Ann Inst Stat Math (2019) 71:1–28
https://doi.org/10.1007/s10463-017-0622-0

Semiparametric efficient estimators in heteroscedastic error models

Mijeong Kim[1] · **Yanyuan Ma**[2]

Received: 8 November 2016 / Revised: 20 September 2017 / Published online: 13 October 2017

Abstract In the mean regression context, this study considers several frequently encountered heteroscedastic error models where the regression mean and variance functions are specified up to certain parameters. An important point we note through a series of analyses is that different assumptions on standardized regression errors yield quite different efficiency bounds for the corresponding estimators. Consequently, all aspects of the assumptions need to be specifically taken into account in constructing their corresponding efficient estimators. This study clarifies the relation between the regression error assumptions and their, respectively, efficiency bounds under the general regression framework with heteroscedastic errors. Our simulation results support our findings; we carry out a real data analysis using the proposed methods where the Cobb–Douglas cost model is the regression mean.

Keywords Heteroscedasticity · Semiparametric method · Standardized regression error · Variance function

Electronic supplementary material The online version of this article (doi:10.1007/s10463-017-0622-0) contains supplementary material, which is available to authorized users.

✉ Mijeong Kim
m.kim@ewha.ac.kr

Yanyuan Ma
yzm63@psu.edu

[1] Department of Statistics, Ewha Womans University, Seoul 03760, Republic of Korea

[2] Department of Statistics, Penn State University, University Park, PA 16802, USA

1 Introduction

Regression models with the form $Y = m(\mathbf{X}, \boldsymbol{\alpha}) + \epsilon$ are among the oldest statistical models studied in statistics. Here, m is a known function and $\boldsymbol{\alpha}$ is an unknown parameter vector that has to be estimated. The most familiar assumption on the regression error ϵ is that $E(\epsilon) = 0$, which leads to the classical mean regression model. It has often been implied—rather than written out explicitly—that the regression error is independent of covariates; that is, $\epsilon \perp \mathbf{X}$. In fact, strictly speaking, we obtain two quite different models based on whether or not the independence assumption holds. While under the independence assumption one would commonly require the regression error to have a mean zero, when there is no independence assumption, a mean regression model would typically require the conditional regression mean to be zero, that is, $E(\epsilon \mid \mathbf{X}) = 0$. This subtle difference actually leads to two quite different models, each having its own estimation and inference procedures and deserving separate studies. Generally, $E(\epsilon \mid \mathbf{X}) = 0$ is the minimum assumption required to justify the term "mean regression model." Since this is a relatively weak assumption, very often other assumptions are added. Jacquez et al. (1968), Bement and Williams (1969), and Fuller and Rao (1978) assumed unequal variances in addition, given the covariates, and studied weighted least squares procedures. Carroll and Ruppert (1982) and Carroll (1982) adopted specific variance functions, assuming the unknown symmetric distribution of ϵ in heteroscedastic linear models. Müller and Zhao (1995) studied a general semiparametric variance function model. Kim and Ma (2012) proposed semiparametric efficient estimators in heteroscedastic nonlinear models under a known variance function. Ma et al. (2006) found the semiparametric efficiency bound in a heteroscedastic partially linear regression model with a nonparametric variance function.

Assumption $E(\epsilon \mid \mathbf{X}) = 0$ motivates us to consider model $Y = m(\mathbf{X}, \boldsymbol{\alpha}) + e^{\sigma(\mathbf{X}, \boldsymbol{\beta})}\epsilon$, which uses parameterization $\mathrm{e}^{\sigma(\mathbf{X}, \boldsymbol{\beta})}$ to ensure that the standard deviation of the regression error is positive. This is a typical regression model with heteroscedastic error modeling heteroscedasticity in a parametric form. Under this framework, ϵ is a standardized regression error that generally satisfies $E(\epsilon \mid \mathbf{X}) = 0$ and $\mathrm{var}(\epsilon \mid \mathbf{X}) = 1$. As mentioned earlier, it is not written out explicitly, but often implied that ϵ and $\mathbf{X}$ are independent of each other. Hall and Carroll (1989) proposed a method for the simultaneous estimation of a variance and a mean function in a parametric heteroscedastic regression model under the implicit assumption of independence of ϵ and $\mathbf{X}$. Lian et al. (2015) proposed a method for the estimation of mean and variance functions in heteroscedastic models when both the functions depend on partially linear single-index models. Fang et al. (2015) extended the model of Lian et al. (2015) to additive partial linear models. Lian et al. (2015) and Fang et al. (2015) assumed that $E(\epsilon) = 0$ and $E(|\epsilon|) = 1$, implicitly confirming the independence of ϵ and $\mathbf{X}$. However, we surprisingly find the efficiency bounds when estimating $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ usually different both with and without the additional independence assumption. In addition, at times we assume for convenience that the distribution of ϵ is symmetric, adding yet another layer of complexity when analyzing the efficiency of any specific estimator.

Since we frequently face these similar yet different regression models with various assumptions on the regression errors, and at times come across efficiency statements in the literature that give no rigorous proofs or careful justifications, we consider it

necessary to study these regression models more carefully and systematically. The main purpose of this study is to investigate four regression models with identical regression mean and variance but different assumptions on the standardized regression error. We perform a semiparametric analysis of these models and derive their optimal efficiency bounds in Sect. 2. We carry out a comparison of these models in Sect. 3, showing that all the aspects of the assumptions need to be specifically taken into account when deriving the optimal estimators. Some numerical illustrations of the efficient estimators are provided in Sect. 4. We conclude the study in Sect. 5. The technical details and proofs of the study are provided in appendix and supplement.

2 Four heteroscedastic regression models

The four regression models with heteroscedastic errors that we consider here have the common form

$$Y = m(\mathbf{X}, \boldsymbol{\alpha}) + e^{\sigma(\mathbf{X},\boldsymbol{\beta})}\epsilon, \tag{1}$$

where $Y \in \mathbb{R}$ is the response variable and $\mathbf{X} \in \mathbb{R}^d$ is a covariate vector. The mean function m is a known function up to the unknown parameter vector $\boldsymbol{\alpha} \in \mathbb{R}^k$, and the heteroscedasticity of the model is reflected in the regression error $e^{\sigma(\mathbf{X},\boldsymbol{\beta})}\epsilon$, where σ is a known function up to the parameter vector $\boldsymbol{\beta} \in \mathbb{R}^l$. Now, ϵ is a "standardized" regression error satisfying

$$E(\epsilon \mid \mathbf{X}) = 0, \quad \text{var}(\epsilon \mid \mathbf{X}) = 1. \tag{2}$$

The model in (1) and (2) is the most basic one considered in this study. It is the first heteroscedastic regression model we consider here.

For notational convenience, we give $\boldsymbol{\theta} = (\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$ as our parameter of interest. Here, our aim is to estimate $\boldsymbol{\theta}$ without imposing any parametric distributional assumption on ϵ. Thus, we consider model (1) along with constraint (2) as our semiparametric model. Assume that $\mathbf{Z} = (Y, \mathbf{X})$. The density of the single observation $\mathbf{z}$ is

$$f_{\mathbf{Z}}(\mathbf{z}, \boldsymbol{\theta}, f_{\mathbf{X}}, f_{\epsilon|\mathbf{X}}) = f_{\epsilon,\mathbf{X}}(\epsilon, \mathbf{x}) = f_{\epsilon|\mathbf{X}}(\epsilon, \mathbf{x}) f_{\mathbf{X}}(\mathbf{x}),$$

where $\boldsymbol{\theta}$ is the finite-dimensional parameter of interest, and $f_{\mathbf{X}}$ and $f_{\epsilon|\mathbf{X}}$ are two infinite-dimensional nuisance parameters. Here, $f_{\epsilon|\mathbf{X}}$ is the conditional probability density function (pdf) of ϵ, given $\mathbf{X}$, and $f_{\mathbf{X}}$ is the pdf of $\mathbf{X}$. We assume that both functions are twice differentiable and have the first four moments.

The assumption on the standardized regression error ϵ can be strengthened in various ways. Specifically, we consider the following four cases; the first case has no assumption other than (1) and (2).

Case 1. (1) and (2).

Case 2. $\epsilon \perp \mathbf{X}$, in addition to (1) and (2).

Case 3. $f_{\epsilon|\mathbf{X}}(\epsilon, \mathbf{x}) = f_{\epsilon|\mathbf{X}}(-\epsilon, \mathbf{x})$, in addition to (1) and (2).

Case 4. $\epsilon \perp \mathbf{X}$, $f_\epsilon(\epsilon) = f_\epsilon(-\epsilon)$, in addition to (1) and (2).

Obviously, Case 1 has the weakest assumption. This case is often referred to as the location scale model. Case 2 additionally has the independence assumption between the standardized regression error and the covariates. Thus, this case has a more stringent assumption than Case 1. Since a more stringent model assumption implies more model structures, and more estimators are available for Case 2, the optimal efficiency bound of Case 2 should improve upon Case 1. Likewise, compared to Case 1, Case 2 further assumes symmetry on the standardized regression error ϵ. Thus, its efficiency bound should further improve upon Case 1. Finally, because Case 4 assumes both symmetry and independence, it should yield the best efficiency bound of all the four cases. We now investigate each of the four cases to illustrate quantitatively the differences of the four optimal estimators and their, respectively, efficiency bounds. Generally, to conceptually derive efficient estimators, one has to derive the efficient score $\boldsymbol{S}_{\text{eff}}$, defined as orthogonal projection of the score function of $\boldsymbol{\theta}$ onto the so-called nuisance tangent space orthogonal complement (Tsiatis 2006). We denote the nuisance tangent space Λ, its orthogonal complement $\Lambda^\perp$. In each of the four models, the nuisance tangent space is the space the score spans with respect to the error distribution; this is assumed to be one of the infinite-dimensional nuisance parameters. To derive Λ mathematically is highly technical and often hard. Operationally, for each model, we first derive Λ, $\Lambda^\perp$, project the score function $\boldsymbol{S}_{\boldsymbol{\theta}}$ onto $\Lambda^\perp$, obtain the efficient score function $\boldsymbol{S}_{\text{eff}}$, and finally use this efficient score function to construct the estimation equation $\sum \boldsymbol{S}_{\text{eff}} = \mathbf{0}$. The root of this estimation equation forms the efficient estimator, whose estimation variability is given as $\{E(\boldsymbol{S}_{\text{eff}}\boldsymbol{S}_{\text{eff}}^{\text{T}})\}^{-1}$. This estimator is known to be minimum among all consistent estimators of $\boldsymbol{\theta}$ (Bickel et al. 1998). For these derivations, we routinely require that the pdf be a sufficiently smooth function of both the random variable and parameter in the neighborhood of the true parameter value. For all the four cases considered here, we obtain efficient estimators; we provide their detailed derivations in appendix (Case 1) and supplement (Cases 2, 3, and 4).

Case 1

Case 1 is the most general model among the four cases. Because this case has the weakest model assumption, it has the smallest class of consistent estimators. In order to find the class of estimators and the semiparametric efficient estimator for this class, we derive the entire nuisance tangent space orthogonal complement $\Lambda^\perp$ and the efficient score function (Tsiatis 2006).

Proposition 1 *The nuisance tangent space Λ and its orthogonal complement space $\Lambda^\perp$ of the Case 1 model are, respectively,*

$$\Lambda = \{\mathbf{h}(\mathbf{x}, \epsilon) : E\{\mathbf{h}(\mathbf{X}, \epsilon)\} = E\{\epsilon \mathbf{h}(\mathbf{X}, \epsilon)|\mathbf{X}\} = E\{t\mathbf{h}(\mathbf{X}, \epsilon)|\mathbf{X}\} = \mathbf{0}\}$$

and

$$\Lambda^\perp = \{\mathbf{g}(\mathbf{x}, \epsilon) : \mathbf{g}(\mathbf{x}, \epsilon) = \mathbf{g}_1(\mathbf{x})\epsilon + \mathbf{g}_2(\mathbf{x})t\},$$

where $t = \epsilon^2 - E(\epsilon^3|\mathbf{X})\epsilon - 1$.

By projecting the score vector with respect to $\boldsymbol{\theta}$, $\boldsymbol{S}_{\boldsymbol{\theta}}(\mathbf{X}, Y)$, onto $\Lambda^{\perp}$, we obtain the efficient score vector $\boldsymbol{S}_{\text{eff}}(\mathbf{X}, Y)$.

Theorem 1 *The efficient score vector in Case 1 is* $\boldsymbol{S}_{\text{eff}}(\mathbf{X}, Y) = (\boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}}^{\mathrm{T}}, \boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}^{\mathrm{T}})^{\mathrm{T}}$, *where*

$$\begin{aligned}\boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}} &= e^{-\sigma(\mathbf{X},\boldsymbol{\beta})}\mathbf{m}'_{\boldsymbol{\alpha}}(\mathbf{X}, \boldsymbol{\alpha})\left\{\epsilon - \frac{E(\epsilon^3|\mathbf{X})}{E(t^2|\mathbf{X})}t\right\},\\ \boldsymbol{S}_{\text{eff},\boldsymbol{\beta}} &= \frac{2t}{E(t^2|\mathbf{X})}\boldsymbol{\sigma}'_{\boldsymbol{\beta}}(\mathbf{X}, \boldsymbol{\beta}),\end{aligned} \tag{3}$$

and $t = \epsilon^2 - E(\epsilon^3|\mathbf{X})\epsilon - 1$. *The optimal efficiency matrix is*

$$\mathbf{M}_1 \equiv E\left(\boldsymbol{S}_{\text{eff}}\boldsymbol{S}_{\text{eff}}^{\mathrm{T}}\right) = \begin{Bmatrix} E\left(\boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}}\boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}}^{\mathrm{T}}\right) & E\left(\boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}}\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}^{\mathrm{T}}\right) \\ E\left(\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}\boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}}^{\mathrm{T}}\right) & E\left(\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}^{\mathrm{T}}\right) \end{Bmatrix},$$

where

$$\begin{aligned}E\left(\boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}}\boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}}^{\mathrm{T}}\right) &= E\left\{\left[1 + \frac{\{E(\epsilon^3|\mathbf{X})\}^2}{E(t^2|\mathbf{X})}\right]e^{-2\sigma(\mathbf{X},\boldsymbol{\beta})}\mathbf{m}'_{\boldsymbol{\alpha}}(\mathbf{X}, \boldsymbol{\alpha})\mathbf{m}'^{\mathrm{T}}_{\boldsymbol{\alpha}}(\mathbf{X}, \boldsymbol{\alpha})\right\},\\ E\left(\boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}}\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}^{\mathrm{T}}\right) &= E\left(\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}\boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}}^{\mathrm{T}}\right)^{\mathrm{T}}\\ &= E\left\{-\frac{2E(\epsilon^3|\mathbf{X})}{E(t^2|\mathbf{X})}e^{-\sigma(\mathbf{X},\boldsymbol{\beta})}\mathbf{m}'_{\boldsymbol{\alpha}}(\mathbf{X}, \boldsymbol{\alpha})\boldsymbol{\sigma}'^{\mathrm{T}}_{\boldsymbol{\beta}}(\mathbf{X}, \boldsymbol{\beta})\right\},\\ E\left(\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}^{\mathrm{T}}\right) &= E\left\{\frac{4}{E(t^2|\mathbf{X})}\boldsymbol{\sigma}'_{\boldsymbol{\beta}}(\mathbf{X}, \boldsymbol{\beta})\boldsymbol{\sigma}'^{\mathrm{T}}_{\boldsymbol{\beta}}(\mathbf{X}, \boldsymbol{\beta})\right\}.\end{aligned}$$

Note that the variance–covariance matrix of the efficient estimator is the inverse of the optimal efficiency matrix $E(\boldsymbol{S}_{\text{eff}}\boldsymbol{S}_{\text{eff}}^{\mathrm{T}})$. In other words, when estimating $\boldsymbol{\theta}$ for the Case 1 model, the minimum possible variance is $E(\boldsymbol{S}_{\text{eff}}\boldsymbol{S}_{\text{eff}}^{\mathrm{T}})^{-1}$.

