Local Polynomial Kernel Regression for Generalized Linear Models and Quasi-Likelihood Functions

JIANQING FAN, NANCY E. HECKMAN and M. P. WAND*

20th November, 1992

Generalized linear models (Wedderburn and Nelder 1972, McCullagh and Nelder 1988) were introduced as a means of extending the techniques of ordinary parametric regression to several commonly-used regression models arising from non-normal likelihoods. Typically these models have a variance that depends on the mean function. However, in many cases the likelihood is unknown, but the relationship between mean and variance can be specified. This has led to the consideration of quasi-likelihood methods, where the conditional log-likelihood is replaced by a quasi-likelihood function. In this article we investigate the extension of the nonparametric regression technique of local polynomial fitting with a kernel weight to these more general contexts. In the ordinary regression case local polynomial fitting has been seen to possess several appealing features in terms of intuitive and mathematical simplicity. One noteworthy feature is the better performance near the boundaries compared to the traditional kernel regression estimators. These properties are shown to carry over to the generalized linear model and quasi-likelihood model. The end result is a class of kernel type estimators for smoothing in quasi-likelihood models. These estimators can be viewed as a straightforward generalization of the usual parametric estimators. In addition, their simple asymptotic distributions allow for simple interpretation and extensions of state-of-the-art bandwidth selection methods.

KEY WORDS: Bandwidth; boundary effects; kernel estimator; local likelihood; logistic regression; nonparametric regression; Poisson regression; Quasi-likelihood.

* Jianqing Fan is Assistant Professor, Department of Statistics, University of North Carolina, Chapel Hill, NC 27599. Nancy E. Heckman is Associate Professor, Department of Statistics, University of British Columbia, Vancouver, Canada, V6T 1Z2. M. P. Wand is Lecturer, Australian Graduate School of Management, University of New South Wales, Kensington, NSW 2033, Australia. During this research Jianqing Fan was visiting the Mathematical Sciences Research Institute under NSF Grant DMS 8505550 and DMS 9203135. Nancy E. Heckman was supported by an NSERC of Canada Grant OGP0007969. During this research M. P. Wand was visiting the Department of Statistics, University of British Columbia and acknowledges the support of that department.

1. INTRODUCTION

Generalized linear models were introduced by Nelder and Wedderburn (1972) as a means of applying techniques used in ordinary linear regression to more general settings. In an important further extension, first considered by Wedderburn (1974), the log-likelihood is replaced by a quasi-likelihood function, which only requires specification of a relationship between the mean and variance of the response. There are, however, many examples where ordinary least squares fails to produce a consistent procedure when the variance function depends on the regression function itself so a likelihood criterion is more appropriate. Variance functions that depend on the mean function occur in logit regression, log linear models and constant coefficient of variation models, for example.

McCullagh and Nelder (1988) give an extensive account of the analysis of parametric generalized linear models. A typical parametric assumption is that some transformation of the mean of the response variable, usually called the link function, is linear in the covariates. In ordinary regression with a single covariate, this assumption corresponds to the scatterplot of the data being adequately fit by a straight line. However, there are many scatterplots that arise in practice that are not adequately fit by straight lines and other parametric curves. This has led to the proposal and analysis of several nonparametric regression techniques (sometimes called *scatterplot smoothers*). References include Eubank (1988), Härdle (1990) and Wahba (1990). The same deficiencies of parametric modeling in ordinary regression apply to generalized linear models.

In this article we investigate the generalization of local polynomial fitting with kernel weights. We are motivated by the fact that local polynomial kernel estimators are both intuitively and mathematically simple which allows for a deeper understanding of their performance. Figure 1 shows how one version of the local polynomial kernel estimator works for a simulated example. The scatterplot in Figure 1a corresponds to 220 simulated Poisson counts generated according to the mean function, shown here by the dot-dash curve. To keep the plot less cluttered, only the average counts of replications are plotted. A local linear kernel estimator of the mean is given by the solid curve. Figure 1b shows how this estimate was constructed. In this figure the scatterplot points correspond to the logarithms of the counts and the dot-dash and solid curves correspond to the log of the mean and its estimate respectively. In this example the link function is the log transformation. The dotted lines show how the estimate was obtained at two particular values of x. At $x = 0.2$ a straight line was fitted to the log counts using a weighted version of the Poisson log-likelihood where the weights correspond to the relative heights of the kernel function which, in this case, is a scaled normal density function centered about

$x = 0.2$ and is shown at the base of the plot.

