Bootstrapping Multiparameter Models, with Applications to Clustered Binary Data

Marc Aerts[1], Gerda Claeskens[2], Geert Molenberghs[1]

[1] Center for Statistics, Limburgs Universitair Centrum, Universitaire Campus, B3590 Diepenbeek, Belgium
[2] Centre for Mathematics and its Applications, School of Mathematical Sciences, Australian National University, Canberra, ACT 0200, Australia

Abstract: It is shown how a one-step semiparametric bootstrap procedure can be applied to multiparameter models in different situations: for testing hypotheses, for the construction of simultaneous confidence intervals based on local polynomial smoothers and for improved estimation and bias correction. The method is illustrated on models for clustered binary data.

Keywords: Bootstrap, Clustered Binary Data, Local Polynomial Smoothing, Multiparameter Models, Testing Hypotheses.

1 Introduction

The bootstrap is a well established statistical methodology nowadays. There are several papers and books showing a multitude of examples where the bootstrap can be implemented and applied succesfully, see e.g. Davison and Hinkley (1997). Here we are interested in applying the bootstrap to clustered binary data, typically modelled by multiparameter likelihood models. There has been considerable interest in bootstrapping generalized linear models (see e.g. Moulton and Zeger 1989) but, to our knowledge, there are not many results on applying the bootstrap to multiparameter models in general. Of course, for fully specified likelihood models, one can always apply the parametric bootstrap. Such an approach has been generalized to pseudolikelihood models and applied to clustered binary data in Aerts and Claeskens (1999a). But often the "true likelihood" is unknown and one might expect a parametric bootstrap to break down if the likelihood model of the data is grossly misspecified. Therefore, a semiparametric bootstrap approach might be preferable. Such a robust method is presented here and it is shown how it can be applied to testing hypotheses, the construction of confidence intervals and to multiparameter local likelihood models. It should be stressed that although we focus attention to *clustered binary response* data, the domain of application of these methods is much broader.

2 A One-Step Bootstap Procedure

Let $\boldsymbol{Y}_i$, $i = 1, \ldots, n$ be independent response variables of length m with (unknown) joint density or discrete probability function (pdf) $g(\boldsymbol{y}_i; \boldsymbol{x}_i)$ where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{im})$ and $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})$, the latter representing a vector of p explanatory variables. In the context of clustered binary data, m corresponds to the size of the cluster.

In general, parametric inference is based on an r dimensional score function $\boldsymbol{\psi}(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{t})$, where the "true" parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)$ is defined as the solution $\boldsymbol{t}$ to $\sum_{i=1}^n E[\boldsymbol{\psi}(\boldsymbol{Y}_i; \boldsymbol{x}_i, \boldsymbol{t})] = \boldsymbol{0}$ where all expectations are w.r.t. the true pdf $g(\boldsymbol{y}_i; \boldsymbol{x}_i)$. Solving the system of equations $\sum_{i=1}^n \boldsymbol{\psi}(\boldsymbol{Y}_i; \boldsymbol{x}_i, \boldsymbol{t}) = \boldsymbol{0}$ leads to the estimator $\widehat{\boldsymbol{\theta}}_n$ for $\boldsymbol{\theta}$.

Within classical maximum likelihood $\boldsymbol{\psi}(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{t}) = (\partial/\partial \boldsymbol{t}) \log f(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{t})$ and, for clustered binary data, $f(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{t})$ represents e.g. the beta-binomial distribution or the conditional model of Molenberghs and Ryan (1999) (MR-model). Note that, in this setting, the assumed pdf $f(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{t})$ might not contain the true structure $g(\boldsymbol{y}; \boldsymbol{x})$. Effects of likelihood misspecification are examined in Molenberghs, Declerck and Aerts (1998). But $\boldsymbol{\psi}(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{t})$ might also represent the pseudolikelihood scores (see Geys, Molenberghs and Ryan 1999) or generalized estimating equations.

