The Annals of Statistics
1994, Vol. 22, No. 4, 1674–1700

ADAPTING FOR THE MISSING LINK[1]

BY S. WEISBERG AND A. H. WELSH

University of Minnesota and Australian National University

We consider the fitting of generalized linear models in which the link function is assumed to be unknown, and propose the following computational method: First, estimate regression coefficients using the canonical link. Then, estimate the link via a kernel smoother, treating the direction in the predictor space determined by the regression coefficients as known. Then reestimate the direction using the estimated link and alternate between these two steps. We show that under fairly general conditions, $n^{1/2}$-consistent estimates of the direction are obtained. A small Monte Carlo study is presented.

1. Introduction. In the usual generalized linear model we observe a response Y which, given the value of a vector covariate X, satisfies

$$\begin{aligned} E(Y \mid X = x) &= g(x^T\beta_0) \\ \operatorname{Var}(Y \mid X = x) &= \sigma^2 V\Big(g(x^T\beta_0)\Big), \end{aligned} \tag{1.1}$$

where β_0 and $\sigma > 0$ are vector and scalar unknown parameters respectively, and g and V are real functions. The function V is called the variance function and is typically treated as known. This is appropriate when the response has a known two parameter exponential family distribution but the model applies more generally to problems involving heteroscedasticity which is a known function of the mean. The function g, called the link function, is usually assumed to be smooth, monotone and hence invertible. In these circumstances, it is more usual to call g^{-1} the link and g the inverse link but this terminology is untenable when g is not invertible. The link is typically treated as known but our interest is in the case of an unknown or missing link. When the link is known, we take β_0 and X to be of dimension $p+1$, including p predictors and an intercept. When the link is unknown, both the magnitude of β_0 and an intercept can be absorbed into g. In this case, X and β_0 will both be of dimension p excluding the intercept. For uniqueness in the unknown link case we may standardise so that $\|\beta_0\| = 1$. We regard σ as a nuisance parameter and concentrate on estimating β_0 and g from n independent observations (Y_i, X_i) on the model (1.1).

The class of generalized linear models (1.1) is quite rich because of the flexibility in the specification of the response distribution and the choice of g. It

Received February 1992; revised October 1993.

[1] This work was substantially completed when the first author was visiting at the Centre for Mathematics and Its Applications, Australian National University, with additional support from NSF Grant 9208819.

AMS 1991 subject *classifications.* 62J12, 62G07.

Key words and phrases. Generalized linear model, kernel smoothing, link function, nonparametric regression, single index models.

includes a variety of regression models including log-linear models for contingency tables, binary response and survival data models, and the single index models of Stoker (1986). See McCullagh and Nelder (1989) for details.

There are at least two good reasons for trying to estimate g. Firstly, an estimate of$g$ may be used to construct diagnostics to study the choice of a particular g by examining how well the resulting model fits the data. Such diagnostics are important in practice as well as in principle: Lord (1980), Wainer (1983) and Azzalini, Bowman and Härdle (1989) have all given examples of real data with binomially distributed responses in which the usual logistic link is inadequate. Secondly, if the chosen function is inadequate, we may obtain a better fit to the data by using a data-driven or estimated function instead, while retaining as much of the structure of the generalized linear model as possible. A fully nonparametric approach requires p-dimensionallocal fitting, which in turn requires extremely large sample sizes, even for modest$p$. In contrast, the method described here uses only a one dimensional smoother.

Ifg were known, we could estimate β_0 by solving

$$0 = n^{-1}\sum_{i=1}^{n} \frac{g'(X_i^T\beta)}{V\left(g(X_i^T\beta)\right)} X_i\left[Y_i - g(X_i^T\beta)\right] \tag{1.2}$$

for β, where in this equation and others for a fixed link g we assume that an intercept term is included in the linear predictor, and β is not scaled. These are the likelihood equations when the response distribution is a two-parameter exponential family distribution. An iteratively reweighted least squares algorithm is widely used for solving this system of equations. Given a current estimate $\tilde{\beta}$, compute

$$\hat{\beta} = \tilde{\beta} + \hat{A}^* n^{-1}\sum_{i=1}^{n} \frac{g'(X_i^T\tilde{\beta})}{V\left(g(X_i^T\tilde{\beta})\right)} X_i\left\{Y_i - g(X_i^T\tilde{\beta})\right\}, \tag{1.3}$$

where $\hat{A}^{*\,-\,1}$ is minus the expected information evaluated at $\tilde{\beta}$,

$$\hat{A}^{*\,-\,1} = n^{-1}\sum_{i=1}^{n} \frac{g'(X_i^T\tilde{\beta})^2}{V\left(g(X_i^T\tilde{\beta})\right)} X_i X_i^T. \tag{1.4}$$

Then, set $\tilde{\beta} = \hat{\beta}$ and iterate to convergence. This iteratively reweighted least squares algorithm solves (1.2) and is identical to the Fisher method of scoring version of the Newton–Raphson algorithm for solving these likelihood equations. When$g$ is the canonical link for a generalized linear model, this algorithm converges very quickly, even for poor starting values.

For the remainder of this article we assume that the link function g is unknown and must be estimated. As pointed out previously, the intercept can be absorbed into g, and so we now view X as a p-vector of covariates, and β is a p-vector of regression coefficients that determine a direction in p-dimensional space. Given an initial estimator $\tilde{\beta}$ of the direction, which may be the result of estimating β_0 with some fixed g, we can easily compute a nonparametric estimate $\widehat{g}$ of g. We use a kernel estimate $\widehat{g}(x, \beta)$ of g, where

$$\widehat{g}(x,\beta) = (nh)^{-1}\sum_{j=1}^{n}\frac{Y_j K_h(x - X_j^T\beta)}{\widehat{\gamma}(x,\beta)}. \tag{1.5}$$

The normalizing constant $\widehat{\gamma}(x, \beta)$ is a density estimate,

$$\widehat{\gamma}(x,\beta) = (nh)^{-1}\sum_{j=1}^{n} K_h(x - X_j^T\beta), \tag{1.6}$$

with $K_h(x) = K(x/h)$ for an appropriate kernel function K, with bandwidth h. To estimate β_0 with the estimated $\widehat{g}$, we also need to estimate g'. In the proofs that follow, we use the estimator $\widehat{g}'(x, \beta)$ given by

$$\widehat{g}'(x,\beta) = (nh^2)^{-1}\sum_{j=1}^{n}\frac{1}{\widehat{\gamma}(x,\beta)}\left\{Y_j L_h(x - X_j^T\beta) - \widehat{g}(x,\beta) L_h(x - X_j^T\beta)\right\}, \tag{1.7}$$

where $L_h(x) = L(x/h)$, for an appropriate kernel function L. It is not necessary to use the same h in both K and L; doing so affords some slight notational simplification. When $L = K'$, $\widehat{g}'$ equals the derivative of $\widehat{g}$ but the proofs allow the flexibility of other choices of L.

We then propose the use of an alternating algorithm, first estimating the direction determined by β, and then the link function g, repeating until some criterion is met. Given $\widehat{g}$ and $\widehat{g}'$, we use the scoring algorithm (1.3) to estimate β_0; given the estimate of β_0, we use the kernel estimate (1.5) and (1.6) to get a new estimate of the link g. It is not hard to implement this algorithm because it involves only univariate smoothing; an implementation using Lisp-Stat [Tierney (1990)] is described in Section 3. It seems to be only slightly more complicated than estimating a parametric class of link functions as in Pregibon (1981, 1982), and less complicated than other nonparametric alternatives. Unlike a fully nonparametric approach, much of the flavor of the parametric fit of the generalized linear model is retained.

Recently, the problem of fitting models like (1.1) with the link function assumed unknown has received considerable attention. Li and Duan (1989), for example, have shown the solution to (1.2) with g incorrectly assumed to be a canonical link will give a consistent estimate of $k\beta_0$ for some k, if the distribution of the x's is sufficiently nice. (Here β_0 excludes the intercept.) What is needed for the Li–Duan result to hold is that the conditional expectation $\mu(v) = E(X \mid X^T\beta_0 = v)$ must be linear in v. This happens for all β_0 if and only if X has an elliptically symmetric distribution [Eaton (1986)]. When the distribution of X is not elliptically symmetric, the condition may hold for particular β_0, and hence the estimate of β_0 from fitting the wrong link may still be consistent. The multiplier k will depend on the true link function, and will be 0 if the true link function is symmetric [Cook and Weisberg (1991)]; this is a case that is likely to be difficult or impossible to fit by any method.

By virtue of the results in Li and Duan (1989), any global method of fitting will fail when the link is incorrectly specified and the distribution of the predictors is sufficiently perverse. For this case, methods based on the local behavior of the regression surface have been shown to give consistent estimates regardless of the distribution of the X's. Härdle and Stoker (1989) point out that β_0 is the normed expected derivative of g and so construct a $n^{1/2}$-consistent nonparametric estimator of β_0 using p-dimensional smoothing. Härdle, Hall and Ichimura (1993) consider a similar problem to ours but minimize a least squares like criterion to estimate both β_0 and h simultaneously. Different approaches using splines and monotone splines have been implemented by Yandell and Green (1986) and Ramsay and Abrahamowicz (1989), respectively. The method proposed here is a combination of global fitting, using (1.3), and local fitting, using (1.4) and (1.5). We show that this method has the same theoretical properties as the local methods proposed in the above references.

In Section 2, we discuss the theoretical aspects of our results, and derive the properties of the estimation method; the proofs are detailed and require a number of preliminary results, included in an Appendix. In Section 3, we discuss the practical problems of implementation of the method, and give the results of a small Monte Carlo experiment. Section 4 contains discussion and further applications.

2. Results. We shall now give the asymptotic behavior of the estimates obtained from the use of the algorithm described in Section 1. We shall consider both one-step estimates and fully iterated ones. One-step estimates are shown in Corollary 1 to be $n^{1/2}$-consistent if the initial estimator of β is $n^{1/2}$-consistent. We then show in Corollaries 2 and 3 that under appropriate conditions, the fully iterated solution to (1.3) does not depend on the starting values, and the estimate of β so obtained will be $n^{1/2}$-consistent. Our results require eight simple conditions which we denote by C, in addition to model (1.1). The first seven conditions are:

(i) $(Y_i, X_i), 1 \le i \le n$, are independent realizations of (Y,X) where (Y,X) have all moments finite.

(ii) The density f of X has three bounded derivatives and the density γ of $X^T\beta_0$ has four bounded derivatives.

(iii) g has four bounded, continuous derivatives on s, the support of $X^T\beta_0$.

