RT-MAE 2005-08

Diagnostic Techniques in Generalized Estimating Equations

by

Maria Kelly Venezuela,
Denise Aparecida Botter
and
Mônica Carneiro Sandoval

Palavras-Chave: Generalized estimating equations, diagnostic techniques, repeated measures, quasi-likelihood methods.
Classificação AMS: 62J20.

- Fevereiro de 2005 -

Diagnostic Techniques in Generalized Estimating Equations

Maria Kelly Venezuela [*]
Denise Aparecida Botter [†]
and
Mônica Carneiro Sandoval[†]

Abstract

Zeger and Liang (1986) proposed a methodology for discrete and continuous longitudinal data that uses the quasi-likelihood approach. For those models, we generalized some diagnostic methods useful in regression models with independent responses as the projection (hat) matrix, the Cook's distance and the standardized residual. Moreover, we use the half-normal probability plot with simulated envelope for checking the adequacy of the fitted model only when the marginal distributions belong to the exponential family. To construct this plot, we simulated correlated outcomes through algorithms describes in statistics literature. Finally, we realized two applications to illustrate the techniques.

Key words: generalized estimating equations, diagnostic techniques, repeated measures, quasi-likelihood methods.

1 Introduction

Zeger and Liang (1986) developed the generalized estimating equations (GEEs) using the quasi-likelihood (Wedderburn, 1974) approach to the analysis of longitudinal data. Liang and Zeger (1986) derived the GEEs from a different and slightly more limited context. They assumed that the marginal distribution of the dependent variable followed a generalized linear model (McCullagh and Nelder, 1989). In both of them, the GEEs are derived

[*] E-mail: mkelly@ime.usp.br

[†] Address for correspondence: Departamento de Estatística, IME - USP, Caixa Postal 66281 (Ag. Cidade de São Paulo), 05315-970, São Paulo - SP - Brazil.

without full specification of the joint distributions and a working correlation matrix for the vector of repeated measurements from each subject need be specified. Moreover, the dependence of the outcomes on the covariates is the primary focus and it is not necessary to specify the working correlation matrix correctly in order to have a consistent estimator of the regression parameters. However, choosing the working correlation close to the true correlation increases the statistical efficiency of the regression parameter estimator.

After this theory, Liang and Zeger's method has been used widely in several areas which have non-Gaussian correlated data in practice. Therefore, in finding an appropriate relationship between correlated response variable and covariates through a linear model, it is important to check that the selected model is adequate to fit the data.

The purpose of this paper is to propose diagnostic measures for any regression analysis with repeated measurements that uses the GEEs methodology. These propose generalize the usual measures in generalized linear models (GLMs) such as the projection (hat) matrix, the Cook's distance and the standardized residual for detecting high leverage points, influential and outlier observations, respectively, and the half-normal probability plot with simulated envelope for checking the adequacy of the adjusted model.

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

2 Generalized Estimating Equation

Let $\mathbf{y}_i$ be mutually independent random vectors, where $\mathbf{y}_i = (y_{i1}, y_{i2}, ..., y_{it_i})^{\mathrm{T}}$ is the $t_i \times 1$ vector of repeated outcome values for the ith subject, and let $\mathbf{X}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{it_i})^{\mathrm{T}}$ be the $t_i \times p$ matrix of covariate values, with $\mathbf{x}_{ij} = (x_{ij1}, ..., x_{ijp})^{\mathrm{T}}$, $i = 1, ..., n$ and $j = 1, ..., t_i$. Assume that the mean and variance of y_{ij} are

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

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

In this way, if $\mathbf{R}_i$ is the $t \times t$ correlation matrix for each $\mathbf{y}_i$, its covariance matrix is given by

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

where $\mathbf{A}_i = \text{diag}\{v(\mu_{i1}), \ldots, v(\mu_{it})\}$ denote a $t \times t$ diagonal matrix. In the context of quasi-likelihood multivariate, $\mathbf{R}_i$ must be a function of the mean $\boldsymbol{\mu}_i$, that is, of $\boldsymbol{\beta}$, with $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{it})^{\mathrm{T}}$. However, the correlation is restrict to the interval $[-1, 1]$ that it increases the complexity of the iteration process.

