Variable Selection for Marginal Longitudinal Generalized Linear Models

Eva Cantoni, Joanna Mills Flemming and Elvezio Ronchetti

Department of Econometrics

University of Geneva

CH - 1211 Geneva 4, Switzerland

Eva.Cantoni@metri.unige.ch, Joanna.Flemming@metri.unige.ch,

Elvezio.Ronchetti@metri.unige.ch

June 2003

Revised September 2004

SUMMARY. Variable selection is an essential part of any statistical analysis and yet has been somewhat neglected in the context of longitudinal data analysis. In this paper we propose a generalized version of Mallows's C_p (GC_p) suitable for use with both parametric and nonparametric models. GC_p provides an estimate of a measure of model's adequacy for prediction. We examine its performance with popular marginal longitudinal models (fitted using GEE) and contrast results with what is typically done in practice: variable selection based on Wald-type or score-type tests. An application to real data further demonstrates the merits of our approach while at the same time emphasizing some important robust features inherent to GC_p.

KEY WORDS: C_p; Generalized estimating equations (GEE); Prediction error; Robustness; Variable selection.

1. INTRODUCTION

Variable (or model) selection is an essential part of any statistical analysis. Although it is often perceived as an extensive search for a single *best* model, it should be viewed as a technique which facilitates the identification of a few *good* models. After all, in many contexts it may not be appropriate to choose a single model. Moreover, one can often achieve better prediction results by aggregating a collection of *good* models in the spirit of bagging; cf. Breiman (1996). This implies that variable selection criteria which allow direct comparisons of models should be preferred to stepwise procedures based on significant testing. Several such criteria exist, including AIC, Schwartz, BIC, and C_p; for an overview see e.g. McQuarrie and Tsai (1998) and Burnham and Anderson (2002). In this paper we focus on such a variable selection technique. It is an extension of Mallows's C_p (Mallows, 1973) and it requires only the data and a procedure from which predicted values can be obtained. The technique does not require a parametric model and can be applied to many different types of models, including those in which the classical assumptions (e.g. independence of variables, normal distribution) do not hold as for instance with binary data (e.g. a subject having a disease or not at a particular point in time) taken over time on the same subject.

We focus on Marginal Longitudinal Generalized Linear Models and develop our variable selection technique for these models. Generalized Linear Models (GLM, McCullagh and Nelder, 1989) and Generalized Estimating Equations (GEE, Liang and Zeger, 1986) are very popular statistical methods which allow us to model a variety of data and properly address the type of situations described above. GLM are a generalization of the regression model for continuous and discrete responses and Marginal Models are extensions of

GLM for correlated data. GEE enable us to fit Marginal Models and are often used for modeling longitudinal data that commonly arise for instance in medical studies and economics. While there have been many novel approaches to analyzing such data, little attention has been paid to the need for appropriate variable selection. In the latest edition of the Analysis of Longitudinal Data (Diggle et al., 2002) a discussion of variable selection techniques has been somewhat neglected, with the exception of a few examples suggesting to the reader, that, in the case of GEE, one can test the significance of covariates using Wald-type test (z-statistics). A notable exception is the recent work by Pan (2001), who developed an Akaike-like criterion for GEE, but with the assumption of working independence. This restriction is not needed in our approach. Horton and Lipsitz (1999) concluded that GEE are well-supported by several software packages with hypothesis testing being particularly well-implemented in some of them. However, it appears that variable selection is restricted to either likelihood ratio or Wald-type tests.

The use of GC_p here for purposes of model selection avoids a stepwise procedure and is based on a measure of predictive error rather than on significance testing.

The paper is organized as follows. In Section 2, we develop a general criterion for prediction and its estimated version which leads to a general and robust C_p statistic. We then derive explicitly this statistic for parametric longitudinal models. In Section 3 we present the results of a simulation study that contrasts our proposal with stepwise procedures and significance testing. Results are examined both in the absence and presence of misspecification of the model. This seems particularly important when investigating tools like the GEE that are used in medical studies where 5% of outlying observations seems to be quite common; cf. Hampel et al. (1986), p. 27. The

results show the favorable performance of the new technique in identifying *good* models. In Section 4 we use our GC_p procedure on real data from a prospective longitudinal study of the process of laryngoscopic endotracheal intubation to identify those features of the process that are most predictive of successful completion. We use a logistic longitudinal model with 9 covariates on 19 subjects performing between 18 and 33 intubations each. Robust and diagnostic features of our approach are further demonstrated. Conclusions and directions for future research are provided in Section 5.

2. DERIVATION OF GC_p

2.1 *General C_p Procedure for Model Selection*

We begin by considering the general setting in which we have only observations y_i, $i = 1, \cdots, K$, and a model, either parametric or nonparametric in form, from which we can obtain predicted values $\hat{y}_i$, $i = 1, \cdots, K$. We define the rescaled weighted predictive squared error

$$\Gamma_P = \sum_{i=1}^{K} E\Big\{ w_i^2 \big(\frac{y_i - \hat{y}_i}{\sigma v_i^{1/2}}\big) \cdot \Big(\frac{\hat{y}_i - Ey_i}{\sigma v_i^{1/2}}\Big)^2 \Big\},$$

where $\hat{y}_i$ is the fitted value for submodel P and Ey_i and $V(y_i) = \sigma^2 v_i$ are the expected value and variance of y_i. The weight function $w_i(\cdot)$ may be chosen so as to achieve a number of different objectives including heteroscedasticity or robustness. If we define the weighted sum of squared residuals by $WSR = \sum_{i=1}^{K} w_i^2(r_i) r_i^2$, where $r_i = \frac{y_i - \hat{y}_i}{\sigma v_i^{1/2}}$, and let $\delta_i = \frac{\hat{y}_i - Ey_i}{\sigma v_i^{1/2}}$, it is easy to show that $\Gamma_P = E(WSR) - \sum_{i=1}^{K} E\{w_i^2(r_i)\epsilon_i^2\} + 2\sum_{i=1}^{K} E\{w_i^2(r_i)\epsilon_i\delta_i\}$, where $\epsilon_i = \frac{y_i - Ey_i}{\sigma v_i^{1/2}}$. This suggests the following generalized version of Mallows's C_p:

$$GC_p = WSR - \sum_{i=1}^{K} E\{w_i^2(r_i)\epsilon_i^2\} + 2\sum_{i=1}^{K} E\{w_i^2(r_i)\epsilon_i\delta_i\}. \qquad (1)$$

