

Semiparametric Analysis of Heterogeneous Data Using Varying-Scale Generalized Linear Models

Minge XIE, Douglas G. SIMPSON, and Raymond J. CARROLL

This article describes a class of heteroscedastic generalized linear regression models in which a subset of the regression parameters are rescaled nonparametrically, and develops efficient semiparametric inferences for the parametric components of the models. Such models provide a means to adapt for heterogeneity in the data due to varying exposures, varying levels of aggregation, and so on. The class of models considered includes generalized partially linear models and nonparametrically scaled link function models as special cases. We present an algorithm to estimate the scale function nonparametrically, and obtain asymptotic distribution theory for regression parameter estimates. In particular, we establish that the asymptotic covariance of the semiparametric estimator for the parametric part of the model achieves the semiparametric lower bound. We also describe bootstrap-based goodness-of-scale test. We illustrate the methodology with simulations, published data, and data from collaborative research on ultrasound safety.

KEY WORDS: Generalized linear regression; Heteroscedasticity; Nonparametric regression; Partially linear model; Semiparametric efficiency; Varying-coefficient model.

1. INTRODUCTION

Varying-coefficient models have been widely studied with the aim of developing flexible nonparametric regression models for various contexts. Hastie and Tibshirani (1993) formulated the broad class of models that have the form

$$E(y_i | \mathbf{x}_i, \mathbf{z}_i) = \mu\{\mathbf{x}_i^T \boldsymbol{\beta}(\mathbf{z}_i)\}, \quad (1)$$

where μ is a given link function, $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^T$ represents the vector of primary covariates, and $\boldsymbol{\beta}(\mathbf{z}_i) = (\beta_1(\mathbf{z}_i), \dots, \beta_p(\mathbf{z}_i))^T$ represents an unknown vector function whose components depend on additional observed variables \mathbf{z}_i . Cai, Fan, and Li (2000) developed rigorous asymptotics using local polynomial regressions to estimate the functions $\beta_j(\mathbf{z}_i)$ (i.e., the varying coefficients), and developed a nonparametric likelihood ratio test of whether in fact the coefficients are varying.

Semiparametric specializations of the model in (1) also have been studied. Hunsberger (1994), Severini and Staniswalis (1994), and Carroll, Fan, Gijbels, and Wand (1997) considered generalized partially linear models. Zhang, Lee, and Song (2002) and Ahmad, Leeahanon, and Li (2005) considered partially linear varying-coefficient models with the identity link and additive errors,

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta}(\mathbf{z}_i) + \mathbf{v}_i^T \boldsymbol{\delta} + \epsilon_i, \quad (2)$$

where $E(\epsilon_i | \mathbf{x}_i, \mathbf{v}_i, \mathbf{z}_i) = 0$ and $\text{var}(\epsilon_i | \mathbf{x}_i, \mathbf{v}_i, \mathbf{z}_i) = \sigma^2(\mathbf{z}_i)$. These authors presented semiparametric estimators that combine nonparametric estimation of the smooth functions $\beta_j(\cdot)$ with root- n -consistent estimation of the parametric components, $\boldsymbol{\delta}$. Ahmad et al. (2005) established the semiparametric efficiency of their estimator of $\boldsymbol{\delta}$ in (2). Lam and Fan (2008) established root- n asymptotic results for semiparametric quasi-likelihood in partially linear generalized varying-coefficient models. Their

results allowed the dimension of the parametric component to increase with n at a certain rate.

In this article we consider a different partitioning of the model elements into parametric and nonparametric components, in which the variation in coefficients is captured by a common multiplicative scaling function w . We call the resulting class of models *varying-scale generalized linear models*. For $i = 1, \dots, n$, consider responses y_i , covariates \mathbf{x}_i and \mathbf{v}_i , and auxiliary variables \mathbf{z}_i . The responses y_i are assumed to follow a structural model of the form

$$E(y_i | \mathbf{x}_i, \mathbf{v}_i, \mathbf{z}_i) = \mu\{w(\mathbf{z}_i) \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{v}_i^T \boldsymbol{\delta}\}, \quad (3)$$

where $\mu(\cdot)$ is a known link function, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ is a vector of p regression parameters subject to scale heterogeneity, $\boldsymbol{\delta} = (\delta_1, \dots, \delta_q)^T$ is a vector of q additional regression parameters, and $w(\cdot)$ is an unknown scaling function assumed only to be smooth in a sense defined later. We consider the exponential family class of error models, so the models are a type of varying-scale generalized linear model. Under regularity conditions and an identifiability constraint, we establish the root- n consistency and asymptotic normality of the semiparametric estimator of $(\boldsymbol{\beta}^T, \boldsymbol{\delta}^T)^T$, and we show that these estimators achieve a semiparametric lower bound for asymptotic variance.

We need to impose a constraint on either $w(\cdot)$ or $\boldsymbol{\beta}$ to ensure identifiability. For example, suppose that we let \mathbf{z}_* denote a reference value for the auxiliary vector \mathbf{z} such as the mean value in the sample, and impose the constraint that $w(\mathbf{z}_*) = 1$. Then $\boldsymbol{\beta}$ is the gradient of $\mu^{-1}\{E(y|\mathbf{x}, \mathbf{z}_*)\}$ as a function of \mathbf{x} at the reference value \mathbf{z}_* . Equivalently, suppose that the first component of $\boldsymbol{\beta}$ is assumed to be nonzero. Then we may set this component equal to 1 in the model, absorbing its sign and magnitude into $w(\cdot)$, which is then unconstrained. We use the latter constraint in the theoretical development. In general, we are able to make parametric inferences about $\boldsymbol{\delta}$ and any function of $\boldsymbol{\beta}$ that depends only on the ratios of its components. For any such function, we obtain root- n -consistent semiparametric-efficient estimates. The approach considered here a combination of local profile likelihood estimation and backfitting to estimate the model. The approach and the large-sample theory

Minge Xie is Associate Professor and Director of Office of Statistical Consulting, Department of Statistics, Rutgers University, Piscataway, NJ 08854 (E-mail: mxie@stat.rutgers.edu). Douglas G. Simpson is Professor and Chair, Department of Statistics, University of Illinois, Champaign, IL 61820 (E-mail: dgs@uiuc.edu). Raymond J. Carroll is Distinguished Professor, Department of Statistics, Texas A&M University, College Station, TX 77843 (E-mail: carroll@stat.tamu.edu). This research was supported in part by National Science Foundation grant SES-0241859 (Xie), National Institute of Biomedical Imaging and Bioengineering grant EB02641 (Simpson), National Cancer Institute grants CA57030 and CA104620 (Carroll), and the Texas A&M Center for Environmental and Rural health through a grant from the National Institute of Environmental Health Sciences (P30-ES09106) (Carroll). The authors thank William O'Brien, Jr. for permission to use the ultrasound data.

adapts and extends results of Severini and Wong (1992), Carroll et al. (1997), and Cai et al. (2000).

Several special cases of the model in (3) are worth noting. If \mathbf{x}_i is a scalar constant, then the model reduces to the generalized partially linear model of Hunsberger (1994) and others. If $\mu(\cdot)$ is the identity link function and x_i is a nonconstant scalar, then the model in (3) coincides with the model in (2). In such cases where β is a scalar, we do not obtain parametric inferences for this parameter, because it is equivalent to a scaling of the nonparametric function $w(\cdot)$. If $\delta = 0$ in (3) and β is a vector, then we obtain a class of *scaled link function models*, fully parametric versions of which were derived for heterogeneous binary and ordinal response data by McCullagh and Nelder (1989, p. 154) and Xie, Simpson, and Carroll (1997) using latent variable constructions. This class of models includes as a special case the generalized linear models with unknown link functions considered by Weisberg and Welsh (1994) and Chiou and Müller (1998) that allow the fixed, unknown link function to be estimated nonparametrically. The resulting model is homoscedastic in that the link function is assumed constant across all observations. In the more general varying-scale model of (3), the effective link function $\mu_i(\cdot)$ varies among different individuals depending on the covariate \mathbf{z}_i . More generally, (3) is a class of varying-coefficient models in which a subset of the covariates have effects that are adjusted in parallel through a nonparametric rescaling function.

Further generalizations are possible in which multiple subsets of the covariates are adjusted in parallel; however, the model in (3) is sufficiently general to develop the fundamental idea of semiparametric-efficient inferences about β and δ after adjusting for the heterogeneity represented by $w(z_i)$.

To illustrate the effect of varying scale, Figure 1(a) presents binary data on the occurrence of ultrasound-induced lung hemorrhage in pigs as a function of age (in weeks) and acoustic pressure in mega-pascals (MPa). The symbols indicate presence or absence of a lesion after exposure. The data come from an experiment described by O'Brien et al. (2003) and are modeled by a semiparametric varying-scale logit model in Section 6 for age-dependent risk of lesion occurrence. Letting ED100p denote the acoustic pressure corresponding to a 100p% risk of lesions, the solid line represents the age-dependent ED50 curve for risk of lesions, and the dashed line represents the ED05 curve. These are contours of the fitted probability surface corresponding to 5% and 50% probability levels. The nonparallelism of the curves is a reflection of the varying scale as a function of age. Figure 1(b) demonstrates the age-dependence of the lesion odds ratio associated with a 1-MPa increase in acoustic pressure. Further details of the analysis are given in Section 6.