(iv) $V > 0$, bounded and has two bounded derivatives on $\{u\colon u = g(t),\ t \in s\}$.

(v) $A^{-1} = E[\{X - \mu(X^T\beta_0)\}\{X - \mu(X^T\beta_0)\}^T g'(X^T\beta_0)^2 V\{g(X^T\beta_0)\}^{-1}]$, where $\mu(v) = E(X \mid X^T\beta_0 = v)$ is positive definite.

(vi) K, L have compact support, bounded derivatives and satisfy

$$\int K(z)\,dz = 1, \qquad \int z^i K(z)\,dz = 0, \qquad i = 1,2, \quad \text{and} \quad \int |z|^3 K(z)\,dz < \infty,$$

$$\int L(z)\,dz = 0, \qquad \int zL(z)\,dz = -1, \qquad \int z^i L(z)\,dz = 0, \qquad i = 2,3,$$

$$\text{and} \quad \int z^4 L(z)\,dz < \infty.$$

(vii) The bandwidth $h = O(n^{-\tau})$, $\frac{1}{6} < \tau < \frac{1}{4}$.

Our arguments require a large number but not all finite moments. There is no particularly neat way to specify the precise finite number we need, so we adopt the simpler approach of assuming that all moments are finite. This is reasonable because the distribution of $Y|X$ would often be assumed to be in or close to the exponential family. We then require smoothness conditions on the underlying functions and conditions on the kernels. Conditions (ii) and (iii) ensure that on appropriate sets, $\mu(v) = E\{X|X^T\beta_0 = v\}$ has three bounded derivatives. Of course, if X has compact support, the boundedness requirements are trivially satisfied under the smoothness conditions.

The kernel conditions are chosen for simplicity. It is not necessary for the kernels to have compact support but his greatly simplifies our arguments. When estimating g, we can use a standard second-order kernel, so K can be a symmetric density function. However, when we estimate β_0, we need to control the bias contributed by the estimates of g and g'. In particular, to control the variability of our estimate of g', we require $n^{1/2}h^2 \to \infty$. If we use second-order kernels, controlling the bias of the estimator of β_0 requires $h^2 = o(n^{-1/2})$ which is incompatible with the previous condition. However, with third-order kernels, the bias requires $h^3 = o(n^{-1/2})$ and the requirements are now compatible. This motivates both the use of third-order kernels and the conditions on h; see Prakasa Rao (1983) for a general definition, and Gasser, Müller and Mammitzsch (1985) for examples. If we had to estimate g but not g', we would only require $h = o(n^{-1/2})$ which is compatible with the use of second-order kernels. To permit elementary arguments, we assume that K and L have compact support and bounded derivatives.

A pointwise central limit theorem for $\hat{g}$ (with a second-order kernel) is proved by Härdle and Stoker (1989). The result shows that the optimal rate of the convergence is attained.

THEOREM 1 [Härdle and Stoker (1989)]. *Suppose conditions* (i) *to* (iii) *hold, and that K is a twice-differentiable symmetric density function with $h \sim n^{-1/5}$. Then for any u such that $\gamma(u) > 0$ and $\tilde{\beta}$ such that $\tilde{\beta} - \beta_0 = O_p(n^{-1/2})$,*

$$n^{2/5}\big(\hat{g}(u,\tilde{\beta}) - g(u)\big) \xrightarrow{d} N\bigg(\big[\tfrac{1}{2}g''(u) + g'(u)\gamma'(u)\gamma(u)^{-1}\big] \times \int z^2K(z)\,dz, V\big(g(u)\big)\gamma(u)^{-1}\int K(z)^2\,dz\bigg).$$

Theorem 1 also holds for an appropriate third-order kernel but the asymptotic bias vanishes. This is unsurprising because with $h \sim n^{-1/5}$ the variance dominates the bias. The optimal choice is actually $h \sim n^{-1/7}$ which leads to the faster rate $n^{3/7}$. Of course, in this case, the expression for the asymptotic bias changes, too.

To facilitate the proof of our major result in Theorem 2, we shall use two further modifications before we incorporate $\widehat{g}$ and g' into the iteratively reweighted least squares algorithm (1.3). First, to ensure that our estimator of β_0 is to first order unaffected by the estimator of g, we center the covariates where they appear as vectors rather than in inner products in the iteratively reweighted least squares algorithm. It turns out that we need to center X_i about $\widehat{\mu}(X_i, \widetilde{\beta})$, where

$$\widehat{\mu}(x, \beta) = (nh)^{-1} \sum_{k=1}^{n} \frac{1}{\widehat{\gamma}(x^T\beta, \beta)} X_k K_h\{(x - X_k)^T\beta\}.$$

We will show in the Appendix that $\widehat{\mu}(x, \beta)$ is an estimator of the p-vector $\mu(x^T\beta_0)$, where $\mu(v) = E\{X \mid X^T\beta_0 = v\}$. Global centering about $E(X)$ leads to additional complications, and since it is unnecessary we will not use it here.

Next, kernel estimates of g, g' and μ are unstable at points for which the denominator $\widehat{\gamma}$ is small, so we exclude from the estimator of β_0 observations for which $\widehat{\gamma} < a$, for $a \downarrow 0$ such that $h \ll a$. This is a common modification to estimators which are functions of kernel regression estimators; see, for example, Härdle and Stoker (1989). While this is usually done by incorporating an appropriate indicator function into the estimator, the discontinuity in the indicator function causes technical difficulties. A simpler approach is to use a smooth version of the indicator function. We use

$$J_a(x) = \begin{cases} 0, & \text{for } x < a, \\ J\{2(x-a)/\theta - 1\}, & \text{for } a \le x < a + \theta, \\ 1, & \text{for } a + \theta \le x \end{cases}$$

for some small fixed $\theta > 0$, where

$$J(x) = \frac{15}{16}\left(\frac{1}{5}x^5 - \frac{2}{3}x^3 + x + \frac{8}{15}\right) \quad \text{for } -1 \le x < 1,$$

but any indicator-like function with two bounded derivatives can be used. The tuning constant a is required to decrease to 0 more slowly than h, which is expressed by:

(viii) $a = O((\log n)^{-1})$.

While it is possible to express this in terms of an algebraic rate of convergence, it is simpler and more convenient to adopt the logarithmic rate.

The modified iteratively reweighted least squares algorithm is then to compute

$$\widehat{\beta} = \widetilde{\beta} + \widehat{A}n^{-1} \sum_{i=1}^{n} J_a\left\{\widehat{\gamma}(X_i^T\widetilde{\beta}, \widetilde{\beta})\right\} \frac{\widehat{g}'(X_i^T\widetilde{\beta}, \widetilde{\beta})}{V\left(\widehat{g}(X_i^T\widetilde{\beta}, \widetilde{\beta})\right)} \times \{X_i - \widehat{\mu}(X_i, \widetilde{\beta})\}\left\{Y_i - \widehat{g}(X_i^T\widetilde{\beta}, \widetilde{\beta})\right\}, \tag{2.1}$$

where

$$\widehat{A}^{-1} = n^{-1}\sum_{i=1}^{n} J_a\left\{\widehat{\gamma}(X_i^T\widetilde{\beta}, \widetilde{\beta})\right\}\frac{\widehat{g}'(X_i^T\widetilde{\beta}, \widetilde{\beta})^2}{V\left(\widehat{g}(X_i^T\widetilde{\beta}, \widetilde{\beta})\right)}$$
$$\times \{X_i - \widehat{\mu}(X_i, \widetilde{\beta})\}\{X_i - \widehat{\mu}(X_i, \widetilde{\beta})\}^T.$$

Then, set $\widetilde{\beta} = \widehat{\beta}$ and iterate to convergence. We can either use a few steps of the algorithm or we can iterate to convergence to decrease the impact of the initial estimate of β_0. In either case, our interest is in the estimates of both g and β_0. Our results show that the estimate of g achieves the optimal rate of convergence, and the estimate of β_0 is $n^{1/2}$-consistent and adaptive in the sense that, to first order, its asymptotic distribution does not depend on the fact that g is estimated rather than known.

Our major result is the following stochastic equicontinuity result from which we deduce the properties of two estimators of β_0.

THEOREM 2. *Suppose conditions C hold. Then for all fixed $B \in (0, \infty)$, as $n \to \infty$,*

$$o_p(1) = \sup_{|\beta - \beta_0| \le n^{-1/2}B} \Bigg| n^{-1/2}\sum_{i=1}^{n}\{X_i - \widehat{\mu}(X_i, \beta)\}\left\{Y_i - \widehat{g}(X_i^T\beta, \beta)\right\}$$
$$\times J_a\left\{\widehat{\gamma}(X_i^T\beta, \beta)\right\}\frac{\widehat{g}'(X_i^T\beta, \beta)}{V\left\{\widehat{g}(X_i^T\beta, \beta)\right\}}$$
$$- n^{-1/2}\sum_{i=1}^{n}\left\{X_i - \mu(X_i^T\beta_0)\right\}\left\{Y_i - g(X_i^T\beta_0)\right\}$$
$$\times J_a\left\{\gamma(X_i^T\beta_0)\right\}\frac{g'(x_i^T\beta_0)}{V\left\{g(X_i^T\beta_0)\right\}} + A^{-1}n^{1/2}(\beta - \beta_0)\Bigg|$$

and

$$o_p(1) = \sup_{|\beta - \beta_0| \le n^{-1/2}B} \Bigg| n^{-1}\sum_{i=1}^{n}\{X_i - \widehat{\mu}(X_i, \beta)\}\{X_i - \widehat{\mu}(X_i, \beta)\}^T$$
$$\times J_a\left\{\widehat{\gamma}(X_i^T\beta, \beta)\right\}\frac{\widehat{g}'(X_i^T\beta, \beta)^2}{V\left\{\widehat{g}(X_i^T\beta, \beta)\right\}} - A^{-1}\Bigg|.$$

The proof of this theorem is contained in the Appendix.

We are now able to prove three important corollaries. These results apply to the unnormalized estimators; it is important for interpretation to realize that the magnitude of the estimators is arbitrary but it is not essential to normalize the estimators to have norm 1. The first describes the behavior of one step of the modified iteratively reweighted least squares algorithm (2.1). As with all one-step estimates, we require a good initial estimator of β_0. A natural candidate initial estimator is that obtained by solving (1.2) with a fixed g, as suggested by the Li and Duan (1989) results, at least for X elliptically symmetric. Of course, from a diagnostic point of view, under the hypothesis that the specified g is the true g, the estimator is $n^{1/2}$-consistent and diagnostics can be based on the one-step modification.

