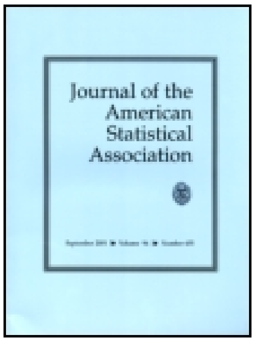

Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Local Estimating Equations

Raymond J. Carroll , David Ruppert & Alan H. Welsh

To cite this article: Raymond J. Carroll , David Ruppert & Alan H. Welsh (1998) Local Estimating Equations, Journal of the American Statistical Association, 93:441, 214-227, DOI: 10.1080/01621459.1998.10474103

To link to this article: https://doi.org/10.1080/01621459.1998.10474103

Published online: 17 Feb 2012.

Local Estimating Equations

Raymond J. CARROLL, David RUPPERT, and Alan H. WELSH

Estimating equations have found wide popularity recently in parametric problems, yielding consistent estimators with asymptotically valid inferences obtained via the sandwich formula. Motivated by a problem in nutritional epidemiology, we use estimating equations to derive nonparametric estimators of a "parameter" depending on a predictor. The nonparametric component is estimated via local polynomials with loess or kernel weighting; asymptotic theory is derived for the latter. In keeping with the estimating equation paradigm, variances of the nonparametric function estimate are estimated using the sandwich method, in an automatic fashion, without the need (typical in the literature) to derive asymptotic formulas and plug-in an estimate of a density function. The same philosophy is used in estimating the bias of the nonparametric function; that is, an empirical method is used without deriving asymptotic theory on a case-by-case basis. The methods are applied to a series of examples. The application to nutrition is called "nonparametric calibration" after the term used for studies in that field. Other applications include local polynomial regression for generalized linear models, robust local regression, and local transformations in a latent variable model. Extensions to partially parametric models are discussed.

KEY WORDS: Asymptotic theory; Bandwidth selection; Local polynomial regression; Logistic regression; Measurement error; Missing data; Nonlinear regression; Partial linear models; Sandwich estimation.

1. INTRODUCTION

A general methodology that has found wide popularity recently, especially in biostatistics, is to estimate parameters via estimating equations. Maximum likelihood estimates, robust regression estimates (Huber 1981), variance function estimates (Carroll and Ruppert 1988), generalized estimating equation (GEE) estimates (Diggle, Liang, and Zeger 1994), marginal methods for nonlinear mixed-effects models (Breslow and Clayton 1993), and indeed most of the estimators used in non-Bayesian parametric statistics are all based on the same technology. If the data are independent observations $(\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n)$, with the Ys possibly vector valued, then a parameter Θ is estimated by solving the estimating equation

$$0 = \sum_{i=1}^{n} \psi(\mathbf{Y}_i, \hat{\Theta}). \qquad (1)$$

We allow Θ to be vector valued, and ψ must have the same dimension as Θ. For example, maximum likelihood estimates are versions of (1) when $\psi(\cdot)$ is the derivative of the log-likelihood function.

One of the reasons that estimating equation methodology has become so popular is that for most estimating equations, the covariance matrix of the parameter estimate can be consistently and nonparametrically estimated using the so-called "sandwich formula" (Huber 1967) described in detail in Section 3.2.

The combination of estimating equations and sandwich covariance matrix estimates thus form a powerful general methodology. In this article we pose the following simple question: How does one proceed if Θ depends in an unknown way on an observable variable Z, so that $\Theta = \Theta(Z)$? The question arises naturally in the context of calibration studies in nutritional epidemiology; Section 2 provides a detailed discussion.

Our aim is to provide methods with the same generality as parametric estimating equations and the sandwich method. Starting only from the parametric estimating equation (1), we propose to develop estimates of $\Theta(Z)$ and use the sandwich method to form consistent and nonparametric estimates of the covariance matrix.

The method that we proposed, called *local estimating equations,* essentially involves estimating $\Theta(Z)$ by local polynomials with local weighting of the estimating equation. The specific application in nutrition is called *nonparametric calibration* because of its roots in nutritional epidemiology calibration studies. This article is concerned primarily with the case where Z is scalar, although in Section 4.2 we describe extensions to the multivariate case and present a numerical example.

In practice, it is often the case that $\Theta(z)$ is a q-dimensional vector, whereas we are often interested in a scalar function of it, say $\alpha(z) = \mathcal{T}\{\Theta(z)\}$. For example, in the nutrition example motivating this research, $\Theta(z)$ is a $q = 6$ dimensional vector of conditional moments of $\mathbf{Y}$ given $Z = z$, and $\alpha(z)$ is the correlation between a component of $\mathbf{Y}$ and another, unobservable random variable.

Our basic method for estimating $\Theta(\cdot)$ involves local polynomials. With superscript (j) denoting a jth derivative with respect to z and with $\mathbf{b}_j = \Theta^{(j)}(z_0)/j!$, the

Raymond J. Carroll is Distinguished University Professor of Statistics and Professor of Nutrition and Toxicology, Department of Statistics, Texas A&M University, College Station, TX 77843. David Ruppert is Professor, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY 14853. Alan H. Welsh is Reader, CMA, Australian National University, Canberra ACT 2601. Carroll's research was supported by grant CA-57030 from the National Cancer Institute. Carroll's research was partially completed while visiting the Institute for Statistics and Ökonometrie, Humboldt University zu Berlin, with partial support from a senior Alexander von Humboldt Foundation research award. Ruppert's research was supported by grants from the National Science Foundation and the National Security Agency. This research is an outgrowth of discussions with Laurence Freedman and Nancy Potischman of the National Cancer Institute concerning nutritional epidemiology; their suggestions are gratefully acknowledged. The authors also wish to acknowledge with thanks Donna Spiegelman and Walter Willett of the Harvard School of Public Health for making the Nurses' Health Trial calibration study data available, and two referees and an associate editor for making many useful suggestions.

Journal of the American Statistical Association
March 1998, Vol. 93, No. 441, Theory and Methods

local polynomial of order p in a neighborhood of z_0 is $\Theta(z) \approx \sum_{j=0}^{p} \mathbf{b}_j(z-z_0)^j$. The local weight for a value of z near z_0 is denoted by $w(z, z_0)$. We then propose to solve in $(\mathbf{b}_0, \ldots, \mathbf{b}_p)$ the $q \times (p+1)$ equations

$$0 = \sum_{i=1}^{n} w(Z_i, z_0)\psi\left\{\mathbf{Y}_i, \sum_{j=0}^{p} \mathbf{b}_j(Z_i - z_0)^j\right\}\mathbf{G}_p^t(Z_i - z_0), \tag{2}$$

where $\mathbf{G}_p^t(v) = (1, v, v^2, \ldots, v^p)$. The final estimates are $\hat{\Theta}(z_0) = \hat{\mathbf{b}}_0$ and $\hat{\alpha}(z_0) = \mathcal{T}\{\hat{\mathbf{b}}_0\}$.

Equations such as (2) are already in common use when $\Theta(z)$ is scalar, although not at the level of generality given here (not being derived from estimating functions). Here are a few examples:

a. Ordinary multivariate-response Nadaraya–Watson kernel regression has $p = 0, \psi(\mathbf{Y}, \mathbf{v}) = \mathbf{Y} - \mathbf{v}$, and $w(z, z_0)$ chosen to be a kernel weight.
b. Local linear regression has $p = 1$ and $\psi(\mathbf{Y}, \mathbf{v}) = \mathbf{Y} - \mathbf{v}$, and if $w(z, z_0)$ is a nearest-neighbor weight, then the result is the loess procedure in S-PLUS (Chambers and Hastie 1992).
c. When the mean and variance of a univariate response Y are related through $E(Y|Z) = \mu\{\Theta(Z)\}$ and $\text{var}(Y|Z) = \sigma^2 V\{\Theta(Z)\}$ for known functions μ and V, local quasi-likelihood regression is based on

$$\psi(Y, x) = \{Y - \mu(x)\}\mu^{(1)}(x)/V(x). \tag{3}$$

With kernel weights, this is the method of Weisberg and Welsh (1994) when $p = 0$ and of Fan, Heckman, and Wand (1995) when $p \geq 1$.

This article is organized as follows. Section 2 describes in detail a problem from nutrition that motivated this work. This problem is easily analyzed in our general local estimating equation framework. Section 3 indicates that local polynomial methods usually have tuning constants that must be set or estimated. If they are to be estimated, then the typical approach is to minimize mean squared error (MSE) which in turn requires estimation of bias and variance functions. It is possible to derive asymptotic theoretical expressions for these functions (indeed, we do so for kernel regression in the Appendix) and then do a "plug-in" operation to obtain an estimate. But following this approach in practice requires density estimation, estimation of higher-order derivatives, and so on, and these complications would limit the range of applications. Instead, we estimate the bias and variance functions empirically, without explicit use of the asymptotic formulas. Bias estimation uses a modification of Ruppert's (1997) empirical bias method, whereas variance estimation can be done by adapting the sandwich formula of Huber (1967) to this context. That the sandwich formula provides consistent variance estimates in this context is not obvious, but in the Appendix we prove this to be the case.

Section 4 deals with a series of examples, including the analysis of nutrient intake data. Section 5 discusses modifications of the algorithm (2). Section 6 presents some concluding remarks. All theoretical details are collected in an Appendix.

Local estimation of parameters for likelihood problems has been previously considered in important work by such authors as Fan and Gijbels (1996), Fan et al. (1995), Hastie and Tibshirani (1990), Kauermann and Tutz (1997), Severini and Staniswallis (1994), Staniswallis (1989), and Tibshirani and Hastie (1987), and these techniques are implemented in S-PLUS for generalized linear models (GLMs). Our methods and this article differ from the local likelihood literature in several ways:

- We do not require a likelihood, but only an unbiased estimating function. Given the popularity of estimating functions in recent statistical work, such work would appear to be of some consequence. Estimating functions allow us to use such techniques as method of moments, robust mean and variance function estimation, Horvitz and Thompson (1952) adjustments for missing data, GEE-type mean and variance function modeling, and so on. A number of our examples, both numerical and theoretical, illustrate the use of nonlikelihood estimating functions.
- Our estimates of variance are straightforward, being nothing more than estimates based on the sandwich method from parametric problems. In particular, one need not compute asymptotic variances in each problem and then estimate the terms in the resulting (often complex) expressions. To the best of our knowledge the use of the parametric sandwich method in general nonparametric regression contexts has not been previously advocated, nor has it been shown theoretically to give consistent estimates of variances. We prove such consistency and derive expressions for bias and variance for kernel weighting. There have been earlier uses of the sandwich formula in special cases of nonparametric regression, however. For example, Ruppert and Wand (1994) gave a sandwich formula for the variance of local polynomial regression estimators, and Gozalo and Linton (1995) used the sandwich formula for an interesting approach to nonparametric regression—local nonlinear regression.
- Our methods allow for estimation of tuning constants such as the span in loess or local bandwidths in kernel weighting. The methods apply at least in principle to all local estimating function–based estimates and hence can be applied in new problems without the need to use asymptotic theory to derive a bias expression, to use additional nonparametric regressions to estimate this expression, or to develop case-by-case tricks to get started.

2. MOTIVATING EXAMPLE

In this section we demonstrate an important problem where $\Theta(z)$ is a vector and $\psi(\cdot)$ arises from an estimating function framework. The assessment and quantification of an individual's usual diet is a difficult exercise but is fundamental to discovering relationships between diet and cancer and to monitoring dietary behavior among individuals and populations. Various dietary assessment instruments have been devised, of which three main types are most commonly used in contemporary nutritional research. The instrument of choice in large nutritional epidemiology studies is the food frequency questionnaire (FFQ). For proper interpretation of epidemiologic studies that use FFQs as the basic dietary instrument, one needs to know the relationship between reported intakes from the FFQ and true usual intake. Such a relationship is ascertained through a substudy, commonly called a calibration study.

The primary aim of a calibration study may vary from case to case. Here we focus on the estimation of the correlation between FFQ intake and usual intake. The variable we use is the % of calories from fat. This correlation can be of crucial interest if the FFQ has been modified extensively from previous versions or is to be used in a new population from which little previous data have been obtained. Very low correlations might persuade the investigators to postpone the main study, pending improvements in the design of the FFQ or in the way it is presented to study participants.

