New Approaches in Applied Statistics
Anuška Ferligoj and Andrej Mrvar (Editors)
Metodološki zvezki, 16, Ljubljana: FDV, 2000

Some Notes on Evaluating the Prediction Error for the Generalized Estimating Equations

Dario Gregori[1]

Abstract

In spite of the frequent use of generalized estimating equations (Liang and Zeger, 1986), in particular for modeling correlated binary data, there has been devoted very small attention by the literature to arguments like model checking, outliers detection and prediction accuracy evaluation. This paper is intended to focus on the latter aspect, discussing the applicability of some common methods to the generalized estimating equation model: *(i) Apparent error*, naive or adjusted according to several criteria (C$_p$, AIC, BIC); *(ii) cross-validation*; *(iii) bootstrap* based methods. The main difficulty in using cross-validation and bootstrap arises from the need of retaining the correlation structure in the data. By sampling clusters instead of observations we retain the correlation present in observations belonging to the same cluster. An advantage of this technique over more model-dependent techniques like bootstrapping residuals is that correlation remains a nuisance term, in line with the spirit of the generalized estimating equations, for which a precise assumption of correlation structure is not needed. Internal and external prediction error are evaluated using the proposed methods with reference to a case study of public health

1 Introduction

Generalized Estimating Equations, also known with the acronym GEE, are a common tool for the analysis of dependent data in medical and biological sciences. Their

[1] Department of Economics and Statistics, University of Trieste, Italy.

This work has been partially supported by a grant MURST 60% 97-98. The author wish to thank an anonymous referee for his/her comments and suggestions, that greatly improved the quality of the manuscript. Responsibility remains of course to the author.

use is still increasing in popularity, both in reaching new fields, like ecology and sociology, and in exploiting new ways of applicability in more traditional fields. There are several reasons to this success: perhaps one of the most important is that GEE allow people to think to situation in which data are dependent in some ways as if they where not. There is no need for people to get involved in complicated modeling of the covariance structure of the data if the main focus of the analysis is in the regression parameters. Indeed, the regression coefficients are interpretable as in the most common regression tools, like linear and (specially) logistic regression. Models based on GEE are however not free of problems: the finite sample behavior is not always well (Gregori and Carmeci, 1996), the model checking is complicated by the introduction of a new levels of analysis, represented by the so-called *clusters* of correlated observations, the correlation structure itself can lead to some computational problems, in particular with large and unequally sized clusters. In spite of this potential pitfalls, very small attention has been given in the statistical literature to the aspects related to model checking for the Generalized Estimating Equations. Influence measures has been proposed by Preisse (Preisse and Qaqish, 1996), missing data treatment has been discussed by Paik (Paik, 1992), but the use of GEE for predictive purposes has not been discussed.

In Section 2 we review generalized Estimating equations introducing some notation. The definition of the relevant quantities for the evaluation of prediction error for GEE models is discussed in Section 3, where cross-validation and bootstrap approach are illustrated in a dependent data framework. Computational aspects and appropriate formulae are given in Section 2. In Section 4 the proposed measures are discusses through an illustrative example on public health data.

2 Generalized estimating equations

Let $Y_i = (Y_{i1}, \ldots, Y_{in_i})'$ be a vector of response values, and $X_i = (x_{i1}, \ldots, x_{in_i})'$ a $n_i \times p$ matrix of covariate values. Let $i = 1, \ldots, K$ index the cluster and let $j = 1, \ldots, n_i$ index the observations. In the terminology of the generalized linear models, the forms of the first two moments for the marginal distribution of Y_{ij} are

$$E(Y_{ij}) = \mu_{ij}, \quad g(\mu_{ij}) = \eta_{ij} = x_{ij}\beta, \quad \operatorname{var}(Y_{ij}) = V_{ij}\theta \tag{2.1}$$

where $g(\mu_{ij})$ is the link function, $V_{ij}(\mu_{ij})$ is a variance function of the mean, β is a $p \times 1$ vector of regression coefficients, and θ is the scale parameter, either known or to be estimated.

Estimates of β are obtained by solving the generalized estimating equations

$$\sum_{i=1}^{K}\left(\frac{\partial\mu_i}{\partial\beta}\right)'(A_iR_i(\alpha)A_i)^{-1}(Y_i-\mu_i)=0 \tag{2.2}$$

where $\frac{\partial\mu_i}{\partial\beta}$ is a $n_i\times p$ matrix, $A_i=\text{diag}(V_{ij}^{\frac{1}{2}})$ is a $n_i\times n_i$ diagonal matrix and $R_i(\alpha)$ is a $n_i\times n_i$ working correlation matrix that depends on an unknown parameter vector α.

Working correlation

The dependences among observations are specified in a variety of ways. The most common specifications for the $\text{corr}(Y_i)$ are

1. $R_i(\alpha)=I$, where I is a $n_i\times n_i$ identity matrix. This corresponds to the working independence assumption
2. $\text{corr}(Y_{is},Y_{it})=\alpha$, with $s\neq t$ is the so- called exchangeable correlation
3. $\text{corr}(Y_{is},Y_{it})=\alpha^{|s-t|}$ with $s\neq t$ is the autoregressive form of the working correlation function
4. $\text{corr}(Y_{is},Y_{it})=\alpha_{st}$, where α is a $\frac{n_i(n_i-1)}{2}\times 1$ vector containing all the pairwise correlations. This corresponds to the unstructured or pairwise working correlation

