Spatial regression with covariate measurement error: A semi-parametric approach

Md Hamidul Huque* Howard D. Bondell† Raymond J. Carroll‡
Louise Ryan§

Abstract
Spatial data have become increasingly common in epidemiology and public health research thanks to rapid advances in GIS (Geographic Information Systems) technology. In health research, for example, it is common for epidemiologists to incorporate geographically indexed data into their studies. In practice, however, the spatially-defined covariates are often measured with error. Naive estimators of regression coefficients are attenuated if measurement error is ignored. Moreover, the classical measurement error theory is inapplicable in the context of spatial modelling because of the presence of spatial correlation among the observations. We propose a semi-parametric regression approach to obtain bias corrected estimates of regression parameters and derive their large sample properties. We evaluate the performance of the proposed method through simulation studies and illustrate using data on Ischemic Heart Disease (IHD). Both simulation and practical application of the proposed method demonstrate that the proposed method can be effective in practice.

Key Words: Attenuation, B-splines, Geostatistics, Measurement error, Penalized least squares, Profile likelihood, SEIFA, Smoothing, Spatial linear model, Spatial regression.

1. Introduction

Rapid growth of high quality Geographic Information Systems (GIS), together with advances in high performance computing environments present a unique opportunity to examine the relationship between risk factors and outcome that vary across time and space. Careful analysis of spatial data can lead to useful explanation of the exposure and disease relationship through natural experimentation (Snow, 1855; Rothman et al., 2008). Spatial analysis of such data helps in understanding the spatial variation of disease, disease clustering, distribution of socio-demographic structure, environmental exposure distribution and its impact on health outcomes.

Analysis of such geo-coded data is complicated by the correlation among neighbouring observations. Regression analysis ignoring this spatial correlation leads to incorrect inference of the estimated regression coefficients by narrowing of associated confidence intervals (Waller and Gotway, 2004). Mixed effect models provide a convenient way of modelling spatial correlations by incorporating spatially defined random effects (Breslow and Clayton, 1993). However, this approach is fully parametric and may be sensitive to model misspecification. Kammann and Wand (2003) studied a semi-parametric formulation of spatial mixed models as a unification of kriging and additive models. Their approach accounts for linear or non-linear covariate effects under the additivity assumption and adjust for spatial correlation by expressing kriging as a linear mixed model.

*School of Mathematical Sciences, University of Technology, Sydney.
†Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA.
‡Department of Statistics, Texas A & M University, College Station, TX 77843-3143, USA.
§School of Mathematical Sciences, University of Technology, Sydney.

However, these advances in statistical methodologies for spatial data analysis are challenged by the presence of measurement error in the covariates as reflected in many epidemiological and socio behavioural studies. For example, in the study of geographical variation of bladder cancer, lung cancer risk might be included in the model as a proxy for smoking exposure (Clayton et al., 1993). In practice, important covariates are often difficult to measure directly from each area or might not present in the original data collection, sample averages of similar covariates from other survey might use as a surrogate measure for the true covariates (Xia and Carlin, 1998). In environmental epidemiology, air pollution level might be approximated by the distance from polluted sites or using the measures available at a few monitoring sites (Carrol et al., l1997). Further examples include geographical mortality studies relating cancer to dietary intakes (Cook and Pocock, 1983; Prentice and Sheppard, 1990).

Many papers have appeared in the literature over the years on covariate measurement error in the context of independent data (Carroll et al., 2006; Fuller, 2009; Ruppert et al., 2009). However, relatively few have addressed the specific context of spatial modelling. Various approaches to adjusting for measurement error bias differ according to the underlying assumptions of the measurement error process, availability of the additional data on the unobserved covariates and theoretical background of the approach, which may be parametric or non-parametric (Guolo, 2008). Bernadinelli et al., (1997) and Xia and Carlin (1998) presented a spatio-temporal analysis of spatially correlated data with errors in the covariates, in the context of disease mapping. They empirically studied several alternative measurement error models using a Gibbs algorithm.

Li et al. (2009) derived asymptotic bias expressions for estimated regression coefficients in the context of a spatial linear mixed model. They showed that the regression estimates obtained from naive use of an error prone covariate attenuates the estimated regression coefficient while variance component estimates are inflated. They proposed the use of a maximum likelihood approach based on the EM algorithm to adjust for measurement error under the assumed error structure. However, their approach is subject to a high computational burden and may lead to spurious results in the presence of outliers or model misspecification (Gryparis et al., 2009; Szpiro et al., 2011). Furthermore, Szpiro et al. (2011) argued that in the presence of spatial correlation, joint modelling becomes challenging as it is very difficult to separate out the spatial correlation between exposure and outcome.

Recently, Huque et al. (2014) have shown that ignoring measurement error and performing a naive analysis attenuates the estimated coefficient towards the null. They showed that the amount of attenuation depends on the strength of spatial correlation in the true covariate of interest. The authors derived expressions for the bias when measurement error is ignored and proposed two different strategies for obtaining consistent estimates: (i) adjusting the estimates using an estimated attenuation factor; and (ii) using an appropriate transformation of the error prone covariate. They showed that bias correction methods using the estimate of the measurement error work reasonably well in obtaining consistent estimates, however, the standard error is underestimated in the case when measurement error variances are estimated from the data. Moreover, their approach is fully parametric. Indeed, Ruppert et al. (2009) argued that penalized splines are the most effective methods for correcting the covariate measurement error in case of independent data. So it is of natural interest to extent the spatial regression model with measurement error to a semiparametric framework.

In this paper we propose a joint modelling approach to assess the relationship between a covariate with measurement error and a spatially correlated outcome in a semiparametric regression context. We estimate the measurement error process and relationship using sub models. Specifically, we expressed the unknown measurement error process as a linear combination of splines basis function, which is then fitted using penalized least squares (Yu and Ruppert, 2002; Xun et al., 2013). Use of penalized least squares makes the estimation of parameters and inference straightforward. We develop an asymptotic theory for estimated parameters and provide a model based and simulation based standard error estimates. Our simulation results reveal that the proposed method works well in obtaining consistent estimate of the true regression coefficient in the presence of measurement error. Our approach is computationally efficient and stable and can be implemented using standard nonlinear least squares software.

The structure of the paper is as follows: Section 2 describes the formulation of the models, estimation and inference procedure. Section 3 presents the data generation process and results from the simulation study. In section 4 we present an application of the proposed method to data on Ischemic Heart Disease (IHD). We conclude with discussion in section 5.