FFQs are thought to often involve a systematic bias (i.e., underreporting or overreporting at the level of the individual). The other two commonly used instruments are the 24-hour food recall and the multiple-day food record (FR). Each of these FRs is more work-intensive and more costly but is thought to involve considerably less bias than a FFQ. At the end of Section 4.1 we comment on this and other issues in nutrition data.

For the ith individual ($i = 1, \ldots, n$), let Q_i denote the intake of a nutrient reported on a FFQ. For the jth ($j = 1, \ldots, m$) replicate on the ith person, let F_{ij} denote the intake reported by a FR, and let T_i denote long-term usual intake for the ith person. A simple model (Freedman, Carroll, and Wax 1991) relating these three is a standard linear errors-in-variables model,

$$Q_i = \beta_0 + \beta_1 T_i + \varepsilon_i; \tag{4}$$

$$F_{ij} = T_i + U_{ij}; \qquad j = 1, \ldots, m. \tag{5}$$

In model (4) deviations from $\beta_0 = 0$ and $\beta_1 = 1$ represent the systematic bias of FFQs, and the U_{ij} are the within individual variation in FRs. All random errors (i.e., εs and Us) are uncorrelated for purposes of this article; see the end of Section 4.1 for more details and further comments.

In measurement error models one wishes to relate a response (in our case, Q) to a predictor (in our case, T). Because of measurement error and other sources of variability, one cannot observe T. Instead, one can observe only a variable (in our case, F) related to T. The measurement error model literature was recently surveyed by Carroll, Ruppert, and Stefanski (1995).

Two studies that we analyze herein fit exactly into this design. The Nurses' Health Study (Rosner, Willett, and Spiegelman 1989), hereafter denoted by NHS, is a calibration study of 168 women, all of whom completed a single FFQ and four multiple-day food diaries ($m = 4$ in our notation). The Women's Interview Survey of Health (WISH) is a calibration study with 271 participants who completed a FFQ and six 24-hour recalls on randomly selected days at least 2 weeks apart ($m = 6$ in our notation). Although different FFQs are used in the two studies, the major difference between them is that the diaries have considerably smaller within-person variability than the 24-hour recalls. For instance, using % calories from fat, a simple component-of-variance analysis suggests that the measurement error in the *mean* of the four diaries in the NHS has variance 3.43 and the variance of usual intake is $\sigma_t^2 = 14.7$; the numbers for the six 24-hour recalls in WISH are 12.9 and 10.8. Thus one can expect that the NHS data will provide considerably more power for estimating effects than the WISH data.

For an initial analysis, we computed ρ_{QT} for each subpopulation formed by the quintiles of age; Section 4.1 provides the computational details. The five estimated correlations were roughly .4, .6, .4, .5, and .8. The five estimated correlations are statistically significantly different ($p < .01$) using a weighted test for equality of means. Note that the highest quintile of age has the highest value of ρ_{QT}. The standard errors of the estimates are approximately .13, except for the highest quintile, for which it is approximately .07.

Such stratified analysis (i.e., defining the strata by age quintiles) can be considered from the viewpoint of nonparametric regression. In each stratum we are estimating a parameter Θ (often multidimensional) and through it a crucial parametric function such as ρ_{QT}. Because these both depend on the stratum, they are more properly labeled as $\Theta(Z_*)$ and $\rho_{QT}(Z_*)$, where Z_* is the stratum level for Z. *Looked at as a function of* Z, this method suggests that $\rho_{QT}(Z)$ is a *discontinuous* function of Z. To avoid the arbitrariness of the categorization, we propose to estimate $\rho_{QT}(Z)$ as a *smooth* function of Z. Our analysis suggests that at least for the NHS, the correlation between the FFQ and usual intake increases with age in a nonlinear fashion.

3. TUNING CONSTANTS

To implement (2), we need a choice of the weight function $w(z, z_0)$. Usually, this weight function will depend on a tuning constant h, and we will write it as $w(z, z_0, h)$. For example, in global bandwidth local regression, h is the bandwidth and $w(z, z_0, h) = h^{-1}K\{(z - z_0)/h\}$, where $K(\cdot)$ is the kernel (density) function. For nearest-neighbor local regression such as loess (Chambers and Hastie 1992, pp. 312–316), h is the span (the percentage of the data to be counted as neighbors of z_0), and $w(z, z_0, h) = K\{|z - z_0|/a(h)d(z_0)\}$, where $d(z_0)$ is the maximum distance from z_0 to the observations in the neighborhood of z_0 governed by the span and $a(h) = 1$ if $h < 1$ and $a(h) = h$ otherwise.

In practice one has two choices for the tuning constant: (a) fixed a priori or determined randomly as a function of the data, and (b) global (independent of z_0) or local. If the tuning constant is global, then one also has the choice of whether it is the bandwidth or the span; for local tuning constants, there is often no essential difference between using a bandwidth and a span. For example, in loess the span h is typically fixed and global; this makes sense, because the nearest-neighbor weighting of loess imposes locality indirectly. In kernel and local polynomial regression, there is a substantial literature for estimating a global bandwidth h, and some work on estimating local bandwidths.

For the purpose of specificity, here we consider local estimation of the tuning constant. If we could determine the bias and variance functions of $\hat{\alpha}(z_0)$, say $\text{bias}(z_0, h, \alpha)$ and $\text{var}(z_0, h, \alpha)$, then we might reasonably choose $h = h(z_0)$ to minimize the mean squared error (MSE) function $\text{MSE}(z_0, h, \alpha) = \text{var}(z_0, h, \alpha) + \text{bias}^2(z_0, h, \alpha)$. To implement this idea, one needs estimates of the bias and variance functions. An associate editor raised the question of whether one would have enough data to estimate a local bandwidth. The answer is often "strictly speaking, no," but there is a compromise between truly local bandwidths and a global bandwidth. Ruppert (1997) proposed smoothing of the MSE function before minimizing to obtain a local bandwidth and then smoothing the local bandwidth. This type of procedure was called a "partial local smoothing rule" by Hall, Marron, and Titterington (1995). Simulation studies by Ruppert (1997) for the smoothed empirical bias bandwidth selection local bandwidth and by Fan and Gijbels (1995, 1996) for another local bandwidth estimator show that local bandwidths can outperform global bandwidths even for moderately small datasets.

The kernel regression literature abounds with ways of estimating the bias and variance functions, usually based on asymptotic expansions. We digress here briefly to discuss this issue; the Appendix contains details of the algebraic arguments. In our general context, the bias and variance of $\hat{\Theta}(z)$ using kernel regression are qualitatively the same as for ordinary local polynomial regression. There are functions $\mathcal{G}_b\{z, K, \Theta(z), p\}$ and $\mathcal{G}_v\{z, K, \Theta(z), p\}$ with the property that in the interior of the support of Z,

$$\begin{aligned} \text{bias}\{\hat{\Theta}(z)\} &\sim h^{p+1}\mathcal{G}_b\{z, K, \Theta(z), p\} \quad \text{if } p \text{ is odd} \\ &\sim h^{p+2}\mathcal{G}_b\{z, K, \Theta(z), p\} \quad \text{if } p \text{ is even} \end{aligned}$$

and

$$\text{cov}\{\hat{\Theta}(z)\} \sim \{nhf_Z(z)\}^{-1}\mathcal{G}_v\{z, K, \Theta(z), p\}.$$

The function $\mathcal{G}_v$ does not depend on the design density. The same is true of $\mathcal{G}_b$ if p is odd, but not if p is even (see Ruppert and Wand 1994 for the case of local polynomial regression and also (A.4) in the Appendix). The actual formulas are given in the Appendix. Results similar to what is known to happen at the boundary in ordinary local polynomial regression can be derived in our context.

For example, if $p = 1$ and $\psi(\mathbf{y}, \mathbf{v}) = \mathbf{y} - \mathbf{v}$ (ordinary multivariate-response local linear regression), then

$$\mathcal{G}_b\{z, K, \Theta(z), 1\} = (1/2)\Theta^{(2)}(z) \int s^2 K(s)\, ds$$

and

$$\begin{aligned} &\mathcal{G}_v\{z, K, \Theta(z), 1\} \\ &\quad = \left\{\int K^2(s)\, ds\right\} \{\mathbf{B}(z)\}^{-1}\mathbf{C}(z)\{\mathbf{B}^t(z)\}^{-1}, \end{aligned}$$

where

$$\mathbf{B}(z) = E\{(\partial/\partial\mathbf{v})\psi(\mathbf{Y}, \mathbf{v})|Z = z\}$$

and

$$\mathbf{C}(z) = E\{\psi(\mathbf{Y}, \mathbf{v})\psi^t(\mathbf{Y}, \mathbf{v})|Z = z\},$$

with both $\mathbf{B}(z)$ and $\mathbf{C}(z)$ evaluated at $\mathbf{v} = \Theta(z)$. In this specific example, if $\mathbf{I}$ is the identity matrix, then $\mathbf{B}(z) = -\mathbf{I}$ and $\mathbf{C}(z) = \text{cov}(\mathbf{Y}|Z = z)$.

We now return to tuning constant estimation. For local regression, one could in principle use the asymptotic expansions to derive bias and variance formulas for $\hat{\alpha}(z_0)$. This is complicated by the facts that (a) the bias depends on higher-order derivatives of $\Theta(z_0)$, (b) if p is even then the bias depends on the design density, and (c) the variance depends on the density of the Zs. Instead of carrying through this line of argument, we instead propose methods that avoid direct use of asymptotic formulas and that are applicable as well to methods other than local regression. Such a goal has already been achieved in the kernel literature for ordinary local polynomial estimation. [See Fan and Gijbels (1995) and Ruppert (1997), the latter of which we use in our more general context.]

3.1 Empirical Bias Estimation

Ruppert (1997) suggested a method of bias estimation that avoids direct estimation of higher-order derivatives arising in asymptotic bias formulas. He termed this the method empirical bias bandwidth selection (EBBS).

The basic idea is as follows. Fix h_0 and z_0, and use as a model for the bias a function $f(h, \gamma)$ known except for the parameters $\gamma = (\gamma_1, \ldots, \gamma_t)$; for example, $f(h, \gamma) = \gamma_1 h^{p+1} + \cdots + \gamma_t h^{p+t}$, where $t \geq 1$, for local pth degree polynomial kernel regression. The model $f(h, \gamma)$ comes from asymptotic theory, which shows that the asymptotic bias has an expansion in powers of h beginning with power $p+1$, assuming that Θ has at least $p+1$ continuous derivatives. For any h_0, form a neighborhood of tuning constants $\mathcal{H}_0$. On a suitable grid of tuning constants h in $\mathcal{H}_0$, say $\{h_1, \ldots, h_K\}$, where $K \geq t+1$, compute the local polynomial estimator $\hat{\alpha}(z_0, h)$, which should be well described as a function of h by $\hat{\alpha}(z_0, h) = \gamma_0 + f(h, \gamma) + o_P(h^{p+t})$, the value $\gamma_0 = \alpha(z_0)$ in the limit. Then let $(\hat{\gamma}_0, \hat{\gamma})$ minimize $\sum_{k=1}^{K}\{\hat{\alpha}(z_0, h_k) - (\hat{\gamma}_0 + f(h_k, \hat{\gamma}))\}^2$. Appealing to asymptotic theory, and if $\mathcal{H}_0$ is small enough, the bias should be well estimated at h_0 by $f(h_0, \hat{\gamma})$.

In practice, the algorithm is defined as follows. For any fixed z_0, set a range $[h_a, h_b]$ for possible local tuning constants. For example, h_a and h_b could be $d(z_0)$ corresponding to spans of .1 and 1.5. Our experience is that the optimal local bandwidth is generally in this range. Then form a geometrically spaced grid of M points,

$$\mathcal{H}_1 = \{h_j : j = 1, \ldots, M, h_1 = h_a, h_M = h_b\}.$$

We have not tried spacing other than geometric, because it seemed intuitive that smaller bandwidths should be more closely spaced.