Estimation

We define $N=\sum_i n_i$, the $N\times 1$ vector $Y=(Y_1',\ldots,Y_K')'$, the $N\times p$ matrix $X=(X_1',\ldots,X_K')'$, assumed to be of full rank and $D=\frac{\partial\eta}{\partial\mu}$ a $N\times N$ diagonal matrix with non zero elements $d_{ij}=\frac{\partial\eta_{ij}}{\partial\mu_{ij}}$. Estimation of β is done with iteratively reweighted least squares by regressing the working response vector $M=X\hat{\beta}+D(Y-\hat{\mu})$ on X with block diagonal weight matrix W. The W_i block corresponding to the i-th cluster is the $n_i\times n_i$ matrix

$$W_i=D_i^{-1}A_i^{-1}R_i^{-1}(\hat{\alpha})A_i^{-1}D_i^{-1},\quad D_i=\text{diag}(d_{i1},\ldots,d_{in_i}) \tag{2.3}$$

Under some regularity conditions it has been shown (Liang and Zeger, 1986) that as $K\to\infty$, $K^{\frac{1}{2}}(\hat{\beta}-\beta)$ is asymptotically multivariate Gaussian with mean vector 0 and covariance matrix given by

$$J_{\hat{\beta}}=\lim_{K\to\infty}KJ_1^{-1}J_2J_1^{-1} \tag{2.4}$$

where

$$J_1=\sum_{i=1}^{K}\left(\frac{\partial\mu_i}{\partial\beta}\right)'[A_iR_i(\alpha)A_i]^{-1}\left(\frac{\partial\mu_i}{\partial\beta}\right) \tag{2.5}$$

$$J_2 = \sum_{i=1}^{K} \left(\frac{\partial \mu_i}{\partial \beta} \right)' [A_i R_i(\alpha) A_i]^{-1} \mathrm{cov}(Y_i) [A_i R_i(\alpha) A_i]^{-1} \left(\frac{\partial \mu_i}{\partial \beta} \right) \tag{2.6}$$

The robust variance estimate of $\hat{\beta}$ is obtained by replacing $\mathrm{cov}(Y_i)$ by $(Y-\hat{\mu})(Y-\hat{\mu})'$ and β, θ, α by their estimates in $J_1^{-1} J_2 J_1^{-1}$. The model is robust in the sense that it consistently estimates $J_{\hat{\beta}}$ even if $R(\alpha)$ is mis-specified.

The estimated adjusted residual vector is then

$$E = D(Y - \hat{\mu}) = M - \hat{\eta} = (I - H)M \tag{2.7}$$

where $H = QW$, $Q = Z(Z'WZ)^{-1}Z'$ and $\hat{\eta} = HM$.

3 Evaluation of the prediction error

The main approach in estimating the prediction error rate is in the definition of a prediction rule to be constructed from a training set of dependent data.

Dropping the cluster index, the training set is defined, as in Section 2, as $Z = (X_1', \ldots, X_K')'$ as the N observations $z_{ij} = (x_{ij}, y_{ij})$ with the x_{ij} being the predictor or feature vector and y_{ij} being the response for the j-th observation in cluster i. In particular, we will use the notation $Q[y, r]$ to indicate the discrepancy between a predicted value r and the actual response y. The short notation

$$Q(z_0, Z) = Q[y_0, r_Z(x_0)] \tag{3.1}$$

is commonly used to indicate the discrepancy between the predicted value and response for a test point $z_0 = (x_0, y_0)$, when using the rule r_Z based on the training set Z.

In the approach we are proposing the bootstrap is based on the idea of treating clusters as the units on which bootstrap is performed. This implies that we will assume that the clusters z_i in the training set are a random sample from some distribution F. This assumption is quite general since it usually depends only on the study design adopted. Notice also that only the information of independence among clusters is required at this stage. The dependency structure within cluster is not modeled, as in the the usual GEE setting.

3.1 Internal (apparent) and external error rate

The prediction error for $r_Z(x_0)$ is defined by

$$\mathrm{err}(Z, F) = E_{0F}\{Q[Y_0, r_Z(x_0)]\} \tag{3.2}$$

where the notation E_{0F} indicates expectation over a new observation (x_0, Y_0) from the population F. This is also known as the estimate of the external error rate, provided that a suitable set of new observations Y_0, x_0 are available. This is however a rather uncommon situation. More often, the error rate is estimated on the basis of the training set (y, x).

Using the short notation g for the observation (i, j), so that $g = 1, \ldots, N$, the *apparent* error rate is estimated by

$$\mathrm{err}(x, \hat{F}) = E_{0\hat{F}}\{Q[Y_0, r_Z(x_0)]\} = \frac{1}{N}\sum_{g=1}^{N} Q[y_g, r_Z(x_g)] \tag{3.3}$$

where the $E_{0\hat{F}}$ simply averages over the N observed cases (x_g, y_g).

Cross-validation

The K-cluster cross-validation estimate is

$$\frac{1}{N}\sum_{g=1}^{N} Q[y_g, r_Z^{-k(g)}(x_g)] \tag{3.4}$$

where $k(g)$ denote the cluster containing observation g, and $r_Z^{-k(g)}(x)$ is the predicted value at x, computed with the $k(g)$-th cluster removed.

