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

Minge Xie, Douglas G. Simpson, and Raymond J. Carroll [1]

Summary

This paper describes a class of heteroscedastic generalized linear regression models in which a subset of the regression parameters are rescaled nonparametrically, and it 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. A bootstrap based goodness of scale test is also described. The methodology is illustrated with simulations, published data and data from collaborative research on ultrasound safety.

Some Key Words: Generalized linear regression; Heteroscedasticity; Nonparametric regression; Partially linear model; Semiparametric efficiency; Varying-coefficient model.

Running Title: Heterogeneity in Generalized Linear Models

[1]Minge Xie is Associate Professor and Director of Office of Statistical Consulting, Department of Statistics, Rutgers University, Piscataway, NJ 08854, mxie@stat.rutgers.edu. Douglas G. Simpson is Professor and Chair, Department of Statistics, University of Illinois, Champaign, IL 61820, dgs@uiuc.edu. Raymond J. Carroll is Distinguished Professor, Department of Statistics, Texas A&M University, College Station, TX 77843, carroll@stat.tamu.edu. This research is partly supported by NSF SES-0241859 (Xie), National Institute of Biomedical Imaging and Bioengineering grant EB02641 (Simpson), National Cancer Institute grants CA57030 and CA104620 (Carroll) and by the Texas A&M Center for Environmental and Rural health via 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.

1 INTRODUCTION

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

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

where μ is a given link function, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ represents the vector of primary covariates, and $\boldsymbol{\beta}(\mathbf{z}_i) = (\beta_1(\mathbf{z}_i), \ldots, \beta_p(\mathbf{z}_i))^{\mathrm{T}}$ 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 they developed a nonparametric likelihood ratio test whether, in fact, the coefficients are varying.

Semiparametric specializations of the model in (1) have been studied as well. 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:

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

where $E(\epsilon_i|\mathbf{x}_i, \mathbf{v}_i, \mathbf{z}_i) = 0$ and $\mathrm{var}(\epsilon_i|\mathbf{x}_i, \mathbf{v}_i, \mathbf{z}_i) = \sigma^2(\mathbf{z}_i)$. 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). Recently, Lam and Fan (2007) established the same type of semiparametric efficiency results in the class of partially linear generalized varying coefficient models. Their results could also accommodate growing number of predictors in the parametric part.

In the present 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 the *varying-scale generalized linear models*. For $i = 1, ..., n$, 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

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

where $\mu(\cdot)$ is a known link function, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^{\mathrm{T}}$ is a vector of p regression parameters subject to scale heterogeneity, $\boldsymbol{\delta} = (\delta_1, ..., \delta_q)^{\mathrm{T}}$ 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 below. 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 $(\boldsymbol{\beta}^{\mathrm{T}}, \delta^{\mathrm{T}})^{\mathrm{T}}$, and we show that these estimators achieve a semiparametric lower bound for asymptotic variance.

It is necessary to impose a constraint on either $w(\cdot)$ or $\boldsymbol{\beta}$ to ensure identifiability. For example, suppose 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. In the theoretical development we employ the latter constraint. In the reparameterized model 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 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, the model in (3) coincides with the model in (2). In such cases where β is a scalar one does not obtain parametric inferences for this parameter, because it is equivalent to a scaling of the nonparametric function $w(\cdot)$. If $\boldsymbol{\delta} = 0$ in (3) and $\boldsymbol{\beta}$ 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, page 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). They allowed 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 via 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 $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ 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 they 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, respectively. The nonparallelism of the curves is a reflection of the varying scale as a function of age. Fig 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.

