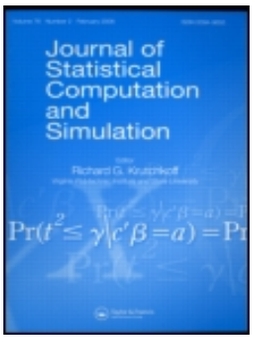

Journal of Statistical Computation and Simulation

ISSN: 0094-9655 (Print) 1563-5163 (Online) Journal homepage: www.tandfonline.com/journals/gscs20

# Diagnostic techniques in generalized estimating equations

Maria Kelly Venezuela, Denise Aparecida Botter & Mônica Carneiro Sandoval

**To cite this article:** Maria Kelly Venezuela, Denise Aparecida Botter & Mônica Carneiro Sandoval (2007) Diagnostic techniques in generalized estimating equations, Journal of Statistical Computation and Simulation, 77:10, 879-888, DOI: 10.1080/10629360600780488

**To link to this article:** https://doi.org/10.1080/10629360600780488

*Journal of Statistical Computation and Simulation*
Vol. 77, No. 10, October 2007, 879–888

Taylor & Francis
Taylor & Francis Group

# Diagnostic techniques in generalized estimating equations

MARIA KELLY VENEZUELA*, DENISE APARECIDA BOTTER and MÔNICA CARNEIRO SANDOVAL

Department of Statistics, Institute of Mathematics and Statistics, University of São Paulo, Caixa Postal 66281, CEP 05311-970, São Paulo, SP, Brazil

*(Received 2 April 2005; in final form 2 May 2006)*

We consider herein diagnostic methods for the quasi-likelihood regression models developed by Zeger and Liang [Zeger, S. L. and Liang, K.-Y., 1986, Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, **42**, 121–130.] to analyse discrete and continuous longitudinal data. Our proposal generalises well-known measures (projection matrix, Cook's distance and standardised residuals) developed for independent responses. Moreover, half-normal probability plots with simulated envelopes were developed for assessing the adequacy of the fitted model when the marginal distributions belong to the exponential family. To obtain such a plot, correlated outcomes were generated by simulation using algorithms described in the literature. Finally, two applications were presented to illustrate the techniques.

*Keywords*: Diagnostic techniques; Generalised estimating equations; Repeated measures; Longitudinal data; Quasi-likelihood methods; Generalised linear models

## 1. Introduction

Zeger and Liang [1] considered the generalised estimating equations (GEEs) to analyse longitudinal data via quasi-likelihood methods. Liang and Zeger [2] derived the GEEs from a different and slightly more limited context. They assumed that the marginal distribution of the dependent variable follows a generalised linear model (GLM). In both articles, the GEEs are derived without fully specifying the joint distribution. The regression coefficients are consistently estimated even in the cases where the correlation structure is misspecified. However, efficiency depends on the working correlation matrix's proximity to the true one [2].

Liang and Zeger's method has been widely used in several areas dealing with non-Gaussian correlated data. In such a context, finding an appropriate relationship between the correlated response variable and the covariates through a linear model requires techniques to assess whether the selected model adequately fits the data.

Tan *et al.* [3] proposed certain diagnostic measures for checking the adequacy of marginal regression models but only in the analysis of correlated binary data. We present an extension

*Corresponding author. Email: mkelly@ime.usp.br

*Journal of Statistical Computation and Simulation*
ISSN 0094-9655 print/ISSN 1563-5163 online © 2007 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/10629360600780488

of such measures valid for repeated measures regression analysis via GEEs in general. In particular, we consider the projection (hat) matrix, Cook's distance, standardised residuals and half-normal probability plots with simulated envelopes [4].

Chang [5], among others, presented a non-parametric test that is a sensitive approach to examine residual values for possible patterns of non-randomness. Pan [6] proposed a modified version (QIC) of Akaike's information criterion that works well for variable and working correlation matrix selection. Finally, Preisser and Qaqish [7] proposed deletion diagnostics for GEEs, but from a slightly different point of view than the one presented here. Their proposal does not allow the construction of half-normal probability plots with simulated envelopes as the expression for covariance matrix of the ordinary residuals does not follow the form required by Atkinson [4].

In section 2, we review GEEs and introduce some notation. Diagnostic measures and graphical methods are derived in section 3. Interpretation of the proposed measures is discussed through two illustrative examples in section 4.

## 2. Generalised estimating equations

Let $y_i = (y_{i1}, y_{i2}, \ldots, y_{it_i})^{\mathrm{T}}$, $i = 1, \ldots, n$, be mutually independent random vectors of repeated outcomes and let $\mathbf{X}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{it_i})^{\mathrm{T}}$ be the $t_i \times p$ matrix of covariate values, with $\mathbf{x}_{ij} = (\mathrm{x}_{ij1}, \ldots, \mathrm{x}_{ijp})^{\mathrm{T}}$, $i = 1, \ldots, n$ and $j = 1, \ldots, t_i$. Assume that the mean and variance of $y_{ij}$ are given by

$$E(\mathbf{y}_{ij}) = \mu_{ij} \quad \text{and} \quad \mathrm{Var}(\mathbf{y}_{ij}) = \phi^{-1} v(\mu_{ij}), \tag{1}$$

where $v(\mu_{ij})$ is a known function of the mean $\mu_{ij}$ and $\phi^{-1}$ is the dispersion parameter. Suppose that the regression model is $\eta_{ij} = g(\mu_{ij}) = \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}$, where $g(\cdot)$ is a link function and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is the vector of unknown parameters to be estimated. To simplify notation, let $t_i = t$ without loss of generality.