The rest of the article is organized as follows. Section 2 presents the estimation framework and a profile local-likelihood algorithm for performing the estimation. Section 3 provides the

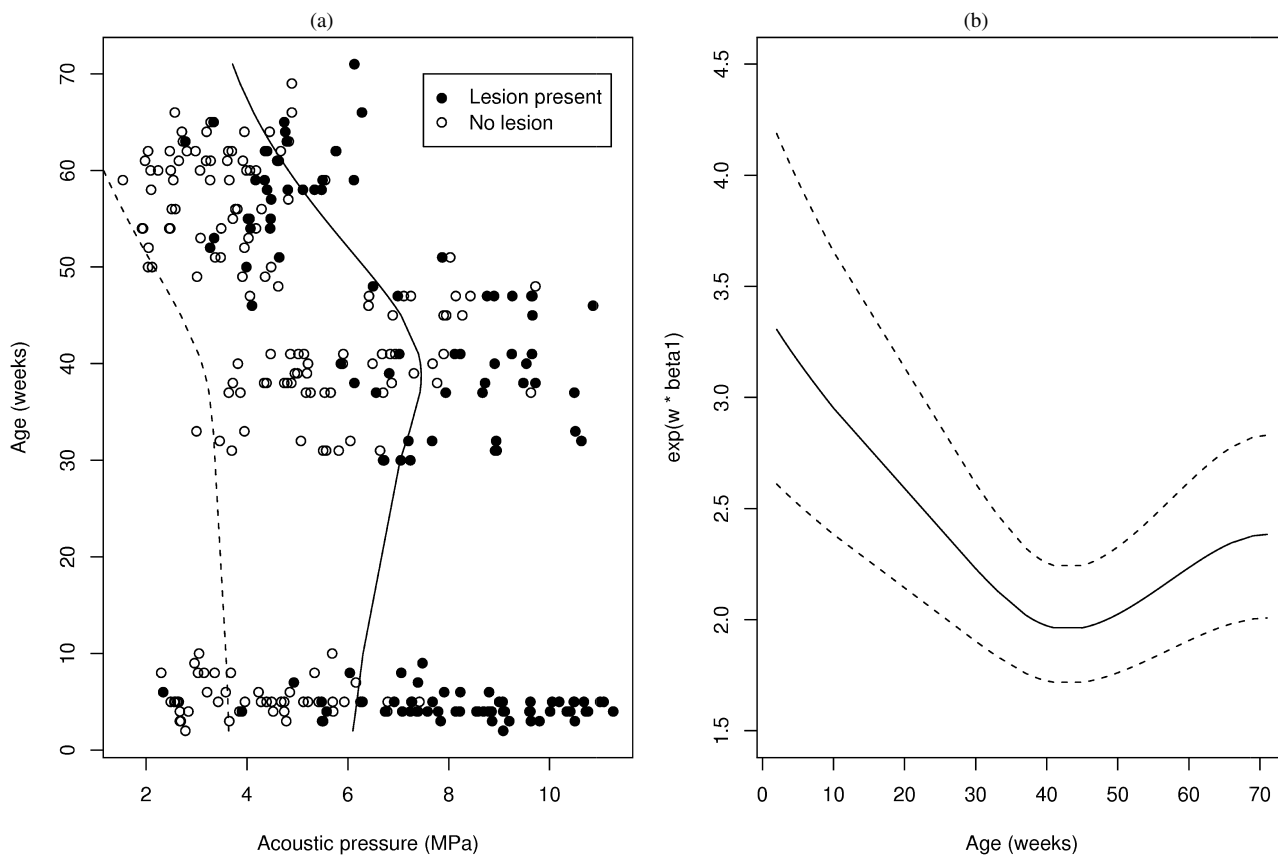


Figure 1. (a) A plot of observed incidence of lesions for pigs exposed to ultrasound. Age is age at time of experimental exposure; acoustic pressure is computed in MPa for the pleural surface of the lung. The solid curve is the age-dependent ED50 (acoustic pressure exposure corresponding to 50% risk), implied by a fitted varying-scale logit model. The dashed curve is the estimated ED05, which corresponds to 5% risk. (b) A plot of the increase odds of risk of lesions, per MPa increase in acoustic pressure, against the age (solid line), together with the corresponding point wise 95% confidence intervals (dotted lines).

main results on the large-sample theory. Section 4 describes a bootstrap-based goodness-of-fit-type of test for varying scales. Section 5 presents simulation studies to illustrate the empirical performance of the proposed estimation and testing methodologies. Section 6 applies the varying-scale modeling approach to data from the ultrasound risk assessment and also to Efron's toxoplasmosis data (Efron 1986). Section 7 discusses further issues. Proofs are given in the Appendix and online supplemental materials.

2. ESTIMATION METHOD

Let $(y_i, \mathbf{x}_i, \mathbf{v}_i, z_i)$, for $i = 1, 2, \dots, n$, be independently and identically distributed copies of random variables $(y, \mathbf{x}, \mathbf{v}, z)$. For simplicity, henceforth we consider the case in which $\mathbf{z} = z$ is one-dimensional. Extension to multivariate \mathbf{z} involves no fundamentally new ideas; see Section 7 for further comments. We develop estimates and asymptotic distribution theory assuming that the conditional density of y_i has the form

$$f(y_i|\psi_i) = \exp[\{y_i\psi_i - a(\psi_i)\}/\phi + b(y_i, \phi)] \quad (4)$$

with respect to a fixed measure, where ψ_i is a twice-differentiable monotone function of

$$\mu_i = \mu\{w(z_i)\mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{v}_i^T\boldsymbol{\delta}\} = \mu(w_i\eta_i + \tilde{\eta}_i), \quad (5)$$

$a(\cdot)$ and $b(\cdot, \cdot)$ are fully specified, ϕ is a possible fixed dispersion parameter, $\eta_i = \mathbf{x}_i^T\boldsymbol{\beta}$, $\tilde{\eta}_i = \mathbf{v}_i^T\boldsymbol{\delta}$, and $w_i = w(z_i)$, where w is an unknown smooth function subject to conditions given in Section 3. Let $\ell(\mu_i, y_i)$ denote the log-likelihood function of the i th observation in (4), and let $\ell(\boldsymbol{\mu}, \mathbf{y}) = \sum_{i=1}^n \ell(\mu_i, y_i)$, the log-likelihood of all the observations. The conditional density in (4) is standard for generalized linear models (McCullagh and Nelder 1989), but the regression model in (5) is more general than previously considered.

We present an alternating local profile-likelihood type of algorithm for carrying out the estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ in (4) and (5), making only smoothness assumptions about the unknown function $w(\cdot)$. The algorithm iteratively cycles between fitting parametric components $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\delta}^T)^T$ and fitting nonparametric scales w_i while holding the other fixed.

Let $\boldsymbol{\beta}_*$, $\boldsymbol{\delta}_*$, and w_* , denote the true parameter values and scale function. The model is identifiable only if $\boldsymbol{\beta}_* \neq \mathbf{0}$, so we assume that the first component of $\boldsymbol{\beta}_*$ is nonzero. Then we may reparameterize, setting the first component equal to 1 and leaving w unconstrained. Thus we assume that $\boldsymbol{\beta}_*^T = (1, \boldsymbol{\beta}^T)$, where $\boldsymbol{\beta}$ is a $(p - 1)$ -dimensional vector. Corresponding to the form of $\boldsymbol{\beta}_*$, write $\mathbf{x}_{i*}^T = (x_{i1}, \mathbf{x}_i^T)$, where in the sequel, \mathbf{x}_i denotes the last $p - 1$ components of the primary covariate vector.

Denote by $w^{(0)}(\cdot)$ the true scale function assuming that $\boldsymbol{\beta}$ has first component set equal to 1. To estimate the function nonparametrically for a fixed $\boldsymbol{\theta}$, we proceed by analogy with the approaches of Carroll et al. (1997), Carroll, Ruppert, and Welsh (1998), and Cai et al. (2000). Given a point z_0 , approximate $w^{(0)}(t)$ in the neighborhood of z_0 by a linear function, $w^{(0)}(t) \approx \lambda_0 + \lambda_1(t - z_0)$. Assume that $w^{(0)}(t)$ is second-order differentiable. The vector $\boldsymbol{\lambda} = (\lambda_0, \lambda_1)^T$ depends on z_0 , the form of the function $w^{(0)}(\cdot)$, and the parameter vector

$\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\delta}^T)^T$. Given prior values for $\boldsymbol{\theta}$ and $w^{(0)}$ write the local likelihood function at z_0 as

$$\begin{aligned} \ell_{LO}(\boldsymbol{\lambda}) &= \ell_{LO}(\boldsymbol{\lambda}|\boldsymbol{\eta}, \tilde{\boldsymbol{\eta}}) \\ &= \frac{1}{n} \sum_{i=1}^n \ell(\mu[\{\lambda_0 + \lambda_1(z_i - z_0)\}\eta_i + \tilde{\eta}_i], y_i) \\ &\quad \times K_b(z_i - z_0), \end{aligned} \quad (6)$$

where $\boldsymbol{\eta} = (\eta_1, \dots, \eta_n)^T$, $\tilde{\boldsymbol{\eta}} = (\tilde{\eta}_1, \dots, \tilde{\eta}_n)^T$, $K_b(\cdot) = K(\cdot/b)/b$, $K(\cdot)$ is a symmetric kernel function, and $b = b_n > 0$ is a bandwidth. Denote by $\hat{\boldsymbol{\lambda}}_\theta = \hat{\boldsymbol{\lambda}}_\theta(z_0) = \{\hat{\lambda}_{0,\theta}(z_0), \hat{\lambda}_{1,\theta}(z_0)\}^T$ the set of values that maximize the local likelihood function $\ell_{LO}(\boldsymbol{\lambda})$ for each given $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\delta}^T)^T$. This $\hat{\boldsymbol{\lambda}}_\theta(z_0)$ is the local maximum likelihood estimate of parameters $\boldsymbol{\lambda}$ at z_0 . For fixed $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\delta}^T)^T$, the nonparametric scale function $w^{(0)}(\cdot)$ at z_0 is estimated by $\hat{w}_\theta(z_0) = \hat{\lambda}_{0,\theta}$. Further background on local likelihood estimation was given by Fan and Gijbels (1996).