2. Model

Suppose that X_i represents the true covariate of interest for spatial location i, $i = 1, ..., n$, and suppose that it is related to an outcome Y_i, according to a spatial linear model:

$$Y_i = \beta_0 + \beta_1 X_i + G_1(S_i) + \epsilon_i, \tag{1}$$

where, the residuals, $\boldsymbol{\epsilon} = (\epsilon_1, ...\epsilon_n)^T \sim N(0, \sigma_\epsilon^2)$ and $G_1(S_i)$ is an unknown function that captures the spatial correlation, for now kept arbitrary. Furthermore, we assume that ϵ_i and $G_1(S_i)$ are independent of each other and of the covariate X (Cressie, 1993).

In the presence of measurement error, measurements on the true covariate X are not available directly, instead an error contaminated version is available. Let W_i be the observed covariate for spatial location i, related to the true covariate X according to a classical measurement error model:

$$W_i = X_i + U_i, \tag{2}$$

where $U_i \sim \left(0, \sigma_u^2\right)$. Note that in the case of independent data, a consistent estimate of the true regression coefficient β_1 can be obtained if a validation data set on the true covariate (X), without measurement error is available (Carroll et al., 2006). However, in the spatial epidemiology such validation data are relatively rare. Instead, we assume that that the true covariate X is smooth and can be characterized by another smooth function, $G_2(S_i)$.

Many choices have been discussed in the literature to approximate these unknown smooth function $G_j(.)$, for j=1,2. These include parametric modelling using an isotopic correlation function that decays as the distance between two individuals increases (Huque et al., 2014). Other approaches include those of conditional auto regressive functions (CAR) (Banerjee et al, 200) or semi-parametric geoadditive models (Kammann and Wand, 2003) are also popular. In this paper we will adopt techniques based on semi-parametric regression.

Within this framework, the unknown smooth functions, $G_j(.)$, for $j = 1, 2$ can be represented as a linear combination of basis function i.e., $G_j(S_i) = B_j^{\mathrm{T}}(S_i)\theta_j$ and estimated by a penalized least squares (Yu and Ruppert, 2002; Xun et al. 2013). Here $B_1(S_i)$ and $B_2(S_i)$ are two sets of basis functions with dimensions $(q_1+4)\times 1$ and $(q_2+4)\times 1$, respectively, where q_1 and q_2 are the corresponding number of knots and θ_1 and θ_2 are vectors of corresponding basis coefficients. We choose thin plate splines because they do not require knots locations, perform reasonably well using a basis of any given lower rank, are reasonably computationally efficient and more importantly rotationally invariant (Wood, 2006; Ruppert et al., 2003).

Therefore, under the above specifications model (1) and (2) can be rewritten as

$$\begin{aligned} Y_i &= B_2^{\mathrm{T}}(S_i)\theta_2\beta_1 + B_1^{\mathrm{T}}(S_i)\theta_1 + \epsilon_i; & (3)\\ W_i &= B_2^{\mathrm{T}}(S_i)\theta_2 + U_i. & (4)\end{aligned}$$

Note that the intercept term β_0 in the model (1) is set to 0, because it is not identifiable in the presence of a nonparametric function $G_1(\cdot)$. Since these equations are linear with respect to a set of unknown parameters, we can use penalized least squares techniques for estimation. However, the parameters of these models are not completely identifiable without additional assumptions.

2.1 Identifiability

From the above models (3) and (4), it is evident that if $B_1(\cdot)\equiv B_2(\cdot)$, then these models are not identifiable because in this case (3) becomes

$$Y_i = B_2^{\mathrm{T}}(S_i)(\theta_2\beta_1+\theta_1)+\epsilon_i.$$

Thus, we can identify only θ_2 and $\theta_2\beta_1+\theta_1$, and cannot separate out β_1 and θ_1. To make these models identifiable, we assume that the asymptotic variability of two sets of basis functions $B_1(.)$ and $B_2(.)$ will be different. i.e., $\Lambda_1\neq\Lambda_2$, where Λ_j for j=1,2, is the limiting value of Λ_{nj}, defined as, $\Lambda_{nj}=\{n^{-1}\sum_{i=1}^{n}B_j(S_i)B_j^{\mathrm{T}}(S_i)+\delta_jD_j\}^{-1}$ with δ_j and D_j and are corresponding penalty parameters and matrices.

2.2 Parameter estimation

In addition to the assumption that $\Lambda_1\neq\Lambda_2$, we also assume that the smoothing parameters are small relative to the sample size, i.e., $n^{1/2}\delta_j\to 0$ for $j=1,2$. This means that, with the large sample size, the estimated regression coefficient obtained using penalized least squares will be close to the OLS estimates. Using penalization in (4) and solving for θ_2, we have

$$\widehat{\theta}_2=\Lambda_{n2}n^{-1}\textstyle\sum_{i=1}^{n}B_2(S_i)W_i, \qquad (5)$$

Similarly, from (3) we can estimate β_1 and θ_1 by minimizing the corresponding penalized sum of squares as

$$\begin{aligned} \widehat{\theta}_1 &= V_n - R_n\widehat{\theta}_2\beta_1 & (6)\\ \widehat{\beta}_1 &= \frac{n^{-1}\sum_{i=1}^{n}Y_i\{B_2^{\mathrm{T}}(S_i)-B_1^{\mathrm{T}}(S_i)R_n\}\widehat{\theta}_2}{\widehat{\theta}_2^{\mathrm{T}}(\mathcal{T}_n-R_n^{\mathrm{T}}\Lambda_{n1}^{-1}R_n)\widehat{\theta}_2}, & (7)\end{aligned}$$

where

$$\begin{aligned} V_n &= \Lambda_{n1}n^{-1}\textstyle\sum_{i=1}^{n}B_1(S_i)Y_i;\\ R_n &= \Lambda_{n1}n^{-1}\textstyle\sum_{i=1}^{n}B_1(S_i)B_2^{\mathrm{T}}(S_i);\\ \mathcal{T}_n &= n^{-1}\textstyle\sum_{i=1}^{n}B_2(S_i)B_2^{\mathrm{T}}(S_i).\end{aligned}$$

Although the above estimator of β_1 was estimated using pseudolikelihood, it is consistent for β_1. In the next section we will establish the asymptotic properties of the estimator.