[–Figure 1 approximately here–]

The rest of the paper is organized as follows. Section 2 presents the estimation framework and a profile local-likelihood algorithm for performing the estimation. Section 3 provides the main results on the large sample theory. Section 4 describes a bootstrap-based goodness of fit type of test for varying scales. Section 5 contains 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 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 the $(y_i, \mathbf{x}_i, \mathbf{v}_i, z_i)$, for $i = 1, 2, \ldots, n$, be independently and identically distributed copies of random variables $(y, \mathbf{x}, \mathbf{v}, z)$. For simplicity, we consider in the rest of the paper 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 the conditional density of y_i has the form

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

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

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

where $a(\cdot)$ and $b(\cdot, \cdot)$ are fully specified, ϕ is a possible fixed dispersion parameter, $\eta_i = \mathbf{x}_i^T\boldsymbol{\beta}$, $\widetilde{\eta}_i = \mathbf{v}_i^T\boldsymbol{\delta}$ 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 loglikelihood 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 loglikelihood 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 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 $\boldsymbol{\beta}_*^{\mathrm{T}} = (1, \boldsymbol{\beta}^{\mathrm{T}})$, where $\boldsymbol{\beta}$ is a $p-1$ dimensional vector. Corresponding to the form of $\boldsymbol{\beta}_*$ write $\mathbf{x}_{i*}^{\mathrm{T}} = (x_{i1}, \mathbf{x}_i^{\mathrm{T}})$, 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 $\boldsymbol{\beta}$ 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 $w^{(0)}(t)$ is second order differentiable. The vector

$\lambda = (\lambda_0, \lambda_1)^T$ depends on z_0, the form of the function $w^{(0)}(\cdot)$ and on the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\delta}^{\mathrm{T}})^{\mathrm{T}}$. Given prior values for $\boldsymbol{\theta}$ and $w^{(0)}$ write the local likelihood function at z_0 as

$$\ell_{LO}(\lambda) = \ell_{LO}(\lambda|\eta, \widetilde{\eta}) = \frac{1}{n}\sum_{i=1}^{n} \ell\Big(\mu[\{\lambda_0 + \lambda_1(z_i - z_0)\}\eta_i + \widetilde{\eta}_i], y_i\Big) K_b(z_i - z_0), \tag{6}$$

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

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

$$\sum_{i=1}^{n}[y_i - \mu\{\widehat{w}_\theta(z_i)\eta_i + \widetilde{\eta}_i\}]\tau\{\widehat{w}_\theta(z_i)\eta_i + \widetilde{\eta}_i\}\{\begin{pmatrix} \widehat{w}_\theta(z_i)\mathbf{x}_i \\ \mathbf{v}_i \end{pmatrix} + \eta_i \widehat{w}^{(1)}_\theta(z_i)\} = 0. \tag{7}$$

Here $\tau(s) = \mu'(s)/a''\{u(s)\}$, $u(s) = \{(a')^{-1} \circ (\mu)\}(s)$, $a(\cdot)$ is 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 $\boldsymbol{\theta}_0$, $\widehat{w}_\theta(z_0)$ is a consistent estimator of $w^{(0)}(z_0)$, and its derivative with respect to $\boldsymbol{\theta}$, $\widehat{w}^{(1)}_\theta(z_0)$, is a consistent estimator of a term related to the projection obtaining the efficient score function; See Section 3 and the Appendix 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

$$\begin{aligned} \widehat{\boldsymbol{\theta}}^{\text{new}} \;=\; \widehat{\boldsymbol{\theta}}^{\text{old}} + \Big(\mathbf{A}_n(\boldsymbol{\theta})\frac{1}{n}\sum_{i=1}^{n}[y_i - \mu\{\widehat{w}_\theta(z_i)\eta_i + \widetilde{\eta}_i\}]\{\begin{pmatrix} \widehat{w}_\theta(z_i)\mathbf{x}_i \\ \mathbf{v}_i \end{pmatrix} \\ + \eta_i \widehat{w}^{(1)}_\theta(z_i)\}\tau\{\widehat{w}_\theta(z_i)\eta_i + \widetilde{\eta}_i\}\Big)\Big|_{\theta=\widehat{\theta}^{\text{old}}}. \end{aligned} \tag{8}$$

where

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

Note that $\mathbf{A}_n^{-1}(\boldsymbol{\theta})|_{\theta=\widehat{\theta}}$ is an estimate of the covariance matrix of the efficient score function $\mathcal{S}^*$; see Section 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 $\widehat{w}_\theta^{\text{new}}(z_i)|_{\theta=\widehat{\theta}^{\text{curr}}}$ and its derivative with respect to $\boldsymbol{\theta}$ by $\widehat{w}_\theta^{(1)}(z_i)|_{\theta=\widehat{\theta}^{\text{cur}}}$.