We propose to resample the score and the differentiated score values. Based on a linear approximation, we define a bootstrap replicate of $\widehat{\boldsymbol{\theta}}_n$ as

$$\widehat{\boldsymbol{\theta}}_n^* = \widehat{\boldsymbol{\theta}}_n - \left(\sum_{i=1}^n \dot{\boldsymbol{\psi}}_i^*(\widehat{\boldsymbol{\theta}}_n) \right)^{-1} \sum_{i=1}^n \boldsymbol{\psi}_i^*(\widehat{\boldsymbol{\theta}}_n) \tag{1}$$

where $(\boldsymbol{\psi}_i^*(\widehat{\boldsymbol{\theta}}_n), \dot{\boldsymbol{\psi}}_i^*(\widehat{\boldsymbol{\theta}}_n)), i = 1, \ldots, n$ is a sample with replacement from the set $\left\{ \left(\boldsymbol{\psi}(\boldsymbol{Y}_i; \boldsymbol{x}_i, \widehat{\boldsymbol{\theta}}_n), (\partial/\partial\boldsymbol{\theta}) \boldsymbol{\psi}(\boldsymbol{Y}_i; \boldsymbol{x}_i, \widehat{\boldsymbol{\theta}}_n) \right), i = 1, \ldots, n \right\}$. A similar linearization idea is used in simulation approaches for the bootstrap, as the linear bootstrap and the one-step bootstrap. For linear models $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, the idea of resampling scores has also been proposed by Hu and Zidek (1995). Inspired by higher order approximations, the linear bootstrap (1) can be considerably improved by a one-step quadratic bootstrap (see Aerts and Claeskens 1999b).

3 Application 1: Hypothesis Testing

Although testing hypotheses is also of great interest in settings with clustered binary data, bootstrap tests have never been studied and applied extensively in this situation. One of the main reasons for this is that for the bootstrap to work, data have to be generated under the restrictions imposed by the specific null hypothesis. Aerts and Claeskens (1999b) show how valid Wald and score tests can be based on the one-step bootstrap by

replacing $\widehat{\boldsymbol{\theta}}_n$ by $\widehat{\boldsymbol{\theta}}_n^{(0)}$ in the rhs of definition (1). This null estimate $\widehat{\boldsymbol{\theta}}_n^{(0)}$ reflects the null hypothesis and the second term of (1) represents the random fluctuation of the bootstrap replicate $\widehat{\boldsymbol{\theta}}_n^*$ around the estimator $\widehat{\boldsymbol{\theta}}_n^{(0)}$.
As an example, consider simulated data as they appear in developmental toxicity studies with rodents. We selected dose levels 0, 0.25, 0.5, 1 and an equal number of 15 litters, assigned to each dose group. 500 datasets were generated from the beta-binomial distribution with $\text{logit}(\pi(d)) = \theta_{10}+\theta_{11}d$, $\text{FisherZ}(\rho(d)) = \theta_{20}$ under the null hypothesis that $H_0 : \theta_{11} = 0$ (no dose effect). Here, for a pregnant rodent exposed to dose d, $\pi(d)$ is the probability that an individual fetus is malformed and $\rho(d)$ represents the intra-litter correlation. For each run, the scores are resampled 1000 times in each dose group separately, denoted by B_1/D for the linear and B_2/D for the quadratic one-step bootstrap method. Resampling the complete set of scores is denoted by B_i/A $(i = 1, 2)$. Finally, B_{it}/D corresponds to resample the data in each dose group. The data were fitted using the pseudolikelihood model and the robust Wald and robust score statistics, testing for no dose effect, were calculated. Some results are shown in Table 1 ($*$ denotes the proportion of significant tests (at 5%) which differs significantly from 5%).

θ_{10}		χ^2	B_1/D	B_2/D	B_{it}/D	B_1/A	B_2/A
-4.0	W_n	10.55*	10.76*	6.96	10.76*	9.28*	6.54
	S_n	6.12	6.75	—	—	5.70	—
-2.5	W_n	7.80*	6.60	5.80	8.20*	6.00	5.20
	S_n	7.40*	5.60	—	—	5.20	—

TABLE 1. Simulated type I errors (as %), significance level 0.05. Data are generated with the beta-binomial model (with $\theta_{20} = 0.2$) and fitted using the pseudolikelihood model. $H_0 : \theta_{11} = 0$.

.