2.3 Asymptotic Theory

Asymptotic theory for the estimator $\widehat{\beta}_1$ is based upon (3) by considering the spatial locations S_i as fixed constants. Following Yu and Ruppert (2002), if $\delta_j \to 0$ as $n \to \infty$, then the bias also tends to 0 and consistency can be established. The asymptotic normality under the assumption $n^{1/2}\delta_j \to 0$ for $j = 1, 2$ is easily established by the following theorem:

<u>Theorem</u> 1 Under the assumption that the smoothing parameters are small relative to the sample size, i.e.,$n^{1/2}\delta_j \to 0$, the estimate of β_1 is consistent and asymptotically normally distributed with

$$n^{1/2}\left(\widehat{\beta}_1 - \frac{\mathcal{A}_n\theta_2}{\theta_2^{\mathrm{T}}\mathcal{C}_n\theta_2}\right) \xrightarrow{d} N\left(0, n^{-1}{\textstyle\sum_{i=1}^{n}}(\sigma_\epsilon^2\mathcal{G}_{ni}^2 + \sigma_u^2\mathcal{H}_{ni}^2)\right), \tag{8}$$

where

$$\begin{array}{rcl} \mathcal{A}_n & = & n^{-1}{\textstyle\sum_{i=1}^{n}}\{G_2(S_i)\beta_1 + G_1(S_i)\}\{B_2(S_i) - R_n^{\mathrm{T}}B_1(S_i)\}^{\mathrm{T}}; \\ \mathcal{C}_n & = & \mathcal{T}_n - R_n^{\mathrm{T}}\Lambda_{n1}^{-1}R_n; \\ \mathcal{D}_{ni} & = & \{B_2(S_i) - R_n^{\mathrm{T}}B_1(S_i)\}^{\mathrm{T}}\theta_2; \\ \mathcal{F}_{ni} & = & \theta_2^{\mathrm{T}}\mathcal{C}_2\Lambda_{n2}B_2(S_i) + B_2^{\mathrm{T}}(S_i)\Lambda_{n2}\mathcal{C}_n\theta_2; \\ \mathcal{G}_{ni} & = & \mathcal{D}_{ni}(\theta_2^{\mathrm{T}}\mathcal{C}_n\theta_2)^{-1}; \\ \mathcal{H}_{ni} & = & \mathcal{A}_n\Lambda_{n2}B_2(S_i)(\theta_2^{\mathrm{T}}\mathcal{C}_n\theta_2)^{-1} - \mathcal{A}_n\theta_2\mathcal{F}_{ni}(\theta_2^{\mathrm{T}}\mathcal{C}_n\theta_2)^{-2}. \end{array} \tag{9}$$

see the Appendix A for proof.

Using this asymptotic expression we can also estimate the standard error of estimated regression coefficient $\widehat{\beta}_1$. The next section will discuss two different methods of obtaining standard error.

2.4 Estimating the standard error of $\widehat{\boldsymbol{\beta}}$

We first consider a model based estimate of standard error using asymptotic theorem in the previous section and then suggest a more robust estimate of standard error using simulation.

2.4.1 Model based standard error

The model based standard errors of $\widehat{\beta}_1$ derived in the previous section can be estimated by substituting the following estimates of σ_ϵ^2 and σ_u^2 into expression (8).

$$\begin{array}{rcl} \widehat{\sigma}_\epsilon^2 & = & \dfrac{\sum_{i=1}^{n}\{Y_i - \widehat{G}_2(S_i)\widehat{\beta}_1 - \widehat{G}_1(S_i)\}^2}{n - 2\mathrm{trace}\{L_1(\delta_1,\delta_2)\} + \mathrm{trace}\{L_1(\delta_1,\delta_2)L_1^{\mathrm{T}}(\delta_1,\delta_2)\}} \\ \widehat{\sigma}_u^2 & = & \dfrac{\sum_{i=1}^{n}\{W_i - \widehat{G}_2(S_i)\}^2}{n - 2\mathrm{trace}\{L_2(\delta_2)\} + \mathrm{trace}\{L_2(\delta_2)L_1^{\mathrm{T}}(\delta_2)\}}, \end{array}$$

where $L_1(\delta_1, \delta_2)$ and $L_2(\delta_2)$ are the smoother matrix corresponding to model (3) and model (4). Define $\boldsymbol{B}_j = \{B_j(S_1), ..., B_j(S_n)\}^{\mathrm{T}}$, for j=1,2 and $\boldsymbol{D}_n = \{D_{n1}, ..., D_{nn})\}^{\mathrm{T}}$, then the smoother matrices has the following expressions (see the Appendix B)

$$L_1(\delta_1, \delta_2) \;=\; n^{-1}\left\{\frac{\boldsymbol{D}_n\boldsymbol{D}_n^{\mathrm{T}}}{\widehat{\theta}_2^{\mathrm{T}}\mathcal{C}_n\widehat{\theta}_2} + \boldsymbol{B}_1\Lambda_{n1}\boldsymbol{B}_1^{\mathrm{T}}\right\} \tag{10}$$

$$L_2(\delta_2) \;=\; n^{-1}\boldsymbol{B}_2\Lambda_{n2}\boldsymbol{B}_2^{\mathrm{T}}. \tag{11}$$

2.4.2 Simulated Standard error

From (7), the expression for $\widehat{\beta}_1$ can be written as (see the Appendix A)

$$\widehat{\beta}_1 \;=\; \frac{\mathcal{A}_n\theta_2 + n^{-1}\sum_{i=1}^{n}\{\mathcal{A}_n\Lambda_{n2}B_2(S_i)U_i + \mathcal{D}_{ni}\epsilon_i\}}{\theta_2^{\mathrm{T}}\mathcal{C}_n\theta_2 + n^{-1}\sum_{i=1}^{n}\mathcal{F}_{ni}U_i} + o_p(n^{-1/2}),$$

where U_i and ϵ_i are the random errors define in model (1) and (2). While these quantities are not directly observed we can estimate the variance of $\widehat{\beta}_1$ by a bootstrap.

Let B be a fairly large number, say 100, and for $b = 1, ..., B$, let $\epsilon_{bi} \sim \text{Normal}(0, \widehat{\sigma}_{\epsilon}^2)$ and $U_{bi} \sim \text{Normal}(0, \widehat{\sigma}_u^2)$ for i = 1, 2,...n. Define the b'th bootstrap estimates of β_1 as

$$\widehat{\beta}_1^b = \frac{\widehat{\mathcal{A}}_n\widehat{\theta}_2 + n^{-1}\sum_{i=1}^{n}\{\widehat{\mathcal{A}}_n\Lambda_{n2}B_2(S_i)U_{bi} + \widehat{\mathcal{D}}_{ni}\epsilon_{bi}\}}{\widehat{\theta}_2^{\mathrm{T}}\widehat{\mathcal{C}}_n\widehat{\theta}_2 + n^{-1}\sum_{i=1}^{n}\widehat{\mathcal{F}}_{ni}U_{bi}},$$

where $\widehat{\mathcal{A}}_n, \widehat{\mathcal{D}}_n, \widehat{\mathcal{C}}_n$ and $\widehat{\mathcal{F}}_{ni}$ can be estimated by substitute the appropriate quantities into expression (9). Then the sample variance of $\widehat{\beta}_1^b$ is a consistent estimate of the variance of $\widehat{\beta}_1$ (Efron and Tibshirani, 1986).