In the first module, (8) is a variant of the iteratively reweighted least square 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 employed. In the examples in Section 6, we use a modified version of the algorithm of Cai, Fan and Li (2000) to update the w estimates.

The semiparametric estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\delta}}$ 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

We develop in this section large sample theory under the framework of the preceding section. These asymptotic results are developed under the regularity conditions listed below. They may not be the weakest possible conditions, but they provide a mild set of sufficient conditions often satisfied in practice. More rigorous treatment of semiparametric efficiency for profile likelihood method can be found in Severini and Wong (1992), and more recently, Lin and Carroll (2006) and Claeskens and Carroll (2007) for local linear method without varying coefficient and Lam and Fan (2007) for the partially linear generalized varying coefficient models.

Throughout we use a symmetric kernel function and assume independent observations following the model defined by equations (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 $\upsilon_2 = \int t^2 K(t)dt$, $\nu_0 = \int\{K(t)\}^2 dt$ and $\nu_2 = \int t^2\{K(t)\}^2 dt$ are all bounded above.

(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, covariate variables $\mathbf{x}$, $\mathbf{v}$ and z respectively. We assume y has finite 4^{th} 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.

In order to state our main results we introduce the following notations. Denote by $\gamma(t) = \mathrm{E}[(\eta^{(0)})^2\,\tau_1\{w^{(0)}(z)\eta^{(0)}+\widetilde{\eta}^{(0)}\}|z=t]$, $\gamma_1(t) = \mathrm{E}([\tau_1\{w^{(0)}(z)\eta^{(0)}+\widetilde{\eta}^{(0)}\}\eta^{(0)}]\mathbf{x}|z=t)$ and $\widetilde{\gamma}_1(t) = \mathrm{E}([\tau_1\{w^{(0)}(z)\eta^{(0)}+\widetilde{\eta}^{(0)}\}\eta^{(0)}]\mathbf{v}|z=t)$, where $\eta^{(0)} = \mathbf{x}^T\boldsymbol{\beta}_0$, $\widetilde{\eta}^{(0)} = \mathbf{v}^{\mathrm{T}}\boldsymbol{\delta}_0$ and $\tau_1(s) = \mu'(s)\tau(s)$. Write $\mathbf{J} = \mathrm{diag}\{1, b\}$, $\mathbf{H} = f_z(z_0)\gamma(z_0)\mathrm{diag}\{1, \upsilon_2\}$ and $\boldsymbol{\Lambda} = f_z(z_0)\gamma(z_0)\mathrm{diag}\{\nu_0, \nu_2\}$, where $\mathrm{diag}\{s_1, \ldots, s_k\}$ represents a $k \times k$ diagonal matrix of elements $s_1, \ldots, s_k$. Also, 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 $\widehat{\lambda}_\theta = \widehat{\lambda}_\theta(z_0)$ that maximizes local likelihood function (6).

