

SOME FLEXIBLE REGRESSION TECHNIQUES FOR COMPLEX SURVEYS

D.R. Bellhouse¹, H. Chipman¹ and J.E. Stafford²

ABSTRACT

Survey sampling is a statistical domain that has been slow to take advantage of flexible regression methods. Two approaches could be used to try to make these methods accessible: adapt the techniques to the complex survey design that has been used, or sample the survey data so that the standard techniques are applicable. In following the former route, we introduce techniques that account for the complex survey structure of the data for the techniques of scatterplot smoothing and additive models. The use of penalized least squares in the sampling context is studied as a tool for the analysis of a general trend in a finite population. We focus on smooth regression with a normal error model. Ties in covariates abound for large scale surveys resulting in the application of scatterplot smoothers to means. The estimation of smooths (for example smoothing splines) is seen to depend on the sampling design only via the sampling weights, meaning that standard software can be used for estimation. Inference for these curves is more challenging, due to correlations induced by the sampling design. We propose and illustrate tests which account for the sampling design. Illustrative examples are given using the Ontario health survey, including scatterplot smoothing, additive models, model diagnostics. In an attempt to approach the problem by appropriate sampling of the survey data file we discuss some of the hurdles that are faced with this approach.

KEY WORDS: Penalized least squares; Scatterplot smoothing; Backfitting; Bootstrap; Binning; Cross validation.

¹ University of Western Ontario; bellhouse@stats.uwo.ca

² University of Toronto, Canada