Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: http://www.tandfonline.com/loi/lssp20

A comparison between bootstrap methods and generalized estimating equations for correlated outcomes in generalized linear models

Michael Sherman & Saskia le Cessie

To cite this article: Michael Sherman & Saskia le Cessie (1997) A comparison between bootstrap methods and generalized estimating equations for correlated outcomes in generalized linear models, Communications in Statistics - Simulation and Computation, 26:3, 901-925, DOI: 10.1080/03610919708813417

To link to this article: http://dx.doi.org/10.1080/03610919708813417

Published online: 27 Jun 2007.

COMMUN. STATIST.—SIMULA., 26(3), 901-925 (1997)

A COMPARISON BETWEEN BOOTSTRAP METHODS AND GENERALIZED ESTIMATING EQUATIONS FOR CORRELATED OUTCOMES IN GENERALIZED LINEAR MODELS

Michael Sherman and Saskia le Cessie

Department of Statistics
Texas A&M University
College Station, Texas 77843
and
Department of Medical Statistics
Leiden University
2300 RC Leiden, The Netherlands

Key Words: correlated data; logistic regression; resampling.

ABSTRACT

We discuss and evaluate bootstrap algorithms for obtaining confidence intervals for parameters in Generalized Linear Models when the data are correlated. The methods are based on a stratified bootstrap and are suited to correlation occurring within "blocks" of data (e.g., individuals within a family, teeth within a mouth, etc.). Application of the intervals to data from a Dutch follow-up study on preterm infants shows the corroborative usefulness of the intervals, while the intervals are seen to be a powerful diagnostic in studying annual measles data. In a simulation study, we compare the coverage rates of the proposed intervals with existing methods (e.g., via Generalized Estimating Equations). In most cases, the bootstrap intervals are seen to perform better than current methods, and are produced in an automatic fashion, so that the user need not know (or have to guess) the dependence structure within a block.

1. INTRODUCTION

In recent years, the generalized linear model (GLM) has become an important and useful generalization of the classical linear model (see, e.g., McCullagh and Nelder (1989)). Logistic regression, log-linear models, and Poisson regression are all examples of a GLM. An important assumption in all GLM's is that of independent observations. If observations are dependent then inferences drawn from parameter estimates and estimated standard errors may not be valid.

The GLM has the general form $E(Y_i) = g^{-1}(x_i'\beta)$. Here $(Y_i, x_i), i = 1, ..., n$, denote a sequence of observations, where Y_i is the response variable and x_i is a vector of fixed covariates believed to have an effect on Y_i. β is an unknown parameter vector and $g(\cdot)$ is the link function, connecting the mean of Y_i, $E(Y_i)$, with $x_i'\beta$. The variance of Y_i is assumed to be $\phi h(E(Y_i))$ for some function $h(\cdot)$ and constant ϕ.

To fix ideas consider the case of logistic regression. Models with binary responses (Y_i =1 or 0) commonly occur in many areas of application. A common model for relating Y_i to $x_i'\beta$ is the logistic for which

$$p_i = P(Y_i = 1|x_i) = \exp(x_i{}'\beta)/(1 + \exp(x_i{}'\beta)).$$

In this case $g(u) = log(u/(1-u))$, $\phi = 1$, and $h(u) = u(1-u)$.

For a given model, inferences about β are based on hypothesis tests and confidence intervals. In the case where the data are independent, the likelihood of the data reduces to the usual product likelihood, and parameter estimates, $\hat{\beta}$, are (usually) obtained by maximum likelihood. Typically, estimated standard errors are obtained from $I(\hat{\beta})$, the observed Fisher information, and confidence intervals are formed by assuming that $\hat{\beta}$ is approximately normally distributed.

If the data are correlated, however, this procedure no longer makes sense because the product likelihood is not correct. For example, if the data consist of families then it is unreasonable to assume that responses within a family are independent.

Parameter estimates will still be consistent (under mild conditions), but estimated standard errors will be incorrect. For example, if the correlation between family members is positive (and the covariates are constant within families) then standard errors derived under the assumption of independence will be too small. There are existing methods adressing this problem, we discuss them in Sections 3 and 4 and examine their performance in the examples and simulations in Sections 5 and 6.

We propose a bootstrap procedure to obtain confidence intervals for β in the presence of dependence within a family. This is a variant of other "whole block" bootstrap algorithms discussed further in Section 3. The basic idea is to resample families instead of individuals. In this fashion, the generated replicates automatically retain the same dependence structure as the original data. The procedure will not only account for dependence but in many cases it also yields confidence intervals which are more accurate than the standard "z-intervals" based on asymptotic normality. Further, the proposed procedure will allow the user to obtain these intervals (that account for dependence within a family) without having to specify the dependence structure (e.g. are correlations between spouses the same as between one parent and a child?). The only requirement is that the mean function is assumed to have the correct form. Because the dependency is unspecified, however, the proposed procedure will not allow the user to determine the structure or measure the strength of the underlying dependence. In the case where the main interest is on parameters in the mean function, however, our proposal is a robust, reliable, general tool for drawing inferences on the parameters in the prescence of nuisance block dependence.

Section 2 discusses the effects of dependence and describes the bootstrap algorithm for generating valid replicates. Also construction of the bootstrap confidence intervals is discussed. Section 3 discusses related methods, while Section 4 describes the method of Generalized Estimating Equations (GEE) for correlated outcomes. In Section 5 we compare different confidence intervals in data from a follow up study on preterm infants and annual measles data. Section 6 compares the coverage accuracy of different confidence intervals in several simulation studies.