Theorem 1. *Suppose the bandwidth $b = O(n^{-\xi})$, $1/6 < \xi < 1/4$, and $\mathbf{H}$ is invertible. 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*

$$\sqrt{nb}\Big\{\mathbf{J}(\widehat{\lambda}_\theta - \lambda) - \frac{b^2}{2}w^{(2)}(z_0)\begin{pmatrix}\upsilon_2\\0\end{pmatrix} - w^{(0)}(z_0)\{\gamma(z_0)\}^{-1}\begin{pmatrix}\{\gamma_1(z_0)\}^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\\0\end{pmatrix}$$
$$-\{\gamma(z_0)\}^{-1}\begin{pmatrix}\{\widetilde{\gamma}_1(z_0)\}^T(\boldsymbol{\delta}-\boldsymbol{\delta}_0)\\0\end{pmatrix} + o_p\Big(\frac{1}{\sqrt{n}}\Big)\Big\} \longrightarrow \mathrm{Normal}(\mathbf{0}, \mathbf{H}^{-1}\boldsymbol{\Lambda}\mathbf{H}^{-1}).$$

The proof of the theorem, together with the asymptotic expansions of $\widehat{\lambda}_\beta$ and $\widehat{\lambda}_\beta^{(1)}$, up to the order of $o_p(n^{-1/2})$, can be found in the online supplemental material for the article. 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 $w^{(0)}(z_0)$, and more.

Denote by $\mu_{w\mathbf{xv}} = \eta^{(0)}(\{\mathbf{m}(z)\}^T, \{\widetilde{\mathbf{m}}(z)\}^T)^T$, 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

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

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 $\mathbf{A}^{-1}$ is positive definite. Let $\widehat{w}_\theta(\cdot)$ and $\widehat{w}_\theta^{(1)}(\cdot)$ be respectively the local maximum likelihood estimator 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 $\|\widehat{\boldsymbol{\theta}}^{\text{new}} - \boldsymbol{\theta}_0\| = O_p(n^{-1/2})$. Also, as $n \to \infty$,*

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

In addition, we can estimate the asymptotic covariance matrix $\mathbf{A}$ consistently by $\mathbf{A}_n(\boldsymbol{\theta})|_{\theta=\widehat{\theta}}$, where

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

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.*

The proofs of Theorems 2 and 3 can be found 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 nonparametric scale, as a means for 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 a unscaled model. Model (M2) covers many parametric scale weight models, including for example polynomial models of order p with $w_i = \psi_0 + \psi_1 z_i + \ldots + \psi_p z_i^p$. Model (M3) is a nonparametric scale weight model. We assume all functions $t_j(\cdot)$ and $w(\cdot)$ have the 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 may also 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 above hypothesis testing problems, we may also be interested in testing any of these varying scale models against a partially linear varying coefficient model:

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

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 bigger 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, respectively. 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 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-square 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 there is 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 the similar test problems in linear regression models.

Here we consider a bootstrap approach facilitate the model testing. This approach is similar to the bootstrap methods described by Cai et al. (2000) and others in 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, utilizing 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^*. Repeating the bootstrap a large number of times (say N), and 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 cut-off value, i.e., $p^* = \frac{1}{N}\sum_{k=1}^{N} \mathbf{1}_{(T_n^* > T_n)}$, where T_n^* is the corresponding T_n value in the k^{th} bootstrap sample.

The empirical evidence in our simulation studies in Section 5 suggests very reasonable performance for this bootstrap procedure. We also see in the simulation evidence of the Wilks phenomenon (chi-square 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 $w(\cdot)$ are modeled parametrically. When $w(\cdot)$ are modeled nonparametrically, we suggest to use the Wald-type or score tests, which are supported by the results 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 following logistic varying scaling model:

$$\mu_i = H\{\delta_0 + \delta_1 v_i + w(z_i)(\beta_0 + x_i\beta_1)\} \text{ with } w(z_i) = 1 + 2\sin(2\pi z_i), \text{ for } i = 1, \ldots, n, \quad (9)$$

where $H(u) = \exp(u)/\{1 + \exp(u)\}$. Let the auxiliary variable z_i follow a uniform distribution on [0,1], and 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 equal to $(0, 0.5, 1)$, variance equal to $(1, 1, 2)$ and correlations equal to $(1/2, 1/\sqrt{2}, 1/3)$ between the first and the second, the first and the third, the second and the third elements, respectively. Then, set $z = \Phi(U_1)$, $x = U_2$ and $v = U_3$. This setting is similar to that in the simulation section of Zhang, Lee and Song (2002) and among others, where trigonometric functions are used for the unknown function in nonparametric regressions, 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 will 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 fittings. Fitting 200 data sets of size $n = 250$ and size $n = 400$ to model (9) with unknown $w(\cdot)$, Figure 2 depicts the average mean integrated squared errors (MISE) of the estimated weight function $\hat{w}(z_i)$ and the average mean squared errors (MSE) of regression parameter estimates, as a function over a range of grid values of bandwidth. It indicates that the performance is fairly robust to the bandwidth choice over a reasonable range.