Fix constants (J_1, J_2) such that $J_1 + J_2 \geq t$. For any $j = 1 + J_1, \ldots, M - J_2$, apply the procedure defined in the previous paragraph with $h_0 = h_j$ and $\mathcal{H}_0 = \{h_k, k =$

$j - J_1, \ldots, j + J_2\}$. This defines $\widehat{\text{bias}}\{\hat{\alpha}(z_0, h_j)\}$. For tuning constants not on the grid $\mathcal{H}_1$, interpolation via a cubic spline is used.

Note that we must set the limits of interesting tuning constants $[h_a, h_b]$ and the four tuning constants (t, M, J_1, J_2). Ruppert (1997) found that $J_1 = 1, (t, J_2) = (1, 1)$ or $(2, 2)$, and M between 12 and 20 give good numerical behavior in the examples that he studied using local polynomial kernel regression.

3.2 Empirical Variance Estimation: The Sandwich Method

It is useful to remember that q is the dimension of $\mathbf{\Theta}, p$ is the degree of the local polynomial, and $\mathbf{G}_p$ is defined just after (2). At this level of generality, the sandwich formula can be used to derive an estimate of the covariance matrix of $(\hat{\mathbf{b}}_0, \ldots, \hat{\mathbf{b}}_p)$. In parametric problems the solution $\hat{\mathbf{\Theta}}$ to (1) has sandwich (often called "robust") covariance matrix estimate $\mathbf{B}_n^{-1}\mathbf{C}_n(\mathbf{B}_n^t)^{-1}$, where

$$\mathbf{C}_n = \sum_{i=1}^{n} \boldsymbol{\psi}(\mathbf{Y}_i, \hat{\mathbf{\Theta}}) \boldsymbol{\psi}^t(\mathbf{Y}_i, \hat{\mathbf{\Theta}})$$

and

$$\mathbf{B}_n = \sum_{i=1}^{n} (\partial / \partial \mathbf{\Theta}^t) \boldsymbol{\psi}(\mathbf{Y}_i, \hat{\mathbf{\Theta}}).$$

The analogous formulas for the solution to (2) are defined as follows. In what follows, if $\mathbf{A}$ is $l \times q$ and $\mathbf{B}$ is $r \times s$, then $\mathbf{A} \otimes \mathbf{B}$ is the Kronecker product, defined as the $lr \times qs$ matrix formed by multiplying individual elements of $\mathbf{A}$ by $\mathbf{B}$; for example, if $\mathbf{A}$ is a 2×2 matrix, then

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} \end{bmatrix}.$$

Let $\boldsymbol{\chi}(\mathbf{y}, \mathbf{v}) = (\partial / \partial \mathbf{v}^t) \boldsymbol{\psi}(\mathbf{y}, \mathbf{v})$. Then the asymptotic covariance matrix of $(\hat{\mathbf{b}}_0, \ldots \hat{\mathbf{b}}_p)$ is estimated by $\{\mathbf{B}_n(z_0)\}^{-1}$ $\mathbf{C}_n(z_0)\{\mathbf{B}_n^t(z_0)\}^{-1}$, where

$$\mathbf{C}_n(z_0) = \sum_{i=1}^{n} w^2(Z_i, z_0) \times [\{\mathbf{G}_p(Z_i - z_0)\mathbf{G}_p^t(Z_i - z_0)\} \otimes (\hat{\boldsymbol{\psi}}_i \hat{\boldsymbol{\psi}}_i^t)] \quad (6)$$

and

$$\mathbf{B}_n(z_0) = \sum_{i=1}^{n} w(Z_i, z_0) \times [\{\mathbf{G}_p(Z_i - z_0)\mathbf{G}_p^t(Z_i - z_0)\} \otimes \hat{\boldsymbol{\chi}}_i], \quad (7)$$

where $\hat{\boldsymbol{\psi}}_i = \boldsymbol{\psi}\{\mathbf{Y}_i, \sum_{j=0}^{p} \hat{\mathbf{b}}_j (Z_i - z_0)^j\}$ and analogously for $\hat{\boldsymbol{\chi}}_i$. In practice, we replace $\sum_{j=0}^{p} \hat{\mathbf{b}}_j (Z_i - z_0)^j$ by $\hat{\mathbf{\Theta}}(Z_i)$. An argument justifying these formulas is sketched in the Appendix. In practice, we multiply the sandwich covariance matrix estimate by $n / \{n - (p + 1)q\}$, an empirical adjustment for loss of degrees of freedom. In a variety of problems that we have investigated (see, e.g., Simpson, Guth, Zhou, Carroll 1996), this empirical adjustment improves coverage probabilities of sandwich-based confidence intervals, when combined with t percentiles with $n - (p+1)q$ df. There is no theoretical justification for this adjustment, however. In specific problems, bias adjustments for the sandwich estimator may be more or less easy to construct. In the case for GLMs, covariance matrix estimators that automatically adjust for leverage and the like already exist (see Hastie and Tibshirani 1990, sec. 6.8.2, and Kauermann and Tutz 1997).

In some problems the sandwich term $\mathbf{C}_n(z_0)$ can be improved on because the covariance matrix of $\boldsymbol{\psi}(\cdot)$ is known partially or fully. For example, if $\boldsymbol{\psi}(\cdot)$ is given by (3), then $E(\boldsymbol{\psi}\boldsymbol{\psi}^t) = \sigma^2\{\mu^{(1)}\}^2/V$, and one would replace $(\hat{\boldsymbol{\psi}}_i \hat{\boldsymbol{\psi}}^t)$ in (6) by $\hat{\sigma}^2\{\hat{\mu}_i^{(1)}\}^2/\hat{V}_i$. In addition, using score-type arguments, one bases work on $\chi(\cdot) = -\{\mu^{(1)}(\cdot)\}^2/V$ and would replace $\hat{\chi}_i$ in (6) by $-\{\hat{\mu}_i^{(1)}\}^2/\hat{V}_i$. We suggest using such additional information when it is available, because the sandwich estimator can be considerably more variable than model-based alternatives. For example, in simple linear regression, sandwich-based estimates of precision are typically at least three times more variable than the usual precision estimates.

The sandwich method in parametric problems does not work in all circumstances, even asymptotically, the most notable exception being the estimate of the median. In this case, if Y is scalar, then $\psi(\mathbf{Y}, x) = I(Y \leq x) - 1/2$, where I is the indicator function. This choice of $\psi(\cdot)$ has zero derivative, and thus (7) equals 0. Alternatives to the sandwich estimators do exist, however, although their implementation and indeed the theory itself needs further investigation. A sandwich-type method was described by Welsh, Carroll, and Ruppert (1994), who used a type of weighted differencing. Alternatively, one can use the so-called "m out of n" resampling method as defined by Politis and Romano (1994), although application of this latter technique requires that one know the rate of convergence of the nonparametric estimator, this being theoretically $(nh)^{1/2}$ for local linear regression. How to choose the level of subsampling m remains an open question.

4. EXAMPLES

In this section we present three example of local estimating equations. Other examples can be found in an early version of this article, available via anonymous ftp at stat.tamu.edu in the directory/pub/rjcarroll/nonparametric. calibration in the file npcal15.ps.

4.1 Nutrition Calibration: NHS and WISH

We used the NHS and WISH data described in Section 2 to understand whether the correlation between a FFQ and usual intake, ρ_{QT}, depends on age, based on the nutrient % calories from fat. Nutrition data with repeated measurements typically have the feature of time trends in total amounts and sometimes in percentages, so that, for example, one might expect reported caloric intake (energy) to decline over time. To take this into account, we ratio adjusted all measurements so that the mean of each FR equals the first. (For an example of ratio adjustment, see Nusser, Carriquiry, Dodd, and Fuller 1996.)

As described previously, i denotes the individual, Q_i and T_i are the nutrient intakes as reported on the FFQ and usual intake, and F_{ij} is the jth replicated FR for the ith individual. The mean of the replicated FRs is $\bar{F}_i$. The unknown parameters in the problem are conveniently characterized as $\Theta = (\theta_1, \ldots, \theta_6)$, where $\theta_1 = E(Q)$, $\theta_2 = E(F) = E(T)$, $\theta_3 = \text{var}(Q)$, $\theta_4 = \text{cov}(Q, F) = \text{cov}(Q, T)$, $\theta_5 = \text{var}(U)$, and $\theta_6 = \text{var}(T)$. Note that for any two replicates F_{ij} and F_{ik} for $j \neq k$, $\theta_6 = \text{cov}(F_{ij}, F_{ik})$. Letting $\mathbf{Y}_i = (Q_i, F_{i1}, \ldots, F_{im})$ be the observed data ($m = 6$ in WISH, $m = 4$ in NHS), the usual method-of-moments estimating function is

$$\psi(\mathbf{Y}_i, \Theta) = \begin{bmatrix} Q_i \\ \bar{F}_i \\ (Q_i - \theta_1)^2 \\ (Q_i - \theta_1)(\bar{F}_i - \theta_2) \\ (m-1)^{-1}\sum_{j=1}^{m}(F_{ij} - \bar{F}_i)^2 \\ \{m(m-1)\}^{-1}\sum_{j=1}^{m}\sum_{k \neq j}^{m}(F_{ij} - \theta_2)(F_{ik} - \theta_2) \end{bmatrix} - \Theta. \tag{8}$$

Numerically, the solution to (2) is easily obtained. Local estimates of $\theta_1(z)$ and $\theta_2(z)$ use nothing more than direct local regression of Q_i and $\bar{F}_i$ on Z_i and once they are plugged into the third–sixth components of $\psi, \{\theta_3(z), \ldots, \theta_6(z)\}$ can also be computed by local least squares; for example, by regressing $(Q_i - \hat{\theta}_1)^2$ on Z_i to obtain $\hat{\theta}_3$. The main parameter of interest is the correlation between Q and $T, \rho_{QT}(z_0) = \theta_4(z_0)\{\theta_3(z_0)\theta_6(z_0)\}^{-1/2}$.

In this example we used nearest-neighbor weights based on the span, as described at the start of Section 3. As in the S-PLUS implementation of loess, we used the tricubed kernel function, which is proportional to $(1 - |v|^3)^3$ for $|v| \leq 1$ and equals 0 elsewhere. For a fixed value of the span, we assessed standard errors by two means. First, we obtained an estimated covariance matrix for $\hat{\Theta}(z_0)$ using the sandwich formula, and then used the delta method to obtain an estimated variance for $\rho_{QT}(z_0), \beta_1(z_0)$, etc. We based the second standard error estimates on the nonparametric bootstrap, with the pairs $(\mathbf{Y}, Z)$ resampled from the data with replacement; we used 500 bootstrap samples. For a range of spans and for a variety of datasets and nutrient variables, the sandwich delta and the bootstrap standard errors were very nearly the same. This is not unexpected, given that the spans used are fairly large. As a theoretical justification, note that if the span is bounded away from 0, then the estimator $\hat{\Theta}(z)$ converges at parametric rates (although to a biased estimate), and the bootstrap and sandwich covariance matrix estimates are asymptotically estimating the same quantity.

Figure 1 shows the value of $\rho_{QT}(\text{age})$ for the NHS % calories from fat for various spans in the range .6–.9 using local quadratic regression. To understand the age distribution in this study, we have also displayed the 10th, 25th, 50th, 75th, and 90th sample percentiles of age. Although there is some variation between the curves for the different values of the span, the essential feature is consistent—namely, that those under age 50 have significantly (in the practical sense) lower correlations than do those over age 50. The statistical significance of this finding can be assessed in various ways. The simplest is to split the data into two populations on the basis of age groups and simply compute $\hat{\rho}_{QT}$ for each population; the estimates are statistically significantly different at a significance level below .02.