The latter two terms comprise the correction term necessary in order to make WSR unbiased for Γ_P. In the selection procedure, $V(y_i) = \sigma^2 v_i$ will be replaced by an "external" estimate typically computed under the full model. It is held fixed while calculating the GC_p of each submodel. In a situation where a "full model" is not available or cannot be fitted, one can select an available large submodel from which to obtain a variance estimate. Notice that for GEE with a logistic regression marginal model, $\sigma^2 = 1$. As previously mentioned, the weights $w_i(\cdot)$ may address heteroscedasticity, robustness or simply be identically one in which case (1) becomes a classical yet generalized version of Mallows's C_p. In the case of robustness, the weights in (1) are different for each model, because an observation can be outlying with respect to one model and have full weight in another. We may select a weighting scheme that has the effect of downweighting the outlying observations with respect to model P and limiting their influence on Γ_P and, therefore, on the model selection procedure. This would not penalize models which do not fit a few outlying observations; cf. Ronchetti and Staudte (1994) in the context of linear regression. A Taylor series expansion of the weights (details are provided in Appendix A) allows us to write the final form of our GC_p statistic as follows:

$$\begin{aligned} GC_p &= WSR - \sum_{i=1}^{K} E\{w_i^2(\epsilon_i)\epsilon_i^2\} + 2\sum_{i=1}^{K} E[\{w_i(\epsilon_i)w_i'(\epsilon_i)\epsilon_i^2 + w_i^2(\epsilon_i)\epsilon_i\}\delta_i] \\ &- \sum_{i=1}^{K} E[\{w_i(\epsilon_i)w_i''(\epsilon_i)\epsilon_i^2 + (w_i'(\epsilon_i))^2\epsilon_i^2 + 4w_i(\epsilon_i)w_i'(\epsilon_i)\epsilon_i\}\delta_i^2]. \end{aligned} \quad (2)$$

To make the definition (2) operational we must be able to compute the latter three terms (comprising the correction). Approximations for these terms can be derived and will depend on the specifics of the model under consideration. In Section 2.2 we compute these terms for longitudinal marginal models.

Models with small values of GC_p will be preferred to others. At this point, the decision of how to proceed will depend on the application at hand. For instance, we may wish to obtain predictions for a few *good* models with small GC_p and then average them in order to draw final conclusions. The weights themselves may also prove very insightful. In the event that they are chosen to address robustness, they can routinely identify outlying observations; cf. Section 4.

2.2 *Computation of GC_p for a Marginal Longitudinal Model*

We now consider a longitudinal data analysis setting, where Y_{it} is the discrete or continuous outcome for subject i at time t, for $i = 1, \cdots, K$ and $t = 1, \cdots, n_i$. For each outcome Y_{it}, we also measure a set of covariates x_{it}. We write $Y_i = (Y_{i1}, \cdots, Y_{in_i})^T$ for the $n_i \times 1$ vector of responses, and $X_i = (x_{i1} \cdots x_{in_i})^T$ for the $n_i \times p$ matrix of covariates of subject i. We suppose that $Corr(Y_i) = A_i^{-1} Var(Y_i) A_i^{-1}$, with $A_i = \text{diag}(v^{1/2}(\mu_{i1}), \cdots, v^{1/2}(\mu_{in_i}))$, and that the subjects (indexed by i) are independent. We model the marginal mean $E(Y_{it}) = \mu_{it}$, and assume that $g(\mu_{it}) = x_{it}^T \beta$ for a known link function g, and that $V(Y_{it}) = \sigma^2 v(\mu_{it})$. For short, we will write v_{it} instead of $v(\mu_{it})$.

An M-estimator $\hat{\beta}_p$ for model P with p parameters is the solution of the estimating equations proposed by Cantoni (2004):

$$\sum_{i=1}^{K} D_i^T \Gamma_i^T V_i^{-1} (\Psi_i - c_i) = 0, \tag{3}$$

where $D_i = D_i(X_i, \beta) = \partial \mu_i / \partial \beta$ is a $n_i \times p$ matrix, and $V_i = V_i(\mu_i, \alpha) = A_i R_i(\alpha) A_i$ is a $n_i \times n_i$ matrix. $R_i(\alpha)$, for an s-parameter α, is said to be the working correlation matrix, as opposed to the "true" correlation matrix $Corr(Y_i) = A_i^{-1} Var(Y_i) A_i^{-1}$. Moreover, $\Psi_i = W_i(Y_i - \mu_i)$ and $c_i = E(\Psi_i)$, where $W_i = W_i(X_i, y_i, \mu_i)$ is a diagonal $n_i \times n_i$ weight matrix containing

weights w_{it} for $t = 1, \cdots, n_i$. These weights may be different to those contained in the definition of our GC_p statistic and can be chosen so as to address a number of different objectives, robustness being one example in which case we refer to Cantoni and Ronchetti (2001) and Cantoni (2004) for a detailed discussion on the choice of weights. Finally, $\Gamma_i = E(\tilde{\Psi}_i - \tilde{c}_i)$ with $\tilde{\Psi}_i = \partial\Psi_i/\partial\mu_i$ and $\tilde{c}_i = \partial c_i/\partial\mu_i$. Note that the classical GEE equations (Liang and Zeger, 1986) are obtained with W_i equal to the identity matrix. Note also that the estimating equation in (3) is a slightly modified version of that in Preisser and Qaqish (1999), in that it includes the matrix Γ_i that makes it optimal in the class of all estimating equations based on $(Y_i - \mu_i)$.

