NONPARAMETRIC ESTIMATION VIA LOCAL ESTIMATING EQUATIONS, WITH APPLICATIONS TO NUTRITION CALIBRATION

R. J. Carroll, David Ruppert, and A. H. Welsh *

January 22, 1997

Abstract

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 formulae and plug-in an estimate of a density function. The same philosophy is used in estimating the bias of the nonparametric function, i.e., we use an empirical method 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 and phrases: Asymptotic Theory; Bandwidth Selection; Local Polynomial Regression; Logistic Regression; Measurement Error; Missing Data; Nonlinear Regression; Partial Linear Models; Sandwich Estimation.

Short title. Local estimating equations.

*R. J. Carroll is Professor of Statistics, Nutrition and Toxicology, Department of Statistics, Texas A&M University, College Station, TX 77843–3143. D. Ruppert is Professor, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY 14853–3801. A. H. Welsh is Reader, Department of Statistics, Australian National University, Canberra ACT 2601. Carroll's research was supported by a grant from the National Cancer Institute. Carroll's research was partially completed while visiting the Institut für Statistik und Ökonometrie, Sonderforschungsbereich 373, Humboldt Universität 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. We also wish to acknowledge with thanks Donna Spiegelman and Walter Willett for making the Nurses' Health Trial calibration study data available to us.

1 INTRODUCTION

A general methodology which 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 estimates (Diggle, Zeger and Liang, 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 $\left(\widetilde{Y}_1, \widetilde{Y}_2, \ldots \widetilde{Y}_n\right)$, with the $\widetilde{Y}$'s possibly vector valued, then a parameter Θ is estimated by solving the estimating equation

$$0 = \sum_{i=1}^{n} \psi(\widetilde{Y}_i, \widehat{\Theta}). \tag{1}$$

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 loglikelihood 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 together 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; see section 2 for 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 we propose, 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. A by-product of the work is a considerable generalization of nonparametric regression methodology. This paper is primarily concerned with the case that 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, while 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 $\widetilde{Y}$ given $Z = z$, while $\alpha(z)$ is the correlation between a component of $\widetilde{Y}$ 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 $b_j = \Theta^{(j)}(z_0)/j!$, the local polynomial of order p in a neighborhood of z_0 is $\Theta(z) \approx \sum_{j=0}^{p} b_j (z - z_0)^j$. The local weight for a value of z near z_0 is denoted by $w(z, z_0)$. We then propose to solve in $(b_0, \ldots, b_p)$ the $q \times (p+1)$ equations

$$0 = \sum_{i=1}^{n} w(Z_i, z_0) \psi \left\{ \widetilde{Y}_i, \sum_{j=0}^{p} b_j (Z_i - z_0)^j \right\} G_p^t (Z_i - z_0) , \tag{2}$$

where $G_p^t(v) = (1, v, v^2, \ldots, v^p)$. The final estimates are $\widehat{\Theta}(z_0) = \widehat{b}_0$ and $\widehat{\alpha}(z_0) = \tau\{\widehat{b}_0\}$.

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 Nadaraya-Watson kernel regression has $p = 0$, $\psi(\widetilde{Y}, v) = \widetilde{Y} - v$ and $w(z, z_0)$ chosen to be a kernel weight;

(b) Local linear regression has $p = 1$, $\psi(\widetilde{Y}, v) = \widetilde{Y} - v$, and if $w(z, z_0)$ is a nearest neighbor weight the result is the LOESS procedure in Splus (Chambers and Hastie, 1992);

(c) When the mean and variance of a univariate response $\widetilde{Y}$ are related through $E(\widetilde{Y}|Z) = \mu\{\Theta(Z)\}$ and $\mathrm{Var}(\widetilde{Y}|Z) = \sigma^2 V\{\Theta(Z)\}$ for known functions μ and V, local quasilikelihood regression is based on

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

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 paper is organized as follows. In section 2, we describe in detail a problem from nutrition which 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 which must be set or estimated. If they are to be estimated, then the typical approach is to minimize mean squared error, 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. However, following

this approach in practice requires density estimation, estimation of higher order derivatives, etc., and these complications would limit the range of applications. Instead, we estimate the bias and variance functions empirically, without explicit use of the asymptotic formulae. Bias estimation uses a modification of Ruppert's (1995) empirical bias method, while 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 involving the analysis of nutrient intake data, transformations to additive models. extensions to missing data and partially parametric models. Section 5 discusses modifications of the algorithm (2). Section 6 has some concluding remarks. All theoretical details are collected into an appendix.

Local estimation of parameters for likelihood problems have been previously considered in important work by such authors as Tibshirani and Hastie (1987), Staniswallis (1989) and Hastie and Tibshirani (1990), and these techniques are implemented in Splus for GLIMs. Our methods and this paper differ from the local likelihood literature in several important 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-Thompson adjustments for missing data, GEE–type mean and variance function modeling, etc. A number of our examples, both numerical and theoretical, illustrate the use of non-likelihood estimating functions.

- Our estimates of variance are exceedingly straightforward, being nothing more than 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. The use of the parametric sandwich method in general nonparametric regression contexts has not to the best of our knowledge 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.

- 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 for asymptotic theory to derive a bias expression, additional nonparametric regressions to

estimate this expression, or the need to develop case-by-case tricks to get started.

2 MOTIVATING EXAMPLE

The purpose of this section is to 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 one that 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 FFQ's as the basic dietary instrument, one needs to know the relationship between reported intakes from the FFQ and true usual intake, defined operationally below. Such a relationship is ascertained through a substudy, commonly called a calibration study.

The primary aim of a calibration study may not be exactly the same in each case. Here we focus on the estimation of the correlation between FFQ intake and usual intake. 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.

FFQ's are thought to often involve a systematic bias (i.e., under- or over-reporting at the level of the individual). The other two instruments that are commonly used are the 24-hour food recall and the multiple-day food record (FR). Each of these FR's 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.

The usual model (Freedman, Carroll and Wax, 1991) relating intake of some nutrient (e.g., % calories from fat) reported on a FFQ (denoted by Q) and intake reported on m FR's (denoted by F) to long-term usual intake (denoted by T) is a standard linear errors-in-variables model

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

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

In model (4), β_1 represents the systematic bias of FFQ's, while the U_{ij} are the within individual

variation in FR's. All random errors, i.e., ϵ's and U's, are uncorrelated for purposes of this paper, see the end of section 4.1 for more details and further comments.

Two studies which we will analyze later fit exactly into this design. The Nurses' Health Study (Rosner, Willett and Spiegelman, 1989), hereafter denoted by NHS, has a calibration study of size $n = 168$ women all of whom completed a single FFQ and four multiple-day food diaries. The Women's Interview Survey of Health, hereafter denoted by WISH, has a calibration study with $n = 271$ participants who completed a FFQ and six 24-hour recalls on randomly selected days at least two weeks apart. While different FFQ's 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 components 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, respectively. One can expect then that the NHS data will provide considerably more power for estimating effects than will WISH.

For an initial analysis, we computed ρ_{QT} for each subpopulation formed by the quintiles of age. The five correlations were, roughly, 0.4, 0.6, 0.4, 0.5, and 0.8; see Figure 1 The five estimates 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 0.13, except for the highest quintile for which it is approximately 0.07.

Such stratified analysis (in this case the strata have been defined by age quintiles) can be looked at through 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}. Since 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 which are 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 apriori 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. 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 purposes of specificity we consider here local estimation of the tuning constant. If we could determine the bias and variance functions of $\widehat{\alpha}(z_0)$, 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 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.

The kernel regression literature abounds with ways of estimating these 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 $\widehat{\Theta}(z)$ 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,

$$\begin{aligned}
\text{bias}\left\{\widehat{\Theta}(z)\right\} &\sim h^{p+1}\mathcal{G}_b\{z, K, \Theta(z), p\} \quad \text{if } p \text{ is odd;} \\
&\sim h^{p+2}\mathcal{G}_b\{z, K, \Theta(z), p\} \quad \text{if } p \text{ is even;} \\
\text{cov}\{\widehat{\Theta}(z)\} &\sim \{nhf_Z(z)\}^{-1}\,\mathcal{G}_v\{z, K, \Theta(z), p\}.
\end{aligned}$$

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 (25) in the appendix.

The actual formulae 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 as well.

For example, if $\widetilde{Y}$ and hence Θ are scalar, $p = 1$ and $\psi(\widetilde{y}, v) = \widetilde{y} - v$ (ordinary local linear regression), then

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

where

$$\begin{aligned}
B(z) &= E\left\{ (\partial/\partial v)\, \psi\left(\widetilde{Y}, v\right) | Z = z \right\}; \\
C(z) &= E\left\{ \psi\left(\widetilde{Y}, v\right) \psi^t\left(\widetilde{Y}, v\right) | Z = z \right\},
\end{aligned}$$

with both $B(z)$ and $C(z)$ evaluated at $v = \Theta(z)$. In this specific example, $B(z) = -1$ and $C(z) = \mathrm{var}(Y|z)$.

We now return to tuning constant estimation. For local regression, one could in principle use the asymptotic expansions to derive bias and variance formulae for $\widehat{\alpha}(z_0)$. 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 Z's. Instead of carrying through this line of argument, we propose instead methods which avoid direct use of asymptotic formulae and which are applicable as well to methods other than local regression.

3.1 Empirical Bias Estimation

Ruppert (1995) suggested a method of bias estimation which avoids direct estimation of higher order derivatives arising in asymptotic bias formulae; the method is called EBBS, for Empirical Bias Bandwidth Selection.