If $\mathbf{R}_i$ denotes the correlation matrix for each $\mathbf{y}_i$, the corresponding covariance matrix is

$$\mathrm{Cov}(\mathbf{y}_i) = \phi^{-1} \mathbf{A}_i^{1/2} \mathbf{R}_i \mathbf{A}_i^{1/2}, \tag{2}$$

where $\mathbf{A}_i = \mathrm{diag}(v(\mu_{i1}), \ldots, v(\mu_{it}))$ is a diagonal matrix.

In general, $\mathbf{R}_i$ is unknown and a practical way to bypass such a problem in the GEE context is to define a working correlation matrix, $\mathbf{R}(\boldsymbol{\alpha})$, which depends on an unknown parameter vector $\boldsymbol{\alpha}$ and is equal for all subjects. Then, the working covariance matrix for $\mathbf{y}_i$ is given by

$$\boldsymbol{\Omega}_i = \phi^{-1} \mathbf{A}_i^{1/2} \mathbf{R}(\boldsymbol{\alpha}) \mathbf{A}_i^{1/2}. \tag{3}$$

It will be equal to $\mathrm{Cov}(\mathbf{y}_i)$ if $\mathbf{R}(\boldsymbol{\alpha})$ is the true correlation matrix for $\mathbf{y}_i$. The related GEEs are

$$\sum_{i=1}^{n} \mathbf{D}_i^{\mathrm{T}} \boldsymbol{\Omega}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}, \tag{4}$$

where $\mathbf{D}_i^{\mathrm{T}} = \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\Lambda}_i$ and $\boldsymbol{\Lambda}_i = \mathrm{diag}(\partial_{\mu_{i1}}/\partial_{\eta_{i1}}, \ldots, \partial_{\mu_{it}}/\partial_{\eta_{it}})$, $i = 1, \ldots, n$ and $j = 1, \ldots, t$.

The parameter estimates are obtained by alternating between the modified Fisher scoring iterative method for $\boldsymbol{\beta}$, as described subsequently, and the moment estimation method for $\boldsymbol{\alpha}$ and $\phi$, as described by Liang and Zeger [2]. The estimates of $\boldsymbol{\alpha}$ and $\phi$ must be recalculated

at each iteration. Given current estimates of the (nuisance) parameters $\boldsymbol{\alpha}$ and $\phi$, the iterative procedure for estimating $\boldsymbol{\beta}$ may be expressed as

$$\hat{\boldsymbol{\beta}}^{(m+1)} = \hat{\boldsymbol{\beta}}^{(m)} + \left\{ \left[ \sum_{i=1}^{n} \hat{\mathbf{D}}_i^{\mathrm{T}} \hat{\boldsymbol{\Omega}}_i^{-1} \hat{\mathbf{D}}_i \right]^{-1} \left[ \sum_{i=1}^{n} \hat{\mathbf{D}}_i^{\mathrm{T}} \hat{\boldsymbol{\Omega}}_i^{-1} (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i) \right] \right\}^{(m)}, \tag{5}$$

where $m = 0, 1, 2, \ldots$ and $\hat{\boldsymbol{\beta}}^{(m)}$ is an initial (arbitrary) estimator. The estimate of $\boldsymbol{\beta}$ is updated by equation (5), evaluating the right-hand side at the current estimates of $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and $\phi$ in the $m$th iteration.

Liang and Zeger [2] showed that under certain regularity conditions, among which $\hat{\boldsymbol{\alpha}}$ and $\hat{\phi}$ are consistent estimators of $\boldsymbol{\alpha}$ and $\phi$, respectively,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \mathbf{J}^{-1}),$$

with

$$\mathbf{J}^{-1} = \lim_{n\to\infty} n \left\{ \sum_{i=1}^{n} \mathbf{D}_i^{\mathrm{T}} \boldsymbol{\Omega}_i^{-1} \mathbf{D}_i \right\}^{-1} \left\{ \sum_{i=1}^{n} \mathbf{D}_i^{\mathrm{T}} \boldsymbol{\Omega}_i^{-1} \operatorname{Cov}(\mathbf{y}_i) \boldsymbol{\Omega}_i^{-1} \mathbf{D}_i \right\} \left\{ \sum_{i=1}^{n} \mathbf{D}_i^{\mathrm{T}} \boldsymbol{\Omega}_i^{-1} \mathbf{D}_i \right\}^{-1}. \tag{6}$$