Case 2

For Case 2, we further add the independence assumption of ϵ and $\mathbf{X}$ to the assumption of Case 1. As with Case 1, we can construct the nuisance tangent space Λ and its orthogonal complement $\Lambda^{\perp}$ for Case 2.

Proposition 2 *The nuisance tangent space* Λ *and its orthogonal complement space* $\Lambda^{\perp}$ *of the model in Case 2 are, respectively,*

$$\Lambda = \{\mathbf{a}(\mathbf{x}) + \mathbf{b}(\epsilon) : E\{\mathbf{a}(\mathbf{X})\} = \mathbf{0}, E\{\mathbf{b}(\epsilon)\} = \mathbf{0}, E\{\epsilon\mathbf{b}(\epsilon)\} = \mathbf{0}, E\{t\mathbf{b}(\epsilon)\} = \mathbf{0}\},$$

and

$$\Lambda^{\perp} = \left\{\mathbf{g}(\mathbf{x}, \epsilon) : E\{\mathbf{g}(\mathbf{X}, \epsilon)|\mathbf{X}\} = \mathbf{0}, E\{\mathbf{g}(\mathbf{X}, \epsilon)|\epsilon\} = \mathbf{c}_1\epsilon + \mathbf{c}_2 t : \mathbf{c}_1, \mathbf{c}_2 \in \mathbb{R}^{k+l}\right\},$$

where $t = \epsilon^2 - E(\epsilon^3)\epsilon - 1$.

Theorem 2 *The efficient score vector in Case 2 is* $\boldsymbol{S}_{\text{eff}}(\mathbf{X}, Y) = (\boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}}^{\mathrm{T}}, \boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}^{\mathrm{T}})^{\mathrm{T}}$, *where*

$$\begin{aligned}
\boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}} &= -\frac{f_\epsilon'(\epsilon)}{f_\epsilon(\epsilon)}\left[e^{-\sigma(\mathbf{X},\boldsymbol{\beta})}\mathbf{m}_{\boldsymbol{\alpha}}'(\mathbf{X},\boldsymbol{\alpha}) - E\left\{e^{-\sigma(\mathbf{X},\boldsymbol{\beta})}\mathbf{m}_{\boldsymbol{\alpha}}'(\mathbf{X},\boldsymbol{\alpha})\right\}\right] \\
&\quad + \left\{\epsilon - \frac{E(\epsilon^3)}{E(t^2)}t\right\} E\left\{e^{-\sigma(\mathbf{X},\boldsymbol{\beta})}\mathbf{m}_{\boldsymbol{\alpha}}'(\mathbf{X},\boldsymbol{\alpha})\right\}, \\
\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}} &= -\left\{\frac{f_\epsilon'(\epsilon)}{f_\epsilon(\epsilon)}\epsilon + 1\right\}\left[\boldsymbol{\sigma}_{\boldsymbol{\beta}}'(\mathbf{X},\boldsymbol{\beta}) - E\left\{\boldsymbol{\sigma}_{\boldsymbol{\beta}}'(\mathbf{X},\boldsymbol{\beta})\right\}\right] + \frac{2t}{E(t^2)}E\left\{\boldsymbol{\sigma}_{\boldsymbol{\beta}}'(\mathbf{X},\boldsymbol{\beta})\right\},
\end{aligned} \tag{4}$$

and $t = \epsilon^2 - E(\epsilon^3)\epsilon - 1$. *The optimal efficiency matrix is*

$$\mathbf{M}_2 \equiv E\left(\boldsymbol{S}_{\text{eff}}\boldsymbol{S}_{\text{eff}}^{\mathrm{T}}\right) = \begin{Bmatrix} E\left(\boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}}\boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}}^{\mathrm{T}}\right) & E\left(\boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}}\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}^{\mathrm{T}}\right) \\ E\left(\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}\boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}}^{\mathrm{T}}\right) & E\left(\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}^{\mathrm{T}}\right) \end{Bmatrix},$$

where

$$\begin{aligned}
E(\boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}}\boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}}^{\mathrm{T}}) &= E\left\{\frac{f_\epsilon'(\epsilon)^2}{f_\epsilon(\epsilon)^2}\right\} E\left\{e^{-2\sigma(\mathbf{X},\boldsymbol{\beta})}\mathbf{m}_{\boldsymbol{\alpha}}'(\mathbf{X},\boldsymbol{\alpha})\mathbf{m}_{\boldsymbol{\alpha}}'^{\mathrm{T}}(\mathbf{X},\boldsymbol{\alpha})\right\} \\
&\quad + \left[1 + \frac{\{E(\epsilon^3)\}^2}{E(t^2)} - E\left\{\frac{f_\epsilon'(\epsilon)^2}{f_\epsilon(\epsilon)^2}\right\}\right] E\left\{e^{-\sigma(\mathbf{X},\boldsymbol{\beta})}\mathbf{m}_{\boldsymbol{\alpha}}'(\mathbf{X},\boldsymbol{\alpha})\right\} \\
&\quad \times E\left\{e^{-\sigma(\mathbf{X},\boldsymbol{\beta})}\mathbf{m}_{\boldsymbol{\alpha}}'(\mathbf{X},\boldsymbol{\alpha})^{\mathrm{T}}\right\}, \\
E\left(\boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}}\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}^{\mathrm{T}}\right) &= E\left(\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}\boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}}^{\mathrm{T}}\right)^{\mathrm{T}} \\
&= E\left\{\frac{f_\epsilon'(\epsilon)^2}{f_\epsilon(\epsilon)^2}\epsilon\right\} E\left\{e^{-\sigma(\mathbf{X},\boldsymbol{\beta})}\mathbf{m}_{\boldsymbol{\alpha}}'(\mathbf{X},\boldsymbol{\alpha})\boldsymbol{\sigma}_{\boldsymbol{\beta}}'^{\mathrm{T}}(\mathbf{X},\boldsymbol{\beta})\right\} \\
&\quad - \left[E\left\{\frac{f_\epsilon'(\epsilon)^2}{f_\epsilon(\epsilon)^2}\epsilon\right\} + \frac{2E(\epsilon^3)}{E(t^2)}\right] E\left\{e^{-\sigma(\mathbf{X},\boldsymbol{\beta})}\mathbf{m}_{\boldsymbol{\alpha}}'(\mathbf{X},\boldsymbol{\alpha})\right\} E\left\{\boldsymbol{\sigma}_{\boldsymbol{\beta}}'(\mathbf{X},\boldsymbol{\beta})^{\mathrm{T}}\right\}, \\
E\left(\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}^{\mathrm{T}}\right) &= E\left\{\frac{f_\epsilon'(\epsilon)^2}{f_\epsilon(\epsilon)^2}\epsilon^2 - 1\right\} E\left\{\boldsymbol{\sigma}_{\boldsymbol{\beta}}'(\mathbf{X},\boldsymbol{\beta})\boldsymbol{\sigma}_{\boldsymbol{\beta}}'^{\mathrm{T}}(\mathbf{X},\boldsymbol{\beta})\right\} \\
&\quad + \left[-E\left\{\frac{f_\epsilon'(\epsilon)^2}{f_\epsilon(\epsilon)^2}\epsilon^2\right\} + 1 + \frac{4}{E(t^2)}\right] E\left\{\boldsymbol{\sigma}_{\boldsymbol{\beta}}'(\mathbf{X},\boldsymbol{\beta})\right\} E\left\{\boldsymbol{\sigma}_{\boldsymbol{\beta}}'(\mathbf{X},\boldsymbol{\beta})^{\mathrm{T}}\right\}.
\end{aligned}$$

Case 3

For Case 3, we strengthen the Case 1 assumption by further assuming symmetry of the standardized regression error distribution; that is, $f_{\epsilon|\mathbf{X}}(\epsilon, \mathbf{x}) = f_{\epsilon|\mathbf{X}}(-\epsilon, \mathbf{x})$. The nuisance tangent space Λ and its orthogonal complement $\Lambda^{\perp}$ for Case 3 are given below.

Proposition 3 *The nuisance tangent space* Λ *and its orthogonal complement space* $\Lambda^{\perp}$ *of the model in Case 3 are, respectively,*

$$\Lambda = \{\mathbf{h}(\mathbf{x}, \epsilon) : E\{\mathbf{h}(\mathbf{X}, \epsilon)\} = E\{t\mathbf{h}(\mathbf{X}, \epsilon)|\mathbf{X}\} = \mathbf{0}, \mathbf{h}(\mathbf{x}, \epsilon) = \mathbf{h}(\mathbf{x}, -\epsilon)\}$$
and
$$\Lambda^{\perp} = \{\mathbf{g}(\mathbf{x}, \epsilon) : \mathbf{g}(\mathbf{x}, \epsilon) = \mathbf{a}(\mathbf{x}, \epsilon) + \mathbf{b}(\mathbf{x})t, \mathbf{a}(\mathbf{x}, \epsilon) + \mathbf{a}(\mathbf{x}, -\epsilon) = \mathbf{0}\},$$

where $t = \epsilon^2 - 1$.

Because of symmetry of the distribution of ϵ conditional on $\mathbf{X}$, we have $t = \epsilon^2 - 1$ in Case 3, whereas $t = \epsilon^2 - E(\epsilon^3|\mathbf{X})\epsilon - 1$ in Case 1.

Theorem 3 *The efficient score vector in Case 3 is* $\boldsymbol{S}_{\text{eff}}(\mathbf{X}, Y) = (\boldsymbol{S}^{\text{T}}_{\text{eff},\boldsymbol{\alpha}}, \boldsymbol{S}^{\text{T}}_{\text{eff},\boldsymbol{\beta}})^{\text{T}}$, *where*

$$\begin{aligned}\boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}} &= -\frac{\partial f_{\epsilon|\mathbf{X}}(\epsilon, \mathbf{X})/\partial\epsilon}{f_{\epsilon|\mathbf{X}}(\epsilon, \mathbf{X})} e^{-\sigma(\mathbf{X},\boldsymbol{\beta})}\mathbf{m}'_{\boldsymbol{\alpha}}(\mathbf{X}, \boldsymbol{\alpha}),\\ \boldsymbol{S}_{\text{eff},\boldsymbol{\beta}} &= \frac{2t}{E(t^2|\mathbf{X})}\boldsymbol{\sigma}'_{\boldsymbol{\beta}}(\mathbf{X}, \boldsymbol{\beta}),\end{aligned} \tag{5}$$

and $t = \epsilon^2 - 1$. *The optimal efficiency matrix is*

$$\mathbf{M}_3 \equiv E\left(\boldsymbol{S}_{\text{eff}}\boldsymbol{S}^{\text{T}}_{\text{eff}}\right) = \begin{Bmatrix} E\left(\boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}}\boldsymbol{S}^{\text{T}}_{\text{eff},\boldsymbol{\alpha}}\right) & \mathbf{0} \\ \mathbf{0} & E\left(\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}\boldsymbol{S}^{\text{T}}_{\text{eff},\boldsymbol{\beta}}\right) \end{Bmatrix},$$

where

$$\begin{aligned}E\left(\boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}}\boldsymbol{S}^{\text{T}}_{\text{eff},\boldsymbol{\alpha}}\right) &= E\left\{\left(\frac{\partial f_{\epsilon|\mathbf{X}}(\epsilon, \mathbf{X})/\partial\epsilon}{f_{\epsilon|\mathbf{X}}(\epsilon, \mathbf{X})}\right)^2 e^{-2\sigma(\mathbf{X},\boldsymbol{\beta})}\mathbf{m}'_{\boldsymbol{\alpha}}(\mathbf{X}, \boldsymbol{\alpha})\mathbf{m}'^{\text{T}}_{\boldsymbol{\alpha}}(\mathbf{X}, \boldsymbol{\alpha})\right\},\\ E\left(\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}\boldsymbol{S}^{\text{T}}_{\text{eff},\boldsymbol{\beta}}\right) &= E\left\{\frac{4}{E(t^2|\mathbf{X})}\boldsymbol{\sigma}'_{\boldsymbol{\beta}}(\mathbf{X}, \boldsymbol{\beta})\boldsymbol{\sigma}'^{\text{T}}_{\boldsymbol{\beta}}(\mathbf{X}, \boldsymbol{\beta})\right\}.\end{aligned}$$

The optimal efficiency matrix of Case 3 is simpler than that of Case 1 because of the symmetry of ϵ, given $\mathbf{X}$.

Case 4

Case 4 assumes independence of the covariates and normalized regression error as well as the symmetry of the normalized regression error distribution; that is, $\epsilon \perp \mathbf{X}$ and $f_{\epsilon}(\epsilon) = f_{\epsilon}(-\epsilon)$. In this case, we obtain the following results.

Proposition 4 *The nuisance tangent space* Λ *and its orthogonal complement space* $\Lambda^{\perp}$ *of the model in Case 4 are, respectively,*

$$\begin{aligned}\Lambda &= \{\mathbf{a}(\mathbf{x}) + \mathbf{b}(\epsilon) : E\{\mathbf{a}(\mathbf{X})\} = \mathbf{0},\, E\{\mathbf{b}(\epsilon)\} = \mathbf{0},\, E\{t\mathbf{b}(\epsilon)\} = \mathbf{0},\, \mathbf{b}(\epsilon) = \mathbf{b}(-\epsilon)\}\\ &\quad \textit{and}\\ \Lambda^{\perp} &= \big\{\mathbf{g}(\mathbf{x},\epsilon) : E\{\mathbf{g}(\mathbf{X},\epsilon)|\mathbf{X}\} = \mathbf{0},\, E\{\mathbf{g}(\mathbf{X},\epsilon)|\epsilon\} = \mathbf{a}(\epsilon) + \mathbf{b}t,\, \mathbf{a}(\epsilon)\\ &\quad + \mathbf{a}(-\epsilon) = \mathbf{0},\, \mathbf{b} \in \mathbb{R}^{k+l}\big\},\end{aligned}$$

where $t = \epsilon^2 - 1$.