The basic idea is as follows. Fix h_0 and z_0, and use as a model for the bias a function $f(h, \gamma_1)$ known except for the parameter γ_1, e.g., $f(h, \gamma_1) = \gamma_1 h^{p+1}$ for local pth degree polynomial kernel regression. For any h_0, form a neighborhood of tuning constants $\mathcal{H}_0$. On a suitable grid of tuning constants h in $\mathcal{H}_0$, compute the local polynomial estimator $\widehat{\alpha}(z_0, h)$, which should be well-described as a function of h by $\widehat{\alpha}(z_0, h) \approx \gamma_0 + f(h, \gamma_1)$, the value $\gamma_0 = \alpha(z_0)$ in the limit. Appealing to asymptotic theory, and if $\mathcal{H}_0$ is small enough, the bias should be well-estimated at h_0 by $f(h_0, \widehat{\gamma}_1)$.

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 0.1 and 1.5, respectively. Our experience is that the optimal local bandwidth is generally in this range. Then form an equally spaced, or perhaps geometrically spaced, grid of M points

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

Fix constants (J_1, J_2). For any $j = 1+J_1, ..., M-J_2$, apply the procedure defined in the previous paragraph with $h_0 = h_j$ and $\mathcal{H}_0 = \{h_k, k = j - J_1, ..., j + J_2\}$. This defines $\widehat{\text{bias}}\{\widehat{\alpha}(z_0, h_j)\}$. For tuning constants not on the grid $\mathcal{H}_1$, interpolation via a cubic spline is used.

Note that we have to set the limits of interesting tuning constants $[h_a, h_b]$ and the three tuning constants (M, J_1, J_2). Ruppert (1995) finds that $J_1 = 1$, $J_2 = 1$, and M between 12 and 20 give good numerical behavior in the examples 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 Θ, p is the degree of the local polynomial, and 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 $(\widehat{b}_0, ...\widehat{b}_p)$. In parametric problems, the solution $\widehat{\Theta}$ to (1) has sandwich (often called "robust") covariance matrix estimate $B_n^{-1} C_n (B_n^t)^{-1}$, where

$$\begin{aligned} C_n &= \sum_{i=1}^{n} \psi(\widetilde{Y}_i, \widehat{\Theta}) \psi^t(\widetilde{Y}_i, \widehat{\Theta}); \\ B_n &= \sum_{i=1}^{n} \left(\partial / \partial \Theta^t\right) \psi(\widetilde{Y}_i, \widehat{\Theta}). \end{aligned}$$

The analogous formulae for the solution to (2) are defined as follows. In what follows, if A is $\ell \times q$ and B is $r \times s$, then $A \otimes B$ is the Kronecker product defined as the $\ell r \times qs$ matrix which is formed by multiplying individual elements of A by B, e.g., if A is a 2×2 matrix,

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

Let $\chi(\widetilde{y}, v) = (\partial / \partial v^t)\psi(\widetilde{y}, v)$. Then the asymptotic covariance matrix of $(\widehat{b}_0, ...\widehat{b}_p)$ is estimated by $\{B_n(z_0)\}^{-1} C_n(z_0) \{B_n^t(z_0)\}^{-1}$, where

$$C_n(z_0) = \sum_{i=1}^{n} w^2(Z_i, z_0) \left[\left\{ G_p(Z_i - z_0) G_p^t(Z_i - z_0) \right\} \otimes (\widehat{\psi}_i \widehat{\psi}_i^t) \right]; \tag{6}$$

$$B_n(z_0) = \sum_{i=1}^{n} w(Z_i, z_0) \left[\left\{ G_p(Z_i - z_0) G_p^t(Z_i - z_0) \right\} \otimes \widehat{\chi}_i \right], \tag{7}$$

where $\widehat{\psi}_i = \psi\{\widetilde{Y}_i, \sum_{j=0}^{p} \widehat{b}_j (Z_i - z_0)^j\}$ and analogously for $\widehat{\chi}_i$. An argument justifying these formulae 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 we have investigated, this little-known empirical adjustment improves coverage probabilities of

sandwich-based confidence intervals, when combined with t-percentiles with $n-(p+1)q$ degrees of freedom.

In some problems, the sandwich term $C_n(z_0)$ can be improved upon because the covariance matrix of $\psi(\cdot)$ is known partially or fully. For example, if $\psi(\cdot)$ is given by (3), then $E(\psi\psi^t) = \sigma^2\{\mu^{(1)}\}^2/V$, and one would replace $(\widehat{\psi}_i\widehat{\psi}^t)$ in (6) by $\widehat{\sigma}^2\{\widehat{\mu}_i^{(1)}\}^2/\widehat{V}_i$. In addition, using score-type arguments one bases work on $\chi(\cdot) = -\{\mu^{(1)}(\cdot)\}^2/V$, and one would replace $\widehat{\chi}_i$ in (6) by $-\{\widehat{\mu}_i^{(1)}\}^2/\widehat{V}_i$. 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 $\widetilde{Y}$ is scalar, $\psi(\widetilde{Y}, x) = I(\widetilde{Y} \leq x) - 1/2$, where I is the indicator function. This choice of $\psi(\cdot)$ has zero derivative, and thus (7) equals zero. Alternatives to the sandwich estimators do exist, however, although their implementation and indeed the theory needs further investigation. A sandwich-type method was described by Welsh, Carroll and Ruppert (1994), who use 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 the 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

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 (1995).

The unknown parameters in the problem are conveniently characterized as $\Theta = (\theta_1, ..., \theta_6)$,

where $\theta_1 = E(Q)$, $\theta_2 = E(F) = E(T)$, $\theta_3 = \mathrm{var}(Q)$, $\theta_4 = \mathrm{cov}(Q,F) = \mathrm{cov}(Q,T)$, $\theta_5 = \mathrm{var}(U)$ and $\theta_6 = \mathrm{cov}(F_1, F_2) = \mathrm{var}(T)$. Letting $\widetilde{Y}_i = (Q_i, F_{i1}, ..., F_{im})$ be the observed data ($m = 6$ in WISH, $m = 4$ in NHS), the usual method of moments estimating function is

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

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 $\overline{F}_i$ on Z_i and once they are plugged into the third–sixth components of ψ, $\{\theta_3(z), ..., \theta_6(z)\}$ can also be computed by local least squares, e.g., by regressing $(Q_i - \widehat{\theta}_1)^2$ on Z_i to obtain $\widehat{\theta}_3$. 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 and the tricubed kernel function, which is proportional to $(1 - |v|^3)^3$ for $|v| \leq 1$ and equals zero elsewhere. For a fixed value of the span, we assessed standard errors by two means. First, an estimated covariance matrix for $\widehat{\Theta}(z_0)$ was obtained using the sandwich formula, and then the delta-method was used to obtain an estimated variance for $\rho_{QT}(z_0)$, $\beta_1(z_0)$, etc. The second standard error estimates are based on the nonparametric bootstrap, with the pairs $(\widetilde{Y}, Z)$ resampled from the data with replacement; we used 500 bootstrap samples. For a range of spans and for a variety of data sets 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 zero, then the estimator $\widehat{\Theta}(z)$ 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 ρ_{QT}(age) for the NHS % Calories from Fat for various spans in the range 0.6 to 0.9 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. While there is some variation between the curves for the different values of the span, the essential feature is remarkably consistent, namely that those under the age of 50 have significantly (in the practical sense) lower correlations than do those aged greater than 50. There are a variety of ways to assess the statistical significance of this finding. The simplest is to split the data into two populations on the basis of age groups and simply compute $\widehat{\rho}_{QT}$ for each population; the estimates are statistically

significantly different at a significance level less that 0.02.

A second test is slightly more involved. We computed the estimate of ρ_{QT}(age) for 16 equally-spaced points on the range from 34 to 59, along with the bootstrap covariance matrix of these 16 estimates. We then tested whether the 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 0.05 for spans equal to 0.7, 0.8 and 0.9.

We also estimated the span, in the following manner. For computational purposes in estimating the span we used eight values of age, and using the methods of section 3 we computed an estimate of the mean squared error using empirical bias estimation ($J_1 = J_2 = 5$, $M = 41$, $h_a = 0.6$, $h_b = 1.0$) and the sandwich method; the estimated span was chosen to minimize the sum over the eight ages of the estimated mean squared error. The estimated span for local linear regression was 0.78, while it was 0.90 for local quadratic regression. We then bootstrapped this process, including the estimation of the span, and found that while the significance level was slightly greater than for a fixed span, but it was still below 0.05.

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 means squared error 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 mean squared errors 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 the 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 FR's in WISH than in NHS.

Finally, we investigated local average, linear, quadratic and cubic regression, with a span of 0.8, 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 end points.

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.

- We have assumed that the errors ϵ_i are independent of U_{ij}. This appears to be roughly the case in these two data sets, although it is not true in other data sets we have studied, e.g., the Women's Health Trial data studied by Freedman, et al. (1991). The model and estimating equation are easily modified in general to account for such correlation when it occurs. Similarly, the model and estimating equation can be modified to take into account a parametric model for correlation among the $U'_{ij}s$, e.g., an AR(1) model. While such correlations exist in these data sets, 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 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 for example Martin, Su, Jones, Lockwood, Tritchler and Boyd (1996). The bias is less crucial for log(total calories) and presumably even less 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 which allow for bias. For example, we changed the FFQ and food record data for those with $22 \leq \text{BMI} \leq 28$ by adding on average 4 to their % Calories from Fat (a 10% change), while for those with BMI > 28 we added on average 7 to their % Calories from Fat (a 20% increase). The adjustments were proportional to FFQ's and food records, and the same adjustment was added to all food records within an individual. These adjustment in effect simulate adjustments to the data which would be made if a strong bias were found in % Calories from Fat for food records. The analysis of the modified data gave correlations which were very similar to those shown in our graphs, i.e., the effect of bias on the correlation estimates was small.