2. BOOTSTRAP ALGORITHMS FOR DEPENDENT DATA IN GLM

To gain some insight into the error incurred by not accounting for dependence, consider the problem of estimating the variance of the sample mean. Assume that there are m families, each of size k. On each individual an observation, Y_i, is made and assume that the common correlation between members within a family is ρ (the same for each family). Then the true variance of the (standardized) sample mean is

$$nVar(\overline{Y}_n) = \sigma^2(1 + \rho(k-1)),$$

where $n = km$. If we use as an estimator for $nVar(\overline{Y}_n)$ the usual sample variance, s_n^2, (i.e., assume independence) then $E(s_n^2) = \sigma^2(1 - (\rho(k-1)/(mk-1)))$. For large correlation, ρ, or for large family size, k, s_n^2 can drastically underestimate the correct variance. This phenomenon has been noted by, e.g., Diggle, Liang, and Zeger (1994).

The proposed bootstrap idea is actually quite simple. Let $(Y_k, x_k)_l$ denote the block of observations $(Y_{k+1}, x_{k+1}), ..., (Y_{k+l}, x_{k+l})$, so that in particular $(Y_0, x_0)_n$ denotes the entire data set. Assume that we have m blocks of observations, each of size k, $(Y_{ik}, x_{ik})_k, i = 0, 1, ..., (m-1)$, so that $mk = n$ (the index "ik" denotes the actual product of i and k). We allow the dependence structure within a block to be arbitrary, and it is allowed to differ from block to block, but observations in different blocks are assumed to be independent. Find the estimate $\hat{\beta}$ by maximizing the likelihood under the assumption of independence. Let $\hat{F}_m$ place mass $1/m$ on each $(Y_{ik}, x_{ik})_k, i = 0, 1, ..., (m-1)$. Resample m times from $\hat{F}_m$, i.e., from the m blocks with replacement, to obtain a bootstrap resample, $(Y_{ik}, x_{ik})_k^*, i = 0, 1, ..., (m-1)$, and obtain the corresponding $\hat{\beta}^*$. Note that this procedure, by resampling blocks of data, retains the correct dependence structure within a block without having to specify the dependence structure. If we repeat this procedure B times, we obtain $\hat{\beta}_1^*, \ldots, \hat{\beta}_B^*$.

If as often may be the case, there are blocks of different sizes, then the algorithm can be modified as follows: let m_i denote the number of blocks of size i, $i = 1, \ldots, I$. For each i resample m_i times with replacement from the m_i blocks and compute $\hat{\beta}^*$ from the $n = \sum_{i=1}^{I} im_i$ resampled observations. This conditioning on block size guarantees a total resample size equal to the original sample size, making the bootstrap replicates "comparable". If, however, I is large it may be more attractive to resample m times from the entire set of blocks. We will call this the "All Block" bootstrap. This algorithm gives a random total sample size, n^*, say, which makes the replicates less comparable. A reasonable approach to make them more comparable is to let the replicate be $(n^*/n)^{1/2}\hat{\beta}^*$, as suggested by Efron and Tibshirani (1993, p.101) for a whole block bootstrap in the time series setting.

Various procedures have been proposed for forming confidence intervals from the $\hat{\beta}^*$'s. Efron and Tibshirani (1993), Chapter 14, describe percentile (P), bias corrected percentiles (BC), and accelerated bias corrected ($\mathrm{BC_a}$) intervals. The P intervals are the easiest to compute but the least accurate, while the ($\mathrm{BC_a}$) intervals require slightly more work but are the most accurate. The difficulties are not daunting, however, requiring only a few more lines of computer code. The "acceleration constant" is estimated using the jackknife estimate of skewness as described in Efron and Tibshirani. We also consider the "bootstrap-t" intervals using the robust GEE estimates of standard error (see Section 4). This is to avoid the second level bootstrapping a nonparametric estimate of standard error would require.

We note that S-Plus routines are available to compute bootstrap confidence intervals. This code can be retrieved as described in Efron and Tibshirani (1993, Appendix).

3. COMPARISON WITH OTHER METHODS

For independent data, bootstrapping regression models was discussed in Freedman (1981) where for the classical linear model it was shown that the standardized distribution of the $\hat{\beta}^*$'s has the correct limiting normal distribution. The bootstrap has been used for dependent data by Freedman (1984) for "linear dynamical models", and by Bose (1988) for autoregression. The main drawback to these methods, as Freedman (1984) emphasizes, is that they assume that the true dependence structure is known.

Model-free bootstrap methods for time series data have been proposed by, e.g., Künsch (1989). These methods are based on resampling blocks of data, similar to the algorithms proposed here. Although, in the time series setup the "moving blocks bootstrap" only makes a partial correction for any fixed block length, while in our setting choosing the block size to be the same as the family size makes the "full" correction. Moulton and Zeger (1989) proposed an alternative bootstrap algorithm for inference in the GLM with longitudinal data, where subjects are measured at the same time points $t, t = 1, ..., T$. Their method generates replicates by resampling the vectors of observations belonging to an individual and obtaining $\hat{\beta}_t^*$ via a "one-step" bootstrap estimation technique at each time $t, t = 1, ..., T$. The $\hat{\beta}_t^*$ can be used to obtain time specific information or can be combined in a suitable way to obtain summary statistics.

Using a parametric likelihood approach is the most efficient way to deal with correlated data, assuming that the likelihood is specified correctly. This paper focusses on situations where the dependence structure between observations is not known and on methods which can be applied without requiring second order moment assumptions. Therefore we will not go into detail about fully parametric models for correlated data. For small blocks, e.g., of size 2 or 3, in the logistic regression model, the full likelihood can be maximized. For larger block sizes, however, maximum likelihood becomes difficult to implement because it is difficult to specify the dependence structure. There are several different ways to model the dependence within a block of size 2 (usual correlation, odds ratio, tetrachoric correlation), although they all give approximately the same parameter estimates. See, e.g., Connolly and Liang (1988), Prentice (1988), Zeger, Liang and Albert (1988), Lipsitz, Laird, and Harrington (1990), Liang, Zeger, and Qaqish (1992), le Cessie and van Houwelingen (1994) and Diggle, Liang, and Zeger (1994) for approaches to logistic regression with correlated outcomes. Thall and Vail (1989) and Diggle, Liang, and Zeger (1994) discuss models for correlated count data.