Theorem 4 *The efficient score vector in Case 4 is* $\boldsymbol{S}_{\text{eff}}(\mathbf{X}, Y) = (\boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}}^{\mathrm{T}}, \boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}^{\mathrm{T}})^{\mathrm{T}}$, *where*

$$\begin{aligned}\boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}} &= -\frac{f_\epsilon'(\epsilon)}{f_\epsilon(\epsilon)} e^{-\sigma(\mathbf{X},\boldsymbol{\beta})} \mathbf{m}_{\boldsymbol{\alpha}}'(\mathbf{X},\boldsymbol{\alpha}),\\ \boldsymbol{S}_{\text{eff},\boldsymbol{\beta}} &= \left\{-\frac{f_\epsilon'(\epsilon)}{f_\epsilon(\epsilon)}\epsilon - 1\right\}\left[\boldsymbol{\sigma}_{\boldsymbol{\beta}}'(\mathbf{X},\boldsymbol{\beta}) - E\{\boldsymbol{\sigma}_{\boldsymbol{\beta}}'(\mathbf{X},\boldsymbol{\beta})\}\right] + \frac{2t}{E(t^2)} E\{\boldsymbol{\sigma}_{\boldsymbol{\beta}}'(\mathbf{X},\boldsymbol{\beta})\},\end{aligned} \tag{6}$$

and $t = \epsilon^2 - 1$. *The optimal efficiency matrix is*

$$\mathbf{M}_4 \equiv E\left(\boldsymbol{S}_{\text{eff}} \boldsymbol{S}_{\text{eff}}^{\mathrm{T}}\right) = \begin{Bmatrix} E\left(\boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}} \boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}}^{\mathrm{T}}\right) & \mathbf{0} \\ \mathbf{0} & E\left(\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}} \boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}^{\mathrm{T}}\right) \end{Bmatrix},$$

where

$$\begin{aligned}E\left(\boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}} \boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}}^{\mathrm{T}}\right) &= E\left\{\frac{f_\epsilon'(\epsilon)^2}{f_\epsilon(\epsilon)^2}\right\} E\left\{e^{-2\sigma(\mathbf{X},\boldsymbol{\beta})} \mathbf{m}_{\boldsymbol{\alpha}}'(\mathbf{X},\boldsymbol{\alpha}) \mathbf{m}_{\boldsymbol{\alpha}}'^{\,\mathrm{T}}(\mathbf{X},\boldsymbol{\alpha})\right\},\\ E\left(\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}} \boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}^{\mathrm{T}}\right) &= E\left\{\frac{f_\epsilon'(\epsilon)^2}{f_\epsilon(\epsilon)^2}\epsilon^2 - 1\right\} E\left\{\boldsymbol{\sigma}_{\boldsymbol{\beta}}'(\mathbf{X},\boldsymbol{\beta}) \boldsymbol{\sigma}_{\boldsymbol{\beta}}'(\mathbf{X},\boldsymbol{\beta})^{\mathrm{T}}\right\}\\ &\quad + \left[-E\left\{\frac{f_\epsilon'(\epsilon)^2}{f_\epsilon(\epsilon)^2}\epsilon^2\right\} + 1 + \frac{4}{E(t^2)}\right] E\left\{\boldsymbol{\sigma}_{\boldsymbol{\beta}}'(\mathbf{X},\boldsymbol{\beta})\right\} E\left\{\boldsymbol{\sigma}_{\boldsymbol{\beta}}'(\mathbf{X},\boldsymbol{\beta})^{\mathrm{T}}\right\}.\end{aligned}$$

3 Comparison of optimal efficiency matrices

Since we have the optimal efficiency matrices for all cases, we can now draw a formal comparison between the cases. Specifically, we can calculate the difference between the covariance matrices and study the difference. We present the results of the theorems and their proofs in supplement.

Comparison of Cases 1 and 2

In general, because the Case 2 model has stronger assumptions than the Case 1 model, the estimators derived in Case 2 are not necessarily consistent with the assumptions of Case 1. Of course, it would not be fair to compare the estimation efficiency of two estimator families when one has consistent and the other has inconsistent estimators. We thus consider exclusively the situation where the Case 2 assumptions are satisfied and compare the optimal efficiencies of the Case 1 and Case 2 estimators. Here, note that the Case 2 model has consistent estimators and specifically takes into account the independence assumption, whereas the Case 1 model also has consistent estimators but ignores the additional independence property. Thus, intuitively, the Case 2 model estimators are larger than the Case 1 model estimators, and we can expect the optimal efficiency matrices to satisfy that $\mathbf{M}_2 - \mathbf{M}_1$ is positive definite.

Theorem 5 *For the Case 2 assumption,* $\mathbf{M}_2 - \mathbf{M}_1 = E(\mathbf{u}\mathbf{u}^{\mathrm{T}})$, *where*

$$\mathbf{u} = \begin{bmatrix} \left\{\epsilon - \frac{E(\epsilon^3)}{E(t^2)}t + \frac{f'_\epsilon(\epsilon)}{f_\epsilon(\epsilon)}\right\}\left[e^{-\sigma(\mathbf{X},\boldsymbol{\beta})}\mathbf{m}'_{\boldsymbol{\alpha}}(\mathbf{X},\boldsymbol{\alpha}) - E\{e^{-\sigma(\mathbf{X},\boldsymbol{\beta})}\mathbf{m}'_{\boldsymbol{\alpha}}(\mathbf{X},\boldsymbol{\alpha})\}\right] \\ \left\{\frac{f'_\epsilon(\epsilon)}{f_\epsilon(\epsilon)}\epsilon + 1 + \frac{2t}{E(t^2)}\right\}\left[\boldsymbol{\sigma}'_{\boldsymbol{\beta}}(\mathbf{X},\boldsymbol{\beta}) - E\{\boldsymbol{\sigma}'_{\boldsymbol{\beta}}(\mathbf{X},\boldsymbol{\beta})\}\right] \end{bmatrix}.$$

Thus, $\mathbf{M}_2 - \mathbf{M}_1$ *is nonnegative definite.*

In general, because $\mathbf{u} \neq \mathbf{0}$, we have $\mathbf{M}_2 \neq \mathbf{M}_1$. The only exception is when both the mean and variance functions, that is, $m(\mathbf{x},\boldsymbol{\alpha})$ and $\sigma(\mathbf{x},\boldsymbol{\beta})$, are constants, or when $\epsilon - tE(\epsilon^3)/E(t^2) + f'_\epsilon(\epsilon)/f_\epsilon(\epsilon) = \epsilon f'_\epsilon(\epsilon)/f_\epsilon(\epsilon) + 1 + 2t/E(t^2) = 0$. The latter relation leads to

$$\frac{f'_\epsilon(\epsilon)}{f_\epsilon(\epsilon)} = -\epsilon,$$

where $f_\epsilon(\epsilon)$ is a standard normal distribution. Thus, we conclude that, in general, the optimal efficiency matrix $\mathbf{M}_2$ is larger than $\mathbf{M}_1$ in terms of positive definiteness, but in the degenerate case when both the mean and variance of the regression function are constants or the normalized error is independent of the covariates and is normally distributed, the two efficiency matrices are identical.

The case of constant mean and variance is of course very special and not where a typical heteroscedastic regression model can be used. On the other hand, a normally distributed standardized regression error is often possible. If this is indeed the case, the additional symmetric assumption of Case 2 will not bring in any efficiency gain.

Comparison of Cases 1 and 3

Similarly, Case 3 also has a stronger assumption than Case 1. We derive a corresponding result for the two estimator classes in terms of optimal efficiency bounds.

Theorem 6 *Under the Case 3 assumption,* $\mathbf{M}_3 - \mathbf{M}_1 = E(\mathbf{u}\mathbf{u}^{\mathrm{T}})$, *where*

$$\mathbf{u} = \begin{bmatrix} \left\{\frac{\partial f_{\epsilon|\mathbf{X}}(\epsilon,\mathbf{X})/\partial\epsilon}{f_{\epsilon|\mathbf{X}}(\epsilon,\mathbf{X})} + \epsilon\right\} e^{-\sigma(\mathbf{X},\boldsymbol{\beta})}\mathbf{m}'_{\boldsymbol{\alpha}}(\mathbf{X},\boldsymbol{\alpha}) \\ \mathbf{0} \end{bmatrix}.$$

Thus, $\mathbf{M}_3 - \mathbf{M}_1$ *is nonnegative definite.*

In general, $\mathbf{u} \neq \mathbf{0}$, except when the mean function $m(\mathbf{x}, \boldsymbol{\alpha})$ is a constant or $\partial f_{\epsilon|\mathbf{X}}(\epsilon, \mathbf{X})/\partial\epsilon / f_{\epsilon|\mathbf{X}}(\epsilon, \mathbf{X}) + \epsilon = 0$. The latter relation implies that ϵ is independent of $\mathbf{X}$ and has a standard normal distribution. Thus, we conclude that the optimal efficiency matrix $\mathbf{M}_3$ is in general larger than $\mathbf{M}_1$ in terms of positive definiteness. Except the degenerated case when $m(\mathbf{x}, \boldsymbol{\alpha})$ is a constant, only when the standardized regression error is normally distributed and independent of the covariates, the symmetric error assumption of Case 3 does not bring in additional gain.

Comparison of Cases 2 and 4

We similarly compare Cases 2 and 4 under the Case 4 assumption, which is stronger than the Case 2 assumption.

Theorem 7 *Under the Case 4 assumption,* $\mathbf{M}_4 - \mathbf{M}_2 = E(\mathbf{u}\mathbf{u}^{\mathrm{T}})$*, where*

$$\mathbf{u} = \begin{bmatrix} \left\{ \frac{f'_\epsilon(\epsilon)}{f_\epsilon(\epsilon)} + \epsilon \right\} E\{e^{-\sigma(\mathbf{X},\boldsymbol{\beta})} \mathbf{m}'_{\boldsymbol{\alpha}}(\mathbf{X}, \boldsymbol{\alpha})\} \\ \mathbf{0} \end{bmatrix}.$$

Therefore, $\mathbf{M}_4 - \mathbf{M}_2$ *is nonnegative definite.*

Because $\mathbf{u} \neq \mathbf{0}$, the optimal efficiency matrix $\mathbf{M}_4$ in general is larger than $\mathbf{M}_2$ in terms of positive definiteness. As with the previous comparisons, the two exceptional situations are when $E\{e^{-\sigma(\mathbf{X},\boldsymbol{\beta})} \mathbf{m}'_{\boldsymbol{\alpha}}(\mathbf{X}, \boldsymbol{\alpha})\} = \mathbf{0}$ and when ϵ has a standard normal distribution. If one of these situations occurs, the symmetry assumption of Case 4 does not bring in additional gain. Note that unlike with the earlier analysis, $E\{e^{-\sigma(\mathbf{X},\boldsymbol{\beta})} \mathbf{m}'_{\boldsymbol{\alpha}}(\mathbf{X}, \boldsymbol{\alpha})\} = \mathbf{0}$ imposes a nontrivial relation between the regression mean function in terms of its derivative and the regression standard deviation function that holds in some practical situations.

Comparison of Cases 3 and 4

Finally, we compare Cases 3 and 4 under the Case 4 assumption.

Theorem 8 *Under the Case 4 assumption,* $\mathbf{M}_4 - \mathbf{M}_3 = E(\mathbf{u}\mathbf{u}^{\mathrm{T}})$*, where*

$$\mathbf{u} = \begin{bmatrix} \mathbf{0} \\ \left\{ \frac{f'_\epsilon(\epsilon)}{f_\epsilon(\epsilon)} \epsilon + 1 + \frac{2t}{E(t^2)} \right\} \left[\boldsymbol{\sigma}'_{\boldsymbol{\beta}}(\mathbf{X}, \boldsymbol{\beta}) - E\left\{ \boldsymbol{\sigma}'_{\boldsymbol{\beta}}(\mathbf{X}, \boldsymbol{\beta}) \right\} \right] \end{bmatrix}.$$

Thus, $\mathbf{M}_4 - \mathbf{M}_3$ *is nonnegative definite.*

Because $\mathbf{u} \neq \mathbf{0}$ in general, $\mathbf{M}_4 \neq \mathbf{M}_3$. Thus, the Case 4 assumption brings in efficiency gain over Case 3. Note that $\mathbf{u} = \mathbf{0}$ only when $\sigma(\mathbf{x}, \boldsymbol{\beta})$ is a constant or $\epsilon f'_\epsilon(\epsilon)/f_\epsilon(\epsilon) + 1 + 2t/E(t^2) = 0$. From the latter relation, $f_\epsilon(\epsilon)$ is a standard normal distribution. Thus, we conclude that the optimal efficiency matrix $\mathbf{M}_4$ in general is larger than $\mathbf{M}_2$ in terms of positive definiteness, but when the model has homoscedastic error or the standardized error is normally distributed, the two efficiency matrices are identical and hence the additional independence assumption of Case 4 does not bring in any additional efficiency gain.

We left out the comparison of Cases 2 and 3 because the assumptions of these two cases do not have a clearly stronger or weaker relation and no definitive conclusion can be drawn in terms of $\mathbf{M}_3 - \mathbf{M}_2$.

4 Numerical results

4.1 Simulations

Using simulations, we show the finite sample performance of semiparametric estimators under the four cases. For all the cases, the model used is

$$Y = 5\exp(0.08X_1 - 0.15X_2) + \exp(0.06X_1 + 0.04X_2)\epsilon.$$

Thus, the mean and log-standard deviation functions are given, respectively, by

$$m(\mathbf{X}, \boldsymbol{\alpha}) = \alpha_0 \exp(\alpha_1 X_1 + \alpha_2 X_2), \quad \text{where } \boldsymbol{\alpha} = (5, 0.08, -0.15)^{\mathrm{T}}, \quad \text{and} \tag{7}$$

$$\sigma(\mathbf{X}, \boldsymbol{\beta}) = \beta_1 X_1 + \beta_2 X_2, \quad \text{where } \boldsymbol{\beta} = (0.06, 0.04)^{\mathrm{T}}. \tag{8}$$

4.1.1 Data generation

We generate X_1 and X_2 from the uniform distributions in (0,10) and (0,15), respectively. We decide the methodology of generating ϵ for the four cases from the corresponding assumptions for ϵ; they are presented below. We generated 1000 data sets with sample size $n = 500$.

Simulation 1. Generate $\epsilon = (e - p(\mathbf{X}))/\sqrt{2p(\mathbf{X})}$, where $e \sim \chi^2\{p(\mathbf{X})\}$ and $p(\mathbf{X}) = (X_1 + X_2) + 0.5$. Then, the probability density function of ϵ, given $\mathbf{X}$, is

$$f_{\epsilon|\mathbf{X}}(\epsilon, \mathbf{x}) = \frac{\sqrt{2p(\mathbf{x})}\left\{\sqrt{2p(\mathbf{x})}\epsilon + p(\mathbf{x})\right\}^{p(\mathbf{x})/2-1}}{2^{p(\mathbf{x})/2}\Gamma\left\{p(\mathbf{x})/2\right\}} \times \exp\left[-\left\{\sqrt{2p(\mathbf{x})}\epsilon + p(\mathbf{x})\right\}\Big/2\right] I\left\{\epsilon > -\sqrt{p(\mathbf{x})/2}\right\}, \tag{9}$$

where $I(\cdot)$ is an indicator function.