- If one had replicated FFQ's, there are many modifications to the basic model which can be made. One might conjecture an entirely different error structure, e.g.,

$$Q_{ij} \quad = \quad \beta_0 + \beta_1 T_i + r_i + \epsilon_{ij};$$

$$
\begin{aligned}
F_{ij} &= T_i + s_i + U_{ij}; \\
\sigma_s^2 &= \sigma_r^2.
\end{aligned}
$$

This model is only identifiable if $\mathrm{corr}(r,s)$ is known. We have fit such models using local method of moments to a large ($n > 400$) data set with repeated FFQ's and using 24-hour recalls for various choices of $\mathrm{corr}(r,s) \leq .5$. The net effect was that such analyses are very different from those based on model (4)–(5); ρ_{QT} increased by a considerable amount, while the local estimates of $\mathrm{var}(T)$ 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 paper can be extended to the multivariate Z case. Suppose that $Z_i = (Z_{i1}, ..., Z_{im})^t$, where the Z_{ij} are scalar. Then, as in Ruppert and Wand (1994), local linear functions are $\Theta(z) = b_0 + b_1(z - z_o)$, where b_0 is a $p \times 1$ vector and b_1 is a $p \times m$ matrix. The generalization of (2) is to solve

$$
0 = \sum_{i=1}^{n} w(Z_i, z_0)\, \psi \left\{ \widetilde{Y}_i, b_0 + b_1(Z_i - z_0) \right\} G_{p,m}(Z_i - z_0), \tag{9}
$$

where $G_{p,m}^t(v) = (1, v^t)$. When Z is multivariate and using kernel weights, the kernel K is multivariate and the bandwidth h is replaced by a positive definite symmetric matrix H. The simplest choice is to restrict H to equal hI for $h > 0$ and I the identity matrix, and in this situation the methods we have discussed for empirical bias and variance estimation apply immediately to the estimates $\widehat{\Theta}(z_0) = \widehat{b}_0$. The application of empirical bias modeling to more general bandwidth matrices is under current investigation.

Extensions to higher order local polynomials require more care. Completely nonparametric functional versions are easy in principle, but the notation is horrendous and practical implementation difficult; see Ruppert and Wand (1994), their section 4. It is much easier to fit additive models, so that if $z = (z_1, ..., z_m)^t$ and $z_0 = (z_{01}, ..., z_{0m})^t$, then $\Theta(z) = b_0 + \sum_{k=1}^{m} \sum_{j=1}^{p} b_{kj}(z_k - z_{0k})^j$; this is identical to (9) when $p = 1$, the extension of (9) to $p > 1$ is immediate.

For an example of (9), we consider a problem in which $\widetilde{Y} = 10 + \log[(m + 0.5)/(10^5 - m + 0.5)]$, where m is the mortality rate per 10^5 males for males dying of lung cancer, as a function of $Z =$ (age class, year). We will call $\widetilde{Y}$ the "adjusted" logit because of the 0.5 offset. The data come from the Australian Institute of Health and are publically 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. Here, N is the size of the relevant subpopulation. The values of the N's are known and will be used later. Since p is small, d is approximately Poisson(Np) and $\text{var}(m) \approx (10^5/N)E(m)$. 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 $\widetilde{Y}$ as a function of its mean and N. Alternatively, we could model the variance of $\widetilde{Y}$ as a function of Z. We will start with the second possibility. If $\Theta = (\theta_1, \theta_2)^t$, the estimating function for mean and variance estimation is just $\psi(\widetilde{Y}, \Theta) = \{\widetilde{Y} - \theta_1, (\widetilde{Y} - \theta_1)^2 - \theta_2\}^t$. However, there are two good reasons for considering a robust analysis. Firstly, there may be concern over the potential for outliers in the response, and secondly, a robust analysis may be numerically more stable. We treat $\tau = \log(\theta_2)$ as the spread parameter (to ensure non-negativity) and use the estimating equation

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

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 parametric estimator is known as "Proposal 2" (Huber, 1981). The spread estimating function can be rewritten as

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

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, i.e., $\widehat{\Theta}_1(z)$ and $\widehat{\Theta}_2(z)$, for the lung cancer mortality data have the following behavior. After some experimentation, the bandwidth matrix was restricted to be of the form $h\,\text{diag}(2, 1)$, and then h was chosen empirically as in Section 3, with a backfitting modification to the basic algorithm (2) described in section 5. However, the results reported below 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, although local quadratic estimates are similar but with somewhat higher peaks in the spread surface. It is clear that the logit of mortality increases non-linearly with age class and that there is at best a very weak year effect which 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 0.5 offset) and the near Poisson variability of m; see the discussion in the final paragraph of this subsection. 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 mostly due to variation in the values of N.

We also tried modeling the variance of $\widetilde{Y}$ as a function of N and the mean of $\widetilde{Y}$. 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 $(N^*)^{1/2}$.

As mentioned earlier, if we assume that

$$\text{var}(m) = (10^5/N)E(m) \tag{10}$$

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

$$\text{std dev}(\widetilde{Y}) \approx \frac{10^5+1}{(E(m)+.5)(10^5-E(m)+.5)}\left\{10^5 E(m)/N\right\}. \tag{11}$$

We checked (10) by dividing the residuals by the right hand 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 Binary Regression: Bladder Cancer

In this example, Y_i is the indicator of bladder cancer, Z_i is the value of a univariate risk factor (the investigators have not given permission to name the variable), and the objective is to model $P(Y_i = 1|Z_i = z)$. While the problem of local logistic regression has been mainly solved, we include this example to show that the methods we propose give reasonable estimates in familiar problems. In Figure 4 we see the linear and quadratic logistic regression fits, the solid and dashed curves, respectively. The two curves differ substantially, indicating that the linear logistic model may not fit well, but the nonmonotonic behavior of the quadratic logistic fit seems odd.

A nonparametric fit can be achieved by local linear logistic regression. Let $H(u) = \{1 + \exp(-u)\}^{-1}$ be the logistic function. Then $\psi(y, \Theta)$ is the score function when Y is Bernoulli with

mean $H(\Theta)$, i.e.,

$$\psi(y,\Theta) = \frac{\partial}{\partial\Theta}\log\left[H^y(\Theta)\left\{1-H(\Theta)\right\}^{1-y}\right] = y - H(\Theta),$$

which is (3) with $\mu = H$ and $V = H(1-H) = H'$. If $\mu(z_0) = P(Y=1|Z=z_0)$, then $\widehat{\mu}(z_0) = H(\widehat{b}_0)$ where

$$0 = \sum_{i=1}^{n} w(Z_i, z_0)\left[Y_i - H\left\{\sum_{j=0}^{p}\widehat{b}_j(Z_i - z_0)\right\}\right] G^t(Z_i - z_0).$$

Fan, Heckman, and Wand (1995) propose a "rough and ready" bandwidth, but consider further development of bandwidth selectors to be a "worthwhile future project." Here we apply the bandwidth methodology developed in Section 3.

The local linear logistic regression fit is given by the dotted line in Figure 4. It differs noticeably from the parametric fits and appears to be the best summary of the data. The nonparametric fit rises quickly, but then levels off around $Z = 0.5$. The sharp increase near the right boundary, say where $Z > 2.5$, is due to the three largest Z values being cancer cases and may be merely a chance phenomenon. The flattening for $0.5 < Z < 2.5$ is supported by a larger amount of data and is likely to be real. The local bandwidth given in the bottom graph of Figure 4 is nearly constant with some increase near the boundaries where the variance of the smooth is higher. This is typical of examples where the function being estimated has no regions of high curvature.

4.4 Nutrition Calibration With Missing Data

In many problems, a component of $\widetilde{Y}$ may be missing, with the probability of missingness depending on Z and the observable components of $\widetilde{Y}$. For example, in some calibration studies, a FFQ is observed for every individual but the FR's can be observed only for a subset, chosen on the basis of the initial FFQ, e.g., to overrepresent those with very high or very low fat in their diet. Let $\widetilde{R}$ be the observable components determining missingness, while Δ and $\pi(\widetilde{R})$ are the indicator and the probability that all components of $(\widetilde{Y}, Z)$ are observed, respectively. If $\pi(\cdot)$ is known, the Horvitz and Thompson (1952) modification of (2) is to reweight the estimating equation and solve

$$0 = \sum_{i=1}^{n}\frac{\Delta_i}{\pi(\widetilde{R}_i)} w(Z_i, z_0)\psi\left\{\widetilde{Y}_i, \sum_{j=0}^{p} b_j\left(Z_i - z_0\right)^j\right\} G_p^t\left(Z_i - z_0\right).$$

The probability function $\pi(\cdot)$ may be unknown in practice, but it can often be estimated at ordinary parametric rates using a flexible parametric model, thus not affecting the asymptotic properties of estimates of $\Theta(z)$.