Bootstrap

To construct a bootstrap estimate of prediction error in the correlated data case, we define as $Z^* = \{(x_1^*, y_1^*), \ldots, (x_K^*, y_K^*)\}$ a bootstrap sample of clusters. Then the estimate of the prediction error $\mathrm{err}(Z, F)$ is defined

$$\mathrm{err}(Z^*, \hat{F}) = \frac{1}{N^*}\sum_{g=1}^{N^*} Q[y_g, r_{Z^*}(x_g)] \tag{3.5}$$

where N^* indicates the total sample size $\sum n_g^*$ for the bootstrap sample (x_i^*, y_i^*) and $g = 1, \ldots, N^*$. In this expression $r_Z^*(x_g)$ is the predicted value at $x = x_g$, based on the model estimated from the bootstrap data set x^*. The notation N^* indicates that the number of observations on each bootstrap replication is not in general equal to the total sample size N, unless of a design completely balanced among clusters. There are several choices proposed to "balance" the study design in order to have approximately the same number of observations in each bootstrap replication (Efron and Tibshirani, 1993). However, the usefulness of such a rounding strategy has not been shown completely, and in any case the asymptotic argument for the convergence of both the GEE estimator and the bootstrap is based on the number of clusters K

tending to infinitum, independently from the cluster size. In addition, more than this "one-shot" estimate, we will focus on the average prediction error

$$E_{\hat{F}}[\mathrm{err}(x^*, \hat{F})] = E_{\hat{F}}\left[\frac{1}{N^*}\sum_{g=1}^{N^*} Q[y_g, r_{Z^*}(x_g)]\right] \tag{3.6}$$

where E_F is the expectation over data sets z with observations z_g F. Expression 3.6 is an ideal bootstrap estimate, corresponding to an infinite number of bootstrap samples. With a finite number B of bootstrap samples, we approximate this letting $r_Z^{*b}(x_g)$ be the predicted value at x_g from the model estimated on the b-th bootstrap sample, with $b = 1, \ldots, B$. Then the approximation to $E_{\hat{F}}[\mathrm{err}(x^*, \hat{F})]$ is

$$E_{\hat{F}}[\mathrm{err}(x^*, \hat{F})] = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{N_b^*}\sum_{g=1}^{N_b^*} Q[y_g, r_{Z^*}(x_g)] \tag{3.7}$$

A more refined bootstrap approach estimates the bias in $\mathrm{err}(Z, \hat{F})$ as an estimator of $\mathrm{err}(Z, F)$ and then corrects $\mathrm{err}(Z, \hat{F})$ by subtracting the estimated bias. The average optimism is defined as

$$\omega(F) = E_F[\mathrm{err}(Z, F) - \mathrm{err}(Z, \hat{F})] \tag{3.8}$$

This is nothing but the average difference between the true prediction error and the apparent error, over data sets Z with observations $Z_g\tilde{F}$; notice that it is usually a positive quantity, since the apparent error rate tends to underestimate the prediction error. The bootstrap estimate of $\omega(F)$ is obtained as

$$\omega(\hat{F}) = E_{\hat{F}}[\mathrm{err}(Z^*, \hat{F}) - \mathrm{err}(Z^*, \hat{F}^*)] \tag{3.9}$$

Here $\hat{F}^*$ is the empirical distribution function of the bootstrap sample Z^*. A practical approximation is

$$\hat{\omega}(\hat{F}) = \frac{1}{B}\left\{\sum_{b=1}^{B}\frac{1}{N_b^*}\sum_{g=1}^{N_b^*} Q[y_g, r_{Z^{*b}}(x_g)] - \sum_{b=1}^{B}\frac{1}{N_b^*}\sum_{g=1}^{N_b^*} Q[y_{gb}^*, r_{Z^{*b}}(x_g^*)]\right\} \tag{3.10}$$

In the above equation, $r_{Z^{*b}}(x_g)$ is the predicted value at x_g^* from the model estimated on the b-th bootstrap sample $b = 1, \ldots, B$ and y_{gb}^* is the response value of the g-th observation for the b-th bootstrap sample. Then, the final estimate of the prediction error is the apparent error plus the downward bias in the apparent error given by

$$\mathrm{err}(Z, \hat{F}) + \omega(\hat{F}). \tag{3.11}$$

3.2 The choice of $Q[\cdot]$

The choice of the discriminating function Q is quite difficult. The most common option, based on the evaluation of the (quasi) deviance function is applicable only under some specific conditions on the structure of the covariance matrix (McCullagh and Nelder, 1989). Other options has been derived and presented in the literature (Hanfelt and Liang, 1995), in particular to avoid problems related to the poor behavior of the Wald test and the known lack of some *ad hoc* goodness of fit tests. An alternative approach could consists in referring to the classical functions used in the literature related to the classification problem, like the residual sum of squares (RSE) or the Pearson statistics. For instance, the latter is an estimate of the overdispersion parameter, and therefore in line with the spirit of the Regal's approach (Hook and Regal, 1992).

In addition, for the dichotomous case where both r and y are either 0 or 1, a very common approach consists in determining the classification function as a rounding up function, having

$$Q[y, r] = \begin{cases} 0 & \text{if } r = y \\ 1 & \text{if } r \neq y \end{cases} \tag{3.12}$$

Typically the rule $r_z(x)$ will predict $y = 0$ if x lies to the lower left of the discriminating function, and $y = 1$ if x lies in the upper right (Efron and Tibshirani, 1996).

In particular, five measures of goodness of fit has been evaluated in the forthcoming case-study: (i) the residual squared error, as defined in Section 3.2 (ii) the Pearson's χ^2 (iii) the quasi-deviance function, constructed under the hypothesis of independence and two round-off function as defined in equation 3.12(iv) based on a threshold of 0.50 of the estimated probability of success, and (v) based on a .75 threshold.

4 Illustration of the methods

Data from the VERO-Chest Study (University of Trieste, Italy) are used to illustrate the use of the diagnostics presented in Section 3.