Next, consider estimation of the regression parameters $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$. Suppose that for a given $\boldsymbol{\theta}$, $\hat{w}_\theta(z_i)$ is obtained from maximizing the local likelihood function (6) at the value z_i , and $\hat{w}_\theta^{(1)}(z_i)$ is its first-order derivative with respect to $\boldsymbol{\theta}$. We consider estimates of $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\delta}^T)^T$ obtained by solving the following equations, which are constructed from an efficient score function as described by, for example, Severini and Wong (1992) and Bickel, Klaassen, Ritov, and Wellner (1993),

$$\begin{aligned} \sum_{i=1}^n [y_i - \mu\{\hat{w}_\theta(z_i)\eta_i + \tilde{\eta}_i\}] \tau\{\hat{w}_\theta(z_i)\eta_i + \tilde{\eta}_i\} \\ \times \left\{ \begin{pmatrix} \hat{w}_\theta(z_i)\mathbf{x}_i \\ \mathbf{v}_i \end{pmatrix} + \eta_i \hat{w}_\theta^{(1)}(z_i) \right\} = 0. \end{aligned} \quad (7)$$

Here $\tau(s) = \mu'(s)/a''\{u(s)\}$, $u(s) = \{(a')^{-1} \circ \mu\}(s)$, $a(\cdot)$ is as defined in (4), and $\mu(\cdot)$ is the link function. In a canonical link model, $\tau(s) \equiv 1$. See the Appendix for a construction of the efficient score function \mathcal{S}^* and the derivation of (7). Note that for any fixed $\boldsymbol{\theta}$ in the neighborhood of the true $\boldsymbol{\theta}_0$, $\hat{w}_\theta(z_0)$ is a consistent estimator of $w^{(0)}(z_0)$, and its derivative with respect to $\boldsymbol{\theta}$, $\hat{w}_\theta^{(1)}(z_0)$, is a consistent estimator of a term related to the projection obtaining the efficient score function; see Section 3 and the online supplemental materials for further details.

The estimating equations (7) can be solved by an iteratively reweighted least squares algorithm. At each step, update the estimates of $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\delta}^T)^T$ by

$$\begin{aligned} \hat{\boldsymbol{\theta}}^{\text{new}} &= \hat{\boldsymbol{\theta}}^{\text{old}} + \left(\mathbf{A}_n(\boldsymbol{\theta}) \frac{1}{n} \sum_{i=1}^n [y_i - \mu\{\hat{w}_\theta(z_i)\eta_i + \tilde{\eta}_i\}] \right. \\ &\quad \times \left\{ \begin{pmatrix} \hat{w}_\theta(z_i)\mathbf{x}_i \\ \mathbf{v}_i \end{pmatrix} + \eta_i \hat{w}_\theta^{(1)}(z_i) \right\} \\ &\quad \left. \times \tau\{\hat{w}_\theta(z_i)\eta_i + \tilde{\eta}_i\} \right) \Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{\text{old}}}, \end{aligned} \quad (8)$$

where

$$\begin{aligned} \mathbf{A}_n^{-1}(\boldsymbol{\theta}) &= \frac{1}{n} \sum_{i=1}^n \left\{ \begin{pmatrix} \hat{w}_\theta(z_i)\mathbf{x}_i \\ \mathbf{v}_i \end{pmatrix} + \eta_i \hat{w}_\theta^{(1)}(z_i) \right\} \\ &\quad \times \left\{ \begin{pmatrix} \hat{w}_\theta(z_i)\mathbf{x}_i \\ \mathbf{v}_i \end{pmatrix} + \eta_i \hat{w}_\theta^{(1)}(z_i) \right\}^T \\ &\quad \times \tau_1\{\hat{w}_\theta(z_i)\eta_i + \tilde{\eta}_i\}. \end{aligned} \quad (9)$$

Note that $\mathbf{A}_n^{-1}(\boldsymbol{\theta})|_{\theta=\hat{\theta}}$ is an estimate of the covariance matrix of the efficient score function \mathcal{S}^* (see Sec. 3).

The proposed estimation method can be implemented by the following generic algorithm, iterating between two modules:

- Estimating regression parameters. Fix the current set of scale variable weights and their first derivatives (with respect to $\boldsymbol{\theta}$), say $\widehat{w}_i^{\text{cur}}$ and $\{\widehat{w}_i^{(1)}\}^{\text{cur}}$, and use (8) to update the estimates of the regression parameters $\widehat{\boldsymbol{\beta}}^{\text{new}}$ and $\widehat{\boldsymbol{\delta}}^{\text{new}}$.
- Estimating scale function. Fix the current estimates of $\boldsymbol{\theta}$, say $\widehat{\boldsymbol{\theta}}^{\text{curr}}$. Maximize (6) to update the estimate of the scale variable function $w(z_i)$ by $\widehat{w}_\theta^{\text{new}}(z_i)|_{\theta=\widehat{\boldsymbol{\theta}}^{\text{curr}}}$ and its derivative with respect to $\boldsymbol{\theta}$ by $\widehat{w}_\theta^{(1)}(z_i)|_{\theta=\widehat{\boldsymbol{\theta}}^{\text{curr}}}$.

In the first module, (8) is a variant of the iteratively reweighted least squares algorithm. In the second module, (6) is essentially fitting a (univariate) varying-coefficient model (Hastie and Tibshirani 1993). Existing algorithms for fitting varying-coefficient models can be used. In the examples in Section 6, we use a modified version of the algorithm of Cai et al. (2000) to update the w estimates.

The semiparametric estimates $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\delta}}$ obtained from this algorithm are root- n consistent and asymptotically efficient under the exponential family modeling assumptions, whereas the estimate $\widehat{w}(z_i)$ has the standard nonparametric asymptotic rate of convergence. The asymptotic results are given in the next section. This asymptotic distribution theory also provides the basis for large-sample semiparametric inferences.

3. SEMIPARAMETRIC EFFICIENCY

In this section we develop large-sample theory under the framework of the preceding section. These asymptotic results are developed under the regularity conditions given herein. These may not be the weakest possible conditions, but they provide a mild set of sufficient conditions often satisfied in practice. Rigorous treatments of semiparametric efficiency for the profile likelihood method in various settings have been given by Severini and Wong (1992), Lin and Carroll (2006), Claeskens and Carroll (2007), and Lam and Fan (2008). We use a similar approach in the present setting.

Throughout we use a symmetric kernel function and assume that independent observations follow the model defined by (4) and (5). We denote the true values of the parameter vectors by $\boldsymbol{\beta}_0$ and $\boldsymbol{\delta}_0$. In addition, we assume the following:

1. For the symmetric kernel function $K(t)$, the terms $v_2 = \int t^2 K(t) dt$, $v_0 = \int \{K(t)\}^2 dt$, and $v_2 = \int t^2 \{K(t)\}^2 dt$ are all finite.
2. The functions $\mu(\cdot)$ and $a(\cdot)$, as well as the true scale function $w^{(0)}(\cdot)$, have continuous third-order derivatives.
3. Let \mathcal{Y} , \mathcal{X} , \mathcal{V} , and \mathcal{Z} be admissible sets of response variable y and covariate variables \mathbf{x} , \mathbf{v} , and z . Assume that y has finite fourth moment. In addition, $\inf\{f_z(t)\} > 0$, where the infimum is over $t \in \mathcal{Z}$ and $f_z(t)$ is the marginal density for z .

To state our main results, we introduce the following notation. Write $\gamma(t) = E[(\eta^{(0)})^2 \tau_1\{w^{(0)}(z)\eta^{(0)} + \widetilde{\eta}^{(0)}\}|z=t]$, $\boldsymbol{\gamma}_1(t) = E[(\tau_1\{w^{(0)}(z)\eta^{(0)} + \widetilde{\eta}^{(0)}\})\mathbf{x}|z=t]$, and $\widetilde{\boldsymbol{\gamma}}_1(t) = E[(\tau_1\{w^{(0)}(z)\eta^{(0)} + \widetilde{\eta}^{(0)}\})\mathbf{v}|z=t]$, where $\eta^{(0)} = \mathbf{x}^T \boldsymbol{\beta}_0$,

$\widetilde{\eta}^{(0)} = \mathbf{v}^T \boldsymbol{\delta}_0$, and $\tau_1(s) = \mu'(s)\tau(s)$. Write $\mathbf{J} = \text{diag}\{1, b\}$, $\mathbf{H} = f_z(z_0)\gamma(z_0)\text{diag}\{1, v_2\}$, and $\mathbf{A} = f_z(z_0)\gamma(z_0)\text{diag}\{v_0, v_2\}$, where $\text{diag}\{s_1, \dots, s_k\}$ represents a $k \times k$ diagonal matrix of elements s_1, \dots, s_k . In addition, define $B_n(r) = \{\boldsymbol{\beta} | \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| < r/\sqrt{n}\}$ and $\widetilde{B}_n(\widetilde{r}) = \{\boldsymbol{\delta} | \|\boldsymbol{\delta} - \boldsymbol{\delta}_0\| < \widetilde{r}/\sqrt{n}\}$ for any fixed constants $r > 0$ and $\widetilde{r} > 0$.

For any given $\boldsymbol{\beta} \in B_n(r)$ and $\boldsymbol{\delta} \in \widetilde{B}_n(\widetilde{r})$, the following theorem provides a \sqrt{nb} -convergence result for the local maximum likelihood estimator (MLE) $\widehat{\boldsymbol{\lambda}}_\theta = \widehat{\boldsymbol{\lambda}}_\theta(z_0)$, which maximizes local likelihood function (6).

