GRAPHICAL DIAGNOSTIC METHODS IN GEE

Pardo, Maria del Carmen and Alonso, Rosa
Complutense University of Madrid, Department of Statistic and O.R.
Plaza de las Ciencias, 3
Madrid (28040), Spain
mcapardo@mat.ucm.es and ralonsos@mat.ucm.es

keywords and phrases: *Generalized estimating equations, outliers, $\phi-$divergence residuals, Q-Q plot.*

1 Introduction

Correlated or clustered data are frequently encountered in medical and biological research. The method of generalized estimating equations (GEE) is often used to analyze these types of data. Liang and Zeger [2] presented this approach, which is an extension of generalized linear models to analysis of longitudinal data via quasi-likelihood methods. It requires the researcher to specify only the model for the marginal mean, the distribution of the dependent variable and a working covariance matrix for the vector of measurements from each subject. The GEE allows for correlation without explicitly defining a model for the origin of the dependency.

Model checking is an important aspect of regression analysis. Therefore GEE approach also needs diagnostic procedures for checking the model's adequacy and for detecting outliers and influential observations. Graphical diagnostic displays can be useful for detecting and examining anomalous features in the fit of a model to data. For correlated binary data, Tan et al. [6] proposed several graphical methods. Oh et al. [3] proposed residual plots to investigate the goodness-of fit of the GEE approach for discrete data. They investigated Pearson, Anscombe and deviance residuals for Poisson and binary responses. In this work, we propose to generalize these plots using a family of residuals based on the ϕ-divergence measures which contains the Pearson and Deviance residuals. The graphical methods are illustrated with a real life example.

2 Overview of GEE

Let $\mathbf{y}_i = (y_{i1}, ..., y_{it_i})^T$, $i = 1, ..., n$, be mutually independent random vectors of repeated outcomes and $\mathbf{X}_i = (\mathbf{x}_{i1}, ..., \mathbf{x}_{it_i})^T$ be the $t_i \times p$ matrix of covariate values for the ith subject, with $x_{ij} = (x_{ij1}, ..., x_{ijp})^T$, $i = 1, ..., n$ and $j = 1, ..., t_i$. The marginal density of y_{ij} is assumed to be a member of a generalized exponential family with a scale parameter γ, so the mean and the variance are given by

$$E(y_{ij}) = \mu_{ij} \quad \text{and} \quad Var(y_{ij}) = \gamma^{-1} v(\mu_{ij})$$

where $v(\mu_{ij})$ is a known function of the mean μ_{ij}.

Suppose that the regression model is $\eta_i = g(\mu_{ij}) = x_{ij}\boldsymbol{\beta}$, where $g(\cdot)$ is a link function and $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ is the $p \times 1$ vector of unknown parameters to be estimated.

The GEE approach consists of two estimation steps:

1. A quasi-likelihood method for estimating regression parameters, $\boldsymbol{\beta}$, which characterize the de-

pendence of outcomes on the covariates

$$\sum_{i=1}^{n}\left(\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}}\right)^T\mathbf{V}_i^{-1}\left(\mathbf{y}_i-\boldsymbol{\mu}_i\right)=\mathbf{0}$$

with $\boldsymbol{\mu}_i=\left(\mu_{i1},...,\mu_{it_i}\right)^T$ and $\mathbf{V}_i$ is a $t_i\times t_i$ covariance matrix of $\mathbf{y}_i$

$$\mathbf{V}_i=\gamma\mathbf{A}_i^{\frac{1}{2}}\mathbf{R}_i\left(\boldsymbol{\alpha}\right)\mathbf{A}_i^{\frac{1}{2}}$$

where $\mathbf{A}_i = diag(v(\mu_{i1}),...,v(\mu_{it_i}))$ is a diagonal matrix and $\mathbf{R}_i\left(\boldsymbol{\alpha}\right)$ is a working correlation matrix for each subgroup repeated outcomes. There are several choices available for longitudinal data, see Zorn [7].

2. A robust moment method for estimating correlation parameters, $\boldsymbol{\alpha}$ and γ, which incorporates the dependence among outcomes. The estimation of correlation parameters is based upon the Pearson residuals.