The *robust, empirical* or *sandwich* variance estimator of $\hat{\boldsymbol{\beta}}$ is given by

$$\left\{ \sum_{i=1}^{n} \hat{\mathbf{D}}_i^{\mathrm{T}} \hat{\boldsymbol{\Omega}}_i^{-1} \hat{\mathbf{D}}_i \right\}^{-1} \left\{ \sum_{i=1}^{n} \hat{\mathbf{D}}_i^{\mathrm{T}} \hat{\boldsymbol{\Omega}}_i^{-1} (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)^{\mathrm{T}} \hat{\boldsymbol{\Omega}}_i^{-1} \hat{\mathbf{D}}_i \right\} \left\{ \sum_{i=1}^{n} \hat{\mathbf{D}}_i^{\mathrm{T}} \hat{\boldsymbol{\Omega}}_i^{-1} \hat{\mathbf{D}}_i \right\}^{-1},$$

which is obtained by replacing $\operatorname{Cov}(\mathbf{y}_i)$ by $(\mathbf{y}_i - \boldsymbol{\mu}_i)(\mathbf{y}_i - \boldsymbol{\mu}_i)^{\mathrm{T}}$ and $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and $\phi$ by their respective estimators. It is robust in the sense that it is a consistent estimator of $\mathbf{J}^{-1}$ even if $\mathbf{R}(\boldsymbol{\alpha})$ is misspecified. If $\mathbf{R}(\boldsymbol{\alpha})$ is correctly specified, the variance estimator of $\hat{\boldsymbol{\beta}}$ will reduce to

$$\left\{ \sum_{i=1}^{n} \hat{\mathbf{D}}_i^{\mathrm{T}} \hat{\boldsymbol{\Omega}}_i^{-1} \hat{\mathbf{D}}_i \right\}^{-1},$$

known as the *naive* or *model-based* variance estimator. Similar robust and naive variance estimates suggests that $\mathbf{R}(\boldsymbol{\alpha})$ is adequate [8].

## 3. Diagnostic techniques

Diagnostic techniques are of great relevance for detecting regression problems and are very well discussed for regression models with independent observations in Paula [9], for example. For such models, the diagonal elements of the projection (hat) matrix, the well-known Cook's distance and the residuals are useful for detecting high leverage, influential and outlying observations, respectively. Another resource for detecting problems in the fit of regression models is the half-normal plot with simulated envelope [4]. However, this plot can only be used when the marginal distributions are known.

Here, we extend such diagnostic measures to evaluate the fit of the regression models with repeated measures described in section 2.

### 3.1 *High leverage, influence and outlying observations*

The most useful way to setup the iterative process outlined in equation (5) is by using the iteratively reweighted least-squares method obtained by employing a pseudo-observation vector $\mathbf{z}_i = \hat{\boldsymbol{\eta}}_i + \hat{\boldsymbol{\Lambda}}_i^{-1}\ (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)$ and a weight matrix $\mathbf{W}_i = \hat{\boldsymbol{\Lambda}}_i\ \hat{\boldsymbol{\Omega}}_i^{-1}\ \hat{\boldsymbol{\Lambda}}_i$. Under this setup, equation (5) reduces to

$$\hat{\boldsymbol{\beta}}^{(m+1)} = \left\{ \left[ \sum_{i=1}^{n} \mathbf{X}_i^{\mathrm{T}} \mathbf{W}_i \mathbf{X}_i \right]^{-1} \left[ \sum_{i=1}^{n} \mathbf{X}_i^{\mathrm{T}} \mathbf{W}_i \mathbf{z}_i \right] \right\}^{(m)}. \tag{7}$$

At convergence, we obtain

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}} \mathbf{W} \mathbf{X}\right)^{-1} \mathbf{X}^{\mathrm{T}} \mathbf{W} \mathbf{z}, \tag{8}$$

where $\mathbf{W} = \mathrm{diag}(\mathbf{W}_1, \ldots, \mathbf{W}_n)$ is a block diagonal weight matrix whose $i$th block corresponds to the $i$th subject, $\mathbf{X} = (\mathbf{X}_1^{\mathrm{T}}, \ldots, \mathbf{X}_n^{\mathrm{T}})^{\mathrm{T}}$ and $\mathbf{z} = (\mathbf{z}_1^{\mathrm{T}}, \ldots, \mathbf{z}_n^{\mathrm{T}})^{\mathrm{T}}$.

The residual vectors defined as the deviation of the observed from the fitted data can be written as

$$\mathbf{r}^* = \mathbf{W}^{1/2}(\mathbf{z} - \hat{\boldsymbol{\eta}}) = \mathbf{W}^{1/2} \hat{\boldsymbol{\Lambda}}^{-1} (\mathbf{y} - \hat{\boldsymbol{\mu}}), \tag{9}$$

where $\mathbf{W}^{1/2}$ is defined as the square root obtained from eigenvalue decomposition of $\mathbf{W}$, $\hat{\boldsymbol{\Lambda}} = \mathrm{diag}(\hat{\boldsymbol{\Lambda}}_1, \ldots, \hat{\boldsymbol{\Lambda}}_n)$, $\mathbf{y} = (\mathbf{y}_1^{\mathrm{T}}, \ldots, \mathbf{y}_n^{\mathrm{T}})^{\mathrm{T}}$ and $\hat{\boldsymbol{\mu}} = (\hat{\boldsymbol{\mu}}_1^{\mathrm{T}}, \ldots, \hat{\boldsymbol{\mu}} 1_n^{\mathrm{T}})^{\mathrm{T}}$.