4.5 Variance-Stabilizing Transformations

In measurement error models, one wishes to relate a response to a predictor X, but because of measurement error and other sources of variability, one cannot observe X. Instead, one can observe only a variable W related to X. The measurement error model literature was recently surveyed by Carroll, Ruppert and Stefanski (1995).

In order to correct for the biases caused by measurement error, essentially all methods are based on the assumption that, perhaps after a transformation, W differs from X only by additive error. Assuming the possibility of replicates, this means in symbols that the replicates (W_{ij}) for $j = 1, ..., m$ are related to X_i by $W_{ij} = X_i + U_{ij}$, where $E(U_{ij}|X_i) = 0$ and $E(U_{ij}^2|X_i) = \sigma^2$ where σ^2 does not depend on X_i. If this model holds, then the within-person sample means and sample variances of the W's are uncorrelated. In this case, it is common to say that the errors are "additive," though "homoscedastic" is perhaps a better choice of adjective.

A simple fully parametric method for transforming to additivity is to consider the parametric family of transformation $h(W, \lambda)$; we work with the power family, so that $h(v, \lambda) = (v^\lambda - 1)/\lambda$ if $\lambda \neq 0$ and $= \log(v)$ otherwise. The idea is to choose λ so that after transformation the sample means and some function of the sample variances, e.g., the sample standard deviations or the log variances, have correlation zero (Ruppert and Aldershof, 1989). With the power transformation family, our numerical experience is that on the usual interval $-2 \leq \lambda \leq 2$, there is a unique value of λ for which the sample correlation equals zero.

This approach can be placed into the framework of estimating functions. Let $\overline{W}(\lambda)$ and $s(\lambda)$ be the sample mean and standard deviation of the transformed data, respectively, and let $\widetilde{Y}$ be the replicates of W. The five unknown parameters $\Theta = (\mu_w, \mu_s, \sigma_w^2, \sigma_s^2, \lambda)$ can be obtained as solutions to (1) when $\psi(\cdot)$ is defined by

$$\psi(\widetilde{Y}, \Theta) = \begin{bmatrix} \overline{W}(\lambda) - \mu_w \\ s(\lambda) - \mu_s \\ \left\{\overline{W}(\lambda) - \mu_w\right\}^2 - \sigma_w^2 \\ \{s(\lambda) - \mu_s\}^2 - \sigma_s^2 \\ \left\{\overline{W}(\lambda) - \mu_w\right\}\{s(\lambda) - \sigma_s\} \end{bmatrix}. \tag{12}$$

In practice, we compute the correlation between $\overline{W}(\lambda)$ and $s(\lambda)$, and use the method of bisection to find λ that is the zero of this correlation (extensive numerical experience has shown that bisection works well with data).

We applied this estimating function locally using (2), both for the WISH data using $W =$ % Calories from Fat ($m = 6$ replicates) and $Z =$ Body Mass Index, and also for $W =$ Cholesterol

and $W = \log(\text{Systolic Blood Pressure} - 50)$ in the Framingham Heart Study data set ($m = 2$, Kannel, et al., 1986) with $Z =$ Age. Previous experience with these data have suggested that % Calories from Fat does not need much if any transformation, while the log transformation for Systolic Blood Pressure perhaps slightly overtransforms the data, although not in any practically significant way. We have previously had no experience with Cholesterol.

The results of the analysis with local linear regression and LOESS weights with a span = 0.7 are given in Figure 6. Since $\lambda = 1$ corresponds to no transformation, here we see that transformation of % Calories from Fat is basically unnecessary and is independent of body mass index. Some transformation of log Systolic Blood Pressure appears necessary, as expected, but there is no evidence that it depends on Age. Cholesterol appears to need transformation, but somewhat unexpectedly the transformation depends on Age. While this finding is somewhat interesting, in the data at hand any measurement error analysis will not depend particularly on the transformation used, since the correlation between the mean transformed response for any two values of $\lambda \in [-0.5, 0.5]$ is quite high (above 0.98), and the same holds for differences.

4.6 Variance Functions and Overdispersion

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

$$\text{var}(Y|X) = \exp\left[\theta_1 + \theta_2 \log\{\mu(\mathcal{B}, X)\}\right]. \tag{13}$$

Here we assume that the mean function is properly determined so that $\mathcal{B}_0 = (\beta_0, \beta_1, \beta_2)$ is to be estimated parametrically. If $\theta_2 = 1$ and $\theta_1 > 1$ then we have overdispersion relative to the Poisson model, while $\theta_2 \neq 2$ 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 θ_2, $\mathcal{B}_0$ is usually estimated by generalized least squares (quasilikelihood). Consistent estimates of $\mathcal{B}_0$ can be obtained using quasilikelihood assuming θ_2 is a fixed value, even if it is not. This well-known fact is often referred to operationally by saying that (13) with fixed θ_2 is a "working" variance model (Diggle, Liang and Zeger, 1994).

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

$$\psi(\widetilde{Y}, \Theta, \widehat{\mathcal{B}}) \quad = \quad \{Y - \mu(\widehat{\mathcal{B}}, X)\}^2 \exp(-\Theta) - 1. \tag{14}$$

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

$$0 = \sum_{i=1}^{n} w(Z_i, z_0)\psi\left\{\widetilde{Y}_i, \sum_{j=0}^{p} b_j\,(Z_i - z_0)^j\,, \widehat{\mathcal{B}}\right\} G_p^t\left(Z_i - z_0, \widehat{\mathcal{B}}\right), \tag{15}$$

Because $\widehat{\mathcal{B}}$ 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 data sets, the esterase assay and hormone assay data sets described in Carroll & Ruppert (1988, Chapter 2), and a simulated data set with $\Theta\{\eta(\mathcal{B}, X)\} = 1.6 + \sin\{\eta(\mathcal{B}, X)\}$, using the same X's 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 since $\theta_2 \approx 1.6$. We used $\theta_2 = 2$ as our working variance model to obtain $\widehat{\mathcal{B}}$ to these three data sets. We fit local linear models weighted using loess with the span allowed to take on values between 0.6 and 2.0 and estimated by the techniques of this paper. We compared 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 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.

4.7 Transformation in Nonlinear Regression: Kinetic Modeling

In fitting a response Y to a predictor Z using a nonlinear regression equation the usual models all assume implicitly that for some known function $f(Z, \mathcal{B})$

$$\text{median}(Y|Z) = f(Z, \mathcal{B}_0).$$

Here $\tilde{Y} = (Y, Z)$. A important example is fitting the Michaelis-Menten model,

$$\text{median}(Y|Z) = (\beta_0 + \beta_1/Z)^{-1}, \tag{16}$$

that is used in kinetic studies, some types of bioassay, and in stock-recruitment analysis of fisheries where it is called the Beverton-Holt model; see Ruppert, Cressie, and Carroll (1989). For concreteness, we will restrict attention to kinetic modeling, though the theory we discuss holds for general

models. A variety of fitting methods are used to estimate the parameters $\mathcal{B}_0 = (\beta_0, \beta_1)$ in this model. One method is nonlinear least squares estimation of Y on Z. Alternatively, since

$$\text{median}(Y^{-1}|Z) = \beta_0 + \beta_1/Z, \tag{17}$$

many researchers regress Y^{-1} on Z^{-1} using ordinary least squares. There is a third alternative, not discussed here, in which Z/Y is regressed on Z.

One should choose between (16) and (17) by using the model where the errors in the regression relationship are most nearly additive. Of course, there is no guarantee that the errors will be nearly additive for either model, and a more flexible approach proceeds as follows. If we define the usual power transformation $h(x, \Theta) = (x^{\Theta} - 1)/\Theta$ for $\Theta \neq 0$ and $= \log(x)$ for $\Theta = 0$, then (16) and (17) are special cases of the model

$$h(Y, \Theta) = h(\beta_0 + \beta_1/Z, \Theta) + \epsilon, \tag{18}$$

for $\Theta = 1$ and $\Theta = -1$, respectively, where ϵ is a symmetric (often normal) random variable independent of Z. The transformation parameter can be estimated by maximum likelihood assuming that ϵ is normally distributed (Ruppert, Cressie and Carroll, 1989), or to minimize tests for skewness or heteroscedasticity (Ruppert and Aldershof, 1989). For given $\mathcal{B}_0$, any of the methods can be recast as solving an equation such as (15), so that Θ, rather than being fixed, depends on Z. The preliminary estimate of $\mathcal{B}_0$ can again be obtained by via a "working" transformation using any of the methods described above, followed by a least absolute values regression in the "working" transformed scale. Least absolute values regression is used because it consistently estimates $\mathcal{B}$ even when $\Theta(z)$ is misspecified; see Carroll and Ruppert (1984).