3 Q-Q plot based on $\phi-$divergence residuals

Oh et al. [3] proposed residual plots to investigate the goodness of fit of the GEE approach, based on Pearson, Anscombe and deviance residuals. The proposed residual plots are based on the quantile-quantile (Q-Q) plots of a χ^2-distribution. We extend Oh et al. [3] Q-Q plot considering the $\phi-$divergence residuals c_{ij}^{ϕ} defined for binary outcomes, by Pardo et al. [4] as

$$c_{ij}^{\phi}=sig\left(y_{ij}-\hat{\mu}_{ij}\right)\sqrt{\frac{2}{\phi''\left(1\right)}}\left\{\hat{\mu}_{ij}\phi\left(\frac{y_{ij}}{\hat{\mu}_{ij}}\right)+\left(1-\hat{\mu}_{ij}\right)\phi\left(\frac{1-y_{ij}}{1-\hat{\mu}_{ij}}\right)\right\}^{1/2}$$

where $\phi\in\Phi^*$ and Φ^* is a class of convex functions $\phi:[0,\infty)\mapsto R\cup\{\infty\}$ such as in $x=1$, $\phi\left(1\right)=0$, $\phi''\left(1\right)>0$ and in $x=0$, $0\phi\left(0/0\right)=0$ and $0\phi\left(p/0\right)=p\lim_{u\to\infty}\phi\left(u\right)/u$. If we consider $\phi\left(x\right)=\frac{1}{2}\left(x-1\right)^2$ and $\phi\left(x\right)=x\log x-x+1$ we get the Pearson and the deviance residuals, respectively.

In this work, we focus on the family of residuals obtained considering the parametric family $\phi=\phi_{(\lambda)}$ proposed by Cressie and Read [1]

$$\phi_{(\lambda)}\left(x\right)=\frac{1}{\lambda\left(\lambda+1\right)}\left(x^{\lambda+1}-\left(x-1\right)\lambda\right);\ \lambda\neq 0,\ \lambda>-1$$

and

$$\phi_{(0)}\left(x\right)=\lim_{\lambda\longrightarrow 0}\phi_{(\lambda)}\left(x\right)=xlogx-x+1.$$

Note that the residuals $c_{ij}^{\phi(0)}$ and $c_{ij}^{\phi(1)}$ are the deviance and Pearson residuals, respectively.

Let $\mathbf{c}_i^{\phi}$ be the $t_i\times 1$ vector of the ith subject residual vector with jth component, c_{ij}^{ϕ}. Let

$\mathbf{M}_i=\mathbf{I}-\mathbf{H}_i$ be the asymptotic covariance matrix of $\mathbf{c}_i^{\phi}$,

where

$$\mathbf{H}_i=\mathbf{W}_i^{\frac{1}{2}}\mathbf{X}_i\mathbf{J}_1\mathbf{X}_i^T\mathbf{W}_i^{\frac{1}{2}}$$

with

$$\mathbf{W}_i = \boldsymbol{\Lambda}_i \mathbf{V}_i^{-1} \boldsymbol{\Lambda}_i, \quad \boldsymbol{\Lambda}_i = diag\left(\frac{d\boldsymbol{\mu}_i}{d\boldsymbol{\eta}_i^T}\right)$$

and

$$\mathbf{J}_1 = \left[\sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)^T \hat{\mathbf{V}}_i{}^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right]^{-1}.$$

When $\mathbf{M}_i$ is known and the mean model is correct, $q_i^{\phi} = \left(\mathbf{c}_i^{\phi}\right)^T \mathbf{M}_i^{-1}\, \mathbf{c}_i^{\phi},\ i = 1, ..., n$, are approximately distributed as a χ^2 with t_i degrees of freedom. That is,

$$q_i^{\phi} = \left(\mathbf{c}_i^{\phi}\right)^T \mathbf{M}_i^{-1} \mathbf{c}_i^{\phi} \sim \chi^2\left(t_i\right).$$

When the number of responses from the same subject is equal to t for all i, we can construct a Q-Q plot easily with observed q_i^{ϕ} using the χ^2-distribution. Let $q_{(1)}^{\phi} \leq ... \leq q_{(n)}^{\phi}$ be ordered values of q_i^{ϕ}. Then, q_i^{ϕ}, is in fact the empirical $100 \times \dfrac{i}{n}$ percentile, and from the χ^2-distribution with t degrees of freedom, we can obtain the corresponding quantiles $\psi_{(1)} \leq ... \leq \psi_{(n)}$. Then, the Q-Q plot is the graph of $\left(q_{(i)}^{\phi}, \psi_{(i)}\right)$ from which we can investigate the model fit and identify outliers.