The data

The VERO-Chest is a prospective multicenter study conducted in 1996 to evaluate the influence of preoperative chest radiography (POCR) on anaesthetic management and to characterize the patient eligible for POCR. 6111 patients entered the study and were submitted to elective surgery, abnormal POCRs were reported in 1116 patients (18.2%). POCR was considered useful for anaesthetic management in 226 patients(5.12%). Male sex, age >51 years, ASA classes >3, coexisting respiratory diseases, and the presence of two or more coexisting diseases were significantly related with the probability of a useful POCR, with wide variations among hospitals. Indeed, the study indicates that in healthy, female, >50-year-old patients, submitted to standard operations, the probability of a useful POCR ranges from 0.2% to 3.5% among hospitals. The probability increases differently in male or elderly subjects, or in the presence of a coexisting respiratory disease, or in ASA classes >3, depending on the particular hospital. For the purposes of illustration, we randomly selected a subset of patients equal to 20% of the original sample size and used them as training set. Other 10% of the patients were randomly selected from the VERO-Chest population and constitute the test set. The marginal distribution of both the training and the test set are quite close each other (Tables 1-2). A major difference lies in the distribution of 'Not useful' x-rays among people with different co-morbidities (53% of not useful x-rays in the training set had more than one co-morbidity whereas only 37% in the test set).

Table 1: Descriptive statistics by xray.

	Yes ($N = 1027$)	No ($N = 53$)	Combined ($N = 1080$)
Sex : Male	46% (473)	58% (31)	47 % (504)
Age	39 56 69	63 70 75	40 56 70
Intervention : Standard	77% (790)	74% (39)	77 % (829)
Minor	13% (137)	6% (3)	13 % (140)
Major	10% (100)	21% (11)	10 % (111)
ASA : 3- 5	15% (152)	60% (32)	17% (184)
Co-morbidity : None	61% (630)	30% (16)	60 % (646)
Cardiac	6% (66)	6% (3)	6 % (69)
Respiratory	3% (35)	8% (4)	4 % (39)
Other	16% (163)	4% (2)	15 % (165)
Several	13% (133)	53% (28)	15 % (161)

a b c represent the lower quartile a, the median b, and the upper quartile c for continuous variables.
N is the number of non–missing values.
Numbers after percents are frequencies.

The distribution of observation among clusters (hospitals) is one of the quantities that should always be explored before performing an analysis of correlated (clustered) data (Louis, 1988). In our samples, the distribution tends to reflect that

Table 2: Descriptive statistics by xray.

	Yes ($N = 573$)	No ($N = 27$)	Combined ($N = 600$)
Sex : Male	45% (259)	59% (16)	46 % (275)
Age	38 56 69	59 64 79	38 57 69
Intervention : Standard	72% (414)	81% (22)	73 % (436)
Minor	16% (93)	4% (1)	16 % (94)
Major	12% (66)	15% (4)	12 % (70)
ASA : 3- 5	14% (79)	52% (14)	16% (93)
Co-morbidity : None	61% (349)	30% (8)	60 % (357)
Cardiac	6% (32)	7% (2)	6 % (34)
Respiratory	2% (10)	19% (5)	2 % (15)
Other	16% (94)	7% (2)	16 % (96)
Several	15% (88)	37% (10)	16 % (98)

a b c represent the lower quartile a, the median b, and the upper quartile c for continuous variables.
N is the number of non–missing values.
Numbers after percents are frequencies.

of the original data set. It has to be noticed however, that the cluster number 15 is equally sized in the training and test set. Another fact is that clusters number 3 and 4 are made of very few observations in the test set. An obvious consequence of this might be the reduction of the observed length of the bootstrap series due to the presence of some not estimable models.

The models

The estimated models are presented in Table 4, assuming exchangeable and unstructured correlation structure for the $R(\alpha)$ matrix. Point estimates of the regression coefficients and the adjusted standard errors do not differ much among the two models. Coefficients significantly different from zero are those related to ASA physical status, co-morbidity (in particular cardiac vs none), intervention type (major vs common) and age. All of them seem indicate an increase of the probability of having some useful results form the x-ray according to the worsening scenario of the patient. This conclusion agrees, beside some small differences not relevant from the point of view of interpretation, with the model estimated on the total population of 611 patients.

Before evaluating the performance of each criterion, it is useful to highlight the target of this kind of analysis: (i) the target of the analysis of the predictive

Table 3: Cluster size in training and test set.

Training set

Hospital	1	2	3	4	5	6	7	8	9	10	
n_i	95	5	9	27	136	9	31	20	74	191	
Hospital	11	12	13	14	15	16	17	18	19	20	21
n_i	55	5	7	73	10	126	51	21	31	75	29

Testset

Hospital	1	2	3	4	5	6	7	8	9	10	
n_i	46	3	4	6	78	10	17	10	33	102	
Hospital	11	12	13	14	15	16	17	18	19	20	21
n_i	41	3	4	33	11	76	35	14	17	39	18

Table 4: GEE estimates of the model based on data in the training set.

	Exchangeable correlation		Unstructured correlation	
Coefficients	Estimate	Robust S.E.	Estimate	Robust S.E.
(Intercept)	*-5.533*	0.792	-5.789	0.802
ASA	*1.307*	0.461	1.439	0.501
Sex	0.139	0.316	0.121	0.327
Co-morbidity: Cardiac	*0.514*	0.302	0.579	0.310
Co-morbidity: Respiratory	0.596	0.643	0.598	0.654
Co-morbidity: Other	-1.480	1.128	-1.730	1.301
Co-morbidity: Several	0.654	0.698	0.732	0.711
Intervention: Minor	-0.696	0.612	-0.701	0.619
Intervention: Major	*0.515*	0.246	0.599	0.275
Age	*0.031*	0.013	0.029	0.011

Table 5: Error rates estimates.

