Bayesian Semiparametric Regression in the Presence of Conditionally Heteroscedastic Measurement and Regression Errors

Author(s): Abhra Sarkar, Bani K. Mallick and Raymond J. Carroll

Source: *Biometrics*, DECEMBER 2014, Vol. 70, No. 4 (DECEMBER 2014), pp. 823-834

Published by: International Biometric Society

Stable URL: https://www.jstor.org/stable/24538366

BIOMETRICS 70, 823–834
December 2014

DOI: 10.1111/biom.12197

Bayesian Semiparametric Regression in the Presence of Conditionally Heteroscedastic Measurement and Regression Errors

Abhra Sarkar, Bani K. Mallick, and Raymond J. Carroll*

Department of Statistics, Texas A&M University, College Station, Texas 77843-3143, U.S.A.
**email:* carroll@stat.tamu.edu

SUMMARY. We consider the problem of robust estimation of the regression relationship between a response and a covariate based on sample in which precise measurements on the covariate are not available but error-prone surrogates for the unobserved covariate are available for each sampled unit. Existing methods often make restrictive and unrealistic assumptions about the density of the covariate and the densities of the regression and the measurement errors, for example, normality and, for the latter two, also homoscedasticity and thus independence from the covariate. In this article we describe Bayesian semiparametric methodology based on mixtures of B-splines and mixtures induced by Dirichlet processes that relaxes these restrictive assumptions. In particular, our models for the aforementioned densities adapt to asymmetry, heavy tails and multimodality. The models for the densities of regression and measurement errors also accommodate conditional heteroscedasticity. In simulation experiments, our method vastly outperforms existing methods. We apply our method to data from nutritional epidemiology.

KEY WORDS: B-splines; Conditional heteroscedasticity; Dirichlet process mixture models; Measurement errors; Regression with errors in covariates; Variance functions.

1. Introduction

We develop a Bayesian semiparametric approach for robust estimation of a regression function when the covariate is measured with error, the density of the covariate, the density of the measurement errors and the density of the regression errors are all unknown, and the variability of both the measurement errors and the regression errors may depend on the associated unobserved value of the covariate through unknown relationships. By "robust" we mean that we avoid restrictive assumptions common in the literature, such as homoscedasticity and normally distributed measurement and regression errors.

The literature on regression with errors in covariates is extensive. A brief review of the existing literature relevant to our problem is presented here. For a more extensive review of the state of the art see Carroll et al. (2006) and Buonaccorsi (2010).

The problem of linear regression in the presence of errors in covariates is vast, and besides the references above also includes the classic text by Fuller (1987). More complex problems have also been studied. Cheng and Riu (2006) studied linear models and considered maximum likelihood, method of moments and generalized least squares estimators for heteroscedastic normally distributed regression and measurement errors. However, they assume that the variances are known and independent of the unobserved value of the covariate. Cook and Stefanski (1994) proposed a simulation-extrapolation (SIMEX) based method that did not make any assumptions about the density of the covariate and the density of the regression errors, but assumes homoscedasticity of both regression and measurement errors: strictly speaking, the latter is assumed to be normally distributed. The SIMEX method also requires the density of the measurement errors to be known. In the presence of replicated surrogates for the unobserved covariate, Devanarayan and Stefanski (2002) relaxed the homoscedasticity assumptions of the SIMEX approach, but the measurement errors are still required to be normally distributed. Carroll, Roeder, and Wasserman (1999a) proposed a Bayesian solution to the problem for normally distributed homoscedastic regression and measurement errors. They modeled the unknown density of the covariate by a finite mixture of normals.

Our focus here is on flexible nonparametric and semiparametric models for all the components. The problem of nonparametric regression with errors in covariates when the regression and measurement errors are both homoscedastic is studied by Fan and Truong (1993), Carroll, Maca, and Ruppert (1999b), Berry, Carroll, and Ruppert (2002), Carroll and Hall (2004) among others. Fan and Truong (1993) studied deconvoluting kernel type estimators when the density of the measurement errors is known. Carroll et al. (1999b) studied SIMEX estimators for the nonparametric regression with errors in covariates problem using three different types of models for the regression function, kernel mixtures, smoothing splines, and penalized truncated polynomial splines, but assuming homoscedastic normally distributed measurement errors. Berry et al. (2002) provided a Bayesian solution to the problem in the presence of normally distributed regression and measurement errors. They also assumed normality of the covariate and modeled the regression function using smoothing splines and penalized mixtures of truncated polynomial splines. Carroll and Hall (2004) considered the problem of estimating a low order estimate of the regression function, rather than the regression function itself. Their method required