A practical way to solve this problem is to define a $t \times t$ working correlation matrix, $\mathbf{R}(\boldsymbol{\alpha})$, that depends on an unknown parameter vector $\boldsymbol{\alpha}$ and is equal for all subjects. So, the $t \times t$ working covariance matrix of $\mathbf{y}_i$ is given by

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

which will be equal to $\text{Cov}(\mathbf{y}_i)$ if $\mathbf{R}(\boldsymbol{\alpha})$ is indeed the true correlation matrix for the $\mathbf{y}_i$´s. The *generalized estimating equations* (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 $\boldsymbol{\Lambda}_i = \text{diag}\{\partial\mu_{i1}/\partial\eta_{i1}, \ldots, \partial\mu_{it}/\partial\eta_{it}\}$ is a $t \times t$ diagonal matrix which will be equal the identity matrix if $\mu_{ij} = \eta_{ij}$ is defined to the model, with $i = 1, ..., n$ and $j = 1, ..., t$.

The estimates of regression coefficients $\boldsymbol{\beta}$, which are our main interest here, are obtained by alternating between the modified Fisher scoring iterative method for $\boldsymbol{\beta}$, as follows, and moment estimation of $\boldsymbol{\alpha}$ and ϕ, as has been described by Liang and Zeger (1986). The estimates of $\boldsymbol{\alpha}$ and ϕ must be recalculated in each iteration. So, given current estimates of these nuisance parameters $\boldsymbol{\alpha}$ and ϕ, the iterative procedure for estimating $\boldsymbol{\beta}$ is

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

where $m = 0, 1, 2, \ldots$ is the iteration number. A current estimate of $\boldsymbol{\beta}$ is updated by equations (5) evaluating the right-hand side at the current estimate of $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and ϕ in the mth iteration.

Liang and Zeger (1986) show that, under regularity conditions and considering $\hat{\boldsymbol{\alpha}}$ and $\hat{\phi}$ as consistent estimates,

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

with

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

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

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

This estimator is obtained by replacing $\mathrm{Cov}(\mathbf{y}_i)$ by $(\mathbf{y}_i-\boldsymbol{\mu}_i)(\mathbf{y}_i-\boldsymbol{\mu}_i)^{\mathrm{T}}$ and $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and ϕ by their respective estimators in (6). It is robust in the sense that is consistently estimates $\mathbf{J}^{-1}$ even if $\mathbf{R}(\boldsymbol{\alpha})$ is misspecified. If $\mathbf{R}(\boldsymbol{\alpha})$ is correctly specified, the variance estimator of $\hat{\boldsymbol{\beta}}$ reduces to

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

which is known by *naive* or *model-based* variance estimator.

Considering that the regression model is correctly specified, the robust variance estimator is always consistent. However, the naive estimator is consistent only if the working correlation matrix is also correctly specified. When the number of subjects is small, say < 20, the naive variance estimator may have better properties (Prentice, 1988) even if $\mathbf{R}(\boldsymbol{\alpha})$ is wrong. This is because the robust variance estimator is asymptotically unbiased, but could be highly biased when the number of subjects is small (Horton and Lipsitz, 1999). When the robust and naive variance estimates are similar, it shows the adequacy of the matrix $\mathbf{R}(\boldsymbol{\alpha})$ to the model (Johnston, 1996).

The GEEs method yields only unbiased estimates for the mean structure if the data are missing completely at random (Rubin, 1976). A general approach for calculating the magnitude of the bias of estimators obtained from standard analysis of estimating equation in the presence of incomplete data was presented by Rotnitzky and Wypij (1994). Several approaches have been proposed to deal with missing data in the framework of the GEEs (Ziegler et al., 1998).

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 (2004), for example. For those models, the diagonal elements of the projection (hat) matrix, the Cook's distance and the residuals are useful for detecting high leverage points, influential and outlier observations, respectively. Another resource for detecting regression problems is the half-normal plots with simulated envelopes (Atkinson, 1985). However, we can use this plot only when the marginal distributions is known.

Tan et al. (1997) proposed those diagnostic methods for checking the adequacy of marginal regression models for analyzing correlated binary data. Here, we present an extension of those diagnostic measures for regression models with repeated measurements as described in Section 2.