2.5 Smoothing parameter selection

Our main objective is to obtain a consistent estimate of the regression parameter β_1 such that it accounts for the measurement error in the covariate. However, selecting a suitable combination of the smoothing parameters (δ_1, δ_2) is a prerequisite to a good model fit. All our discussion so far has assumed that these parameters are fixed and known. In this section we discuss how to find suitable values of δ_1 and δ_2 such that the mean square error for the model (3) is minimum along with the unknown smooth covariate, X, should be approximate as close as possible.

To choose smoothing parameters that minimize the mean square error (prediction error), three common approaches (a) Generalized cross validation (GCV); (b) Mallow's C_p; and (c) Akaike Information criterion (AIC) have been discussed in the literature (Ruppert et al. 2003). Among these methods, minimization of GCV is more attractive because of its advantages in terms of invariance and computations (Wood, 2006).

In our simulations, we found that minimizing GCV scores for both δ_1 and δ_2 result in biased estimates of the regression coefficient, $\widehat{\beta}_1$. Our simulation also suggest that (see next section) estimates obtained by minimizing GCV scores for δ_1 and minimizing the following criteria for δ_2

$$H(\delta_2) = \frac{n^{-1}\sum_{i=1}^{n}\{W_i - \widehat{W}_i\}^4}{\{1 - n^{-1}\text{trace}\{\widehat{L}_2(\delta_2)\}\}^2}$$

where L_2 is the smoother matrix defined in section 2.4, works quite well in estimating the regression parameter, $\widehat{\beta}_1$. That is, we first obtain an estimate of δ_2 by minimizing

$H(\delta_2)$ and then substitute this value of δ_2 in (5) to get an estimate of θ_2. We then use these estimates of $\widehat{\delta}_2$ and $\widehat{\theta}_2$ in (10) to obtain an expression for $L_1(\delta_1)$. Finally, we minimize the standard GCV score to get an optimum value of δ_1, i.e., we minimize

$$GCV(\delta_1) = \frac{n^{-1}\sum_{i=1}^{n}\{Y_i - \widehat{Y}_i\}^2}{\{1 - n^{-1}\mathrm{trace}\{\widehat{L}_1(\delta_1, \widehat{\delta}_2)\}\}^2};$$

where L_1 is defined in section 2.4. The resulting (δ_1, δ_2) combination appears to work well in terms of estimating β_1 as seen in the next section.

3. Simulation

In this section we will discuss a simulation study designed to evaluate the finite sample properties of our proposed method in the presence of covariate measurement error in spatial linear regression.

3.1 Data generation

We simulated 500 sample locations randomly within an 80 × 80 rectangular grid. Specifically, the i^{th} random sample location S_i was generated by simulating two coordinates (e.g., latitude and longitude) from a Uniform[0,80] distribution. Given a set of simulated S_i's, the unobserved true covariate X was generated with mean 0 and covariance matrix Σ_X, where Σ_X was assumed to have an exponential correlation structure with unit variance. This implies that the correlation between two observations with distance h units apart is $exp(-h/\tau_x)$, where τ_x is the range parameter. We considered three different range parameters ($\tau_x = 1, 5, 10$) resulting in minimal, moderate and high correlation among the values of X's.

The observed error-prone versions, W, of the true covariate were generated by adding independent Gaussian noise with variance σ_U^2 to X. Outcome data, Y, were then generated according to equation (1), the slope and intercept parameter are taken as $(\beta_0, \beta_1)^T = (1, 2)^T$ and the variance component was generated using a similar exponential correlation structure as Σ_X, but with different range parameters. We also add a random Gaussian noise to the residual error variance (a so called nugget effect, Clark 2009). The variance parameter for this independent residual error was taken as 0.5.

We incorporate spatial correlation in the residual error by using a similar range parameters as of covariate X. Thus, we have 9 different combination of range parameters $(\tau_X, \tau_{\boldsymbol{\epsilon}})$ allowing for variation in the degree of spatial correlation in the covariates and in the error of the model Y on X. To generate exponential spatial correlation for our simulated data and in model fitting, we used the *nlme* package (Pinheiro et al., 2013). To extract the covariate matrices from the object corresponding to each lme fit we used the *mgcv* package (Wood, 2006) in ***R***.

3.2 Generating bi-variate splines basis function

We now describe the steps used to fit our proposed semi-parametric model. We generated two sets of basis function $B_1(\cdot)$ and $B_2(\cdot)$ using bivariate thin plate spline regression basis with 125 and 150 knots for response and covariate model, respectively. In particular, we used the *"tps.cov"* function (Ngo and Wand, 2004) in ***R*** to generate a bivariate thin plate splines regression basis. The number of knots were different to make the model identifiable, see Section 2.1. The number of knots for the response model were analogous to the default number of knots [max{20,min(n/4,150)}] suggested by Ruppert et al. (2003). For the covariate model we increased the deafult number of knots by 20%. Given a fixed value of the number of knots, knot positions were automatically selected using the cluster separation method *"clara"* (Kaufman and Rousseeuw, 2005) in ***R*** (R Core Team, 2013). Specifically, this method selects k representative objects in the data set, where k is the number of knots. The remaining objects are then assigned to the nearest representative object to form a cluster. The representative objects are selected in such a way that the average distance of the representative objects to all other objects in the same cluster is minimized. These optimal representative objects are also known as *"medoid"*, which serve as knots for the splines basis functions.

3.3 Performance of the proposed method

The average of the regression parameter estimates along with estimated standard errors from the 1000 replication are presented in Table 1, assuming a sample size of 500 and measurement error variance σ_U^2=0.2.

Table 1: Simulation results using different combinations of range parameters. Reported numbers are averaged over 1000 simulations with 500 observations per simulation and measurement error variance 0.2.

	Naive analysis			Estimated standard error		
Range*	LME	GAM	Proposed	Empirical†	Model based	Simulated
$(\tau_X, \tau_{\boldsymbol{\epsilon}})$	$\widehat{\beta}$	$\widehat{\beta}$	$\widehat{\beta}$	se($\widehat{\beta}$)	se($\widehat{\beta}$)	se($\widehat{\beta}$)
(1,1)	1.654	1.666	2.067	0.109	0.214	0.229
(1,5)	1.647	1.662	2.069	0.121	0.212	0.227
(1,10)	1.651	1.661	2.067	0.114	0.210	0.225
(5,1)	1.580	1.608	2.026	0.077	0.111	0.110
(5,5)	1.483	1.584	2.028	0.104	0.110	0.110
(5,10)	1.484	1.580	2.028	0.112	0.109	0.109
(10,1)	1.422	1.462	1.971	0.072	0.091	0.090
(10,5)	1.263	1.437	1.973	0.105	0.090	0.089
(10,10)	1.263	1.432	1.973	0.119	0.089	0.088

Range*- $(\tau_X, \tau_{\boldsymbol{\epsilon}})$values of the range parameter following exponential correlation in X and the error term in the model on Y respectively
Empirical† - Standard deviation of the 1000 simulated $\widehat{\beta}_1$'s.