Figure 1. (a) Local linear kernel estimate of the conditional mean E(Y|X=x) where Y|X=x is Poisson. The data are simulated and consist of 20 replications at each of 11 points. The plus signs show the average of each set of replications. The solid curve is the estimate, the dot-dash curve is the true mean. (b) Illustration of how the kernel smoothing is done to estimate ln E(Y|X=x) at points x=0.2 and 0.7. In each case the estimate is obtained by fitting a line to the data by maximum weighted conditional log-likelihood. The weights are with respect to the kernel function centered about x, shown at the base of the figure for these two values of x (dotted curves). The sold curve is the estimate, the dot-dash curve is the true log(mean).

The estimate at this point is the height of the line above $x = 0.2$. For estimation at the second point, $x = 0.7$, the same principle is applied with the kernel function centered about $x = 0.7$. This process is repeated at all points x at which an estimate is required. The estimate of the mean function itself is obtained by applying the inverse link function, in this case exponentiation, to the kernel smooth in Figure 1b.

There have been several other proposals for extending nonparametric regression estimators to the generalized case. Extensions of smoothing spline methodology have been studied by Green and Yandell (1985), O'Sullivan, Yandell and Raynor (1986) and Cox and O'Sullivan (1990). Tibshirani and Hastie (1987) based their generalization on the "running lines" smoother. Staniswalis (1989) carried out a similar generalization of the Nadaraya-Watson kernel estimator (Nadaraya 1964, Watson 1964) which is equivalent to local constant fitting with a kernel weight.

In an important further extension of the generalized linear model one only needs to model the conditional variance of the response variable as an arbitrary but known function of the conditional mean. This has led to the proposal of quasi-likelihood methods (Wedderburn 1974, McCullagh and Nelder 1988, Chapter 9). Optimal properties of the quasi-likelihood methods have received considerable attention in the literature. See, for example, Cox (1983) and Godambe and Heyde (1987) and references therein.

The kernel smoothing ideas described above can be easily extended to the case where a quasi-likelihood function is used, so we present our results at this level of generality. If the distribution of the responses is from an exponential family then quasi-likelihood estimation with a correctly specified variance function is equivalent to maximum likelihood estimation. Thus, results for exponential family models follow directly from those for quasi-likelihood estimation. Sevirini and Staniswalis (1992) considered quasi-likelihood estimation using locally constant fits.

In the case of normal errors, quasi-likelihood and least squares techniques coincide.

Thus, the results presented here are a generalization of those for the local least squares kernel estimator for ordinary regression considered by Fan (1992a,b) and Ruppert and Wand (1992). In the ordinary regression context these authors showed that local polynomial kernel regression has several attractive mathematical properties. This is particularly the case when the polynomial is of odd degree since the asymptotic bias near the boundary of the support of the covariates can be shown to be of the same order of magnitude as that of the interior. Since, in applications, the boundary region will often include 20% or more of the data, this is a very appealing feature. This is not the case for the Nadaraya-Watson kernel estimator since it corresponds to degree zero fitting. In addition, the asymptotic bias of odd degree polynomial fits at a point x depends on x only through a higher order derivative of the regression function itself, which allows for simple interpretation and expressions for the asymptotically optimal bandwidth. We are able to show that these properties carry over to generalized linear models.