Among others papers that discuss the diagnostic techniques for GEEs, Chang (2000) presents a nonparametric test that is a sensitive approach to examining residual values for possible patterns of non-randomness, Pan (2001) proposes a modified model-selection criterion QIC based in Akaike's Information Criterion that works well in variable selection and selecting the working correlation matrix, and finally, we cite here Preisser and Qaqish (1996) that propose deletion diagnostics for GEEs, but in a subtle different point of view as we present in the next section.

3.1 High Leverage, Influence and Outlier Points

A most useful way to view the iterative process outlined in (5) is by the method of iteratively reweighted least squares. This is obtained by employing the pseudo-observation vector $\mathbf{z}$ and the weight matrix $\mathbf{W}$, upon which it becomes

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

where $\hat{\mathbf{W}}_i = \hat{\mathbf{\Lambda}}_i\hat{\mathbf{\Omega}}_i^{-1}\hat{\mathbf{\Lambda}}_i$ and $\mathbf{z}_i = \hat{\eta}_i + \hat{\mathbf{\Lambda}}_i^{-1}(\mathbf{y}_i - \hat{\mu}_i)$.

At convergence of this iterative procedure, we may write

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

where $\hat{\mathbf{W}} = \mathrm{diag}(\hat{\mathbf{W}}_1, \ldots, \hat{\mathbf{W}}_n)$ is a block diagonal weight matrix whose ith block corresponding to the ith subject, $\mathbf{X} = (\mathbf{X}_1^{\mathrm{T}}, \ldots, \mathbf{X}_n^{\mathrm{T}})^{\mathrm{T}}$ and $\mathbf{z} = (\mathbf{z}_1^{\mathrm{T}}, \ldots, \mathbf{z}_n^{\mathrm{T}})^{\mathrm{T}}$, all of them with dimensions $N \times N$, $N \times p$ and $N \times 1$, respectively, with $N = nt$.

In this sense, the residuals defined by the deviation of the observed data from the fit can be written as

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

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

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

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

The matrix $\mathbf{H}$ is symmetric ($\mathbf{H}^T = \mathbf{H}$) and idempotent ($\mathbf{HH} = \mathbf{H}$), so $\mathrm{r}(\mathbf{H}) = \mathrm{tr}(\mathbf{H}) = p$, where $\mathrm{r}(\mathbf{H})$ is the rank of $\mathbf{H}$ and p is the number of regression coefficients.

The elements of $\mathbf{r}^*$ own different variances, which are difficult to compare themself, so we define the standardized residual associated with y_{ij} by

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

where $\mathbf{e}_{(ij)}$ is the $t \times 1$ vector with the jth element 1 and all the others 0, and h_{ij} is the jth diagonal element of $\mathbf{H}_i$, $i = 1, ..., n$ and $j = 1, ..., t$.

The ordinary residual in (9) can also be written as $\mathbf{r}^* = (\mathbf{I}-\mathbf{H})\hat{\mathbf{W}}^{1/2}\mathbf{z}$. Then, considering that $\hat{\mathbf{W}}^{1/2}\mathbf{z}$ plays the role of the outcome vector, we can interpret $\mathbf{H}$ as the hat matrix in the same way as in the normal linear regression, where $\hat{\mathbf{W}}$ is the identity matrix. This view allows us to use the diagonal elements of $\mathbf{H}$ to detect high leverage point, as well as Paula (2004) presents to GLMs and Tan et al. (1997) propose to logistic regression with correlated responses.

A large value of h_{ij} indicates that the influence of the covariate measurements of this observation may be unduly large. Assuming that all points exercise the same influence on the regression parameter estimates, we hope each element h_{ij} is around the average $\mathrm{tr}(\mathbf{H})/N = p/N$, so the point that $\mathrm{h}_{ij} \geq 2p/N$ can be considered a high leverage point. As another guideline to identify outlying subjects, we can use the average of the h_{ij}s within a subject to identify leverage subjects. Namely, the ith subject can exercise a large influence on fitted model, if

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

Graphically, we can make a plot of h_{ij} versus i, where i indicates the index of the subject, with $i = 1, ..., n$ and $j = 1, ..., t$. If the interest is check the influence of each subject, then we build a plot of $\mathrm{h}_{i.}$ versus i, $i = 1, ..., n$.