4.8 Partially Parametric Models

The overdispersion (section 4.6) and kinetic modeling (section 4.7) examples both contained a parametric part $\mathcal{B}_0$ and a nonparametric part $\Theta(\cdot)$. The "working" estimation methods used for the parametric parts were carefully chosen so that $\widehat{\mathcal{B}}$ was consistent and asymptotically normally distributed with variance of order n^{-1} *even if* $\Theta(\cdot)$ *was completely misspecified.*

The kinetic modeling problem is a good example of what happens when 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 $\widehat{\Theta}(\cdot)$ is to perform nonlinear least squares regression of $h(Y, \widehat{\Theta})$ on $h\left\{(\beta_0 + \beta_1/Z), \widehat{\Theta}\right\}$. The resulting estimate of $\mathcal{B}_0$ is consistent only if $\widehat{\Theta}(\cdot)$ is consistent

(and vice-versa!). The estimate solves in $\mathcal{B}$ an estimating equation of the form

$$0 = \sum_{i=1}^{n} \chi\left\{\widetilde{Y}_i, \mathcal{B}, \widehat{\Theta}(\cdot)\right\},$$

where in this instance $\chi(\cdot)$ is the nonlinear least squares normal equation in the transformed scale $\widehat{\Theta}(\cdot)$. The natural approach to use then is to solve the following equations

$$0 = \sum_{i=1}^{n} \chi\left\{\widetilde{Y}_i, \mathcal{B}, \widehat{\Theta}(\cdot)\right\}; \tag{19}$$

$$0 = \sum_{i=1}^{n} w(Z_i, z_0)\psi\left\{\widetilde{Y}_i, \mathcal{B}, \Theta(\cdot)\right\} G_p^t\left(Z_i - z_0\right), \tag{20}$$

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

As we have described it, solving (19)–(20) 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 (20) and assuming that $\widehat{\mathcal{B}}$ is fixed at its estimated value. Asymptotically valid estimates of the covariance matrix of $\widehat{\mathcal{B}}$ remain an open problem, although in some cases they can be derived using the methods of Carroll, Fan, Gijbels and Wand (1995).

The backfitting algorithm has a well-known feature. We confine our remarks to local regression, but they 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 $n^{1/2}(\widehat{\mathcal{B}} - \mathcal{B}_0)$ has variance of order 1 but has bias of the order $(nh^4)^{1/2}$, so that to get 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 $\widehat{\mathcal{B}}$ does not disappear. If one is even going to worry about this problem (we know of no commercial program which does, nor of any practical examples in which the bias problem is of real concern), the usual solution is to undersmooth in some way. For example, one might use a standard bandwidth but set $p \geq 2$ in (2) and (20).

Some problems allow for a somewhat more elegant solution to the bias problem, specifically when (19)–(20) are formed as the derivatives of a *single* optimization criterion. None of the estimators we have described in this paper have this form except the kinetic modeling example when all parameters are estimated by maximum likelihood under the assumption that the errors are normally distributed (section 4.7). Optimization of a single criterion basically means a likelihood specification. When it 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

$\ell\{\mathcal{B}, \Theta(\cdot)\}$. For fixed $\mathcal{B}$, let $\widehat{\Theta}(\cdot, \mathcal{B})$ be the local estimator derived by maximizing the likelihood in Θ with $\mathcal{B}$ fixed. Nonparametric likelihood maximizes $\ell\{\mathcal{B}, \widehat{\Theta}(\cdot, \mathcal{B})\}$ as a function of $\mathcal{B}$. In contrast, backfitting fixes the current $\widehat{\Theta}(\cdot, \mathcal{B})$ and updates the estimate of $\mathcal{B}$ by maximizing $\ell\{\alpha, \widehat{\Theta}(\cdot, \mathcal{B})\}$ in α. Nonparametric likelihood can be more difficult to implement than backfitting, especially in our context when $\Theta(\cdot)$ is multivariate. It is however easy to implement if Θ is scalar, $\widetilde{Y} = (Y, X, Z)$, and Y follows a GLIM with mean $f\{\Theta(Z) + X^t\mathcal{B}\}$; see Severini and Staniswalis (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 as variance function nonparametrically, one would often first estimate the mean function, say $\Theta_1(z)$, form squared residuals $\{\widetilde{Y} - \widehat{\Theta}_1(Z_i)\}^2$, and then regress these squared residuals on Z nonparametrically to obtain $\widehat{\Theta}_2(z_0)$, the variance estimate at a given z_0. In this context, strict application of (2) is different, since it is based on squared pseudo-residuals $\{\widetilde{Y} - \sum_{j=0}^{p} \widehat{\Theta}^{(j)}(z_0)(Z_i - z_0)^j/j!\}^2$. In addition, one would often use different tuning constants at each step, but (2) assumes use of the same tuning constant.

The previous example, as well as the nonparametric calibration problem, is an example 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 $\Theta = (\Theta_1, \Theta_2)$. By the two-stage process, we mean that the first component can be estimated without reference to the first, with weight function w_1 and estimating function ψ_1, so that we solve

$$0 = \sum_{i=1}^{n} w_1(Z_i, z_0)\psi_1 \left\{ \widetilde{Y}_i, \sum_{j=0}^{p} b_{j,1} (Z_i - z_0)^j \right\} G_p^t (Z_i - z_0) . \tag{21}$$

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

At the second stage there is a second weight function w_2 and a second estimating function ψ_2, and we solve

$$0 = \sum_{i=1}^{n} w_2(Z_i, z_0)\psi_2 \left\{ \widetilde{Y}_i, \widehat{\Theta}_1(Z_i), \sum_{j=0}^{p} b_{j,2} (Z_i - z_0)^j \right\} G_p^t (Z_i - z_0) . \tag{22}$$

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

The asymptotic covariance matrix of $\{\widehat{\Theta}_1(z_0), \widehat{\Theta}_2(z_0)\}$ defined by (21)–(22) is estimated by applying the sandwich method to the estimating equation

$$0 = \sum_{i=1}^{n} \begin{bmatrix} w_1(Z_i, z_0)\psi_1 \left\{ \widetilde{Y}_i, \sum_{j=0}^{p} b_{j,1} (Z_i - z_0)^j \right\} \\ w_2(Z_i, z_0)\psi_2 \left\{ \widetilde{Y}_i, \sum_{j=0}^{p} b_{j,1} (Z_i - z_0)^j , \sum_{j=0}^{p} b_{j,2} (Z_i - z_0)^j \right\} \end{bmatrix} G_p^t (Z_i - z_0) . \tag{23}$$

If $c_{p.i} = G_p(Z_i - z_0)G_p^t(Z_i - z_0)$, the sandwich formulae are

$$\begin{aligned} B_n(z_0) &= \sum_{i=1}^{n} \begin{Bmatrix} w_1(Z_i, z_0)c_{p.i} \otimes \widehat{\chi}_{i11} & 0 \\ w_2(Z_i, z_0)c_{p.i} \otimes \widehat{\chi}_{i21} & w_2(Z_i, z_0)c_{p.i} \otimes \widehat{\chi}_{i21} \end{Bmatrix} \\ C_n(z_0) &= \sum_{i=1}^{n} \begin{Bmatrix} w_1^2(Z_i, z_0)c_{p.i} \otimes \widehat{\psi}_{i1}\widehat{\psi}_{i1}^t & w_1(Z_i, z_0)w_2(Z_i, z_0)c_{p,i} \otimes \widehat{\psi}_{i1}\widehat{\psi}_{i2}^t \\ w_1(Z_i, z_0)w_2(Z_i, z_0)c_{p.i} \otimes \widehat{\psi}_{i2}\widehat{\psi}_{i1}^t & w_2^2(Z_i, z_0)c_{p.i} \otimes \widehat{\psi}_{i2}\widehat{\psi}_{i2}^t \end{Bmatrix} , \end{aligned}$$

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

Tuning constant estimation in multistage problems may also need to be adjusted. 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}, while 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, while in general estimating h_2 requires a two-dimensional EBBS. However, in both the variance function problem as well as nonparametric calibration, the effect on Θ_2 due to estimating Θ_1 is nil asymptotically, and standard EBBS can be used at each stage without modification.

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

$$\begin{aligned} 0 &= \sum_{i=1}^{n} w_1(Z_i, z_0)\psi_1 \left\{ \widetilde{Y}_i, \sum_{j=0}^{p} b_{j,1} (Z_i - z_0)^j , \widehat{\Theta}_2(Z_i) \right\} G_p^t (Z_i - z_0) ; \\ 0 &= \sum_{i=1}^{n} w_2(Z_i, z_0)\psi_2 \left\{ \widetilde{Y}_i, \widehat{\Theta}_1(Z_i), \sum_{j=0}^{p} b_{j,2} (Z_i - z_0)^j \right\} G_p^t (Z_i - z_0) . \end{aligned}$$

For example, this is the procedure 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

$$\begin{aligned} 0 &= \sum_{i=1}^{n} w_1(Z_i, z_0)\psi_1 \left\{ \widetilde{Y}_i, \sum_{j=0}^{p} b_{j,1} (Z_i - z_0)^j , \sum_{j=0}^{p} b_{j,2} (Z_i - z_0)^j \right\} G_p^t (Z_i - z_0) ; \\ 0 &= \sum_{i=1}^{n} w_2(Z_i, z_0)\psi_2 \left\{ \widetilde{Y}_i, \sum_{j=0}^{p} b_{j,1} (Z_i - z_0)^j , \sum_{j=0}^{p} b_{j,2} (Z_i - z_0)^j \right\} G_p^t (Z_i - z_0) . \end{aligned}$$

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 Θ 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, e.g., a bandwidth h. A suitable value of h can be found by minimizing an estimate of the mean square error. The latter is found by estimating variance using the "sandwich formula" (or more efficient modifications described earlier) and estimating bias empirically as in Ruppert (1995).

We have applied this methodology to nonparametric calibration in nutritional studies, binary regression with the indicator of bladder cancer as the response, variance-stabilizing transformations, and robust modeling of lung cancer mortality rates. Other possible application, e.g., to overdispersion and to transformation in nonlinear regression, have been discussed.