As $\mathrm{Cov}(\mathbf{z}) = \hat{\boldsymbol{\Lambda}}^{-1}\, \mathrm{Cov}(\mathbf{y}) \hat{\boldsymbol{\Lambda}}^{-1} \cong \mathbf{W}^{-1}$, we have $\mathrm{Cov}(\mathbf{r}^*) \cong (\mathbf{I} - \mathbf{H})$, where $\mathbf{I}$ is the identity matrix and $\mathbf{H}$ is the block diagonal matrix given by $\mathrm{diag}(\mathbf{H}_1, \ldots, \mathbf{H}_n)$, where

$$\mathbf{H}_i = \mathbf{W}_i^{1/2} \mathbf{X}_i (\mathbf{X}^{\mathrm{T}} \mathbf{W} \mathbf{X})^{-1} \mathbf{X}_i^{\mathrm{T}} \mathbf{W}_i^{1/2}, \quad i = 1, \ldots, n. \tag{10}$$

$\mathbf{H}$ is symmetric and idempotent, so that $\mathrm{r}(\mathbf{H}) = \mathrm{tr}(\mathbf{H}) = p$, where $\mathrm{r}(\mathbf{H})$ is the rank of $\mathbf{H}$.

The elements of $\mathbf{r}^*$ have different variances; consequently, the standardized residual associated to $y_{ij}$ is defined by

$$(\mathrm{r}_{\mathrm{SD}})_{ij} = \frac{\mathbf{e}_{(ij)}^{\mathrm{T}} \mathbf{W}^{1/2} \hat{\boldsymbol{\Lambda}}^{-1} (\mathbf{y} - \hat{\boldsymbol{\mu}})}{\sqrt{1 - h_{ij}}}, \tag{11}$$

where $\mathbf{e}_{(ij)}$ is a vector with 1 at the position of observation $(ij)$ lexicographically ordered and 0 at the remaining positions and $h_{ij}$ is the $j$th diagonal element of $\mathbf{H}_i$, $i = 1, \ldots, n$ and $j = 1, \ldots, t$.

The ordinary residual in equation (9) can also be written as $\mathbf{r}^* = (\mathbf{I} - \mathbf{H})\mathbf{W}^{1/2}\mathbf{z}$. Then, considering that $\mathbf{W}^{1/2}\mathbf{z}$ plays the role of the outcome vector, we can interpret $\mathbf{H}$ as the hat matrix in the same way as in normal linear regression, where $\mathbf{W}$ is the identity matrix. This allows us to use the diagonal elements of $\mathbf{H}$ to detect high-leverage point, as in Paula [9] for GLMs and Tan *et al.* [3] for logistic regression with correlated responses.

A large value of $h_{ij}$ indicates that $\mathbf{x}_{ij}$ has a large influence on its fitted value $\hat{\mathbf{y}}_{ij}$. When all points have the same influence on the regression parameter estimates, we expect that $h_{ij}$ is around the average $\mathrm{tr}(\mathbf{H})/N = p/N$, so that points with $h_{ij} \geq 2p/N$ can be considered as high-leverage points. As another guideline to identify outlying subjects, we can use the average of the $h_{ij}$'s within a subject to identify high-leverage subjects. Namely, we consider

the $i$th subject as having a large influence on the fitted model if

$$h_i = \frac{1}{t}\sum_{j=1}^{t} h_{ij} = \frac{\mathrm{tr}(\mathbf{H}_i)}{t} \geq \frac{2p}{N}.$$

Graphically, we can plot $h_{ij}$ versus $i$, with $i = 1, \ldots, n$ and $j = 1, \ldots, t$. If the goal is to determine the influence of each subject, then we plot $h_i$ versus $i$, $i = 1, \ldots, n$.

Outliers can be detected by plotting the standardized residuals $(r_{\mathrm{SD}})_{ij}$ versus the indices $i$, where $i = 1, \ldots, n$ and $j = 1, \ldots, t$. Outliers are observations that differ greatly from the fitted value and not having high leverage.

The influence of each observation on the regression coefficients can be assessed by Cook's distance [10]. This measure is the distance between the estimated regression coefficients using all response values ($\hat{\boldsymbol{\beta}}$) and those estimated without observation $\mathbf{y}_{ij}$($\hat{\boldsymbol{\beta}}_{(ij)}$), with $i = 1, \ldots, n$ and $j = 1, \ldots, t$.