Under the usual regularity conditions for M-estimators (Huber, 1981) the estimator defined as the solution of (3) is asymptotically normally distributed with asymptotic variance $M^{-1}QM^{-1}$, where

$$M = \lim_{K\to\infty} \frac{1}{K} \sum_{i=1}^{K} D_i^T \Gamma_i^T V_i^{-1} \Gamma_i D_i$$

and

$$Q = \lim_{K\to\infty} \frac{1}{K} \sum_{i=1}^{K} D_i^T \Gamma_i^T V_i^{-1} Var(\Psi_i) V_i^{-1} \Gamma_i D_i.$$

For such longitudinal marginal models and writing $\psi(\epsilon_{it}) = w(\epsilon_{it}) \cdot \epsilon_{it}$, GC_p from (2) becomes:

$$GC_p = WSR - \sum_{i=1}^{K} \sum_{t=1}^{n_i} E\{\psi^2(\epsilon_{it})\} + t_1 - t_2, \tag{4}$$

where

$$t_1 \approxeq \frac{2}{\sigma K} \sum_{i=1}^{K} Tr\{M^{-1} E(D_i^T Z_i a_i^T A_i^{-1} D_i)\}, \tag{5}$$

with $Z_i = \Gamma_i^T V_i^{-1}(\Psi_i - c_i)$, $a_i = (a_{i1} \dots a_{in_i})^T$ and $a_{it} = \psi(\epsilon_{it})\psi'(\epsilon_{it})$, and

$$t_2 \approxeq \frac{1}{\sigma^2 K^2} \sum_{i=1}^{K} Tr\Big[E\Big\{B_i A_i^{-1} D_i M^{-1} \big(\sum_{j=1}^{K} D_j Z_j Z_j^T D_j\big) M^{-1} D_i^T A_i^{-1}\Big\}\Big], \tag{6}$$

with $B_i = \text{diag}(b_{i1} \dots b_{in_i})$ and $b_{it} = \psi(\epsilon_{it})\psi''(\epsilon_{it}) - \psi^2(\epsilon_{it})/\epsilon_{it}^2 + (\psi'(\epsilon_{it}))^2$. Computations are provided in Appendix B. Notice that the weighting scheme included in the GC_p definition ignores the within subject correlation. However, the technique we use is designed for situations where this correlation is fairly weak; see Cantoni (2004). Notice also that the expectations in (5) and (6) can be easily evaluated by Monte Carlo. Essentially one generates $nsim$ block correlated outcomes $Y = (Y_1, \dots, Y_K)$ following the algorithm described in Emrich and Piedmonte (1991) and then uses these to compute empirical expectations as approximations to those expectations appearing in t_1 and t_2 above. Note that one uses the current estimates $\hat{\beta}_p$ and $\hat{\alpha}$. We took $nsim = 100$ and found the approximations worked quite well.

If the weights in (2) are chosen to be identically one, we obtain

$$GC_p = \sum_{i=1}^{K}\sum_{t=1}^{n_i} r_{it}^2 - \sum_{i=1}^{K} n_i + 2\sum_{i=1}^{K}\sum_{t=1}^{n_i} E(\epsilon_{it}\delta_{it}), \tag{7}$$

where we have used the fact that $E(\epsilon_{it}^2) = 1$, and where we notice that the term t_2 is exactly zero in this case. Moreover, following the same reasoning leading to (5), we obtain in this case

$$GC_p = \sum_{i=1}^{K}\sum_{t=1}^{n_i} r_{it}^2 - \sum_{i=1}^{K} n_i + \frac{2}{\sigma^2 K}\sum_{i=1}^{K} Tr(M^{-1}D_i^T A_i^{-2} D_i), \tag{8}$$

which can be computed directly without simulation. Hereafter, we refer to (8) as our classical GC_p, and to (4)-(6) as our weighted or robust GC_p.

3. SIMULATION STUDY

We have considered a simulation study that encompasses two designs, two sample sizes and ten techniques. Each combination has been tested on both noncontaminated and contaminated data.

For $i = 1, \ldots, K$ ($K = 15$ and $K = 30$) independent subjects and $t = 1, \ldots, n_i = 10$, we consider a marginal longitudinal model (see Section 2.2) with logistic link, where the linear predictor is $\beta_0 + x_{it}^T\beta$, with x_{it} and β of dimension 5. The response Y_{it} is binary (0 or 1).

Design I: The explanatory variables $x_{it}^T = (D_1, D_2, C_1, C_2, C_3)$ are: a dummy variable D_1, e.g. sex, coded 0 or 1 with probability of each equal to 0.5, a three level variable D_2 with probabilities 0.35, 0.15 and 0.5 respectively, and 3 continuous variables C_1, C_2, and C_3 generated independently according to standard normal distributions. The true values of the parameters are $\beta_0 = 0.5$ and $\beta = (1, 0, 0.5, 0.5, 0)$, meaning that the true model generating the data is defined by the intercept, D_1, C_1, and C_2. The correlation within subjects is exchangeable: for each i, $Corr(Y_{it}, Y_{it'}) = \alpha = 0.1$, for $t \neq t'$.

Design II: Here we replace C_3 above with the interaction variable $I_1 = D_1 * C_1$ and $\beta = (1, 0, 0.5, 0, 0.5)$, which defines the generating model as D_1, C_1, and I_1 plus intercept. All other aspects remain the same as in Design I.

In both designs the full model contains all available variables.

We simulated 100 replications of block-correlated binary responses. This sample represents our noncontaminated data. To obtain a slightly contaminated dataset we flipped (from 0 to 1, or vice-versa) 5% randomly chosen responses from each replication contained in the noncontaminated data. This contamination is reflective of what may occur in practice, when in a few cases a zero might be recorded as a one or vice versa. In addition, it is compatible with the fact that we consider observation downweighting, as opposed to cluster downweighting.