When fitting local polynomials an important choice that has to be made is the degree of the polynomial. Boundary bias considerations indicate that one should, at the least, fit local lines. However, for estimation of regions of high curvature of the true function, such as at peaks and valleys, local line estimators can have a substantial amount of bias. This problem can be alleviated by fitting higher degree polynomials such as quadratics and cubics, although there are costs in terms of increased variance and computational complexity that need to be considered. Nevertheless, in many examples that we have tried we have noticed that gains can often be made using higher degree fits, and this is our main motivation for extending beyond linear fits.

In Section 2 we present some notation for the generalized linear model and quasi-likelihood functions. Section 3 deals with the locally weighted maximum quasi-likelihood approach to local polynomial fitting. We discuss the problem of choosing the bandwidth in Section 4. A real data example is presented in Section 5 and a summary of our findings and further discussion is given in Section 6.

2. GENERALIZED LINEAR MODELS AND QUASI-LIKELIHOOD FUNCTIONS

In our definition of generalized linear models and quasi-likelihood functions we will follow the notation of McCullagh and Nelder (1988). Let $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ be a set of independent random pairs where, for each i, Y_i is a scalar response variable and $\mathbf{X}_i$ is an $\mathbb{R}^d$-valued vector of covariates having density f with support $\text{supp}(f) \subseteq \mathbb{R}^d$. Let $(\mathbf{X}, Y)$ denote a generic member of the sample. Then we will say that the conditional density of

Y given $\mathbf{X} = \mathbf{x}$ belongs to a *one-parameter exponenial family* if

$$f_{Y|\mathbf{X}}(y|\mathbf{x}) = \exp\{y\theta(\mathbf{x}) - b(\theta(\mathbf{x})) + c(y)\}$$

for known functions b and c. The function θ is usually called the *canonical* or *natural* parameter. In parametric generalized linear models it is usual to model a transformation of the regression function $\mu(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$ as linear in $\mathbf{x}$, that is,

$$\eta(\mathbf{x}) = \beta_0 + \sum_{i=1}^{d} \beta_{1i} x_i \quad \text{where} \quad \eta(\mathbf{x}) = g(\mu(\mathbf{x}))$$

and g is the *link* function. If $g = (b')^{-1}$ then g is called the *canonical link* since $b'(\theta(\mathbf{x})) = \mu(\mathbf{x})$. A further noteworthy result for the one-parameter exponential family is $\text{var}(Y|\mathbf{X} = \mathbf{x}) = b''(\theta(\mathbf{x}))$.

A simple example of this set-up arises when the Y_i are binary variables, in which case it is easily shown that

$$f_{Y|\mathbf{X}}(y|\mathbf{x}) = \exp\{y \ln(\mu(\mathbf{x})/(1 - \mu(\mathbf{x}))) + \ln(1 - \mu(\mathbf{x}))\}$$

The canonical parameter in this case is $\theta(\mathbf{x}) = \text{logit}(\mu(\mathbf{x}))$ and the logit function is the canonical link. However, one can assume that $\eta(\mathbf{x}) = g(\mu(\mathbf{x}))$ is linear where g is any increasing function on $(0, 1)$. Common examples include $g(u) = \Phi^{-1}(u)$, where Φ is the standard normal distribution function, and $g(u) = \ln\{-\ln(1-u)\}$. These are often referred to as the *probit* link and the *complementary log-log* link respectively.

The conditional density $f_{Y|\mathbf{X}}$ can be written in terms of η as

$$f_{Y|\mathbf{X}}(y|\mathbf{x}) = \exp[y(g \circ b')^{-1}(\eta(\mathbf{x})) - b\{(g \circ b')^{-1}(\eta(\mathbf{x}))\} + c(y)]$$

where $\circ$ denotes function composition.