4. GENERALIZED ESTIMATING EQUATIONS

Another model-based method for analyzing data of the type considered here is through Generalized Estimation Equations, or GEE (Liang and Zeger (1986)). Their method also accounts for the dependency within a block of data, without fully specifying the dependence structure. Instead of using the true likelihood function, the so-called generalized estimation equations are solved. These equations have the form:

$$\sum_{i=1}^{m} \frac{\partial \mu_i(\beta)'}{\partial \beta} V_i^{-1}(Y_i - \mu_i(\beta)) = 0,$$

where Y_i is the vector of responses in block i and $\mu_i(\beta) = E(Y_i)$, and where V_i is some approximation to $var(Y_i)$. Common choices for V_i are the independence structure where the correlation matrix is the identity matrix, and the exchangable structure, corresponding to equal correlations between all observations in a block. Even if the covariance structure is misspecified, the GEE approach yields consistent estimates of β under mild regularity conditions. Notice that if the independence structure is used, the GEE estimates are the same as the estimates obtained when maximizing the likelihood assuming independence. These are the estimates used in the bootstrap approach of Section 2 and this approach can therefore be regarded as a bootstrap analog of GEE with an independence working correlation. For further relationships between bootstrap methods and GEE, see, e.g., Moulton and Zeger (1989).

One drawback to the GEE method, however is that the user has to specify a correlation structure in order to perform the estimation. The naive estimate of $var(\hat{\beta})$

is H_1^{-1} with $H_1 = \sum_{i=1}^{m} \frac{\partial \mu_i(\hat{\beta})}{\partial \beta}' V_i^{-1} \frac{\partial \mu_i(\hat{\beta})}{\partial \beta}$. Even if V_i is chosen correctly, confidence intervals from this method will still be based on the assumption of approximate normality. There is a robust estimate of $var(\hat{\beta})$, the so called robust or sandwich estimator (Royall (1986)), that uses an empirical estimate $V_{0i} = (Y_i - \mu_i(\hat{\beta}))'(Y_i - \mu_i(\hat{\beta}))$ of $var(Y_i)$. This estimator has the form

$$H_1^{-1} H_2 H_1^{-1}$$

with $H_2 = \sum_{i=1}^{m} \frac{\partial \mu_i(\hat{\beta})}{\partial \beta}' V_i^{-1} V_{0i} V_i^{-1} \frac{\partial \mu_i(\hat{\beta})}{\partial \beta}$. However, this estimator is usually downwards biased and the bias can be substantial if the sample size is small or when the covariates are skewed.

To illustrate this bias problem, we look at a simple situation, the linear regression model with independent errors.

$$Y_i = \beta_0 + \beta_1 x_i + e_i$$

with $e_i \sim N(0,1)$ and independent. In matrix notation: $Y = X\beta + e$. The GEE estimator of β (under the assumption of independent blocks of size 1) is the usual least squares estimator:

$$\hat{\beta} = (X'X)^{-1} X'Y.$$

Then the true variance of $\hat{\beta}$ is

$$var(\hat{\beta}) = (X'X)^{-1} X' cov(Y) X (X'X)^{-1} = (X'X)^{-1}.$$

In the robust variance estimator cov (Y) is estimated using the empirical covariances. For the linear regression model the robust variance estimator is

$$var_{rob}(\hat{\beta}) = (X'X)^{-1} X' \begin{bmatrix} (y_1 - \hat{y}_1)^2 & & \emptyset \\ & \ddots & \\ \emptyset & & (y_n - \hat{y}_n)^2 \end{bmatrix} X(X'X)^{-1}.$$

The expectation of this expression is

$$E[var_{rob}(\hat{\beta})] = (X'X)^{-1} X' \begin{bmatrix} (1 - h_1) & & \emptyset \\ & \ddots & \\ \emptyset & & (1 - h_n) \end{bmatrix} X(X'X)^{-1},$$

with h_i the i-th diagonal element of the hat matrix $H = X(X'X)^{-1}X'$ And this expression equals

$$E[var_{rob}(\hat{\beta})] = (X'X)^{-1} - (X'X)^{-1} \begin{bmatrix} \sum h_i & \sum x_i h_i \\ \sum x_i h_i & \sum x_i^2 h_i \end{bmatrix} (X'X)^{-1}.$$

We see that the expected value of the robust variance estimator of $\hat{\beta}$ is too small. The deviation depends on $\sum h_i, \sum x_i h_i$, and $\sum x_i^2 h_i$. The first term equals 2, the number of parameters in the model, the latter two terms will be large when there are outliers in the x-space.

In general, estimating the covariance of Y using the empirical covariances typically results in a downwards bias. That this bias can be substantial can be seen, e.g., in the first simulation experiment in Table 3.

5. EXAMPLES

We present two examples and compare "standard" and GEE confidence intervals with bootstrap intervals in each. The first is an example using logistic regression where most of the intervals are in reasonable agreement. The second uses a more complicated model and we can see that the bootstrap algorithm not only can obtain informative confidence intervals, but can be used as a powerful diagnostic tool.

5.1 The POPS Data

The data for this example come from a Dutch follow-up study on preterm infants (Verloove and Verwey (1988)). Data were collected on 1338 infants, born in the Netherlands in 1983 with a gestational age of less than 32 completed weeks and/or a birth weight of less than 1500 grams. A problem occurring when modelling the different binary outcome variables, like mortality after 28 days, of these data with logistic regression, is that the assumption of independent observations does not hold. A large subset of the infants occur from multiple births and these infants could well respond more similar to each other than to other infants. We analyze the outcomes for children coming from multiple births. Specifically, there are 107 complete sets of twins, and 62 twins whose twin sibling is not included in the study, 5 sets of complete triplets, 6 pairs of infants coming from triplets, and 7 triplets with neither of the two remaining triplets in the study, and one infant coming from a quadruplet. Thus, there are 70 blocks of size 1, 113 blocks of size 2, and 5 blocks of size 3.