For the classical approach, we estimated the parameters according to (3) with $W_i = I$ (and therefore $\Gamma_i = I$ and $c_i = 0$), which reproduces the

Liang and Zeger (1986) equations, and investigate five variable selection procedures: 1) GC_p: identify the model with smallest GC_p as defined by (7); 2) z-test: fit the full model and retain all the variables for which the Wald test gives a p-value lower than 0.05; 3) z-stepwise: backward stepwise selection procedure based on Wald test with cutoff on the p-value set at 0.1; 4) score-test: fit the full model and retain all the variables for which the score-type test as per formulae (3)-(5) of Heritier and Ronchetti (1994) with $\boldsymbol{\psi}(\mathbf{z}_i, T_n) = D_i^T V_i^{-1}(Y_i - \mu_i)$ gives a p-value lower than 0.05; and 5) score-stepwise: backward stepwise selection procedure based on score-type test with cutoff on the p-value set at 0.1. For the robust approach, the regression parameters are estimated according to (3) where $W_i = \text{diag}(w_{i1}, \ldots, w_{in_i})$, with $w_{it} = c/|\epsilon_{it}|$ if $\epsilon_{it} > c$ and $w_{it} = 1$ otherwise (c=1.5). The exchangeable correlation parameter α is estimated by a generalized version of the method of moments, as defined in Cantoni (2004) ($k = 2.4$). Five robust variable selection counterparts of the classical techniques are used: our robust GC_p as defined by (4) with $\psi = \psi_c(r) = r\min(1, c/|r|)$ being the Huber's function ($c = 1.5$); a robust z-test and a robust stepwise procedure, both based on a Wald-type test of the form $\hat{\beta}^2_{rob}/\widehat{Var}(\hat{\beta}_{rob})$; a robust score-type test and a robust stepwise procedure, both based on a score-type test as per formulae (3)-(5) of Heritier and Ronchetti (1994) with $\boldsymbol{\psi}(\mathbf{z}_i, T_n) = D_i^T \Gamma_i^T V_i^{-1}(\Psi_i - c_i)$. Note that all the test statistics used (classical and robust) are distributed as χ^2_1 under the null hypothesis, see Heritier and Ronchetti (1994).

As pointed out by a referee, although the classical procedures 2) and 4) are often used, they are a "bad habit" and can be unreliable, because, among other things, the choice of working dependence model can impact point estimates and significance levels. Moreover, Wald's tests in logistic regression for independent observations are known to behave in an aberrant

manner (Hauck and Donner, 1977). We can expect to observe a similar behavior for dependent data.

[Table 1 about here.]

[Table 2 about here.]

The results of the simulation study are displayed in Table 1 for Design I and in Table 2 for Design II, where we report the percentage of selected models that match the true generating model (*true*), that contain additional variables (*extra*), that miss variable(s) used to generate the model (*missing*) and finally, that do not fall into one of these categories (*others*). According to the definition in Shao (1993), p. 487, a selected model is considered *good* if it contains the true model generating the data. On the other hand, *incorrect* models are those where at least one variable used to generate the data is missing. We therefore also report a summary of the *good* models.

Let us first comment on the results of Table 1 and Table 2 with reference to the number of *good* models. We see that, based on this measure, in all situations GC_p is the technique performing best, with the two stepwise procedures sometimes comparable (when K is large and data are not contaminated), the z-test never performing as well and the score-test even worse. This should not come as much of a surprise since GC_p is a technique allowing us to compare all possible models. Moreover, both z-test and score-test (as well as the two corresponding stepwise procedures to an extent) suffer from the fact that the estimated coefficients of a fitted model are not independent (even when the x's are independent).

¿From the robustness point of view, we see that, when faced with contaminated data the classical GC_p appears robust by design (at least for Design I),

whereas the "success rate" of all the other procedures drops considerably, illustrating that all of these variable selection procedures are more affected by outlying observations than GC_p. The results further indicate that even though classical GC_p is robust by design, its performance on contaminated data can be improved by using its robust version. On the other hand, the robust versions of z-test, score-test and the two stepwise procedures cannot handle contaminated data any better than their classical counterparts.

The results for Design II show a similar pattern to those of Design I, with the major differences being that the "success rates" are lower (due to the more difficult task of dealing with interaction effects), and the superiority of GC_p even more evident.

Performances (not surprisingly) drop for smaller sample sizes (smaller K), with score-test and z-test and the two stepwise procedures being much more heavily affected. This is likely a consequence of the fact that all of these techniques rely on asymptotic results. GC_p, on the other hand, does not rely on asymptotics, but rather on a Taylor series expansion which apparently has a less detrimental effect. Notice also that if we consider *good* models, GC_p tends to be more liberal in the sense that it chooses more *extra* models.

It is interesting to take a closer look at the distribution of the classical and robust GC_p statistics for *good* models in the contaminated setting, for Design I and $K = 30$ for example. We identify the models with a five-letter sequence of T (=True) and F (=False), according to the inclusion of the corresponding parameters. For instance, the model that generated the data is TFTTF. Then, among the $2^5(= 32)$ possible models there are 3 other *good* models, namely TFTTT, TTTTF and TTTTT.

[Figure 1 about here.]

Figure 1 shows a boxplot of the values of classical and robust GC_p for these 4 models. It appears clearly that not only are the values of the robust GC_p in median smaller than those of its classical counterpart but also that their variability is lower. This confirms that the robust technique is more stable than its classical counterpart and should be preferred in the presence of misspecification of the model.

4. APPLICATION ON REAL DATA

Many healthcare professionals are trained in direct laryngoscopic endotracheal intubation (LEI), a potentially life saving procedure. We examine data from a prospective longitudinal study on LEI at Dalhousie University, previously analyzed in Mills, Field and Dupuis (2002). Variable selection is an important step as the model(s) chosen will include only those covariates significant in predicting successful completion of LEI.

