COMPUTATIONAL STATISTICS & DATA ANALYSIS

ELSEVIER Computational Statistics & Data Analysis 39 (2002) 57–74

www.elsevier.com/locate/csda

A modified score function estimator for multinomial logistic regression in small samples

Shelley B. Bull[a,*], Carmen Mak[b], Celia M.T. Greenwood[c]

[a] Samuel Lunenfeld Research Institute and Department of Public Health Sciences, University of Toronto, Toronto, Ont. Canada

[b] Schering Plough Research Institute, Department of Statistics, Kenilworth, NJ, USA

[c] Department of Epidemiology and Biostatistics, McGill University, Montreal, Canada

Received 1 January 2001; received in revised form 1 June 2001; accepted 1 June 2001

Abstract

Logistic regression modelling of mixed binary and continuous covariates is common in practice, but conventional estimation methods may not be feasible or appropriate for small samples. It is well known that the usual maximum likelihood estimates (MLEs) of the log-odds-ratio parameters are biased in finite samples, and there is a non-zero probability that an MLE is infinite, i.e., does not exist. In this paper, we extend the approach proposed by Firth (Biometrika 80 (1993) 27) for bias reduction of MLEs in exponential family models to the multinomial logistic regression model, and consider general regression covariate types. The method is based on a suitable modification of the score function that removes first order bias. We apply the method in the analysis of two datasets: one is a study of disease prognosis and the other is a disease prevention trial. In a series of simulation studies in small samples, the modified-score estimates for binomial and trinomial logistic regressions had mean bias closer to zero and smaller mean squared error than other approaches. The modified-score estimates have properties that make them attractive for routine application in logistic regressions of binary and continuous covariates, including the advantage that they can be obtained in samples in which the MLEs are infinite.

Keywords: Asymptotic bias; Bayesian estimates; Bias reduction; Continuous covariate; Infinite estimates; Jeffreys' prior; Odds ratio; Polychotomous logistic regression; Polytomous logistic regression; Sparse data

* Corresponding author. Samuel Lunenfeld Research Institute, Suite 850, 600 University Avenue, Toronto, Ont. Canada M5G 1X5. Tel.: +416-586-8245; fax: +416-586-8404.

E-mail address: bull@mshri.on.ca (S.B. Bull).

PII: S 0167-9473(01)00048-2

1. Introduction

Methods for logistic regression modelling of nominal categorical responses based on the multinomial logistic likelihood are now generally available in standard statistical packages, and have been applied in the analysis of case-control studies with multiple case or multiple control groups, and in randomized trials and cross-sectional surveys with categorical responses. One of the concerns of investigators is the valid estimation of model parameters in the finite sample sizes encountered in practice. In finite samples, the usual maximum likelihood estimates (MLEs) of the log odds ratios are biased, and the bias increases as the ratio of the number of observations to the number of parameters (n-to-p ratio) decreases (Cordeiro and McCullagh, 1991; Bull et al., 1997). This is of particular concern when there are several response categories and multiple covariates because the number of parameters can become large.

We consider an alternative estimation method for small samples based on a modification of the score function that removes first order bias and is equivalent to penalizing the likelihood by Jeffreys' prior (Firth, 1993). We extend the modified score function method to the multinomial logistic regression model with nominal response categories, and compare the modified estimates to the usual MLEs and to the MLEs corrected by an estimate of the asymptotic bias. As the sample size increases, the modified-score estimates become equivalent to the usual MLEs. As systematic small sample comparisons of this approach have not been reported previously for binomial or multinomial logistic regression models, we also present a Monte Carlo simulation study in which we compare the mean bias and mean squared error (MSE) of the modified estimates to the MLEs, and to the MLEs corrected by the estimated asymptotic bias. Considering the same series of logistic regression models studied previously (Bull et al., 1997), we find that the modified-score estimates are competitive and often superior to the other approaches.

2. Methods for small-sample analysis

The small-sample properties of the logistic regression MLEs can be improved by the general approach of Cox and Snell (1968) which uses higher order terms in a Taylor series expansion of the log-likelihood to approximate the asymptotic bias and obtain bias-corrected MLEs (Anderson and Richardson, 1979; Schaefer, 1983; Copas, 1988; Cordeiro and McCullagh, 1991; Bull et al., 1997). When the magnitude of the linear predictor is small, Cordeiro and McCullagh (1991) showed that the effect of bias correction in generalized linear models is to shrink the MLEs toward the origin by a factor that depends on the ratio of the number of parameters to the sample size and on the magnitude of the underlying log-odds-ratio parameter. The asymptotic bias correction method reduces bias and MSE (Anderson and Richardson, 1979; Schaefer, 1983), but in small samples tends to correct beyond the true value (Bull et al., 1997). Bias reduced estimates can also be obtained by jackknife methods, with or without full iteration, but the fully iterated estimates, particularly, are also over