COROLLARY 1. *Suppose conditions C hold and that there is an initial estimator $\tilde{\beta}$ such that $\tilde{\beta} - \beta_0 = O_p(n^{-1/2})$. Then*

$$n^{1/2}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, A).$$

The second corollary describes the behavior of a root of the system of equations

$$(2.2)\qquad \begin{aligned} 0 = \sum_{i=1}^{n} \{X_i - \hat{\mu}(X_i, \beta)\}\left\{Y_i - \hat{g}(X_i^T\beta, \beta)\right\} \\ \times J_a\left\{\hat{\gamma}(X_i^T\beta, \beta)\right\}\frac{\hat{g}'(X_i^T\beta, \beta)}{V\left\{\hat{g}(X_i^T\beta, \beta)\right\}}, \end{aligned}$$

which corresponds to fully iterating the modified iteratively reweighted least squares algorithm (2.1). The statement of the corollary reflects the fact that there may be multiple roots.

COROLLARY 2. *Suppose conditions C hold. Then a solution $\bar{\beta}$ to the set of equations* (2.2) *which satisfies $\beta - \beta_0 = O_p(n^{-1/2})$ exists in probability. Moreover,*

$$n^{1/2}(\bar{\beta} - \beta_0) \xrightarrow{d} N(0, A).$$

PROOF. Since the right-hand side of (2.2) is continuous in β, the first part of the result will follow from (6.3.4) of Ortega and Rheinboldt (1973), page 163, if we can show that in probability

$$\begin{aligned} \sum_{i=1}^{n}(\beta - \beta_0)^T \{X_i - \hat{\mu}(X_i, \beta)\}\left\{Y_i - \hat{g}(X_i^T\beta, \beta)\right\} \\ \times J_a\left\{\hat{\gamma}(X_i^T\beta, \beta)\right\}\frac{\hat{g}'(X_i^T\beta, \beta)}{V\left\{\hat{g}(X_i^T\beta, \beta)\right\}} < 0 \end{aligned}$$

for $|\beta - \beta_0| = n^{-1/2}B$, for some $B < \infty$. The argument uses Theorem 2 to approximate the left-hand side of the inequality by

$$\begin{aligned} \sum_{i=1}^{n}(\beta - \beta_0)^T\left\{X_i - \mu(X_i^T\beta_0)\right\}\left\{Y_i - g(X_i^T\beta_0)\right\}J_a\left\{\gamma(X_i^T\beta_0)\right\}\frac{g'(X_i^T\beta_0)}{V\left\{g(X_i^T\beta_0)\right\}} \\ - n(\beta - \beta_0)^T A^{-1}(\beta - \beta_0). \end{aligned}$$

The first term is like B times a random variable which is bounded in probability and the second is like a constant times B^2 so the desired inequality holds for B large enough. The details of the argument are given in the proof of Theorem 5.1 of Welsh (1989) so are omitted. The second part of the result follows by applying Corollary 1 with $\tilde{\beta} = \bar{\beta}$. □

In practice, we need not center the X's so we solve

$$0 = \sum_{i=1}^{n} X_i\left\{Y_i - \hat{g}(X_i^T\beta, \beta)\right\} J_a\left\{\hat{\gamma}(X_i^T\beta, \beta)\right\} \frac{\hat{g}'(X_i^T\beta, \beta)}{V\left\{\hat{g}(X_i^T\beta, \beta)\right\}} \tag{2.3}$$

instead of (2.2). As shown in the Appendix, the conclusion of Theorem 2 should be replaced by

$$\begin{aligned} o_p(1) = \sup_{|\beta - \beta_0| \le n^{-1/2}B} \Bigg| & n^{-1/2} \sum_{i=1}^{n} X_i\left\{Y_i - \hat{g}(X_i^T\beta, \beta)\right\} \\ & \times J_a\left\{\hat{\gamma}(X_i^T\beta, \beta)\right\} \frac{\hat{g}'(X_i^T\beta, \beta)}{V\left\{\hat{g}(X_i^T\beta, \beta)\right\}} \\ & - n^{-1/2} \sum_{i=1}^{n} \left\{X_i + \mu(X_i^T\beta_0)\right\}\left\{Y_i - g(X_i^T\beta_0)\right\} \\ & \times J_a\left\{\gamma(X_i^T\beta_0)\right\} \frac{g'(X_i^T\beta_0)}{V\left\{g(X_i^T\beta_0)\right\}} + A^{-1}n^{1/2}(\beta - \beta_0)\Bigg|. \end{aligned}$$

Arguing as in the proof of Corollary 2, we obtain the following result.

COROLLARY 3. *Suppose conditions C hold. Then a solution $\bar{\beta}$ to the set of equations* (2.3) *which satisfies $\bar{\beta} - \beta_0 = O_p(n^{-1/2})$ exists in probability. Moreover*,

$$n^{1/2}(\bar{\beta} - \beta_0) \xrightarrow{d} N(0, AB^*A),$$

where

$$B^* = E\left\{X + \mu(X^T\beta_0)\right\}\left\{X + \mu(X^T\beta_0)\right\}^T \frac{g'(X^T\beta_0)^2}{V\left(g(X^T\beta_0)\right)}.$$

The corollaries show that it is possible to adapt for β_0 when g is unknown in the sense that the asymptotic distributions of our estimators are the same as when g is known. This is a useful property because it simplifies the use of existing software to incorporate the estimated g. The second part of Theorem 2 shows that we can in fact estimate the asymptotic variance and so make asymptotically valid inferences.

3. Implementation and simulation.

3.1. *Implementation.* Application of the methodology described here requires software for the scoring algorithm, for fitting a kernel smoother, and then using the result of the smooth to estimate g. The standard package GLIM cannot be used for this purpose for two reasons. First, GLIM uses g^{-1} as the link function, and requires that g be monotone and invertible. As is evident from (1.3), it is unnecesary to invert g to use the scoring algorithm. Second, GLIM does not provide a way to do the smoothing.

The system Lisp-Stat with the "glim prototype" [Tierney (1990, 1991)] provides a very congenial setting for the computations that need to be done here. Our algorithm works as follows:

(a) Specify a generalized linear model using the glim prototype with the canonical link, and fit that model to get an estimated vector of regression parameters $\widetilde{\beta}$, including an intercept.

(b) Estimate the link g by smoothing the two-dimensional plot of $\widetilde{\eta} = X\widetilde{\beta}$ versus the response Y. The intercept can be included in $\widetilde{\eta}$, or it can be ignored since only the direction, not the magnitude, is important. The link function g and its derivative g' are estimated from (1.6) and (1.7), respectively, over a grid covering the range of $\widehat{\eta}$. Lacking optimality results, the bandwidth h is chosen visually using a slide bar to vary h to obtain an estimate $\widehat{g}$ that matches the data, and $\widehat{g}'$ that is fairly smooth.

(c) Reestimate β via the scoring algorithm, using the values of $\widehat{g}$ and $\widehat{g}'$ computed by interpolation over the grid obtained in (b). Points whose estimated linear predictor are outside the range of $\widehat{g}$ will have $\widehat{g}' = 0$, and will be effectively ignored in the refitting, roughly corresponding to removing points whose fitted density is too small. In Lisp-Stat the fitting can be done using the "newtonmax" function, which maximizes a function via the Newton–Raphson procedure. This function can be used for the scoring algorithm by explicitly computing the gradient and $\widehat{A}$, rather than letting the function compute numerical estimates of the gradient and Hessian. We do not normalize $\widehat{\beta}$ to length 1, although this could be done after the iteration is finished.

(d) Alternate between (b) and (c) until a stopping criterion is met.

3.2. *Example.* We shall use the wool data from Box and Cox (1964) to illustrate the methodology. This is a three-by-three-by-three experiment with three quantitative factors, specimen length, amplitude of loading cycle and load, with the levels of each factor equally spaced. The response is the number of cycles to failure of a sample of wool. Figure 1 is a plot of $\{\widehat{\eta}, y\}$ for the fit with normal errors and the identity link, using a first-order model. Also shown on Figure 1 is the estimated $\widehat{g}$. The value of the smoothing parameter h can be chosen visually

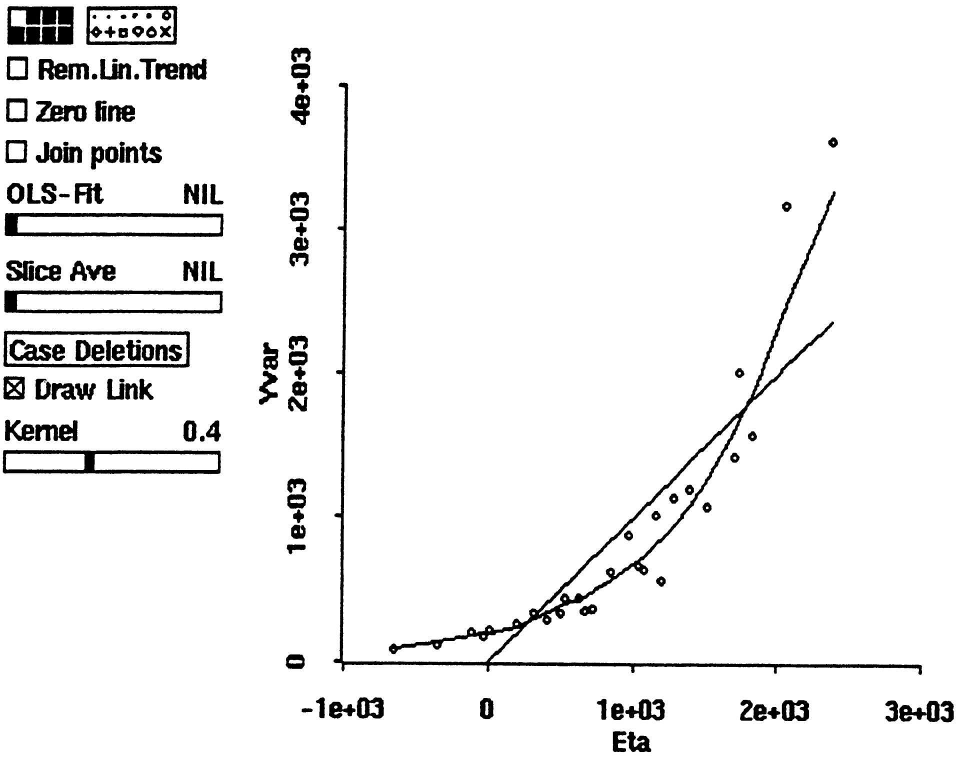

□ Rem.Lin.Trend
□ Zero line
□ Join points
OLS-Fit NIL
Slice Ave NIL
Case Deletions
☒ Draw Link
Kernel 0.4