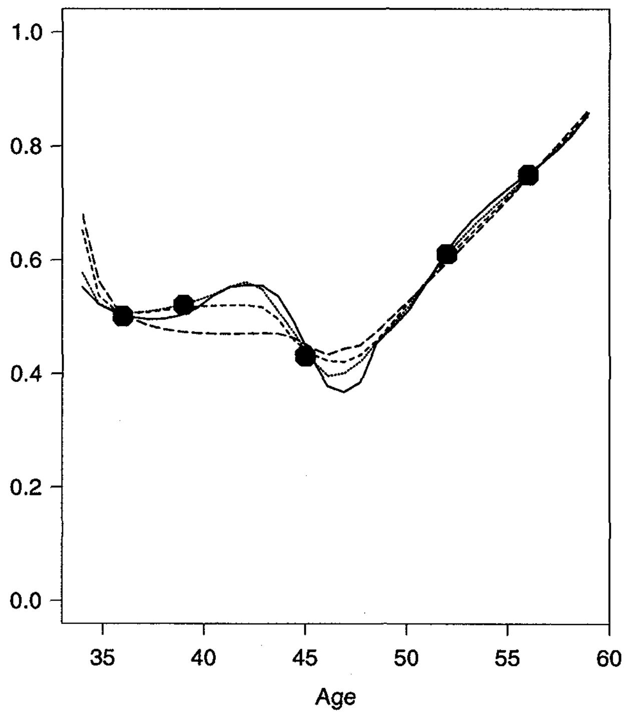

Figure 1. Nurses' Health Study Estimating ρ_{QT} According to Sensitivity to the Choice of Span. Percent calories from fat by using local quadratic regressions with 10th, 25th, 50th, 75th, and 90th percentiles of age. ——, span = .6; · · ·, span = .7; – – –, span = .8; — — —, span = .9.

A second test is slightly more involved. We computed the estimate of $\rho_{QT}(\text{age})$ for 16 equally spaced points on the range from 34–59, along with the bootstrap covariance matrix of these 16 estimates. We then tested whether the estimates were the same using Hotelling's T^2 test and tests for linear and quadratic trend using weighted least squares. As expected after inspection of Figure 1, the linear and quadratic tests had significance levels below .05 for spans in [.7, .9].

We also estimated the span, in the following manner. For computational purposes, we used eight values of age, and using the methods of Section 3 we computed an estimate of the MSE using empirical bias estimation ($J_1 = J_2 = 5, M = 41, h_a = .6, h_b = 1.0$) and the sandwich method; we chose the estimated span to minimize the sum over the eight ages of the estimated MSE. The estimated span was .78 for local linear regression and .90 for local quadratic regression. We then bootstrapped this process, including the estimation of the span, and found that although the significance level was slightly greater than that for a fixed span, it was still below .05.

Because the empirical bias estimate has the tuning constants (M, J_1, J_2), there is still some art to estimating the span. We studied the sensitivity of the estimated span and the estimated average MSE to these tuning constants, and found that the results did not depend too heavily on them as long as J_1 and J_2 were increased with increasing values of M. For example, the estimated average MSEs for local linear regression in three cases—$(M, J_1, J_2) = (41, 5, 5)$, $(101, 13, 13)$, and $(201, 25, 25)$—were calculated, and there was little difference between the three MSEs. However, fixing J_1 and J_2 while increasing M resulted in quite variable bias estimates.

We repeated the estimation process for WISH. There is no evidence of an age effect on ρ_{QT} in WISH. This may be due to the different population or the different FFQ, but may just as well be due to the much larger measurement error in the FRs in WISH than in NHS.

Finally, we investigated local average, linear, quadratic, and cubic regression, with a span of .8; see Figure 2, where we also display the five estimates of ρ_{QT} based on the quintiles of the age distribution. Given the variability in the estimates, the main difference in the methods occurs for higher ages, where the local average regression is noticeably different from the others and from the quintile analysis. Our belief is that this difference arises from the well-known bias of local averages at endpoints.

We redid this analysis using kernel instead of loess weights with locally estimated bandwidths. The results of the two analyses were similar and are not displayed here.

Finally, we comment on issues specific to nutrition:

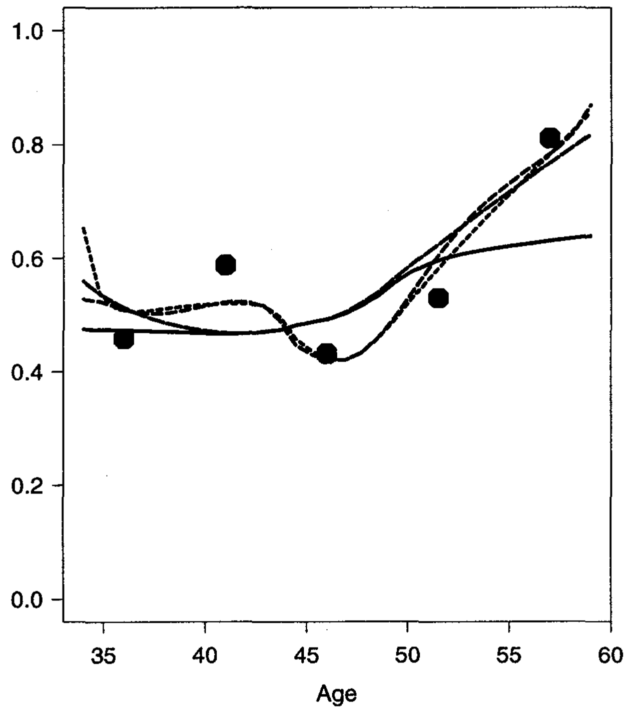

Figure 2. Nurses' Health Study (NHS) Estimating ρ_{QT} With Span .8 and for Local Average, Linear Quadratic, and Cubic Regression With Results From Quintile Analysis. ——, local average; — — —, local linear; - - -, local quadratic; – – –, local cubic.

- We have assumed that the errors ε_i are independent of U_{ij}. This appears to be roughly the case in these two datasets, although it is not true in other datasets that we have studied; for example, the Women's Health Trial data studied by Freedman et al. (1991). The model and the estimating equation are easily modified in general to account for such correlation when it occurs. Similarly, the model and the estimating equation can be modified to take into account a parametric model for correlation among the U_{ij}s; for example, an AR(1) model. Although such correlations exist in these datasets, they are relatively small and should not have a significant impact on the results.
- The method of moments (8) is convenient and easy to compute. In various asymptotic calculations and numerical examples, we have found that it is effectively equivalent to normal-theory maximum likelihood.
- There is emerging evidence from biomarker studies that food records such as those used in NHS are biased for total caloric intake, with those having high body mass index (BMI) underreporting total caloric intake by as much as 20% (see, e.g., Martin, Su, Jones, Lockwood, Tritchler and Boyd 1996). The bias is less crucial for log(total calories) and presumably even less so for the variable used in our analysis, % calories from fat, although no biomarker data exist to verify our conjecture. Despite our belief that this variable is not much subject to large biases explainable by BMI, we have performed various sensitivity analyses that allow for bias. For example, we changed the FFQ and food record data for those with $22 \leq$ BMI ≤ 28 by adding on average 4 to their % calories from fat (a 10% change), whereas for those with BMI > 28 we added on average 7 to their % calories from fat (a 20% increase). The adjustments were proportional to FFQs and food records, and the same adjustment was added to all food records of an individual. These adjustment in effect simulate adjustments to the data that would be made if a strong bias were found in % calories from fat for food records. The analysis of the modified data gave correlations very similar to those shown in our graphs; that is, the effect of bias on the correlation estimates was small.
- If one had replicated FFQs, then many modifications to the basic model could be made. One might conjecture an entirely different error structure; for example,

$$Q_{ij} = \beta_0 + \beta_1 T_i + r_i + \varepsilon_{ij}, \qquad F_{ij} = T_i + s_i + U_{ij},$$

$$\sigma_s^2 = \sigma_r^2.$$

This model is identifiable only if $\operatorname{corr}(r, s)$ is known. We have fit such models using local method of moments to a large $(n > 400)$ dataset with repeated FFQs and using 24-hour recalls for various choices of $\operatorname{corr}(r, s) \leq .5$. The net effect was that such analyses are very different from those based on model (4)–(5); ρ_{QT} increased by a considerable amount, whereas the local estimates of $\operatorname{var}(T)$ as a function of age became much smaller. Of course, the point is that analyses of such complex models are relatively easy using our local estimating function approach.

4.2 Multivariate Z: Lung Cancer Mortality Rates

The methods of this article can be extended to the multivariate $\mathbf{Z}$ case. Suppose that $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{im})^t$, where the Z_{ij} are scalar. Then, following Ruppert and Wand (1994), local linear functions are $\mathbf{\Theta}(\mathbf{z}) = \mathbf{b}_0 + \mathbf{b}_1(\mathbf{z} - \mathbf{z}_0)$, where $\mathbf{b}_0$ is a $p \times 1$ vector and $\mathbf{b}_1$ is a $p \times m$ matrix. The generalization of (2) is to solve

$$0 = \sum_{i=1}^{n} w(\mathbf{Z}_i, \mathbf{z}_0)\psi\{\mathbf{Y}_i, \mathbf{b}_0 + \mathbf{b}_1(\mathbf{Z}_i - \mathbf{z}_0)\}\mathbf{G}_m(\mathbf{Z}_i - \mathbf{z}_0), \tag{9}$$

where $\mathbf{G}_m^t(\mathbf{v}) = (1, \mathbf{v}^t)$. When $\mathbf{Z}$ is multivariate and using kernel weights, the kernel K is multivariate and the bandwidth h is replaced by a positive-definite symmetric matrix $\mathbf{H}$. The simplest choice is to restrict $\mathbf{H}$ to equal $h\mathbf{I}$ for $h > 0$ and with $\mathbf{I}$ the identity matrix, and in this situation the methods we have discussed for empirical bias and variance estimation apply immediately to the estimates $\hat{\mathbf{\Theta}}(\mathbf{z}_0) = \hat{\mathbf{b}}_0$. The application of empirical bias modeling to more general bandwidth matrices is currently under investigation.

Extensions to higher-order local polynomials require more care. Completely nonparametric functional versions are easy in principle, but the notation is complex and practical implementation difficult (see Ruppert and Wand 1994, sec. 4). It is much easier to fit "local additive" models, so that if $\mathbf{z} = (z_1, \ldots, z_m)^t$ and $\mathbf{z}_0 = (z_{01}, \ldots, z_{0m})^t$, then $\mathbf{\Theta}(\mathbf{z}) = \mathbf{b}_0 + \sum_{k=1}^{m}\sum_{j=1}^{p} \mathbf{b}_{kj}(\mathbf{z}_k - \mathbf{z}_{0k})^j$; this is identical to (9) when $p = 1$, and the extension of (9) to $p > 1$ is immediate. We use the term "local additive" to warn the reader that we are not considering a globally additive model in the sense of Hastie and Tibshirani (1990), where $\mathbf{\Theta}(\mathbf{z}) = \mathbf{\Theta}_1(z_1) + \cdots + \mathbf{\Theta}_m(z_m)$ for all $\mathbf{z}$ and some functions $\mathbf{\Theta}_1, \ldots, \mathbf{\Theta}_m$. Globally additive models are outside the scope of this article but would be quite useful when m is larger than 2 or 3 and the "curse of dimensionality" comes into play.