Simulation 2. Generate $\epsilon = (e - q)/\sqrt{2q}$, where $e \sim \chi^2(q)$ and $q = 13$. Then, ϵ has the following probability density function:

$$f_\epsilon(\epsilon) = \frac{\sqrt{2q}\left(\sqrt{2q}\epsilon + q\right)^{(q/2-1)}}{2^{q/2}\Gamma\left(q/2\right)} \exp\left\{-(\sqrt{2q}\epsilon + q)\Big/2\right\} I\left(\epsilon > -\sqrt{q/2}\right), \tag{10}$$

where $I(\cdot)$ is an indicator function.

Simulation 3. Generate ϵ from the different distributions according to $\mathbf{x}$.

$$\epsilon|\mathbf{X} \sim \begin{cases} \text{Unif}(-\sqrt{3}, \sqrt{3}), & \text{if } \mathbf{x} \in A, \\ \text{GN}\left(0, \sqrt{\frac{\Gamma(1/k)}{\Gamma(3/k)}}, k\right), \ k = 1.7, & \text{otherwise,} \end{cases}$$

where $A = \{(0, 5) \times (0, 7.5),\ (5, 10) \times (7.5, 15)\}$. Here, $(a_1, a_2) \times (b_1, b_2)$ denotes the rectangular area with $x_1 \in (a_1, a_2)$ and $x_2 \in (b_1, b_2)$. GN$(0, s, k)$ stands for the generalized normal distribution with scale parameter s and shape parameter k. Its probability density function is given by

$$\frac{k}{2s\Gamma(1/k)} e^{-(|\epsilon|/s)^k}. \tag{11}$$

Simulation 4. Generate ϵ from Logistic$(0, \sqrt{3}/\pi)$.

In Simulation 1, we generate ϵ from a standardized Chi-squared distribution whose degree of freedom depends on $\mathbf{X}$. This satisfies the error properties $E(\epsilon|\mathbf{X}) = 0$ and $\text{var}(\epsilon|\mathbf{X}) = 1$ of Case 1. For Simulation 2, we used a standardized Chi-squared distribution whose degree of freedom is independent of $\mathbf{X}$. This ensures that ϵ depends on the covariates in Simulation 1 but is independent in Simulation 2. In other words, the data in Simulation 2 satisfy the assumption of Case 2. For Simulation 3, we used a standardized generalized normal distribution (Nadarajah 2005), with both parameters depending on the covariates. Here, we set $s = \sqrt{\Gamma(1/k)/\Gamma(3/k)}$ in (11) so that $E(\epsilon|\mathbf{X}) = 0$ and $\text{var}(\epsilon|\mathbf{X}) = 1$. Thus, Simulation 3 fulfills the condition of Case 3. Finally, for Simulation 4, we generated ϵ from Logistic$(0, \sqrt{3}/\pi)$. The generation strategies used here ensure that ϵ is symmetrically distributed for both Simulations 3 and 4, is dependent on the covariates in Simulation 3, and is independent in Simulation 4. Thus, Simulation 3 belongs to Case 4.

4.1.2 Estimation

We implement all the four estimators corresponding to the assumed four error structures for four simulations, yielding a total of 16 sets of results. Some of the assumptions match the data generation procedure and hence are ideal. Some of the assumptions mismatch the data generation procedure and hence lead to either an inconsistent estimator when the assumptions are stronger than the true data property, or an inefficient estimator when the assumptions are weaker than the true data property. Specifically, we implement the methods of Cases 1–4 as shown below.

1. *The Case 1 method* This method is based on the weakest assumption of ϵ, taking only the conditional mean and variance. Thus, to implement the method of Case 1, we adopt the standardized Chi-squared distribution family with degree of freedom $p(\mathbf{x})$ for the model $f_{\epsilon|\mathbf{X}}(\epsilon, \mathbf{x})$, as given in (9). From this, we calculate $E(\epsilon^3|\mathbf{X})$ and $E(t^2|\mathbf{X})$ for the efficient score function (3), regardless of how the ϵ_i's are generated. We then construct the efficient score function from (3) and proceed to obtain the semiparametric efficient estimator for each simulation. We select the

degree of freedom $p(\mathbf{x})$ for Simulations 1, 2, and 4 using $p(\mathbf{x}) = x_1 + x_2 + 0.5$. This $p(\mathbf{x})$ is the true degree of freedom for Simulation 1. For the estimation of Simulation 3, we use the degree of freedom $p(\mathbf{x}) = 10(x_1 + x_2 + 1)$, for an acceptable performance.

2. *The Case 2 method* This method assumes that ϵ and $\mathbf{X}$ are independent, although $f_\epsilon(\epsilon)$ is not necessarily symmetric. To implement the model of Case 2, we use a standardized Chi-squared distribution $f_\epsilon(\epsilon)$ with degree of freedom q, as given in (10). Using this distribution, we calculate $E(\epsilon^3)$, $E(t^2)$, and $f'_\epsilon(\epsilon)/f_\epsilon(\epsilon)$, to form the efficient score function (4). We then proceed to obtain the semiparametric efficient estimator for each simulation. For Simulations 1 and 2, we set $q = 13$, which is the true parameter for Simulation 2. For Simulations 3 and 4, we use $q = 21$ and $q = 180$, respectively.
3. *The Case 3 method* This method assumes $f_{\epsilon|\mathbf{X}}(\epsilon, \mathbf{x})$ to be a symmetric function of ϵ, although ϵ and X are not necessarily independent. To reflect this distribution property in Simulations 1, 2, and 4, we adopt the generalized normal distribution $\mathrm{GN}(0, s, k)$ for $f_{\epsilon|\mathbf{X}}(\epsilon, \mathbf{x})$. Here, we set the scale parameter $s = \sqrt{\Gamma(1/k)/\Gamma(3/k)}$ to ensure the unit variance property; we then set different values for the shape parameter k in (11). By letting k depend on $\mathbf{X}$ through $k = 0.06(X_1 + X_2) + 1.5$, we establish the dependence between ϵ and $\mathbf{X}$. We calculate $E(t^2|\mathbf{X})$ and $f'_{\epsilon|\mathbf{X}}(\epsilon, \mathbf{X})/f_{\epsilon|\mathbf{X}}(\epsilon, \mathbf{X})$ based on the above model, to form the efficient score function (5). For Simulation 3, we illustrate the performance of the efficient estimator by adopting the mixture of the generalized normal distribution that generated the data.
4. *The Case 4 method* This method is based on the symmetric density $f_\epsilon(\epsilon)$, which does not involve $\mathbf{x}$. For the estimations in Simulations 1, 2, and 4, we use the Logistic$(0, \sqrt{3}/\pi)$ distribution for symmetric $f_\epsilon(\epsilon)$. Note that Logistic $(0, \sqrt{3}/\pi)$ is the true density for Simulation 4. For Simulation 3, we implement the GN$(0, \sqrt{\Gamma(1/k)/\Gamma(3/k)}, k)$ model for $f_\epsilon(\epsilon)$, where $k = 5.4$.

For a detailed description of the efficient score function for the calculation of each method, see Section S.11.

The results of the four methods used for Simulations 1–4 are given in Tables 1, 2, 3, and 4, respectively. From the results, we make the following observations. First, in terms of estimation consistency, (1) only the estimator of Case 1 is consistent for Simulation 1; (2) the estimators of both Cases 1 and 2 are consistent for Simulation 2; (3) the estimators of both Cases 1 and 3 are consistent for Simulation 3; and (4) all the estimators are consistent for Simulation 4. The separation of consistent and inconsistent results is best reflected in the coverage of the 95% confidence intervals. When an estimator is inconsistent, it has a large bias compared to its estimation standard error, leading to a much lower coverage rate than the nominal level. This observation confirms that to achieve consistency, we need to make only right or weaker assumptions on the error distribution. Any stronger assumptions made during the estimation procedure that do not satisfy the true error structure can lead to inconsistencies.

Further, compared to the other methods used for the same simulation settings, we observe that the estimation variabilities are minimized for the method of Case 1 in Simulation 1, Case 2 in Simulation 2, Case 3 in Simulation 3, and Case 4 in Simulation

Table 1 Results of Simulation 1

Estimator		Bias	Bias (%)	Var	Var1	95% cov (%)	95% CI
Case 1							
$\widehat{\alpha}_0$	5.0030	0.0030	0.06	0.0327	0.0300	95.8	(4.6898, 5.3538)
$\widehat{\alpha}_1$	0.0799	−0.0001	−0.14	$3.25e^{-5}$	$3.18e^{-5}$	95.9	(0.0693, 0.0909)
$\widehat{\alpha}_2$	−0.1499	0.0001	−0.06	$5.34e^{-5}$	$5.07e^{-5}$	94.1	(−0.1638, −0.1352)
$\widehat{\beta}_1$	0.0594	−0.0006	−1.02	$9.73e^{-5}$	$8.53e^{-5}$	93.7	(0.0399, 0.0788)
$\widehat{\beta}_2$	0.0401	0.0001	0.36	$3.88e^{-5}$	$3.59e^{-5}$	93.8	(0.0276, 0.0526)
Case 2							
$\widehat{\alpha}_0$	5.1229	0.1229	2.46	0.0278	0.0261	88.2	(4.8038, 5.4518)
$\widehat{\alpha}_1$	0.0772	−0.0028	−3.45	$2.86e^{-5}$	$2.80e^{-5}$	91.4	(0.0670, 0.0877)
$\widehat{\alpha}_2$	−0.1507	−0.0007	0.45	$5.13e^{-5}$	$4.74e^{-5}$	94.1	(−0.1648, −0.1364)
$\widehat{\beta}_1$	0.0599	−0.0001	−0.16	$7.22e^{-5}$	$6.26e^{-5}$	93.7	(0.0429, 0.0750)
$\widehat{\beta}_2$	0.0425	0.0025	6.20	$2.88e^{-5}$	$2.63e^{-5}$	91.4	(0.0318, 0.0529)
Case 3							
$\widehat{\alpha}_0$	4.8337	−0.1663	−3.33	0.0474	0.0449	82.5	(4.4554, 5.2788)
$\widehat{\alpha}_1$	0.0835	0.0035	4.32	$4.39e^{-5}$	$4.29e^{-5}$	91.1	(0.0707, 0.0964)
$\widehat{\alpha}_2$	−0.1435	0.0065	−4.34	$6.42e^{-5}$	$5.95e^{-5}$	85.9	(−0.1597, −0.1275)
$\widehat{\beta}_1$	0.0594	−0.0006	−1.02	$1.04e^{-4}$	$8.91e^{-5}$	93.2	(0.0394, 0.0787)
$\widehat{\beta}_2$	0.0402	0.0002	0.53	$4.33e^{-5}$	$3.82e^{-5}$	93.0	(0.0270, 0.0535)
Case 4							
$\widehat{\alpha}_0$	4.8689	−0.1311	−2.62	0.0435	0.0416	86.7	(4.5238, 5.3038)
$\widehat{\alpha}_1$	0.0822	0.0022	2.80	$4.17e^{-5}$	$4.23e^{-5}$	93.6	(0.0697, 0.0946)
$\widehat{\alpha}_2$	−0.1559	−0.0059	3.94	$6.57e^{-5}$	$6.40e^{-5}$	90.5	(−0.1712, −0.1396)
$\widehat{\beta}_1$	0.0594	−0.0006	−0.96	$9.54e^{-5}$	$8.95e^{-5}$	94.5	(0.0398, 0.0775)
$\widehat{\beta}_2$	0.0414	0.0014	3.47	$4.03e^{-5}$	$3.88e^{-5}$	94.2	(0.0289, 0.0536)

For the data corresponding to the model of Case 1, the table uses the estimation methods of Cases 1–4. The estimators median, the estimators median bias, the sample variance (Var) of 1000 estimators, and the median of 1000 estimated variances (Var1) are presented. The results are based on 500 and 1000 simulations, where $\boldsymbol{\alpha} = (5, 0.08, -0.15)^{\mathrm{T}}$ and $\boldsymbol{\beta} = (0.06, 0.04)^{\mathrm{T}}$, respectively
The bolded method shows better result than other methods

4. This confirms the efficiency results by which the error properties should be exploited to maximize the benefits of those properties. In these cases, note that the estimated variances are very close to the sample variances, leading to a better 95% coverage rate compared to the other estimation methods.

Third, when a weaker assumption is made with regard to the error structure, such as for the methods of (1) Case 1 in Simulations 2, 3, and 4 and the methods of (2) Cases 1, 2, and 3 in Simulation 4, the resulting estimators retain their consistency. In addition, the inferences are reasonably good: The estimated variances are close to the sample variances, and the 95% confidence interval coverage rates are also close to the nominal level. The cost of such "wasteful" practices affects estimation efficiency;

Table 2 Results of Simulation 2

Estimator		Bias	Bias (%)	Var	Var1	95% cov (%)	95% CI
Case 1							
$\widehat{\alpha}_0$	5.0013	0.0013	0.03	0.0437	0.0407	94.7	(4.6343, 5.4238)
$\widehat{\alpha}_1$	0.0802	0.0002	0.19	$3.90e^{-5}$	$3.84e^{-5}$	94.7	(0.0677, 0.0925)
$\widehat{\alpha}_2$	−0.1506	−0.0006	0.41	$5.27e^{-5}$	$5.50e^{-5}$	95.6	(−0.1650, −0.1367)
$\widehat{\beta}_1$	0.0594	−0.0006	−0.95	$9.44e^{-5}$	$8.25e^{-5}$	93.6	(0.0401, 0.0782)
$\widehat{\beta}_2$	0.0401	0.0001	0.26	$4.00e^{-5}$	$3.47e^{-5}$	94.0	(0.0273, 0.0519)
Case 2							
$\widehat{\alpha}_0$	5.0032	0.0032	0.06	0.0386	0.0345	94.3	(4.6509, 5.4020)
$\widehat{\alpha}_1$	0.0800	−0.0000	−0.03	$3.55e^{-5}$	$3.55e^{-5}$	94.0	(0.0682, 0.0920)
$\widehat{\alpha}_2$	−0.1508	−0.0008	0.56	$4.87e^{-5}$	$4.87e^{-5}$	95.7	(−0.1644, −0.1372)
$\widehat{\beta}_1$	0.0593	−0.0007	−1.20	$6.35e^{-5}$	$6.36e^{-5}$	94.4	(0.0446, 0.0747)
$\widehat{\beta}_2$	0.0400	0.0000	0.07	$2.83e^{-5}$	$2.83e^{-5}$	92.6	(0.0288, 0.0503)
Case 3							
$\widehat{\alpha}_0$	4.8672	−0.1328	−2.66	0.0507	0.0495	88.4	(4.4492, 5.3332)
$\widehat{\alpha}_1$	0.0828	0.0028	3.56	$4.42e^{-5}$	$4.51e^{-5}$	93.1	(0.0692, 0.0954)
$\widehat{\alpha}_2$	−0.1440	0.0060	−3.97	$5.91e^{-5}$	$6.09e^{-5}$	87.2	(−0.1601, −0.1303)
$\widehat{\beta}_1$	0.0594	−0.0006	−1.02	$9.92e^{-5}$	$8.58e^{-5}$	93.0	(0.0404, 0.0781)
$\widehat{\beta}_2$	0.0399	−0.0001	−0.19	$4.42e^{-5}$	$3.73e^{-5}$	92.6	(0.0261, 0.0525)
Case 4							
$\widehat{\alpha}_0$	4.9514	−0.0486	−0.97	0.0538	0.0514	93.1	(4.5369, 5.4438)
$\widehat{\alpha}_1$	0.0809	0.0009	1.07	$4.56e^{-5}$	$4.75e^{-5}$	95.2	(0.0668, 0.0933)
$\widehat{\alpha}_2$	−0.1575	−0.0075	4.98	$6.43e^{-5}$	$6.70e^{-5}$	86.8	(−0.1735, −0.1422)
$\widehat{\beta}_1$	0.0593	−0.0007	−1.11	$8.72e^{-5}$	$8.59e^{-5}$	94.6	(0.0414, 0.0778)
$\widehat{\beta}_2$	0.0403	0.0003	0.63	$4.08e^{-5}$	$3.80e^{-5}$	94.0	(0.0270, 0.0520)

For the data corresponding to the model of Case 2, the table uses the estimation methods of Cases 1–4. The estimators median, the estimators median bias, the sample variances (Var) of 1000 estimators, and the median of 1000 estimated estimator variances (Var1) are presented. The results are based on 500 and 1000 simulations, where $\boldsymbol{\alpha} = (5, 0.08, -0.15)^{\mathrm{T}}$ and $\boldsymbol{\beta} = (0.06, 0.04)^{\mathrm{T}}$, respectively
The bolded method shows better result than other methods

that is, they have large estimation variability compared to the corresponding efficient estimators.