	App. Err.	Ext. Err.	CV^{21}	B^{21}	B^n
Exchangeable correlation					
RSE	0.042	0.102	0.081	0.099	0.182
Pearson	2.61	2.98	2.7	2.60	2.70
D^Q	342.703	298.776	390.5	392.7	390.887
C50	0.103	0.181	0.129	0.118	0.149
C75	0.237	0.310	0.198	0.210	0.245
Unstructured correlation					
RSE	0.058	0.120	0.119	0.089	0.143
Pearson	2.61	2.86	2.78	2.64	2.67
D^Q	349.745	312.889	352.329	360.943	599.781
C50	0.115	0.122	0.156	0.144	0.377
C75	0.241	0.290	0.222	0.237	0.487

capability of a model is the estimation of the error rate, for which the evaluation of the apparent error rate provide only a biased information. The external error rate is, on the other side, an unbiased estimator of this quantity, but it needs a test sample which is not always available, and (ii) therefore, the need to have a procedure able to reduce the bias in the apparent error is of great importance. Following this lines, the estimates of the five criteria (Table 5) have been evaluated for the training set and presented as an estimate of the true error rate. The external error rate has been estimated on the test set of 600 observations.

The apparent error rate is half the size of the external error rate when measured using the common measure of RSE. The same happens for the $C50$ criterion. The $C75$ and the Pearson criteria provide an estimate of the apparent error which is closer to the estimate for the external error rate. It has to be noticed that the quasi-deviance function provides an estimate of the apparent error which is more conservative that the external error itself. This is not surprising since this is the only measure incorporating information about the distribution function inside each cluster. We the evaluated these measures using both the cross-validation and the bootstrap criteria.

The re-sampling statistics

The bias corrected bootstrap has been calculated both using the resamplig-clusters approach described in Section 3.1 and the *naive* bootstrap obtained re-sampling

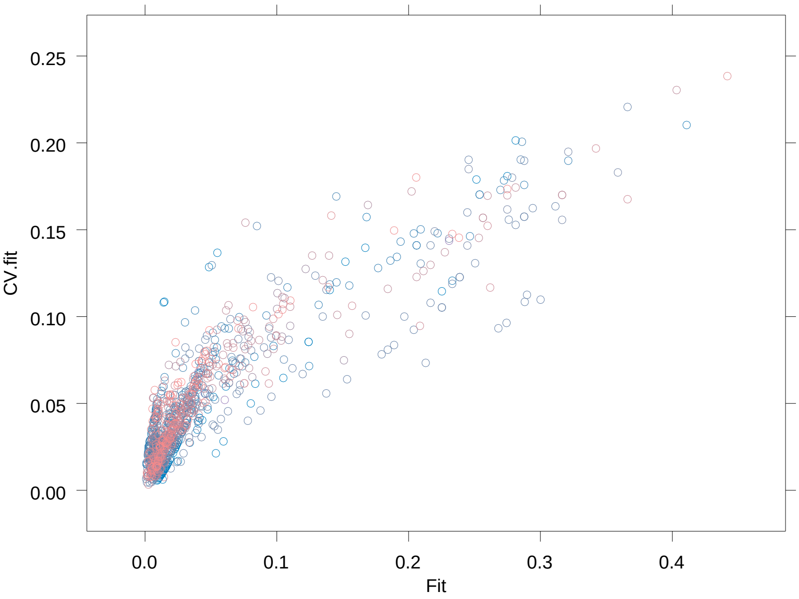

Figure 1: GEE model fit (exchangeable correlation). Cross validation versus model fit.

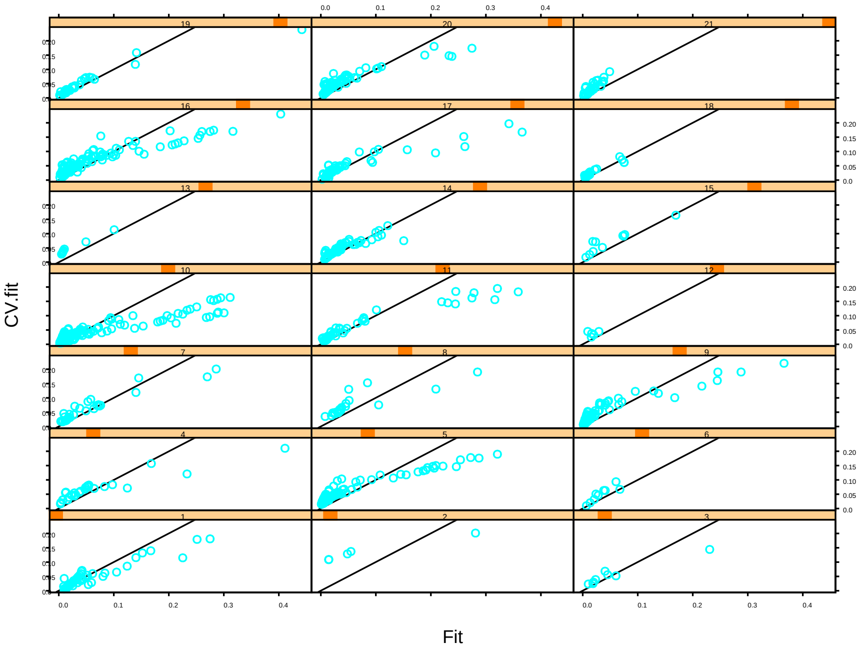

Figure 2: GEE model fit (exchangeable correlation). Cross validation versus model fit for each cluster.