4 Application 2: Bootstrapping Local Likelihood Estimators

Aerts and Claeskens (1997) and Claeskens and Aerts (1999) studied local polynomial likelihood in the context of clustered binary data. Definition (1) can be modified by including kernel weights $(K((x_i - x)/h)$ for $p = 1$ and K a density) and an extra term in the rhs representing a bias correction. Details and consistency results for this local version of the one-step bootstrap are given in Claeskens and Aerts (1999). There it is also indicated how simulation of the bootstrap distribution allows for the construction of simultaneous confidence intervals in a finite number of grid points.

As an illustration, consider data from the Wisconsin diabetes study. Both eyes of each of 720 younger onset diabetic persons were examined for the presence of macular edema. See Klein, Klein, Moss, Davis, and DeMets (1984) for more details. So the response data are $\boldsymbol{y}_i = (y_{i1}, y_{i2})$ with y_{ij} the binary response value of eye $j = 1, 2$ of person i. We will study the probability of macular edema as a function of the patient's systolic blood pressure, hereby taking the clustered nature of the data into account, as indeed the response values of both eyes are likely to be correlated. The simultaneous and pointwise 90% confidence intervals for the probability of macular edema and for the intra-person correlation, are given in Figure 1.

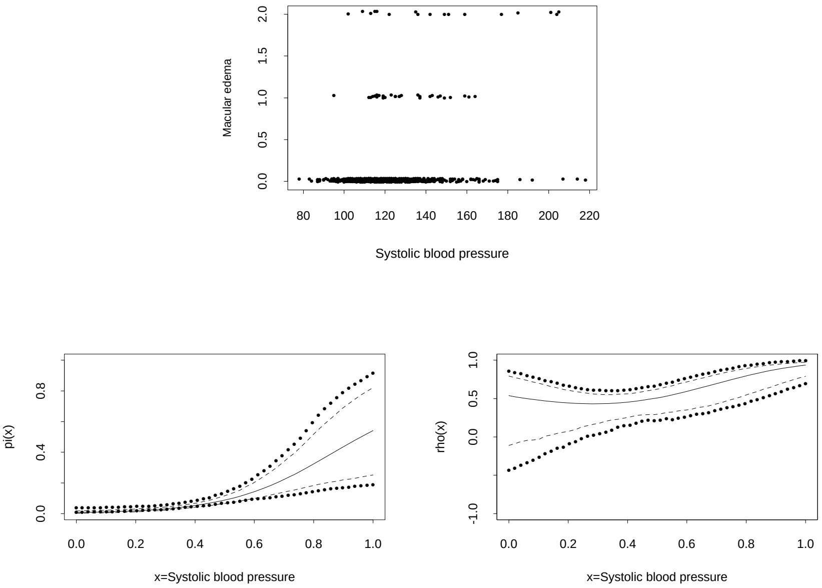

FIGURE 1. The Wisconsin diabetes data (top panel), Simultaneous and pointwise 90 % confidence intervals for the probability of macular edema (left bottom panel) and the intra-person correlation (right bottom panel).

5 Application 3: Bias Correction and Double Bootstrap

Although the ML estimator $\widehat{\boldsymbol{\theta}}_n$ is asymptotically unbiased, the quadratic one-step bootstrap procedure can be used for finite sample bias correction.

In practical applications a large number, say B, resamples are taken, resulting in a set of B bootstrap estimators $\widehat{\boldsymbol{\theta}}_n^{*1}, \dots, \widehat{\boldsymbol{\theta}}_n^{*B}$. From this set a bias corrected estimator is defined as $\widehat{\boldsymbol{\theta}}_n^{bc} = 2\widehat{\boldsymbol{\theta}}_n - \frac{1}{B}\sum_{i=1}^{B} \widehat{\boldsymbol{\theta}}_n^{*i}$.
Simulations show that this bias correction might even decrease the variance. Using a double bootstrap procedure, Aerts, Claeskens and Molenberghs (1999) study the distribution of $\widehat{\boldsymbol{\theta}}_n^{bc}$ and define a bootstrap based variance estimator for $\widehat{\boldsymbol{\theta}}_n^{bc}$.
Table 2 shows that the quadratic one-step bootstrap slope estimator is quite able to estimate the finite sample bias. The settings in this simulation were as follows. We generated 2000 data sets of size $n = 10$ and $n = 25$, for each value of x, from a logistic regression model logit$\{P(Y = 1)\} = \beta_0 + \beta_1 x$, with (β_0, β_1) equal to (-1,-1), (-2.5, 1) or (-2.5, 2), and $x = 0, 0.25, 0.5$ and 1. For each of these 2000 data sets we constructed 1000 one-step quadratic bootstrap replicates, the latter were used to obtain the bias corrected estimates $(\hat\beta_0^{bc}, \hat\beta_1^{bc})$.
An important observation is that the bias correction even decreases the variance, as the simulated standard deviation $\sigma(\hat\beta_0^{bc})$ and $\sigma(\hat\beta_1^{bc})$ are, for all settings in this study, smaller than the corresponding simulated values of $\sigma(\hat\beta_0)$ and $\sigma(\hat\beta_1)$, respectively.