In this example, gestational age and birth weight are used as covariates, and the outcome Y is 1 if an infant has died within 28 days after birth. Thus the model is $p_i = P[Y_i = 1|x_i] = \exp(\beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2)/(1 + \exp(\beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2))$, where x_{i1} and x_{i2} denote the gestational age and birth weight, respectively, for the i'th observation. The gestational age is the same for all children from one multiple birth; the birth weight can differ. These data have been analyzed previously by le Cessie and van Houwelingen (1994).

Table 1 shows ten 90% confidence intervals for β_1 and β_2: the percentile and BC intervals (the BC$_a$ intervals were identical to the BC intervals because the acceleration constant, a, was close to 0 for both covariates) based on the algorithm in Section 2 with resampling conditional on the block size, the interval based on ordinary maximum likelihood (ML) ignoring dependence, three GEE intervals, one based on a working independence assumption, and two based on exchangeable correlation (one naive, one robust) and asymptotic normality, the bootstrap-t interval, using the GEE estimate and robust standard error in bootstrap resamples, and percentile, BC, and bootstrap-t intervals based on the All Block bootstrap. The bootstrap intervals are based on B=1000 replicates. The GEE estimates were computed using the SAS/IML software macro for Longitudinal Data Analysis (Karim and Zeger (1988)).

The intervals based on ordinary ML are narrow relative to the others. This is due to estimated standard errors which are probably too small, due to the positive correlation among twins. The bootstrap percentile intervals seem to be shifted to the left. This is due to the fact that for positive (negative) parameters the ML estimate tends to be biased up (down) (see, e.g. McCullagh and Nelder (1989)) and

TABLE I

Table 1: POPS data. Estimated standard errors, 90 % confidence intervals and length of these intervals for the two covariates gestational age and birth weight.

	Gestational age			Birth weight		
Method	se	90 % CI	width	se	90 % CI	width
Ordinary Logistic Reg.	.102	(-.724,-.388)	.34	.068	(-.197,.028)	.23
Bootstrap, Percentile	.122	(-.779,-.390)	.39	.081	(-.212,.052)	.26
Bias Corrected (BC)		(-.745,-.354)	.39		(-.219,.049)	.27
Bootstrap-t		(-.742,-.368)	.37		(-.215,.039)	.25
All Block, Percentile	.119	(-.796,-.399)	.40	.081	(-.209,.057)	.27
Bias Corrected (BC)		(-.762,-.372)	.39		(-.215,.046)	.26
Bootstrap-t		(-.742,-.360)	.38		(-.218,.040)	.26
GEE-robust, ind.	.115	(-.745,-.367)	.38	.077	(-.211,.043)	.25
GEE-naive-exc.	.108	(-.723,-.367)	.36	.071	(-.202,.032)	.23
GEE-robust-exc.	.117	(-.738,-.353)	.38	.077	(-.212,.042)	.25

the percentile intervals do not account for this bias. The 5th and 95th percentile of the bootstrap-t statistic were -1.67 and 1.64 for gestational age and -1.61 and 1.71 for birth weight, suggesting that skewness of the estimate is not too big a problem in this example.

The robust GEE intervals and the BC intervals basically agree, which is reassuring. The main conclusion then is that the BC intervals corroborate the GEE intervals derived from modeling the dependence, giving more strength to the conclusions.

5.2 The Measles Data

Annual measles data were collected for each of 15 counties in the United States between 1985 and 1991. For each county the annual number of preschoolers with measles was recorded as well as factors possibly related to measles incidence, like immunization rate, density of preschoolers per county, etc. The data given here are preliminary.

In this example, we study the relationship between the mean annual measles incidence for preschoolers (i.e., total cases of preschool measles divided by the total number of preschoolers in a particular year) and the immunization rate in two year old children. The immunization rate was measured in 1986 and is assumed to be constant during the period 1985-1991. The data are listed in the appendix. There seems to be a strong (negative) relationship between the immunization rate and the incidence of measles, but there is also wide variability in measles incidence from year to year. For example, in county 6, the county with the lowest immunization rate, there are four years with a large numbers of preschool measles, but also three years with a negligible number of cases.

A formal framework for modelling the data is as follows. Let X_i be the immunization rate for county i, $i = 1, \dots, 15$. Let Y_{ij} be the number of preschool measles in county i in year j, $j = 1, \dots, 7$, and let n_{ij} be the total number of preschool children in county i, year j. The incidence in county i in year j is then $I_{ij} = Y_{ij}/n_{ij}$. The mean incidence $E[I_{ij}] = \lambda_{ij}$ depends on the immunization rate. Since λ_{ij} is greater or equal to 0, we model $\log(\lambda_{ij})$ on a linear scale:

$$\log(\lambda_{ij}) = \beta_0 + \beta_1 X_i. \tag{1}$$

In these data dependence arises from the potential serial correlation between years within a county. The number of measles cases varies wildly from year to year so that the dependence is not easily modeled, for example, by an autoregressive process. For this reason, we use the bootstrap for dependent data. In this case the bootstrap resamples from the 15 counties. For each bootstrap resample, the parameters were estimated by maximizing the independent likelihood using generalized linear model methodology with Y_{ij} as a Poisson response variable and $\log(n_{ij})$ as the "offset". The Percentile, BC, and BC$_a$ intervals are all based on B=1000 bootstrap replicates.

We compared the bootstrap approach with the ML estimates, assuming that the observations are independent Poisson variables and with two GEE approaches, assuming model (1) for the mean incidence. In the first GEE approach, the ML estimate is used and the robust variance is computed, using an independent working

TABLE II

Table 2: Measles data. Estimated standard errors and 90 % confidence intervals for immunization rate.