A total of 438 LEI were analyzed. We let $Y_{ij} = 1$ if trainee i performs a complete LEI in less than 30 seconds on trial j, and 0 otherwise. The correlation between observations on the same trainee was taken to be exchangeable, consistent with the findings of the original analysis (Mills et al., 2002). We judge trainees based on the following 9 covariates (our full model): whether the head and neck were in optimal position (NECKFLEX and EXTOA); whether they inserted the scope properly (PROPLGSP); whether they performed the lift successfully (PROPLIFT); whether there was appropriate request for help (ASKAS); whether there was unsolicited intervention by the attending anesthesiologist (HELP); the number of complications (COMPS) and the trainee's handedness (TRHAND) and sex (TRGEND). All covariates are binary with the exception of COMPS which is ordinal. 19 trainees performed anywhere from 18 to 33 trials. A categorical covariate TRIALCAT

was also defined: 1 for trials 1-5, 2 for trials 6 through 10, and so forth.

Our classical GC_p procedure selects the model containing covariates TRIALCAT, PROPLGSP, PROPLIFT, ASKAS, HELP and COMPS whereas both the classical tests and stepwise approaches select a subset of these particular covariates. Such behavior is consistent with that observed in our simulation results where we saw the tests and stepwise approaches favoring (often incorrectly) smaller models. The robust procedures are quite insightful. Robust GC_p as defined by (4) with $\psi = \psi_c(r) = r \min(1, c/|r|)$ being the Huber's function ($c = 1.5$), selects the model that includes two extra covariates (NECKFLEX and EXTOA) when compared to the model selected by the classical GC_p procedure. Robust tests and robust stepwises again select a subset of these particular covariates: the z and score stepwise approaches exclude EXTOA, the robust z-test excludes both EXTOA and COMPS, while the robust score test excludes in addition NECKFLEX and ASKAS, yielding in fact the same model as the classical score stepwise.

To summarize, each robust technique selects a larger (or equal) model than its classical counterpart. In addition, of the five techniques, GC_p tends to select larger models. Moreover, it is for small values of K (here $K = 19$) that we see the major benefits in using the robust version of GC_p; cf. Table 1, $K = 15$, contaminated data.

[Figure 2 about here.]

Our robust GC_p is designed to automatically downweight outlying data thereby reducing their influence on model selection. Figure 2 shows the weights associated with each observation in the GC_p formula (4) for the model selected by the robust GC_p procedure. These weights allow one to identify the outlying observations. There are 9 observations (corresponding to 2%,

identified in Figure 2) that were heavily downweighted by the procedure (weight less than 0.4).

The disagreement between results obtained using robust and classical GC_p suggest that the outlying data have a significant impact on the model chosen. For this reason the model selected by the robust GC_p procedure is to be preferred in this case.

5. CONCLUSIONS

Model selection is an important part of any statistical analysis. In requiring only observations and a model from which predicted values can be obtained, GC_p can be applied to a wide range of statistical problems. Its design also makes it a welcome addition to areas where model selection procedures are mainly based on Wald-type tests, score-type tests or stepwise procedures all of which must be used with caution.

In choosing to focus on longitudinal marginal modeling, much has been learned about the performance of the z-test, score-test and corresponding stepwise procedures, and important comparisons drawn with that of GC_p. GC_p performs as well as the stepwise procedures and much better than both the z-test and score-test, in identifying good models. GC_p goes on to exceed all approaches when faced with contamination as often occurs in practice.

Work in progress includes the application of GC_p to nonparametric techniques (such as GAM) and nonparametric extensions to longitudinal data.

Acknowledgements

The authors thank the Editor, the Associate Editor, two referees and Chris Field for insightful comments. The financial support of the Swiss National Science Foundation (Project # 1214-66989) is also gratefully acknowledged.

References

Breiman, L. (1996). Bagging predictors. *Machine Learning* **24**, 123–140.

Burnham, K. P. and Anderson, D. R. (2002). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer-Verlag Inc, New York.

Cantoni, E. (2004). A robust approach to longitudinal data analysis. *Canadian Journal of Statistics* **32**, 169–180.

Cantoni, E. and Ronchetti, E. (2001). Robust inference for generalized linear models. *Journal of the American Statistical Association* **96**, 1022–1030.

Diggle, P. J., Heagerty, P., Liang, K.-Y. and Zeger, S. L. (2002). *Analysis of Longitudinal Data*. Oxford University Press, New York.

Emrich, L. J. and Piedmonte, M. R. (1991). A method for generating high-dimensional multivariate binary variates. *The American Statistician* **45**, 302–304.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association* **69**, 383–393.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York.

Hauck, W. W. and Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis (Corr: V75 p482). *Journal of the American Statistical Association* **72**, 851–853.

Heritier, S. and Ronchetti, E. (1994). Robust bounded-influence tests in general parametric models. *Journal of the American Statistical Association* **89**, 897–904.

Horton, N. J. and Lipsitz, S. R. (1999). Review of software to fit generalized

estimating equation regression models. *The American Statistician* **53**, 160–169.

Huber, P. J. (1981). *Robust Statistics*. Wiley, New York.

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

Mallows, C. L. (1973). Some comments on C_p. *Technometrics* **15**, 661–675.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall, London, second edition.

McQuarrie, A. D. R. and Tsai, C.-L. (1998). *Regression and time series model selection*. World Scientific Publications, Singapore.

Mills, J. E., Field, C. A. and Dupuis, D. J. (2002). Marginally specified generalized linear mixed models: A robust approach. *Biometrics* **58**, 727–734.

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

Preisser, J. S. and Qaqish, B. F. (1999). Robust regression for clustered data with applications to binary regression. *Biometrics* **55**, 574–579.

Ronchetti, E. and Staudte, R. G. (1994). A robust version of Mallows' C_p. *Journal of the American Statistical Association* **89**, 550–559.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association* **88**, 486–494.

