Journal of Statistical Research *https://doi.org/10.47302/jsr.2021550106*
2021, Vol. 55, No. 1, pp. 59-73 *ISSN 0256 - 422 X*

MODEL MISSPECIFICATION UNDER THE PARTIALLY LINEAR SINGLE INDEX MODEL

GRACE Y. YI*

Department of Statistical and Actuarial Sciences
Department of Computer Science, University of Western Ontario
1151 Richmond Street North, London, Ontario, Canada N6A 5B7

Email: gyi5@uwo.ca

WENQING HE

Department of Statistical and Actuarial Sciences, University of Western Ontario
1151 Richmond Street North, London, Ontario, Canada N6A 5B7

Email: whe@stats.uwo.ca

SUMMARY

The partially linear single index model has greater flexibility than linear regression models in facilitating the relationship between a continuous response and a set of covariates. This model allows not only linear dependence but also nonlinear dependence of the response variable on the covariates. Such a flexibility is, however, achieved at the price of losing the closed-form estimators of linear regression models. In this paper, we describe an estimation procedure using the spline approach to handle the nonlinear unknown function in the partially linear single index model. To explore the robustness of the partially linear single index model, we establish consistency results for the model parameters in the linear form under certain model misspecification. We identify several important settings with model misspecification where consistent results for the model parameters in the linear form are still retained. Those settings include cases with spurious covariates, covariates omission, covariate measurement error, and misspecfying the distribution of the noise term in the model. Further, we stress the importance of the independence assumption imposed for the noise term and the regressors, the assumption that is often overlooked in the literature. We illustrate, using an example of measurement error models, that the negligence of this independence assumption can yield biases results which would not be the case otherwise. Numerical studies confirm the satisfactory performance of the proposed method under a variety of settings.

*Keywords and phrases:*Covariate measurement error, covariate omission, model misspecification, partially linear single index model, spline approach, spurious covarites.

AMS Classification: 62F10, 62J10.

* Corresponding author

1 Introduction

Linear regression models have been widely used in applications to characterize the relationship between a continuous response and its associated covariates. Under such models, the least squares estimation method is commonly employed to estimate the model parameters, and the resulting estimators are of an analytical form.

Although linear regression models are convenient to use, they can be restrictive in that it only facilitates covariate effects in the linear form. When more sophisticated covariate effects arise from applications, the linear structure is inadequate to reflect the complex dependence of the response on the covariates. To allow for flexible dependence structures in modeling, various extensions of linear regression models were proposed by including a nonlinear function of covariates in modeling. To name a few, Härdle and Stoker (1989) and Powell et al. (1989) investigated single index models. Carroll et al (1997) and Xia and Härdle (2006) studied generalized partially linear single index models. Lu et al. (2006) considered the partially linear single index model for survival data. Cai et al. (2007) investigated partially linear regression structures for multivariate survival data. He and Yi (2020) examined partially linear single index accelerated failure time (AFT) models.

Including a nonlinear function of covariates in the modeling offers more flexibility to delineate the relationship between the response and covariates. This flexibility, however, comes at the price of the complications of developing inferential procedures. Usual estimation procedures for linear regression models cannot be directly applied and the resulting estimators usually do not have a closed-form like the least squares estimators for linear regression models. In this paper, we consider the partially linear single index model and describe an estimation procedure using the spline method to approximate the nonlinear function in the model. The asymptotic distribution of the resulting estimators is studied. To further understand the performance of the estimation procedure, we explore the robustness of the proposed method to model misspecification both analytically and numerically.

This paper contains several contributions. First, we describe an easily implemented procedure for estimating the parameters in the partially linear single index model. Secondly, we examine model misspecification effects and establish consistency of the resulting estimators for the parameters in the linear form under certain scenarios. Thirdly, we identify useful scenarios of model misspecification, and our explorations offer new insights into settings with spurious covariates, covariates omission, covariate measurement error, and misspecfying the distribution of the noise term in the model. Finally, we stress the importance of the independence assumption between the noise term and the regressors, an assumption that is critical but often overlooked in the literature. We demonstrate that biased results would be produced if this assumption were to be ignored, which would not be the case otherwise.

The remainder of the paper is organized as follows. Basic notation and the partially linear single index model are introduced in Section 2. In Section 3, we describe an estimation procedure using the spline approach to handle the nonlinear function in the model, and study the asymptotic distribution of the resulting estimators. In Section 4, we explore robustness of the proposed estimation method under model misspecification. Numerical studies are reported in Section 5, and concluding remarks are given in the last section.

2 Notation and Model Setup

For $i = 1, 2, \ldots, n$, let Y_i denote the response variable for subject i, and let $\mathbf{x}_i$ and $\mathbf{z}_i$ denote the covariates of subject i, where the $\mathbf{x}_i$ are linearly related with Y_i, and the $\mathbf{z}_i$ are non-linearly associated with Y_i. Consider that the relationship between the response variable Y_i and the covariates is characterized by the *partially linear single index model*

$$Y_i = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i + \theta(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{z}_i) + \sigma e_i, \tag{2.1}$$

where e_i is the noise term with a given probability density function $f(e_i)$, $\sigma > 0$ is a scale parameter, $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are unknown regression parameters, and $\theta(\cdot)$ is an unknown smooth function.

The model form (2.1) covers commonly used regression models as special cases (e.g., Xia and Härdle 2006). Setting the $\theta(\cdot)$ function to be the identity function gives the linear regression model where all the covariates are linked with Y_i via the linear form. If $\boldsymbol{\beta}$ is constrained to be zero and the $\theta(\cdot)$ function is left unspecified, then model (2.1) recovers the single index model (e.g., Xia 2006).