The detection of outlier observations using a graphical representation can be made by the plot of standardized residual $(\mathrm{r}_{SD})_{ij}$ versus the index i, where $i = 1, ..., n$ and

$j = 1, ..., t$. An outlier observation occurs when it differs too much from the fitted value and has not a so high leverage.

In addiction to the diagnostic, the influence observation occurs when the difference between this observation and the fitted value is unduly large and presents a high leverage. The influence of each observation on regression coefficients can be assessed by Cook's distance. This measure shows the distance between the estimate regression coefficients using all responses values ($\hat{\beta}$) and without the observation y_{ij} ($\hat{\beta}_{(ij)}$), with $i = 1, ..., n$ and $j = 1, ..., t$.

The idea of the one-step approximation for $\hat{\beta}_{(ij)}$ presents by Pregibon (1981) is applied here for the GEEs estimator in (7), which is given by

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

Then, Cook's distance with the deletion of the observation y_{ij} is defined by

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

The plot of $(CD)_{ij}$ against the index i, with $i = 1, ..., n$ and $j = 1, ..., t$, indicates an influential observation when it presents a discrepant value compared with the others points.

All of the diagnostic statistics discussed above involve estimated correlation parameters, $\mathbf{R}(\hat{\alpha})$, and thus may not be accurate when those estimates are not close to the true values.

The measures cited above were found using the ordinary residual, $\mathbf{r}^*$, and making analogies between $\hat{\beta}$ given by (8) and the equation to estimate β in the case of linear normal regression. Whatever, Preisser and Qaqish (1996) proposed their measures making an analogy of $\hat{\eta} = \mathbf{X}\hat{\beta}$ with the predictor of the linear normal regression which is given by $\hat{\eta} = \hat{\mu} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{H}\mathbf{y}$. So, in the case of GEEs, the authors define $\mathbf{H}$ that is the projection matrix given by

$$\hat{\eta} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\hat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\hat{\mathbf{W}}\mathbf{z} = \mathbf{H}\mathbf{z},$$

where $\mathbf{H}$ is here an asymmetric and idempotent matrix.

Following this purpose, the ordinary residual is given by $\mathbf{r}^* = \mathbf{z} - \hat{\eta} = (\mathbf{I} - \mathbf{H})\mathbf{z}$, and the expression of the covariance matrix is different of $(\mathbf{I} - \mathbf{H})$. This situation makes impossible the construction of the half-normal probability plot with simulated envelope, which procedure consists basically in generating residuals with mean zero and covariance matrix equal $(\mathbf{I} - \mathbf{H})$ (Atkinson, 1985).

3.2 Simulated envelope

The half-normal plots with simulated envelopes are useful for identifying outlier observations and for examining the adequacy of the fitted model, even if the distribution of the residuals is not known (Neter et al., 1996).

The steps of the algorithm presented below show that the construction of the half-normal probability plot with simulated envelope is simple since we know how to generate correlated variables. So, to provide this, we can use the function rmvnorm of S-Plus or R softwares for generating multivariate Gaussian distribution (Venables and Ripley, 1999). Park et al. (1996) propose an algorithm for generating a random vector of correlated binary variables and Park and Shin (1998) develop an algorithm for generating a vector of dependent Poisson or gamma variables. Both of these papers provide a simple algorithm for generating a set of nonnegatively correlated variables of arbitrary dimension. For generating correlated variables with others distributions, we suggest the use of copulas (Nelsen, 1999).

A simulated envelope for a half-normal probability plot of the absolute residuals is constructed to models with repeated measures in the following way:

1. For each subject i, $i = 1, ..., n$, simulate a $t \times 1$ vector of responses using the means vector and the correlations matrix estimates, based on the model fitted to the original data $\mathbf{y}$.

2. Fit to the simulated responses in the first step, the same model to $\mathbf{y}$ with the same covariates.

3. Compute the set of standardized residuals as in the equation (11) and order them.

4. Repeat steps 1 through 3 more 24 times independently. Here, let $(\mathrm{r}_{SD})_{lm}$ be the lth ordered absolute value of standardized residual belonging to mth simulation, $l = 1, ..., N$ e $m = 1, ..., M$, where $M = 25$.