FIG. 1. *Plot of* $\{\widehat{\eta}, y\}$ *for the wool data.*

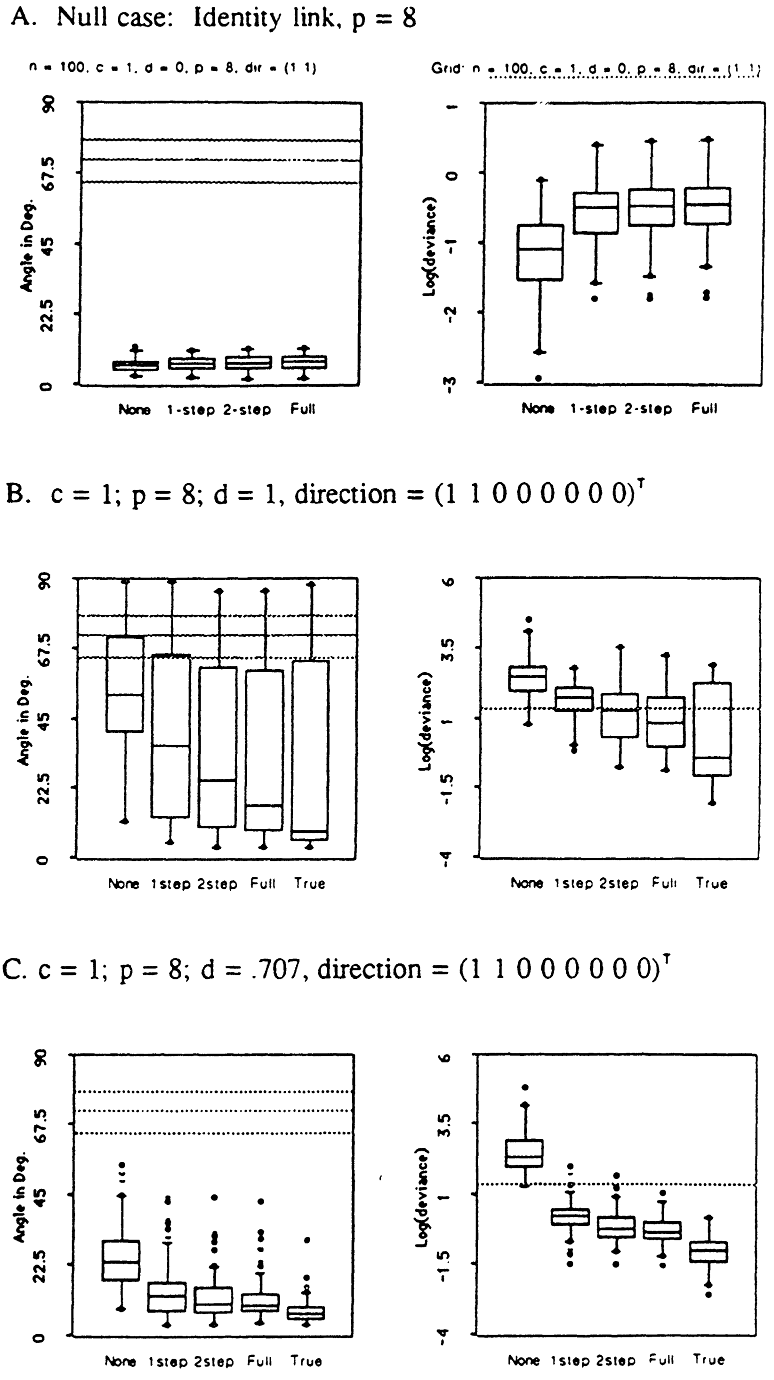
using the slide bar at the left of the plot. One could also view a second graph that plots $\{\widehat{\eta}, \widehat{g}'\}$ as h is varied, or superimpose a scaled version of this curve on Figure 1. Given $\widehat{g}$, coefficients can be reestimated. Fitting with the estimated $\widehat{g}$ reduces the deviance from 5.5 million to 0.9 million. The angle between the fitted direction from the canonical fit and the fit using $\widehat{g}$ is only about 1°, so in this example it is clear that little will be gained by further iteration, as the estimate of g will not change. If we then fit a larger model that includes interactions between the three predictors using $\widehat{g}$ as if it were the true link, the deviance is reduced by about 40,000, or about 4%, a relatively small reduction. Using the canonical link, fitting these three interactions reduces the deviance by about 3.4 million or about 62%, suggesting that the interactions are indeed very important.

Estimating the link function is not without cost. Since a scale and location factor can be absorbed into the link, the estimated coefficient vector $\widehat{\beta}$ estimates a direction in p-dimensional space, but the magnitudes of individual coefficients in $\widehat{\beta}$ cannot be related directly to rates of change in the predictors. Ratios of coefficient estimates, however, are still meaningful so we could take the point of view of Li and Duan (1989) and examine ratios of estimates. Alternatively, we could adopt a graphical solution (in which case the magnitude of $\widehat{\beta}$ is unimportant) or, after estimating g, treat $\widehat{g}$ as fixed and then estimate β and its length. Finally, we can view $\widehat{g}$ simply as a powerful diagnostic for exploring the empirical validity of fixed links and for suggesting alternative links.

An alternative approach to the wool data problem is to transform y using the methodology of Box and Cox (1964). This approach will in this example lead to essentially the same answer as we obtain. The approach suggested here is more general, however, because it does not require a monotone link function and does not restrict g to a specific class of functions indexed by a small number of parameters.

3.3. *Simulation.* To illustrate the methodology, we did a small simulation. The built-in Lisp-Stat kernel-smooth function provides only second-order kernels, so we used the second-order quartic kernel in the simulations. One can only expect that using a third-order kernel would give even better results. To automate bandwidth selection, we chose h via cross validation, as suggested by Härdle (1990), page 159, using the quasi-likelihood implicit in (1.1) to get the measure of fit. In the simulation, the derivative $\hat{g}'$ was obtained by numerical differentiation, effectively setting $L = K'$.

We used an example similar to one used by Härdle, Hall and Ichimura (1993). We consider n observations on p predictors $X_1, \dots, X_p$, such that each X_j is independently distributed according to an exponential distribution with mean 0.5. For the link function, we take a quadratic, $g(\eta, c, d) = c(1 + (\eta - d)^2)$. The conditional distribution of $Y \mid \{(X_1, \dots, X_p) = x\}$ is taken to be normal with mean $g(x^T\beta, c, d)$ and standard deviation 0.2. The example in Härdle, Hall and Ichimura is limited to the case $p = 2$, and sets the predictors to be uniform. This example was chosen for several reasons. First, by choosing the predictors to be independent, the correct direction vector is clearly defined. Second, since the distributions of the x's is not elliptical, the Li–Duan theorem will not apply for all β; however, it does apply for some β, so we can compare cases in which fitting the wrong link gives a consistent estimate of β to cases in which the estimate is not consistent. In the simulations reported below, we have set $p = 8$ and $c = 1$; qualitatively similar results are obtained for smaller dimension p and for higher signal-to-noise ratio (larger values of c). All simulations are based on 100 replications, each with $n = 100$ observations. On each replication, two smmary statistics are reported: the angle between the estimate of β and its true value (thus ignoring the intercept and the scale factor that depends on the link), and the squared difference between the final fitted values $\widehat{g}(x^T\widehat{\beta})$ and the true values of $g(x^T\beta, c, d)$. This latter statistic is designed to give a measure of agreement between the g and its estimate. These were computed for the initial fit of the canonical link, for a one-step procedure of estimating the link once and then reestimating β once, for a two-step estimate, for a fully iterated estimate and finally for a fit using the true link function. The number of iterations required for a convergence criterion to be met rarely exceeded 4. All results here are summarized by parallel boxplots, with the goodness of fit statistic in log scale.

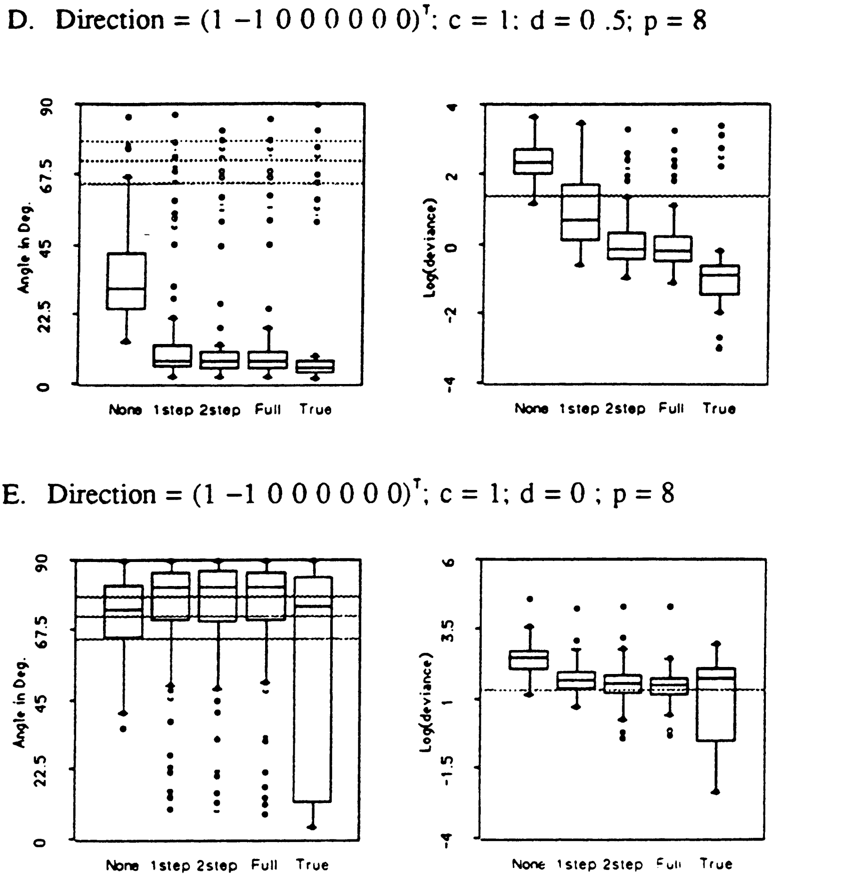


A. Null case: Identity link, p = 8

B. c = 1; p = 8; d = 1, direction = (1 1 0 0 0 0 0 0)ᵀ

C. c = 1; p = 8; d = .707, direction = (1 1 0 0 0 0 0 0)ᵀ

FIG. 2. A. *Null case: Identity link*, $p = 8$. B. $c = 1; p = 8; d = 1$, *direction* = $(1 1 0 0 0 0 0 0)^T$. C. $c = 1$; $p = 8; d = .707$, *direction* = $(1 1 0 0 0 0 0 0)^T$.