Method	se	90 % CI
Bootstrap percentile	.043	(-.214, -.090)
Bias Corrected		(-.188, -.086)
BC accelerated		(-.182, -.080)
Bootstrap-t		(-.135, -.045)
All Block, Percentile		(-.245, -.052)
ML (independent)	.002	(-.111, -.105)
GEE robust-ind.	.016	(-.139, -.077)
GEE naive-exc.	.022	(-.144, -.072)
GEE robust-exc.	.016	(-.134, -.082)

correlation. We also used the GEE approach of Thall and Vail (1990). They proposed a Poisson model with overdispersion for longitudinal count data. In this model it is assumed that there are random county effects which act multiplicatively on the mean of Y_i (we calculate both naive and robust intervals under this exchangeable correlation structure). For a related approach, see Breslow and Clayton (1993).

Table 2 shows the results. The estimated slope (for the independence ML and GEE setups) is $\hat{\beta}_1 = -.108$.

We see in this example that the confidence intervals vary widely across the different methods. Which can we trust? Due to the dependence within a country from year to year, the ordinary ML results seem unreliable. Note that the percentile intervals are the most skewed left interval while the bootstrap-t is the most skewed right. This is a common phenomenon, although the extent to which it happens here is cause for concern.

To gain a better understanding of why the GEE and bootstrap intervals differ consider the empirical distribution of the 1000 bootstrap replicates $\hat{\beta}_1^*$ pictured in Figure 1. The pronounced bimodal distribution is somewhat suspicious. Notice that about 1/3 of the total mass is in the left mode. In this example there are 15 counties

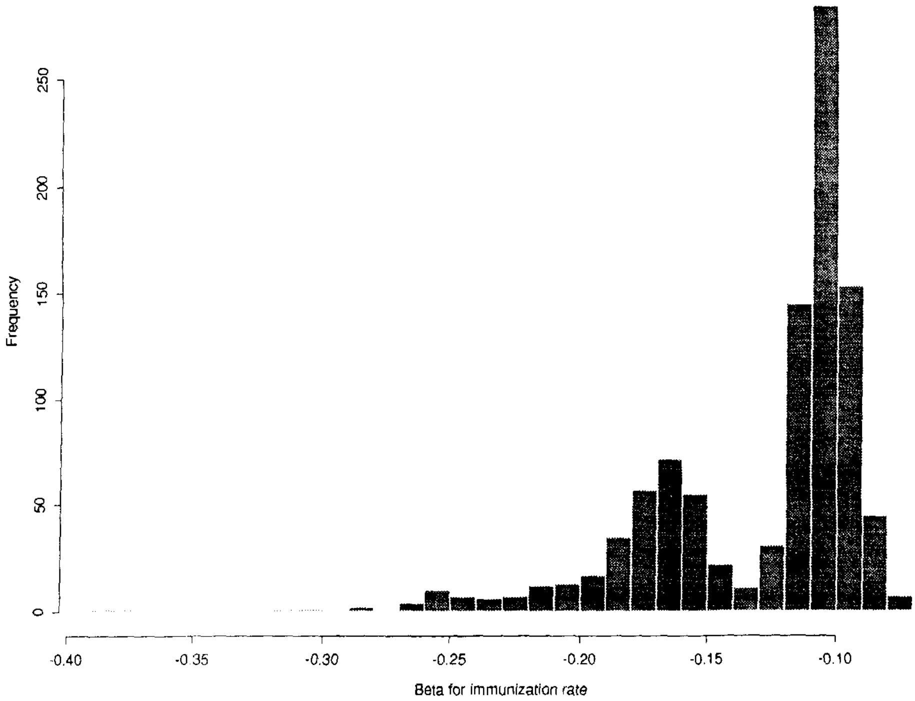

FIG. 1

Histogram of the 1000 bootstrap replicates $\hat{\beta}_1^*$ in Example 5.2.

and the probability that any specific county is excluded from a bootstrap resample is $(1-(1/15))^{15} \sim e^{-1}$ (approximately 1/3). Thus one may suspect that the parameter estimate is high if a specific county is included in a bootstrap resample and low if it is excluded. Due to the complicated nature of the model it is not easy to tell which, if any of the counties are influential by inspection of the data.

To explore this question, the 1000 $\hat{\beta}_1^*$'s were ranked in increasing order. It turns out that county number 6 was present in only 2 of the resamples leading to the 333 lowest $\hat{\beta}_1^*$'s and was in 658 of the remaining 667. Further the mean of the 660 bootstrap replicates in which this county appeared was -.104, while the mean of those without this county was -.180, corresponding nicely with the observed modes in Figure 1. Notice that this county was also on the far left side of the covariate space. Thus it appears that it was highly influential in this analysis and a model less sensitive to this influential point may give a better fit to the data. So although, quoting Efron (1987), "small sample nonparametric confidence intervals are far from well understood", we can still gain important diagnostic information from the bootstrap distribution based on a relatively small sample.

6. SIMULATION EXPERIMENTS

The following simulations are designed in order to compare the coverage accuracy of the proposed bootstrap confidence intervals with each other and with competing procedures. To make the simulations informative we consider the normal and logistic setups, in a variety of settings, varying dependence structure and spacing of covariates.

6.1 Simulation Experiment 1

In the first simulation data are generated from a linear regression model $Y_i = \beta_0 + \beta_1 x_i + e_i$, where $\beta_0 = 0$, $\beta_1 = 1$. In this situation the exact variance of the least squares estimator, $\hat{\beta}_1$, is known and we compare it with the bootstrap estimator, the GEE-robust estimator, and the least squares estimator under an i.i.d. assumption. We also compare the coverage rate of the bootstrap intervals, bootstrap t-intervals, GEE-robust confidence intervals, using an exchangable working correlation, and the least squares confidence interval under the assumption of independence. In the linear regression model the estimates of β_0 and β_1 are unbiased so there is no need for bias corrected bootstrap methods here.