5. Calculate the minimum, median (or mean) and maximum of the smallest absolute values of residuals for all simulations, that is, of the values $(\mathrm{r}_{SD})_{1m}$, $m = 1, ..., 25$.

6. Repeat the last step for the second small absolute values of standardized residuals $(\mathrm{r}_{SD})_{2m}$. After that, repeat this step for the third small values $(\mathrm{r}_{SD})_{3m}$, and so forth, until the biggest absolute values of standardized residuals $(\mathrm{r}_{SD})_{Nm}$, $m = 1, ..., 25$. In the end of this step, we got three $N \times 1$ vectors of the minimum, median (or mean) and maximum values of the standardized residuals.

7. Finally, plot these values and the ordered absolute values of the original fitted stan-

dardized residuals 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.

Large deviations of points from the medians (or means) of the simulated values or the occurrence of points near to or outside the simulated envelope, are indications that the fitted model is not appropriate. However, they do not provide enough information on how to improve the fit of the model. If there are outlier points, they will appear at the top right of the half-normal probability plot as points separated from the others.

Atkinson (1985) suggests by using $M = 19$ simulations, there is one chance in 20, or 5 percent, that the largest absolute standardized residual from the original data set lies outside the simulated envelope when the fitted model is correct. The value $M = 25$ simulations is suggested by Tan et al. (1997), which we adopt in this paper.

4 Applications

As illustration we analyze two data set applying the theories developed in §2 and §3.

4.1 Application using Gaussian Distribution

In this example, the data was obtained from Lima and Sañudo (1997). The aim was to verify the learning process of a certain task, which was developed by 40 volunteers. Each volunteer practice the task in 8 blocks of attempts. The response variable named Absolute Error was fitted by a regression model considering normal distribution and identity (canonical) link. The parameters involved in this model were: β_0, intercept and β_1, slope - block effect. This model was fitted using the working correlation matrix under the unstructured, exchangeable and AR(1) structures. This last one presented the lowest standardized residuals and was preferable to fit the response variable.

Table 1 shows the estimates of the regression, dispersion and correlation parameters of the fitted model using the working correlation matrix under the AR(1) structure. Through the generalized Wald test statistic proposed by Rotnitzky and Jewell (1990), we concluded the block effect is significant (P-value $< 0,001$).

Table 1: Parameter estimates of the normal regression model using AR(1) structure to the logarithm Absolute Error variable.

Parameter		Estimate	Robust S.E.	Naive S.E.
β_0	*Intercept*	3,850	0,067	0,068
β_1	*Block*	-0,051	0,010	0,013
ϕ^{-1}	*Dispersion*	0,173		
α	*Correlation*	0,531		

Applying the diagnostic techniques presented in § 3, we calculated the Cook's distance and the standardized residual for each pair of subject (volunteer) i and block of attempts j, $i = 1, \ldots, 40$ and $j = 1, \ldots, 8$. Figure 1a and Figure 1b show the values of these both measures, respectively. In the first one, the observations of the subjects 1, 33 and 39 connected to the first block of attempts are higher than the other observations, indicating to be a possible influence points. In the set of data, the values of the Absolute Error from the subjects 1, 33 and 39 connected to the first block of attempts were 5.0, 5.1 and 4.9, respectively. These values are not similar to the values of the other subjects of the first block, whose mean is 3.7. It indicates that the Cook's distance measure evidenced distinguished observations correctly. Figure 1b do not show any residual far away the others. Notice that this example does not have quantitative covariates and because of that, we do not utilize the projection matrix to detect high leverage observations.

The half-normal probability plot with simulated envelope (Figure 2) does not indicate any observation outside the simulated envelope. Then, we conclude that the normal regression model is adequate to fit the Absolute Error response.