While model (2.1) delineates various types of covariate structures, parameter $\boldsymbol{\alpha}$ is unidentifiable unless certain constraints are imposed on $\boldsymbol{\alpha}$. By convention, we assume $||\boldsymbol{\alpha}|| = 1$ and $\alpha_1 > 0$, where α_1 is the first coordinate of $\boldsymbol{\alpha}$ (e.g., Carroll et al. 1997; Yi, He and Liang 2009). Since $\theta(\cdot)$ is assumed unknown, an intercept can be accommodated in this function, therefore, no intercept appears in (2.1) explicitly.

Finally, we emphasize that an implicit but critical assumption is often required in specifying model (2.1). When $\{\mathbf{x}_i, \mathbf{z}_i\}$ are fixed by the design, no independence assumption between the noise term e_i and the regressors $\{\mathbf{x}_i, \mathbf{z}_i\}$ is needed. But when $\{\mathbf{x}_i, \mathbf{z}_i\}$ are treated as random variables, the noise term e_i needs to be assumed to be *independent* of the covariates $\{\mathbf{x}_i, \mathbf{z}_i\}$; if this independence assumption is not perceived feasible, then the marginal probability density function $f(e_i)$ in (2.1) needs to be replaced by the conditional probability density function of e_i, given $\{\mathbf{x}_i, \mathbf{z}_i\}$. This assumption becomes quite subtle in the presence of model misspecification, as illustrated in Section 4.

3 Inference Procedure

Let $\boldsymbol{\eta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\alpha}^{\mathrm{T}}, \sigma)^{\mathrm{T}}$. Write $m_i = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i + \theta(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{z}_i)$ and $e_i = (y_i - m_i)/\sigma$. Then the model form (2.1), together with the comment in the last paragraph of Section 2, gives that the probability density function for Y_i, given $\{\mathbf{x}_i, \mathbf{z}_i\}$, is

$$f_{y|(x,z)}(y_i|\mathbf{x}_i, \mathbf{z}_i) = \frac{1}{\sigma} f(e_i).$$

Suppose a random sample consists of the observed data $\{\{y_i, \mathbf{x}_i, \mathbf{z}_i\} : i = 1, \ldots, n\}$. Then the likelihood contributed from subject i is $L_i(\boldsymbol{\eta}) = \sigma^{-1} f(e_i)$, leading to the log-likelihood $\ell_i(\boldsymbol{\eta}) =$

$\log f(e_i) - \log\sigma$. Therefore, the score functions contributed from subject i are given by

$$\begin{aligned}
\frac{\partial \ell_i(\boldsymbol{\eta})}{\partial \boldsymbol{\beta}} &= -\frac{f'(e_i)}{f(e_i)} \times \frac{\mathbf{x}_i}{\sigma}, \\
\frac{\partial \ell_i(\boldsymbol{\eta})}{\partial \boldsymbol{\alpha}} &= -\frac{f'(e_i)}{f(e_i)} \times \frac{\theta'(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{z}_i)\mathbf{z}_i}{\sigma}, \\
\frac{\partial \ell_i(\boldsymbol{\eta})}{\partial \sigma} &= -\frac{f'(e_i)}{f(e_i)} \times \frac{e_i}{\sigma} - \frac{1}{\sigma},
\end{aligned} \tag{3.1}$$

where $f'(\cdot)$ and $\theta'(\cdot)$ denote the derivatives of $f(\cdot)$ and $\theta(\cdot)$, respectively. Define

$$\frac{\partial \ell_i(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = \left(\frac{\partial \ell_i(\boldsymbol{\eta})}{\partial \boldsymbol{\beta}^{\mathrm{T}}}, \frac{\partial \ell_i(\boldsymbol{\eta})}{\partial \boldsymbol{\alpha}^{\mathrm{T}}}, \frac{\partial \ell_i(\boldsymbol{\eta})}{\partial \sigma} \right)^{\mathrm{T}}.$$

If $\theta(\cdot)$ were known, estimation of the parameters $\boldsymbol{\eta}$ would proceed directly by solving the equations

$$\sum_{i=1}^{n} \frac{\partial \ell_i(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = 0 \tag{3.2}$$

for $\boldsymbol{\eta}$. Here and elsewhere, 0 represents a zero vector, a zero matrix , or the real number zero whose meaning is clear from the context. Function $\theta(\cdot)$ is, however, unknown, so directly working with (3.2) is impossible to obtain an estimate of parameter $\boldsymbol{\eta}$.

To handle the unknown function $\theta(\cdot)$, we use the spline approach to specify function $\theta(\cdot)$ as a linear combination of a finite number of basis spline functions:

$$\theta(u) = \sum_{k=1}^{r} \nu_k M_k(u; \zeta, J), \tag{3.3}$$

where the $M_k(u; \zeta, J)$ are piecewise polynomial functions with order J, defined on the knot sequence ζ with K interior knots; the ν_k are the parameters to be estimated; and r is the number of basis functions, determined by the knot sequence ζ and the order J (e.g., He and Yi 2020).

With expression (3.3), we employ the likelihood method to estimate the associated model parameters. Let $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_r)^{\mathrm{T}}$, $\boldsymbol{\gamma} = (\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\nu}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}}, \sigma)^{\mathrm{T}}$, and $m_{i\mathrm{P}} = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i + \sum_{k=1}^{r} \nu_k M_k(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{z}_i; \zeta, J)$. The log-likelihood of the sample data is then given by

$$\ell(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\gamma}),$$

where

$$\ell_i(\boldsymbol{\gamma}) = f\left(\frac{y_i - m_{i\mathrm{P}}}{\sigma} \right) - \log\sigma.$$

Maximizing $\ell(\boldsymbol{\gamma})$ with respect to parameter $\boldsymbol{\gamma}$ gives an estimate of $\boldsymbol{\gamma}$. Let $\widehat{\boldsymbol{\gamma}} = (\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}, \widehat{\boldsymbol{\nu}}^{\mathrm{T}}, \widehat{\boldsymbol{\beta}}^{\mathrm{T}}, \widehat{\sigma})^{\mathrm{T}}$ denote the resulting estimator of $\boldsymbol{\gamma}$.