[–Figure 2 approximately here–]

Silverman (1986, pp. 45-46) suggested an empirical formula to compute the bandwidth in density estimation, which is also closely related to the default choice of bandwidth in R (Venables and Ripley, 2002, pp. 127). Under our current simulation setting, the empirical bandwidth choice would be around 0.10. Although it may not be the optimal choice, this empirical bandwidth choice seems all right 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 further computing burden (especially in bootstrap) and theoretical complication of cross validation we use Silverman's empirical formula to select our bandwidths in the rest of the paper.

Let us consider the performance of various varying scale models in parameter estimation. The varying scale forms considered are: (a) unscaled model $w(z_i) \equiv 1$, (b) quadratic model $w(z_i) = \psi_0 + \psi_1 z_i + \psi_2 z_i^2$, (c) 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 before, 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 the estimates of other parameters can be directly compared to their true values.

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

$$\mu_i = H\{\delta_1 v_i + \beta_0(z_i) + x_i\beta_1(z_i)\} \text{ with varying coefficients } \beta_0(z_i) \text{ and } \beta_1(z_i). \tag{10}$$

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), respectively. The regression parameter δ_1 is estimated by solving a semiparametrically efficient estimating equation similar to (but slightly simpler than) equations (7); see, also, Lam and Fan (2007). If model (10) were a Gaussian model with an identity link function, this estimator of δ_1 would be the equivelant to the efficient estimators considered by Ahmad, Leelahanon and Li (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 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 model selection viewpoint, we would likely prefer the simpler varying scale model models, which also have nice parametric interpretations and rot-n inference on β. We also have examined the boxplots of mean squared errors (MSE) of the parameter estimates (not shown in the paper), which point to the same conclusion.

[–Figure 3 approximately here–]

We use the testing procedures described in Section 4 to test between various models. The first half of Table 1 on the left hand side summarizes the p-values of testing four pairs of nested models: the quadratic varying scale model (b) versus the 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). They are calculated from 120 simulated data sets of size $n = 400$ from the true varying scale model (9). At significant level $\alpha = .05$, most of the time we would reject the null quadratic model against the alternative 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, since the null model under H_0 is roughly true in the last three tests, the p-value function theory (see, e.g., Fraser, 1991) suggests that the second, third and forth values in the last row should be roughly around 5% . Also, since in the first test the alternative cubic model is roughly true, the first number, 90%, in the last row is roughly the power of the test (see, e.g., Beran, 1986).

To further study the performance of these tests, we consider alternatively the data is in fact not from a varying scale model, but from a partially linear regression varying scale model (10) with true coefficients

$$\delta_1 = 2.00,\ \beta_0(z_i) = -3.5 + 1.25\{1 + 2\sin(2\pi z_i)\} \text{ and } \beta_1(z_i) = 1.25\{1 + 2\cos(2\pi z_i)\}$$

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 test the varying scale models (c) or (d) against the partially linear varying coefficient model (e), most of the time the p-values are less than 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) nonparametric 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 the QQ-plots (not shown in the paper) suggest that a^* is consistent with a normal distribution, providing clear evidence of 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, Zhang, Zhang (2001).