4.2 Application using Poisson Distribution

This example, reported by Montgomery et al. (2001, p.215), is a biomedical example with 30 subjects (rats) that have had a leukemic condition induced. Three chemotherapy types drugs were used, so we have 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

$$\log \mu_{ij} = \beta_l + \beta_4 W_{ij} + \beta_5 R_{ij}, \tag{13}$$

where i, j and l index subject, time period and drug, respectively, with $i = 1, \ldots, 30$,

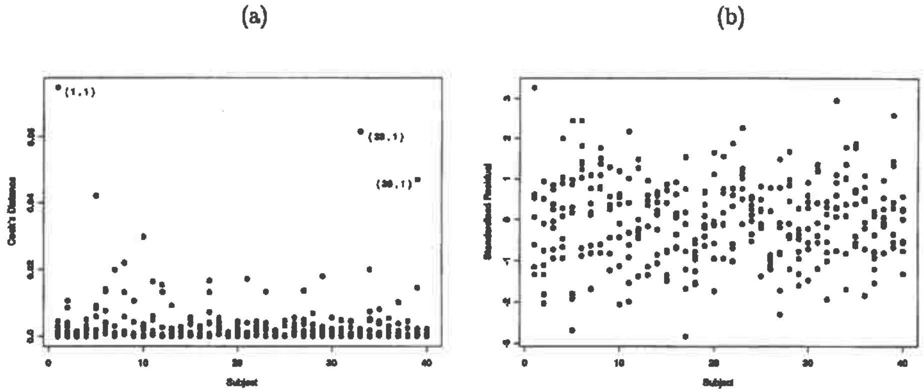

Figure 1: Plot of the Cook's distance (a) and plot of the standardized residual (b) for the normal regression model using AR(1) structure to the Absolute Error variable.

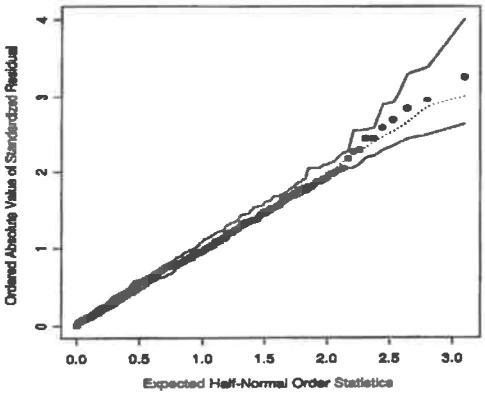

Figure 2: Half-normal probability plot with simulated envelope for the normal regression model using AR(1) structure to the Absolute Error variable.

$j = 1, \ldots, 4$ and $l = 1, 2, 3$. Here, the first ten rats ($i = 1, \ldots, 10$) used the drug 1, the rats indexed by $i = 11, \ldots, 20$ used the drug 2 and the last ten rats ($i = 21, \ldots, 30$) used the drug 3.

This model was fitted using unstructured, exchangeable and AR(1) correlation structures. In the first one, the regression estimates did not achieve the convergence after 50 iterations. The AR(1) structure was chosen to explicate the correlation among the observations of the same subject and drug because it got the lowest residual. Besides, the algorithm for generating a vector of dependent Poisson described by Park and Shin (1998) failed to the fitted model with exchangeable correlation structure, what it made impossible the construction of the half-normal plot with simulated envelope to this structure.

Table 2 shows the estimates of the regression and correlation parameters of the fitted model as in (13) using the working correlation matrix under the AR(1) structure. Through the generalized Wald test statistic proposed by Rotnitzky and Jewell (1990), all the regression parameters are highly significant (P-value $< 0,001$, for each parameter).

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

Parameter		Estimate	Robust S.E.	Naive S.E.
β_1	*Drug* 1	3,0120	0,0778	0,0315
β_2	*Drug* 2	3,2315	0,0976	0,0891
β_3	*Drug* 3	3,1363	0,1540	0,1075
β_4	W	-0,0305	0,0051	0,0045
β_5	R	0,0221	0,0065	0,0073
α	*Correlation*	0,9227		

Applying the diagnostic techniques presented in § 3, we calculated the measures $\mathbf{h}_{ij}$ and $\mathbf{h}_{i\cdot}$ to verify, respectively, if the observation (i, j) or the subject i is a high leverage, $i = 1, \ldots, 30$ and $j = 1, \ldots, 4$. Figure 3a and Figure 3b give these both measures, respectively. In this first one, apparently six observations are a high leverage cases: $(1, 4)$, $(3, 1)$ and $(6, 1)$ connected to the drug 1, $(16, 1)$ and $(16, 4)$ connected to the drug 2 and $(23, 4)$ connected to the drug 3. However, Figure 3b shows no subject as a possible high leverage.