For an example of (9), we consider a problem in which $Y = 10 + \log[(R + .5)/(10^5 - R + .5)]$, where R is the mortality rate per 10^5 males for males dying of lung cancer, as a function of $\mathbf{Z}$ = (age class, year). We call Y the "adjusted" logit because of the .5 offset. The data come from the Australian Institute of Health and are publicly available. The age classes are represented by their midpoints, which are (2, 7, 12, 17, 22, 27, 32, 37, 42, 47, 52, 62, 67, 72, 77, 82, 87), and the years run from 1950–1992 inclusive. For each age class and year subpopulation, we can treat the number of deaths per 10^5 males as being $(d/N) \times 10^5$, where d, the total number of deaths in the subpopulation due to lung cancer, is binomial (N, π) with π the probability of death for an individual and N is the size of the relevant subpopulation. The values of the Ns are known and are used later. Because p is small, d is approximately Poisson(Np) and $\text{var}(R) \approx (10^5/N)E(R)$. In this case, the logit and the log transformation are similar; we use the former to maintain comparability with other work currently being done on these data. We could model the variance of Y as a function of its mean and of N. Alternatively, we could model the variance of Y as a function of $\mathbf{Z}$. We start with the second possibility. If $\mathbf{\Theta} = (\theta_1, \theta_2)^t$, then the estimating function for mean and variance estimation is just $\psi(Y, \mathbf{\Theta}) = \{Y - \theta_1, (Y - \theta_1)^2 - \theta_2\}^t$. There are two good reasons for considering a robust analysis, however. First, there may be concern over the potential for outliers in the response; second, a robust analysis may be numerically more stable. We treat $\tau = \log(\theta_2)$ as the spread parameter (to ensure nonnegativity) and use the estimating equation

$$\psi(Y, \mathbf{\Theta}) = \begin{bmatrix} g\{(Y - \theta_1)/\exp(\tau)\} \\ g^2\{(Y - \theta_1)/\exp(\tau)\} - \int g^2(v)\phi(v)\,dv \end{bmatrix},$$

where $g(v) = g(-v) = v$ if $0 \le v \le c$ and $= c$ if $v > c$, $\phi(v)$ is the standard normal density function and c is a tuning constant controlling the amount of robustness desired; $c = 1.345$ is standard. In the robustness literature, the parameter estimator is known as "proposal 2" (Huber 1981). The spread estimating function can be rewritten as

$$g^2\{\exp(\log|Y - \theta_1| - \tau)\} - \int g^2(v)\phi(v)\,dv,$$

which expresses the spread equation in the form of a location equation. Consideration of the function $g^2\{\exp(x)\} - \int g^2(v)\phi(v)\,dv$ suggests that we simplify the procedure further by replacing it by the much simpler function g with $c = 2$ to increase the efficiency of spread estimation. This is in accordance with the procedure developed by Welsh (1996).

The response and spread surfaces, $\hat{\Theta}_1(z)$ and $\hat{\Theta}_2(z)$, for the lung cancer mortality data are shown in Figures 3a and 3b as surface plots and as contour plots. After some experimentation, the bandwidth matrix was restricted to be of the form h diag(2, 1), and then h was chosen empirically as in Section 3, with a backfitting modification to the basic algorithm (2) described in Section 5. But the results reported here are stable over a range of bandwidth matrices, the main effect of substantial increases in bandwidths being to reduce the ripple and peak in the response and spread surfaces at high ages and early years. Local linear fitting was used in Figures 3a and 3b; local quadratic estimates are similar but with somewhat higher peaks in the spread surface. It is clear that the logit of mortality increases nonlinearly with age class and that there is at best a very weak year effect that shows increased mortality in recent years in the highest age classes. The spread surface shows a ridge of high variability in age classes 20–40 with generally lower variability at both extremes. A delta-method analysis shows that this ridge is due to the logit transformation (with the .5 offset) and the near-Poisson variability of R; see the discussion in the final paragraph of this section. There is also high variability in the highest age classes for the earlier years. This is also the only evidence of a year effect on the variability. The roughness of the spread surface is due mostly to variation in the values of N.

We also modeled the variance of Y as a function of N and the mean of Y. Let N^* be the value of N for a given age class and year divided by the mean of all the N's. Let e^* be the "population size–adjusted residual," defined as the residual for that age class and year times $(N^*)^{1/2}$. Figure 3c plots the absolute values of the e^*s versus the fitted values. Figure 3d plots a local linear fit to the data in 3c. To estimate the spread for a given age class and year, one divides the fitted value from 3d by $(N^*)^{1/2}$. The peak in 3d corresponds to the ridge in 3a.

As mentioned earlier, if we assume that

$$\operatorname{var}(R) = (10^5/N)E(R), \tag{10}$$

then this ridge can be explained by a delta-method calculation showing that

$$\mathrm{SD}(Y) \approx \frac{10^5+1}{(E(R)+.5)(10^5-E(R)+.5)}\,\{10^5E(R)/N\}. \tag{11}$$

We checked (10) by dividing the residuals by the right side of (11), squaring, and then smoothing these squared "standardized residuals" against the fitted values and N. The resulting surface, not included here to save space, was nearly constantly equal to 1, supporting (10).

4.3 Variance Functions and Overdispersion

Problems involving count and assay data are often concerned with overdispersion. For example, if $\mathbf{Y} = (Y, X)$, then the mean of Y might be modeled as $\mu(\mathcal{B}, X)$ and its variance might have the form

$$\operatorname{var}(Y|X) = \exp[\theta_1 + \theta_2 \log\{\mu(\mathcal{B}, X)\}]. \tag{12}$$

Here we assume that the mean function is properly determined so that $\mathcal{B}$ is to be estimated parametrically. If $\theta_2 = 1$ and $\theta_1 > 1$, then we have overdispersion relative to the Poisson model, whereas $\theta_2 \neq 2$ means a departure from the gamma model. In general, we are asking how the variance function depends on the logarithm of the mean. For given $\theta_2, \mathcal{B}$ is usually estimated by generalized least squares (quasi-likelihood). Consistent estimates of $\mathcal{B}$ can be obtained using quasi-likelihood assuming that θ_2 is a fixed value, even if it is not. This well-known fact is often referred to operationally by saying that (12) with fixed θ_2 is a "working" variance model (Diggle et al. 1994).

The problem then is one of variance function estimation, where if $\eta(\mathcal{B}, X) = \log\{\mu(\mathcal{B}, X)\}$, then we believe that the variances are of the form $\exp[\Theta\{\eta(\mathcal{B}, X)\}]$ for some function $\Theta(\cdot)$. Our objective now is to find a suitable estimator of $\boldsymbol{\Theta}(\cdot)$. In a population the variance is $\exp(\Theta)$, which is estimated using the estimating function

$$\psi(\mathbf{Y}, \Theta, \hat{\mathcal{B}}) = \{Y - \mu(\hat{\mathcal{B}}, X)\}^2 \exp(-\Theta) - 1. \tag{13}$$

Estimating Θ as a function of $Z = \eta(\hat{\mathcal{B}}, X)$ is accomplished by using (2) in the obvious manner, namely

$$0 = \sum_{i=1}^{n} w(Z_i, z_0)\psi \times \left\{\mathbf{Y}_i, \sum_{j=0}^{p} b_j(Z_i - z_0)^j, \hat{\mathcal{B}}\right\} G_p^t(Z_i - z_0, \hat{\mathcal{B}}), \tag{14}$$

with ψ given by (13). Because $\hat{\mathcal{B}}$ estimates $\mathcal{B}_0$ at parametric rates, asymptotically there is no effect due to estimating $\mathcal{B}_0$ on the estimate of $\Theta(z)$.

We applied this analysis to three datasets, the esterase assay and hormone assay datasets described by Carroll and Ruppert (1988, chap. 2) and a simulated dataset with $\Theta\{\eta(\mathcal{B}, X)\} = 1.6 + \sin\{\eta(\mathcal{B}, X)\}$, using the same Xs and estimates of $\mathcal{B}$ as in the esterase assay. The model for the mean in all three cases is linear. Previous analyses suggested that the esterase assay data were reasonably well described by a gamma model, with the hormone assay less well described as such because $\theta_2 \approx 1.6$. We used $\theta_2 = 2$ as our working variance model to obtain $\hat{\mathcal{B}}$ for these three datasets. We fit local linear models weighted using loess with the span allowed to take on values between .6 and 2.0 and estimated by the techniques of this article. Figure 4 compares the fitted variance functions divided by the gamma model variance function and rescaled on the horizontal axis to fit on the same plot. Through the range of the data, the deviation from the gamma function is only a factor of about 35% for the esterase assay, indicating a good fit for this model. The hormone assay deviates from the gamma model somewhat more, with variances ranging over factors of two. Both have estimated spans greater than 1.0, indicating that the linear model is a reasonable fit; the hormone

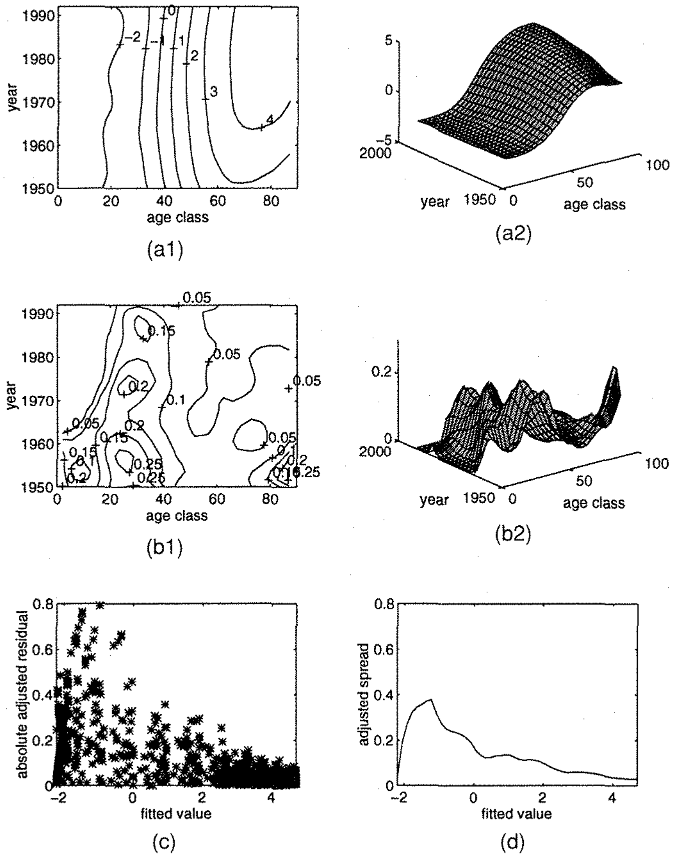

Figure 3. Lung Cancer Mortality Rates; Estimates of the Response Surface and Spread. (a1) and (a2) Adjusted logit of mortality; (b1) and (b2) spread; (c) absolute residuals multiplied by $(N^*)^{1/2}$; (d) local linear fit to the scatterplot in (c).

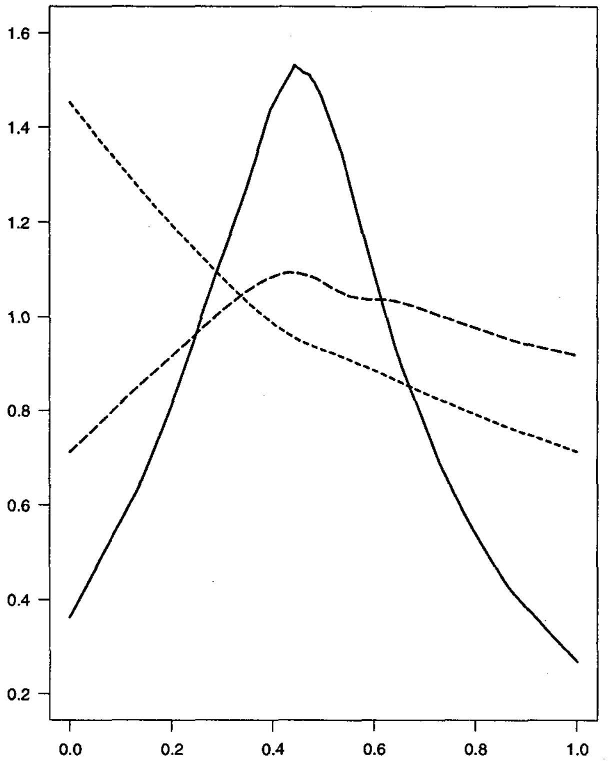

Figure 4. Esterase Assay (– – –), Hormone Assay (- - -), and Simulated (——) Datasets. Estimated discrepancies from the gamma variance model, rescaled. On the vertical axis is the fitted variance function divided by the gamma model variance function. The horizontal axis is $Z = \eta(\mathcal{B}, X) = \log\{\mu(\mathcal{B}, X)\}$.

data simply have a value $\theta_2 < 2$. The simulated data show the sine-type behavior from which they were generated, and a much smaller estimated span (.7).