D. Direction = (1 −1 0 0 0 0 0 0)ᵀ; c = 1; d = 0 .5; p = 8

E. Direction = (1 −1 0 0 0 0 0 0)ᵀ; c = 1; d = 0 ; p = 8

FIG. 2. D. *Direction* = $(1\ -1\,0\,0\,0\,0\,0\,0)^T$; $c = 1$; $d = 0.5$; $p = 8$. E. *Direction* = $(1\ -1\,0\,0\,0\,0\,0\,0)^T$; $c = 1$; $d = 0$; $p = 8$.

A. *Null case.* To calibrate the results, we first fit using an identity link, with $p = 8$ and $c = 1$. The results are given in the first row of Figure 2, the first column for angles, and the second for the goodness of fit statistic. As expected, all the angles are very small, typically in the range 5° to 9°. The initial fit of the canonical link, given the label "None" in the figure, does somewhat better, giving a median angle of about 6°, compared to about 7° for the iterated estimates. The lack of fit statistic is about twice the size for the iterated estimates as it is for the canonical fit, showing the cost of estimating a link when the null link is in fact true.

B. *Direction* = $((1\,1\,0\,0\,0\,0\,0\,0)^T$; $c = 1$; $d = 1$; $p = 8$. For this particular direction, the Li–Duan theorem does hold, since $E(X \mid X^T\beta = v) = (v/2\ v/2\,0\,0\,0\,0\,0\,0)^T$, which is linear in v, but the choice of centering constant makes estimating the angle quite difficult. The results are summarized in Figure 2B. In the angles boxplots, the three dashed lines correspond to the median and quartiles of the angle between β and a randomly chosen direction. While the estimators are quite variable, the iterated estimates, with medians in the range 17° to 37°, generally offer a clear improvement over the initial estimate based on fitting the canonical link, with median of 53°; the initial estimate is only slightly preferable to choosing a random estimate of the direction. Fitting with the correct link has a smaller median angle than the iterated estimates, but is more variable. The iterated estimates also give a substantial reduction in the lack of fit statistics (the dashed line corresponds to a mean of the lack of fit statistic of .04, the nominal error in the data).

C. *Direction* = $(1\,1\,0\,0\,0\,0\,0\,0)^T$; $c = 1$; $d = .707$; $p = 8$. This example is very similar to case B, except that the centering constant d is smaller, making the fitting much less challenging for the initial estimate. This is clearly shown in the boxplots. The initial estimate has a median angle of 23°, while the iterated estimates have medians of about 10°, and fitting with the correct link has median angle of about 7°. There is little to be gained here by increasing beyond one iteration, and not much loss for estimating the link.

D. *Direction* = $(1\,-1\,0\,0\,0\,0\,0\,0)^T$; $c = 1$; $d = 0.5$; $p = 8$. This is a case where the Li–Duan theorem does not apply, as $E(X \mid X^T\beta_0 = v)$ is easily shown to be nonlinear in v, with the first component increasing with v and the second component decreasing with v. The summary of the simulation is given as Figure 2d. In spite of the failure of the Li–Duan theorem, the initial estimate does substantially better than random, with a median angle of about 31°. This is the case where the iterated estimates do remarkably well: The median angle decreases to about 7°, while it is about 5° for the fit using the true link. Little improvement on the angle is gained after the first iteration; however, reestimation of the link a second time decreases the lack of fit statistics by a factor of about 2.

E. *Direction* = $(1\,-1\,0\,0\,0\,0\,0\,0)^T$; $c = 1$; $d = 0$; $p = 8$. As with case D, the Li–Duan theorem does not apply, but now the link function centers the linear predictor at its mean. Since the distribution of the linear predictor is symmetric and the link is symmetric about 0, a result in Cook, Hawkins and Weisberg (1992) suggests that no consistent estimate of the direction is possible, and we can expect all the methods to do very poorly. This is reflected in Figure 2E, where the uniterated estimate behaves like choosing a direction at random, and the iterated estimates do somewhat worse than random. Fitting with the link known generally also does poorly, although it will occasionally be reasonably accurate, and is certainly better than choosing a direction at random.

In summary, the simulations indicate that in most cases little will be lost by estimating a link after fitting a canonical link. This method seems to improve estimates in most circumstances, with the angles and the link function reasonably well determined. One somewhat surprising result is the general success of the one-step estimates, even in challenging conditions. A good procedure might be to fit the canonical link, estimate the link nonparametrically, reestimate the direction given the link and then reestimate the link given the direction, corresponding to 1.5 iterations.

4. Discussion. In models with a discrete response, the use of smoothers to get a better picture of the data has been proposed by a number of authors. For example, for binary regression Copas (1983) has suggested that a plot of a single predictor x versus the kernel smooth of y on x may be much more informative than the plot of x versus y. Fowlkes (1987) defines residuals based on comparing a parametric fit to a smoothed fit, and suggests that these may be useful in examining diagnostics for the need to transform predictors, and interactions, and the like. Azzalini, Bowman and Härdle (1989) provide a method for testing the adequacy of a proposed parametric link by comparing it to nonparametric smoothed links; le Cressie and van Houwelingen (1991) provide an alternative goodness of fit procedure based on smoothing residuals rather than the link function.

More development of the computational method is required for routine use of this methodology. The algorithm proposed uses a version of Fisher scoring, which has only linear convergence in generalized linear models with noncanonical links [Smyth (1987)]. In some examples, convergence in the scoring step has been very slow. Improvements may be possible by considering the use of a computational algorithm that has quadratic convergence, such as Newton–Raphson.

We have chosen to avoid restriction to monotone links in the interests of enhanced flexibility. First, there is no particular reason in a model like (1.1) to assume that the link is monotone. Second, when our method is used diagnostically, it is useful to allow nonmonotone g because nonmonotonicity may highlight serious deficiencies in the data or model. Finally, nonmonotone links do occur in practice as in the example published by Azzalini, Bowman and Härdle (1989).

Using the estimation method described here, one can use a nonparametric smoothed link both as a basis of inference and as a basis for diagnostics, without the requirement of a p-dimensional smoother. For example, one could use the test procedure proposed by Azzalini, Bowman and Härdle (1989) but use the fitting procedure proposed here. The methodology given by Fowlkes (1987) can also be used. However, further work in this area is likely to be required to make full use of the methods described in this paper.

APPENDIX

Proof of Theorem 2. Our arguments are informed by those of Härdle, Hall and Ichimura (1993) which in turn are closely related to those used in projection pursuit regression by Hall (1989) and Chen (1991). To simplify the notation, let $\mathcal{B} = \{\beta \in R^p: |\beta - \beta_0| \leq n^{-1/2}B\}$ and $\mathcal{X} = \{x \in R^p: \gamma(x^T\beta_0) \geq a\}$. For $\beta \in \mathcal{B}$ and $x \in \mathcal{X}$, set $g(u,\beta) = E(Y \mid X^T\beta = u) = E\{g(X^T\beta_0) \mid X^T\beta = u\}$,

$$\Delta(x,\beta) = \hat{g}(x^T\beta,\beta) - g(x^T\beta,\beta) - \hat{g}(x^T\beta_0,\beta_0) + g(x^T\beta_0),$$
$$D(x) = \hat{g}(x^T\beta_0,\beta_0) - g(x^T\beta_0)$$

and

$$\delta(x,\beta)=g\left(x^T\beta,\beta\right)-g\left(x^T\beta_0\right),$$

so that we can write

$$\widehat{g}\left(x^T\beta,\beta\right)=g\left(x^T\beta_0\right)+\Delta(x,\beta)+D(x)+\delta(x,\beta).$$

Also, let

$$\begin{aligned} w(x,\beta)={}&J_a\left\{\widehat{\gamma}\left(x^T\beta,\beta\right)\right\}\widehat{g}'\left(x^T\beta,\beta\right)V\left\{\widehat{g}\left(x^T\beta,\beta\right)\right\}^{-1}\\ &-J_a\left\{\gamma\left(x^T\beta_0\right)\right\}g'\left(x^T\beta_0\right)V\left\{g\left(x^T\beta_0\right)\right\}^{-1}\end{aligned}$$

and

$$\Omega_i=X_i-\mu\left(X_i^T\beta_0\right).$$

Then, for $\beta\in\mathcal{B}$, write

$$\begin{aligned} &n^{-1/2}\sum_{i=1}^n\{X_i-\widehat{\mu}(X_i,\beta)\}\left\{Y_i-\widehat{g}\left(X_i^T\beta,\beta\right)\right\}J_a\left\{\widehat{\gamma}\left(X_i^T\beta,\beta\right)\right\}\\ &\quad\times\widehat{g}'\left(X_i^T\beta,\beta\right)V\left\{\widehat{g}\left(X_i^T\beta,\beta\right)\right\}^{-1}\\ &\quad-n^{-1/2}\sum_{i=1}^n\Omega_i\left\{Y_i-g\left(X_i^T\beta_0\right)\right\}J_a\left\{\gamma\left(X_i^T\beta_0\right)\right\}g'\left(X_i^T\beta_0\right)\\ &\quad\times V\left\{g\left(X_i^T\beta_0\right)\right\}^{-1}+A^{-1}n^{1/2}(\beta-\beta_0)=\sum_{m=1}^{10}T_m,\end{aligned}$$

where

$$\begin{aligned} T_1&=n^{-1/2}\sum_{i=1}^n\Omega_i\left\{Y_i-g\left(X_i^T\beta_0\right)\right\}w(X_i,\beta),\\ T_2&=-n^{-1/2}\sum_{i=1}^n\Omega_i\delta(X_i,\beta)J_a\left\{\widehat{\gamma}\left(X_i^T\beta,\beta\right)\right\}\widehat{g}'\left(X_i^T\beta,\beta\right)V\left\{\widehat{g}\left(X_i^T\beta,\beta\right)\right\}^{-1}\\ &\quad+A^{-1}n^{1/2}(\beta-\beta_0),\\ T_3&=-n^{-1/2}\sum_{i=1}^n\Omega_iD(X_i)w(X_i,\beta),\\ T_4&=-n^{-1/2}\sum_{i=1}^n\Omega_i\Delta(X_i,\beta)J_a\left\{\widehat{\gamma}\left(X_i^T\beta,\beta\right)\right\}\widehat{g}'\left(X_i^T\beta,\beta\right)V\left\{\widehat{g}\left(X_i^T\beta,\beta\right)\right\}^{-1},\end{aligned}$$