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 $(\xi_1, ..., \xi_p)$, a regression cubic spline has the form

$$\Theta(z, b_0, ..., b_{p+3}) = b_0 + b_1 z + b_2 z^2 + b_3 z^3 + \sum_{j=1}^{p} b_{j+3}(z - \xi_j)_+^3,$$

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

$$0 = \sum_{i=1}^{n} \psi \left\{ \widetilde{Y}_i, \Theta(Z_i, b_0, ..., b_{p+3}) \right\} G_{p,s}(Z_i),$$

where $G_{p,s}^t(z) = (1, z, z^2, z^3, (z - \xi_1)_+^3, ..., (z - \xi_p)_+^3)$. The interesting issue here is the selection of the knots, a problem of considerable interest in the broad context and one we are currently working on for estimating functions. The regression splines outlined above may have an advantage since the knots can be chosen on a component-wise basis.

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. (1995), "Generalized Partially Linear Single-Index Models," Preprint.

Carroll, R. J., and Ruppert, D. (1984), "Power transformations when fitting theoretical models to data," *Journal of the American Statistical Association*, 79, 321–328.

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: Wadsworth.

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

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.

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: John Wiley and Sons.

Kannel, W. B., Neaton, J. D., Wentworth, D., Thomas, H. E., Stamler, J., Hulley, S. B., and Kjelsberg, M. O. (1986), "Overall and Coronary Heart Disease Mortality Rates in Relation to Major Risk Factors in 325,348 Men Screened for MRFIT," *American Heart Journal*, 112, 825–836.

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*, to appear.

Politis, D. N., and Romano, J. P. (1994), "Large Sample Confidence Regions Based on Subsamples Under Minimal Conditions," *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. (1995), "Empirical-Bias Bandwidths for Local Polynomial Nonparametric Regression and Density Estimation," TR1137, School of Operations Research and Industrial Engineering, Cornell. (http://gamble.orie.cornell.edu/trlist/trlist.html)

Ruppert, D., and Aldershof, B. (1989), "Transformations to Symmetry and Homoscedasticity" *Journal of the American Statistical Association*, 84, 437–446.

Ruppert, D., Cressie, N., and Carroll, R. J. (1989), "A Transformation/Weighting Model for Estimating Michaelis-Menten Parameters," *Biometrics*, 45, 637–656.

Ruppert, D., Wand, M.P., Holst, U., and Hössjer, O. (1995), "Local polynomial variance function estimation," TR1132, School of Operations Research and Industrial Engineering, Cornell University. (http://gamble.orie.cornell.edu/trlist/trlist.html)

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

Severini, T. A., and Staniswalis, J. G. (1994), "Quasilikelihood 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, 1768–1802.

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

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

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

Welsh, A. H. (1996), "Robust estimation of smooth regression and spread functions and their derivatives," *Statistica Sinica*, to appear.

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

7 APPENDIX

7.1 Bias and Variance for Local Polynomial Estimation

Here we give a brief derivation of bias and variance formulae 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 Ruppert and Wand (1994) and Fan, Heckman and Wand (1995).

A useful simplification is to let the unknown parameters be $a_j = h^j \Theta^{(j)}(z_0)/j!$, see the appendix of Fan et al. (1995).

For any $p \times q$ matrix $C = (c_1, \ldots, c_\ell)^t$, where c_j is a $q \times 1$ vector, define $\mathrm{vec}(C) = (c_1^t, \ldots, c_\ell^t)^t$. 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 $C(z_0) = E\left[\psi\left\{\widetilde{Y}, \Theta(z_0)\right\}\psi^t\left\{\widetilde{Y}, \Theta(z_0)\right\}|Z = z_0\right]$ and $B(z_0) = E\left[\chi\left\{\widetilde{Y}, \Theta(z_0)\right\}|Z = z_0\right]$, where $\chi(\widetilde{Y}, v) = (\partial/\partial v^t)\psi(\widetilde{Y}, v)$. Define

$$\mathcal{L}_n(a_0, \ldots, a_p) = n^{-1}\sum_{i=1}^{n} K_h(Z_i - z_0)\ \text{vec}\left[G_{p,h}(Z_i - z_0) \otimes \psi^t\left\{\widetilde{Y}_i, \sum_{j=0}^{p} a_j\left(Z_i - z_0\right)^j/h^j\right\}\right],$$

where $G_{p,h}(v) = (1, v/h, v^2/h^2, \ldots, v^p/h^p)^t$. We are solving $0 = \mathcal{L}_n\left(\widehat{a}_0, \ldots, \widehat{a}_p\right)$, with $\widehat{a}_j = h^j\widehat{\Theta}^{(j)}(z_0)/j!$

By a Taylor series expansion, we find that the estimates are asymptotically equivalent to

$$\begin{pmatrix} \widehat{a}_0 - a_0 \\ \vdots \\ \widehat{a}_p - a_p \end{pmatrix} \approx -\left\{B_*(z_0)\right\}^{-1}\mathcal{L}_n(a_0, \ldots, a_p), \tag{24}$$

where

$$B_*(z_0) = \frac{\partial}{\partial(a_0^t, \ldots, a_p^t)}\mathcal{L}_n(a_0, \ldots, a_p).$$

It is helpful to keep in mind the following aspect:

$$\mathcal{L}_n(a_0, \ldots, a_p) = n^{-1}\sum_{i=1}^{n} K_h(Z_i - z_0)\begin{bmatrix} \psi\left\{\widetilde{Y}_i, \sum_{j=0}^{p} a_j(Z_i - z_0)^j/h^j\right\} \\ \{(Z_i - z_0)/h\}\,\psi\left\{\widetilde{Y}_i, \sum_{j=0}^{p} a_j(Z_i - z_0)^j/h^j\right\} \\ \vdots \\ \{(Z_i - z_0)/h\}^p\,\psi\left\{\widetilde{Y}_i, \sum_{j=0}^{p} a_j(Z_i - z_0)^j/h^j\right\} \end{bmatrix}.$$

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

It is easily seen that

$$B_*(z_0) \xrightarrow{p} f_Z(z_0)\left\{D_p(\mu) \otimes B(z_0)\right\},$$

where $D_p(\mu)$ is the $(p+1) \times (p+1)$ matrix with $(j,k)^{th}$ element $\mu_K(j + k - 2)$.

It is also easily shown that

$$\text{cov}\left\{\mathcal{L}_n(a_0, \ldots a_p)\right\} \sim (nh)^{-1} f_Z(z_0)\left\{D_p(\gamma) \otimes C(z_0)\right\},$$

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

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

$$\begin{aligned} &E\mathcal{L}_n(a_0, \ldots, a_p) \\ &\quad = -\int K_h(z - z_0) f_{Y|Z}(\widetilde{y}|z) f_Z(z) \\ &\qquad \times\text{vec}\left(G_{p,h}(z - z_0) \otimes \left[\psi\left\{\widetilde{y}, \Theta(z)\right\} - \psi\left\{\widetilde{y}, \sum_{j=0}^{p} a_j(z - z_0)^j/h^j\right\}\right]^t\right) d\widetilde{y}dz \\ &\quad \approx -\int K_h(z - z_0) f_{Y|Z}(\widetilde{y}|z) f_Z(z) \\ &\qquad \times\text{vec}\left(G_{p,h}(z - z_0) \otimes \left[\chi\left\{\widetilde{y}, \Theta(z)\right\}\left\{\Theta(z) - \sum_{j=0}^{p}(z - z_0)^j\Theta^{(j)}(z_0)/j!\right\}\right]^t\right) d\widetilde{y}dz. \end{aligned}$$

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

$$E\mathcal{L}_n(a_0,\ldots,a_p) \approx A_{1h} + A_{2h},$$

where

$$\begin{aligned} A_{kh} &= -\frac{h^{p+k}}{(p+k)!}\int K(x) f_{Y|Z}(\widetilde{y}|z_0+xh) f_Z(z_0+xh) \\ &\qquad \text{vec}\left(G_{p,1}(x)\otimes\left[\chi\left\{\widetilde{y},\Theta(z_0+xh)\right\}\Theta^{(p+k)}(z_0)x^{p+k}\right]^t\right) d\widetilde{y}dx \\ &= -\frac{h^{p+k}}{(p+k)!}\int K(x) f_Z(z_0+xh)\ \text{vec}\left[G_{p,1}(x)\otimes\left\{B\left(z_0+hx\right)\Theta^{(p+k)}(z_0)x^{p+k}\right\}^t\right] dx. \end{aligned}$$

Clearly,

$$A_{2h} \approx \frac{-h^{p+2} f_Z(z_0)}{(p+2)!}\ \text{vec}\left[D_\mu(p+2)\otimes\left\{B(z_0)\Theta^{(p+2)}(z_0)\right\}^t\right],$$

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

$$\begin{aligned} A_{1h} &\approx \frac{-h^{p+1}}{(p+1)!}\int K(x)\ \text{vec}\left[x^{p+1}G_{p,1}(x)\otimes\left\{Q(z_0+hx)\Theta^{(p+1)}(z_0)\right\}^t\right] dx \\ &\approx -\frac{h^{p+1} f_Z(z_0)}{(p+1)!}\ \text{vec}\left[D_\mu(p+1)\otimes\left\{B(z_0)\Theta^{(p+1)}(z_0)\right\}^t\right] \\ &\quad -\frac{h^{p+2}}{(p+1)!}\int K(x)\ \text{vec}\left[x^{p+2}G_{p,1}(x)\otimes\left\{Q^{(1)}(z_0)\Theta^{(p+1)}(z_0)\right\}^t\right] dx \\ &\approx -\frac{h^{p+1}}{(p+1)!}\ \text{vec}\left[D_\mu(p+1)\otimes\left\{f_Z(z_0)B(z_0)\Theta^{(p+1)}(z_0)\right\}^t\right] \\ &\quad -\frac{h^{p+2}}{(p+1)!}\ \text{vec}\left[D_\mu(p+2)\otimes\left\{Q^{(1)}(z_0)\Theta^{(p+1)}(z_0)\right\}^t\right]. \end{aligned}$$