Theorem 1. Suppose that the bandwidth $b = O(n^{-\xi})$, $1/6 < \xi < 1/4$, and \mathbf{H} is nonsingular. Under conditions 1–3, and for any given $\boldsymbol{\beta} \in B_n(r)$ and $\boldsymbol{\delta} \in \widetilde{B}_n(\widetilde{r})$, we have

$$\begin{aligned} & \sqrt{nb} \left\{ \mathbf{J}(\widehat{\boldsymbol{\lambda}}_\theta - \boldsymbol{\lambda}) - \frac{b^2}{2} w^{(2)}(z_0) \begin{pmatrix} v_2 \\ 0 \end{pmatrix} \right. \\ & \quad - w^{(0)}(z_0) \{\gamma(z_0)\}^{-1} \begin{pmatrix} \{\boldsymbol{\gamma}_1(z_0)\}^T (\boldsymbol{\beta} - \boldsymbol{\beta}_0) \\ 0 \end{pmatrix} \\ & \quad \left. - \{\gamma(z_0)\}^{-1} \begin{pmatrix} \{\widetilde{\boldsymbol{\gamma}}_1(z_0)\}^T (\boldsymbol{\delta} - \boldsymbol{\delta}_0) \\ 0 \end{pmatrix} + o_p\left(\frac{1}{\sqrt{n}}\right) \right\} \\ & \rightarrow \text{Normal}(\mathbf{0}, \mathbf{H}^{-1} \mathbf{A} \mathbf{H}^{-1}). \end{aligned}$$

The proof of the theorem, together with the asymptotic expansions of $\widehat{\boldsymbol{\lambda}}_\beta$ and $\widehat{\boldsymbol{\lambda}}_\beta^{(1)}$, up to the order of $o_p(n^{-1/2})$, are provided in the online supplemental material. The theorem implies that for any fixed $\boldsymbol{\theta}$ in the neighborhood of the true $\boldsymbol{\theta}_0$, $\widehat{w}_\theta(z_0)$ is a consistent estimator of $w^{(0)}(z_0)$.

Write $\mu_{w\mathbf{x}\mathbf{v}} = \eta^{(0)}(\{\mathbf{m}(z)\}^T, \{\widetilde{\mathbf{m}}(z)\}^T)^T$, where $\mathbf{m}(z) = w^{(0)}(z)\{\gamma(z)\}^{-1}\boldsymbol{\gamma}_1(z)$ and $\widetilde{\mathbf{m}}(z) = \{\gamma(z)\}^{-1}\widetilde{\boldsymbol{\gamma}}_1(z)$, and

$$\begin{aligned} \mathbf{A}^{-1} = E \left[\left\{ \begin{pmatrix} w^{(0)}(z)\mathbf{x} \\ \mathbf{v} \end{pmatrix} - \mu_{w\mathbf{x}\mathbf{v}} \right\} \left\{ \begin{pmatrix} w^{(0)}(z)\mathbf{x} \\ \mathbf{v} \end{pmatrix} - \mu_{w\mathbf{x}\mathbf{v}} \right\}^T \right. \\ \left. \times \tau_1\{w^{(0)}(z)\eta^{(0)} + \widetilde{\eta}^{(0)}\} \right]. \end{aligned}$$

The next two theorems state that the estimators from the estimating equations (7) for the regression parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\delta}^T)^T$ are root- n consistent, asymptotically normally distributed, and asymptotically efficient.

Theorem 2. Suppose that \mathbf{A}^{-1} is positive definite. Let $\widehat{w}_\theta(\cdot)$ and $\widehat{w}_\theta^{(1)}(\cdot)$ be the local MLE and its first derivative with respect to $\boldsymbol{\theta}$. Then, under conditions 1–3, a solution $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}^{\text{new}}$ to estimation equations (7) exists in probability and satisfies $\|\widehat{\boldsymbol{\theta}}^{\text{new}} - \boldsymbol{\theta}_0\| = O_p(n^{-1/2})$. In addition, as $n \rightarrow \infty$,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}^{\text{new}} - \boldsymbol{\theta}_0) \xrightarrow{d} \text{Normal}(0, \mathbf{A}).$$

We also can estimate the asymptotic covariance matrix \mathbf{A} consistently by $\mathbf{A}_n(\boldsymbol{\theta})|_{\theta=\hat{\theta}}$, where

$$\begin{aligned} \mathbf{A}_n^{-1}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n \left\{ \begin{pmatrix} \widehat{w}_\theta(z_i)\mathbf{x}_i \\ \mathbf{v} \end{pmatrix} + \eta_i \widehat{w}_\theta^{(1)}(z_i) \right\} \\ \times \left\{ \begin{pmatrix} \widehat{w}_\theta(z_i)\mathbf{x}_i \\ \mathbf{v} \end{pmatrix} + \eta_i \widehat{w}_\theta^{(1)}(z_i) \right\}^T \\ \times \tau_1\{\widehat{w}_\theta(z_i)\eta_i + \widetilde{\eta}_i\}. \end{aligned}$$

Theorem 3. Under the regularity conditions specified in Theorem 2, the matrix \mathbf{A}^{-1} is the information lower bound for the parametric part. Thus the estimator obtained from (7) is asymptotically efficient.

Proofs of Theorems 2 and 3 are given in the Appendix and the online supplemental material.

4. BOOTSTRAP TEST OF VARYING SCALE

Within the varying-scale modeling framework, it is useful to be able to test parametrically scaled models versus a nonparametric scale as a means of testing goodness of fit. We consider three levels of complexity in the varying-scale model:

- (M1) $w_i \equiv 1$
- (M2) $w_i = \sum_{j=0}^p \psi_j t_j(z_i)$, for known functions $t_j(\cdot)$ and unknown parameters ψ_j
- (M3) $w_i = w(z_i)$, for unknown smooth function w .

The scale weight model (M1) corresponds to an unscaled model. Model (M2) covers many parametric-scale weight models, including polynomial models of order p with $w_i = \psi_0 + \psi_1 z_i + \dots + \psi_p z_i^p$. Model (M3) is a nonparametric scale weight model. We assume that all functions $t_j(\cdot)$ and $w(\cdot)$ have second-order derivatives. Model (M1) is nested within model (M2), and model (M2) is typically nested within model (M3). It may be of interest to test a hypothesis of a constant scale model (M1) versus a varying-scale model, with scale weight form being either parametric (M2) or nonparametric (M3). It also may be of interest to test a hypothesis of a parametric scale weight model (M2) against the nonparametric scale weight model (M3).

In addition to the foregoing hypothesis-testing problems, we also may be interested in testing any of these varying-scale models against a partially linear varying-coefficient model,

$$(M4) \ E(y_i | \mathbf{x}_i, \mathbf{v}_i, \mathbf{z}_i) = \mu \{ \mathbf{x}_i^T \boldsymbol{\beta}(\mathbf{z}_i) + \mathbf{v}_i^T \boldsymbol{\delta} \}.$$

In particular, we might want to test the varying-scale model (3) with a scale weight form in either (M3) or (M2) against the partially linear varying-coefficient model (M4), which has a separate nonparametric varying coefficient. Note that models (M2) and (M3) are nested within the larger partially linear varying-coefficient model (M4).

A heuristic testing procedure is based on the likelihood ratio statistic $T_n = 2\{\ell(\widehat{\boldsymbol{\mu}}_{H_1}, \mathbf{y}) - \ell(\widehat{\boldsymbol{\mu}}_{H_0}, \mathbf{y})\}$, where ℓ_{H_0} and ℓ_{H_1} are the regular likelihood functions and $\widehat{\boldsymbol{\mu}}_{H_0}$ and $\widehat{\boldsymbol{\mu}}_{H_1}$ are the maximum likelihood estimates (or the local maximum likelihood estimates in nonparametric case) of $\boldsymbol{\mu}$ under the corresponding null and alternative hypotheses, H_0 and H_1 . In the tests with parametric models in both H_0 and H_1 [e.g., non-scale model (M1) versus parametric scale model (M2)], the test based on T_n is just the regular likelihood ratio test and is straightforward.

In the tests that involve the nonparametric forms [e.g., model (M3) or model (M4) or both], the standard chi-squared approximation fails (because the effective number of parameters tends to infinity). In this case T_n becomes the so-called “generalized likelihood ratio” test statistic and the so-called “Wilks phenomenon.” (See Fan, Zhang, and Zhang 2001 for a theoretical treatment of such problems in varying-coefficient models.) Eubank and Hart (1992) and Aerts, Claeskens, and Hart (1999) studied similar test problems in linear regression models.

Here we consider a bootstrap approach facilitate model testing. This approach is similar to the bootstrap methods described by Cai et al. (2000) and others in the varying-coefficient models literature. Suppose that \widehat{w}_i is an estimate of the scale weight function at z_i under the varying-scale model under H_0 . We bootstrap n sets of $(\mathbf{x}_i^*, \mathbf{v}_i^*, z_i^*, w_i^*)$ from the n sets of $(\mathbf{x}_i, \mathbf{v}_i, z_i, \widehat{w}_i)$; then simulate y_i^* from the model under the null hypothesis H_0 , using the estimates under H_0 . When the varying-scale model under H_0 is a fully parametric model, this simulation is exactly the same as the parametric bootstrap method. We fit the simulated data to both models under H_0 and H_1 , and compute the test statistic T_n^* . Repeating the bootstrap numerous times (say, N), the N values of T_n^* can be used to compute the distribution of the test statistic under H_0 . The p value is the percentile of this simulated null distribution, where T_n is the cutoff value (i.e., $p^* = \frac{1}{N} \sum_{k=1}^N \mathbf{1}(T_n^* > T_n)$).

The empirical evidence in our simulation studies in Section 5 suggests very reasonable performance for this bootstrap procedure. It also shows evidence of the Wilks phenomenon (chi-squared approximation) on the bootstrapped likelihood ratio test statistic T_n^* .

In hypothesis-testing problems of regression parameters $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$, we can use either the likelihood ratio or Wald-type or score tests when the $w(\cdot)$'s are modeled parametrically. When the $w(\cdot)$'s are modeled nonparametrically, we suggest using Wald-type or score tests, which are supported by the results given in Theorem 2 of Section 3.