Three different standard error estimates along with the average of estimated regression coefficients based on 1000 simulations are presented in Table 1. These include, empirical standard errors i.e., taking the sample standard deviation of the 1000 simulated regression coefficient estimates, average of the estimated standard errors and average of the simulated standard errors defined in section (2.4). The first column of table 1 specifies the combination of range parameters $(\tau_X, \tau_{\boldsymbol{\epsilon}})$ to characterize the 9 different combinations of spatial correlation in the covariate X and in the error for the model Y given X. The 2nd and 3rd columns list the naive estimates (i.e., ignoring measurement error) fitted with linear mixed models and generalized additive model, respectively. The fourth column presents the estimated regression parameters based on proposed method. The last three columns of this table presents the estimated standard error based on empirical calculation, model based and average of the simulated standard error. The simulated standard errors were obtained by taking 100 bootstrap samples. The standard errors were then averaged over 1000 replications.

Our results confirm that as expected, both the linear mixed model and generalized additive model attenuate the estimated regression coefficient towards the null hypothesis of no effect when an error prone covariate is used. Instead, the proposed bias correction method performs well even if the degree of bias for linear mixed model or generalized additive model with error prone covariate varies (range: 1.65-1.26 and 1.67-1.43, respectively) with the strength of the spatial correlation structure. Both model based and simulated estimates of the standard error are consistent when there is moderate to high correlation in the covariates. However, both of these standard error are over estimated when there is low correlation in the covariate X. This makes sense because the smooth spatial surface in X is non-identifiable in that setting.

To evaluate the performance of the proposed method under small samples, we also conducted simulations with sample size of 250 and 100. The results are given in Table 2.

Table 2: Simulation results using different combinations of range parameters and sample sizes. Reported numbers are averaged over 1000 simulations with measurement error variance 0.2.

	Sample size 250				Sample Size 100			
		Estimated standard error				Estimated standard error		
Range*	Coef.	I	II	III	Coef.	I	II	III
$(\tau_X, \tau_{\boldsymbol{\epsilon}})$	$\widehat{\beta}$	se($\widehat{\beta}$)	se($\widehat{\beta}$)	se($\widehat{\beta}$)	$\widehat{\beta}$	se($\widehat{\beta}$)	se($\widehat{\beta}$)	se($\widehat{\beta}$)
(1,1)	1.944	0.17	0.309	0.404	1.712	0.208	0.454	3.679
(1,5)	1.946	0.197	0.309	0.402	1.714	0.234	0.453	3.671
(1,10)	1.947	0.202	0.307	0.401	1.715	0.242	0.452	3.671
(5,1)	1.998	0.116	0.185	0.194	1.822	0.172	0.317	0.585
(5,5)	2.000	0.145	0.184	0.194	1.824	0.197	0.316	0.588
(5,10)	2.000	0.155	0.183	0.192	1.824	0.207	0.315	0.588
(10,1)	1.959	0.105	0.148	0.152	1.833	0.165	0.258	0.327
(10,5)	1.962	0.138	0.147	0.152	1.834	0.192	0.257	0.326
(10,10)	1.963	0.153	0.146	0.15	1.835	0.206	0.255	0.324

Range*- $(\tau_X, \tau_{\boldsymbol{\epsilon}})$values of the range parameter following exponential correlation in X and the error term in the model on Y respectively.
Estimated standard errors: I= Empirical, II= Model Based, III= Simulated.

With the size of 250 samples our proposed methods still provides very consistent estimates of the true regression coefficient. However, with small sample cases (say, n=100) the estimates are attenuated and variance becomes inflated.

4. Application

4.1 Analysis of Ischemic Heart Disease Data

We applied our proposed methodology to re-analyse data on Ischemic Heart Disease (IHD). One of the key objectives of the analysis is to ascertain whether there is any relationship between IHD rates with socio-economic status of the patient population. These data were collected from all hospitals in New South Wales (NSW), Australia between July 1, 1994 to June 30, 2002. A detailed description of the data has been given elsewhere (Burden et al., 2005). Briefly, patients who were admitted to the hospitals via the emergency room and discharged with IHD were defined as acute IHD cases. Data also includes patient age, gender and geographic location reported via postcode of residence. Data from 579 postcodes were included in the analysis. IHD event data were linked with the Census data which contains age and gender-specific population counts. SEIFA (Socio-Economic Indexes For Areas) scores and centroid co-ordinates (latitude and longitude) for each postcode were obtained from Australian Bureau of Statistics (ABS). We calculated age-sex adjusted standardized incidence ratios (SIR) by dividing the observed number of IHD cases by the age-sex adjusted expected IHD cases (Breslow and Day, 1987). Since SEIFA indexes are calculated using the principal component analysis which only accounts for about 30 percent of the total variation, it is likely that the SEIFA score is subject to substantial measurement error (Huque et al., 2014).

The results of our analysis are given in Table 3.

Table 3: Analysis of Ischemic Heart Disease Data in NSW, Australia under different specification of measurement error

Methods	Estimates for SEIFA		
		model based	simulated
	$\hat{\boldsymbol{\beta}}$	se($\hat{\boldsymbol{\beta}}$)	se($\hat{\boldsymbol{\beta}}$)
Ordinary Least Squares	-0.062	0.014	—
Generalized additive model	-0.145	0.014	—
Proposed semiparametric approach	-0.281	0.045	0.047
Huque et al. (2014) approach			
Method I: Method of Moments	-0.377	0.041	—
Method II: Transformation of covariate	-0.278	0.015	—

The naive analysis ignoring spatial correlation, suggests a significant protective effect associated with higher SEIFA values ($\hat{\boldsymbol{\beta}}_{SEIFA}$=-0.062, SE=0.014). Our proposed semi-parametric approach that account for measurement error in the covariates result in an estimated slope parameter β_1 of -0.281. The model based and simulated standard errors were estimated as 0.045 and 0.047, respectively. Thus, accounting for the measurement error in the covariate results in a very strong protective effect of higher SEIFA scores on Ischemic Heart Disease rates.