Adapting the derivations of He and Yi (2020), we readily show that under regularity conditions, $\sqrt{n}(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})$ has the asymptotical normal distribution with mean zero and a covariance matrix $(\Gamma\Sigma)^{-1}(\Gamma\Sigma\Gamma^{\mathrm{T}})(\Gamma\Sigma)^{-1\mathrm{T}}$, where

$$\Sigma = E\left\{-\frac{\partial^2 \ell_i(\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^{\mathrm{T}}}\right\} \quad \text{and} \quad \Gamma = \begin{pmatrix} I_q - \boldsymbol{\alpha}\boldsymbol{\alpha}^{\mathrm{T}} & 0 \\ 0 & I_p \end{pmatrix}.$$

Here q and p are the dimension of $\boldsymbol{\alpha}$ and of $(\boldsymbol{\nu}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}}, \sigma)^{\mathrm{T}}$, respectively; and I_a stands for the $a \times a$ identity matrix for a positive integer a.

Finally, we comment that this estimation method hinges on the choice of basis spline functions and the knot sequence. While variation due to the selection of knots and basis spline functions is not built into the estimation, this method has been broadly used in the literature because of the appeal of easy implementation. One may view this inference procedure as conditional analysis, given the specified basis spline functions and the knot sequence. Quadratic or cubic spline functions with $J = 3$ or 4 are usually a viable choice for many applications. The number of interior knots is often set on the scale of $O(n^{1/5})$ for a quadratic spline approach, and the knot positions are determined so that the data fall in each interval with roughly equally probabilities. Discussions on these aspects can be found in Wang et al. (2014) and He and Yi (2020), among others.

4 Model Misspecification

4.1 Analytic development

The validity of the estimation method described in Section 3 relies on the correct specification of the distribution form for the noise term e_i. It is thereby important to understand how the method may be affected when the distribution of e_i is misspecified.

For the technical reason, here we treat $\mathbf{x}_i$ as centered random variables with $E(\mathbf{x}_i) = 0$ (the $\mathbf{z}_i$ are treated as random variables as well), in the same lines as Gould and Lawless (1998) and He and Lawless (2005). Assume that the true model generating the data is given by (2.1) where the probability density function of the noise term e_i is $f(e_i)$. When fitting the data, a working model is specified as

$$Y_i = \boldsymbol{\beta}^{*\mathrm{T}}\mathbf{x}_i + \theta^*(\boldsymbol{\alpha}^{*\mathrm{T}}\mathbf{z}_i) + \sigma^* e_i^*, \tag{4.1}$$

where the asterisk is added to the symbols to show their possible difference from the corresponding quantities in the true model (2.1). Here the noise term e_i^* is independent of $\{\mathbf{x}_i, \mathbf{z}_i\}$, the distribution of e_i^* is misspeicified as $f^*(e_i^*)$, and $\theta^*(\cdot)$ is a *user-specified* function.

Let $\boldsymbol{\eta}^* = (\boldsymbol{\beta}^{*\mathrm{T}}, \boldsymbol{\alpha}^{*\mathrm{T}}, \sigma^*)^{\mathrm{T}}$. Write $e_i^* = \big(y_i - \boldsymbol{\beta}^{*\mathrm{T}}\mathbf{x}_i - \theta^*(\boldsymbol{\alpha}^{*\mathrm{T}}\mathbf{z}_i)\big)/\sigma^*$ and $\phi(e_i^*) = \log f^*(e_i^*)$. With the independence assumption between the noise term e_i^* and $\{\mathbf{x}_i, \mathbf{z}_i\}$ in the working model (4.1) and using the result of Yi (2017, Problem 5.6(a)), we obtain that the log-likelihood contributed from subject i from the working model (4.1) is

$$\ell_i^*(\boldsymbol{\eta}^*) = -\log\sigma^* + \phi(e_i^*).$$

Maximizing $\sum_{i=1}^{n} \ell_i^*(\boldsymbol{\eta}^*)$ with respect to the parameter $\boldsymbol{\eta}^*$ yields the estimate $\widehat{\boldsymbol{\eta}}^* = (\widehat{\boldsymbol{\beta}}^{*\mathrm{T}}, \widehat{\boldsymbol{\alpha}}^{*\mathrm{T}}, \widehat{\sigma}^*)^{\mathrm{T}}$.

Adapting the arguments of White (1982), He and Lawless (2005) and Yi and Reid (2010), we can show that under the assumptions A1 - A3 of White (1982), the estimator $\hat{\boldsymbol{\eta}}^*$ converges in probability to a unique limit $\boldsymbol{\eta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\alpha}^{\mathrm{T}}, \sigma)^{\mathrm{T}}$ which is the solution to the equation

$$E_T \left\{ \frac{\partial \ell_i^*(\boldsymbol{\eta}^*)}{\partial \boldsymbol{\eta}^*} \right\} = 0, \tag{4.2}$$

where the expectation E_T is taken with respect to the distribution under the true model (2.1). Specifically,

$$\begin{aligned} E_T \left\{ \frac{\partial \ell_i^*(\boldsymbol{\eta}^*)}{\partial \boldsymbol{\beta}^*} \right\} &= -\frac{1}{\sigma^*} E_T\{\mathbf{x}_i \phi'(e_i^*)\} = 0, \\ E_T \left\{ \frac{\partial \ell_i^*(\boldsymbol{\eta}^*)}{\partial \boldsymbol{\alpha}^*} \right\} &= -\frac{1}{\sigma^*} E_T\{\mathbf{z}_i \theta^{*\prime}(\boldsymbol{\alpha}^{*\mathrm{T}}\mathbf{z}_i)\phi'(e_i^*)\} = 0, \\ E_T \left\{ \frac{\partial \ell_i^*}{\partial \sigma^*} \right\} &= -\frac{1}{\sigma^*} E_T\{e_i^* \phi'(e_i^*) + 1\} = 0. \end{aligned} \tag{4.3}$$