4.4 Partially Parametric Models

The overdispersion example in Sec. 4.3 contained a parametric part $\mathcal{B}_0$ and a nonparametric part $\Theta(\cdot)$. The "working" estimation method used for the parametric part was chosen so that $\hat{\mathcal{B}}$ was consistent and asymptotically normally distributed with variance of order n^{-1} *even if* $\Theta(\cdot)$ *was completely misspecified.* In other problems, an estimation method for $\mathcal{B}_0$ is chosen whose validity depends on correctly specifying or consistently estimating $\Theta(\cdot)$. An alternative estimator for $\mathcal{B}_0$ given a version $\hat{\Theta}(\cdot)$ is to solve in $\mathcal{B}$ the estimating equation $0 = \sum_{i=1}^{n} \mathbf{\Lambda}\{\mathbf{Y}_i, \mathcal{B}, \hat{\Theta}(\cdot)\}$. The natural approach to use then is to solve the equations

$$0 = \sum_{i=1}^{n} \mathbf{\Lambda}\{\mathbf{Y}_i, \mathcal{B}, \hat{\Theta}(\cdot)\} \tag{15}$$

and

$$0 = \sum_{i=1}^{n} w(Z_i, z_0)\psi\{\mathbf{Y}_i, \mathcal{B}, \Theta(\cdot)\}\mathbf{G}_p^t(Z_i - z_0) \tag{16}$$

for $z_0 = Z_1, \ldots, Z_n$, where $\Theta(z_0) = \sum_{j=0}^{p} \mathbf{b}_j(Z_i - z_0)^j$.

As we have described it, solving (15)–(16) simultaneously is a form of backfitting. One fixes the current estimate of $\mathcal{B}_0$ and obtains an updated estimate of $\Theta(\cdot)$, reverses the process, and then iterates. Asymptotically valid inferences for $\Theta(z)$ are obtained using only (16) and assuming that $\hat{\mathcal{B}}$ is fixed at its estimated value. Asymptotically valid estimates of the covariance matrix of $\hat{\mathcal{B}}$ remain an open problem, although in some cases they can be derived (see Carroll, Fan, Gijbels, and Wand 1997 for single-index models and Severini and Staniswallis 1994 for partial linear models).

The backfitting algorithm has a well-known feature. We confine our remarks to local regression, but these remarks hold for other types of fitting methods as well (Hastie and Tibshirani 1990, pp. 154–155). Specifically, in local linear regression, if the bandwidth is h, then $n^{1/2}(\hat{\mathcal{B}} - \mathcal{B}_0)$ has variance of order 1 but has bias of the order $(nh^4)^{1/2}$, so that getting an asymptotic normal limit distribution with zero bias requires that $nh^4 \to 0$. Unfortunately, "optimal" kernel bandwidth selectors for given $\mathcal{B}$ are typically of the order $h \sim n^{1/5}$, in which case $nh^4 \to \infty$ and the bias in the asymptotic distribution of $\hat{\mathcal{B}}$ does not disappear. If one is even going to worry about this problem (we know of no commercial program that does, nor of any practical examples in which the bias problem is of real concern), then the usual solution is to undersmooth in some way.

Some problems allow for a somewhat more elegant solution to the bias problem, specifically when (15)–(16) are formed as the derivatives of a *single* optimization criterion. None of the estimators that we have described in this article has this form. Optimization of a single criterion basically means a likelihood specification. When this occurs, nonparametric likelihood as described by Severini and Wong (1992) can be applied to make the bias problem disappear, at least in principle, as follows. Let the data likelihood be $l\{\mathcal{B}, \Theta(\cdot)\}$. For fixed $\mathcal{B}$, let $\hat{\Theta}(\cdot, \mathcal{B})$ be the local estimator derived by maximizing the likelihood in Θ with $\mathcal{B}$ fixed. Nonparametric likelihood maximizes $l\{\mathcal{B}, \hat{\Theta}(\cdot, \mathcal{B})\}$ as a function of $\mathcal{B}$. In contrast, backfitting fixes the current $\hat{\Theta}(\cdot, \mathcal{B})$ and updates the estimate of $\mathcal{B}$ by maximizing $l\{\alpha, \hat{\Theta}(\cdot, \mathcal{B})\}$ in α. Nonparametric likelihood can be more difficult to implement than backfitting, especially in our context when $\Theta(\cdot)$ is multivariate. But it is easy to implement if Θ is scalar, $\mathbf{Y} = (Y, X, Z)$, and Y follows a GLM with mean $f(\{\Theta(Z) + X^t\mathcal{B}\}$. (See Severini and Staniswallis 1994 for the ordinary kernel regression case.)

5. MODIFICATIONS OF THE ALGORITHM

The method suggested in (2) requires that all components of $\Theta(z_0)$ be estimated simultaneously. This may be undesirable in some contexts. For example, when estimating a variance function nonparametrically, one often would first estimate the mean function, say $\Theta_1(z)$, form squared residuals $\{\mathbf{Y} - \hat{\Theta}_1(Z_i)\}^2$, and then regress these squared residuals on Z nonparametrically to obtain $\hat{\Theta}_2(z_0)$, the variance estimate at a given z_0. In this context strict application of (2) is different, because it is based on squared pseudoresiduals $\{\mathbf{Y} - \sum_{j=0}^{p} \hat{\Theta}^{(j)}(z_0)(Z_i - z_0)^j/j!\}^2$. In addition, one would often use different tuning constants at each step, but (2) assumes use of the same tuning constant.

The aforementioned example, as well as the nonparametric calibration problem, are examples of a multistage process, where components of $\Theta(\cdot)$ are estimated first and then plugged into the estimating equation for further components. Such problems are easily handled by a slight modification of our approach.

We illustrate the idea in a two-stage context, so that $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)$. By the two-stage process we mean that the first component can be estimated without reference to the second, with weight function w_1 and estimating function $\boldsymbol{\psi}_1$, so that we solve

$$0 = \sum_{i=1}^{n} w_1(Z_i, z_0)\boldsymbol{\psi}_1 \times \left\{ \mathbf{Y}_i, \sum_{j=0}^{p} \mathbf{b}_{j,1}(Z_i - z_0)^j \right\} \mathbf{G}_p^t(Z_i - z_0). \quad (17)$$

The estimate is $\hat{\boldsymbol{\Theta}}_1(z_0) = \hat{\mathbf{b}}_{0,1}(z_0)$.

At the second stage there is a second weight function w_2 and a second estimating function $\boldsymbol{\psi}_2$, and we solve

$$0 = \sum_{i=1}^{n} w_2(Z_i, z_0)\boldsymbol{\psi}_2 \times \left\{ \mathbf{Y}_i, \hat{\boldsymbol{\Theta}}_1(Z_i), \sum_{j=0}^{p} \mathbf{b}_{j,2}(Z_i - z_0)^j \right\} \mathbf{G}_p^t(Z_i - z_0). \quad (18)$$

The estimate is $\hat{\Theta}_2(z_0) = \hat{b}_{0,2}(z_0)$.

The asymptotic covariance matrix of $\{\hat{\Theta}_1(z_0), \hat{\Theta}_2(z_0)\}$ defined by (17)–(18) is estimated by applying the sandwich method to the estimating equation

$$0 = \sum_{i=1}^{n} \left[\begin{array}{l} w_1(Z_i, z_0)\boldsymbol{\psi}_1 \left\{ \mathbf{Y}_i, \sum_{j=0}^{p} \mathbf{b}_{j,1}(Z_i - z_0)^j \right\} \\ w_2(Z_i, z_0)\boldsymbol{\psi}_2 \left\{ \mathbf{Y}_i, \sum_{j=0}^{p} \mathbf{b}_{j,1}(Z_i - z_0)^j, \right. \\ \qquad \left. \sum_{j=0}^{p} \mathbf{b}_{j,2}(Z_i - z_0)^j \right\} \end{array} \right] \times \mathbf{G}_p^t(Z_i - z_0). \quad (19)$$

If $\mathbf{c}_{p \cdot i} = \mathbf{G}_p(Z_i - z_0)\mathbf{G}_p^t(Z_i - z_0)$, the sandwich formulas are

$$B_n(z_0) = \sum_{i=1}^{n} \left\{ \begin{array}{ll} w_1(Z_i, z_0)\mathbf{c}_{p.i} \otimes \hat{\boldsymbol{\chi}}_{i11} & 0 \\ w_2(Z_i, z_0)\mathbf{c}_{p.i} \otimes \hat{\boldsymbol{\chi}}_{i21} & w_2(Z_i, z_0)\mathbf{c}_{p \cdot i} \otimes \hat{\boldsymbol{\chi}}_{i21} \end{array} \right\}$$

and

$$C_n(z_0) = \sum_{i=1}^{n} \left\{ \begin{array}{c} w_1^2(Z_i, z_0)\mathbf{c}_{p \cdot i} \otimes \hat{\boldsymbol{\psi}}_{i1}\hat{\boldsymbol{\psi}}_{i1}^t \\ w_1(Z_i, z_0)w_2(Z_i, z_0)\mathbf{c}_{p \cdot i} \otimes \hat{\boldsymbol{\psi}}_{i1}\hat{\boldsymbol{\psi}}_{i2}^t \\ w_1(Z_i, z_0)w_2(Z_i, z_0)\mathbf{c}_{p \cdot i} \otimes \hat{\boldsymbol{\psi}}_{i2}\hat{\boldsymbol{\psi}}_{i1}^t \\ w_2^2(Z_i, z_0)\mathbf{c}_{p \cdot i} \otimes \hat{\boldsymbol{\psi}}_{i2}\hat{\boldsymbol{\psi}}_{i2} \end{array} \right\},$$

where $\boldsymbol{\chi}_i$ is made up of the elements $\boldsymbol{\chi}_{ijk}$ for $j, k = 1, 2$. In practice, one might replace $\sum_{j=0}^{p} \hat{\mathbf{b}}_{j,k}(Z_i - z_0)^j$ by $\hat{\boldsymbol{\Theta}}_k(Z_i)$.

Tuning constant estimation in multistage problems also may need adjustment. For example, using kernels with bandwidth h_k at stage k, for odd-powered polynomials the bias at stage 1 is of course of the order h_1^{p+1}, whereas at stage 2 it is $c_1(z_0)h_1^{p+1} + c_2(z_0)h_2^{p+1}$. Standard EBBS can be used to estimate h_1 at stage 1, whereas in general estimating h_2 requires a two-dimensional EBBS. But in both the variance function problem and nonparametric calibration, the effect on $\boldsymbol{\Theta}_2$ due to estimating $\boldsymbol{\Theta}_1$ is nil asymptotically, and standard EBBS can be used at each stage without modification.

In general problems, via backfitting one can use different weight functions and tuning constants to estimate each component of $\boldsymbol{\Theta}(z)$. For example, one might iterate between solving the two equations (with estimated tuning constants)

$$0 = \sum_{i=1}^{n} w_1(Z_i, z_0) \times \boldsymbol{\psi}_1 \left\{ \mathbf{Y}_i, \sum_{j=0}^{p} \mathbf{b}_{j,1}(Z_i - z_0)^j, \hat{\boldsymbol{\Theta}}_2(Z_i) \right\} \mathbf{G}_p^t(Z_i - z_0)$$

and

$$0 = \sum_{i=1}^{n} w_2(Z_i, z_0) \times \boldsymbol{\psi}_2 \left\{ \mathbf{Y}_i, \hat{\boldsymbol{\Theta}}_1(Z_i), \sum_{j=0}^{p} \mathbf{b}_{j,2}(Z_i - z_0)^j \right\} \mathbf{G}_p^t(Z_i - z_0).$$

This is the procedure that we used in the lung cancer mortality example.

We conjecture that the asymptotic variance of these backfitted estimates can be estimated consistently by applying the sandwich formula to the equations

$$0 = \sum_{i=1}^{n} w_1(Z_i, z_0) \times \boldsymbol{\psi}_1 \left\{ \mathbf{Y}_i, \sum_{j=0}^{p} \mathbf{b}_{j,1}(Z_i - z_0)^j, \sum_{j=0}^{p} \mathbf{b}_{j,2}(Z_i - z_0)^j \right\} \times \mathbf{G}_p^t(Z_i - z_0)$$

and

$$0 = \sum_{i=1}^{n} w_2(Z_i, z_0) \times \boldsymbol{\psi}_2 \left\{ \mathbf{Y}_i, \sum_{j=0}^{p} \mathbf{b}_{j,1}(Z_i - z_0)^j, \sum_{j=0}^{p} \mathbf{b}_{j,2}(Z_i - z_0)^j \right\} \times \mathbf{G}_p^t(Z_i - z_0).$$

This idea can be shown to work in the case of robust estimation of a mean and variance function, as in the lung cancer mortality example.