5. Discussion

In this paper, we develop a semi-parametric framework to obtain an unbiased estimate of the true regression coefficients when covariates are measured with error in spatial modelling settings. We develop an asymptotic theory for the model fit and provide a model based and simulation based standard error estimates. We show that our biased corrected estimate of the regression coefficient is consistent and asymptotically normal. Our empirical simulation results confirm that ignoring measurement error and conducting naive analysis using both generalized additive model and linear mixed model attenuates the estimated regression coefficient towards the null hypothesis of no effect. Our results concur with the results of Huque et al. (2014) who showed that the attenuation depends on the degree of spatial correlation in both $\boldsymbol{X}$ and the assumed random error from the regression model.

Our proposed semi-parametric bias correction method performs very well in obtaining consistent estimates. Our proposed method provides comparable estimates of the regression parameters to the transformation of covariate methods described by Huque et al. (2014) when applied to Ischemic Heart Disease (IHD) data. Our approach is computationally efficient and stable because it involves direct estimation using least squares and can be implemented using standard nonlinear least squares software.

Although Huque et al. (2014) and Li et al. (2009) reported similar results for the bias associated with regression analysis involving covariate measurement error, their approaches largely depends on the knowledge of the true covariate measurement error variance. Even though these authors proposed a sensitivity analysis to be carried out in practice, Huque et al. (2014) reported under estimation of standard error when measmeasurement error variances are estimated from the data. Additionally, the method proposed by Li et al. (2009) are based on E-M algorithm and difficult to apply, especially in situations involving large data sets. Moreover, both of these methods are based on linear mixed model formulation with no direct account to the generalized additive model settings.

Our proposed method is an important addition to the existing literature that addresses the issue of covariate measurement error in an additive model framework. This approach is also robust because it neither assumes that the covariate measurement error is known nor depends on any particular kind of spatial correlation structure. Our proposed method requires large samples (n $\approx$ 250 or more) to yield reliable results. In many applications, for examples, in air pollution data set, often sample size is quite large to effectively address the issues of measurement error with mixed model softwares due to the computational burden. In such situation, our proposed methods would be helpful for practitioners to obtain unbiased estimates and valid inferences.

Our heart disease example demonstrated a substantial increase in the rates of IHD as the level of SEIFA measured at the postcode level decreased. The magnitude of the effect increased after adjusting for measurement error. Our results are consistent with the result using similar analysis of Huque et al. (2014). The relationship between low socio-economic status and increased health outcome has also been observed in various social epidemiological research domains (see systematic review by Pickett and Pearl 2001). However, interpreting these results as applying at the individual level may result in ecological bias (Sheppard, 2003). Eventhough, caution is needed when interpreting group level covariates to the individuals level outcomes, in many research areas, group-level data are the only available source for analysis. Air pollution epidemiology provides a classic example, because individual measurements of air pollution studies are rarely collected and instead are estimated based on neighbourhood monitoring and other sources (Sheppard et al., 2012). Consequently, air pollution exposures are typically measured with error, and it would be useful to consider the impact of this error on subsequent effect size estimates.

In our simulation, we have considered only a single covariate measured with error in a spatial linear mixed model with Gaussian error. It would be of interest to explore the effect of covariate measurement error in the presence of multiple covariates and also omitted covariates. Future work should also consider extensions of our formulation to the setting of spatial generalized linear mixed model with non-Gaussian outcomes. However, such explorations are beyond the scope of this present paper.

In correlated data settings, for examples, in environmental epidemiology, with the increasing popularity of the semi parametric models/multilevel model to account for the observed data correlations, it is important that practitioners be aware of the consequences of measurement error. Furthermore, it is useful to quantify its effect on the exposure-outcome relationship prior to drawing potentially spurious conclusions regarding the relationship between the exposure of interest and outcome.

REFERENCES

Banerjee, S., Gelfand, A. E., and Carlin, B. P. (2004). *Hierarchical modeling and analysis 21 for spatial data*, Florida, U.S.A: Chapman and Hall/CRC.

Bernadinelli, L., Pascutto, C., Best, N., and Gilks, W. (1997). "Disease mapping with errors in covariates," *Statistics in Medicine*, 16, 741-752.

Breslow, N., and Day, N. (1987). *Statistical Methods in Cancer Research. Volume II–The Design and Analysis of Cohort Studies*, International Agency for Research on Cancer, New York, U.S.A.:Oxford University Press.

Breslow, N. E., and Clayton, D. G. (1993)." Approximate inference in generalized linear mixed models," *Journal of the American Statistical Association*, 88, 9-25.

Burden, S., Guha, S., Morgan, G., Ryan, L., Sparks, R., and Young, L. (2005). "Spatio-temporal analysis of acute admissions for ischemic heart disease in nsw, australia," *Environmental and Ecological Statistics*, 12, 427-448.

Carroll, R., Chen, R., George, E., Li, T., Newton, H., Schmiediche, H., and Wang, N. (1997). "Ozone exposure and population density in harris county, texas," *Journal of the American Statistical Association*, 92, 392-404.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective*, Florida, U.S.A.: Chapman and Hall/CRC.

Clark, I. (2009). " Statistics or geostatistics? Sampling error or nugget effect? ", *In Fourth World Conference on Sampling & Blending*, The Southern African Institute of Mining and Metallurgy, 13-18.

Clayton, D. G., Bernardinelli, L., and Montomoli, C. (1993). "Spatial correlation in ecological analysis", *International Journal of Epidemiology*, 22, 1193-1202.

Cook, D. G., and Pocock, S. J. (1983). "Multiple regression in geographical mortality studies, with allowance for spatially correlated errors", *Biometrics*, 39, 361-371.

Cressie, N. (1993). *Statistics for Spatial Data.* New York, U.S.A: Wiley.

Efron, B., and Tibshirani, R. (1986)." Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy", *Statistical Science*, 1, 54-75.

Elliott, P. and Wartenberg, D. (2004). "Spatial epidemiology: current approaches and future 10 challenges", *Environmental Health Perspectives*, 112, 998-1006.

Eilers, P. H., and Marx, B. D. (1996). "Flexible smoothing with b-splines and penalties," *Statistical science*, 11, 89-121.

Fuller, W. (1987).*Measurement Error Models.* New York, U.S.A.: John Wiley & Sons.

Gryparis, A., Paciorek, C. J., Zeka, A., Schwartz, J., and Coull, B. A. (2009). "Measurement error caused by spatial misalignment in environmental epidemiology," *Biostatistics*, 10, 258-274.

Guolo, A. (2008). "Robust techniques for measurement error correction: a review," *Statistical Methods in Medical Research*, 17, 555-580.

Huque, M. H., Bondell, H.D., and Ryan, L. (2014), " On the impact of covariate measurement error on spatial regression modelling," *Environmetrics*, 25, To appear

Kammann, E., and Wand, M. P. (2003). "Geoadditive models", it Journal of the Royal Statistical Society: Series C (Applied Statistics), 52, 1-18.