Now we examine what parameter value would be a solution of (4.3). Subtracting $\boldsymbol{\beta}^{*\mathrm{T}}\mathbf{x}_i + \theta^*(\boldsymbol{\alpha}^{*\mathrm{T}}\mathbf{z}_i)$ from both sides of the true model (2.1), we obtain that

$$Y_i - \boldsymbol{\beta}^{*\mathrm{T}}\mathbf{x}_i - \theta^*(\boldsymbol{\alpha}^{*\mathrm{T}}\mathbf{z}_i) = (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^{\mathrm{T}}\mathbf{x}_i + \{\theta^*(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{z}_i) - \theta(\boldsymbol{\alpha}^{*\mathrm{T}}\mathbf{z}_i)\} + \sigma e_i.$$

Combining this with (4.1) gives

$$e_i^* = \frac{(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^{\mathrm{T}}\mathbf{x}_i + \{\theta(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{z}_i) - \theta^*(\boldsymbol{\alpha}^{*\mathrm{T}}\mathbf{z}_i)\} + \sigma e_i}{\sigma^*}. \tag{4.4}$$

Next, we examine that

$$\begin{aligned} E_T \left\{ \frac{\partial \ell_i^*(\boldsymbol{\eta}^*)}{\partial \boldsymbol{\beta}^*} \right\} &= E_{\mathbf{x}_i} \left\{ E_{\mathbf{z}_i|\mathbf{x}_i} \left(E_{Y_i|(\mathbf{x}_i,\mathbf{z}_i)} \left[-\frac{1}{\sigma^*}\{\mathbf{x}_i \phi'(e_i^*)\} \right] \right) \right\} \\ &= E_{\mathbf{x}_i} \left(E_{\mathbf{z}_i|\mathbf{x}_i} \left[\mathbf{x}_i E_{Y_i|(\mathbf{x}_i,\mathbf{z}_i)} \left\{ -\frac{1}{\sigma^*}\phi'(e_i^*) \right\} \right] \right), \end{aligned} \tag{4.5}$$

where the expectations $E_{Y_i|\mathbf{x}_i,\mathbf{z}_i}$, $E_{\mathbf{z}_i|\mathbf{x}_i}$ and $E_{\mathbf{x}_i}$ are evaluated with respect to the conditional distribution of Y_i given $\{\mathbf{x}_i, \mathbf{z}_i\}$, the conditional distribution of $\mathbf{z}_i$ given $\mathbf{x}_i$, and the marginal distribution of $\mathbf{x}_i$, respectively.

If $\boldsymbol{\beta}^* = \boldsymbol{\beta}$, then (4.4) shows that e_i^* is free of $\mathbf{x}_i$, and thus, (4.5) becomes

$$\begin{aligned} & E_{\mathbf{x}_i} \left(\mathbf{x}_i E_{\mathbf{z}_i|\mathbf{x}_i} \left[E_{Y_i|(\mathbf{x}_i,\mathbf{z}_i)} \left\{ -\frac{1}{\sigma^*}\phi'(e_i^*) \right\} \right] \right) \\ &= E_{\mathbf{x}_i}(\mathbf{x}_i) \cdot E_{\mathbf{z}_i} \left[E_{Y_i|\mathbf{z}_i} \left\{ -\frac{1}{\sigma^*}\phi'(e_i^*) \right\} \right] \\ &= 0, \end{aligned}$$

where the last step comes from that $\mathbf{x}_i$ is centralized to have zero expectation; the expectations $E_{Y_i|\mathbf{z}_i}$ and $E_{\mathbf{z}_i}$ are evaluated with respect to the conditional distribution of Y_i given $\mathbf{z}_i$ and the marginal distribution of $\mathbf{z}_i$, respectively.

Since $\boldsymbol{\beta}$ is a solution to Equation 4.3 and the solution to Equation 4.3 is unique (White, 1982), we conclude that

$$\boldsymbol{\beta}^* = \boldsymbol{\beta},$$

which implies that the estimator $\widehat{\boldsymbol{\beta}}^*$ obtained from the working model (4.1) is still a consistent estimator of $\boldsymbol{\beta}$.

Theorem 1. *Suppose that $E(\mathbf{x}_i) = 0$ and that there is a unique solution to (4.3), then the estimator produced by the misspecified working model (4.1) is still a consistent estimator for the parameter $\boldsymbol{\beta}$ of the true model (2.1).*

Model misspecification is generally expected to yield inconsistent estimation results, especially when using likelihood-based methods. But Theorem 1 uncovers a situation where consistency of estimating some model parameters is not affected when the distribution of the noise term e_i in the true model (2.1) is misspecfified. While the consistency is retained for estimation of the linear parameter $\boldsymbol{\beta}$ under the working model (4.1), this property does not necessarily hold for estimation of the nonlinear parameter $\boldsymbol{\alpha}$ or the scale parameter σ. It is key to retain the linearity in $\mathbf{x}_i$ and the additivity of the terms when specifying a working model to ensure consistent results for estimation of $\boldsymbol{\beta}$. Furthermore, the independence assumption for the noise term in a working model is also critical, as demonstrated at the end of the next subsection.

4.2 Useful settings of model misspecification

Theorem 1 enables us to use model (2.1) in a broader scope. If the primary interest is in inference about the linear parameter $\boldsymbol{\beta}$, one may take a working model to fit the data as long as it retains the additive structure of model (2.1) with $\mathbf{x}_i$ centered and appearing in linearity, together with the independence requirement for the noise term and the covaraites; the nonlinear function $\theta(\cdot)$ and the distribution of the noise term in the working model can be misspecified. We now examine several scenarios which are pertinent to misspecifying the nonlinear structure and/or the distribution of the noise term, spurious covariates, covariates omission, and covariates with measurement error.