We have thus shown that asymptotically,

$$\begin{aligned} \text{bias}\left(\widehat{a}_0^t,\widehat{a}_1^t,\ldots,\widehat{a}_0^t\right)^t &= h^{p+1}\left\{D_p(\mu)\otimes B(z_0)\right\}^{-1}\ \text{vec}\left[D_\mu(p+1)\otimes\left\{B(z_0)\Theta^{(p+1)}(z_0)\right\}^t\right]/(p+1)! \\ &\quad + h^{p+2}\left\{D_p(\mu)\otimes B(z_0)\right\}^{-1} s(z_0) + \mathcal{O}(h^{p+3}), \end{aligned} \tag{25}$$

where

$$s(z_0) = \text{vec}\left[D_\mu(p+2)\otimes\left\{\frac{B(z_0)\Theta^{(p+2)}(z_0)}{(p+2)!} + \frac{Q^{(1)}(z_0)\Theta^{(p+1)}(z_0)}{f_Z(z_0)(p+1)!}\right\}^t\right].$$

The variance is

$$\{nhf_Z(z_0)\}^{-1}\left\{D_\mu(p)\otimes B(z_0)\right\}^{-1}\left\{D_\gamma(p)\otimes C(z_0)\right\}\left\{D_\mu(p)\otimes B(z_0)\right\}^{-t}\{1+o(1)\}.$$

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

$$\{D_\mu(p) \otimes B(z_0)\}^{-1} \text{ vec } \left[D_\mu(p+1) \otimes \left\{ B(z_0)\Theta^{(p+1)}(z_0) \right\}^t \right]$$

equals zero, which is clearly the case since $\mu_K(r) = 0$ 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.

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 which stems from terms ignored in the linearizing approximation (24). Since h is chosen so that the *squared* bias from smoothing is of order $(nh)^{-1}$, bias terms of order $(nh)^{-1}$ will be 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.

7.2 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 $(a_0, ..., a_p)$. Noting that $B_*(z_0)$ in (24) equals $n^{-1}B_n(z_0)$ in (7), it suffices to show that $n^{-1}C_n(z_0)$ defined in (6) has limiting covariance matrix $(nh)^{-1}f_Z(z_0)\{D_p(\gamma) \otimes C(z_0)\}$, which is easily established. This completes the argument.

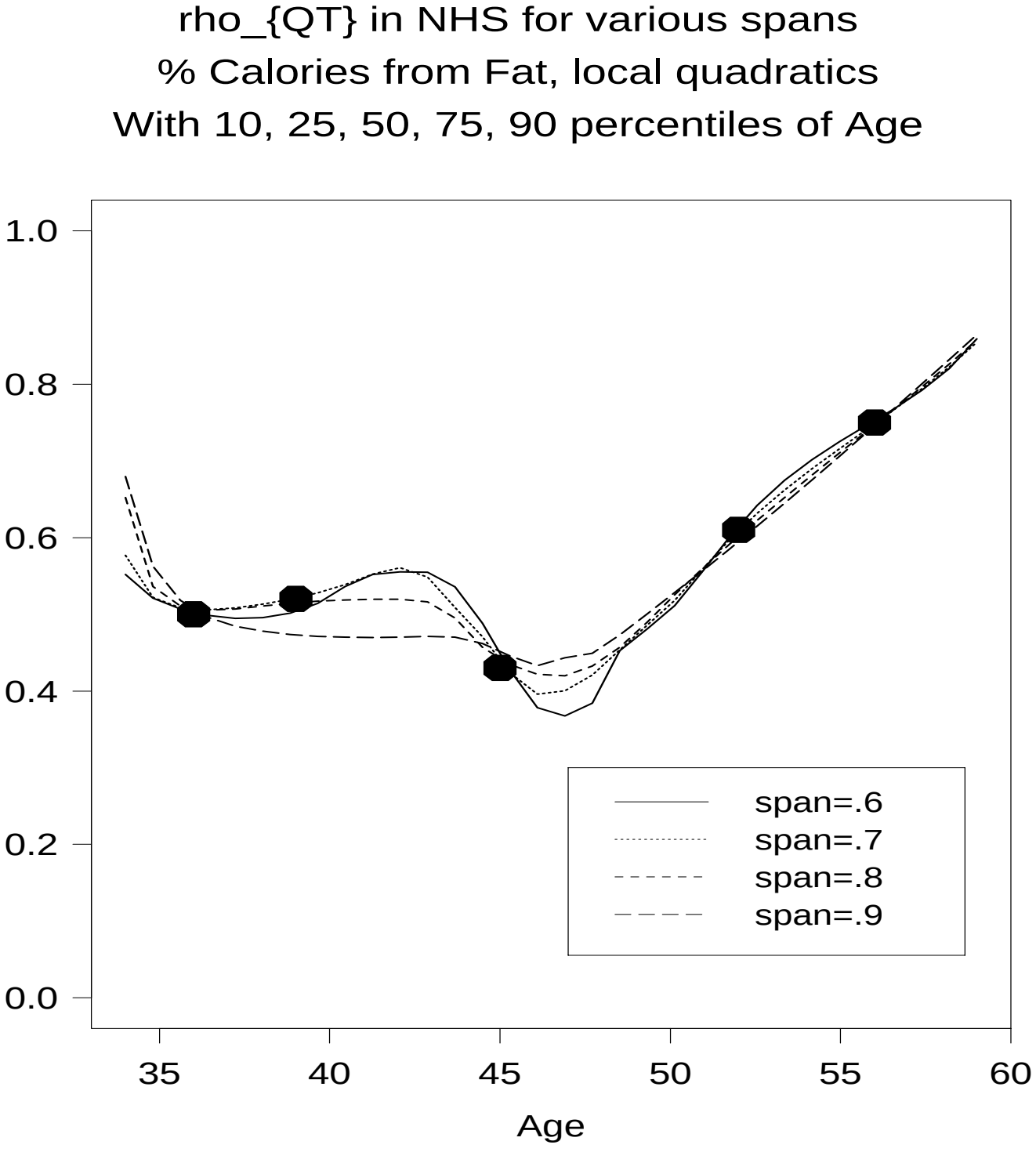

Figure 1: *Nurses' Health Study, estimating ρ_{QT}, sensitivity to the choice of span.*

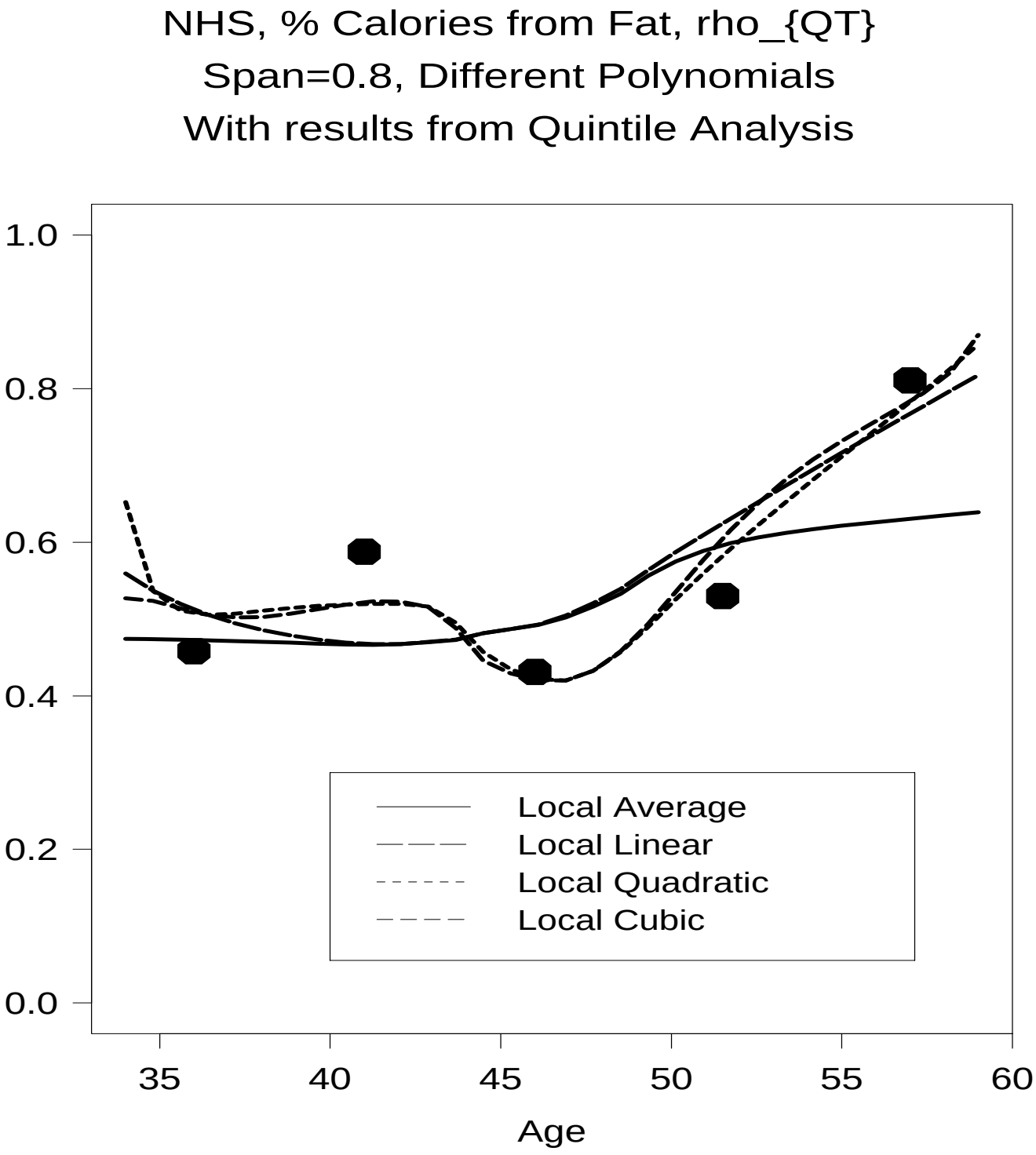

Figure 2: *Nurses' Health Study (NHS) estimating ρ_{QT} with span* 0.8 *and for local average, linear quadratic and cubic regression.*