The idea in the one-step approximation for $\hat{\boldsymbol{\beta}}_{(ij)}$ presented by Pregibon [11] is applied here to the GEEs estimator in equation (7) and is given by

$$\hat{\boldsymbol{\beta}}^{(1)}_{(ij)} = \hat{\boldsymbol{\beta}} - \frac{[\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}]^{-1}[\mathbf{X}^{\mathrm{T}}\mathbf{W}^{1/2}\mathbf{e}_{(ij)}][\mathbf{e}^{\mathrm{T}}_{(ij)}\mathbf{W}^{1/2}\hat{\boldsymbol{\Lambda}}^{-1}(\mathbf{y} - \hat{\boldsymbol{\mu}})]}{1 - h_{ij}}.$$

Cook's distance with the deletion of the observation $\mathbf{y}_{ij}$ is defined as

$$(\mathrm{CD})_{ij} = \frac{1}{p}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(ij)})^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(ij)}) = (\mathrm{r_{SD}})^2_{ij}\frac{h_{ij}}{p(1 - h_{ij})}. \tag{12}$$

In the plot of $(\mathrm{CD})_{ij}$ versus the index $i$, with $i = 1, \ldots, n$ and $j = 1, \ldots, t$, influential observations are discrepant values compared with the others. Cook's distance for subject $i$ is given by $(\mathrm{CD})_i = \sum_{j=1}^{t}(\mathrm{CD})_{ij}$, $i = 1, \ldots, n$.

As all the diagnostic statistics discussed earlier involve the estimated correlation parameters, they may not be accurate when the estimates are not close to the true values.

### 3.2 *Simulated envelope*

Half-normal plots with simulated envelopes are useful for identifying outliers and examining the adequacy of the fitted model, even when the distribution of the residuals is not known [12].

A simulated envelope for a half-normal probability plot of the absolute residuals is constructed as follows:

1. For each subject $i$, $i = 1, \ldots, n$, simulate a $t \times 1$ vector of responses using the estimates mean vector and the correlation matrix, on the basis of the model fitted to the original data $\mathbf{y}$.
2. For the simulated responses in the first step, fit the same model using the same covariates.
3. Compute the set of standardised residuals given in equation (11) and order them.
4. Repeat the first three steps 24 times, independently. Here, let $(\mathrm{r_{SD}})_{lm}$ be the $l$th-ordered absolute value of the standardised residual belonging to $m$th step, $l = 1, \ldots, N$ and $m = 1, \ldots, M$, where $M = 25$.
5. Compute the minimum, median (or mean) and maximum of the smallest absolute values of residuals for all steps, that is, $(\mathrm{r_{SD}})_{1m}$, $m = 1, \ldots, 25$.
6. Repeat the last step for the second smallest absolute values of standardised residuals $(\mathrm{r_{SD}})_{2m}$. Next, repeat this step for the third smallest values $(\mathrm{r_{SD}})_{3m}$, and so forth, until the largest

absolute values of the standardised residuals $(r_{SD})_{Nm}$, $m = 1, \ldots, 25$. At the end of this step, three $N \times 1$ vectors were obtained: one with the minimum, one with the median (or mean) and the other with maximum values of the standardised residuals.

7. Finally, plot the values obtained in step 6 and the ordered absolute values of the standardised residuals from the original fit against the half-normal scores

$$\Phi^{-1}\left(\frac{l + N - 1/8}{2N + 1/2}\right),$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

The occurrence of points falling outside of the simulated envelope indicates that the fitted model is not appropriate. If there are outliers, they will appear at the top side of the half-normal probability plot, separated from the other points.

The half-normal probability plot with simulated envelope is easily constructed once correlated variables are generated. This may be easily accomplished by the function **rmvnorm** available in S-Plus or R designed to generate multivariate Gaussian distributions [13]. Park *et al.* [14] proposed an algorithm to generate a random vector of correlated binary variables and Park and Shin [15] developed an algorithm to generate a vector of dependent Poisson or gamma variables. These authors provide a simple algorithm for generating a set of non-negatively correlated vectors with arbitrary dimension. To generate correlated variables under other distributions, we suggest the use of copulas [16].

## 4. Applications

As an illustration, two data sets were analysed applying the methods developed in sections 2 and 3.

### 4.1 *Application using a Gaussian distribution*

The data were obtained from Lima and Sañudo [17]. The objective was to verify the learning process of a certain task. Each of the 40 volunteers completed the task in eight different attempts. Considering normal distribution and identity (canonical) link, the natural logarithm mean of the response variable was fitted by

$$\mu_{ij} = \mathbf{x}_{ij}^{\mathrm{T}}\hat{\boldsymbol{\beta}},$$

where $\mathbf{x}_{i,j} = (1\, j)^{\mathrm{T}}$ and $\hat{\boldsymbol{\beta}} = (\beta_0, \beta_1)^{\mathrm{T}}$, with $i = 1, \ldots, 40$ and $j = 1, \ldots, 8$.