Scenario 1: *Misspecifying the partially linear single index model as an usual linear regression model*

If the true model is (2.1), but we use the ordinary linear regression model (i.e., setting $\theta(\cdot)$ to be an identity function) as a working model to fit the data with $\mathbf{x}_i$ centered, then the resulting estimator of $\boldsymbol{\beta}$ is still consistent. This is because no matter what forms of $\theta(\cdot)$ and $\theta^*(\cdot)$ are, e_i^* in (4.5) is free of $\mathbf{x}_i$ when taking $\boldsymbol{\beta}^* = \boldsymbol{\beta}$; and this ensures a zero value of (4.5) in combination of $E(\mathbf{x}_i) = 0$.

Scenario 2: *Omitting covariates $\mathbf{z}_i$ in the partially linear single index model*

When applying model (2.1) to analyze data, we may neglect some important covariates and do not include them in the model. For example, consider a situation where all the $\mathbf{z}_i$ covariates are ignored, i.e., the working model is taken as

$$Y_i = \boldsymbol{\beta}^{*\mathrm{T}}\mathbf{x}_i + \sigma^* e_i^*,$$

with covariates $\mathbf{z}_i$ omitted, then the estimator $\widehat{\boldsymbol{\beta}}^*$ for the effects of covariate $\mathbf{x}_i$ is still consistent. This result complements to that obtained by Struthers and Kalbfleisch (1986) who considered misspeficiation under the proportional hazard model for censored data.

Scenario 3: *Including additionally unimportant covariates in the partially linear single index model*

Opposite to Scenario 2, we may fit the data by blindly including additionally unimportant covariates when using (2.1). A close examination of the derivations in Section 4.1 shows that the resulting estimator for $\boldsymbol{\beta}$ is still consistent, as long as the term $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i$ is in an additive relationship with other covariates and the noise term in the working model is assumed to be independent of the regressors.

Scenario 4: *Impact of Berkson measurement error on the partially linear single index model*

Consider a setting where $\mathbf{x}_i$ is subject to measurement error, and the $\mathbf{z}_i$ covariates may or may not contain measurement error. Suppose $\mathbf{x}_i^*$ is an observed measurement of $\mathbf{x}_i$ which is delineated by the relationship

$$\mathbf{x}_i = \mathbf{x}_i^* + \epsilon_i, \tag{4.6}$$

where error term ϵ_i is independent of $\{Y_i, \mathbf{x}_i^*, \mathbf{z}_i\}$ and has mean zero; this is the so-called Berkson error model used in the literature of measurement error problems (e.g, Carroll et al. 2006; Yi 2017).

If we disregard the difference between the observed measurement $\mathbf{x}_i^*$ and the true covariate value $\mathbf{x}_i$, and conduct a naive analysis by using model (2.1), then we essentially use a working model

$$Y_i = \boldsymbol{\beta}^{*\mathrm{T}}\mathbf{x}_i^* + \theta^*(\boldsymbol{\alpha}^{*\mathrm{T}}\mathbf{z}_i) + \sigma^* e_i^*, \tag{4.7}$$

where the asterisk is added to the symbols to show their possible differences from the corresponding parameters in the true model (2.1), and e_i^* is assumed to be independent of the regressors $\{\mathbf{x}_i^*, \mathbf{z}_i\}$.

In contrast, substituting (4.6) into the true model (2.1) gives

$$Y_i = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i^* + \theta(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{z}_i) + \sigma\widetilde{e}_i^*, \tag{4.8}$$

where $\widetilde{e}_i^* = \boldsymbol{\beta}^{\mathrm{T}}\epsilon_i/\sigma + e_i$, which is independent of the regressors $\{\mathbf{x}_i^*, \mathbf{z}_i\}$. Comparing (4.8) to the working model (4.7) shows that $\boldsymbol{\beta}^* = \boldsymbol{\beta}$, suggesting that using the working model (4.7) to the surrogate data $\{\{Y_i, \mathbf{x}_i^*, \mathbf{z}_i\} : i = 1, \ldots, n\}$ still yields a consistent estimator for $\boldsymbol{\beta}$. This result offers a new angle to view inference about data with measurement error from the perspective of model misspecification. It generalizes the discussion of the Berkson error effects under linear regression models (e.g., Carroll et al. 2006; Yi 2017) to partially linear single index models.

Caution: We emphasize the importance of the independence assumption between e_i and $\{\mathbf{x}_i, \mathbf{z}_i\}$ required in the working model (4.1). This condition is often implicitly imposed but not explicitly stated by many authors when discussing model misspecification. Here we illustrate that this assumption cannot be ignored; otherwise, incorrect conclusions would arise.

To be specific, we consider Scenario 4 with the Berkson model (4.6) replaced by the classical additive measurement error model (e.g., Carroll et al. 2006; Yi 2017):

$$\mathbf{x}_i^* = \mathbf{x}_i + \epsilon_i^*, \tag{4.9}$$

where the error term ϵ_i^* is independent of $\{Y_i, \mathbf{x}_i, \mathbf{z}_i\}$ and has mean zero.

Suppose that disregarding the difference between the observed measurement $\mathbf{x}_i^*$ and the true covariate value $\mathbf{x}_i$, we conduct a naive analysis by applying model (2.1) to the error-prone data $\{(Y_i, \mathbf{x}_i^*, \mathbf{z}_i) : i = 1, \ldots, n\}$:

$$Y_i = \boldsymbol{\beta}^{**\mathrm{T}}\mathbf{x}_i^* + \theta^{**}(\boldsymbol{\alpha}^{**\mathrm{T}}\mathbf{z}_i) + \sigma^{**}e_i^{**}, \tag{4.10}$$

where the double asterisk is added to the symbols to show their possible differences from the corresponding parameters in the true model (2.1), and e_i^{**} is assumed to be independent of the regressors $\{\mathbf{x}_i^*, \mathbf{z}_i\}$.

Substituting (4.9) into (4.10) gives a working model

$$Y_i = \boldsymbol{\beta}^{**\mathrm{T}}\mathbf{x}_i + \theta^{**}(\boldsymbol{\alpha}^{**\mathrm{T}}\mathbf{z}_i) + \sigma^{**}\widetilde{e}_i^{**}, \tag{4.11}$$

where $\widetilde{e}_i^{**} = \boldsymbol{\beta}^{**\mathrm{T}}\epsilon_i^*/\sigma^{**} + e_i^{**}$.