Table 1: Hypothesis Testing: p-values for various tests

True Model	Varying Scale Model (9)				Varying Coefficient Model (10)			
Tests (H_0/H_1)	(b)/(c)	(c)/(d)	(c)/(e)	(d)/(e)	(b)/(c)	(c)/(d)	(c)/(e)	(d)/(e)
Min	.0000	.0010	.0120	.0050	.0000	.0090	.0000	.0000
1st Qu.	.0000	.2176	.2522	.1852	.0000	.0718	.0000	.0010
Median	.0000	.3277	.5601	.3292	.0000	.1460	.0000	.0080
3rd Qu.	.0000	.5018	.7364	.5045	.0000	.2620	.0030	.0754
Max	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, and (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 1000 bootstrap samples in each simulated data.

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 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 bigger and 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 on pigs to assess risk of lesions in the lung due to focused ultrasound as a function of the ultrasound energy level (acoustic pressure in megapascals or 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 scatter plot 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 see whether the damage in lung is related to

the acoustic pressure (in MPa) and age (in weeks). Furthermore, at any given age (in weeks), it is also of interest to know the increase risk (odds ratio) of lesions with a unit increase in the acoustic pressure. We also need to find out the effective dose (ED) level of the acoustic pressure that may result in lesions.

Since in addition to the main age effect, a complicated interaction effect of age on the effect of acoustic pressure is expected, we fit to the data the following varying scale models

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

Here, the binary response y_j indicates whether the jth pig has lesions in the lung or not, v_j is age (in weeks) and x_j is the acoustic pressure in mega-pascals (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 from 2 to 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 $w(z_*) = 1$ for a reference value z_*. Also, $e^{w_j\beta_1}$ is the increase odds of risk per MPa increase in the acoustic pressure at age v_j, and its value is not affected by the choice of the constraint either.

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 modeled nonparametrically. Constraining $\beta_0 = 1$, it follows from the results of preceding sections 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.

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	0.0655	0.7080
Standard Error	(0.7575)	(0.0077)	(0.0706)

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 0.541. We also compared the varying scale model (11) to the simpler parametric model:

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

where $\gamma = (\gamma_0, \gamma_1, \gamma_2, \gamma_{12})^T$ 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, in Table 3, and indicated a significant deviation from the fully parametric model, so the expansion to the semiparametric model was warranted.

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

Fitted Model	Null Model	Deviance	Diff. of Deviances	p-value*
Parametric model	-	271.99		
Varying scale model	Parametric model	254.34	17.65	.000

*Based on 1000 bootstrap runs.

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, respectively. 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 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). The plot shows that per MPa increase in acoustic pressure causes more than twice the increase of risk in lesions. Also, the risk increase is the largest (more than three times) in young pigs less than 10 weeks, followed by the older pigs, and least in middle aged pigs. See O'Brien et al. (2003) for further discussion.

6.2 Application to Toxoplasmosis Data

The toxoplasmosis data (Efron, 1986, page 710, Table 1) contains the proportions of subjects testing positive for the disease 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 jth 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) re-analyzed 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, $z_j = (n_j - \bar{n})/\{\sum_{i=1}^{34}(n_j - \bar{n})^2/33\}^{1/2}$ 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 further detail under several dispersion models using double binomial model based as well as quasi-likelihood based inferences. The final model that they concluded 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 jth city, ϕ is an unknown overdispersion parameter and it is the same across all 34 cities.

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

$$\mu_j = \exp(w_j\eta_j)/\{1 + \exp(w_j\eta_j)\} \quad \text{with } \eta_j = \beta_0 + \beta_1 X_j + \beta_2 X_j^2 + \beta_3 X_j^3, \tag{13}$$

and keep the same overdispersion model $\text{var}(y_j) = n_j\mu_j(1 - \mu_j)/\phi$. Here, the scale weight $w_j = w(z_j)$, z_j 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 in Efron (1986) and Xie, Simpson and Carroll (1997).