Kaufman, L., and Rousseeuw, P. (2005). *Finding Groups in Data: An Introduction to Cluster Analysis*, New Jersey, U.S.A.:John Wiley & Sons.

Li, Y., Tang, H., and Lin, X. (2009). "Spatial linear mixed models with covariate measurement errors".*Statistica Sinica* 19, 1077-1093.

Ngo, L. andWand, M. P. (2004). "Smoothing with mixed model software", *Journal of Statistical Software*, 9, 1-54.

Pickett, K. E., and Pearl, M. (2001). "Multilevel analysis of neighbouring socioeconomic context and health outcomes: a critical review". *Journal of Epidemiology & Community Health*,55,111-122.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team (2013).*nlme: Linear and Nonlinear Mixed Effects Models*, R package, version 3,1-109.

Prentice, R. L., and Sheppard, L. (1990). "Dietary fat and cancer: consistency of the epidemiologic data, and disease prevention that may follow from a practical reduction in fat consumption", *Cancer Causes & Control*, 1, 81-97.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Rothman, K.J., Greenland, S., and Lash, T.L. (2008). *Modern Epidemiology*, 3rd Edition. Philadelphia, PA: Lippincott, Williams & Wilkins.

Ruppert, D., Wand, M., and Carroll, R. J. (2009). "Semiparametric regression during 2003-2007".*Electronic Journal of Statistics*, 3, 1193-1256.

Ruppert, D., Wand, P., and Carroll, R. (2003). *Semiparametric Regression.* New York, U.S.A.: Cambridge University Press.

Sheppard, L. (2003). "Insights on bias and information in group-level studies", *Biostatistics*, 4, 265-278.

Sheppard, L., Burnett, R. T., Szpiro, A. A., Kim, S.-Y., Jerrett, M., Pope III, C. A., & Brunekreef, B. (2012). "Confounding and exposure measurement error in air pollution epidemiology Air Quality", *Atmosphere & Health*, 5, 203-216.

Snow, J. (1855). *On the mode of communication of cholera*, London, England: John Churchill.

Szpiro, A. A., Sheppard, L., and Lumley, T. (2011). "Efficient measurement error correction with spatially misaligned data", *Biostatistics*, 12, 610-623.

Waller, L. A., and Gotway, C. A. (2004). *Applied spatial statistics for public health data*, New Jersey, U.S.A.: John Wiley & Sons.

Wansbeek, T. J. and Meijer, E. (2000). *Measurement error and latent variables in econo-metrics*, North-Holland, Amsterdam: Elsevier.

Wood, S. (2006). *Generalized additive models: an introduction with R*,Florida, U.S.A.: Chapman and Hall/CRC.

Xia, H. and Carlin, B. P. (1998). "Spatio-temporal models with errors in covariates: mapping ohio lung cancer mortality", *Statistics in Medicine*, 17, 2025-2043.

Xun, X., Cao, J., Mallick, B. K., Maity, A., and Carroll, R. J. (2013). "Parameter estimation of partial differential equation models", *Journal of the American Statistical Association*, 108, 1009-1020.

Yu, Y. and Ruppert, D. (2002). "Penalized spline estimation for partially linear single-index models", *Journal of the American Statistical Association*, 97, 1042-1054.

Appendix

Appendix A

Consider that S_i's as fixed constants and recall that, $Y_i = G_2(S_i)\beta_1 + G_1(S_i) + \epsilon_i$. Substituting the expression for Y_i into the numerator of (7) and simplifying using the expression from (9), we have

$$\begin{aligned}
&n^{-1}\textstyle\sum_{i=1}^{n} Y_i\{B_2^{\mathrm{T}}(S_i) - B_1^{\mathrm{T}}(S_i)R_n\}\widehat{\theta}_2 \\
&= n^{-1}\textstyle\sum_{i=1}^{n}(B_2^{\mathrm{T}}(S_i)\theta_2\beta_1 + B_1^{\mathrm{T}}(S_i)\theta_1 + \epsilon_i)\{B_2^{\mathrm{T}}(S_i) - B_1^{\mathrm{T}}(S_i)R_n\}\widehat{\theta}_2 \\
&= n^{-1}\textstyle\sum_{i=1}^{n}(B_2^{\mathrm{T}}(S_i)\theta_2\beta_1 + B_1^{\mathrm{T}}(S_i)\theta_1)\{B_2^{\mathrm{T}}(S_i) - B_1^{\mathrm{T}}(S_i)R_n\}\widehat{\theta}_2 + \\
&n^{-1}\textstyle\sum_{i=1}^{n}\epsilon_i\{B_2^{\mathrm{T}}(S_i) - B_1^{\mathrm{T}}(S_i)R_n\}\widehat{\theta}_2 \\
&= \mathcal{A}_n\widehat{\theta}_2 + n^{-1}\textstyle\sum_{i=1}^{n}\mathcal{D}_{ni}\epsilon_i \\
&= \mathcal{A}_n\theta_2 + \mathcal{A}_n(\widehat{\theta}_2 - \theta_2) + n^{-1}\textstyle\sum_{i=1}^{n}\mathcal{D}_{ni}\epsilon_i.
\end{aligned}$$

Applying (5) to the above equation, we have

$$\mathcal{A}_n\theta_2 + n^{-1}\textstyle\sum_{i=1}^{n}\{\mathcal{A}_n\Lambda_{n2}B_2(S_i)U_i + \mathcal{D}_{ni}\epsilon_i\} + o_p(n^{-1/2}).$$

Again, the denominator of (7) is

$$\widehat{\theta}_2^{\mathrm{T}}(\mathcal{T}_n - R_n^{\mathrm{T}}\Lambda_{n1}^{-1}R_n)\widehat{\theta}_2 = \widehat{\theta}_2^{\mathrm{T}}\mathcal{C}_n\widehat{\theta}_2$$

Now applying (5), the denominator becomes,

$$\begin{aligned}
&(\theta_2 + n^{-1}\textstyle\sum_{i=1}^{n}\Lambda_{n2}B_2(S_i)U_i) + o_p(n^{-1/2})^{\mathrm{T}}\mathcal{C}_n(\theta_2 + n^{-1}\textstyle\sum_{i=1}^{n}\Lambda_{n2}B_2(S_i)U_i) + o_p(n^{-1/2}) \\
&= \theta_2^{\mathrm{T}}\mathcal{C}_n\theta_2 + n^{-1}\textstyle\sum_{i=1}^{n}\theta_2^{\mathrm{T}}\mathcal{C}_n\Lambda_{n2}B_2(S_i)U_i + n^{-1}\textstyle\sum_{i=1}^{n}U_i^{\mathrm{T}}B_2(S_i)^{\mathrm{T}}\Lambda_{n2}\mathcal{C}_n\theta_2 \\
&+(n^{-1}\textstyle\sum_{i=1}^{n}\Lambda_{n2}B_2(S_i)U_i)^{\mathrm{T}}(n^{-1}\textstyle\sum_{i=1}^{n}\Lambda_{n2}B_2(S_i)U_i) + o_p(n^{-1/2}) \\
&= \theta_2^{\mathrm{T}}\mathcal{C}_n\theta_2 + n^{-1}\textstyle\sum_{i=1}^{n}\mathcal{F}_{ni}U_i + o_p(n^{-1/2}).
\end{aligned}$$