Appendix

A. Derivation of the correction term in the formula for GC_p

We derive the correction term $-\sum_{i=1}^{k} E\{w_i^2(r_i)\epsilon_i^2\} + 2\sum_{i=1}^{K} E\{w_i^2(r_i)\epsilon_i\delta_i\}$. Consider first the Taylor expansion of the weights $w_i(r_i)$ around $w_i(\epsilon_i)$:

$$w_i(r_i) = w_i(\epsilon_i) + (r_i - \epsilon_i)w'(\epsilon_i) + \frac{1}{2}(r_i - \epsilon_i)^2 w''(\epsilon_i) + \frac{1}{6}(r_i - \epsilon_i)^3 w'''(\epsilon_i^*)$$

for some ϵ_i^* lying between r_i and ϵ_i. It is assumed that the weight function w_i is even and three times differentiable. The above expansion leads us to an approximate expression for $w_i^2(r_i)$ from which we obtain equation (2).

B. Derivation and computation of t_1 and t_2

In the setting of Section 2.2, (2) becomes

$$\begin{aligned} GC_p &= \sum_{i=1}^{K}\sum_{t=1}^{n_i} w_{it}^2(r_{it})r_{it}^2 - \sum_{i=1}^{K}\sum_{t=1}^{n_i} E\{w_{it}^2(\epsilon_{it})\epsilon_{it}^2\} \\ &+ 2\sum_{i=1}^{K}\sum_{t=1}^{n_i} E\{\big(w_{it}(\epsilon_{it})w_{it}'(\epsilon_{it})\epsilon_{it}^2 + w_{it}^2(\epsilon_{it})\epsilon_{it}\big)\delta_{it}\} \qquad (9) \\ &- \sum_{i=1}^{K}\sum_{t=1}^{n_i} E[\big\{w_{it}(\epsilon_{it})w_{it}''(\epsilon_{it})\epsilon_{it}^2 + \{w_{it}'(\epsilon_{it})\}^2\epsilon_{it}^2 + 4w_{it}(\epsilon_{it})w_{it}'(\epsilon_{it})\epsilon_{it}\big\}\delta_{it}^2] \end{aligned}$$

where $r_{it} = (y_{it} - \hat{y}_{it})/(\sigma v_{it}^{1/2})$, $\epsilon_{it} = (y_{it} - Ey_{it})/(\sigma v_{it}^{1/2})$ and $\delta_{it} = (\hat{y}_{it} - Ey_{it})/(\sigma v_{it}^{1/2})$. Notice that Ey_{it} and $\sigma^2 v_{it}$ are the expected value and variance under the full model. By defining $\psi(\epsilon_{it}) = w(\epsilon_{it}) \cdot \epsilon_{it}$, we can rewrite (9) as

$$GC_p = \sum_{i=1}^{K}\sum_{t=1}^{n_i} w_{it}^2(r_{it})r_{it}^2 - \sum_{i=1}^{K}\sum_{t=1}^{n_i} E\{\psi^2(\epsilon_{it})\} + t_1 - t_2,$$

where

$$t_1 = 2\sum_{i=1}^{K}\sum_{t=1}^{n_i} E\{\psi(\epsilon_{it})\psi'(\epsilon_{it})\delta_{it}\} \qquad (10)$$

and

$$t_2 = \sum_{i=1}^{K}\sum_{t=1}^{n_i} E\Big[\Big\{\psi(\epsilon_{it})\psi''(\epsilon_{it}) - \frac{\psi^2(\epsilon_{it})}{\epsilon_{it}^2} + (\psi'(\epsilon_{it}))^2\Big\}\delta_{it}^2\Big]. \qquad (11)$$

At this point we use the structure of the marginal model form in order to compute the expectations in (10) and (11). Let $\delta_{it} = (\hat{y}_{it} - Ey_{it})/(\sigma v_{it}^{1/2}) = \{g^{-1}(x_{it}^T\hat{\beta}) - g^{-1}(x_{it}^T\beta)\}/(\sigma v_{it}^{1/2})$. A Taylor expansion of $g^{-1}(x_{it}^T\hat{\beta})$ about $g^{-1}(x_{it}^T\beta)$ results in the following approximation: $\delta_{it} \cong \frac{1}{\sigma v_{it}^{1/2}} \frac{\partial g^{-1}(\theta)}{\partial\theta}\Big|_{\theta=x_{it}^T\beta} x_{it}^T(\hat{\beta}-$

β). The influence function (Hampel, 1974, Hampel et al., 1986) of the estimator defined by (3) is given by $IF\{(X,y);\hat{\beta},F_\beta\} = M^{-1}D^T\Gamma^T V^{-1}(\Psi - c)$ for a generic observation (X, y). We therefore have for $K \to \infty$:

$$\hat{\beta} - \beta \approxeq \frac{1}{K}\sum_{j=1}^{K} IF\{(X_j,y_j);\hat{\beta},F_\beta\} = \frac{1}{K}\sum_{j=1}^{K} M^{-1}D_j^T\Gamma_j^T V_j^{-1}(\Psi_j - c_j),$$

which implies

$$\delta_i = \frac{1}{\sigma}A_i^{-1}D_i M^{-1}\frac{1}{K}\sum_{j=1}^{K} D_j^T\Gamma_j^T V_j^{-1}(\Psi_j - c_j) \tag{12}$$

for $\delta_i = (\delta_{i1}\dots\delta_{in_i})^T$. We can now express t_1 of equation (10) as $t_1 \approxeq 2\sum_{i=1}^{K} E\{Tr(\delta_i a_i^T)\}$. Substituting our expression for δ_i from equation (12) into the above and recognizing that all $(j \neq i)$ terms are 0 by independence, we obtain the following expression for t_1:

$$t_1 \approxeq \frac{2}{\sigma K}\sum_{i=1}^{K} E(Tr[A_i^{-1}D_i M^{-1}D_i^T\Gamma_i^T V_i^{-1}\{W_i(y_i - \mu_i) - c_i\}a_i^T]),$$

and by using the properties of expectation and trace

$$t_1 \approxeq \frac{2}{\sigma K}\sum_{i=1}^{K} Tr\{M^{-1}E(D_i^T Z_i a_i^T A_i^{-1}D_i)\}, \tag{13}$$

where $Z_i = \Gamma_i^T V_i^{-1}\{W_i(y_i - \mu_i) - c_i\}$ $a_i = (a_{i1}\dots a_{in_i})^T$ and $a_{it} = \psi(\epsilon_{it})\psi'(\epsilon_{it})$.

