Semiparametric Regression Splines in Matched Case-Control Studies

Inyoung Kim*, Raymond J. Carroll†and Noah Cohen‡

Abstract

We develop semiparametric methods for matched case-control studies using regression splines. Three methods are developed: an approximate crossvalidation scheme to estimate the smoothing parameter inherent in regression splines, as well as Monte Carlo Expectation Maximization (MCEM) and Bayesian methods to fit the regression spline model. We compare the approximate cross-validation approach, MCEM and Bayesian approaches using simulation, showing that they appear approximately equally efficient, with the approximate cross-validation method being computationally the most convenient. An example from equine epidemiology that motivated the work is used to demonstrate our approaches.

Keywords: Bayesian method, Cross-Validation, Matched case-control, Monte Carlo EM, Penalized regression splines.

Running Title: Splines in Matched Case-Control Studies

*Inyoung Kim (kiy@yumc.yonsei.ac.kr), Cancer Metastasis Research Center, Yonsei University, 134Sinchon-dong, Seodaemun-gu, Seoul 120-749, Korea

†Raymond J. Carroll (carroll@stat.tamu.edu), Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA.

‡Noah Cohen (ncohen@cvm.tamu.edu). Department of Large Animal Medicine and Surgery, Texas A& M University, College Station, TX 77843-3143, USA.

1 Introduction

This paper concerns semiparametric regression in matched case–control studies. Matched case–control studies are based upon the classical prospective logistic regression model, with a binary outcome Y (case-control status), covariate (Z, X) and stratum level S and the model

$$\Pr(Y = 1|Z, X) = H\{Z^T\beta_0 + \beta_1 X + q(S)\}, \tag{1}$$

where $H(\bullet)$ is the logistic distribution function and $q(\bullet)$ is an arbitrary function which includes the intercept and unknown effects of the strata. The classical matched case-control study begins with the model (1), but by conditioning on the fixed number of cases in the stratum, any stratum effect is removed, i.e. $q(\bullet)$ disappears (Hosmer and Lemeshow, 1989).

The purpose of this paper is to generalize model (1) into the matched case–control study to allow the effect of X to be modeled nonparametrically. The resulting prospective model is

$$\Pr(Y = 1|Z, X) = H\{Z^T\beta_0 + m(X) + q(S)\}. \tag{2}$$

We model the function $m(X)$ via penalized regression splines, see Eilers and Marx (1996) and Ruppert (2002).

This paper is motivated by an example from equine epidemiology. Equine colic is an important cause of disease and death in horses, frightening for the animal's human companions, and of considerable financial import. Despite the high prevalence of colic, there is not a huge literature on the epidemiology of this disease. We report here results of one of the few large studies extant in the literature (Cohen, et al., 1999), a 1:1 matched case–control study of 498 pairs of horses with colic and their controls. The study design was stratified, with the pairs matched by their veterinarian and the month of examination. Cohen, et al. (1999) reported that horse age was highly associated with risk of colic. It is the purpose of this section to investigate the shape of this relationship.

A simple analysis showed that there was a statistically significant quadratic effect of horse age, suggesting an interesting and somewhat unexpected curvature in the data. This led us to attempt to understand the shape of the age effect in a more detailed way. Our approach, based on model (2), is to fit a quadratic regression spline to the data. The analysis

effectively confirms the quadratic relationship.

2 Methods

We consider several methods to fit our semiparametric regression spline model (2) in the matched case–control study. In general they depend on a smoothing parameter, which can be estimated either by cross-validation or using a mixed model formulation (Coull, et al. 2001). In the latter case, certain of the regression spline parameters are treated as if they are random with a variance related to the smoothing parameter.

Our first approach is to use cross-validation to choose the smoothing parameter. Rather than use the computationally expensive direct cross-validation, we instead develop a computationally convenient approximate method, based on an expansion for the leave-one-out method.

The second approach is to start with a mixed model formulation, where as described above certain parameters of the regression spline are treated as if they were random. We first develop a Monte Carlo EM (MCEM) algorithm in this context, see McCulloch (1997). Our third approach is the natural Bayesian counterpart to the MCEM algorithm, using Gibbs sampling.

3 Application

We did small simulations comparing the various methods. Our simulations suggest that they appear approximately efficient in terms of mean squared error, with the approximate cross-validation method being computationally the fastest, while the Bayesian method carries with it posterior credible intervals. Finally, we apply our approaches to the equine epidemiology example that motivated this work.

REFERENCES

Cohen, N. D., Gibbs, P. G. and Woods, A. M. (1999). Dietary and other management factors associated with colic in horses. *Journal of American Veterinary Medical Association*, 215, 53-60.

Coull, B. A., Ruppert, D. and Wand, M. P. (2001). Simple incorporation of interactions into additive models. *Biometrics*, 57, 539-545.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statistical Science*, 11, 89-121.

Hosmer, D. W. and Lemeshow, S. (1989). *Applied Logistic Regression.* New York: Wiley, New York.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, 92 (437), 162-70.