Finally, for each simulation, the estimators with stronger assumptions on the error structure have smaller estimation variance. For example, in all the tables, the Case 1 method has the largest variance, while the Case 4 method has the smallest variance. The Case 2 and Case 3 methods fall between the Case 1 and Case 4 methods, while the relation between the Case 2 and Case 3 methods is not conclusive. This is a direct consequence of the indefinite relation between the Case 3 and Case 4 assumptions. Thus, to minimize the estimation variability, we need to utilize as much properties of

Table 3 Results of Simulation 3

Estimator		Bias	Bias (%)	Var	Var1	95% cov (%)	95% CI
Case 1							
$\widehat{\alpha}_0$	5.0183	0.0183	0.37	0.0555	0.0524	93.2	(4.5425, 5.5187)
$\widehat{\alpha}_1$	0.0795	−0.0005	−0.62	$4.82e^{-5}$	$4.54e^{-5}$	93.2	(0.0657, 0.0939)
$\widehat{\alpha}_2$	−0.1501	−0.0001	0.09	$6.67e^{-5}$	$6.11e^{-5}$	94.2	(−0.1678, −0.1357)
$\widehat{\beta}_1$	0.0594	−0.0006	−0.99	$7.12e^{-5}$	$6.78e^{-5}$	94.0	(0.0422, 0.0752)
$\widehat{\beta}_2$	0.0401	0.0001	0.28	$3.10e^{-5}$	$3.03e^{-5}$	94.5	(0.0294, 0.0510)
Case 2							
$\widehat{\alpha}_0$	5.0595	0.0595	1.19	0.0656	0.0502	91.6	(4.5351, 5.5775)
$\widehat{\alpha}_1$	0.0779	−0.0021	−2.58	$5.99e^{-5}$	$4.46e^{-5}$	90.6	(0.0628, 0.0936)
$\widehat{\alpha}_2$	−0.1503	−0.0003	0.20	$9.95e^{-5}$	$6.48e^{-5}$	91.4	(−0.1698, −0.1307)
$\widehat{\beta}_1$	0.0583	−0.0017	−2.84	$1.81e^{-4}$	$7.04e^{-5}$	85.5	(0.0330, 0.0876)
$\widehat{\beta}_2$	0.0405	0.0005	1.29	$8.69e^{-5}$	$3.24e^{-5}$	86.7	(0.0237, 0.0588)
Case 3							
$\widehat{\alpha}_0$	5.0095	0.0095	0.19	0.0085	0.0064	93.7	(4.8250, 5.1925)
$\widehat{\alpha}_1$	0.0798	−0.0002	−0.28	$1.86e^{-5}$	$1.69e^{-5}$	92.6	(0.0708, 0.0877)
$\widehat{\alpha}_2$	−0.1502	−0.0002	0.16	$1.54e^{-5}$	$1.45e^{-5}$	94.4	(−0.1577, −0.1424)
$\widehat{\beta}_1$	0.0595	−0.0005	−0.77	$6.32e^{-5}$	$6.04e^{-5}$	93.2	(0.0437, 0.0747)
$\widehat{\beta}_2$	0.0400	0.0000	0.01	$2.75e^{-5}$	$2.69e^{-5}$	95.1	(0.0297, 0.0502)
Case 4							
$\widehat{\alpha}_0$	5.0219	0.0219	0.44	0.0522	0.0339	92.4	(4.5772, 5.4496)
$\widehat{\alpha}_1$	0.0792	−0.0008	−1.01	$8.98e^{-5}$	$4.83e^{-5}$	88.9	(0.0603, 0.0979)
$\widehat{\alpha}_2$	−0.1502	−0.0002	0.16	$1.53e^{-4}$	$8.78e^{-5}$	89.5	(−0.1766, −0.1262)
$\widehat{\beta}_1$	0.0588	−0.0012	−1.92	$1.41e^{-4}$	$7.04e^{-5}$	83.3	(0.0347, 0.0807)
$\widehat{\beta}_2$	0.0410	0.0010	2.45	$6.04e^{-5}$	$3.11e^{-5}$	84.6	(0.0253, 0.0556)

For the data corresponding to the model of Case 3, the table uses the estimation methods of Cases 1–4. The estimators median, the estimators median bias, the sample variances (Var) of 1000 estimators, and the median of 1000 estimated variances (Var1) are presented. The results are based on 500 and 1000 simulations, where $\boldsymbol{\alpha} = (5, 0.08, -0.15)^{\mathrm{T}}$ and $\boldsymbol{\beta} = (0.06, 0.04)^{\mathrm{T}}$, respectively
The bolded method shows better result than other methods

the error structure as possible. On the other hand, one has to be careful not to impose structures not satisfied by the error distribution, since it could lead to inconsistencies.

4.2 Data analysis

We analyze a data set of 145 US electricity producers in 1955. Nerlove (1963) suggested the following Cobb–Douglas cost function to model the economic scale.

$$\log\left(\frac{C}{P_F}\right) = \alpha_0 + \alpha_1 \log Q + \alpha_2 \log\left(\frac{P_L}{P_F}\right) + \alpha_3 \log\left(\frac{P_K}{P_F}\right) + u, \tag{12}$$

Table 4 Results of Simulation 4

Estimator		Bias	Bias (%)	Var	Var1	95% cov (%)	95% CI
Case 1							
$\widehat{\alpha}_0$	5.0087	0.0087	0.17	0.0611	0.0540	93.5	(4.5205, 5.4841)
$\widehat{\alpha}_1$	0.0803	0.0003	0.39	$5.12e^{-5}$	$4.83e^{-5}$	94.7	(0.0666, 0.0946)
$\widehat{\alpha}_2$	−0.1503	−0.0003	0.23	$7.36e^{-5}$	$6.61e^{-5}$	94.1	(−0.1679, −0.1345)
$\widehat{\beta}_1$	0.0596	−0.0004	−0.62	$1.23e^{-4}$	$1.11e^{-4}$	94.4	(0.0373, 0.0806)
$\widehat{\beta}_2$	0.0396	−0.0004	−1.00	$5.57e^{-5}$	$4.86e^{-5}$	92.5	(0.0249, 0.0548)
Case 2							
$\widehat{\alpha}_0$	5.0037	0.0037	0.07	0.0543	0.0506	94.9	(4.5547, 5.4602)
$\widehat{\alpha}_1$	0.0801	0.0001	0.12	$4.55e^{-5}$	$4.50e^{-5}$	94.9	(0.0672, 0.0942)
$\widehat{\alpha}_2$	−0.1504	−0.0004	0.27	$6.55e^{-5}$	$6.04e^{-5}$	94.8	(−0.1673, −0.1358)
$\widehat{\beta}_1$	0.0599	−0.0001	−0.16	$1.25e^{-4}$	$1.02e^{-4}$	93.3	(0.0373, 0.0809)
$\widehat{\beta}_2$	0.0394	−0.0006	−1.44	$5.46e^{-5}$	$4.48e^{-5}$	92.7	(0.0252, 0.0550)
Case 3							
$\widehat{\alpha}_0$	4.9949	−0.0051	−0.10	0.0484	0.0483	94.9	(4.5836, 5.4238)
$\widehat{\alpha}_1$	0.0800	0.0000	0.05	$4.25e^{-5}$	$4.38e^{-5}$	96.0	(0.0678, 0.0929)
$\widehat{\alpha}_2$	−0.1505	−0.0005	0.36	$6.72e^{-5}$	$6.30e^{-5}$	95.1	(−0.1672, −0.1354)
$\widehat{\beta}_1$	0.0595	−0.0005	−0.90	$1.09e^{-4}$	$9.83e^{-5}$	93.6	(0.0383, 0.0793)
$\widehat{\beta}_2$	0.0398	−0.0002	−0.51	$4.88e^{-5}$	$4.33e^{-5}$	93.0	(0.0251, 0.0537)
Case 4							
$\widehat{\alpha}_0$	5.0006	0.0006	0.01	0.0459	0.0459	94.6	(4.5852, 5.4022)
$\widehat{\alpha}_1$	0.0802	0.0002	0.29	$4.00e^{-5}$	$4.14e^{-5}$	94.5	(0.0683, 0.0929)
$\widehat{\alpha}_2$	−0.1506	−0.0006	0.39	$5.64e^{-5}$	$5.55e^{-5}$	95.3	(−0.1657, −0.1364)
$\widehat{\beta}_1$	0.0598	−0.0002	−0.28	$9.72e^{-5}$	$9.58e^{-5}$	95.1	(0.0394, 0.0775)
$\widehat{\beta}_2$	0.0398	−0.0002	−0.55	$4.37e^{-5}$	$4.24e^{-5}$	94.3	(0.0265, 0.0529)

For the data corresponding to the Case 4 model, the table uses the estimation methods of Cases 1–4. The estimators median, the estimators median bias, the sample variances (Var) of 1000 estimators, and the median of 1000 estimated variances (Var1) are presented. The results are based on 500 and 1000 simulations, where $\boldsymbol{\alpha} = (5, 0.08, -0.15)^{\mathrm{T}}$ and $\boldsymbol{\beta} = (0.06, 0.04)^{\mathrm{T}}$, respectively
The bolded method shows better result than other methods

where C is the cost, Q output, P_K capital, P_L labor, P_F fuel, and u the error term. Assume that $Y = \log(C/P_F)$, $X_1 = \log Q$, $X_2 = \log(P_L/P_F)$, and $X_3 = \log(P_K/P_F)$. Then, (12) can be simplified as

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + u. \tag{13}$$

Economic theory requires that α_1, α_2, and α_3 be positive, because the cost increases when the output, labor, and capital increase. However, we find a negative $\widehat{\alpha}_3$ in Table 5 for the ordinary least squares (OLS) estimation. Since the adjusted R^2 is 0.93, the mean model in (13) represents the data sufficiently.

Table 5 Regression estimations for the US electricity data

	(a) OLS	(b) Case 1	(c) Case 2	(d) Case 3	(e) Case 4
$\widehat{\alpha}_0$	− 4.6858 (0.7837)	− 6.2103 (0.1836)	− 6.0962 (0.1369)	− 6.9715 (0.2667)	− 6.3525 (0.2144)
$\widehat{\alpha}_1$	0.7207 (0.0003)	0.8641 (0.0004)	0.8551 (0.0003)	0.9459 (0.0036)	0.8628 (0.0003)
$\widehat{\alpha}_2$	0.5940 (0.0418)	0.4874 (0.0095)	0.4974 (0.0064)	0.4569 (0.0009)	0.4540 (0.0113)
$\widehat{\alpha}_3$	− 0.0085 (8.4253)	0.0950 (0.0080)	0.0852 (0.0058)	0.1141 (4.04e^{-5})	0.1255 (0.0094)
$\widehat{\beta}_0$		0.0499 (0.0002)	0.0301 (0.0003)	0.0726 (0.0003)	0.0592 (0.0002)
$\widehat{\beta}_1$		− 0.8568 (0.0215)	− 0.5814 (0.0399)	− 1.1940 (0.0395)	− 0.9573 (0.0264)
$\widehat{\beta}_2$		− 0.5126 (0.0121)	− 0.4098 (0.0192)	− 0.8869 (0.0269)	− 0.5883 (0.0146)
$\widehat{\beta}_3$		0.2953 (0.0292)	− 0.0358 (0.0245)	0.3963 (0.0356)	0.3064 (0.0300)

The parameter estimates and their variances for each estimation method are presented. The estimated variances are given in parentheses. (a) OLS estimation. (b) Semiparametric estimation: Case 1 with the mixed asymmetric distribution assumption of ϵ, given $\mathbf{X}$. (c) Semiparametric estimation: Case 2 with the Chi-squared distribution assumption of ϵ, $df = 8.17$. (d) Semiparametric estimation: Case 3 with the mixed generalized normal distribution assumption of ϵ, given $\mathbf{X}$. (e) Semiparametric estimation: Case 4 with the normal distribution assumption of ϵ

The OLS is the efficient estimator for $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \alpha_3)^{\mathrm{T}}$ when u is normally distributed and homoscedastic. We thus further test the normality of u in (13). We check the Shapiro–Wilk and Anderson–Darling tests using the null hypothesis H_0; "the regression error has a normal distribution." The Shapiro–Wilk test statistic is 0.92 with the p value $2.25e^{-7}$, and the Anderson–Darling test statistic is 2.19 with the p value $1.41\mathrm{e}^{-5}$ for (13). In addition, we carry out White's test to check for the null hypothesis of homoscedasticity, to obtain the test statistic 70.81 with the p value $1.06e^{-11}$. Thus, we conclude that the data set does not satisfy the normal error assumption and reveals heteroscedasticity. Therefore, we proceed to further investigate the error distribution by allowing for heteroscedastic nonnormal error, retaining the Cobb–Douglas cost function. Specifically, we consider the following heteroscedastic error model.

$$Y = m(\mathbf{X}, \boldsymbol{\alpha}) + e^{\sigma(\mathbf{X}, \boldsymbol{\beta})}\epsilon, \quad E(\epsilon|\mathbf{X}) = 0, \quad \mathrm{Var}(\epsilon|\mathbf{X}) = 1, \tag{14}$$

where $m(\mathbf{X}, \boldsymbol{\alpha}) = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3$ and $\sigma(\mathbf{X}, \boldsymbol{\beta}) = \beta_0 X_1^2 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$.

Figure 1 presents the scatter plots of the (1) Y, (2) OLS residuals, and (3) log of the OLS residual squares against the covariates. These scatter plots clearly show the relation between X_1 and Y. In the residual plot of X_1, we see pattern changes near the median value of X_1. From the plot of the OLS residual squares log against the covariates, we can find the shape of the $\sigma(\mathbf{X}, \boldsymbol{\beta})$. Because of the slight curve pattern with regard to X_1, we propose the model in (14) for $\sigma(\mathbf{X}, \boldsymbol{\beta})$, which includes the term X_1^2.

