}_0)$	$SE(\hat{\beta}_0)$	MSE	$E(\hat{\beta}_0)$	$SE(\hat{\beta}_0)$	MSE
5%	8.493574	0.689463	40.055071	1.614313	0.331758	0.455831
10%	9.788645	0.962589	58.478701	1.765065	0.204619	0.523307
20%	10.525858	0.877176	70.050502	2.356166	0.470975	0.724548
30%	10.635832	0.570803	71.449706	3.071576	0.974599	1.054251
			$\hat{\beta}_1$			
ε	$E(\hat{\beta}_1)$	$SE(\hat{\beta}_1)$	MSE	$E(\hat{\beta}_1)$	$SE(\hat{\beta}_1)$	MSE
5%	0.309635	0.108437	0.418891	1.052717	0.059291	0.014543
10%	0.194983	0.044670	0.568586	1.011790	0.074132	0.009602
20%	0.097441	0.077894	0.729016	0.918343	0.038557	0.002349
30%	0.092040	0.031461	0.733153	0.809874	0.101063	0.029211

	$\hat{\beta}_2$					
ε	$E(\hat{\beta}_2)$	$SE(\hat{\beta}_2)$	MSE	$E(\hat{\beta}_2)$	$SE(\hat{\beta}_2)$	MSE
5%	0.281547	0.064601	0.478304	0.947537	0.021808	0.000986
10%	0.159566	0.060507	0.660655	0.967633	0.065401	0.004284
20%	0.087528	0.031481	0.779957	0.900333	0.028748	0.005696
30%	0.067109	0.035320	0.816673	0.825752	0.057643	0.024164

5. Concluding Remark

In this paper, we have shown that our proposed procedure can minimize the effect of outliers on parameter estimation; IRLTS can produce a relatively efficient and consistent estimator compared to the classical GEE (IRLS). Based on the MSE, IRLTS performs much better than the classical GEE for gamma and Poisson models.

Acknowledgement

13

The authors would like to thank the anonymous referees for their valuable suggestions and comments that led to a considerably improved manuscript.

References

[1] A. Abebe, J. W. McKean, J. D. Kloke and Y. Bilgic, Iterated reweighted rank-based estimates for GEE models, Technical Report, 2014.

[2] P. S. Gill, A robust mixed linear model analysis for longitudinal data, Stat. Med. 19 (2000), 975-987.

[3] G. Johnston and M. Stokes, Repeated measures analysis with discrete data using the SAS system, SUGI Proceeding, SAS Institute Inc., Cary, NC, 1996.

[4] S. H. Jung and Z. Ying, Rank-based regression with repeated measurements data, Biometrika 90 (2003), 732-740.

[5] K. Y. Liang and S. L. Zeger, Longitudinal data analysis using generalized linear models, Biometrika 73 (1986), 13-22.

[6] P. McCullagh and J. A. Nelder, Generalized Linear Models, Chapman and Hall, London, 1989.

[7] K. Nisa and N. Herawati, Robust estimation of generalized estimating equation when data contain outliers, INSIST 2 (2017), 1-5.

[8] B. F. Qaqish and J. S. Preisser, Resistant fits for regression with correlated outcomes: an estimating equations approach, J. Statist. Plann. Inference 75(2) (1999), 415-431.

[9] R. J. Rousseeuw and A. M. Leroy, Robust Regression and Outlier Detection, Wiley, New York, 1987.

[10] P. J. Rousseeuw and K. van Driessen, Computing LTS regression for large data sets, Data Mining and Knowledge Discovery 12 (2006), 29-45.

[11] Y. G. Wang and M. Zhu, Rank-based regression for analysis of repeated measures, Biometrika 93 (2006), 459-464.

A ROBUST PROCEDURE FOR GEE MODEL

ORIGINALITY REPORT

22%

SIMILARITY INDEX

PRIMARY SOURCES

1	Bahjat F. Qaqish, John S. Preisser. "Resistant fits for regression with correlated outcomes an estimating equations approach", Journal of Statistical Planning and Inference, 1999 Crossref	60 words — 2%
2	Ming Wang. "Generalized Estimating Equations in Longitudinal Data Analysis: A Review and Recent Developments", Advances in Statistics, 2014 Crossref	59 words — 2%
3	en.wikipedia.org Internet	59 words — 2%
4	www.stats.uwaterloo.ca Internet	47 words — 2%
5	www.stat.auckland.ac.nz Internet	46 words — 2%
6	Wei Pan. "Akaike's Information Criterion in Generalized Estimating Equations", Biometrics, 2004 Crossref	42 words — 2%
7	repository.unhas.ac.id Internet	40 words — 2%
8	Keon Lee, S.K.. "On generalized multivariate decision tree by using GEE", Computational Statistics and Data Analysis, 20050615 Crossref	31 words — 1%
9	www.m-hikari.com	

Internet

29 words — 1%

10 Bijan Pirnia, Kambiz Pirnia, Rozita Ershad Sarabi, Parastoo Malekanmehr, Leila Soltani, Alireza Zahiroddin, Paria Sadeghi. "A Double-Blind Randomized Controlled Trial in Effectiveness of Parent-Child Interaction Therapy on Psychological Indicator and Cortisol Level in Children of Caregiver with Cancer", International Journal of Cancer Management, 2019
Crossref

27 words — 1%

11 Pan, W.. "Sample Size and Power Calculations with Correlated Binary Data", Controlled Clinical Trials, 200104
Crossref

15 words — 1%

12 Alireza Hadayeghi, Brian Malone, Jeff Suggett, Jeffrey Reid. "Identification of Intersections with Promise for Red Light Camera Safety Improvement", Transportation Research Record: Journal of the Transportation Research Board, 2007
Crossref

14 words — 1%

13 doaj.org
Internet

12 words — < 1%

14 www.nice.org.uk
Internet

11 words — < 1%

15 www2.sas.com
Internet

11 words — < 1%

16 You-Gan Wang. "Robust Estimating Functions and Bias Correction for Longitudinal Data Analysis", Biometrics, 9/2005
Crossref

10 words — < 1%

17 Torti, F.. "Benchmark testing of algorithms for very robust regression: FS, LMS and LTS", Computational Statistics and Data Analysis, 201208
Crossref

8 words — < 1%

18	www.archive.org Internet	8 words — < 1%
19	Khoirin Nisa, Netti Herawati. "Robust Estimation of Generalized Estimating Equation when Data Contain Outliers", INSIST, 2017 Crossref	6 words — < 1%

EXCLUDE QUOTES	ON	EXCLUDE MATCHES	OFF
EXCLUDE BIBLIOGRAPHY	ON		