4 Example for Binary responses

Koch et al. [5] presented a longitudinal study comparing a new drug and a standard drug for patients with both mild and severe mental depression. Patients in each diagnosis group are randomly assigned to two drugs and their condition is evaluated as normal (N) or abnormal (A) at the end of one week, two weeks, and four weeks of continuous treatment. Koch et al. [5] used time $(0,\ 1,\ 2)$, the logs to base 2 of week numbers $(1,\ 2,\ 4)$ for time as covariates. This dataset is shown in Table 1.

Table 1. Depression data

Tabulation of Responses for 3 times for diagnoses and treatment										
		Response profile at week 1 vs 2 vs week 4								
Diag	Treatment	NNN	NNA	NAN	NAA	ANN	ANA	AAN	AAA	Total
Mild	Standard	16	13	9	3	14	4	15	6	80
Mild	New drug	31	0	6	0	22	2	9	0	70
Severe	Standard	2	2	8	9	9	15	27	28	100
Severe	New drug	7	2	5	2	31	5	32	6	90

We consider the same models proposed by Oh et al. [3]

Model 1: $\log it(\mu_{ij}) = \beta_0,$
Model 2: $\log it(\mu_{ij}) = \beta_0 + \beta_1 \cdot Trt,$
Model 3: $\log it(\mu_{ij}) = \beta_0 + \beta_1 \cdot Trt + \beta_2 \cdot Diag,$
Model 4: $\log it(\mu_{ij}) = \beta_0 + \beta_1 \cdot Trt + \beta_2 \cdot Diag + \beta_3 \cdot Time$

We fit a GEE model with an exchangeable correlation structure.

Figures 1-4 show the Q-Q plots based on $\phi_{(-1/2)}-$divergence residuals, $\phi_{(0)}-$divergence residuals (Deviance residuals), $\phi_{(2/3)}-$divergence residuals and $\phi_{(1)}-$divergence residuals (Pearson residuals) for Model 4, Model 3, Model 2 and Model 1, respectively.
Note that the Q-Q plot based on $\phi_{(-1/2)}-$divergence residuals has an erratic behavior. Model 4 is the closest to the $y = x$. Although the Q-Q plots for $\phi_{(2/3)}-$divergence residuals and $\phi_{(0)}-$divergence residuals (Deviance residuals) in Figure 1, are more apart from $y = x$ line than that of the Pearson residuals ($\phi_{(1)}-$divergence residuals), they consistently show than the Model 4 is the most suitable.

Figure 1. Q-Q plots for the depression data for Model 4.