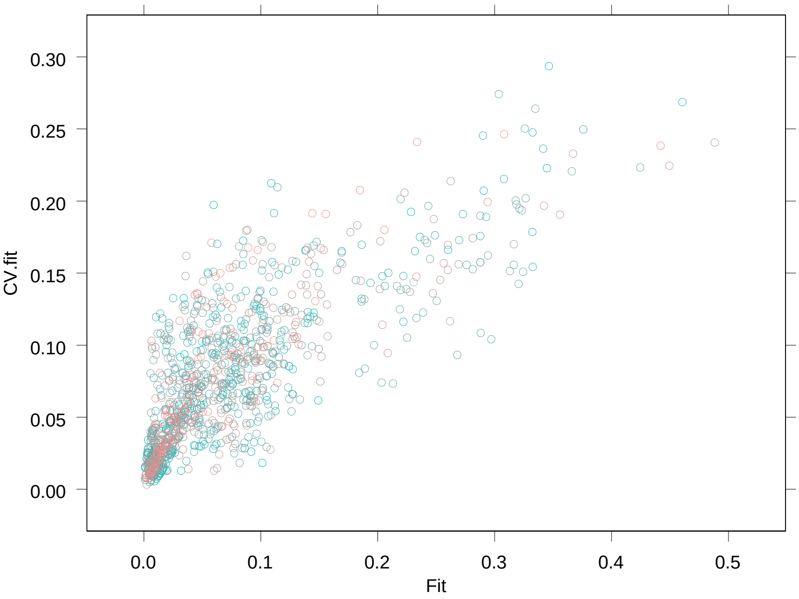

Figure 3: GEE model fit (unstructured correlation). Cross validation versus model fit.

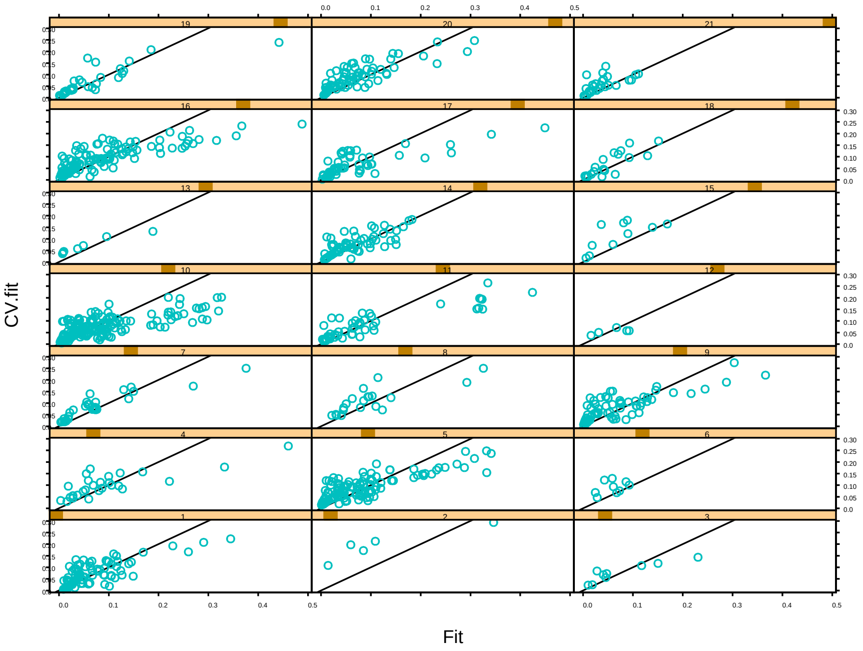

Figure 4: GEE model fit (unstructured correlation). Cross validation versus model fit for each cluster.

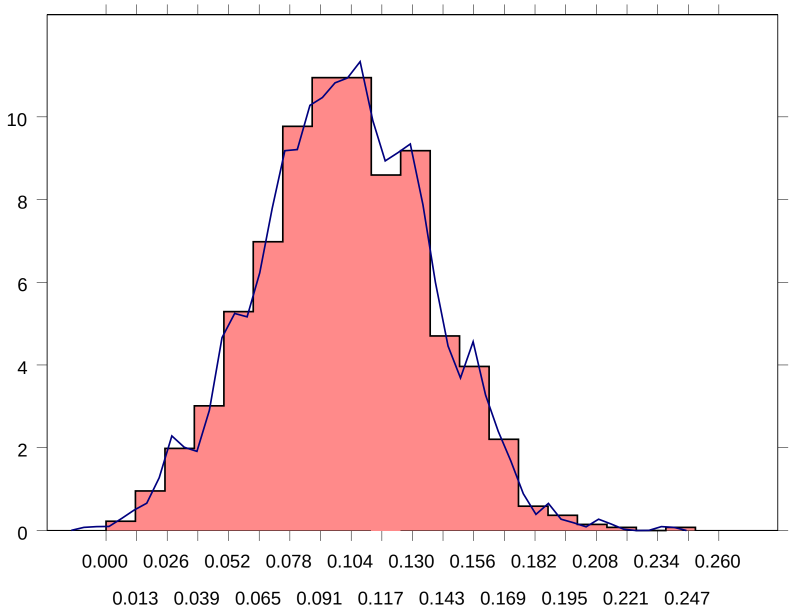

Figure 5: Bootstrap estimates of prediction error (apparent + optimism) for the GEE model with exchangeable correlation. Cluster bootstrap (B=1000).