$$T_5 = -n^{-1/2}\sum_{i=1}^{n}\Omega_i D(X_i)J_a\Big\{\gamma(X_i^T\beta_0)\Big\}g'(X_i^T\beta_0)V\Big\{g(X_i^T\beta_0)\Big\}^{-1},$$

$$T_6 = -n^{-1/2}\sum_{i=1}^{n}\Big\{\widehat{\mu}(X_i,\beta)-\mu(X_i^T\beta_0)\Big\}\Big\{Y_i-g(X_i^T\beta_0)\Big\}J_a\Big\{\gamma(X_i^T\beta_0)\Big\}$$
$$\times g'(X_i^T\beta_0)V\Big\{g(X_i^T\beta_0)\Big\}^{-1},$$

$$T_7 = -n^{-1/2}\sum_{i=1}^{n}\Big\{\widehat{\mu}(X_i,\beta)-\mu(X_i^T\beta_0)\Big\}\Big\{Y_i-g(X_i^T\beta_0)\Big\}w(X_i,\beta),$$

$$T_8 = n^{-1/2}\sum_{i=1}^{n}\Big\{\widehat{\mu}(X_i,\beta)-\mu(X_i^T\beta_0)\Big\}\Delta(X_i,\beta)J_a\Big\{\widehat{\gamma}(X_i^T\beta,\beta)\Big\}$$
$$\times \widehat{g}'(X_i^T\beta,\beta)V\Big\{\widehat{g}(X_i^T\beta,\beta)\Big\}^{-1},$$

$$T_9 = n^{-1/2}\sum_{i=1}^{n}\Big\{\widehat{\mu}(X_i,\beta)-\mu(X_i^T\beta_0)\Big\}D(X_i)J_a\Big\{\widehat{\gamma}(X_i^T\beta,\beta)\Big\}$$
$$\times \widehat{g}'(X_i^T\beta,\beta)V\Big\{\widehat{g}(X_i^T\beta,\beta)\Big\}^{-1},$$

$$T_{10} = n^{-1/2}\sum_{i=1}^{n}\Big\{\widehat{\mu}(X_i,\beta)-\mu(X_i^T,\beta_0)\Big\}\delta(X_i,\beta)$$
$$\times J_a\Big\{\widehat{\gamma}(X_i^T\beta,\beta)\Big\}\widehat{g}'(X_i^T\beta,\beta)V\Big\{\widehat{g}(X_i^T\beta,\beta)\Big\}^{-1}.$$

The first part of the theorem will follow once we prove that $\sup_{|\beta-\beta_0|\le n^{-1/2}B}|T_m| \xrightarrow{p} 0, 1 \le m \le 10$; the proof of the second part is straightforward and is omitted.

The results required to complete the proof of Theorem 2 are based on expressions and bounds for $\Delta(x,\beta)$, $D(x,\beta)$, $w(x,\beta)$ and $\widehat{\mu}(x,\beta)-\mu(x^T\beta_0)$ for $\beta\in\mathcal{B}$ and $x\in\mathcal{X}$ which are obtained by combining results for $\widehat{\gamma}(x^T\beta,\beta)$, $\widehat{g}(x^T\beta,\beta)$ and $\widehat{g}'(x^T\beta,\beta)$. The key technical steps in obtaining the bounds are (1) the establishment of pointwise mean square error bounds by evaluating means and variances and (2) the use of Rosenthal's inequality [Hall and Heyde (1980), page 23] to establish that the bounds hold uniformly on $\beta\in\mathcal{B}$ and $x\in\mathcal{X}$. This approach increases the pointwise bound by n^{ξ} for any $\xi>0$. The argument is detailed in Härdle, Hall and Ichimura (1993). While we often apply the argument to sums of independent random variables, we require the full generality of Rosenthal's inequality when we represent T_1, T_5 and T_6 as martingales.

We begin with four lemmas which are proved using standard moment calculations and Rosenthal's inequality.

LEMMA 1. *For any* $\xi>0$,

$$\sup_{x\in\mathcal{X}}B(x,\beta_0) = O_p\big(n^{\xi}h^3+n^{-1/2+\xi}h^{1/2}\big),$$
$$\sup_{\beta\in\mathcal{B}}\sup_{x\in\mathcal{X}}|B(x,\beta)-B(x,\beta_0)| = O_p\big(n^{-1/2+\xi}h\big)$$

and

$$\sup_{\beta \in \mathcal{B}} \sup_{x \in \mathcal{X}} |C(x,\beta)| = O_p\left(n^{\xi}h^3 + n^{-1/2+\xi}h^{1/2}\right),$$

where

$$B(x,\beta) = (nh)^{-1}\sum_{j=1}^{n}\left\{g\left(X_j^T\beta\right) - g\left(x^T\beta\right)\right\}K_h\{(x-X_j)^T\beta\}$$

and

$$C(x,\beta) = (nh)^{-1}\sum_{j=1}^{n}\left\{X_j g\left(X_j^T\beta\right) - \mu\left(x^T\beta_0\right)g\left(x^T\beta_0\right)\right\}K_h\{(x-X_j)^T\beta\}.$$

LEMMA 2. *For any* $\xi > 0$,

$$\begin{aligned}\sup_{\beta \in \mathcal{B}} \sup_{x \in \mathcal{X}} \Bigg| &\hat{\gamma}\left(x^T\beta,\beta\right) - \gamma\left(x^T\beta_0\right) \\ &- (nh)^{-1}\sum_{j=1}^{n}\left[K_h\left((x-X_j)^T\beta\right) - EK_h\left((x-X)^T\beta\right)\right] \\ &- h^{-1}E\left[K_h\left((x-X)^T\beta\right) - K_h\left((x-X)^T\beta_0\right)\right]\Bigg| = O_p\left(n^{\xi}h^3\right)\end{aligned}$$

and

$$\sup_{\beta \in \mathcal{B}} \sup_{x \in \mathcal{X}} \left|\hat{\gamma}\left(x^T\beta,\beta\right) - \gamma\left(x^T\beta_0\right)\right| = O_p\left(n^{\xi}h^3 + n^{-1/2+\xi}h^{-1/2}\right).$$

Note from Lemma 2 that

$$\sup_{\beta \in \mathcal{B}} \sup_{x \in \mathcal{X}} \hat{\gamma}\left(x^T\beta,\beta\right)^{-1} = O(\log n),$$

whenever

$$\sup_{\beta \in \mathcal{B}} \sup_{x \in \mathcal{X}} \gamma\left(x^T\beta_0\right)^{-1} = O(\log n).$$

LEMMA 3. *For any* $\xi > 0$,

$$\begin{aligned}\sup_{\beta \in \mathcal{B}} \sup_{x \in \mathcal{X}} \Bigg| &P(x,\beta) - \gamma\left(x^T\beta_0\right)g'\left(x^T\beta_0\right) -`\gamma'\left(x^T\beta_0\right)g\left(x^T\beta_0\right) \\ &- \left(nh^2\right)^{-1}\sum_{j=1}^{n}\left[g\left(X_j^T\beta_0\right)L\{(x-X_j)^T\beta/h\} - Eg\left(X^T\beta_0\right)L\{(x-X)^T\beta/h\}\right] \\ &- h^{-2}Eg\left(X^T\beta_0\right)\left[L\{(x-X)^T\beta/h\} - L\{(x-X)^T\beta_0/h\}\right]\Bigg| \\ = {} &O_p\left(n^{\xi}h^3\right)\end{aligned}$$

and

$$\sup_{\beta \in \mathcal{B}} \sup_{x \in \mathcal{X}} \left| P(x, \beta) - \gamma(x^T\beta_0)g'(x^T\beta_0) - \gamma'(x^T\beta_0)g(x^T\beta_0) \right|$$
$$= O_p\left(n^{\xi}h^3 + n^{-1/2+\xi}h^{-3/2}\right),$$

where

$$P(x, \beta) = n^{-1}h^{-2}\sum_{j=1}^{n} g(X_j^T\beta_0)L_h\{(x - X_j)^T\beta\}.$$

Lemma 4. *For any* $\xi > 0$,

$$\sup_{\beta \in \mathcal{B}} \sup_{x \in \mathcal{X}} \Bigg| Q(x, \beta) - \gamma'(x^T\beta_0) - n^{-1}h^{-2}\sum_{j=1}^{n}\left[L_h((x - X_j)^T\beta) - EL_h((x - X)^T\beta)\right]$$
$$- h^{-2}E\left[L_h((x_i - X)^T\beta) - L_h((x - X)^T\beta_0)\right]\Bigg| = O_p(n^{\xi}h^3)$$

and

$$\sup_{\beta \in \mathcal{B}} \sup_{x \in \mathcal{X}} \left| Q(x, \beta) - \gamma'(x^T\beta_0) \right| = O_p\left(n^{\xi}h^3 + n^{-1/2+\xi}h^{-3/2}\right),$$

where

$$Q(x, \beta) = n^{-1}h^{-2}\sum_{j=1}^{n} L_h\{(x - X_j)^T\beta\}.$$

Lemmas 1 and 2 enable us to describe the behavior of Δ and D, and hence $\widehat{g}$.