6. DISCUSSION

We have extended estimating equation theory to cases where the parameter vector $\boldsymbol{\Theta}$ is not constant but rather depends on a covariate Z. The basic idea is to solve the estimating equation locally at each value of z using weights that for the ith case decrease with the distance between z and the observed Z_i. The weights depend on a tuning parameter; for example, a bandwidth h. A suitable value of h can be found by minimizing an estimate of the MSE. The latter if found by estimating variance using the "sandwich formula" (or more efficient modifications described earlier) and estimating bias empirically (as in Ruppert 1997).

We have applied this methodology to nonparametric calibration in nutritional studies, robust modeling of lung cancer mortality rates, and overdispersion. We have focused on local weighted polynomials. Regression splines could also be used in this context and appear to have considerable promise. Given a set of knots $(\xi_1, \ldots, \xi_p)$, a regression cubic spline has the form

$$\begin{aligned}\Theta(z, \mathbf{b}_0, \ldots, \mathbf{b}_{p+3})\\ &= \mathbf{b}_0 + \mathbf{b}_1 z + \mathbf{b}_2 z^2 + \mathbf{b}_3 z^3 + \sum_{j=1}^{p} \mathbf{b}_{j+3}(z - \xi_j)_+^3,\end{aligned}$$

where $v_+ = v$ if $v > 0$ and equals 0 otherwise. If regression splines are used, then (2) becomes

$$0 = \sum_{i=1}^{n} \psi\{\mathbf{Y}_i, \Theta(Z_i, \mathbf{b}_0, \ldots, \mathbf{b}_{p+3})\}\mathbf{G}_{p,s}(Z_i),$$

where $\mathbf{G}_{p,s}^t(z) = (1, z, z^2, z^3, (z-\xi_1)_+^3, \ldots, (z-\xi_p)_+^3)$. The interesting issue here is the selection of the knots, a problem of considerable interest in the broad context and one on which we are currently working for estimating functions. The regression splines outlined earlier may have an advantage, because the knots can be chosen on a componentwise basis. An alternative to knot selection would be to penalize the knot coefficients, as Eilers and Marx (1996) and Ruppert and Carroll (1997) have suggested for nonparametric regression.

The associate editor has noted that the estimating equation (2) is implicitly adapting to the component of $\Theta(z)$ that has the least amount of smoothness. In principle, one could allow different bandwidths for each component, or even different orders of the local polynomial, and the sandwich variance estimator would still apply. Also, in principle, the EBBS methodology can be used to estimate many different bandwidths. It is not at all clear to us, however, how to decide which components of $\Theta(z)$ are more or less smooth.

Finally, a referee has noted that local polynomial methods need not be range preserving. For example, consider the case where the Ys are all positive and thus the regression function of Y on Z is necessarily positive. Even in ordinary nonparametric local linear kernel estimation, the fitted regression function need not be positive, whereas for local averages the fitted function will be positive. In many cases, appropriate reformulation of the model will preserve ranges. For example, consider binary regression. If one runs an ordinary local linear regression of Y on Z ignoring the binary nature of Y, then it may happen that fitted probabilities do not fall in the unit interval. But if the binary regression is based on the likelihood score (3) where μ is the logistic function (i.e., local logistic regression as in Fan et al. 1995), then the fitted probabilities will necessarily fall in the unit interval. Similarly, for positive Ys, one could use the local model where the log of the mean function is a polynomial.

APPENDIX: ASYMPTOTICS

A.1 Bias and Variance for Local Polynomial Estimation

Here we give a brief derivation of bias and variance formulas for local polynomial estimation of order p in the interior of the support of Z. The methods use to derive the calculations roughly parallel those of Fan et al. (1995) and Ruppert and Wand (1994). The regularity conditions necessary include the smoothness conditions on $\Theta(z)$ and $f_Z(\cdot)$ of Fan et al. (1995), with the smoothness conditions on $\psi(\cdot)$ guaranteeing that it is at least twice continuously differentiable and the regularity conditions from estimating function theory assuring that a consistent sequence of solutions to (2) exists. A useful simplification is to let the unknown parameters be $\mathbf{a}_j = h^j \Theta^{(j)}(z_0)/j!$ (see the appendix of Fan et al. 1995).

For any $p \times q$ matrix $\mathbf{C} = (\mathbf{c}_1, \ldots, \mathbf{c}_l)^t$, where $\mathbf{c}_j$ is a $q \times 1$ vector, define $\mathrm{vec}(\mathbf{C}) = (\mathbf{c}_1^t, \ldots, \mathbf{c}_l^t)^t$. Define $\mu_K(r) = \int z^r K(z)\,dz$ and $\gamma_K(r) = \int z^r K^2(z)\,dz$. Assume that K is symmetric about 0, so that $\mu_K(r) = \gamma_K(r) = 0$ if r is odd.

Let $\mathbf{C}(z_0) = E[\psi\{\mathbf{Y}, \Theta(z_0)\}\psi^t\{\mathbf{Y}, \Theta(z_0)\}|Z = z_0]$ and $\mathbf{B}(z_0) = E[\chi\{\mathbf{Y}, \Theta(z_0)\}|Z = z_0]$, where $\chi(\mathbf{Y}, \mathbf{v}) = (\partial/\partial\mathbf{v}^t)$ $\psi(\mathbf{Y}, \mathbf{v})$. Define

$$\begin{aligned}\mathcal{L}_n(\mathbf{a}_0, \ldots, \mathbf{a}_p)\\ &= n^{-1}\sum_{i=1}^{n} K_h(Z_i - z_0)\\ &\quad \times \mathrm{vec}\left[\mathbf{G}_{p,h}(Z_i - z_0) \otimes \psi^t\left\{\mathbf{Y}_i, \sum_{j=0}^{p} \mathbf{a}_j (Z_i - z_0)^j/h^j\right\}\right],\end{aligned}$$

where $\mathbf{G}_{p,h}(v) = (1, v/h, v^2/h^2, \ldots, v^p/h^p)^t$. We are solving $0 = \mathcal{L}_n(\hat{\mathbf{a}}_0, \ldots, \hat{\mathbf{a}}_p)$, with $\hat{\mathbf{a}}_j = h^j \hat{\Theta}^{(j)}(z_0)/j!$ By a Taylor series expansion, we find that the estimates are asymptotically equivalent to

$$\begin{pmatrix} \hat{\mathbf{a}}_0 - \mathbf{a}_0 \\ \vdots \\ \hat{\mathbf{a}}_p - \mathbf{a}_p \end{pmatrix} \approx -\{\mathbf{B}_*(z_0)\}^{-1}\mathcal{L}_n(\mathbf{a}_0, \ldots, \mathbf{a}_p), \qquad \text{(A.1)}$$

where

$$\mathbf{B}_*(z_0) = \frac{\partial}{\partial(\mathbf{a}_0^t, \ldots, \mathbf{a}_p^t)}\,\mathcal{L}_n(\mathbf{a}_0, \ldots, \mathbf{a}_p).$$

It is helpful to keep in mind the following aspect:

$$\begin{aligned}\mathcal{L}_n(\mathbf{a}_0, \ldots, \mathbf{a}_p)\\ &= n^{-1}\sum_{i=1}^{n} K_h(Z_i - z_0)\\ &\quad \times \begin{bmatrix} \psi\left\{\mathbf{Y}_i, \sum_{j=0}^{p} \mathbf{a}_j (Z_i - z_0)^j/h^j\right\} \\ \{(Z_i - z_0)/h\}\psi\left\{\mathbf{Y}_i, \sum_{j=0}^{p} \mathbf{a}_j (Z_i - z_0)^j/h^j\right\} \\ \vdots \\ \{(Z_i - z_0)/h\}^p\psi\left\{\mathbf{Y}_i, \sum_{j=0}^{p} \mathbf{a}_j (Z_i - z_0)^j/h^j\right\} \end{bmatrix}.\end{aligned}$$

(A.2)

Also note that the **a**s are vectors, of the same length as Θ and ψ. The calculations are easier to follow if this expanded form is used.

It is easily seen that

$$\mathbf{B}_*(z_0) \xrightarrow{p} f_Z(z_0)\{\mathbf{D}_p(\mu) \otimes \mathbf{B}(z_0)\}, \qquad \text{(A.3)}$$

where $\mathbf{D}_p(\mu)$ is the $(p+1) \times (p+1)$ matrix with (j,k)th element $\mu_K(j+k-2)$. It is also easily shown that

$$\operatorname{cov}\{\mathcal{L}_n(\mathbf{a}_0,\dots \mathbf{a}_p)\} \sim (nh)^{-1} f_Z(z_0)\{\mathbf{D}_p(\gamma) \otimes \mathbf{C}(z_0)\},$$

where $\mathbf{D}_p(\gamma)$ is the $(p+1) \times (p+1)$ matrix with (j,k)th element $\gamma_K(j+k-2)$.

Finally, note that because $E[\psi\{\mathbf{Y}, \boldsymbol{\Theta}(Z)\}|Z] = 0$,

$$\begin{aligned} &E\mathcal{L}_n(\mathbf{a}_0,\dots,\mathbf{a}_p) \\ &\quad = -\int K_h(z-z_0) f_{Y|Z}(\mathbf{y}|z) f_Z(z) \\ &\qquad \times \operatorname{vec}\left(\mathbf{G}_{p,h}(z-z_0) \otimes \left[\psi\{\mathbf{y}, \boldsymbol{\Theta}(z)\}\right.\right. \\ &\qquad\qquad \left.\left. - \psi\left\{\mathbf{y}, \sum_{j=0}^{p} \mathbf{a}_j (z-z_0)^j/h^j\right\}\right]^t\right) d\mathbf{y}\, dz \\ &\quad \approx -\int K_h(z-z_0) f_{Y|Z}(\mathbf{y}|z) f_Z(z) \\ &\qquad \times \operatorname{vec}\left(\mathbf{G}_{p,h}(z-z_0) \otimes \left[\chi\{\mathbf{y}, \boldsymbol{\Theta}(z)\}\right.\right. \\ &\qquad\qquad \left.\left. \times \left\{\boldsymbol{\Theta}(z) - \sum_{j=0}^{p} (z-z_0)^j \boldsymbol{\Theta}^{(j)}(z_0)/j!\right\}\right]^t\right) d\mathbf{y}\, dz. \end{aligned}$$

But $\boldsymbol{\Theta}(z) - \sum_{j=0}^{p} (z-z_0)^j \boldsymbol{\Theta}^{(j)}(z_0)/j! = (z-z_0)^{p+1}\boldsymbol{\Theta}^{(p+1)}(z_0)/(p+1)! + (z-z_0)^{p+2}\boldsymbol{\Theta}^{(p+2)}(z_0)/(p+2)! + \mathcal{O}\{(z-z_0)^{p+3}\}$. Hence, to terms of order $\{1 + \mathcal{O}(h)\}$,

$$E\mathcal{L}_n(\mathbf{a}_0,\dots,\mathbf{a}_p) \approx \mathbf{A}_{1h} + \mathbf{A}_{2h},$$

where

$$\begin{aligned} &\mathbf{A}_{kh} \\ &\quad = -\frac{h^{p+k}}{(p+k)!} \int K(x) f_{Y|Z}(\mathbf{y}|z_0+xh) f_Z(z_0+xh) \\ &\qquad \times \operatorname{vec}(\mathbf{G}_{p,1}(x) \otimes [\chi\{\mathbf{y}, \boldsymbol{\Theta}(z_0+xh)\}\boldsymbol{\Theta}^{(p+k)}(z_0) x^{p+k}]^t)\, d\mathbf{y}\, dx \\ &\quad = -\frac{h^{p+k}}{(p+k)!} \int K(x) f_Z(z_0+xh) \\ &\qquad \times \operatorname{vec}[\mathbf{G}_{p,1}(x) \otimes \{B(z_0+hx)\boldsymbol{\Theta}^{(p+k)}(z_0) x^{p+k}\}^t]\, dx. \end{aligned}$$