the N units as if they where independent. The cross-validation and the cluster bootstrap perform always better than the simple apparent error estimates, in particular when using the common RSE criterion . The *naive* bootstrap seems unstable and providing misleading information, getting worse as the the complexity of the statistics increases. It has to be noticed, however, that all bootstrap models, but in particular those based on an unstructured working correlation matrix, have a high percentage of models not reaching convergence, for instance about one third of them. Cross-validation seems in particular more stable than other approaches with respect to different choices of the working correlation structure. The Figures 1-2 and 3-4 do not show any particular difference in the fit between the exchangeable and the unstructured model.

The asymptotic distribution of the bootstrap estimates has been checked, both for the B^{21} and the B^n with two working correlations, exchangeable and unstructured. The B^{21} has a distribution close to the gaussian curve for both correlation structures (Figures 5 and 6). The same does not apply to the B^n, which is performing very poorly, in particular when the unstructured correlation is chosen (Figures 7 and 8).

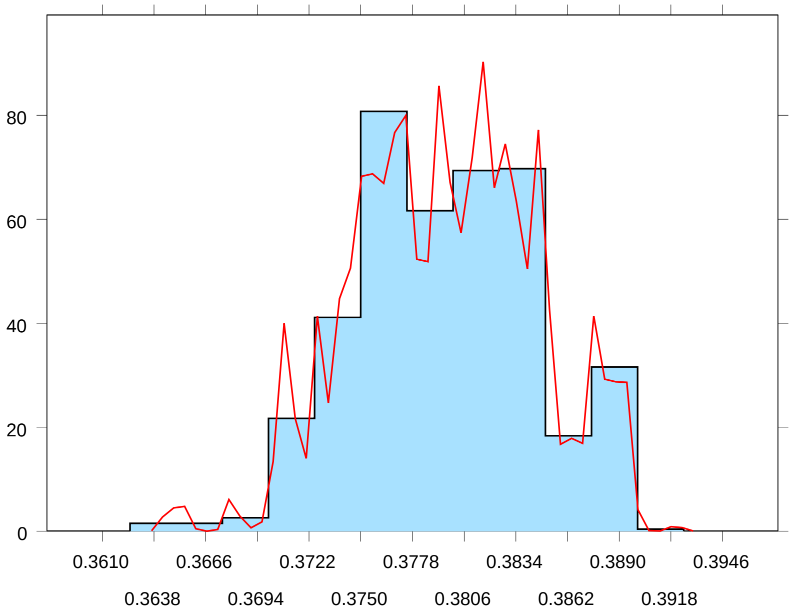

Figure 6: Bootstrap estimates of prediction error (apparent + optimism) for the GEE model with exchangeable correlation. "Naive" bootstrap (B=1000).

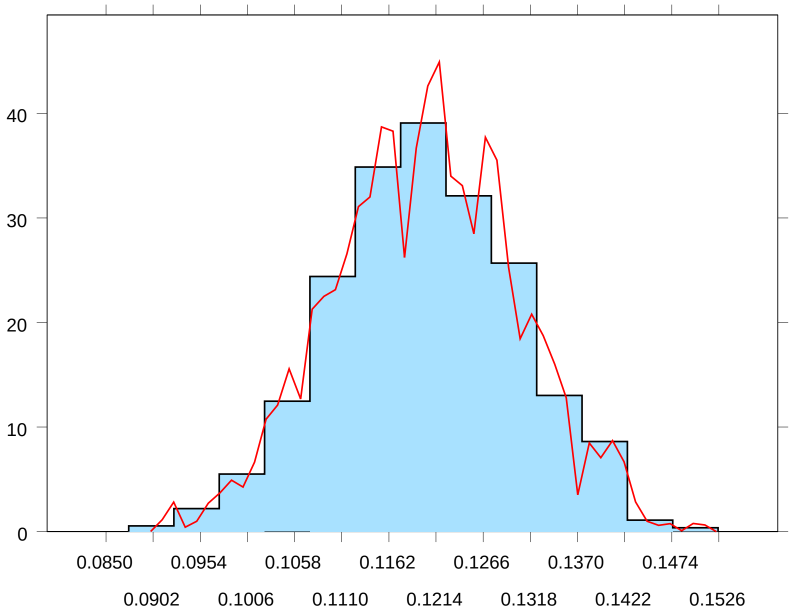

Figure 7: Bootstrap estimates of prediction error (apparent + optimism) for the GEE model with unstructured correlation. Cluster bootstrap (B=1000).