To detect influence and outlier observations, we calculated the Cook's distance and the standardized residual, which are presented, respectively, in Figure 3c and Figure 3d. Both of them do not show any distinguished observation compared with the others. The half-normal probability plot with simulated envelope (Figure 4) does not indicate any observation outside the simulated envelope. So, we can conclude that the poisson regression under 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.

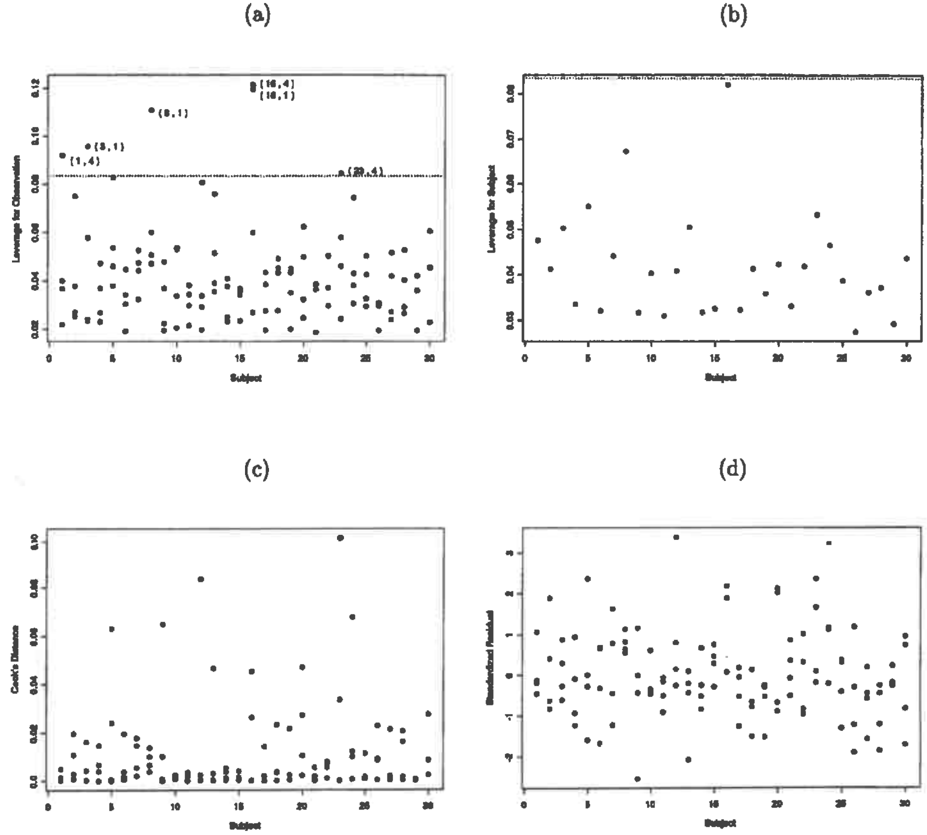

Figure 3: Plot of the leverage for each observation (a), plot of the leverage for each subject (b), plot of the Cook's distance (c) and plot of the standardized residual (d) for the Poisson regression model using AR(1) structure.

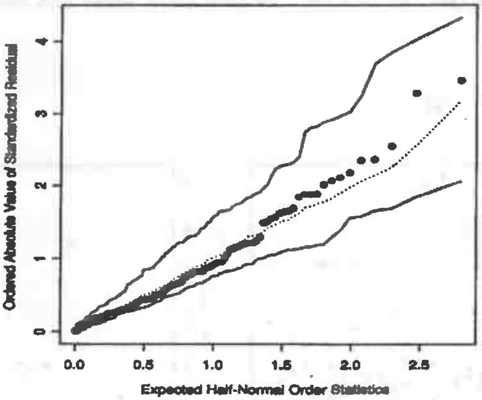

Figure 4: Half-normal probability plot with simulated envelope for the Poisson regression model using AR(1) structure.

References

Atkinson, A. C. (1985). *Plots, Transformations and Regressions*, Oxford Statistical Science Series, Oxford.

Chang, Y.-C. (2000). Residuals analysis of the generalized linear models for longitudinal data, *Statistics in Medicine* **19**: 1277–93.