	$\beta_0 = -1$		$\beta_0 = -2.5$		$\beta_0 = -2.5$	
	$\beta_1 = -1$		$\beta_1 = 1$		$\beta_1 = 2$	
	$n = 10$	$n = 25$	$n = 10$	$n = 25$	$n = 10$	$n = 25$
$E(\hat\beta_1)$	-1.265	-1.070	0.889	1.027	2.152	2.096
$E(\hat\beta_1^{bc})$	-1.043	-0.988	0.852	1.001	1.959	2.022
$\sigma(\hat\beta_1)$	1.511	0.795	1.555	0.908	1.303	0.779
$\sigma(\hat\beta_1^{bc})$	1.314	0.747	1.297	0.836	1.119	0.735
$\frac{MSE(\hat\beta_1^{bc})}{MSE(\hat\beta_1)}$	0.734	0.875	0.701	0.846	0.728	0.878

TABLE 2. Simulated mean, standard deviation and mean squared error values of original and bias corrected slope estimators.

Acknowledgements: This project was partly supported by the NATO Collaborative Research Grant 950648. The research of Gerda Claeskens was supported by the Fund for Scientific Research - Flanders (Belgium) (FWO). We thank Professor R. Klein of the University of Wisconsin, Madison, for kindly providing this data set (NIH grant EY 03083, Wisconsin Diabetic Retinopathy Study).

References

Aerts, M. and Claeskens, G. (1997). Local polynomial estimators in multiparameter likelihood models, *Journal of the American Statistical Association*, **92**, 1536–1545.

Aerts, M. and Claeskens, G. (1999a). Bootstrapping pseudolikelihood models for clustered binary data, *Annals of the Institute of Statistical Mathematics*, **51**, 515–530.

Aerts, M. and Claeskens, G. (1999b). Bootstrap tests for misspecified models, with application to clustered binary data", Submitted.

Aerts, M., Claeskens, G. and Molenberghs, G. (1999). A note on the quadratic bootstrap and improved estimation in logistic regression, Technical Report, Limburgs Universitair Centrum, Diepenbeek.

Claeskens, G. and Aerts, M. (2000). Bootstrapping local polynomial estimators in likelihood-based models, *Journal of Statistical Planning and Inference*, **86**, 63–80.

Davison, A.C. and Hinkley, D.V. (1997). *Bootstrap Methods and Their Application.* Cambridge: Cambridge University Press.

Geys, H., Molenberghs, G. and Ryan, L.M. (1999). Pseudo-likelihood modelling of multivariate outcomes in developmental toxicology, *Journal of the American Statistical Association*, **94**, 734–745.

Hu, F. and Zidek, J. (1995). A bootstrap based on the estimation equations of the linear model, *Biometrika*, **82**, 263–275.

Klein, R., Klein, B.E.K., Moss, S.E., Davis, M.D. and DeMets, D.L. (1984). The Wisconsin epidemiologic study of diabetic retinopaty: II. Prevalence and risk of diabetic retinopathy when age at diagnosis is less than 30 years, *Archives of Ophthalmology,* **102**, 520–526.

Molenberghs, G., Declerck, L. and Aerts, M. (1998). Misspecifying the likelihood for clustered binary data, *Computational Statistics and Data Analysis,* **26**, 327–349.

Molenberghs, G. and Ryan, L.M. (1999). Likelihood inference for clustered multivariate binary data, *Environmetrics*, **10**, 279–300.

Moulton, L.H. and Zeger, S.L. (1989). Analyzing repeated measures on generalized linear models via the bootstrap, *Biometrics*, **45**, 381–394.