The results are compared for four different sample sizes and in two different settings:

1. $Y_{ij} = \beta_0 + \beta_1 x_{ij} + e_{ij}$, where i indicates the block, j indicates which observation within a block. There is a group random effect: $e_{ij} = c_i + z_{ij}$ with: $c_i \sim N(0,1), z_{ij} \sim N(0,1)$, and c_i, z_{ij} independent. The covariate x is constant within a group but varies between groups, fixed and equidistant on [-1,1]. The group size is constant and equals 10. The number of groups varies,

2. As for 1, but now x varies within groups. Within each group the x's are fixed and equidistant on [-1,1].

In each simulation series there are 500 replications with B=2000 bootstrap resamples for each. This ensures that each estimated coverage probability has a standard error of less than 1.4 % (for nominal 90 % confidence intervals). The results are in Table 3 and 4.

In both settings the bootstrap variance estimate is as close or closer to the true value than the robust sandwich estimate, most dramatically for small sample sizes. For the smaller sample sizes the bootstrap estimate is on the average too large, while the robust estimate is on the average too small.

The coverage rate of the bootstrap intervals is closer to 90% than the robust intervals in most cases, but they still undercover in most cases. The bootstrap-t intervals have a coverage rate which is closer to 90% than the robust intervals.

Due to the dependency among observations, ordinary least squares, which does not account for dependence, leads in setting 1 to confidence intervals which undercover and in setting 2 to confidence intervals which overcover. This is because standard errors, which ignore the dependence, will be too small when the covariate x is constant within a group (setting 1), but too large if the covariate changes within a group (setting 2). To see this, consider paired observations where the first subject of a pair received a treatment (x=1) while the second one did not receive the treatment (x=0). Then a paired approach (accounting for dependence) estimates the effect of x with a smaller standard error, than the unpaired approach (which ignores dependence).

TABLE III

Table 3: Simulation Experiment 1. Linear regression model. True variance and mean estimated variance of $\hat{\beta}_1$.

	Sample[1]	True Var.	Least Squ.		
	Size			Bootstrap	Robust
Setting 1	100	.270	.045	.292	.194
Random effect	150	.193	.033	.194	.157
x varies	200	.149	.026	.148	.129
between groups	250	.122	.021	.118	.107
group size=10					
Setting 2	100	.025	.047	.023	.022
Random effect	150	.016	.032	.015	.015
x varies	200	.012	.024	.011	.011
within groups	250	.010	.019	.009	.009
group size = 10					

1) Sample size is total number of observations (group size times number of groups)

TABLE IV

Table 4: Simulation Experiment 1. Linear regression model. Observed coverage probabilities of 90 % confidence intervals for β_1, median length of the interval is in parentheses.

	Sample Size[1]	Least Squ.	Percentile	Robust	Boot-t
Setting 1	100	.490(.70)	.842(1.66)	.780(1.38)	.888(2.36)
random effect	150	.502(.60)	.870(1.39)	.826(1.26)	.908(1.69)
x varies	200	.484(.53)	.874(1.23)	.840(1.15)	.902(1.40)
between groups	250	.470(.48)	.866(1.11)	.854(1.06)	.894(1.23)
group size = 10					
Setting 2	100	.972(.72)	.812(.48)	.810(.48)	.892(.59)
random effect	150	.958(.59)	.854(.40)	.866(.40)	.884(.46)
x varies	200	.978(.51)	.870(.35)	.870(.35)	.898(.38)
within groups	250	.972(.46)	.852(.31)	.856(.31)	.878(.34)
group size = 10					

1) Sample size is total number of observations (group size times number of groups)

In both situations, the bootstrap-t intervals have the best coverage. It should be noted, however, that the bootstrap-t intervals are the widest, and occasionally have "wild" endpoints.

6.2 Simulation Experiment 2

In this experiment data are generated from a logistic regression model with $p_i = P[Y_i = 1|x_i] = \exp(\beta_0 + x_i\beta_1)/(1 + \exp(\beta_0 + x_i\beta_1))$, where $\beta_0 = 0$, $\beta_1 = 3$. Here the parameter estimates for β_1 are biased (upwards) and we compute the bias corrected bootstrap confidence intervals as well. For the 3rd and 4th setting we define tetrachoric correlations as follows:

Assume that $(Y_1, \ldots, Y_k)$ are realizations of the continuous latent variables $(Z_1, \ldots, Z_k)$, where $(Z_1, \ldots, Z_k)$ is distributed multivariate normal with mean vector 0 and correlation matrix Σ. The elements of Σ are a measure of the dependence among the Y_i's and are called tetrachoric correlations (Pearson (1900)). Specifically, the outcome $Y_i, i = 1, \ldots, k$, equals 1 if $Z_i < g_i$, where $g_i = \Phi^{-1}(p_i)$ (Φ denotes the cumulative normal distribution function) Y_i and Y_j have tetrachoric correlation equal to the (i, j) element of Σ.

We considered the following settings:

1. Observations Y_i are independent, x's are fixed and equidistant on [-1,1],
2. Observations Y_i are independent, x's come from the skewed design; $x = 2z^2+1$, where z's are fixed and equidistant on [0,1],
3. We have blocks of size 2 (k=2), with tetrachoric correlation equal to .5. The x's are fixed and equidistant on [-1,1],
4. As for 3., but now the group size is $k = 10$, all tetrachoric correlations within a block equal to .5.

Again all bootstrap confidence intervals are based on B=2000 bootstrap resamples and the number of simulations is 500. Table 5 gives the resulting estimated

TABLE V

Table 5: Simulation Experiment 2. Logistic regression model. Observed coverage probabilities for 90 % confidence intervals of β_1, median interval length is in parentheses.