- *Case 1 method* We consider different distributions on ϵ according to $\mathbf{X}$ as follows.

$$\epsilon|\mathbf{X} \sim \begin{cases} \mathrm{GN}\left(0, \sqrt{\frac{\Gamma(1/p_1)}{\Gamma(3/p_1)}}, p_1\right), & p_1 = 1.95, \quad \text{where } X_1 \le \mathrm{median}(X_1), \\ \text{standardized } \chi^2(p_2), & p_2 = 18.29, \quad \text{where } X_1 > \mathrm{median}(X_1). \end{cases}$$

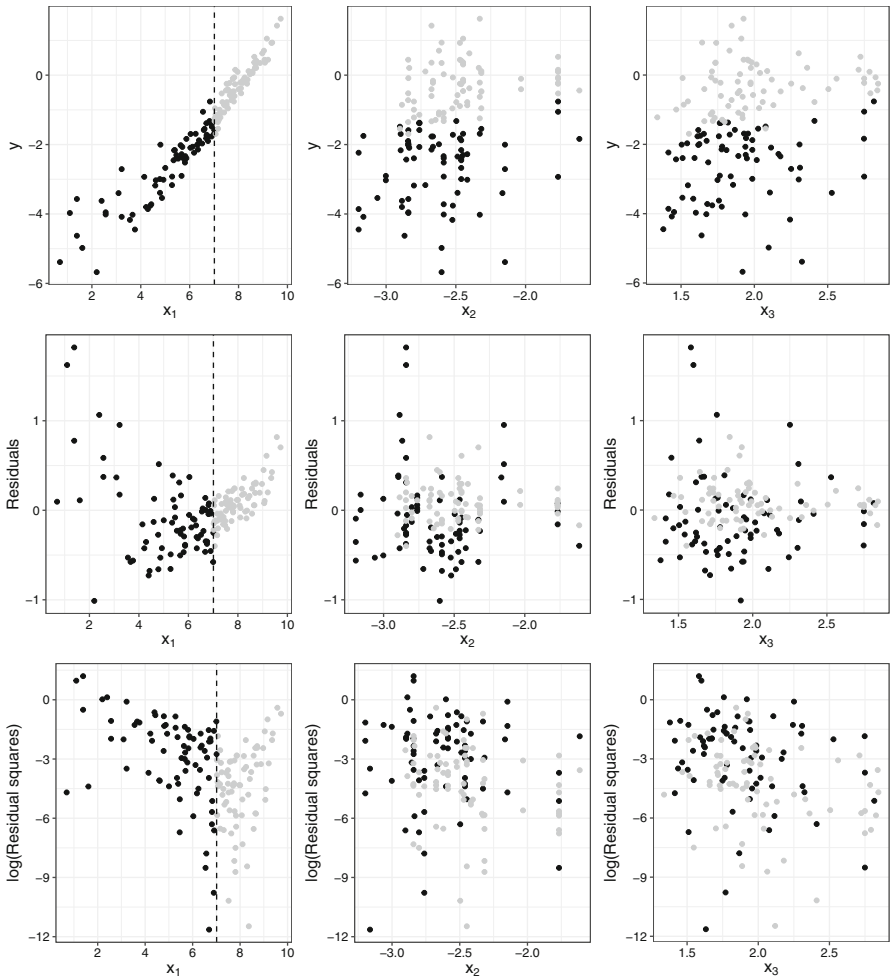

Fig. 1 US electricity data. We denote $Y = \log(C/P_F)$, $X_1 = \log Q$, $X_2 = \log(P_L/P_F)$, and $X_3 = \log(P_K/P_F)$, where C is the cost, Q output, P_K capital, P_L labor, and P_F the fuel. The scatter plots of the (1) Y, (2) OLS residuals, and (3) log (OLS residual squares) against the covariates are given. In all the plots, the data with $X_1 \leq$ median (X_1) are shown in black color and those with $X_1 >$ median (X_1) are shown in gray color. In some plots, the dotted line is drawn on the median $(X_1) = 7.01$

We obtain the degree of freedom for the above mixed distribution by minimizing the difference between the sample and theoretical skewness. Sample skewness (Joanes and Gill 1998) is calculated as

$$\frac{\sqrt{n(n-1)}}{n-2}\frac{m_3}{m_2^{3/2}},$$

where $m_3 = \sum_{i=1}^{n}(e_i - \bar{e})^3/n$, $m_2 = \sum_{i=1}^{n}(e_i - \bar{e})^2/n$, $e_i = \{y_i - m(\mathbf{x}_i, \widehat{\boldsymbol{\alpha}})\}$ $/e^{\sigma(\mathbf{x}_i, \widehat{\boldsymbol{\beta}})}$, and $\bar{e}$ is the mean of e_i's.

- *Case 2 method* We use a standardized Chi-squared distribution with $q = 8.17$ in (10). This q value minimizes the difference between the sample and theoretical skewness.
- *Case 3 method* The distribution of ϵ, given X, is assumed to be symmetric. For a flexible symmetric distribution, we use a generalized normal distribution.

$$\epsilon|\mathbf{X} \sim \begin{cases} \text{GN}\left(0, \sqrt{\frac{\Gamma(1/k)}{\Gamma(3/k)}}, k\right), & k = 0.95, \quad \text{where } X_1 \le \text{median}(X_1), \\ \text{GN}\left(0, \sqrt{\frac{\Gamma(1/k)}{\Gamma(3/k)}}, k\right), & k = 14.19, \quad \text{where } X_1 > \text{median}(X_1). \end{cases}$$

That is, the distribution of ϵ conditional on $\mathbf{X}$ is given by

$$f_{\epsilon|\mathbf{X}}(\epsilon, \mathbf{X}) = \frac{u(\mathbf{X})}{2\Gamma(1/u(\mathbf{X}))\sqrt{\Gamma(1/u(\mathbf{X}))/\Gamma(3/u(\mathbf{X}))}} \times \exp\left[-\left\{\frac{|\epsilon|}{\sqrt{\Gamma(1/u(\mathbf{X}))/\Gamma(3/u(\mathbf{X}))}}\right\}^{u(\mathbf{X})}\right],$$

where $u(\mathbf{X}) = k_1 I\{X_1 \le \text{median}(X_1)\} + k_2 I\{X_1 > \text{median}(X_1)\}$. Here, we obtain $k_1 = 0.95$ and $k_2 = 14.19$ by minimizing the difference between the sample and theoretical skewness.
- *Case 4 method* we use $N(0, 1)$ for $f_\epsilon(\epsilon)$.

The parameter estimates and their variances for Cases 1–4 are given in Table 5. Note that we obtain a positive $\widehat{\alpha}_3$ in (14) with these methods, whereas $\widehat{\alpha}_3$ is negative from OLS. Thus, our model results reflect the economic theory more appropriately. The Q–Q plots and histograms of the standardized residuals obtained for Cases 1, 2, 3, and 4 are given in Figs. 2, 3, 4, and 5, respectively . For Case 1, we assume two different distributions on ϵ according to $\mathbf{X}$. Figure 2 shows that the Q–Q plots are very close to a straight line. Also, the histograms of $\widehat{\epsilon}$ show a reasonable fit to the corresponding distribution. On the other hand, Figs. 3, 4, and 5 give slightly less convincing results in that the Q–Q plots deviate from a straight line to different extents, and/or the histograms of the residuals do not fit the distributions well. Hence, we consider the Case 1 method the most appropriate of the four different methods.

5 Discussion

We developed semiparametric efficient estimators for the heteroscedastic model under four different error scenarios and studied the issue of misspecifying the standardized error distribution through underassuming, overassuming, and misassuming some error properties. The overall message obtained is that error assumptions play an important role in mean regression in terms of both efficiency and consistency. When information is available on the error properties, such information should be taken into account when constructing efficient estimators. On the other hand, it is risky to assume arbitrary structure of error distribution without careful consideration because an incorrect assumption on an error can lead to inconsistent estimation and misleading results.

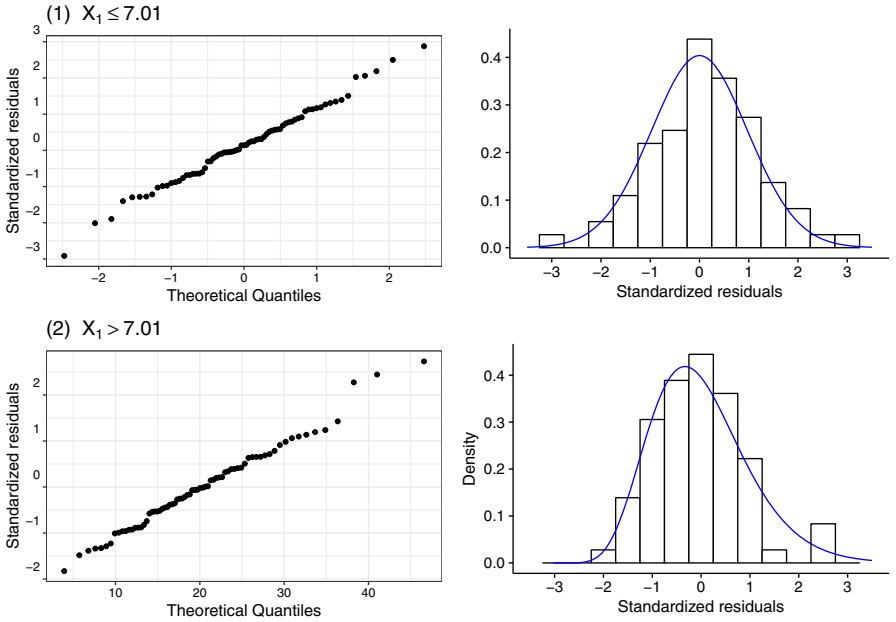

Fig. 2 Q–Q plots and the histograms of standardized residuals obtained from Case 1 for the US electricity data. The left-hand side of the figure gives the $\widehat{\epsilon}$ versus theoretical quantiles of the assumed distribution. The right-hand side shows the assumed density curve overlaid on each histogram of $\widehat{\epsilon}$

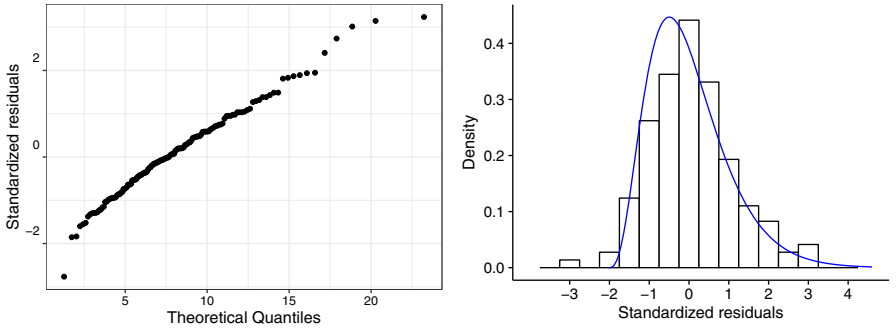

Fig. 3 Q–Q plot and the histogram of standardized residuals obtained from Case 2 for the US electricity data. The left-hand side gives the $\widehat{\epsilon}$ versus theoretical quantiles of the assumed distribution. The right-hand side shows the assumed density curve overlaid on the histogram of $\widehat{\epsilon}$

We proposed different estimation methods for the different error assumptions of a regression model. Our methods can be applied practically to the regression model based on an economic theory such as the Cobb–Douglas function, as shown in the data example. Since we work with the parametric model for the mean and variance functions, we need to have suitable parametric forms. Often, the mean model can be based on practical or scientific knowledge, whereas the variance model relies more on statistical analysis. An alternative to the model we considered here is a semiparametric

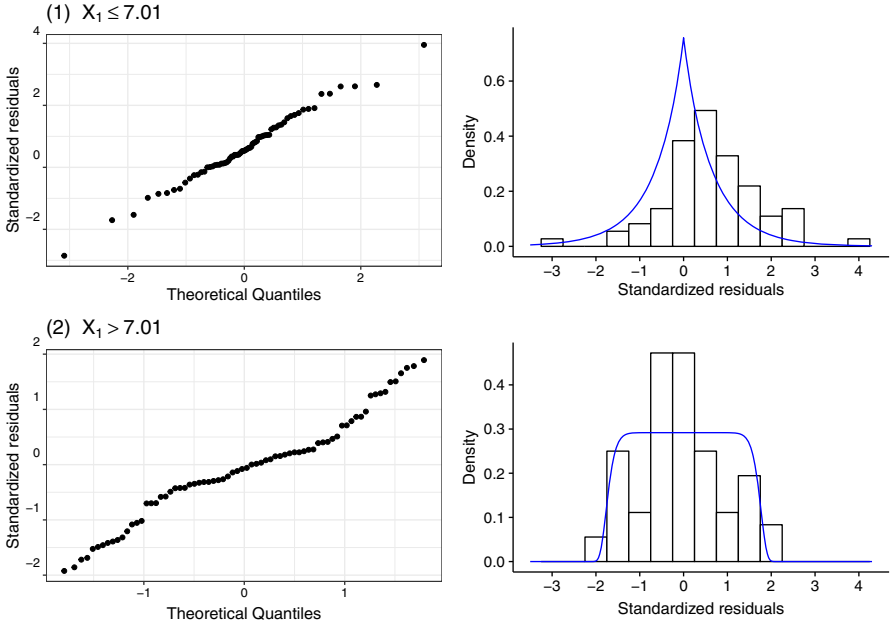

Fig. 4 Q–Q plots and the histograms of standardized residuals obtained from Case 3 for the US electricity data. The left-hand side gives the $\widehat{\epsilon}$ versus theoretical quantiles of the assumed distribution. The right-hand side shows the assumed density curve overlaid on each histogram of $\widehat{\epsilon}$

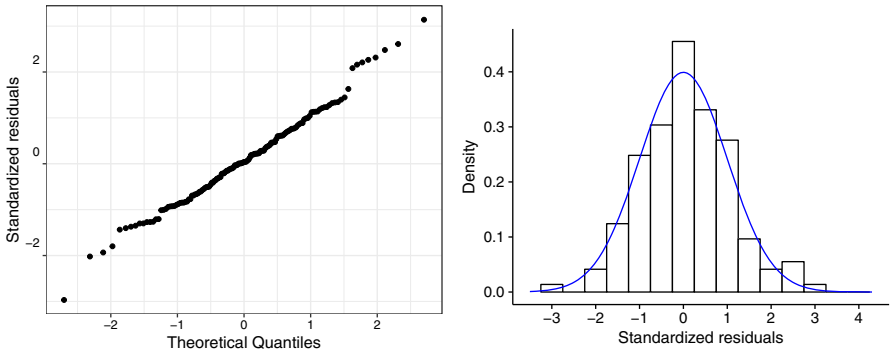

Fig. 5 Q–Q plot and the histogram of standardized residuals obtained from Case 4 for the US electricity data. The left-hand side gives the $\widehat{\epsilon}$ versus theoretical quantiles of the assumed distribution. The right-hand side shows the assumed density curve overlaid on the histogram of $\widehat{\epsilon}$

variance model along with the parametric mean model, which is a much richer class and worthy of an in-depth and thorough investigation. Characterizing and modeling standardized errors also form an important aspect of model building. In case the standardized error and covariates are independent, the Case 2 and Case 4 methods, for example, can benefit by taking advantage of this. However, in case the independent assumption does not hold, it could be difficult to further characterize and model the distribution of the standardized error, given the covariates. This would especially be

the case if the number of observation is small and the number of covariates is large. These issues can be interesting and useful extensions for a future work.