When (i) $w_j \equiv 1$, the varying-scale model (13) is the same as the one concluded in Ganio and Schafer (1992). Besides this, we consider three addition scale weight w_j forms: (ii) quadratic $w_j = \psi_0 + \psi_1 z_j + \psi_2 z_j^2$, (iii) cubic $w_j = \psi_0 + \psi_1 z_j + \psi_2 z_j^2 + \psi_3 z_j^3$, and (iv) 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 non-scaled model corresponds to the model concluded by in Ganio and Schafer (1992), and the numbers in last column is from Efron's double-binomial fit (Efron, 1986). It is clear that the cubic regression on rainfall is highly significant across all models.

TABLE 4: Parameter estimates of varying-scale models for Toxoplasmosis data.

Parameter	Non-scaled	Quadratic	Cubic	Nonparametric	Double-Binomial
β_0	.099	.003	$-.040$	$-.053$	$-.071$
	(0.142)	(0.132)	(.057)	(.04)	(.14)
β_1	$-.448$	$-.829$	$-.484$	$-.657$	$-.620$
	(.216)	(.267)	(.136)	(.07)	(.23)
β_2	$-.187$	$-.216$	$-.165$	$-.220$	$-.170$
	(.127)	(.106)	(.075)	(.044)	(.11)
β_3	.213	.298	.203	.304	.272
	(.089)	(.092)	(.057)	(.033)	(.09)

The numbers in the Parentheses are the standard errors. Constraint $w(z_0) = 1$ are set at $z_0 = -0.06$, 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 contains the model deviances of all four models (i)-(iv), as well as three respective tests of the three varying-scale models (ii)-(iv) versus the ordinary, over-dispersion logistic model (i). Likelihood ratio statistics T_n were adjusted for overdispersion by modifying the statistic as $\tilde{T}_n = T_n\widehat{\phi}$, where $\widehat{\phi}$ is the estimated overdispersion parameter under the alternative model. For the tests between the parametric models, the standard chi-square asymptotics apply. Both chi-square asymptotic based and bootstrap based p-values are obtained. For the test involving nonparametric varying scale model, only bootstrap p-value is obtained. Since 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^* the as described in Section 4. Then, to incorporate overdispersion, modify the test statistic $\tilde{T}_n^* = T_n^*\widehat{\phi}$, and compute the p-value by $\tilde{p}^* = \frac{1}{N}\sum_{k=1}^{N}\mathbf{1}_{(\tilde{T}_n^* > T_n)}$. Here, $\widehat{\phi}$ is an (external) consistent estimator of the overdispersion parameter ϕ. At significance level

$\alpha = .05$, neither the parametric quadratic and nor the cubic scale models offers a significant improvement over the simple overdispersion model (i). But, the nonparametric scale model in (iv) with bootstrap testing leads to a significant result: the bootstrap p-value is less than $\alpha = .05$, indicating significant improvement over the non-scaled model (i). It suggests that the nonparametric scaled link model have captured some of the heterogeneity in the toxoplasmosis data.

TABLE 5: Deviance based tests of unscaled logistic model versus varying-scale models (Toxoplasmosis Data)

Scale Models	Null Model	Deviance	Diff. of Deviances	Diff. of DF	p-value
Non-scaled	-	62.605			
Quadratic	Non-scaled	55.297	7.308	2	.109 (.154*)
Cubic	Non-scaled	50.817	11.788	3	.060 (.095*)
Nonparametric	Non-scaled	49.341	13.264	-	.037*

* Indicates the p-value was computed from 1000 bootstrap samples. All other p-values were computed using the chi-square approximation.

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

$$\mu_j = \exp(\eta_j)/\{1+\exp(\eta_j)\} \quad \text{with } \eta_j = \beta_0(z_j) + \beta_1(z_j)X_j + \beta_2(z_j)X_j^2 + \beta_3(z_j)X_j^3, \quad (14)$$

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 0.167, after adjusting for overdipersion. It 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 tackle 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 are clearly possible, such as allowing a finite number of different levels of scaling, at the cost of additional complexity in the analysis.

