Fitting Generalized Linear Models to sample survey data

Alastair Scott and Thomas Lumley

University of Auckland, New Zealand

Abstract

Data from large complex surveys like NHANES are being used increasingly to build regression models. To give some idea of the extent of this, a call to Google Scholar comes up with more than 30,000 papers containing both "NHANES" and "regression model". Unfortunately complexities such as variable selection probabilities and multi-stage sampling mean that the assumptions underlying standard statistical methods for model-building are not even approximately valid for survey data. The problem of parameter estimation has been largely solved through the use of weighted estimating equations, and software for fitting GLMs to survey data is now available in most major statistical packages. The big gap in the output from these packages is an analogue of the deviance and related quantities like AIC. It turns out to be straightforward to extend the results in Rao & Scott (1984) for loglinear models in contingency tables to arbitrary GLMs. We show that the asymptotic distribution of the log-likelihood ratio is a linear combination of chi-squared random variables whose coefficients are eigenvalues of a matrix product that does not involve the inverse of the estimated covariance matrix. We then use results from Scott & Styan (1985) to obtain usable approximations to this asymptotic distribution using only information that is routinely available in large public-release surveys.