Supplementary materials

Supplement to "Semiparametric efficient estimators in heteroscedastic error models"

We provide comprehensive proofs of Propositions 2–4 and Theorems 2–8, supporting the theory. We also describe in detail the procedures of constructing the efficient score functions of the methods of Cases 1–4 used in simulations.

Acknowledgements Mijeong Kim was supported by a Ewha Womans University Research Grant of 2015 and a National Research Foundation of Korea (NRF) grant funded by the Korean Government (NRF-2017R1C1B5015186). Yanyuan Ma was supported by National Science Foundation DMS-1608540.

Appendix

A.1 Proof of Proposition 1

First, note that the construction of t ensures the property

$$E(\epsilon t|\mathbf{X}) = E\left(\epsilon^3|\mathbf{X}\right) - E\left(\epsilon^3|\mathbf{X}\right) E\left(\epsilon^2|\mathbf{X}\right) - E\left(\epsilon|\mathbf{X}\right) = E\left(\epsilon^3|\mathbf{X}\right) - E\left(\epsilon^3|\mathbf{X}\right) = 0,$$

which is crucial for the following proof.

Following the semiparametric consideration in (1), we estimate the parameters $\boldsymbol{\theta}$, while the nuisance parameter is the distribution $f_{\epsilon|\mathbf{X}}(\epsilon, \mathbf{x})$, which is subject to the constraints $E(\epsilon|\mathbf{X}) = 0$ and $E(\epsilon^2|\mathbf{X}) = 1$. Note that we do not assume any parametric distribution model on ϵ. Thus, (1) is a semiparametric model. We first write out the joint distribution of $(\epsilon, \mathbf{X})$

$$f_{\epsilon,\mathbf{X}}(\epsilon, \mathbf{x}) = f_{\epsilon|\mathbf{X}}(\epsilon, \mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) = \eta_1(\mathbf{x})\eta_2(\epsilon, \mathbf{x}),$$

where $\eta_1(\cdot)$ and $\eta_2(\cdot)$ denote the infinite-dimensional nuisance parameter function. Since $\eta_1(\cdot)$ and $\eta_2(\cdot)$ are probability density functions, we have

$$\int \eta_1(\mathbf{x})\mathrm{d}\mathbf{x} = 1, \quad \text{and} \quad \int \eta_2(\epsilon, \mathbf{x})\mathrm{d}\epsilon = 1 \quad \text{for all } \mathbf{x}. \tag{15}$$

In the Hilbert space $\mathcal{H}$ formed by all the mean zero finite variance functions, the nuisance tangent space of a semiparametric model is the mean squared closure of all the parametric submodel nuisance tangent spaces. A parametric submodel is defined as a parametric model included by the original semiparametric model and includes the true density $f(\epsilon, \mathbf{X}; \boldsymbol{\theta}_0, \boldsymbol{\gamma}_0)$. A nuisance tangent space of the parametric model $f(\epsilon, \mathbf{X}; \boldsymbol{\theta}, \boldsymbol{\gamma})$ is a linear space spanned by the nuisance score vector

$S_{\boldsymbol{\gamma}} = \partial \log f(\epsilon, \mathbf{X}; \boldsymbol{\theta}_0, \boldsymbol{\gamma})/\partial \boldsymbol{\gamma}|_{\boldsymbol{\gamma}_0}$. From the above general concept, we first derive the nuisance tangent space Λ. From Condition (2), we have

$$\int \epsilon \eta_2(\epsilon, \mathbf{x}) \mathrm{d}\epsilon = 0 \ \text{ for all } \ \mathbf{x}, \quad \text{and} \quad \int \epsilon^2 \eta_2(\epsilon, \mathbf{x}) \mathrm{d}\epsilon = 1 \quad \text{for all } \ \mathbf{x}. \tag{16}$$

From (15) and (16), it follows that

$$\begin{aligned}
\Lambda_1 &= \{\mathbf{a}(\mathbf{x}) : E\{\mathbf{a}(\mathbf{X})\} = \mathbf{0}\}, \\
\Lambda_2 &= \{\mathbf{b}(\mathbf{x}, \epsilon) : E\{\mathbf{b}(\mathbf{X}, \epsilon)|\mathbf{X}\} = E\{\epsilon \mathbf{b}(\mathbf{X}, \epsilon)|\mathbf{X}\} = E\{\epsilon^2 \mathbf{b}(\mathbf{X}, \epsilon)|\mathbf{X}\} = \mathbf{0}\} \\
&= \{\mathbf{b}(\mathbf{x}, \epsilon) : E\{\mathbf{b}(\mathbf{X}, \epsilon)|\mathbf{X}\} = E\{\epsilon \mathbf{b}(\mathbf{X}, \epsilon)|\mathbf{X}\} = E\{t \mathbf{b}(\mathbf{X}, \epsilon)|\mathbf{X}\} = \mathbf{0}\}.
\end{aligned}$$

In the above, we use

$$\begin{aligned}
E\{t\mathbf{b}(\mathbf{X}, \epsilon)|\mathbf{X}\} &= E\{\epsilon^2 \mathbf{b}(\mathbf{X}, \epsilon)|\mathbf{X}\} - E(\epsilon^3|\mathbf{X}) E\{\epsilon \mathbf{b}(\mathbf{X}, \epsilon)|\mathbf{X}\} - E\{\mathbf{b}(\mathbf{X}, \epsilon)|\mathbf{X}\} \\
&= E\{\epsilon^2 \mathbf{b}(\mathbf{X}, \epsilon)|\mathbf{X}\} = \mathbf{0}.
\end{aligned}$$

Combining Λ_1 and Λ_2, the nuisance tangent space Λ can be written as

$$\begin{aligned}
\Lambda &= \Lambda_1 \oplus \Lambda_2 \\
&= \{\mathbf{h}_1(\mathbf{x}) + \mathbf{h}_2(\mathbf{x}, \epsilon) : E\{\mathbf{h}_1(\mathbf{X})\} = E\{\mathbf{h}_2(\mathbf{x}, \epsilon)|\mathbf{X}\} = E\{\epsilon \mathbf{h}_2(\mathbf{x}, \epsilon)|\mathbf{X}\} \\
&\quad = E\{t\mathbf{h}_2(\mathbf{X}, \epsilon)|\mathbf{X}\} = \mathbf{0}\} \\
&= \{\mathbf{h}(\mathbf{x}, \epsilon) : E\{\mathbf{h}(\mathbf{x}, \epsilon)\} = E\{\epsilon \mathbf{h}(\mathbf{x}, \epsilon)|\mathbf{X}\} = E\{t\mathbf{h}(\mathbf{X}, \epsilon)|\mathbf{X}\} = \mathbf{0}\}.
\end{aligned}$$

In order to find $\Lambda^{\perp}$, we let $\mathbf{K} = \{\mathbf{g}(\mathbf{x}, \epsilon) : \mathbf{g}(\mathbf{x}, \epsilon) = \mathbf{g}_1(\mathbf{x})\epsilon + \mathbf{g}_2(\mathbf{x})t\}$ and have $\Lambda^{\perp} = \mathbf{K}$ by showing that $\mathbf{K} \subset \Lambda^{\perp}$ and $\Lambda^{\perp} \subset \mathbf{K}$.

For arbitrary $\mathbf{h}(\mathbf{x}, \epsilon) \in \Lambda$ and $\mathbf{g}(\mathbf{x}, \epsilon) = \mathbf{g}_1(\mathbf{x})\epsilon + \mathbf{g}_2(\mathbf{x})t \in \mathbf{K}$,

$$\begin{aligned}
&E\{\mathbf{h}(\mathbf{X}, \epsilon)^{\mathrm{T}} \mathbf{g}(\mathbf{X}, \epsilon)\} \\
&\quad = E\left[E\{\mathbf{h}(\mathbf{X}, \epsilon)^{\mathrm{T}} \mathbf{g}(\mathbf{X}, \epsilon)|\mathbf{X}\}\right] \\
&\quad = E\left[E\{\epsilon \mathbf{h}(\mathbf{X}, \epsilon)|\mathbf{X}\}^{\mathrm{T}} \mathbf{g}_1(\mathbf{X})\right] + E\left[E\{t\mathbf{h}(\mathbf{X}, \epsilon)|\mathbf{X}\}^{\mathrm{T}} \mathbf{g}_2(\mathbf{X})\right] = 0.
\end{aligned}$$

Hence, $\mathbf{g}(\mathbf{x}, \epsilon) = \mathbf{g}_1(\mathbf{x})\epsilon + \mathbf{g}_2(\mathbf{x})t \in \Lambda^{\perp}$. Thus, $\mathbf{K} \subset \Lambda^{\perp}$.

Now, we will show $\Lambda^{\perp} \subset \mathbf{K}$.

For arbitrary $\mathbf{h}(\mathbf{x}, \epsilon) \in \Lambda^{\perp}$, we can decompose $\mathbf{h}(\mathbf{x}, \epsilon)$ into $\mathbf{h}(\mathbf{x}, \epsilon) = \mathbf{r}_1(\mathbf{x}, \epsilon) + \mathbf{r}_2(\mathbf{x}, \epsilon)$, where

$$\begin{aligned}
\mathbf{r}_1(\mathbf{X}, \epsilon) &= \mathbf{h}(\mathbf{X}, \epsilon) - E\{\epsilon \mathbf{h}(\mathbf{X}, \epsilon)|\mathbf{X}\}\epsilon - \frac{E\{t\mathbf{h}(\mathbf{X}, \epsilon)|\mathbf{X}\}}{E(t^2|\mathbf{X})} t, \\
\mathbf{r}_2(\mathbf{X}, \epsilon) &= E\{\epsilon \mathbf{h}(\mathbf{X}, \epsilon)|\mathbf{X}\}\epsilon + \frac{E\{t\mathbf{h}(\mathbf{X}, \epsilon)|\mathbf{X}\}}{E(t^2|\mathbf{X})} t.
\end{aligned}$$

It is obvious that $\mathbf{r}_2(\mathbf{x}, \epsilon) \in \mathbf{K}$. Since $\mathbf{r}_2(\mathbf{x}, \epsilon) \in \mathbf{K} \subset \Lambda^{\perp}$ and $\mathbf{h}(\mathbf{x}, \epsilon) \in \Lambda^{\perp}$, we have $\mathbf{r}_1(\mathbf{x}, \epsilon) = \mathbf{h}(\mathbf{x}, \epsilon) - \mathbf{r}_2(\mathbf{x}, \epsilon) \in \Lambda^{\perp}$. For $\mathbf{r}_1$, we can easily verify that

$$E\{\mathbf{r}_1(\mathbf{X}, \epsilon)\} = E\{\epsilon \mathbf{r}_1(\mathbf{X}, \epsilon)|\mathbf{X}\} = E\{t\mathbf{r}_1(\mathbf{X}, \epsilon)|\mathbf{X}\} = \mathbf{0}.$$

This indicates that $\mathbf{r}_1(\mathbf{x}, \epsilon) \in \Lambda$. Since $\mathbf{r}_1(\mathbf{x}, \epsilon)$ is also an element of $\Lambda^{\perp}$, it follows that $\mathbf{r}_1(\mathbf{x}, \epsilon) = \mathbf{0}$, and we obtain $\mathbf{h}(\mathbf{x}, \epsilon) = \mathbf{r}_2(\mathbf{x}, \epsilon) \in \mathbf{K}$. We next show that $\mathbf{h}(\mathbf{x}, \epsilon) \in \mathbf{K}$ for arbitrary $\mathbf{h}(\mathbf{x}, \epsilon) \in \Lambda^{\perp}$. Thus, $\Lambda^{\perp} \subset \mathbf{K}$.

Consequently, we have

$$\Lambda^{\perp} = \{\mathbf{g}(\mathbf{x}, \epsilon) : \mathbf{g}(\mathbf{x}, \epsilon) = \mathbf{g}_1(\mathbf{x})\epsilon + \mathbf{g}_2(\mathbf{x})t\}.$$

□

A.2 Proof of Theorem 1

In model (1), we obtain the score functions of $\boldsymbol{\theta} = (\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$ as

$$\begin{aligned}
\boldsymbol{S}_{\boldsymbol{\alpha}} &= \frac{\partial \log f_{\mathbf{X},Y}(\mathbf{X}, y)}{\partial \boldsymbol{\alpha}} = -\frac{\partial f_{\epsilon|\mathbf{X}}(\epsilon, \mathbf{X})/\partial \epsilon}{f_{\epsilon|\mathbf{X}}(\epsilon, \mathbf{X})} e^{-\sigma(\mathbf{X}, \boldsymbol{\beta})} \mathbf{m}'_{\boldsymbol{\alpha}}(\mathbf{X}, \boldsymbol{\alpha}), \\
\boldsymbol{S}_{\boldsymbol{\beta}} &= \frac{\partial \log f_{\mathbf{X},Y}(\mathbf{X}, y)}{\partial \boldsymbol{\beta}} = -\frac{\partial f_{\epsilon|\mathbf{X}}(\epsilon, \mathbf{X})/\partial \epsilon}{f_{\epsilon|\mathbf{X}}(\epsilon, \mathbf{X})} \epsilon \boldsymbol{\sigma}'_{\boldsymbol{\beta}}(\mathbf{X}, \boldsymbol{\beta}) - \boldsymbol{\sigma}'_{\boldsymbol{\beta}}(\mathbf{X}, \boldsymbol{\beta}).
\end{aligned}$$

By projecting the above score vectors onto $\Lambda^{\perp}$, we can find the efficient score vectors of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

First, $\boldsymbol{S}_{\mathrm{eff},\boldsymbol{\alpha}}$ and $\boldsymbol{S}_{\mathrm{eff},\boldsymbol{\beta}}$ are the form of $\mathbf{g}_1(\mathbf{x})\epsilon + \mathbf{g}_2(\mathbf{x})t$. Thus, $\boldsymbol{S}_{\mathrm{eff}} \in \Lambda^{\perp}$.