Using the same arguments we can express t_2 in equation (11) as $t_2 \approxeq \sum_{i=1}^{K} E\{Tr(B_i\delta_i\delta_i^T)\}$, where $B_i = \mathrm{diag}(b_{i1}\dots b_{in_i})$ and b_{it} is defined in Section 2.2. Substituting our expression for δ_i from equation (12) into that above and removing those terms whose expectation is zero (by independence), we obtain the following expression for t_2:

$$t_2 \approx \frac{1}{\sigma^2 K^2}\sum_{i=1}^{K} Tr\Big[E\Big\{B_i A_i^{-1}D_i M^{-1}\big(\sum_{j=1}^{K} D_j Z_j Z_j^T D_j\big)M^{-1}D_i^T A_i^{-1}\Big\}\Big]. \tag{14}$$

Figure 1. Distribution of GC_p for the *good* models.

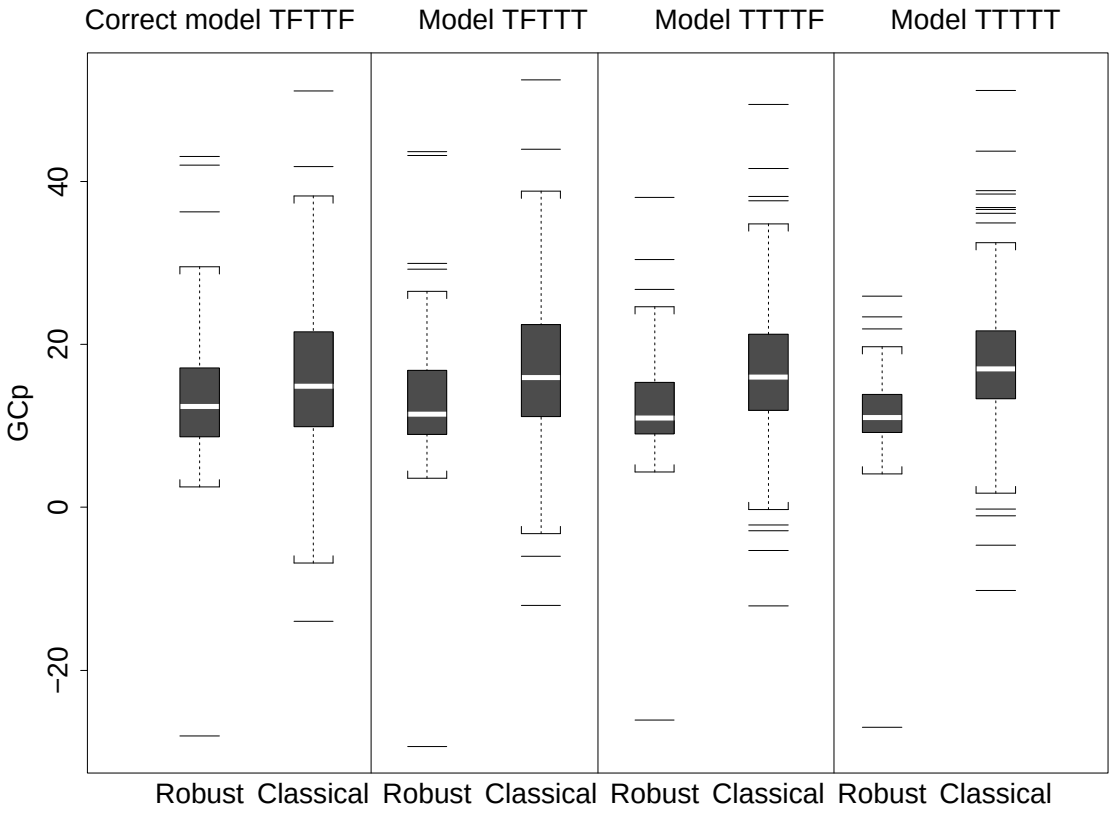

Figure 2. Observation weights in (4) for the model selected by robust GC_p.