5. SIMULATION STUDIES

In this section we use simulation studies to illustrate the empirical performance of the proposed estimation and testing methodologies. Consider the logistic varying-scale model

$$\mu_i = H\{\delta_0 + \delta_1 v_i + w(z_i)(\beta_0 + x_i \beta_1)\}$$

$$\text{with } w(z_i) = 1 + 2 \sin(2\pi z_i) \text{ for } i = 1, \dots, n, \quad (9)$$

where $H(u) = \exp(u) / \{1 + \exp(u)\}$. Let the auxiliary variable z_i follow a uniform distribution on $[0, 1]$, and let the covariate variables x_i and v_i be normally distributed. In particular, in our simulations we simulate (U_1, U_2, U_3) from a three-dimensional multivariate normal distribution with mean $(0, .5, 1)$, variance $(1, 1, 2)$, and correlations $(1/2, 1/\sqrt{2}, 1/3)$ between the first and second elements, the first and third elements, and the second and third elements. Then set $z = \Phi(U_1)$, $x = U_2$, and $v = U_3$. This setting is similar to that in the simulations of Zhang et al. (2002), among others, in which trigonometric functions are used for the unknown function in nonparametric regressions, the auxiliary variable is uniformly distributed, and covariates are normally distributed. Assume the true regression coefficients $\delta_0 = -3.50$, $\delta_1 = 2.00$, $\beta_0 = 1.25$, and $\beta_1 = 1.00$. We repeatedly simulate data sets of size $n = 250$ and data sets of size $n = 400$ from this model. We use these simulated data and additional simulations to illustrate the performance of various parametric and nonparametric varying-scale models. The simulated data sets contain four columns: the responses y and the covariates x , v , and z .

In fitting a nonparametric varying-scale model, we need to select bandwidth as in any nonparametric model fitting. Fitting 200 data sets of sizes $n = 250$ and 400 to model (9) with

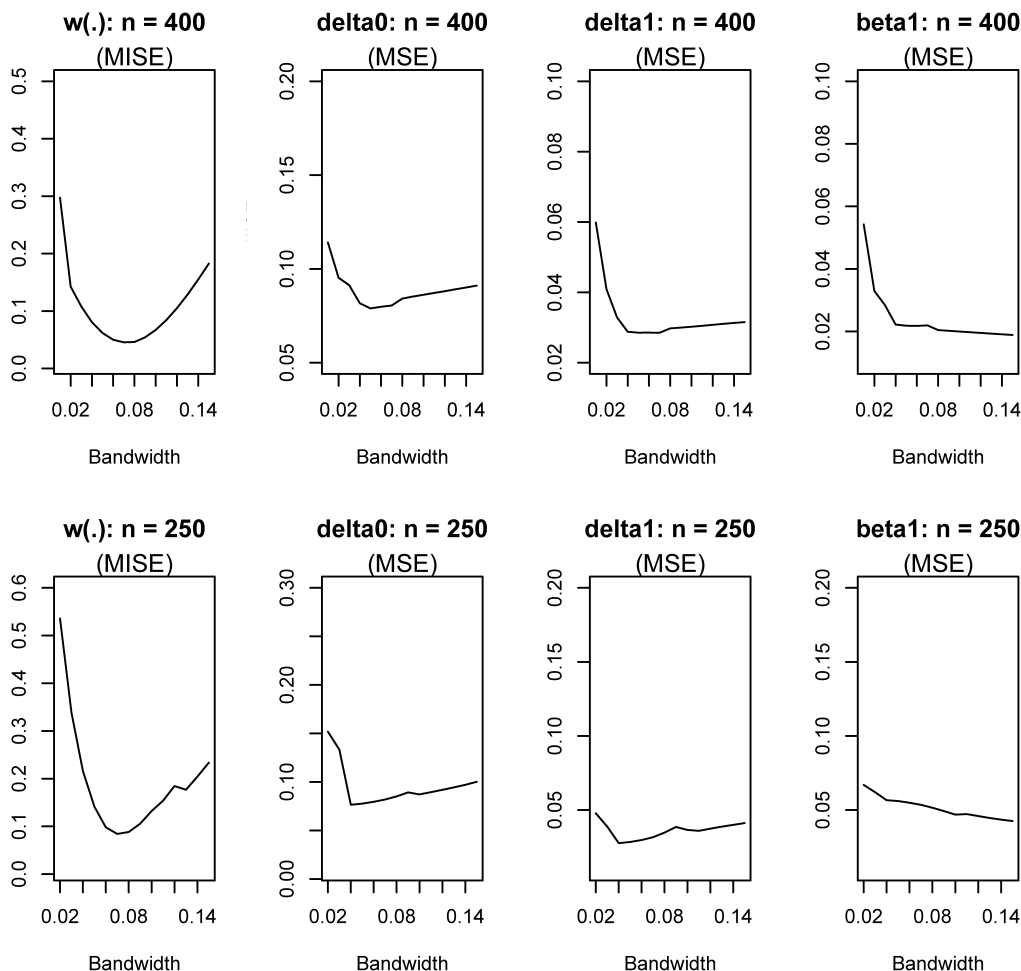


Figure 2. The average MISE of the nonparametric varying scale estimates \hat{w} and the average MSE of the parameter estimates $\hat{\delta}_0$, $\hat{\delta}_1$, and $\hat{\beta}_1$ against bandwidth values from .01 to .15. The first row is based on 200 simulated data of size $n = 400$; the second row, on 200 simulated data of size $n = 250$.

unknown $w(\cdot)$, Figure 2 depicts the average mean integrated squared errors (MISEs) of the estimated weight function $\hat{w}(z_i)$ and the average mean squared errors (MSEs) of regression parameter estimates as a function over a range of grid values of bandwidth. The figure shows that the performance is fairly robust to the bandwidth choice over a reasonable range.

Silverman (1986, pp. 45–46) suggested an empirical formula to compute the bandwidth in density estimation, which also is closely related to the default choice of bandwidth in R (Venables and Ripley 2002, pp. 127). In our current simulation setting, the empirical bandwidth choice would be around .10. Although it may not be the optimal choice, this empirical bandwidth choice seems to be acceptable in terms of estimating the both nonparametric and parametric elements of the varying-scale models. Note that the main focus of a varying-scale model is on the parametric elements. To avoid adding to the computing burden (especially in bootstrap) and theoretical complications of cross-validation, henceforth we use Silverman’s empirical formula to select our bandwidths.

We now consider the performance of various varying-scale models in parameter estimation. We study the following varying-scale forms: (a) the unscaled model $w(z_i) \equiv 1$, (b) the quadratic model $w(z_i) = \psi_0 + \psi_1 z_i + \psi_2 z_i^2$, (c) the cubic model

$w(z_i) = \psi_0 + \psi_1 z_i + \psi_2 z_i^2 + \psi_3 z_i^3$, and (d) the nonparametric $w(z_i)$ for unknown form of $w(\cdot)$. As discussed earlier, the scaling weight function is unique only up to a positive constant multiplier, and we can place a constraint on either the scale function $w(\cdot)$ or the parameters. For convenience, we set a constraint that $\beta_0 \equiv 1.25$ (true value), so that we can directly compare the estimates of other parameters to their true values.

In addition to the varying-scale models, we also fitted (e) a partial linear varying-coefficient model,

$$\mu_i = H\{\delta_1 v_i + \beta_0(z_i) + x_i \beta_1(z_i)\}$$

$$\text{with varying coefficients } \beta_0(z_i) \text{ and } \beta_1(z_i). \quad (10)$$

Here $\beta_0(z_i)$ and $\beta_1(z_i)$ correspond to $\delta_0 + w(z_i)\beta_0$ and $w(z_i)\beta_1$ in model (9). The regression parameter δ_1 is estimated by solving a semiparametrically efficient estimating equation similar to, but slightly simpler than, (7) (see also Lam and Fan 2008). If model (10) were a Gaussian model with an identity link function, then this estimator of δ_1 would be the equivalent to the efficient estimators considered by Ahmad et al. (2005).

Models (a)–(e) are nested sequentially from simplest to most complex. Figure 3 shows boxplots of model deviance estimates and estimates of all identifiable regression parameters. The unscaled and quadratic varying-scale models perform worse than