Horton, N. J. and Lipsitz, S. R. (1999). Review of software to fit generalized estimating equation regression models, *The American Statistician* **53**: 160–169.
URL: *http://www.amstat.org/publications/tas/horton.pdf*

Johnston, G. (1996). Repeated measures analysis with discrete data using the sas system, SAS Institute Inc.
URL: *http://academic.son.wisc.edu/rdsu/pdf/gee.pdf*

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

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.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd edn, Chapman and Hall, London.

Montgomery, D. C., Myers, R. H. and Vining, G. (2001). *Generalized Linear Models: with Applications in Engineering and the Sciences*, John Wiley & Sons, New York.

Nelsen, R. (1999). *An Introduction to Copulas*, Springer, New York.

Neter, J., Kutner, M. H., Naschstheim, C. J. and Wasserman, W. (1996). *Applied Linear Statistical Models*, IE McGraw Hill, Chicago.

Pan, W. (2001). Akaike´s information criterion in generalized estimating equations, *Biometrics* **57**: 120–125.

Park, C. G., Park, T. and Shin, D. W. (1996). A simple method for generating correlated binary variates, *The American Statistician* **50**: 306–310.

Park, C. G. and Shin, D. W. (1998). An algorithm for generating correlated random variables in a class of infinitely divisible distributions, *J. Statist. Comput. Simul.* **61**: 127–139.

Paula, G. A. (2004). Modelos de regressão com apoio computacional, *Notas de aulas*, Departamento de Estatística, Universidade de São Paulo.
URL: *http://www.ime.usp.br/~giapaula/mlgs.html*

Pregibon, D. (1981). Logistic regression diagnostics, *Annals of Statistics* **9**: 705–724.

Preisser, J. S. and Qaqish, B. F. (1996). Deletion diagnostics for generalised estimating equations, *Biometrika* **83**: 551–562.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation, *Biometrics* **44**: 1033–1048.

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.

Rotnitzky, A. and Wypij, D. (1994). A note on the bias of estimators with missing data, *Biometrika* **50**: 1163–1170.

Rubin, R. M. (1976). Inference and missing data, *Biometrika* **63**: 581–592.

Tan, M., Qu, Y. and Kutner, M. H. (1997). Model diagnostics for marginal regression analysis of correlated binary data, *Commun. Statist. - Simula.* **26**: 539–558.

Venables, W. N. and Ripley, B. D. (1999). *Modern Applied Statistics with S-Plus*, 3 edn, Springer-Verlag, New York.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the gauss-newton method, *Biometrika* **61**: 31–38.

Zeger, S. L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes, *Biometrics* **42**: 121–130.

Ziegler, A., Kastner, C. and Blettner, M. (1998). The generalised estimating equations: An annoted bibliography, *Biometrical Journal* 40(2): 115–139.

ÚLTIMOS RELATÓRIOS TÉCNICOS PUBLICADOS

2005-01 - DE SOUZA BORGES,W., GUSTAVO ESTEVES, L., WECHSLER, S. Process Parameters Estimation in The Taguchi On-Line Quality Monitoring Procedure For Attributes. 2005.19p. (RT-MAE-2005-01)

2005-02 - DOS ANJOS,U.,KOLEV, N. Copulas with Given Nonoverlapping Multivariate Marginals. 2005.09p. (RT-MAE-2005-02)

2005-03 - DOS ANJOS,U.,KOLEV, N. Representation of Bivariate Copulas via Local Measure of Dependence. 2005.15p. (RT-MAE-2005-03)

2005-04 - BUENO, V. C., CARMO, I. M. A constructive example for active redundancy allocation in a k-out-of-n:F system under dependence conditions. 2005. 10p. (RT-MAE-2005-04)

2005-05 - BUENO, V. C., MENEZES, J.E. Component importance in a modulated Markov system. 2005. 11p. (RT-MAE-2005-05)

2005-06 - BASAN, J. L., BRANCO, M.D´E., BOLFARINE, H. A skew item response model. 2005. 20p. (RT-MAE-2005-06)

2005-07 - KOLEV, N., MENDES, B. V. M., ANJOS, U. Copulas: a Review and recent developments. 2005. 46p. (RT-MAE-2005-07)

The complete list of "Relatórios do Departamento de Estatística", IME-USP, will be sent upon request.

Departamento de Estatística
IME-USP
Caixa Postal 66.281
05311-970 - São Paulo, Brasil