If we regarded model (4.11) as a working model of the form (4.1) by neglecting the requirement of the independence between the noise term and the regressors, then applying the derivations in Section 4.1 would yield that $\boldsymbol{\beta}^{**} = \boldsymbol{\beta}$, saying that fitting error-prone data $\{\{Y_i, \mathbf{x}_i^*, \mathbf{z}_i\} : i = 1, \ldots, n\}$ with the model (4.10) would still yield a consistent estimator of $\boldsymbol{\beta}$ in the true model (2.1).

This seemingly rational claim is, however, incorrect. The reason is that the term $\widetilde{e}_i^{**}$ in (4.11) is not ensured to be independent of the regressors $\{\mathbf{x}_i, \mathbf{z}_i\}$, a condition required by the working model (4.1). In the next section, we conduct numerical studies to demonstrate that fitting error-prone data $\{\{Y_i, \mathbf{x}_i^*, \mathbf{z}_i\} : i = 1, \ldots, n\}$ with the model (4.10) would output biased estimates of $\boldsymbol{\beta}$. This discussion and the numerical results to be reported for Scenario V in Section 5.2 align with the well known fact that ignoring classical additive error in covariates under the linear regression model produces biased estimates of the model parameters (e.g., Carroll et al. 2006; Yi 2017), and they further extend the development from the usual linear regression model to the partially linear single index model. The discussion here also gives us a new angle to view the independence assumption between the noise term and covariates, required when specifying a working model for inference, by rephrasing it in the framework of covariates subject to classical additive measurement error.

5 Numerical Studies

We conduct simulation studies to numerically evaluate the performance of the estimation method in Section 3 as well as the impact of model misspecification discussed in Section 4. Five hundred

simulations are run for each of the following parameter configurations, and the sample sizes $n = 100$ and $n = 500$ are considered.

Table 1: Simulation Results for Assessing the Performance of the Estimation Method

	$\theta(u) = \sin(\cdot)$						$\theta(u) = u$					
	$n = 100$			$n = 500$			$n = 100$			$n = 500$		
	Bias	SE	MSE	Bias	SE	MSE	Bias	SE	MSE	Bias	SE	MSE
β	0.004	0.085	0.007	-0.001	0.036	0.001	0.004	0.082	0.007	0.001	0.035	0.001
α_1	-0.055	0.239	0.060	-0.018	0.133	0.018	-0.046	0.187	0.037	-0.032	0.151	0.024
α_2	-0.055	0.237	0.059	-0.007	0.126	0.016	-0.039	0.197	0.040	-0.035	0.159	0.026
α_3	-0.039	0.230	0.054	-0.018	0.127	0.017	-0.019	0.209	0.044	-0.007	0.193	0.037

5.1 Performance of the proposed estimator

For $i = 1, \ldots, n$, covariate x_i is generated from the standard normal distribution $N(0, 1)$, and covariates z_{ij} are independently generated from the uniform distribution $U[0, 1]$ for $j = 1, 2, 3$. The response measurements are generated from the model

$$Y_i = \beta x_i + \theta(\alpha_1 z_{i1} + \alpha_2 z_{i2} + \alpha_3 z_{i3}) + \sigma e_i, \tag{5.1}$$

where the error distribution of e_i is taken as the standard extreme value with the cumulative distribution function $F(e_i) = 1 - \exp\{-\exp(e_i)\}$, and σ is set as 1.5. We set $\beta = 0.3$ and $\alpha_1 = \alpha_2 = \alpha_3 = 1/\sqrt{3}$, and consider two function forms for the $\theta(\cdot)$ function: $\theta(u) = u$, or $\theta(u) = \sin[\pi(u - 1.355\sqrt{3}/6)/(1.645\sqrt{3}/3)]$, as in Carroll et al. (1997).

We fit the simulated data to model (2.1) and apply the estimation method described in Section 3 to estimate the model parameters, where the $M_k(u; \zeta, J)$ in (3.3) are taken as M-spline basis functions. Specifically, setting the order $J = 3$ and the number of interior knots K to be 4 gives $r = K + J$ piecewise quadratic polynomial basis functions $M_k(u; \zeta, J)$ for (3.3). The four interior knots are determined by examining the numerical range of $\boldsymbol{\alpha}_i^{\mathrm{T}}\mathbf{z}_i$ to yield roughly the equal number of observations in each interval. To facilitate the constraints of $\|\boldsymbol{\alpha}\| = 1$ and $\alpha_1 > 0$ in the estimation procedure, we reparameterize the components of $\boldsymbol{\alpha}$ using the polar coordinate transformation described by He and Yi (2020), where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)^{\mathrm{T}}$. In all the simulations, initial values of β and $\boldsymbol{\alpha}$ are taken as the estimates obtained from the linear model with $\theta(\cdot)$ set as the identity function, and initial values for the parameters pertaining to estimation of $\theta(u)$ in (3.3) are set as zero.

Table 1 reports the simulation results for estimation of the parameters β and $\boldsymbol{\alpha}$, including the average difference between the estimates and the true values (Bias), the empirical standard deviation (SE) of the estimates, and the mean squared error (MSE). Estimation of β incurs very small finite sample biases regardless of the form of the $\theta(\cdot)$ function or the sample size, though a large sample size further reduces the finite sample biases slightly; as the sample size increases, both SE and MSE

decrease. Estimation of the $\boldsymbol{\alpha}$ parameters incurs larger finite sample biases than that of β, but the magnitudes of those biases still appear to fall in reasonable ranges. The form of the $\theta(\cdot)$ function and the sample size have noticeable effects on both Bias and SE (and hence, on MSE as well) for estimation of the $\boldsymbol{\alpha}$ parameters. As the sample size increases, the estimation results of the $\boldsymbol{\alpha}$ parameters improve.