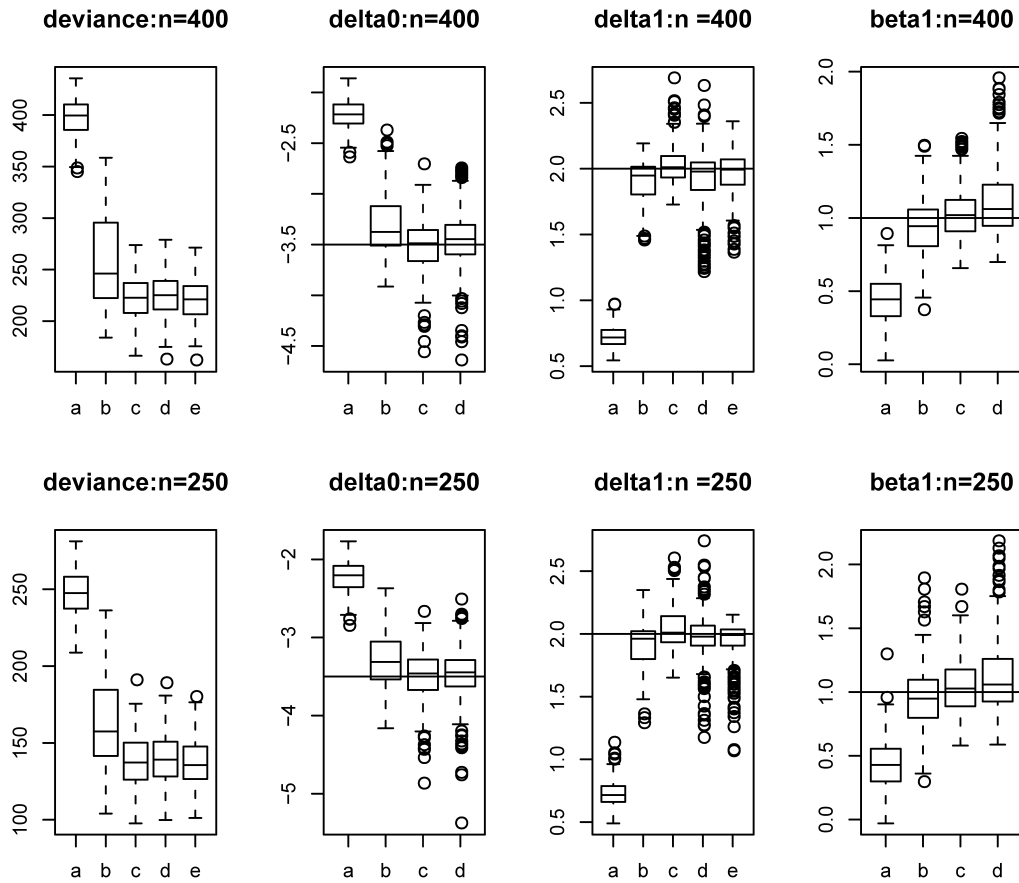


Figure 3. Side-by-side boxplots of model deviances and regression parameter estimates $\hat{\delta}_0$, $\hat{\delta}_1$, and $\hat{\beta}_1$ for fitting the (a) unscaled, (b) quadratic, (c) cubic, (d) nonparametric varying-scale models, and (e) the partially linear varying-coefficient model. The first row is based on 600 simulated data of size $n = 400$; the second row, on 600 simulated data of size $n = 250$.

the others. In each of the cubic nonparametric varying-scale models and the varying-coefficient model, the parameter estimates are more or less on the target. Clearly, both the cubic and nonparametric scale forms, as well as the varying-coefficient model, can more or less recover the shape of the true scale weight function in model (9), but the larger (or more complex) partially linear varying-coefficient model does not appear to give better results over the smaller (or simpler) cubic and nonparametric varying-scale models. From a model selection standpoint, we likely would prefer the simpler varying-scale model forms, which also have nice parametric interpretations and root- n inference on β . We also have examined the boxplots

of MSEs of the parameter estimates (not shown in the article), which point to the same conclusion.

We use the testing procedures described in Section 4 to test between various models. The first half of Table 1 summarizes the p values of testing four pairs of nested models: the quadratic varying-scale model (b) versus the cubic varying-scale model (c), the cubic varying-scale model (c) versus the nonparametric scale model (d), the nonparametric scale model (d) versus the partially linear varying-coefficient model (e). We calculate these from 120 simulated data sets of size $n = 400$ from the true varying-scale model (9). At significance level $\alpha = .05$, most often we would reject the null quadratic model against the al-

Table 1. Hypothesis testing: p values for various tests

True model	Varying-scale model (9)				Varying-coefficient model (10)			
	(b)/(c)	(c)/(d)	(c)/(e)	(d)/(e)	(b)/(c)	(c)/(d)	(c)/(e)	(d)/(e)
Tests (H_0/H_1)								
Minimum	.0000	.0010	.0120	.0050	.0000	.0090	.0000	.0000
First quarter	.0000	.2176	.2522	.1852	.0000	.0718	.0000	.0010
Median	.0000	.3277	.5601	.3292	.0000	.1460	.0000	.0080
Third quarter	.0000	.5018	.7364	.5045	.0000	.2620	.0030	.0754
Maximum	1.000	.9616	.9951	.7630	1.000	.8710	.3671	.3704
<5%	90.0%	3.4%	5.9%	5.1%	93.3%	19.2%	90.0%	70.3%

NOTE: The models compared are the (b) quadratic, (c) cubic, (d) nonparametric varying-scale models, and (e) the partially linear varying-coefficient model. Except for the test of model (b) versus (c), the p values are computed based on 1,000 bootstrap samples in each simulated data set.

ternative cubic model; however, about 95% of time, we would conclude the simpler null varying-scale models in the last three tests. The numbers reported in the last row are consistent with the theoretical developments on p value functions and testing powers. In fact, because the null model under H_0 is roughly true in the final three tests, the p value function theory (see, e.g., Fraser 1991) suggests that the second, third, and fourth values in the last row should be roughly around 5%. Also, because in the first test the alternative cubic model is roughly true, the first number in the last row, 90%, is roughly the power of the test (see, e.g., Beran 1986).

To further study the performance of these tests, we consider the alternative that the data are not from a varying-scale model, but from a partially linear regression varying-scale model (10) with true coefficients

$$\begin{aligned} \delta_1 &= 2.00, \\ \beta_0(z_i) &= -3.5 + 1.25\{1 + 2 \sin(2\pi z_i)\}, \end{aligned}$$

and

$$\beta_1(z_i) = 1.25\{1 + 2 \cos(2\pi z_i)\}.$$

Using 120 data sets of size 400 simulated from this model, we test the same four pairs of the varying-scale or varying-coefficient models. The second half of Table 1 summarizes the p values of these tests. When testing the varying-scale models (c) or (d) against the partially linear varying coefficient model (e), most often the p values are <5%, suggesting to reject the null varying-scale models at significant level $\alpha = 5\%$. Again from Beran (1986), we know that the last two numbers in the last row are roughly the powers of the corresponding tests. The results in Table 1 also show that in this setting, we are likely to separate the (c) cubic from the (b) quadratic varying-scale models but unlikely to distinguish the (c) cubic and the (d) non-parametric varying-scale models.

Using the bootstrap samples, we draw a QQ plot in each sample set by plotting the standardized bootstrap test statistic $a^* = (T_n^* - \bar{T}_n^*) / (2\bar{T}_n^*)^{1/2}$ against the standard normal quantiles. Here \bar{T}_n^* is the bootstrap sample mean of T_n^* . Almost all of the QQ plots (not shown here) suggest that a^* is consistent with a normal distribution, providing clear evidence of the Wilks phenomenon. When testing a parametric model against a varying-coefficient model, the Wilks phenomenon suggests that T_n can be standardized to asymptotically a standard normal by its mean and variance; the mean of T_n goes to infinity, and its variance is about the twice of the mean (see Fan et al. 2001).

The empirical evidence in simulation studies demonstrates that the proposed estimation and testing procedures work well in these settings. They can be used to estimate parameters and to select between varying-scale models or between a varying-scale model and a varying-coefficient model. A varying-scale model focuses on parametric components of the model, which has nice parametric interpretations. If in a test there is no significant difference between a varying-scale model and a larger, more complex varying-coefficient model, then the more parsimonious varying-scale model is preferred.

6. APPLICATIONS

6.1 Application to Ultrasound Risk Assessment Data

O'Brien et al. (2003) reported on an experiment in pigs to assess the risk of lesions in the lung due to focused ultrasound as a function of the ultrasound energy level (acoustic pressure in MPa). As secondary issue in this study was to investigate the age dependence of the risk. Figure 1(a) shows the observed incidence of lesions for pigs exposed to ultrasound beams focused at the lung surface. The scatterplot shows the age and acoustic pressure exposure level for each pig, and the plotting symbol indicates whether or not a lesion was present. A regression analysis of interest is to evaluate whether the lung damage is related to the acoustic pressure (in MPa) and age (in weeks). Furthermore, at any given age (in weeks), it also is of interest to know the increased risk (odds ratio) of lesions with a unit increase in the acoustic pressure, as well as the effective dose (ED) level of the acoustic pressure that may result in lesions.

Because in addition to the main age effect, a complicated interaction effect of age on the effect of acoustic pressure is expected, we fit the varying-scale model

$$\text{logit}\{E(y_j)\} = \delta_0 + v_j\delta_1 + (\beta_0 + x_j\beta_1)w_j. \quad (11)$$

To the data here the binary response y_j indicates whether the j th pig has lesions in the lung or not, v_j is age (in weeks), and x_j is the acoustic pressure (in MPa). The varying-scale weight $w_j = w(z_j)$ is an unknown nonparametric function of the age variable, where for convenience we use a transformed age variable $z_j = (v_j - \min(v)) / \max(v)$ (with values between 0 and 1) instead of the original age v_j (with values between 2 and 70). To ensure the parameters are fully identifiable, we set $\beta_0 = 1$. The result of the significance test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ is not affected whether we use constraint $\beta_0 = 1$ or constraint $w(z_*) = 1$ for a reference value z_* . Moreover, $e^{w_j\beta_1}$ is the increased odds of risk per MPa increase in the acoustic pressure at age v_j , and its value also is not affected by the choice of the constraint.

Because of the coefficient $w(z_j)$ in the varying-scale model (11), both the main effect for age and the interaction between age and acoustic pressure are nonlinear and are modeled non-parametrically. Constraining $\beta_0 = 1$, it follows from the preceding results that parametric root- n inferences can be applied to all of the remaining parameters δ_0 , δ_1 , and β_1 . These estimates and their standard errors, computed from the semiparametric large-sample theory, are given in Table 2.

The linear effect for age and the interaction term representing the age-dependent effect of acoustic pressure are highly significant. We considered replacing the linear function $\delta_0 + v_j\delta_1$ in (11) by the quadratic $\delta_0 + v_j\delta_1 + v_j^2\delta_2$, but the test of $H_0 : \delta_2 = 0$ was not significant, with a p value of .541. We also compared the varying-scale model (11) to the simpler parametric model,

$$\text{logit}\{E(y_j)\} = \gamma_0 + \gamma_1 v_j + \gamma_2 x_j + \gamma_{12} v_j x_j, \quad (12)$$

Table 2. Parameter estimates and standard errors in the varying-scale model for ultrasound risk

Parameter	δ_0	δ_1	β_1
Estimate	-8.6205	.0655	.7080
Standard error	(.7575)	(.0077)	(.0706)

Table 3. Bootstrap likelihood ratio test of parametric model versus semiparametric model for the ultrasound data

Fitted model	Null model	Deviance	Difference of deviances	<i>p</i> value*
Parametric model		271.99		
Varying-scale model	Parametric model	254.34	17.65	0

*Based on 1,000 bootstrap runs.

where $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2, \gamma_{12})^T$ are unknown parameters. Model (12) is nested within model (11) provided linear functions are included in the space of w . A bootstrap likelihood ratio test was performed as described in Section 5. The results, given in Table 3, indicate a significant deviation from the fully parametric model, so the expansion to the semiparametric model was warranted.