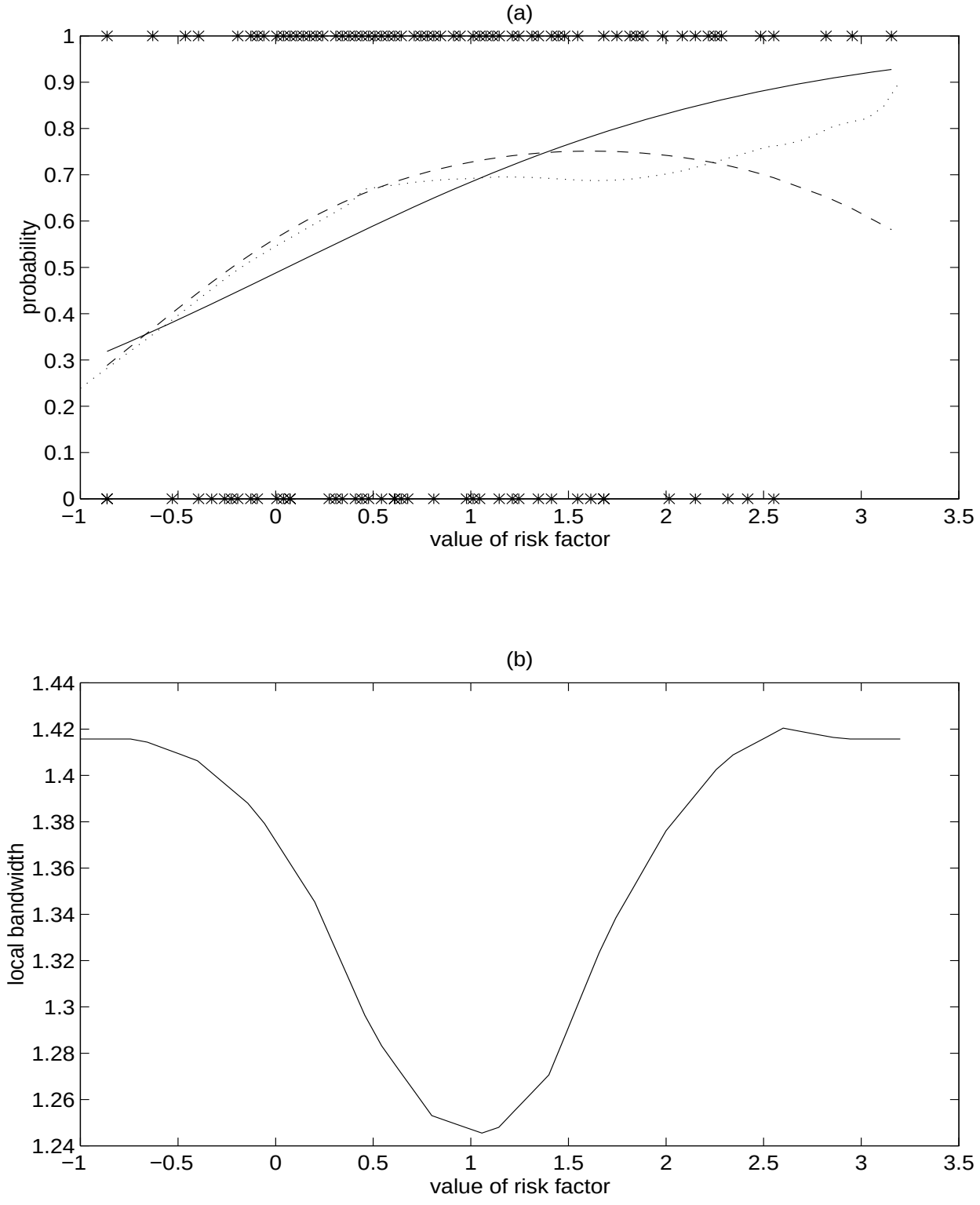

Figure 4: *Bladder cancer. (a) Linear logistic (solid), quadratic logistic (dashed), and local linear logistic (dotted) fits. (b) Local bandwidth for local logistic fit. The scale of the y-axis overemphasizes changes in the bandwidth, which, in effect, is nearly constant.*

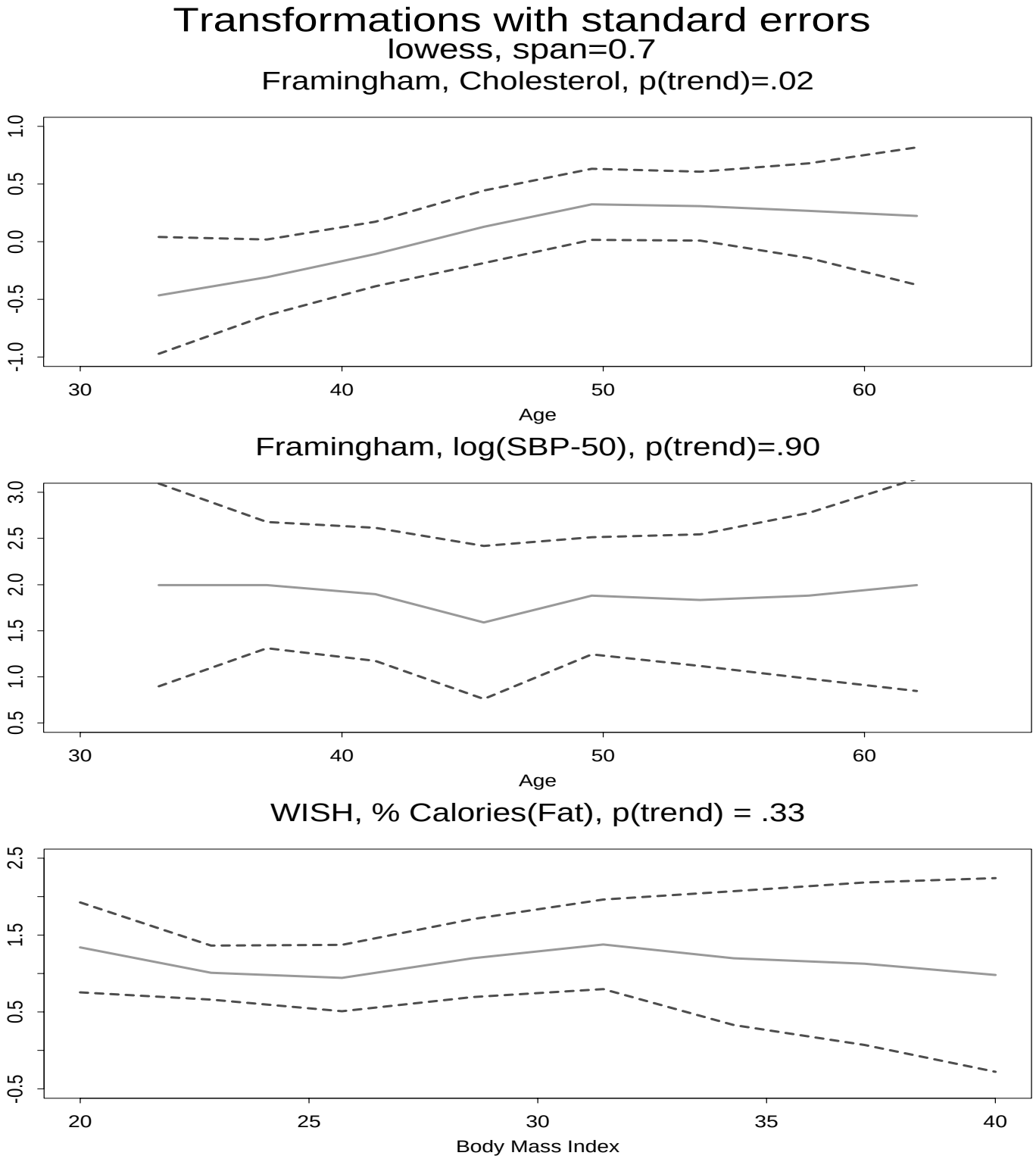

Figure 6: *Analysis of local power transformation to make mean and standard deviation of replicates locally uncorrelated. Plots of Box-Cox power transformation parameter to additivity (solid) with standard errors (dashed).*