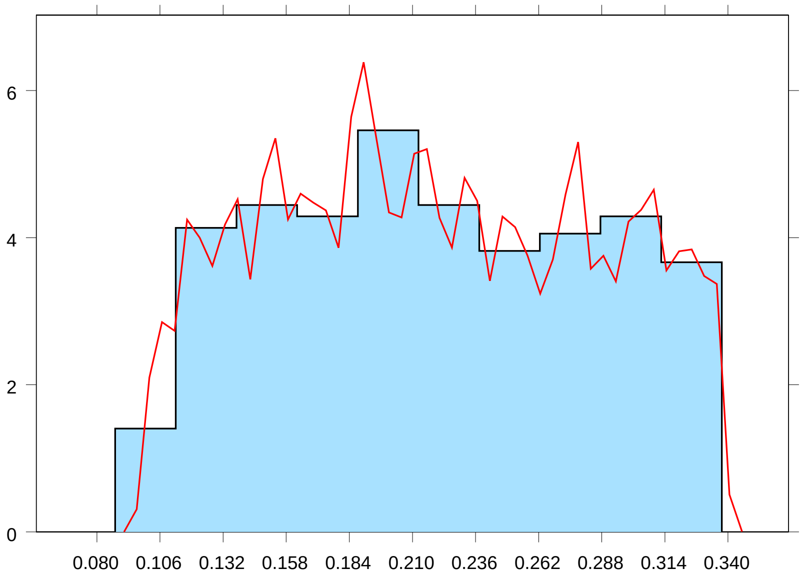

Figure 8: Bootstrap estimates of prediction error (apparent + optimism) for the GEE model with unstructured correlation. ”Naive” bootstrap (B=1000).

5 Conclusive remarks

The question we had in mind in writing this paper was: is it possible to get more information about the appropriateness of a GEE model using re-sampling techniques? In fact, it has been pointed out that cross-validation is not a panacea for estimating the predicted value or to select a model (Cressie, 1991); indeed it cannot provide information about the fact that the model is correct, but only that it is not clearly incorrect. However, a critical use of this technique can provide an useful insight in situations where the typical techniques fail to provide a valid indication (Schumacher, 1995). The classical criteria like for instance the AIC (Akaike, 1973) and the BIC (Schwarz, 1978) are not applicable in situations where data are dependent (Cressie, 1991). Alternative re-sampling schemes has been proposed for more specific data structure, in particular for time-series sequences (Hesterberg, 1997). Other ways of bootstrapping are based on the idea that a suitable model can be written so that the resulting underlying structure is in terms of i.i.d. components. This idea, proposed by several authors (Freedman and Peters, 1984; Solow, 1985) is highly dependent on the estimation procedure used for the model and works fine in case of no dependence of higher moments (Cressie, 1991) where the GEE (one) are known to have troubles (Carmeci and Gregori, 1995). In case of the semi-parametric model, the ideal method for estimating predictive accuracy should retain all the nice properties of GEE models. In fact, we observed that cross-validation in particular, but also bootstrap methods can be of some help in evaluating models, especially when no other samples are available to serve as test set. It is in addition evident that retaining information about correlation is fundamental, as already stated in previous studies (Freedman and Peters, 1984). How to choose among the various criteria is not entirely clear. From the evidence coming out of this analysis, it seems that simple methods, like Pearson over-dispersion estimate or round-off threshold functions are the most suitable. In particular for the latter, a deeper analysis to determine the sensitivity to the threshold used in rounding-off is needed.

References

[1] Akaike, H. (1973): Information Theory as an extension of the maximum likelihood principle. In Petrov, B. N. and Csaki, F. (Eds.), *Second International Symposium on Infromation Theory.* Akademiai Kaido, Budapest, 267-281.

[2] Carmeci, Gaetano and Gregori, Dario (1995): Higher order moment assumptions for GEE models: a simulation study. *21st European Meeting of Statisticians, Programme and Abstract Book*, 127, University of Aahrus, for Bernoulli Society.

[3] Cressie, Noel A.C. (1991): *Statistics for Spatial Data.* New York: John Wiley & Sons.

[4] Efron, B. and Tibshirani, R. (1996): *Cross-Validaiton and the Bootstrp: Estimating the Error Rate of a Prediction Rule.*

[5] Efron, Bradley and Tibshirani, Robert, J. (1993): *An Introduction to the Bootstrap.* Monographs on Statistics and Applied Probability, **57**. New York: Chapman & Hall.

[6] Freedman, D.A. and Peters, S.C. (1984): Bootstrapping a Regression Equation: Some Empirical Results. *Journal of the American Statistical Association*, **79**, 97-106.

[7] Gregori, Dario and Carmeci, Gaetano (1996): Some notes on small sample behavior of GEE. *Statistica Applicata*, **8**, 1-26.

[8] Hanfelt, John J. and Liang, Kung Yee (1995): Approximate likelihood ratios for general estimating functions. *Biometrika*, **82**, 461-77.

[9] Hesterberg, Tim (1997): *Matched-Block Bootstrap for Long Memory Processes.* Research Report 66. MathSoft Inc., Seattle (WA), USA.

[10] Hook, E. B. and Regal, R. R. (1992): The value of capture-recapture methods even for apparent exhaustive surveys. The need for adjustment for source of ascertainment intersection in attempted complete prevalence studies. *American Journal of Epidemiology*, **135**, 1060-1067.

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

[12] Louis, T.A. (1988): General methods for analysing repeated measures. *Statistics in Medicine*, **7**, 29-45.

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

[14] Paik, Myunghee C. (1992): Parametric Variance Function Estimation for Non-Normal Repeated Measurement Data. *Biometrics*, **48**, 19-30.

[15] Preisse, John S. and Qaqish, Bahjat F. (1996): Deletion Diagnostics for Generalized Estimating Equations. *Biometrics*, **83**, 551-562.

[16] Schumacher, M. (1995): Resampling and cross-validation techniques: a tool to reduce bias caused by model building. ISCB (Ed.), *Final Programme and Abstracts.* 37. International Society of Clinical Biostatistics.

[17] Schwarz, M. (1978): Estimating the dimension of a model. *Annals of Statistics*, **6**, 461-464.

[18] Solow, A.R. (1985): Bootstrapping correlated data. *Journal of the International Society for Mathematical Geology*, **17**, 769-775.