Figure 1(a) represents the nonparametric varying-scale model (11) by the calculated contours for 5% and 50% risk of lesions. For a given age, these curves give the ED05 and ED50. The plot clearly shows the nonlinear age dependence of the risk of lesions. The acoustic pressures corresponding to 50% risk are much higher than the levels used in conventional human diagnostic applications. Figure 1(b) plots the increased odds of risk of lesions, per MPa increase in acoustic pressure, against the age (solid line), together with the corresponding pointwise 95% confidence intervals (dotted lines). The plot shows that per MPa increase in acoustic pressure more than doubles the risk in lesions. In addition, the risk increase is the highest (more than three fold) in young pigs (age < 10 weeks), intermediate in older pigs, and lowest in middle-aged pigs (see O'Brien et al. 2003 for further discussion).

6.2 Application to Toxoplasmosis Data

The toxoplasmosis data set (Efron 1986, p. 710, table 1) contains the proportions of subjects testing positive for toxoplasmosis in 34 cities of El Salvador, the annual rainfall in the 34 cities, and the sample sizes of tested subjects. Efron (1978) used an ordinary logistic regression model to model the positive incidence rate as a function of rainfall in the j th city μ_j and found that a cubic regression on rainfall $\text{logit}(\mu_j) = \beta_0 + \beta_1 X_j + \beta_2 X_j^2 + \beta_3 X_j^3$ was highly significant, where $X_j = (x_j - \bar{x}) / \{\sum_{i=1}^{34} (x_i - \bar{x})^2\}^{1/2}$ and x_j is the annual rainfall in

the j th city. Efron (1986) reanalyzed the data using a binomial double-exponential model with the same cubic model for the incidence rate and a quadratic model for the overdispersion parameter $\phi_j = 1.25 / \{1 + \exp(-\psi_0 - \psi_1 z_j - \psi_2 z_j^2)\}$; this gave a method for modeling heterogeneity and overdispersion. Here $z_j = (n_j - \bar{n}) / \{\sum_{i=1}^{34} (n_i - \bar{n})^2 / 33\}^{1/2}$ and n_j is the sample size of tested subjects in the j th city. Efron (1986) found that the “effective size” $n_j \phi_j$ was quite different from the actual sample size n_j for many cities. Ganio and Schafer (1992) proposed a diagnostic tool for testing overdispersion in binomial and Poisson models. They studied the toxoplasmosis data in more detail under several dispersion models using double-binomial model-based as well as quasi-likelihood-based inferences. Their final model is the “simple” ordinary overdispersion logistic model with $\text{logit}(\mu_j) = \beta_0 + \beta_1 X_j + \beta_2 X_j^2 + \beta_3 X_j^3$ and $\text{var}(y_j) = n_j \mu_j (1 - \mu_j) / \phi$, where y_j is the total number of incidences in the j th city and ϕ is an unknown overdispersion parameter that is the same across all 34 cities.

We extend the model of Ganio and Schafer (1992) for the positive incidence rate to a varying-scale model,

$$\mu_j = \exp(w_j \eta_j) / \{1 + \exp(w_j \eta_j)\}$$

$$\text{with } \eta_j = \beta_0 + \beta_1 X_j + \beta_2 X_j^2 + \beta_3 X_j^3, \quad (13)$$

and keep the same overdispersion model, $\text{var}(y_j) = n_j \mu_j (1 - \mu_j) / \phi$. Here the scale weight $w_j = w(z_j)$, with z_j as defined in the previous paragraph, is a function of the sample size n_j of tested subjects in the j th city. The model defined by (13) adapts for heterogeneity in the data by adjusting the samples size in each city through the scale weight w_j , whose function is very much similar to the effective sample size described by Efron (1986) and Xie et al. (1997).

When (a) $w_j \equiv 1$, the varying-scale model (13) is the same as that of Ganio and Schafer (1992). Besides this, we consider three addition scale weight w_j forms: (b) quadratic $w_j = \psi_0 + \psi_1 z_j + \psi_2 z_j^2$, (c) cubic $w_j = \psi_0 + \psi_1 z_j + \psi_2 z_j^2 + \psi_3 z_j^3$, and (d) nonparametric regression $w_j = w(z_j)$ for unknown form of $w(\cdot)$. To avoid complications, in the case of nonparametric scale, we assume that the overdispersion parameter is fixed or estimated from an external source. Under this assumption, the semiparametric inference results developed in Section 3, except for the efficiency result, still hold.

Table 4 lists the parameter estimates of the cubic regression model for the incidence rate. The first column of nonscaled model corresponds to the model of Ganio and Schafer (1992), and the last column is from Efron’s double-binomial fit (Efron

Table 4. Parameter estimates of varying-scale models for the toxoplasmosis data

Parameter	Nonscaled	Quadratic	Cubic	Nonparametric	Double-binomial
β_0	.099 (.142)	.003 (.132)	-.040 (.057)	-.053 (.04)	-.071 (.14)
β_1	-.448 (.216)	-.829 (.267)	-.484 (.136)	-.657 (.07)	-.620 (.23)
β_2	-.187 (.127)	-.216 (.106)	-.165 (.075)	-.220 (.044)	-.170 (.11)
β_3	.213 (.089)	.298 (.092)	.203 (.057)	.304 (.033)	.272 (.09)

NOTE: The numbers in parentheses are standard errors.

Table 5. Deviance-based tests of unscaled logistic model versus varying-scale models

Scale models	Null model	Deviance	Difference of deviances	Difference of degrees of freedom	<i>p</i> value
Nonscaled		62.605			
Quadratic	Nonscaled	55.297	7.308	2	.109 _(.154*)
Cubic	Nonscaled	50.817	11.788	3	.060 _(.095*)
Nonparametric	Nonscaled	49.341	13.264		.037*

*The *p* value was computed from 1,000 bootstrap samples. All other *p* values were computed using the chi-squared approximation.

1986). It is clear that the cubic regression on rainfall is highly significant across all models. Constraint $w(z_0) = 1$ is set at $z_0 = -.06$, which corresponds to the sample size of the eighth city, $n_8 = 19$ (the closest to the mean sample size of the 34 cities).

Table 5 presents the model deviances of all four models (a)–(d), as well as three respective tests of the three varying-scale models (b)–(d) versus the ordinary, overdispersion logistic model (a). Likelihood ratio statistics T_n were adjusted for overdispersion by modifying the statistic as $\tilde{T}_n = T_n \hat{\phi}$, where $\hat{\phi}$ is the estimated overdispersion parameter under the alternative model. For the tests among the parametric models, the standard chi-squared asymptotics apply. Both chi-squared asymptotic based and bootstrap-based *p* values are obtained. For the test involving the nonparametric varying-scale model, only bootstrap *p* value is obtained. Because overdispersion exists, the bootstrap method in Section 4 is modified as follows. First, obtain the bootstrap samples and compute the bootstrap likelihood ratio test statistic T_n^* as described in Section 4. Then, to incorporate overdispersion, modify the test statistic $\tilde{T}_n^* = T_n^* \hat{\phi}$ and compute the *p* value by $\tilde{p}^* = \frac{1}{N} \sum_{k=1}^N \mathbf{1}_{\{\tilde{T}_n^* > T_n\}}$. Here $\hat{\phi}$ is an (external) consistent estimator of the overdispersion parameter ϕ . At significance level $\alpha = .05$, neither the parametric quadratic nor the cubic scale models offers a significant improvement over the simple overdispersion model (a). But the nonparametric scale model in (d) with bootstrap testing leads to a significant result: The bootstrap *p* value is $< \alpha = .05$, indicating significant improvement over the nonscaled model (a). This suggests that the nonparametric scaled link model has captured some of the heterogeneity in the toxoplasmosis data.

We also compared model (13) with the fully nonparametric varying-coefficient model,

$$\mu_j = \exp(\eta_j) / \{1 + \exp(\eta_j)\}$$

$$\text{with } \eta_j = \beta_0(z_j) + \beta_1(z_j)X_j + \beta_2(z_j)X_j^2 + \beta_3(z_j)X_j^3, \quad (14)$$

where all four coefficients $\beta_0(z_j)$, $\beta_1(z_j)$, $\beta_2(z_j)$, and $\beta_3(z_j)$ are unknown smooth functions of z_j . Based on the bootstrap testing method described in Section 4, the *p* value is .167, after adjusting for overdispersion. This suggests that the more parsimonious varying-scale model (13) is adequate for these data.

7. DISCUSSION

The varying-scale model (3) provides an effective approach to tackling variations in magnitudes of regression coefficients for heteroscedastic data. It has a regular regression term, while

allowing the other to have coefficients with varying magnitudes for different observations. Further extensions, such as allowing a finite number of different levels of scaling, at the cost of additional complexity in the analysis, are clearly possible.

Although the theoretical results given in Section 3 can be directly extended to multivariate \mathbf{z} variables, there remains the challenge of the ‘‘curse of dimensionality’’ in fitting multivariate nonparametric regression in the literature. To avoid such a problem in practice, we may use generalized additive models to model the nonparametric scale variable function $w(\cdot)$. We may also use a single-index model for the scale variable function, that is, let $w(\mathbf{z}) = w(\mathbf{z}^T \boldsymbol{\alpha})$ with a constraint $\|\boldsymbol{\alpha}\| = 1$. (See, e.g., Li 1991; Härdle, Hall, and Ichimura 1993; Carroll et al. 1997 for single-index models and an interpretation of the parameters $\boldsymbol{\alpha}$.) In either case, the proposed algorithm and theoretical results will remain the same or can be extended in a straightforward manner.