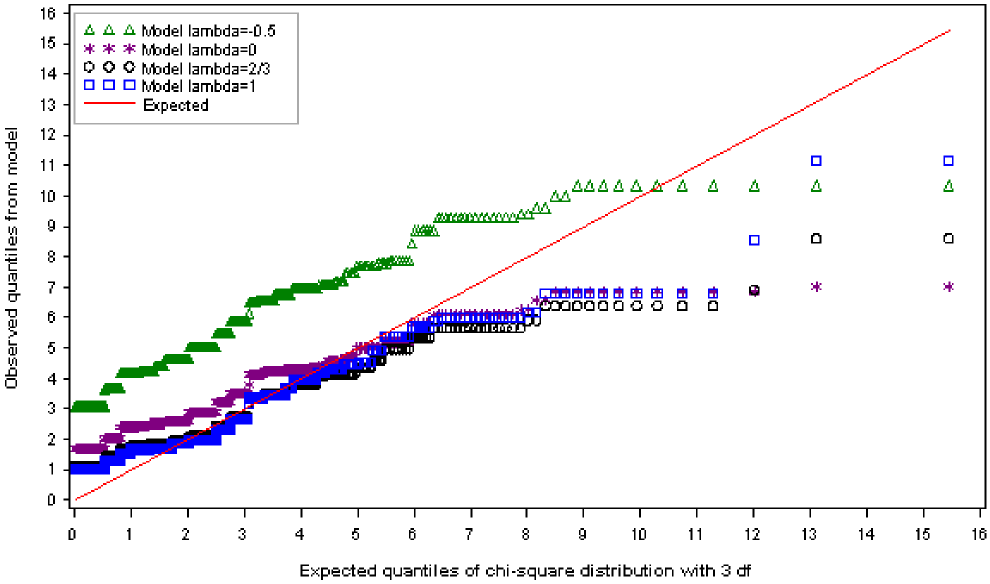
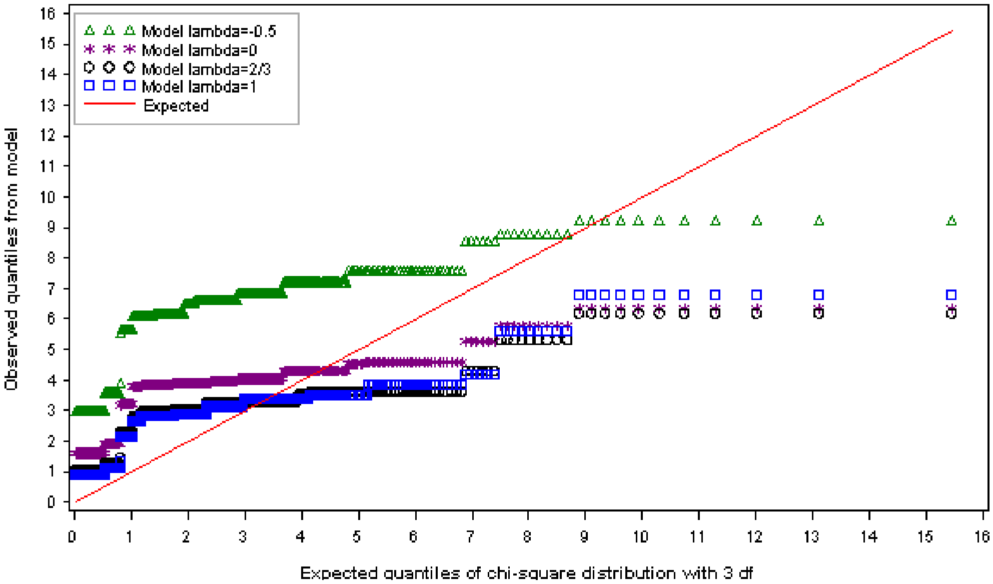


Figure 2. Q-Q plots for the depression data for Model 3.



Therefore, Pearson residuals provide a closer χ^2-distribution when the model fits the data well. The second best alternative is to use $\phi_{(2/3)}-$divergence residuals. However, to detect poor models, the use of a Q-Q plot based on $\phi_{(2/3)}-$divergence residuals is most effective as can be seen in Figure 2 than that based on Pearson residuals. Also both residuals are quite similar for Models 3 and 4 (see Figures 3 and 4).

In summary, the new Q-Q plot based on $\phi_{(2/3)}-$divergence residuals is at least as efficient as the Q-Q plot based on Pearson residuals. Nevertheless, further research is necessary.

Figure 3. Q-Q plots for the depression data for Model 2.