Set.	Sample Size[1]	ML	Perc.	Robust	BC	BCa	Boot-t
1	25[2]	.958(4.2)	.763(8.5)	.867(3.8)	.841(5.1)	.869(4.9)	.827(3.9)
	50	.932(2.8)	.856(3.4)	.906(2.7)	.894(2.9)	.912(2.9)	.894(2.7)
	75	.918(2.2)	.862(2.5)	.908(2.2)	.894(2.3)	.910(2.3)	.896(2.2)
	100	.918(1.9)	.876(2.0)	.898(1.9)	.894(1.9)	.906(1.9)	.894(1.9)
2	25[3]	.934(4.0)	.766(inf)[4]	.844(3.8)	.834(6.6)	.846(6.1)	.746(3.9)
	50	.930(2.8)	.838(3.6)	.896(2.7)	.880(3.0)	.904(2.9)	.888(2.7)
	75	.912(2.2)	.852(2.5)	.898(2.2)	.886(2.3)	.902(2.3)	.898(2.1)
	100	.900(1.9)	.864(2.1)	.898(1.9)	.884(2.0)	.898(2.0)	.900(1.9)
3	50	.866(2.7)	.808(4.2)	.870(3.0)	.848(3.4)	.868(3.4)	.858(3.1)
	100	.866(1.9)	.856(2.3)	.882(2.1)	.864(2.2)	.876(2.1)	.868(2.1)
	200	.854(1.3)	.862(1.5)	.888(1.5)	.878(1.5)	.882(1.5)	.890(1.5)
4	100	.642(2.1)	.766(5.4)	.798(2.9)	.806(4.3)	.786(5.1)	.864(3.8)
	200	.636(1.4)	.818(2.7)	.832(2.2)	.832(2.5)	.830(2.7)	.848(2.5)

1) Sample size is total number of observations (group size times number of groups)
2) In 3 simulations the maximum likelihood estimate was infinite. These three simulations are deleted.
3) In 12 simulations the maximum likelihood estimate was infinite.
4) The median length of the Percentile interval in setting 2, n=25 was infinite because in more than half of the simulations the upper interval endpoint was infinite.

coverage rates (for β_1) and median interval width from the 500 simulations in each setting.

We first discuss coverage: In settings 1 and 2, where the observations are independent we see that the bootstrap percentile method badly undercovers, with the BC performing somewhat better. This phenomenon was noted by, for example, Schenker (1985), for statistics with skewed distributions (of which $\hat{\beta}_1$ is an example). The BC$_a$ intervals are competitive with the GEE intervals (using robust standard errors under an assumed exchangeable correlation structure). The bootstrap-t intervals have good coverage probabilities except for the smallest sample size. One problem is that $\hat{\beta}_1$ is infinite when the group $Y=0$ and the group $Y=1$ can be completely separated based on the covariate x. When samples sizes are small this happens occasionally.

This problem is exacerbated when bootstrapping. In the first simulation, with $n = 25$ the average fraction of resamples with infinite estimates was 13%. This yields no problems when constructing bootstrap percentile or BC intervals, although it can occasionally happen that one of the interval endpoints is infinite. However $t_i^* = (\hat{\beta}_i^* - \hat{\beta})/se(\hat{\beta}_i^*)$ is no longer defined, and to construct bootstrap-t intervals we used an ad hoc approach using only those resamples where the estimate is not infinite. This procedure induces a small bias against the bootstrap-t, however.

The results in setting 3 are very similar. The robust-GEE intervals, BC$_a$ intervals and bootstrap-t interval coverage rates are comparable. In setting 4 the dependence is much heavier. The ML intervals now badly undercover because the estimated standard errors are too small. In this setting the bootstrap-t intervals outperform the other intervals.

In considering interval width, we see that not only do the percentile intervals undercover but they are wider than the competitors, the BC intervals perform somewhat better, but they are still wider than the BC$_a$ intervals. This gives additional incentive to use the BC$_a$ intervals. In all cases the BC$_a$ intervals are wider than the Robust intervals and thus give less precise information. The bootstrap-t, on the other hand, has more comparable width to the Robust intervals, while generally having better coverage.

7. CONCLUSIONS

We have discussed and evaluated bootstrap algorithms which account for correlated outcomes in generalized linear models and compared their performance to existing methods. Two examples show the corroborative usefullness and diagnostic potential of the bootstrap algorithms.

Through simulation experiments we have compared the bootstrap approach with existing methods. In the first simulation experiment, the bootstrap-t intervals are superior. For small sample sizes the bootstrap over-estimates the variance of the parameter estimates while the robust variance estimate is too small. Nevertheless, the bootstrap estimate of variance is closer to the true variance than the robust variance estimator.

When the parameter estimators are biased, as in the logistic model, the bootstrap percentile method does not work well, and a bias correction method should be used. In the logistic model there is the additional problem of infinite parameter estimates. Due to this problem, the bootstrap-t did not work well when the sample sizes are small. When the number of blocks is small, but the total sample sizes is moderate, the bootstrap-t method is, in general, superior to other methods considered.

APPENDIX: DATA IN EXAMPLE 5.2

Data for the measles example. 'county' denotes the county number, 'cases' the number of preschool measles cases, 'rate' the immunization rate, and 'children' the total number of preschoolers.

county	cases	rate	children	year
1	1	72	46917	85
1	19	72	48334	86
1	3	72	49620	87
1	0	72	51170	88
1	5	72	53233	89
1	4	72	55291	90
1	1	72	57025	91
2	8	68	47123	85
2	0	68	47799	86
2	1	68	48672	87
2	0	68	49726	88
2	26	68	51391	89
2	16	68	53420	90
2	0	68	55191	91
3	13	65	101479	85
8	3	88	88142	89
8	1	88	88413	90
8	1	88	88802	91
9	9	80	93806	85
9	18	80	95551	86
9	1	80	97042	87
9	0	80	99123	88
9	2	80	102098	89
9	28	80	105362	90
9	2	80	108362	91
10	0	69	52173	85
10	0	69	52075	86
10	0	69	51385	87
10	0	69	50492	88
10	0	69	49555	89

(continued)