Now, we verify that $\boldsymbol{S}_{\boldsymbol{\theta}} - \boldsymbol{S}_{\mathrm{eff}} \in \Lambda$. $\boldsymbol{S}_{\boldsymbol{\theta}} - \boldsymbol{S}_{\mathrm{eff}}$ is given by

$$\begin{aligned}
\boldsymbol{S}_{\boldsymbol{\alpha}} - \boldsymbol{S}_{\mathrm{eff},\boldsymbol{\alpha}} &= e^{-\sigma(\mathbf{X}, \boldsymbol{\beta})} \mathbf{m}'_{\boldsymbol{\alpha}} \left\{ -\frac{\partial f_{\epsilon|\mathbf{X}}(\epsilon, \mathbf{X})/\partial \epsilon}{f_{\epsilon|\mathbf{X}}(\epsilon, \mathbf{X})} - \epsilon + \frac{E(\epsilon^3|\mathbf{X})}{E(t^2|\mathbf{X})} t \right\}, \\
\boldsymbol{S}_{\boldsymbol{\beta}} - \boldsymbol{S}_{\mathrm{eff},\boldsymbol{\beta}} &= \boldsymbol{\sigma}'_{\boldsymbol{\beta}}(\mathbf{X}, \boldsymbol{\beta}) \left\{ -\frac{\partial f_{\epsilon|\mathbf{X}}(\epsilon, \mathbf{X})/\partial \epsilon}{f_{\epsilon|\mathbf{X}}(\epsilon, \mathbf{X})} \epsilon - 1 - \frac{2t}{E(t^2|\mathbf{X})} \right\}.
\end{aligned}$$

We verify $\boldsymbol{S}_{\boldsymbol{\theta}} - \boldsymbol{S}_{\mathrm{eff}} \in \Lambda$ by showing the following.

$$E(\boldsymbol{S}_{\boldsymbol{\theta}} - \boldsymbol{S}_{\mathrm{eff}}) = \mathbf{0}, \quad E\{\epsilon(\boldsymbol{S}_{\boldsymbol{\theta}} - \boldsymbol{S}_{\mathrm{eff}})|\mathbf{X}\} = \mathbf{0}, \quad E\{t(\boldsymbol{S}_{\boldsymbol{\theta}} - \boldsymbol{S}_{\mathrm{eff}})|\mathbf{X}\} = \mathbf{0}.$$

We can check the details as follows.

$$\begin{aligned}
E(\boldsymbol{S}_{\boldsymbol{\alpha}} - \boldsymbol{S}_{\mathrm{eff},\boldsymbol{\alpha}}|\mathbf{X}) &= e^{-\sigma(\mathbf{X}, \boldsymbol{\beta})} \mathbf{m}'_{\boldsymbol{\alpha}}(\mathbf{X}, \boldsymbol{\alpha}) E\left\{ -\frac{\partial f_{\epsilon|\mathbf{X}}(\epsilon, \mathbf{X})/\partial \epsilon}{f_{\epsilon|\mathbf{X}}(\epsilon, \mathbf{X})} - \epsilon + \frac{E(\epsilon^3|\mathbf{X})}{E(t^2|\mathbf{X})} t|\mathbf{X} \right\} = \mathbf{0}, \\
E(\boldsymbol{S}_{\boldsymbol{\beta}} - \boldsymbol{S}_{\mathrm{eff},\boldsymbol{\beta}}|\mathbf{X}) &= \boldsymbol{\sigma}'_{\boldsymbol{\beta}}(\mathbf{X}, \boldsymbol{\beta}) E\left\{ -\frac{\partial f_{\epsilon|\mathbf{X}}(\epsilon, \mathbf{X})/\partial \epsilon}{f_{\epsilon|\mathbf{X}}(\epsilon, \mathbf{X})} \epsilon - 1 - \frac{2t}{E(t^2|\mathbf{X})} |\mathbf{X} \right\} = \mathbf{0}.
\end{aligned}$$

The above results imply that $E(\boldsymbol{S}_{\boldsymbol{\theta}} - \boldsymbol{S}_{\text{eff}}) = \mathbf{0}$.

$$
\begin{aligned}
E\{\epsilon(\boldsymbol{S}_{\boldsymbol{\alpha}} - \boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}})|\mathbf{X}\} &= e^{-\sigma(\mathbf{X},\boldsymbol{\beta})}\mathbf{m}'_{\boldsymbol{\alpha}}(\mathbf{X},\boldsymbol{\alpha}) \\
&\quad E\left\{-\frac{\partial f_{\epsilon|\mathbf{X}}(\epsilon,\mathbf{X})/\partial\epsilon}{f_{\epsilon|\mathbf{X}}(\epsilon,\mathbf{X})}\epsilon - \epsilon^2 + \frac{E(\epsilon^3|\mathbf{X})}{E(t^2|\mathbf{X})}\epsilon t|\mathbf{X}\right\} = \mathbf{0}, \\
E\{t(\boldsymbol{S}_{\boldsymbol{\alpha}} - \boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}})|\mathbf{X}\} &= e^{-\sigma(\mathbf{X},\boldsymbol{\beta})}\mathbf{m}'_{\boldsymbol{\alpha}}(\mathbf{X},\boldsymbol{\alpha}) \\
&\quad E\left\{-\frac{\partial f_{\epsilon|\mathbf{X}}(\epsilon,\mathbf{X})/\partial\epsilon}{f_{\epsilon|\mathbf{X}}(\epsilon,\mathbf{X})}t - \epsilon t + \frac{E(\epsilon^3|\mathbf{X})}{E(t^2|\mathbf{X})}t^2|\mathbf{X}\right\} = \mathbf{0}, \\
E\{\epsilon(\boldsymbol{S}_{\boldsymbol{\beta}} - \boldsymbol{S}_{\text{eff},\boldsymbol{\beta}})|\mathbf{X}\} &= \boldsymbol{\sigma}'_{\boldsymbol{\beta}}(\mathbf{X},\boldsymbol{\beta})E\left\{-\frac{\partial f_{\epsilon|\mathbf{X}}(\epsilon,\mathbf{X})/\partial\epsilon}{f_{\epsilon|\mathbf{X}}(\epsilon,\mathbf{X})}\epsilon^2 - \epsilon - \frac{2\epsilon t}{E(t^2|\mathbf{X})}|\mathbf{X}\right\} = \mathbf{0}, \\
E\{t(\boldsymbol{S}_{\boldsymbol{\beta}} - \boldsymbol{S}_{\text{eff},\boldsymbol{\beta}})|\mathbf{X}\} &= \boldsymbol{\sigma}'_{\boldsymbol{\beta}}(\mathbf{X},\boldsymbol{\beta})E\left\{-\frac{\partial f_{\epsilon|\mathbf{X}}(\epsilon,\mathbf{X})/\partial\epsilon}{f_{\epsilon|\mathbf{X}}(\epsilon,\mathbf{X})}\epsilon t - t - \frac{2t^2}{E(t^2|\mathbf{X})}|\mathbf{X}\right\} = \mathbf{0}.
\end{aligned}
$$

The above equations verify that $\boldsymbol{S}_{\boldsymbol{\theta}} - \boldsymbol{S}_{\text{eff}} \in \Lambda$.

Now, we find the optimal efficiency matrix. First, we calculate $\boldsymbol{S}_{\text{eff}}\boldsymbol{S}_{\text{eff}}^{\text{T}}$.

$$
\boldsymbol{S}_{\text{eff}}\boldsymbol{S}_{\text{eff}}^{\text{T}} = \begin{Bmatrix} \boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}}\boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}}^{\text{T}} & \boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}}\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}^{\text{T}} \\ \boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}\boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}}^{\text{T}} & \boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}^{\text{T}} \end{Bmatrix},
$$

where

$$
\begin{aligned}
\boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}}\boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}}^{\text{T}} &= \left\{\epsilon - \frac{E(\epsilon^3|\mathbf{X})}{E(t^2|\mathbf{X})}t\right\}^2 e^{-2\sigma(\mathbf{X},\boldsymbol{\beta})}\mathbf{m}'_{\boldsymbol{\alpha}}(\mathbf{X},\boldsymbol{\alpha})\mathbf{m}'^{\text{T}}_{\boldsymbol{\alpha}}(\mathbf{X},\boldsymbol{\alpha}), \\
\boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}}\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}^{\text{T}} &= \frac{2te^{-\sigma(\mathbf{X},\boldsymbol{\beta})}}{E(t^2|\mathbf{X})}\left\{\epsilon - \frac{E(\epsilon^3|\mathbf{X})}{E(t^2|\mathbf{X})}t\right\}\mathbf{m}'_{\boldsymbol{\alpha}}(\mathbf{X},\boldsymbol{\alpha})\boldsymbol{\sigma}'^{\text{T}}_{\boldsymbol{\beta}}(\mathbf{X},\boldsymbol{\beta}), \\
\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}^{\text{T}} &= \frac{4t^2}{\{E(t^2|\mathbf{X})\}^2}\boldsymbol{\sigma}'_{\boldsymbol{\beta}}(\mathbf{X},\boldsymbol{\beta})\boldsymbol{\sigma}'^{\text{T}}_{\boldsymbol{\beta}}(\mathbf{X},\boldsymbol{\beta}).
\end{aligned}
$$

Since $E(\boldsymbol{S}_{\text{eff}}\boldsymbol{S}_{\text{eff}}^{\text{T}}) = E\{E(\boldsymbol{S}_{\text{eff}}\boldsymbol{S}_{\text{eff}}{}^{\text{T}}|\mathbf{X})\}$, we have each block inside $E(\boldsymbol{S}_{\text{eff}}\boldsymbol{S}_{\text{eff}}^{\text{T}})$ as follows.

1. $E(\boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}}\boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}}^{\text{T}})$ is equivalent to

$$
E\left(E\left[\left\{\epsilon - \frac{E(\epsilon^3|\mathbf{X})}{E(t^2|\mathbf{X})}t\right\}^2|\mathbf{X}\right]e^{-2\sigma(\mathbf{X},\boldsymbol{\beta})}\mathbf{m}'_{\boldsymbol{\alpha}}(\mathbf{X},\boldsymbol{\alpha})\mathbf{m}'^{\text{T}}_{\boldsymbol{\alpha}}(\mathbf{X},\boldsymbol{\alpha})\right).
$$

From the above expression, we can rewrite the part $E\left[\left\{\epsilon - \frac{E(\epsilon^3|\mathbf{X})}{E(t^2|\mathbf{X})}t\right\}^2|\mathbf{X}\right]$ as

$$
E\left[\left\{\epsilon - \frac{E(\epsilon^3|\mathbf{X})}{E(t^2|\mathbf{X})}t\right\}^2|\mathbf{X}\right] = 1 + \frac{\{E(\epsilon^3|\mathbf{X})\}^2}{E(t^2|\mathbf{X})}.
$$

Thus, $E(\boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}}\boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}}^{\mathrm{T}})$ becomes

$$E\left(\left[1+\frac{\{E(\epsilon^3|\mathbf{X})\}^2}{E(t^2|\mathbf{X})}\right]e^{-2\sigma(\mathbf{X},\boldsymbol{\beta})}\mathbf{m}'_{\boldsymbol{\alpha}}(\mathbf{X},\boldsymbol{\alpha})\mathbf{m}'^{\mathrm{T}}_{\boldsymbol{\alpha}}(\mathbf{X},\boldsymbol{\alpha})\right).$$

2. $E(\boldsymbol{S}_{\text{eff},\boldsymbol{\alpha}}\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}^{\mathrm{T}})$ can be rewritten as

$$\begin{aligned}&E\left[E\left\{\epsilon t-\frac{E(\epsilon^3|\mathbf{X})}{E(t^2|\mathbf{X})}t^2|\mathbf{X}\right\}\frac{2}{E(t^2|\mathbf{X})}e^{-\sigma(\mathbf{X},\boldsymbol{\beta})}\mathbf{m}'_{\boldsymbol{\alpha}}(\mathbf{X},\boldsymbol{\alpha})\boldsymbol{\sigma}'^{\mathrm{T}}_{\boldsymbol{\beta}}(\mathbf{X},\boldsymbol{\beta})\right]\\&=E\left\{-\frac{2E(\epsilon^3|\mathbf{X})}{E(t^2|\mathbf{X})}e^{-\sigma(\mathbf{X},\boldsymbol{\beta})}\mathbf{m}'_{\boldsymbol{\alpha}}(\mathbf{X},\boldsymbol{\alpha})\boldsymbol{\sigma}'^{\mathrm{T}}_{\boldsymbol{\beta}}(\mathbf{X},\boldsymbol{\beta})\right\}.\end{aligned}$$

3. $E(\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}\boldsymbol{S}_{\text{eff},\boldsymbol{\beta}}^{\mathrm{T}})$ is equivalent to

$$E\left(\frac{4E(t^2|\mathbf{X})}{\{E(t^2|\mathbf{X})\}^2}\boldsymbol{\sigma}'_{\boldsymbol{\beta}}(\mathbf{X},\boldsymbol{\beta})\boldsymbol{\sigma}'^{\mathrm{T}}_{\boldsymbol{\beta}}(\mathbf{X},\boldsymbol{\beta})\right)=E\left\{\frac{4}{E(t^2|\mathbf{X})}\boldsymbol{\sigma}'_{\boldsymbol{\beta}}(\mathbf{X},\boldsymbol{\beta})\boldsymbol{\sigma}'^{\mathrm{T}}_{\boldsymbol{\beta}}(\mathbf{X},\boldsymbol{\beta})\right\}.$$

Theorem 1 is the immediate consequence of the above calculations. □

References

Bement, T. R., Williams, J. S. (1969). Variance of weighted regression estimators when sampling errors are independent and heteroscedastic. *Journal of the American Statistical Association*, *64*, 1369–1382.

Bickel, P. J., Klaassen, C. A., Ritov, Y., Wellner, J. A. (1998). *Efficient and adaptive estimation for semiparametric models*. New York: Springer.

Carroll, R. J. (1982). Adapting for heteroscedasticity in linear models. *The Annals of Statistics*, *10*, 1224–1233.

Carroll, R. J., Ruppert, D. (1982). Robust estimation in heteroscedastic linear models. *The Annals of Statistics*, *10*, 429–441.

Fang, Y., Lian, H., Liang, H., Ruppert, D. (2015). Variance function additive partial linear models. *Electronic Journal of Statistics*, *9*, 2793–2827.

Fuller, W. A., Rao, J. (1978). Estimation for a linear regression model with unknown diagonal covariance matrix. *The Annals of Statistics*, *6*, 1149–1158.

Hall, P., Carroll, R. (1989). Variance function estimation in regression: The effect of estimating the mean. *Journal of the Royal Statistical Society*: *Series B (Methodological)*, *51*, 3–14.

Jacquez, J. A., Mather, F. J., Crawford, C. R. (1968). Linear regression with non-constant, unknown error variances: Sampling experiments with least squares, weighted least squares and maximum likelihood estimators. *Biometrics*, *24*, 607–626.

Joanes, D., Gill, C. (1998). Comparing measures of sample skewness and kurtosis. *Journal of the Royal Statistical Society*: *Series D (The Statistician)*, *47*, 183–189.

Kim, M., Ma, Y. (2012). The efficiency of the second-order nonlinear least squares estimator and its extension. *Annals of the Institute of Statistical Mathematics*, *64*, 751–764.

Lian, H., Liang, H., Carroll, R. J. (2015). Variance function partially linear single-index models. *Journal of the Royal Statistical Society*: *Series B (Statistical Methodology)*, *77*, 171–194.

Ma, Y., Chiou, J.-M., Wang, N. (2006). Efficient semiparametric estimator for heteroscedastic partially linear models. *Biometrika*, *93*, 75–84.

Müller, H.-G., Zhao, P.-L. (1995). On a semiparametric variance function model and a test for heteroscedasticity. *The Annals of Statistics*, *23*, 946–967.

Nadarajah, S. (2005). A generalized normal distribution. *Journal of Applied Statistics*, *32*, 685–694.

Nerlove, M. (1963). *Returns to scale in electricity supply. En "Measurement in economics-studies in mathematical economics and econometrics in memory of Yehuda Grunfeld" edited by Carl F. Christ*. Palo Alto: Stanford University Press.
Tsiatis, A. (2006). *Semiparametric theory and missing data*. New York: Springer.