Table 1 shows the estimated regression, dispersion and correlation parameters of the fitted model using the AR(1) structure working correlation matrix. Using the generalised Wald test

Table 1. Parameter estimates of the normal regression model using AR(1) structure.

| | Parameter | Estimate | Robust S.E. | Naive S.E. |
|---|---|---|---|---|
| $\beta_0$ | Intercept | 3.850 | 0.067 | 0.068 |
| $\beta_1$ | Slope | −0.051 | 0.010 | 0.013 |
| $\phi^{-1}$ | Dispersion | 0.173 | | |
| $\alpha$ | Correlation | 0.531 | | |

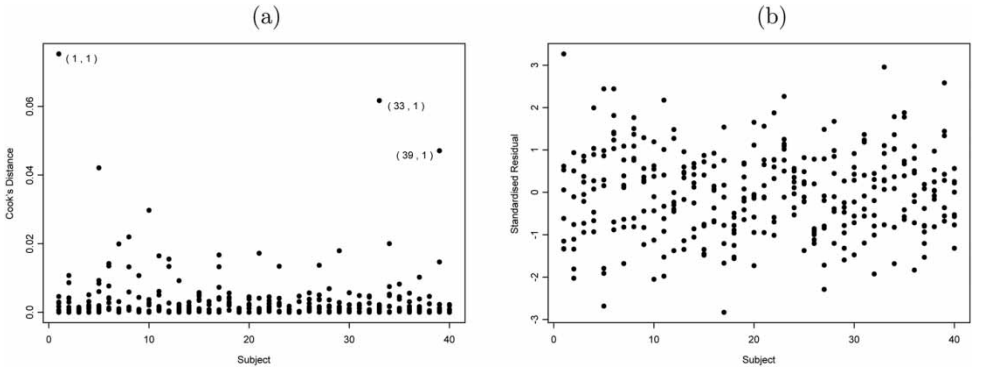


Figure 1. Plot of Cook's distance (a) and plot of the standardised residual (b) for the normal regression model using AR(1) structure.

statistic proposed by Rotnitzky and Jewell [18], we concluded for a significant attempt effect ($P < 0.001$).

Cook's distance and the standardised residual were computed for each pair of subject (volunteer) $i$ and attempt $j$, $i = 1, \ldots, 40$ and $j = 1, \ldots, 8$. In figure 1a, the observations of the subjects 1, 33 and 39 related to the first attempt (which are 5.0, 5.1 and 4.9, respectively) present higher values than the other observations (whose mean is 3.7), suggesting that they are influential points. This indicates that Cook's distance measure distinguishes these observations correctly. Figure 1b shows no any residual distinct from the others. It is noteworthy that this example presents no quantitative covariates; consequently, we do not use the projection matrix to detect high-leverage observations.

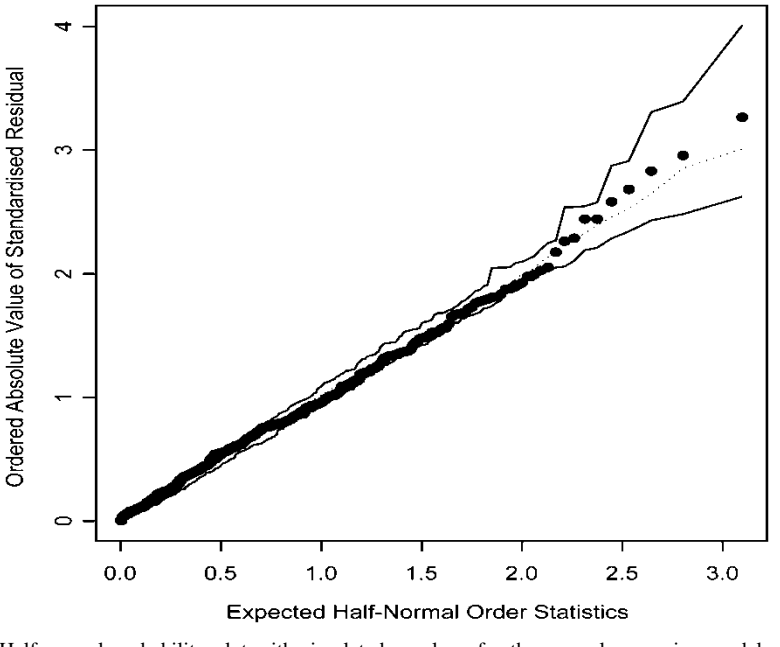


Figure 2. Half-normal probability plot with simulated envelope for the normal regression model using AR(1) structure.

The half-normal probability plot with simulated envelope (figure 2) indicates no observation outside the simulated envelope. Thus, it was concluded that the normal regression model adequately fits the data.

### 4.2 *Application using a Poisson distribution*

The second example, reported by Montgomery *et al.* [19, p. 215], is a biomedical one with 30 subjects (rats) having an induced leukemic condition. Three chemotherapy type drugs were used, with 10 subjects for each drug. White ($W$) and red ($R$) blood cell counts were collected as covariates and the response is the number of cancer cell colonies. The data were collected on each subject at four different time periods. Poisson responses using a log (canonical) link were assumed. Thus,

$$\log \mu_{ij} = \beta_l + \beta_4 W_{ij} + \beta_5 R_{ij},$$

where $i$, $j$ and $l$ are index subject, time period and drug, respectively, with $i = 1, \ldots, 30$, $j = 1, \ldots, 4$ and $l = 1, 2, 3$. Here, the first 10 rats ($i = 1, \ldots, 10$) used drug 1, the following 10 rats ($i = 11, \ldots, 20$) used drug 2 and the last 10 rats ($i = 21, \ldots, 30$) used drug 3.