Appendix Continued

county	cases	rate	children	year
3	257	65	104772	86
3	114	65	108710	87
3	3	65	113203	88
3	5	65	118453	89
3	397	65	123888	90
3	370	65	128463	91
4	1	81	106798	85
4	0	81	106163	86
4	0	81	105756	87
4	0	81	106038	88
4	35	81	107101	89
4	8	81	108811	90
4	0	81	110754	91
5	6	65	270171	85
5	4	65	272265	86
5	2	65	269601	87
5	184	65	265986	88
5	729	65	265161	89
5	43	65	266885	90
5	5	65	271107	91
6	1	50	41126	85
6	354	50	41426	86
6	1	50	41983	87
6	0	50	42736	88
6	102	50	43842	89
6	30	50	45229	90
6	98	50	46548	91
7	0	65	127411	85
7	3	65	129602	86
7	1	65	132260	87
7	0	65	135503	88
7	13	65	139514	89
7	136	65	143145	90
7	612	65	145446	91
8	1	88	89081	85
8	0	88	89118	86
8	0	88	88766	87
8	2	88	88308	88
10	0	69	48384	90
10	0	69	47342	91
11	6	70	55459	85
11	43	70	56600	86
11	1	70	57415	87
11	0	70	58883	88
11	0	70	61014	89
11	136	70	64929	90
11	0	70	70411	91
12	1	65	92189	85
12	6	65	94727	86
12	0	65	97188	87
12	1	65	100216	88
12	7	65	103438	89
12	58	65	106854	90
12	1	65	110292	91
13	8	69	127809	85
13	197	69	133299	86
13	32	69	138222	87
13	23	69	143049	88
13	17	69	149032	89
13	26	69	155575	90
13	28	69	161370	91
14	0	71	70433	85
14	0	71	70250	86
14	1	71	69948	87
14	1	71	69947	88
14	17	71	70448	89
14	1	71	70933	90
14	0	71	71440	91
15	2	60	79480	85
15	0	60	79931	86
15	0	60	80141	87
15	0	60	80493	88
15	382	60	81419	89
15	359	60	82333	90
15	1	60	83066	91

ACKNOWLEDGEMENTS

This work has been supported by NIH grant T32-CA09667 and NCI grant CA72015-01 (Michael Sherman) and a NATO-science fellowship (Saskia le Cessie).

The authors would like to thank Martin Tanner and Hans van Houwelingen for helpful discussions.

BIBLIOGRAPHY

Bose, A. (1988). "Edgeworth Correction By Bootstrap in Autoregression" *Annals of Statistics*, 16, 1709-1722.

Breslow, N.E. and Clayton, D.G. (1993). "Approximate Inference in Generalized Linear Mixed Models" *Journal of the American Statistical Association* 88, 9-25.

Connolly, M.A. and Liang, K.Y. (1988). "Conditional Logistic Regression Models for Correlated Binary Data" *Biometrika* 75, 501-506.

Diggle, P.J., Liang, K.-Y. and Zeger, S.L. (1994). *Analysis of Longitudinal Data,* Oxford: Oxford University Press.

Efron, B. (1987). "Better Bootstrap Confidence Intervals" *Journal of the American Statistical Association* 82, 171-185.

Efron, B. and Tibshirani, R. (1993). *Introduction to the Bootstrap,* New York: Chapman and Hall.

Freedman, D.A. (1981). "Bootstrapping Regression Models" *Annals of Statistics* 9, 1218-1228.

Freedman, D.A. (1984). "On Bootstrapping Two-stage Least Squares Estimates in Stationary Linear Models" *Annals of Statistics* 12, 827-842.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion,* New York: Springer-Verlag.

Karim, M.R. and Zeger, S.L. (1988). "GEE a SAS Macro for Longitudinal Data Analysis" *Technical report #674,* Department of Biostatistics, John Hopkins University.

Künsch, H. (1989). "The Jackknife and the Bootstrap for General Stationary Observations" *Annals of Statistics* 17, 1217-1241.

le Cessie, S. and van Houwelingen, J.C. (1994). "Logistic Regression for Correlated Binary Data" *Applied Statistics (JRSS-C)* 43, 95-108.

Liang, K.-Y. and Zeger, S.L. (1986). "Longitudinal Data Analysis Using Generalized Linear Models" *Biometrika* 73, 13-22.

Liang, K.-Y., Zeger, S.L. and Qaqish, B. (1992). "Multivariate Regression Analysis for Categorical Data" *Journal of the Royal Statistical Society (with discussion)* B 54, 3-40.

Lipsitz, S.R., Laird, N.M. and Harrington, D.P. (1990). "Maximum Likelihood Regression Methods for Paired Binary Data" *Statistics in Medicine* 9, 1517-1525.

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

Moulton, L.H. and Zeger, S. (1989). "Analyzing Repeated Measures on Generalized Linear Models via the Bootstrap" *Biometrics* 45, 381-394.

Pearson, K. (1900). "Mathematical Contribution to the Theory of Evolution VII. On the Correlation of Characters not Quantitatively Measurable" *Philosophical Transactions of the Royal Society of London* A, 195, 1-47.

Prentice, R.L. (1988). "Correlated Binary Regression with Covariates Specific to each Binary Observation" *Biometrics* 44, 1033-1048.

Royall, R.M. (1986). "Model Robust Confidence Intervals Using Maximum Likelihood Estimators" *International Statistical Review* 54, 221-226.

Schenker, N. (1985). "Qualms About Bootstrap Confidence Intervals" *Journal of the American Statistical Association* 80, 360-361.

Thall, P.F. and Vail, S.C. (1990) "Some Covariance Models for Longitudinal Count Data with Overdispersion" *Biometrics* 46, 657-671.

Verloove, S.P. and Verwey, R.Y. (1988). *Project on Preterm and Small-for-gestational Age Infants in the Netherlands, 1983 (Thesis, University of Leiden).* University Microfilms International, Ann Arbor, Michigan, USA, no 8807276.

Zeger, S.L., Liang, K.Y. and Albert, P.S. (1988). "Models for Longitudinal Data : a Generalized Estimation Equation Approach" *Biometrics* 44, 1049-1060.

Received November, 1996; Revised December, 1996.