Lemma 5.

$$\sup_{\beta \in \mathcal{B}} \sup_{x \in \mathcal{X}} |\Delta(x, \beta)| = o_p\left(n^{-1/2}\right).$$

Proof. Since for any $x, w \in \mathcal{X}$ we have

$$g(w^T\beta_0) - g(x^T\beta, \beta) = g(w^T\beta) - g(x^T\beta) + (\beta - \beta_0)^T$$
$$\times \left\{\mu(x^T\beta_0)g'(x^T\beta_0) - wg'(w^T\beta_0)\right\} + O_p(n^{-1}\log n),$$

it follows that for any $x \in \mathcal{X}$,

$$\begin{aligned}\Delta(x,\beta) = (nh)^{-1}\sum_{j=1}^{n}\Big\{Y_j - g\big(X_j^T\beta_0\big)\Big\}\Big[&K_h\{(x-X_j)^T\beta\} \\ &- K_h\{(x-X_j)^T\beta_0\}\Big]\hat{\gamma}\big(x^T\beta_0,\beta_0\big)^{-1} \\ - (nh)^{-1}\sum_{j=1}^{n}\Big\{Y_j - g\big(X_j^T\beta_0\big)\Big\}&K_h\{(x-X_j)^T\beta\} \\ \times \Big\{\hat{\gamma}\big(x^T\beta,\beta\big) - \hat{\gamma}\big(x^T\beta_0,\beta_0\big)\Big\}&\Big\{\hat{\gamma}\big(x^T\beta,\beta\big)\hat{\gamma}\big(x^T\beta_0,\beta_0\big)\Big\}^{-1} \\ + B(x,\beta)\Big\{\hat{\gamma}\big(x^T\beta,\beta\big) - \hat{\gamma}\big(x^T\beta_0,\beta_0\big)\Big\}&\Big\{\hat{\gamma}\big(x^T\beta,\beta\big)\hat{\gamma}\big(x^T\beta_0,\beta_0\big)\Big\}^{-1} \\ - \big\{B(x,\beta) - B(x,\beta_0)\big\}&\hat{\gamma}\big(x^T\beta_0,\beta_0\big)^{-1} \\ - (\beta-\beta_0)^T C(x,\beta)&\hat{\gamma}\big(x^T\beta,\beta\big)^{-1} + o_p\big(n^{-1/2}\big).\end{aligned}$$

Now, apply Lemmas 1 and 2 and the fact that for each fixed $\beta \in \mathcal{B}$ and $x \in \mathcal{X}$,

$$\begin{aligned}\sup_{\beta\in\mathcal{B}}\sup_{x\in\mathcal{X}} E&\left[(nh)^{-1}\sum_{j=1}^{n}\Big\{Y_j - g\big(X_j^T\beta_0\big)\Big\}\Big[K_h\{(x-X_j)^T\beta\} - K_h\{(x-X_j)^T\beta_0\}\Big]\right]^2 \\ &= O\big(n^{\xi-2}h^{-1}\big)\end{aligned}$$

and

$$\begin{aligned}\sup_{\beta\in\mathcal{B}}\sup_{x\in\mathcal{X}} E&\left[(nh)^{-1}\sum_{j=1}^{n}\Big\{Y_j - g\big(X_j^T\beta_0\big)\Big\}K_h\{(x-X_j)^T\beta\}\right]^2 \\ &= O\big(n^{\xi-1}h^{-1}\big).\end{aligned}$$

□

LEMMA 6.

$$\begin{aligned}\sup_{x\in\mathcal{X}}&\left|D(x) - (nh)^{-1}\sum_{j=1}^{n}\Big\{Y_j - g\big(X_j^T\beta_0\big)\Big\}K_h\{(x-X_j)^T\beta_0\}\gamma\big(x^T\beta_0\big)^{-1}\right| \\ &= o_p\big(n^{-1/2}\big)\end{aligned}$$

and, for any $\xi > 0$,

$$\sup_{x\in\mathcal{X}}|D(x)| = O_p\big(n^{-1/2+\xi}h^{-1/2}(\log n)\big).$$

The next lemma follows immediately from Lemmas 5 and 6.

LEMMA 7.

$$\sup_{\beta\in\mathcal{B}}\sup_{x\in\mathcal{X}}\left|\widehat{g}(x^T\beta,\beta)-g(x^T\beta_0)-(nh)^{-1}\sum_{j=1}^{n}\left\{Y_j-g(X_j^T\beta_0)\right\}\right.$$
$$\left.\times K_h\{(x-X_j)^T\beta_0\}\gamma\left(X_i^T\beta_0\right)^{-1}-\delta(x,\beta)\right|=o_p\left(n^{-1/2}\right)$$

and, for any $\xi > 0$,

$$\sup_{\beta\in\mathcal{B}}\sup_{x\in\mathcal{X}}\left|\widehat{g}(x^T\beta,\beta)-g(x^T\beta_0)\right|=O_p\left(n^{-1/2}(\log n)^2+n^{-1/2+\xi}h^{-1/2}\right).$$

We need a similar result for $\widehat{g}'(x^T\beta,\beta)-g'(x^T\beta_0)$ to complete the description of $w(x,\beta)$. The required results are given in Lemmas 8 and 9.

LEMMA 8.

$$\sup_{\beta\in\mathcal{B}}\sup_{x\in\mathcal{X}}\left|\widehat{g}'(x^T\beta,\beta)-g'(x^T\beta_0)\right.$$
$$-\left(nh^2\right)^{-1}\sum_{k=1}^{n}\left\{Y_k-g(X_k^T\beta_0)\right\}L_h((x-X_k)^T\beta_0)\gamma\left(x^T\beta_0\right)^{-1}$$
$$+\left\{\widehat{g}(x^T\beta,\beta)-g(x^T\beta_0)\right\}\gamma'\left(x^T\beta_0\right)\gamma\left(x^T\beta_0\right)^{-1}$$
$$-\left\{P(x,\beta)-\gamma\left(x^T\beta_0\right)g'\left(x^T\beta_0\right)-\gamma'\left(x^T\beta_0\right)g\left(x^T\beta_0\right)\right\}\gamma\left(x^T\beta_0\right)^{-1}$$
$$+\gamma\left(x^T\beta_0\right)g'\left(x^T\beta_0\right)\left\{\widehat{\gamma}\left(x^T\beta,\beta\right)-\gamma\left(x^T\beta_0\right)\right\}\gamma\left(x^T\beta_0\right)^{-2}$$
$$\left.+g\left(x^T\beta_0\right)\left\{Q(x,\beta)-\gamma'\left(x^T\beta_0\right)\right\}\gamma\left(x^T\beta_0\right)^{-1}\right|=o_p\left(n^{-1/2}\right)$$

and, for any $\xi > 0$,

$$\sup_{\beta\in\mathcal{B}}\sup_{x\in\mathcal{X}}\left|\widehat{g}'\left(x^T\beta,\beta\right)-g'\left(x^T\beta_0\right)\right|$$
$$=O_p\left(n^{\xi}h^3(\log n)+n^{-1/2}(\log n)^3+n^{-1/2+\xi}h^{-3/2}(\log n)\right).$$

PROOF. Apply Lemmas 2, 3, 4 and 7 to $\widehat{g}'(x^T\beta,\beta)-g'(x^T\beta_0)$. □

LEMMA 9.

$$\sup_{\beta\in\mathcal{B}}\sup_{x\in\mathcal{X}}\bigg|w(x,\beta)+J_a\Big\{\gamma\big(x^T\beta_0\big)\Big\}g'\big(x^T\beta_0\big)\times\Big\{\widehat{g}(x^T\beta,\beta)-g(x^T\beta_0)\Big\}V'\Big\{g(x^T\beta_0)\Big\}V\Big\{g(x^T\beta_0)\Big\}^{-2}-J_a\Big\{\gamma(x^T\beta_0)\Big\}\Big\{\widehat{g}'(x^T\beta,\beta)-g'(x^T\beta_0)\Big\}V\Big\{g(x^T\beta_0)\Big\}^{-1}-\Big\{\widehat{\gamma}(x^T\beta,\beta)-\gamma(x^T\beta_0)\Big\}J_a'\Big\{\gamma(x^T\beta_0)\Big\}g'(x^T\beta_0)\times V\Big\{g(x^T\beta_0)\Big\}^{-1}\bigg|=o_p\big(n^{-1/2}\big)$$

and, for any $\xi > 0$,

$$\sup_{\beta\in\mathcal{B}}\sup_{x\in\mathcal{X}}|w(x,\beta)|=O_p\big(n^{\xi}h^3+n^{-1/2+\xi}h^{-3/2}\big).$$

PROOF. Apply Lemmas 2, 6 and 8. □

The last two preliminary lemmas describe the behavior of $\widehat{\mu}$.

LEMMA 10. *For any* $\xi > 0$,

$$\sup_{\beta\in\mathcal{B}}\sup_{x\in\mathcal{X}}\bigg|(nh)^{-1}\sum_{j=1}^{n}X_jK_h\big((x-X_j)^T\beta\big)-\mu(x^T\beta_0)\gamma(x^T\beta_0)-(nh)^{-1}\sum_{j=1}^{n}\Big[X_jK_h\big((x-X_j)^T\beta\big)-EXK_h\big((x-X)^T\beta\big)\Big]-h^{-1}EX\Big[K_h\big((x-X)^T\beta\big)-K_h\big((x-X)^T\beta_0\big)\Big]\bigg|=O\big(n^{\xi}h^3\big)$$

and

$$\sup_{\beta\in\mathcal{B}}\sup_{x\in\mathcal{X}}\bigg|(nh)^{-1}\sum_{j=1}^{n}X_jK_h\big((x-X_j)^T\beta\big)-\mu(x^T\beta_0)\gamma(x^T\beta_0)\bigg|=O_p\big(n^{\xi}h^3+n^{-1/2+\xi}h^{-1/2}\big).$$

LEMMA 11.

$$
\begin{aligned}
\sup_{\beta\in\mathcal{B}}\sup_{x\in\mathcal{X}}\Bigg|\,&\widehat{\mu}(x^T\beta,\beta)-\mu(x^T\beta_0)\\
&+(nh)^{-1}\sum_{j=1}^{n}\Big[X_jK_h\big((x-X_j)^T\beta\big)-EXK_h\big((x-X)^T\beta\big)\Big]\\
&+h^{-1}EX\Big[K_h\big((x-X)^T\beta\big)-K_h\big((x-X)^T\beta_0\big)\Big]\gamma\big(x^T\beta_0\big)^{-1}\\
&-\mu\big(x^T\beta_0\big)\Bigg\{(nh)^{-1}\sum_{j=1}^{n}\Big[K_h\big((x-X_j)^T\beta\big)\\
&\qquad\qquad-EK_h\big((x-X)^T\beta\big)\Big]\gamma\big(x^T\beta_0\big)^{-1}\\
&-\mu\big(x^T\beta_0\big)h^{-1}E\Big[K_h\big((x-X)^T\beta\big)-K_h\big((x-X)^T\beta_0\big)\Big]\\
&\qquad\qquad\times\gamma\big(x^T\beta_0\big)^{-1}\Bigg|=o_p\big(n^{-1/2}\big)
\end{aligned}
$$

and, for any $\xi>0$,

$$
\sup_{\beta\in\mathcal{B}}\sup_{x\in\mathcal{X}}\big|\widehat{\mu}(x^T\beta,\beta)-\mu(x^T\beta_0)\big|=O_p\Big(\big(n^{\xi}h^3+n^{-1/2+\xi}h^{-1/2}\big)(\log n)^2\Big).
$$