Table 2 shows the estimated regression and correlation parameters of the fitted model using the AR(1) structure working correlation matrix. Using the generalised Wald-test statistic proposed by Rotnitzky and Jewell [18], it can be noted that all regression parameters are highly significant ($P < 0.001$, for each parameter).

Applying the diagnostic techniques presented in section 3, the measures $h_{ij}$ and $h_i$ were computed to verify whether the observation $(i, j)$ or the subject $i$, respectively, are high-leverage cases, $i = 1, \ldots, 30$ and $j = 1, \ldots, 4$. Figures 3a and b, respectively, provide these two measures. In figure 3a, apparently, six observations are high leverage: (1, 4), (3, 1) and (6, 1) related to the drug 1, (16, 1) and (16, 4) related to the drug 2 and (23, 4) related to the drug 3. However, figure 3b shows no subject as a high-leverage case.

To detect influential and outlier observations, Cook's distance and the standardised residual were computed and are presented, respectively, in figures 3c and 3d. Neither figures show any observation distinct from each other. The half-normal probability plot with simulated envelope (figure 4) indicates no observation outside the simulated envelope. Therefore, it can be concluded that the Poisson regression with AR(1) correlation structure is adequate to explain the relation between the number of cancer cell colonies and the covariates white and red blood cell counts.

Table 2. Parameter estimates of the Poisson regression model using AR(1) structure.

| | Parameter | Estimate | Robust S.E. | Naive S.E. |
|---|---|---|---|---|
| $\beta_1$ | Drug 1 | 3.0120 | 0.0778 | 0.0315 |
| $\beta_2$ | Drug 2 | 3.2315 | 0.0976 | 0.0891 |
| $\beta_3$ | Drug 3 | 3.1363 | 0.1540 | 0.1075 |
| $\beta_4$ | $W$ | −0.0305 | 0.0051 | 0.0045 |
| $\beta_5$ | $R$ | 0.0221 | 0.0065 | 0.0073 |
| $\alpha$ | Correlation | 0.9227 | | |

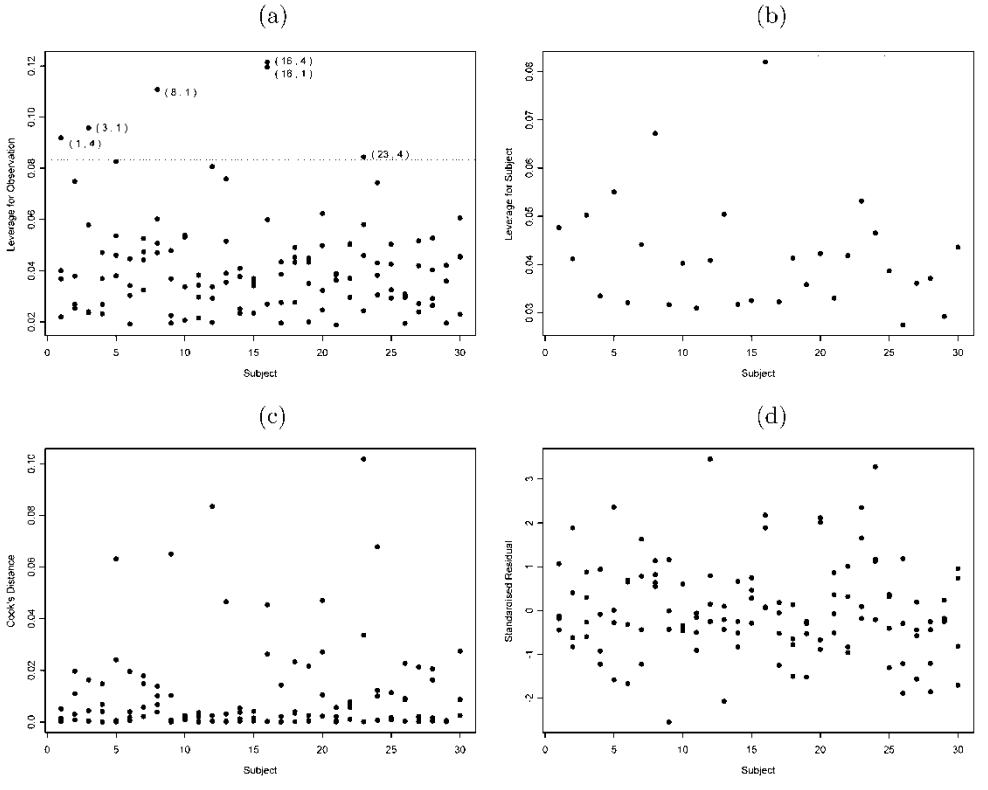


Figure 3. Plot of the leverage for each observation (a) plot of the leverage for each subject (b) plot of Cook's distance (c) and plot of the standardised residual (d) for the Poisson regression model using AR(1) structure.