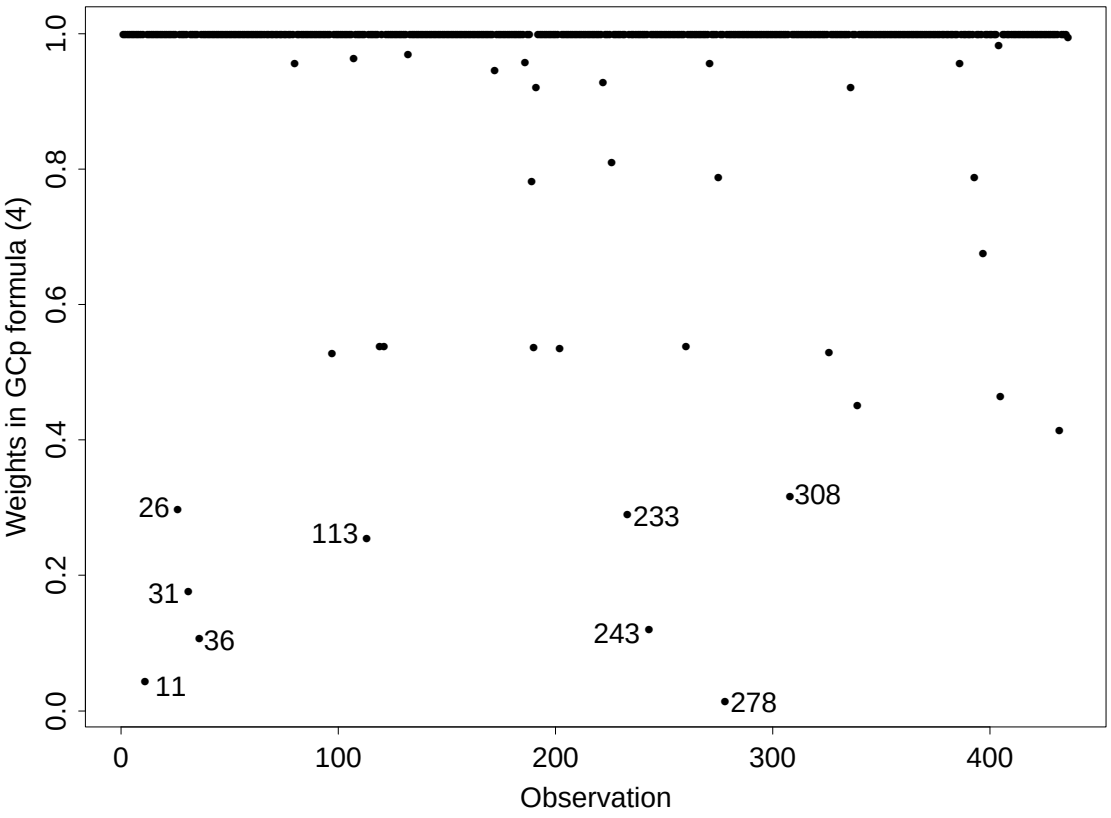

Table 1

Simulations results for design I. Percentage of models in each category.

		non contaminated									
	model	classical					robust				
		GC_p	z	z-step.	score	sc.-step.	GC_p	z	z-step.	score	sc.-step.
$K = 30$	**good**	**90**	**75**	**87**	**73**	**88**	**89**	**75**	**84**	**69**	**84**
$K = 30$	extra	29	14	22	10	22	56	11	22	8	21
$K = 30$	true	61	61	65	63	66	33	64	62	61	63
$K = 30$	missing	4	22	9	24	9	2	18	10	28	11
$K = 30$	others	6	3	4	3	3	9	7	6	3	5
$K = 15$	**good**	**62**	**29**	**48**	**15**	**33**	**61**	**30**	**43**	**10**	**32**
$K = 15$	extra	27	7	11	0	9	42	6	10	0	3
$K = 15$	true	35	22	37	15	24	19	24	33	10	29
$K = 15$	missing	23	55	36	74	51	11	50	38	66	46
$K = 15$	others	15	16	16	11	16	28	20	19	24	22

		contaminated									
	model	classical					robust				
		GC_p	z	z-step.	score	sc.-step.	GC_p	z	z-step.	score	sc.-step.
$K = 30$	**good**	**82**	**60**	**73**	**54**	**67**	**91**	**63**	**72**	**56**	**69**
$K = 30$	extra	29	8	19	5	13	67	8	16	6	11
$K = 30$	true	53	52	54	49	54	24	55	56	50	58
$K = 30$	missing	13	34	20	40	24	2	33	22	38	22
$K = 30$	others	5	6	7	6	9	7	4	6	6	9
$K = 15$	**good**	**42**	**17**	**28**	**6**	**19**	**63**	**14**	**25**	**5**	**16**
$K = 15$	extra	21	5	7	1	4	48	4	10	0	3
$K = 15$	true	21	12	21	5	15	15	10	15	5	13
$K = 15$	missing	32	56	46	61	60	8	57	44	59	58
$K = 15$	others	26	27	26	33	21	29	29	31	36	26

Table 2

Simulations results for design II. Percentage of models in each category.

		non contaminated									
	model	classical					robust				
		GC_p	z	z-step.	score	sc.-step.	GC_p	z	z-step.	score	sc.-step.
$K = 30$	**good**	**42**	**13**	**29**	**5**	**11**	**75**	**18**	**31**	**5**	**12**
$K = 30$	extra	18	2	9	1	5	54	3	12	1	6
$K = 30$	true	24	11	20	4	6	21	15	19	4	6
$K = 30$	missing	33	69	52	59	48	5	64	49	56	48
$K = 30$	others	25	18	19	36	41	20	18	20	39	40
$K = 15$	**good**	**21**	**8**	**15**	**1**	**7**	**43**	**3**	**9**	**1**	**3**
$K = 15$	extra	5	2	4	1	2	30	2	6	0	2
$K = 15$	true	16	6	11	0	5	13	1	3	1	1
$K = 15$	missing	38	64	58	41	43	8	62	67	36	55
$K = 15$	others	41	28	27	55	50	49	35	24	63	42

		contaminated									
	model	classical					robust				
		GC_p	z	z-step.	score	sc.-step.	GC_p	z	z-step.	score	sc.-step.
$K = 30$	**good**	**22**	**9**	**17**	**3**	**10**	**67**	**14**	**21**	**4**	**8**
$K = 30$	extra	5	2	5	0	2	53	2	6	1	2
$K = 30$	true	17	7	12	3	8	14	12	15	3	6
$K = 30$	missing	45	71	56	65	52	7	64	58	63	55
$K = 30$	others	33	20	27	32	38	26	22	21	33	37
$K = 15$	**good**	**7**	**5**	**8**	**0**	**3**	**31**	**3**	**5**	**0**	**3**
$K = 15$	extra	4	2	5	0	1	22	2	4	0	2
$K = 15$	true	3	3	3	0	2	9	1	1	0	1
$K = 15$	missing	55	58	73	32	49	18	56	72	28	53
$K = 15$	others	38	37	19	68	48	51	41	23	72	44