PROOF. The result follows immediately from Lemmas 2 and 10. □

We are now able to complete the proof of Theorem 2. We have

$$
\sup_{\beta\in\mathcal{B}}|T_3|=o_p(1)\quad\text{and}\quad\sup_{\beta\in\mathcal{B}}|T_4|=o_p(1)
$$

by Lemmas 6 and 10 and 5 and 9, respectively. Next, note that the derivatives of J_a are nonzero only on $(a, a+\theta)$ so that with probability which can be made arbitrarily close to 1, $w(x,\beta)$ is nonzero only on the set $x\in\mathcal{X}$. Since $\delta(x,\beta) = \{x-\mu(x^T\beta)\}^T(\beta-\beta_0)g'(x^T\beta_0)+O(n^{-1}\log n)$ uniformly on $\mathcal{X}\times\mathcal{B}$, we have

$$
\begin{aligned}
\sup_{\beta\in\mathcal{B}}|T_2| = \sup_{\beta\in\mathcal{B}}\Bigg|\,&n^{-1/2}\sum_{i=1}^{n}\Omega_i\delta(X_i,\beta)J_a\Big\{\widehat{\gamma}\big(X_i^T\beta,\beta\big)\Big\}\widehat{g}'\big(X_i^T\beta,\beta\big)V\Big\{\widehat{g}\big(X_i^T\beta,\beta\big)\Big\}^{-1}\\
&-n^{-1}\sum_{i=1}^{n}\Omega_i\Omega_i^TJ_a\Big\{\gamma\big(X_i^T\beta_0\big)\Big\}g'\big(X_i^T\beta_0\big)^2V\Big\{g\big(X_i^T\beta_0\big)\Big\}^{-1}\\
&\qquad\qquad\times n^{1/2}(\beta-\beta_0)\Bigg|\\
\le{}& Bn^{-1}\sum_{i=1}^{n}\big|\Omega_i\Omega_i^Tg'\big(X_i^T\beta_0\big)\big|\sup_{\beta\in\mathcal{B}}\sup_{x\in\mathcal{X}}|w(x,\beta)|+o_p(1)\\
={}&o_p(1)
\end{aligned}
$$

by Lemma 9.

Next, write

$$\begin{aligned} T_5 &= n^{-3/2}h^{-1}\sum_{i=1}^{n}\Omega_i J_a\Big\{\gamma(X_i^T\beta_0)\Big\}g'(X_i^T\beta_0)V\Big(g(X_i^T\beta_0)\Big)^{-1} \\ &\qquad \times\Big\{\widehat{\gamma}(X_i^T\beta_0,\beta_0)^{-1}-\gamma(X_i^T\beta_0)^{-1}\Big\} \\ &\qquad \times\sum_{j=1}^{n}\Big\{Y_j-g(X_j^T\beta_0)\Big\}K_h((X_i-X_j)^T\beta_0) \\ &\quad +n^{-1/2}\sum_{j=1}^{n}\Big\{Y_i-g(X_j^T\beta_0)\Big\}(nh)^{-1}\sum_{i=1}^{n}\Omega_i K_h((X_i-X_j)^T\beta_0) \\ &\qquad \times J_a\Big\{\gamma(X_i^T\beta_0)\Big\}g'(X_i^T\beta_0)\Big/\Big\{V\Big(g(X_i^T\beta_0)\Big)\gamma(X_i^T\beta_0)\Big\}+o_p(1) \\ &= S_1+S_2+o_p(1), \end{aligned}$$

say. Then

$$\sup_{x\in\mathcal{X}}\left|n^{-1/2}\sum_{j=1}^{n}\Big\{Y_j-g(X_j^T\beta_0)\Big\}K_h((x-x_j)^T\beta_0)\right|=O_p(n^{\xi}h^{1/2}),$$

so $|S_1|=o_p(1)$. Also, let (Z,W) be independent of but with the same distribution as (Y,X). Then for any fixed t, let

$$\begin{aligned} u(y,x,z,w) &= \Big(z-g(w^T\beta_0)\Big)h^{-1}t^T\Big\{x-\mu(x^T\beta_0)\Big\}K_h((x-w)^T\beta_0) \\ &\qquad \times J_a\Big\{\gamma(x^T\beta_0)\Big\}g'(x^T\beta_0)V\Big\{g(x^T\beta_0)\Big\}^{-1}\gamma(x^T\beta_0)^{-1}. \end{aligned}$$

Since $Eu(y,x,Z,W)+Eu(Z,W,y,x)=0$ because $E(Z\,|\,W=w)=g(w^T\beta_0)$ and $E(W\,|\,W^T\beta_0=w^T\beta_0)=\mu(w^T\beta_0)$, we can write

$$S_2=n^{-3/2}\sum_{i<j}\{u(Y_i,X_i,Y_j,X_j)+u(Y_j,X_j,Y_i,X_i)\}+n^{-3/2}\sum_{j=1}^{n}u(Y_i,X_i,Y_i,X_i),$$

where $\sum_{i<j}\{u(Y_i,X_i,Y_j,X_j)+u(Y_j,X_j,Y_i,X_i)\}$ is a martingale with respect to $\mathcal{F}_n=\sigma\{(Y_1,X_1),\dots,(Y_n,X_n)\}$. Hence

$$\begin{aligned} &E\Big\{n^{-3/2}\sum_{i<j}\{u(y_i,x_i,y_j,x_j)+u(y_j,x_j,y_i,x_i)\}\Big\}^2 \\ &\qquad =\{n^{-3}n(n-1)/2\}E\{u(y_i,x_i,y_j,x_j)+u(y_j,x_j,y_i,x_i)\}^2 \end{aligned}$$

and it follows that $T_5=o_p(1)$.

To show that $\sup_{\beta\in\mathcal{B}}|T_1|=o_p(1)$, use Lemmas 2, 3, 4, 7, 8 and 9 to approximate $w(x,\beta)$ and then apply arguments similar to those used in the treatment of T_5. Finally, we can show that $\sup_{\beta\in\mathcal{B}}|T_m|=o_p(1)$ for $m=6,7,8,9$ and 10 by simple modifications of these arguments.

Acknowledgments. We are grateful to the editor, associate editor and referees for their helpful comments. Peter Hall, Dennis Cook and Shane Pederson made important contributions to some phases of this work.

REFERENCES

AZZALINI, A., BOWMAN, A. and HÄRDLE, W. (1989). On the use of nonparametric regression for model checking. *Biometrika* **76** 1–12.

BOX, G. E. P. and COX, D. R. (1964). An analysis of transformations (with discussion). *J. Roy. Statist. Soc. Ser. B* **26** 211–246.

CHEN, H. (1991). Estimation of a projection–pursuit type regression model. *Ann. Statist.* **19** 142–157.

COOK, R. D., HAWKINS, D. M. and WEISBERG, S. (1992). Comparison of model misspecification diagnostics using residuals from least mean of squares and least median of squares fits. *J. Amer. Statist. Assoc.* **87** 419–424.

COOK, R. D. and WEISBERG, S. (1991). Comment on "Sliced inverse regression for dimension reduction," by K. C. Li. *J. Amer. Statist. Assoc.* **86** 328–333.

COPAS, J. (1983). Plotting *p* against *x*. *J. Roy. Statist. Soc. Ser. C* **32** 25–31.

EATON, M. L. (1986). A characterization of spherical distributions. *J. Multivariate Anal.* **20** 272–276.

FOWLKES, E. B. (1987). Some diagnostics for binary logistic regression via smoothing. *Biometrika* **74** 503–515.

GASSER, T., MÜLLER, H.-G. and MAMMITZSCH, V. (1985). Kernels for nonparametric curve estimation. *J. Roy. Statist. Soc. Ser. B* **47** 238–252.

HALL, P. (1989). On projection pursuit regression. *Ann. Statist.* **17** 573–588.

HALL, P. and HEYDE, C. C. (1980). *Martingale Limit Theory and Its Application.* Academic, New York.

HÄRDLE, W. (1990). *Applied Nonparametric Regression.* Cambridge Univ. Press.

HÄRDLE, W., HALL, P. and ICHIMURA, H. (1993). Optimal smoothing in single index models. *Ann. Statist.* **21** 157–178.

HÄRDLE, W. and STOKER, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.* **84** 986–995.

LE CRESSIE, S. and VAN HOUWELINGEN, J. C. (1991). A goodness-of-fit test for binary regression models, based on smoothing methods. *Biometrics* **47** 1267–1282.

LI, K.-C. and DUAN, N. (1989). Regression analysis under link violation. *Ann. Statist.* **17** 1009–1052.

LORD, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Erlbaum, Hillsdale, NJ.

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalised Linear Models*, 2nd ed. Chapman and Hall, London.

ORTEGA, J. M. and RHEINBOLDT, W. C. (1973). *Iterative Solution of Nonlinear Equations in Several Variables.* Academic, New York.

PRAKASA RAO, B. L. S. (1983). *Nonparametric Functional Estimation.* Academic, New York.

PREGIBON, D. (1981). Logistic regression diagnostics. *Ann. Statist.* **9** 705–724.

PREGIBON, D. (1982). Score tests in GLIM. In *GLIM82: Proceedings of the International Conference on Generalized Linear Models. Lecture Notes in Statist.* **14**. Springer, New York.

RAMSAY, J. O. and ABRAHAMOWICZ, M. (1989). Binomial regression with monotone splines: a psychometric application. *J. Amer. Statist. Assoc.* **84** 906–915.

SMYTH, G. K. (1987). Curvature and convergence. In *Proceedings of the Statistical Computing Section* 278–283. Amer. Statist. Assoc., Alexandria, VA.

STOKER, T. M. (1986). Consistent estimation of scaled coefficients. *Econometrica* **54** 1461–1481.

TIERNEY, L. (1990). *Lisp-Stat.* Wiley, New York.

TIERNEY, L. (1991). Generalized linear models in Lisp-Stat. Technical Report 557, School of Statistics, Univ. Minnesota.

WAINER, H. (1983). Pyramid power: searching for an error in test scoring with 830,000 helpers. *Amer. Statist.* **37** 87–91.

WELSH, A. H. (1989). On M-processes and M-estimation. *Ann. Statist.* **17** 337–361. [Correction (1990) **18** 1500.]

YANDELL, B. S. and GREEN, P. (1986). Semi-parametric generalised linear model diagnostics. In *Proceedings of the Statistical Computing Section* 48–53. Amer. Statist. Assoc., Alexandria, VA.

DEPARTMENT OF APPLIED STATISTICS
UNIVERSITY OF MINNESOTA
1994 BUFORD AVENUE
ST PAUL, MINNESOTA 55108

DEPARTMENT OF STATISTICS
THE FACULTIES
AUSTRALIAN NATIONAL UNIVERSITY
CANBERRA ACT 0200
AUSTRALIA