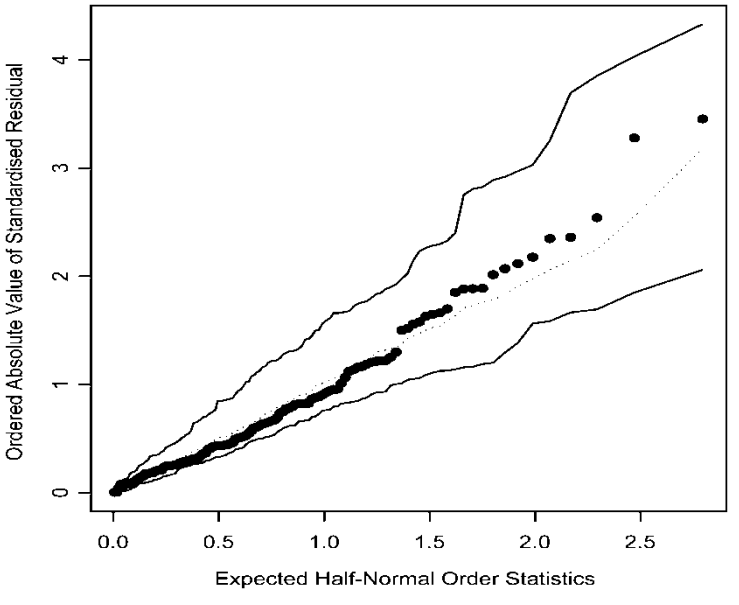


Figure 4. Half-normal probability plot with simulated envelope for the Poisson regression model using AR(1) structure.

## Acknowledgements

We are grateful to Professor Júlio da Motta Singer of the Institute of Mathematics and Statistics, University of São Paulo, for revision and useful comments on a previous version of the article. The research was partially supported by FAPESP – Brazil (01/00098-3).

## References

[1] Zeger, S.L. and Liang, K.-Y., 1986, Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, **42**, 121–130.
[2] Liang, K.-Y. and Zeger, S.L., 1986, Longitudinal analysis using generalized linear models. *Biometrika*, **73**, 13–22.
[3] 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.
[4] Atkinson, A.C., 1985, *Plots, Transformations and Regressions* (Oxford: Oxford Statistical Science Series).
[5] Chang, Y.-C., 2000, Residuals analysis of the generalized linear models for longitudinal data. *Statistics in Medicine*, **19**, 1277–1293.
[6] Pan, W., 2001, Akaike's information criterion in generalized estimating equations. *Biometrics*, **57**, 120–125.
[7] Preisser, J.S. and Qaqish, B.F., 1996, Deletion diagnostics for generalised estimating equations. *Biometrika*, **83**, 551–562.
[8] Johnston, G., 1996, Repeated measures analysis with discrete data using the SAS system, SAS Institute Inc. Cary, NC. Available online at: URL: http://academic.son.wisc.edu/rdsu/pdf/gee.pdf.
[9] Paula, G.A., 2004, Modelos de regressão com apoio computacional, Notas de aulas, Departamento de Estatística, Universidade de São Paulo. Available online at: URL: http://www.ime.usp.br/∼giapaula/mlgs.html.
[10] Cook, R.D., 1977, Detection of infuential observations in linear regressions. *Technometrics*, **19**, 15–18.
[11] Pregibon, D., 1981, Logistic regression diagnostics. *Annals of Statistics*, **9**, 705–724.
[12] Neter, J., Kutner, M.H., Naschstheim, C.J. and Wasserman, W., 1996, *Applied Linear Statistical Models* (Chicago: IE McGraw Hill).
[13] Venables, W.N. and Ripley, B.D., 1999, *Modern Applied Statistics with S-Plus* (3rd edn) (New York: Springer-Verlag).
[14] Park, C.G., Park, T. and Shin, D.W., 1996, A simple method for generating correlated binary variates. *The American Statistician*, **50**, 306–310.
[15] Park, C.G. and Shin, D.W., 1998, An algorithm for generating correlated random variables in a class of infinitely divisible distributions. *Journal of Statistical Computation and Simulation*, **61**, 127–139.
[16] Nelsen, R., 1999, *An Introduction to Copulas* (New York: Springer).
[17] Lima, A.C.P. and Sañudo, A., 1997, Transferência entre tarefas sincronizatórias com diferentes níveis de complexidade. Technical Report RAE-CEA-9702, IME – USP, São Paulo.
[18] Rotnitzky, A. and Jewell, N.P., 1990, Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika*, **77**, 485–497.
[19] Montgomery, D.C., Myers, R.H. and Vining, G., 2001, *Generalized Linear Models: With Applications in Engineering and the Sciences* (New York: John Wiley & Sons).