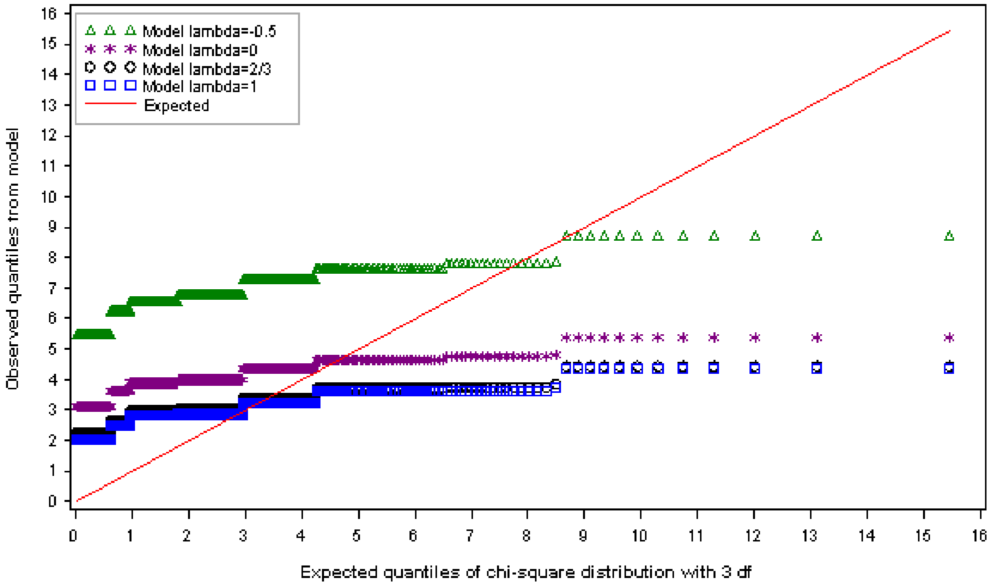
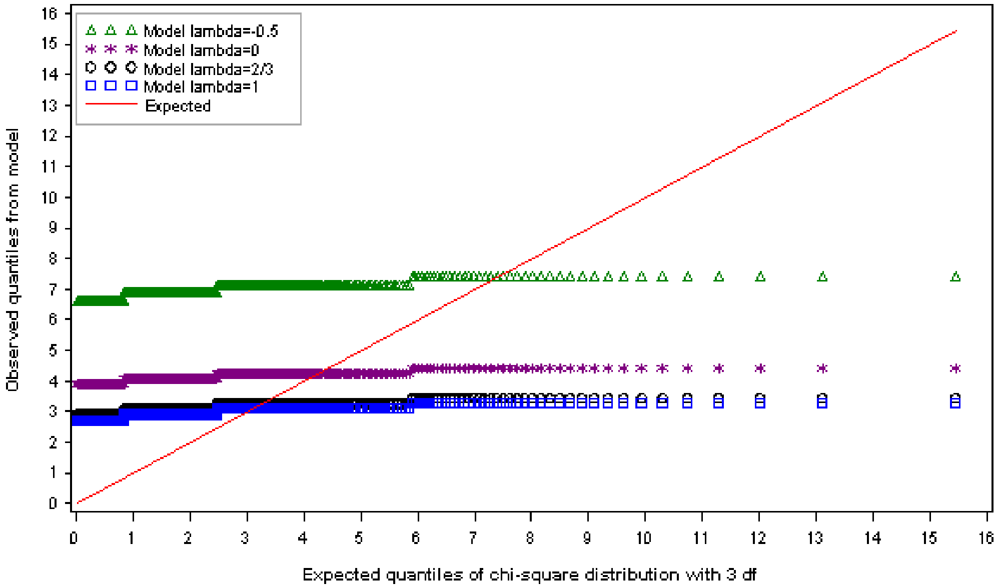


Figure 4. Q-Q plots for the depression data for Model 1.



Acknowledgments

This work was partially supported by Grant MTM2009-06997.

References

[1] Cressie, N. and Read, T.R.C. (1984). Multinomial goodness-of-fit test. *Journal of the Royal Statistical Society, Serie B*, 46, 440-464.

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

[3] Oh, S., Carriere, K.C. and Park, T. (2008). Model diagnostic plots for repeated measures data using the generalized estimating equations approach. *Computational Statistics and Data Analysis*, 53, 222-232.

[4] Pardo, J.A., Pardo, L. and Pardo, M.C. (2006). Testing in logistic regression models based on $\phi-$divergence measure. *Journal of Statistical Planning and Inference*, 136, 982-1006.

[5] Koch, G.G., Landis, J.R., Freeman, J.L., Freeman, D.H., Lehnen, R.G. (1977). A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics* 33, 133-158.

[6] Tan, M., Qu, Y. and Kutner, M.H. (1997). Model diagnostics for marginal regression analysis of correlated binary data. *Communication in Statistics Simulation*, 26, 539-558.

[7] Zorn, C.J.W. (2001). Generalized Estimating Equation Models for Correlated Data: A Review with Applications. *American Journal of Political Science*, 45, (2), 470-490.