Clearly,

$$\mathbf{A}_{2h} \approx \frac{-h^{p+2} f_Z(z_0)}{(p+2)!} \operatorname{vec}[\mathbf{D}_\mu(p+2) \otimes \{\mathbf{B}(z_0)\boldsymbol{\Theta}^{(p+2)}(z_0)\}^t],$$

where $\mathbf{D}_\mu(L) = \{\mu_K(L), \mu_K(L+1), \dots, \mu_K(L+p)\}^t$. If we define $\mathbf{Q}(z) = f_Z(z)\mathbf{B}(z)$ with first derivative $\mathbf{Q}^{(1)}(z)$, then it also follows that

$$\begin{aligned} &\mathbf{A}_{1h} \\ &\quad \approx \frac{-h^{p+1}}{(p+1)!} \int K(x) \operatorname{vec}[x^{p+1}\mathbf{G}_{p,1}(x) \\ &\qquad\qquad \otimes \{\mathbf{Q}(z_0+hx)\boldsymbol{\Theta}^{(p+1)}(z_0)\}^t]\, dx \\ &\quad \approx -\frac{h^{p+1} f_Z(z_0)}{(p+1)!} \operatorname{vec}[\mathbf{D}_\mu(p+1) \otimes \{\mathbf{B}(z_0)\boldsymbol{\Theta}^{(p+1)}(z_0)\}^t] \\ &\qquad - \frac{h^{p+2}}{(p+1)!} \int K(x) \operatorname{vec}[x^{p+2}\mathbf{G}_{p,1}(x) \\ &\qquad\qquad \otimes \{\mathbf{Q}^{(1)}(z_0)\boldsymbol{\Theta}^{(p+1)}(z_0)\}^t]\, dx \\ &\quad \approx -\frac{h^{p+1}}{(p+1)!} \operatorname{vec}[\mathbf{D}_\mu(p+1) \otimes \{f_Z(z_0)\mathbf{B}(z_0)\boldsymbol{\Theta}^{(p+1)}(z_0)\}^t] \\ &\qquad - \frac{h^{p+2}}{(p+1)!} \operatorname{vec}[\mathbf{D}_\mu(p+2) \otimes \{\mathbf{Q}^{(1)}(z_0)\boldsymbol{\Theta}^{(p+1)}(z_0)\}^t]. \end{aligned}$$

Thus we have shown that asymptotically,

$$\begin{aligned} &\operatorname{bias}(\hat{\mathbf{a}}_0^t, \hat{\mathbf{a}}_1^t, \dots, \hat{\mathbf{a}}_0^t)^t \\ &\quad = h^{p+1}\{\mathbf{D}_p(\mu) \otimes \mathbf{B}(z_0)\}^{-1} \\ &\qquad \times \operatorname{vec}[\mathbf{D}_\mu(p+1) \otimes \{\mathbf{B}(z_0)\boldsymbol{\Theta}^{(p+1)}(z_0)\}^t]/(p+1)! \\ &\qquad + h^{p+2}\{\mathbf{D}_p(\mu) \otimes \mathbf{B}(z_0)\}^{-1}\mathbf{s}(z_0) + \mathcal{O}(h^{p+3}), \end{aligned} \qquad \text{(A.4)}$$

where

$$\begin{aligned} &\mathbf{s}(z_0) \\ &\quad = \operatorname{vec}\left[\mathbf{D}_\mu(p+2)\right. \\ &\qquad\qquad \left. \otimes \left\{\frac{\mathbf{B}(z_0)\boldsymbol{\Theta}^{(p+2)}(z_0)}{(p+2)!} + \frac{\mathbf{Q}^{(1)}(z_0)\boldsymbol{\Theta}^{(p+1)}(z_0)}{f_Z(z_0)(p+1)!}\right\}^t\right]. \end{aligned}$$

The variance is

$$\begin{aligned} &\{nhf_Z(z_0)\}^{-1}\{\mathbf{D}_\mu(p) \otimes \mathbf{B}(z_0)\}^{-1} \\ &\qquad \times \{\mathbf{D}_\gamma(p) \otimes \mathbf{C}(z_0)\}\{\mathbf{D}_\mu(p) \otimes \mathbf{B}(z_0)\}^{-t}\{1 + o(1)\}. \end{aligned}$$

The only thing left to show is that if p is even, then the bias is of order $\mathcal{O}(h^{p+2})$—that is, the first element in

$$\{\mathbf{D}_\mu(p) \otimes \mathbf{B}(z_0)\}^{-1}\operatorname{vec}[\mathbf{D}_\mu(p+1) \otimes \{\mathbf{B}(z_0)\boldsymbol{\Theta}^{(p+1)}(z_0)\}^t]$$

equals 0, which is clearly the case because $\mu_K(r) = 0$ if r is odd. For z_0 on the boundary of the support of Z, the terms of order h^{p+1} dominate, and the bias is of that order.

It is useful for theoretical purposes to note that we have actually shown the following. Denote (A.2) by $\mathcal{F}(z_0)$, (A.3) by $\mathcal{T}(z_0)$ and (A.4) by $\mathcal{S}(z_0)$. Then we have shown that

$$\begin{pmatrix} \hat{\mathbf{a}}_0 - \mathbf{a}_0 \\ \vdots \\ \hat{\mathbf{a}}_p - \hat{\mathbf{a}}_p \end{pmatrix} \approx \mathcal{S}(z_0) - \{\mathcal{T}(z_0)\}^{-1}\mathcal{F}(z_0). \qquad \text{(A.5)}$$

Remark. In the application of parametric estimating equations, unless the equations are linear in the parameter there is typically a bias of order n^{-1}, which, however, is negligible compared to the standard deviation. Similarly, there will be a bias of order $(nh)^{-1}$ here that stems from terms ignored in the linearizing approximation (A.1). Because h is chosen so that the *squared* bias from smoothing is of order $(nh)^{-1}$, bias terms of order $(nh)^{-1}$ are ignored here. (However, see Ruppert, Wand, Holst, and Hössjer 1995 for a method of correcting the order $(nh)^{-1}$ bias due to estimation of the mean when a variance function is estimated.)

A.2 The Sandwich Formula

Here we sketch a justification for the sandwich formula (6)–(7), using the notation established previously in this Appendix. We continue to work with the parameterization $(\mathbf{a}_0, \ldots, \mathbf{a}_p)$. Noting that $\mathbf{B}_*(z_0)$ in (A.1) equals $n^{-1}\mathbf{B}_n(z_0)$ in (7), it suffices to show that $n^{-1}\mathbf{C}_n(z_0)$ defined in (6) has limiting covariance matrix $(nh)^{-1}f_Z(z_0)\{\mathbf{D}_p(\gamma) \otimes \mathbf{C}(z_0)\}$, which is easily established. This completes the argument.

[Received July 1996. Revised August 1997.]

REFERENCES

Breslow, N. E., and Clayton, D. G. (1993), "Approximate Inference in Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 88, 9–36.

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1997), "Generalized Partially Linear Single-Index Models," *Journal of the American Statistical Association*, 92, 477–489.

Carroll, R. J., and Ruppert, D. (1988), *Transformation and Weighting in Regression*, London: Chapman and Hall.

Carroll, R. J., Ruppert, D., and Stefanski, L. A. (1995), *Measurement Error in Nonlinear Models*, London: Chapman and Hall.

Chambers, J. M., and Hastie, T. J. (1992), *Statistical Models in S*, Pacific Grove, CA: Wadsworth.

Diggle, P. J., Liang, K. Y., and Zeger, S. (1994), *Analysis of Longitudinal Data*, Oxford, U.K.: Oxford University Press.

Eilers, P. H. C., and Marx, B. D. (1996), "Flexible Smoothing With B-Splines and Penalties" (with discussion), *Statistical Science*, 11, 89–121.

Fan, J., and Gijbels, I. (1995), "Data-Driven Bandwidth Selection in Local Polynomial Fitting: Variable Bandwidth and Spatial Adaptation," *Journal of the Royal Statistical Society*, Ser. B, 57, 371–394.

——— (1996), *Local Polynomial Modelling and Its Applications*, London: Chapman and Hall.

Fan, J., Heckman, N. E., and Wand, M. P. (1995), "Local Polynomial Kernel Regression for Generalized Linear Models and Quasi-Likelihood Functions," *Journal of the American Statistical Association*, 90, 141–150.

Freedman, L. S., Carroll, R. J., and Wax, Y. (1991), "Estimating the Relationship Between Dietary Intake Obtained From a Food Frequency Questionnaire and True Average Intake," *American Journal of Epidemiology*, 134, 510–520.

Gozalo, P., and Linton, O. (1995), "Using Parametric Information in Nonparametric Regression," Working Paper 95-40, Brown University, Dept. of Economics.

Hall, P., Marron, J. S., and Titterington, D. M. (1995), "On Partial Local Smoothing Rules for Curve Estimation," *Biometrika*, 82, 575–588.

Hastie, T. J., and Tibshirani, R. (1990), *Generalized Additive Models*, London: Chapman and Hall.

Horvitz, D. G., and Thompson, D. J. (1952), "A Generalization of Sampling Without Replacement From a Finite Universe," *Journal of the American Statistical Association*, 47, 663–685.

Huber, P. J. (1967), "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions," *Proceedings of the 5th Berkeley Symposium*, 1, 221–233.

Huber, P. J. (1981), *Robust Statistics*, New York: Wiley.

Kauermann, G., and Tutz, G. (1997), "On Model Diagnostics and Bootstrapping in Varying Coefficient Models," unpublished manuscript.

Martin, L. J., Su, W., Jones, P. J., Lockwood, G. A., Tritchler, D. L., and Boyd, N. F. (1996), "Comparison of Energy Intakes Determined by Food Records and Doubly Labeled Water in Women Participating in a Dietary Intervention Trial," *American Journal of Clinical Nutrition*, 63, 483–490.

Nusser, S. M., Carriquiry, A. L., Dodd, K. W., and Fuller, W. A. (1996), "A Semiparametric Transformation Approach to Estimating Usual Intake Distributions," *Journal of the American Statistical Association*, 91, 1440–1449.

Politis, D. N., and Romano, J. P. (1994), "Large Sample Confidence Regions Based on Subsamples Under Minimal Conditions," *The Annals of Statistics*, 22, 2031–2050.

Rosner, B., Willett, W. C., and Spiegelman, D. (1989), "Correction of Logistic Regression Relative Risk Estimates and Confidence Intervals for Systematic Within-Person Measurement Error," *Statistics in Medicine*, 8, 1051–1070.

Ruppert, D. (1997), "Empirical-Bias Bandwidths for Local Polynomial Nonparametric Regression and Density Estimation," submitted to *Journal of the American Statistical Association*, 92, 1049–1062.

Ruppert, D., and Carroll, R. J. (1997), "Penalized Regression Splines," manuscript. Available at http://www.orie.cornell.edu/davidr/papers/index.html.

Ruppert, D., and Wand, M. P. (1994), "Multivariate Locally Weighted Least Squares Regression," *The Annals of Statistics*, 22, 1346–1370.

Ruppert, D., Wand, M. P., Holst, U., and Hössjer, O. (1997), "Local Polynomial Variance Function Estimation," *Technometrics*, 39, 262–273.

Severini, T. A., and Staniswallis, 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, 1768–1802.

Simpson, D. G., Guth, D., Zhou, H., and Carroll, R. J. (1996), "Interval Censoring and Marginal Analysis in Ordinal Regression," *Journal of Agricultural, Biological and Environmental Statistics*, 1, 354–376.

Staniswallis, J. G. (1989), "The Kernel Estimates of a Regression Function in Likelihood-Based Models," *Journal of the American Statistical Association*, 84, 276–283.

Tibshirani, R., and Hastie, T. (1987), "Local Likelihood Estimation," *Journal of the American Statistical Association*, 82, 559–567.

Weisberg, S., and Welsh, A. H. (1994), "Estimating the Missing Link Function," *The Annals of Statistics*, 22, 1674–1700.

Welsh, A. H. (1996), "Robust Estimation of Smooth Regression and Spread Functions and Their Derivatives," *Statistica Sinica*, 6, 347–366.

Welsh, A. H., Carroll, R. J., and Ruppert, D. (1994), "Fitting Heteroscedastic Regression Models," *Journal of the American Statistical Association*, 89, 100–116.