Then, by a Taylor series expansion,

$$\begin{aligned}
\widehat{\beta}_1 &= \frac{\mathcal{A}_n\theta_2 + n^{-1}\sum_{i=1}^{n}\{\mathcal{A}_n\Lambda_{n2}B_2(S_i)U_i + \mathcal{D}_{ni}\epsilon_i\}}{\theta_2^{\mathrm{T}}\mathcal{C}_n\theta_2 + n^{-1}\sum_{i=1}^{n}\mathcal{F}_{ni}U_i} + o_p(n^{-1/2}) \\
&= \frac{\mathcal{A}_n\theta_2}{\theta_2^{\mathrm{T}}\mathcal{C}_n\theta_2} + \frac{n^{-1}\sum_{i=1}^{n}\{\mathcal{A}_n\Lambda_{n2}B_2(S_i)U_i + \mathcal{D}_{ni}\epsilon_i\}}{\theta_2^{\mathrm{T}}\mathcal{C}_n\theta_2} \\
&\qquad -\frac{\mathcal{A}_n\theta_2}{(\theta_2^{\mathrm{T}}\mathcal{C}_n\theta_2)^2}n^{-1}\textstyle\sum_{i=1}^{n}\mathcal{F}_{ni}U_i + o_p(n^{-1/2}).
\end{aligned}$$

Thus

$$\widehat{\beta}_1 - \frac{\mathcal{A}_n\theta_2}{\theta_2^{\mathrm{T}}\mathcal{C}_n\theta_2} = n^{-1}\textstyle\sum_{i=1}^{n}(\mathcal{G}_{ni}\epsilon_i + \mathcal{H}_{ni}U_i) + o_p(n^{-1/2}).$$

Now considering the fact from (5) that $n^{-1}\sum_{i=1}^{n}B_2(S_i)B_2^{\mathrm{T}}(S_i) = \Lambda_{n2}^{-1} + o(n^{-1/2})$ and using this in the expression for $\mathcal{A}_n\theta_2$, we have

$$\begin{aligned}
\mathcal{A}_n\theta_2 &= n^{-1}\textstyle\sum_{i=1}^{n}\beta_1\theta_2^{\mathrm{T}}B_2(S_i)\{B_2^{\mathrm{T}}(S_i) - B_1^{\mathrm{T}}(S_i)R_n\}\theta_2 \\
&\qquad +\theta_1^{\mathrm{T}}n^{-1}\textstyle\sum_{i=1}^{n}B_1(S_i)\{B_2^{\mathrm{T}}(S_i) - B_1^{\mathrm{T}}(S_i)R_n\}\theta_2 \\
&= \beta_1\theta_2^{\mathrm{T}}\mathcal{C}_n\theta_2 + \theta_1^{\mathrm{T}}\{\Lambda_{n1}^{-1} - n^{-1}\textstyle\sum_{i=1}^{n}B_1(S_i)B_1^{\mathrm{T}}(S_i)\}R_n\theta_2 \\
&= \beta_1\theta_2^{\mathrm{T}}\mathcal{C}_n\theta_2 + o_p(n^{-1/2}).
\end{aligned}$$

Therefore,

$$\frac{\mathcal{A}_n\theta_2}{\theta_2^{\mathrm{T}}\mathcal{C}_n\theta_2} = \beta_1 + o_p(n^{-1/2})$$

Hence,

$$n^{1/2}(\widehat{\beta}_1 - \frac{\mathcal{A}_n\theta_2}{\theta_2^{\mathrm{T}}\mathcal{C}_n\theta_2}) \sim \mathrm{Normal}(0, \sigma^2),$$

where $\sigma^2 = n^{-1}\sum_{i=1}^n(\sigma_\epsilon^2\mathcal{G}_{ni}^2 + \sigma_u^2\mathcal{H}_{ni}^2)$.

Thus, $\hat{\beta}_1$ is a consistent estimate for β_1.

Appendix B

From equation 3, we have

$$\begin{aligned}
\widehat{Y}_i &= B_2^{\mathrm{T}}(S_i)\widehat{\theta}_2\widehat{\beta}_1 + B_1^{\mathrm{T}}(S_i)\widehat{\theta}_1 \\
&= B_2^{\mathrm{T}}(S_i)\widehat{\theta}_2\widehat{\beta}_1 + B_1^{\mathrm{T}}(S_i)[V_n - R_n\widehat{\theta}_2\widehat{\beta}_1] \\
&= [B_2(S_i) - B_1^{\mathrm{T}}(S_i)R_n]^{\mathrm{T}}\widehat{\theta}_2\widehat{\beta}_2 + B_1^{\mathrm{T}}(S_i)V_n \\
&= [B_2(S_i) - B_1^{\mathrm{T}}(S_i)R_n]^{\mathrm{T}}\widehat{\theta}_2\left(\frac{n^{-1}\sum_{i=1}^n Y_i\{B_2^{\mathrm{T}}(S_i) - B_1^{\mathrm{T}}(S_i)R_n\}\widehat{\theta}_2}{\widehat{\theta}_2^{\mathrm{T}}\mathcal{C}_n\widehat{\theta}_2}\right) \\
&\qquad + B_1^{\mathrm{T}}(S_i)\Lambda_{n1}n^{-1}\sum_{i=1}^n B_1(S_i)Y_i \\
&= \frac{D_{ni}n^{-1}\sum_{i=1}^n D_{ni}Y_i}{\widehat{\theta}_2^{\mathrm{T}}\mathcal{C}_n\widehat{\theta}_2} + B_1^{\mathrm{T}}(S_i)\Lambda_{n1}n^{-1}\sum_{i=1}^n B_1(S_i)Y_i.
\end{aligned}$$

Similarly from equation 4, we have

$$\begin{aligned}
\widehat{W_i} &= B_2^{\mathrm{T}}(S_i)\widehat{\theta}_2 \\
&= B_2^{\mathrm{T}}(S_i)\Lambda_{n2}n^{-1}\sum_{i=1}^n B_2(S_i)W_i.
\end{aligned}$$