APPENDIX: EFFICIENT SCORE AND PROOF OF THEOREM 3

Here we provide a construction of the efficient score function and sketch the proof of the semiparametric lower bound of Theorem 3. Proofs of Theorems 1 and 2 are provided in the online supplemental material for this article.

Construction of Efficient Score Function \mathcal{S}^*

Let $g(\mathbf{x}, \mathbf{v}, z)$ be the joint density of $(\mathbf{x}, \mathbf{v}, z)$. The joint density of $(y, \mathbf{x}, \mathbf{v}, z)$ is

$$f(y, \mathbf{x}, \mathbf{v}, z) = \exp[\{y\psi - a(\psi)\}/\phi + b(y, \phi)]g(\mathbf{x}, \mathbf{v}, z), \quad (\text{A.1})$$

where $\psi = u(w^{(0)}(z)\eta^{(0)} + \tilde{\eta}^{(0)})$. Define $P = \{\text{model (A.1) with given } \boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0^T, \boldsymbol{\delta}_0^T)^T \text{ and density function } g(\cdot)\}$. Then, by the standard argument (e.g., Bickel et al. 1993), the tangent space of the nonparametric model P is $\{[y - \mu\{w^{(0)}(z)\eta^{(0)} + \tilde{\eta}^{(0)}\}] \tau\{w^{(0)}(z)\eta^{(0)} + \tilde{\eta}^{(0)}\} \eta^{(0)} a^*(z) \mid \text{for all } a^* \in L_2\}$. Thus the efficient score function is $\mathcal{S}^* = \mathcal{S} - \{\text{projection of } \mathcal{S} \text{ onto } P\}$, where $\mathcal{S} = [y - \mu\{w^{(0)}(z)\eta^{(0)} + \tilde{\eta}^{(0)}\}] \tau\{w^{(0)}(z)\eta^{(0)} + \tilde{\eta}^{(0)}\} \{w^{(0)}(z)\mathbf{x}\}$ is the score function for $\boldsymbol{\theta}_0$. Note that the MSE $E\{|y - \mu\{w^{(0)}(z)\eta^{(0)} + \tilde{\eta}^{(0)}\}] \tau\{w^{(0)}(z)\eta^{(0)} + \tilde{\eta}^{(0)}\} \{w^{(0)}(z)\mathbf{x}\} - \eta^{(0)} a^*(z)\}|^2$ achieves its minimum when $a^*(z) = (\{\mathbf{m}(z)\}^T, \{\tilde{\mathbf{m}}(z)\}^T)^T$. Thus the efficient score is

$$\mathcal{S}^* = [y - \mu\{w^{(0)}(z)\eta^{(0)} + \tilde{\eta}^{(0)}\}] \tau\{w^{(0)}(z)\eta^{(0)} + \tilde{\eta}^{(0)}\} \times \left\{ \begin{pmatrix} w^{(0)}(z)\mathbf{x} \\ \mathbf{v} \end{pmatrix} - \eta^{(0)} \begin{pmatrix} \mathbf{m}(z) \\ \tilde{\mathbf{m}}(z) \end{pmatrix} \right\}. \quad (\text{A.2})$$

For each $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$, replacing $w^{(0)}(\cdot)$, $\mathbf{m}(\cdot)$ and $\tilde{\mathbf{m}}(\cdot)$ by their estimators in \mathcal{S}^* leads to estimating equations (7).

Proof of Theorem 3

The form of the efficient score function S^* is given in (A.2). The Fisher information lower bound is $E\{S^*S^{*T}\}$, which is equal to A^{-1} (see Bickel et al. 1993). The theorem follows.

[Received February 2006. Revised December 2007.]

REFERENCES

- Aerts, M., Claeskens, G., and Hart, J. D. (1999), "Testing the Fit of a Parametric Function," *Journal of the American Statistical Association*, 94, 869–879.
- Agresti, A. (1996), *An Introduction to Categorical Data Analysis*, New York: Wiley.
- Ahmad, I., Leelahanon, S., and Li, Q. (2005), "Efficient Estimation of Semiparametric Partially Linear Varying-Coefficient Model," *The Annals of Statistics*, 33, 258–283.
- Beran, R. (1986), "Simulated Power Function," *The Annals of Statistics*, 14, 151–173.
- Bickel, P., Klaassen, C., Ritov, Y., and Wellner, J. (1993), *Efficient and Adaptive Inference for Semiparametric Models*, Baltimore: Johns Hopkins University Press.
- Breslow, N. E., and Holubkov, R. (1997), "Weighted Likelihood, Pseudo-Likelihood and Maximum Likelihood Methods for Logistic Regression Analysis of Two-Stage Data," *Statistics in Medicine*, 16, 103–116.
- Cai, Z., Fan, J., and Li, R. Z. (2000), "Efficient Estimation and Inferences for Varying-Coefficient Models," *Journal of American Statistical Association*, 95, 888–902.
- Carroll, R. J. (1982), "Adapting for Heteroscedasticity in Linear Models," *The Annals of Statistics*, 10, 1224–1233.
- Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1997), "Generalized Partially Linear Single-Index Models," *Journal of American Statistical Association*, 92, 477–489.
- Carroll, R. J., Ruppert, D., and Welsh, A. (1998), "Local Estimating Equations," *Journal of American Statistical Association*, 93, 214–227.
- Chiou, J. M., and Müller, H. G. (1998), "Quasi-Likelihood Regression With Unknown Link and Variance Functions," *Journal of the American Statistical Association*, 93, 1376–1387.
- Claeskens, G., and Carroll, R. J. (2007), "Post-Model Selection Inference in Semiparametric Models," *Biometrika*, 94, 249–265.
- Efron, B. (1978), "Regression and ANOVA With Zero–One Data: Measures of Residual Variation," *Journal of the American Statistical Association*, 73, 113–121.
- (1986), "Double-Exponential Families and Their Use in Generalized Linear Regression," *Journal of the American Statistical Association*, 81, 709–721.
- Eubank, R. L., and Hart, J. D. (1992), "Testing Goodness-of-Fit in Regression via Order Selection Criteria," *The Annals of Statistics*, 20, 1412–1425.
- Fan, J., and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, New York: Chapman & Hall.
- Fan, J., Zhang, C., and Zhang, J. (2001), "Generalized Likelihood Ratio Statistics and Wilks Phenomenon," *The Annals of Statistics*, 29, 153–193.
- Fraser, D. A. S. (1991), "Statistical Inference: Likelihood to Significance," *Journal of the American Statistical Association*, 86, 258–265.
- Ganio, L. M., and Schafer, D. W. (1992), "Diagnostics for Overdispersion," *Journal of the American Statistical Association*, 87, 795–804.
- Härdle, W., Hall, P., and Ichimura, H. (1993), "Optimal Smoothing of Single Index Models," *The Annals of Statistics*, 21, 157–178.
- Hastie, T., and Tibshirani, R. (1993), "Varying-Coefficient Models," *Journal of the Royal Statistical Society, Ser. B*, 55, 757–796.
- Hu, F., and Rosenberger, W. F. (2000), "Analysis of Time Trends in Adaptive Designs With Applications to a Neurophysiology Experiment," *Statistics in Medicine*, 19, 2067–2075.
- Hunsberger, S. (1994), "Semiparametric Regression in Likelihood-Based Models," *Journal of the American Statistical Association*, 89, 1354–1365.
- Johansen, S. (1984), *Functional Relations, Random Coefficients, and Nonlinear Regression With Application to Kinetic Data*, New York: Springer-Verlag.
- Lam, C., and Fan, J. (2008), "Profile-Kernel Likelihood Inference With Diverging Number of Parameters," *The Annals of Statistics*, in press.
- Lehmann, E. L. (1983), *Theory of Point Estimation*, New York: Wiley.
- Li, K.-C. (1991), "Sliced Inverse Regression for Dimension Reduction," *Journal of the American Statistical Association*, 86, 316–342.
- Lin, X., and Carroll, R. J. (2006), "Semiparametric Estimation in General Repeated Measures Problems," *Journal of the Royal Statistical Society, Ser. B*, 68, 68–88.
- McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman & Hall.
- O'Brien, W. D., Simpson, D. G., Ho, M.-H., Miller, R. J., Frizzell, L. A., and Zachary, J. F. (2003), "Suprathreshold Behavior and Threshold Estimation of Ultrasound-Induced Lung Hemorrhage in Pigs: Role of Age Dependency," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 50, 153–169.
- Ortega, J. M., and Rheinboldt, R. (1973), *Iterative Solution of Nonlinear Equations in Several Variables*, New York: Academic Press.
- Severini, T. A., and Staniswalis, J. G. (1994), "Quasi-Likelihood Estimation in Semiparametric Models," *Journal of the American Statistical Association*, 89, 501–511.
- Severini, T. A., and Wong, W. H. (1992), "Profile Likelihood and Conditionally Parametric Models," *The Annals of Statistics*, 20, 1768–802.
- Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, London: Chapman & Hall.
- Venables, W. N., and Ripley, B. D. (2002), *Modern Applied Statistics With S-PLUS* (4th ed.), New York: Springer.
- Weisberg, S., and Welsh, A. H. (1994), "Adapting for the Missing Link," *The Annals of Statistics*, 22, 1674–1700.
- Xie, M., Simpson, D. G., and Carroll, R. J. (1997), "Scaled Link Functions for Heterogeneous Ordinal Response Data," in *Modeling Longitudinal and Spatially Correlated Data: Methods, Applications, and Future Directions*, eds. T. Gregoire et al., New York: Springer-Verlag, pp. 23–36.
- Zhang, W., Lee, S.-Y., and Song, X. (2002), "Local Polynomial Fitting in Semivarying Coefficient Model," *Journal of Multivariate Analysis*, 82, 166–188.