In summary, the estimation method described in Section 3 performs well, regardless of the linear or nonlinear form of the smooth function $\theta(\cdot)$. As expected, the performance gets better as the sample size increases.

5.2 Evaluation with model misspecification

To confirm the consistency results of estimating the β parameter in the presence of model misspecification discussed in Section 4, we conduct simulation studies for five scenarios corresponding to those discussed in Section 4.2.

In each scenario we consider the combinations of different specification of the distribution for the noise term and the form of the single index function $\theta(\cdot)$. In the data generation step, we set the scale parameter σ to be 0.83 and 1.50, respectively, when using the standard logistic and the standard extreme value distributions for the noise term. The single index function $\theta(\cdot)$ is set to be either the sine or the identity function, as described in Section 5.1. The values of $\boldsymbol{\alpha}$ and β and the sample size are set as those in Section 5.1.

In Scenario I, we generate covariates x_i and the z_{ij} in the same way as in Section 5.1. Then we generate response measurements from the model (5.1) by letting the noise term e_i follow the standard logistic distribution. But we fit the simulated data using the model (5.1) with the noise term e_i assumed to follow the standard extreme value distribution.

In Scenario II, we consider the possibility of omitting some covariates when fitting the model. Specifically, we generate covariates x_i and the z_{ij} in the same way as in Section 5.1. Next, we independently generate a new covariate, denoted w_i, from the standard normal distribution $N(0,1)$ for $i = 1, \ldots, n$. Then we generate response measurements from the model

$$Y_i = \beta x_i + \theta(\alpha_1 z_{i1} + \alpha_2 z_{i2} + \alpha_3 z_{i3}) + \beta_w w_i + \sigma e_i,$$

where $\beta_w = 0.25$; the function $\theta(\cdot)$, σ, and the distribution of the noise term e_i are specified as in the aforementioned descriptions. We fit the simulated data using the model (5.1) with the noise term e_i assumed to follow the standard extreme value distribution.

In Scenario III, we generate covariates x_i in the same way as in Section 5.1. Then we simulate response measurements from the model (5.1) with $\theta(\cdot) = 0$, where σ and the distribution of the noise term e_i are specified as in the aforementioned descriptions. But we fit the simulated data using the model (5.1) with the noise term e_i assumed to follow the standard extreme value distribution.

In Scenario IV, we first generate x_i^* from $N(0,1)$ for $i = 1, \ldots, n$, and then generate x_i from the Berkson model $x_i = x_i^* + u_i^*$, where u_i^* is independent of x_i^* and other variables, and $u_i^* \sim N(0, \sigma_e^{*2})$ with $\sigma_e^* = 0.25, 0.50$ or 1.00 to reflect an increasing degree of measurement error. The covariates z_{ij} are independently generated in the same way as in Section 5.1. Then we simulate response measurements in the same way as in Scenario 1 with the aforementioned specification of

the function $\theta(\cdot)$, the distribution of noise term e_i, and the σ value. But we fit the model (5.1) with the noise term e_i assumed to follow the standard extreme value distribution to the data $\{\{Y_i, x_i^*, \mathbf{z}_i\} : i = 1, \ldots, n\}$. This reflects a case where the covariate x_i is subject to measurement error which follows a Berkson model, but such an error is ignored when fitting the model.

In Scenario V, we generate covariates x_i and z_{ij} in the same way as in Section 5.1. Then we generate a surrogate measurement x_i^* of x_i using the model $x_i^* = x_i + u_i$, where u_i is independent of x_i and other variables, and $u_i \sim N(0, \sigma_u^2)$ with $\sigma_u = 0.25, 0.50$ or 1.00 to reflect an increasing degree of measurement error. Response measurements are generated in the same way as in Scenario IV. But we fit the model (5.1) with e_i assumed to follow the standard extreme value distribution to the data $\{\{Y_i, x_i^*, \mathbf{z}_i\} : i = 1, \ldots, n\}$. This scenario reflects a case considered at the end of Section 4.2, where we illustrate the importance of imposing the independence assumption between the noise term and the regressors when using the working model (4.1).

Table 2 reports the simulation results for estimation of the β parameter, including the average difference between the estimates and the true values (Bias), the empirical standard deviation (SE) of the estimates, and the mean squared error (MSE). As discussed in Section 4.2, under model misspecification considered in Scenarios I-IV, using the working model (4.1) still yields a consistent estimator of β, and this is confirmed by the very small finite sample biases of the estimates of β. As expected, when the sample size increases, biases of estimating β tends to decrease, together with decreasing empirical standard errors and mean squared errors. Regarding the results for Scenario V, finite sample biases are reasonably small when the measurement error degree is minor (e.g., $\sigma_u = 0.25$). But when the magnitude of measurement error becomes moderate or large, considerable finite sample biases are observed, which clearly demonstrates the point emphasized at the end of Section 4.2.

6 Discussion

While the linear regression model is widely used in applications to facilitate the dependence of the response on associated covariates, such a model is inadequate to accommodating nonlinear dependence structures. Consequently, partially linear single index models become useful due to its ability of incorporating both linear and nonlinear relationship between the response and covariates. In this paper, we describe a method using the spline technique to estimate the model parameters and establish the asymptotic result of the induced estimators. The validity and performance of the method are further assessed for model misspecification. We identify a number of scenarios for obtaining a consistent estimator of the linear model parameter $\boldsymbol{\beta}$ even in the presence of model misspecification. The findings have important implications which enlarge the usage scope of the partially linear single index model.