Although the theoretical results 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 might use generalized additive models (GAM's) to model the nonparametric scale variable function $w(\cdot)$. We may also use a single index model for the scale variable function, i.e., let $w(\mathbf{z}) = w(\mathbf{z}^{\mathrm{T}}\boldsymbol{\alpha})$ with a constraint $||\boldsymbol{\alpha}|| = 1$. See, e.g., Li (1991), Härdle, Hall, and Ichimura (1993), or 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

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

Construction of Efficient Score Function $\mathcal{S}^*$. Let $\mathrm{g}(\mathbf{x}, \mathbf{v}, z)$ be the joint density of ($\mathbf{x}$, $\mathbf{v}$, z). The joint density of $(y, \mathbf{x}, \mathbf{v}, z)$ is

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

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

$$\mathcal{S}^* = [y - \mu\{w^{(0)}(z)\eta^{(0)} + \widetilde{\eta}^{(0)}\}]\tau\{w^{(0)}(z)\eta^{(0)} + \widetilde{\eta}^{(0)}\}\Big\{\begin{pmatrix} w^{(0)}(z)\mathbf{x} \\ \mathbf{v} \end{pmatrix} - \eta^{(0)}\begin{pmatrix} \mathbf{m}(z) \\ \widetilde{\mathbf{m}}(z) \end{pmatrix}\Big\}. \tag{A.2}$$

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

Proof of Theorem 3. The form of the efficient score function $\mathcal{S}^*$ is given in equation (A.2). The Fisher information lower bound is $\mathrm{E}\{\mathcal{S}^*\mathcal{S}^{*T}\}$, which is equal to $\mathbf{A}^{-1}$; see Bickel et al. (1993). The theorem follows.

REFERENCES

Aerts, M., Claeskens, G. and Hart, J.D. (1999). Testing the fit of a parametric function. *Journal of American Statistical Association*, 94, 869-879.

Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. Wiley, New York.

Ahmad, I., Leelahanon, S., and Li, Q. (2005). Efficient estimation of semiparametric partially linear varying coefficient model. *Annals of Statistics*. 33, 258-283.

Beran, R. (1986). Simulated power function. *Annals of Statistics*. 14, 151-173.

Bickel, P., Klaasen, 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. *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.

Efron, B. (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 *Annals of Statistics*, 20, 1412-1425.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*. Chapman & Hall. New York.

Fan, J., Zhang, C. and Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *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. *Annals of Statistics*, 21, 157-178.

Hastie, T., Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society*. Series 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.* Lecture Notes in Statistics, 22, Springer-Verlag, New York.

Lam, C. and Fan, J. (2007). Profile-Kernel Likelihood Inference with Diverging Number of Parameters. *Annals of Statistics*. In press.

Lehmann, E. L. (1983). *Theory of Point Estimation*. Wiley, New York.

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. *J. R. Statist. Soc. B*, 68, 68-88.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models* (2nd edition). London: Chapman and Hall.

O'Brien, W. D., Simpson, D. G., Ho, M.-H., Miller, R. J., Frizzell, L. A. and Zachary, J. F. (2003). Superthreshold 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.* Academic, New York.

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. *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 edition. Springer, New York.

Weisberg, S. and Welsh, A. H. (1994). Adapting for the missing link. *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*, Ed. T. Gregoire et al., Springer-Verlag, Lecture Notes in Statistics 122, 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.

Figure 1. Figure 1(a) is 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 megaPascals (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. Figure 1(b) is 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).

Figure 2. The Figure plots the average MISE of the nonparametric varying scale estimates $\widehat{w}$, the average MSE of the parameter estimates $\hat{\delta}_0$, $\hat{\delta}_1$ and $\hat{\beta}_1$ againgst bandwidth values from 0.01 to 0.15. The first row is based on 200 simulated data of size $n = 400$, and the second row is based on 200 simulated data of size $n = 250$.

Figure 3. These are 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 $n = 400$, and the second row is based on 600 simulated data of size $n = 250$.

use the direct method instead

use the direct method instead

use the direct method instead