It is interesting to note that the discussion in Section 4 applies to the partially single index AFT models considered by He and Yi (2020) if the censoring percentage is zero. When the censoring proportion is nonzero, using model (2.1) to handle survival data is generally vulnerable to model misspecification (where the response variable is set as the logarithm of a survival time); misspecification for the distribution of the noise term e_i usually yields inconsistent estimation of $\boldsymbol{\beta}$. It may be

Table 2: Simulation Results for Estimation of β with Misspecification of Models

	Data Generation		$n = 100$			$n = 500$		
Scenarios	$\theta(u)$	Distribution	Bias	SE	MSE	Bias	SE	MSE
I	$\sin(\cdot)$	Logistic	0.008	0.204	0.042	0.002	0.115	0.013
	u	Logistic	-0.012	0.208	0.043	0.004	0.113	0.013
II	$\sin(\cdot)$	Logistic	-0.004	0.208	0.043	0.005	0.110	0.012
	u	Ext Val	-0.002	0.083	0.007	-0.001	0.038	0.001
	u	Logistic	-0.003	0.211	0.045	0.005	0.115	0.013
III	0	Ext Val	0.000	0.072	0.005	0.001	0.030	0.001
	0	Logistic	0.020	0.203	0.041	0.008	0.110	0.012
IV($\sigma_e = 0.25$)	$\sin(\cdot)$	Logistic	0.009	0.204	0.042	0.007	0.110	0.012
	u	Ext Val	0.005	0.081	0.007	0.001	0.035	0.001
	u	Logistic	0.011	0.204	0.042	0.008	0.110	0.012
IV($\sigma_e = 0.50$)	$\sin(\cdot)$	Logistic	0.005	0.222	0.049	0.002	0.116	0.013
	u	Ext Val	0.007	0.083	0.007	0.002	0.038	0.001
	u	Logistic	0.010	0.203	0.041	0.004	0.114	0.013
IV($\sigma_e = 1.00$)	$\sin(\cdot)$	Logistic	0.005	0.226	0.051	0.001	0.111	0.012
	u	Ext Val	0.003	0.095	0.009	0.000	0.035	0.001
	u	Logistic	0.007	0.218	0.047	0.002	0.109	0.012
V($\sigma_u = 0.25$)	$\sin(\cdot)$	Logistic	-0.011	0.198	0.039	-0.017	0.107	0.012
	u	Ext Val	-0.002	0.077	0.006	-0.019	0.034	0.001
	u	Logistic	-0.010	0.194	0.038	-0.017	0.104	0.011
V($\sigma_u = 0.50$)	$\sin(\cdot)$	Logistic	-0.055	0.184	0.037	-0.058	0.101	0.014
	u	Ext Val	-0.057	0.072	0.008	-0.059	0.034	0.005
	u	Logistic	-0.053	0.179	0.035	-0.057	0.100	0.013
V($\sigma_u = 1.00$)	$\sin(\cdot)$	Logistic	-0.155	0.155	0.048	-0.149	0.078	0.028
	u	Ext Val	-0.151	0.061	0.027	-0.149	0.027	0.023
	u	Logistic	-0.145	0.147	0.043	-0.148	0.079	0.028

interesting to create pseudo-responses by synthesizing censored observations (e.g., using the discussion by Lu and Cheng 2007), and then investigate how censoring and model misspecification may interplay.

Acknowledgements

This research is partially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC). Yi is Canada Research Chair in Data Science (Tier 1). Her research was undertaken, in part, thanks to funding from the Canada Research Chairs Program.

References

Cai,J., Fan, J., Jiang, J, and Zhou, H. (2007), "Partially linear hazard regression for multivariate survival data," *Journal of the American Statistical Association*, 102, 538-551.

Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. (1997), "Generalized partially linear single-index models," *Journal of the American Statistical Association*, 92, 477-489.

Gould, A. and Lawless, J. F. (1988), "Consistency and efficiency of regression coefficient estimates in location-scale models," *Biometrika*, 75, 535-540.

Härdle, W. and Stoker, T. M. (1989), "Investigating smooth multiple regression by the method of average derivative," *Journal of the American Statistical Association*, 84, 986-995.

He, W. and Lawless, J. F. (2005), "Bivariate location-scale models for regression analysis, with applications to lifetime data," *Journal of Royal Statistical Society, (Ser. B)*, 67, 63-78.

He, W. and Yi, G. Y. (2020), "Parametric and semiparametric estimation methods for survival data under a flexible class of models," *Lifetime Data Analysis*, 26, 369-388.

Lu, X., Chen, G., Song, X.-K. and Singh, R. S. (2006), "A class of partially linear single-index survival models," *Canadian Journal of Statistics*, 34, 99-116.

Lu, X. and Cheng, T. (2007), "Randomly censored partially linear single-index models," *Journal of Multivariate Analysis*, 98, 1895-1922.

Powell, J. L., Stock, J. H. and Stoker, T. M. (1989), "Semiparametric estimation of index coefficients," *Econometrica*, 57, 1403-1430.

Struthers, C. A. and Kalbfleisch, J. D. (1986), "Misspecified proportional hazard models," *Biometrika*, 73, 363-369.

Wang, L., Xue, L., Qu, A. and Liang, H. (2014), "Estimation and model selection in generalized additive partial linear models for correlated data with diverging number of covariates," *The Annals of Statistics*, 42, 592-624.

White, H. (1982), "Maximum likelihood estimation of misspecified models," *Econometrica*, 50, 1-25.

Xia, Y. (2006), "Asymptotic distributions for two estimators of the single-index model," *Econ. Theory*, 22, 1112-1137.

Xia, Y. and Härdle, W. (2006), "Semi-parametric estimation of partially linear single-index models," *Journal of Multivariate Analysis*, 97, 1162-1184.

Yi, G. Y. (2017), *Statistical Analysis with Measurement Error or Misclassification: Strategy, Method and Application*. Springer.

Yi, G. Y., He, W., and Liang, H. (2009), "Analysis of correlated binary data under partially linear single-index logistic models," *Journal of Multivariate Analysis*, 100, 278-290.

Yi, G. Y. and Reid, N. (2010), "A note on mis-specified estimating functions," *Statistica Sinica*, 20, 1749-1769.

Received: March 2, 2021

Accepted: April 7, 2021