There are many practical circumstances where even though the full likelihood is unknown, one can specify the relationship between the mean and variance. In this situation, estimation of the mean can be achieved by replacing the conditional log-likelihood $\ln f_{Y|\mathbf{X}}(y|\mathbf{x})$ by a *quasi-likelihood function* $Q(\mu(\mathbf{x}), y)$. If the conditional variance is modeled as $\text{var}(Y|\mathbf{X} = \mathbf{x}) = V(\mu(\mathbf{x}))$ for some known positive function V then the corresponding quasi-likelihood function $Q(w, y)$ satisfies

$$\frac{\partial}{\partial w} Q(w, y) = \frac{y - w}{V(w)} \tag{2.1}$$

(Wedderburn 1974, McCullagh and Nelder 1988, Chapter 9). The quasi-score (2.1) possesses properties similar to those of the usual likelihood score function. That is, it satisfies the first two moment conditions of Bartlett's identities. Quasi-likelihood methods behave analogously to the usual likelihood methods and thus are reasonable substitutes when the likelihood function is not available. Note the log-likelihood of the one-parameter exponential family is a special case of a quasi-likelihood function with $V = b'' \circ (b')^{-1}$.

3. LOCALLY WEIGHTED MAXIMUM QUASI-LIKELIHOOD

3.1 Single Covariate Case

Because of its generality we will present our results in the quasi-likelihood context. Results for the exponential family and generalized linear models follow as a special case. We will introduce the idea of local polynomial fitting for the case where $d = 1$. Multivariate extensions will be discussed in Section 3.2.

As we mentioned in the preceding section, the typical parametric approach to estimating $\mu(x)$ is to assume that $\eta(x) = g(\mu(x))$ is a linear function of x, $\eta(x) = \beta_0 + \beta_1 x$ say. An obvious extension of this idea is to suppose that $\eta(x) = \beta_0 + \ldots + \beta_p x^p$, that is, $\eta(x)$ is a pth degree polynomial in x. One would then estimate $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T$ by maximizing the quasi-likelihood

$$\sum_{i=1}^{n} Q(g^{-1}(\beta_0 + \ldots + \beta_p X_i^p), Y_i). \tag{3.1}$$

To enhance the flexibility of the fitted curve, the number of parameters $(p+1)$ is typically large. However, this means that the estimated coefficients are subject to large variability. Also there is a greater threat of numerical instability caused by an ill-conditioned design matrix. A more appealing approach is to estimate $\eta(x)$ by *locally* fitting a low-degree polynomial at x, (e.g. $p = 1$ or 3). This involves centering the data about x and weighting the quasi-likelihood in such a way that it places more emphasis on those observations closest to x. The estimate of $\eta(x)$ is then the height of the local polynomial fit.

There are many possible strategies for weighting the conditional quasi-likelihood, see Hastie and Tibshirani (1990, Chapter 2) for several of these. Because of their mathematical simplicity we will use kernel weights. Let K be a symmetric probability density having compact support and let $K_h(z) = K(z/h)/h$ be a rescaling of K. The parameter $h > 0$ is usually called the *bandwidth* or *window width* and in the mathematical analysis will be taken to be a sequence that depends on n. Each observation in the quasi-likelihood is given the weight $K_h(X_i - x)$. The local polynomial kernel estimator of $\eta(x)$ is then given

by

$$\hat{\eta}(x; p, h) = \hat{\beta}_0$$

where $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)^T$ maximizes

$$\sum_{i=1}^{n} Q(g^{-1}(\beta_0 + \ldots + \beta_p(X_i - x)^p), Y_i) K_h(X_i - x). \tag{3.2}$$

We will assume that the maximizer exists. This can be easily verified for standard choices of Q. The conditional mean $\mu(x)$ can then be estimated by applying the inverse of the link function to give

$$\hat{\mu}(x; p, h) = g^{-1}(\hat{\eta}(x; p, h)).$$

Note that in the case $p = 0$, the estimator for $\mu(x)$ is simply a weighted average:

$$\hat{\mu}(x; 0; h) = \sum_{i=1}^{n} K_h(X_i - x) Y_i \Big/ \sum_{i=1}^{n} K_h(X_i - x). \tag{3.3}.$$

Although this estimator is intuitive, it nevertheless suffers serious drawbacks such as large biases and low efficiency (Fan 1992b).

Figure 1 and its accompanying discussion in Section 1 show how simple $\hat{\eta}(\cdot; p, h)$ and $\hat{\mu}(\cdot; p, h)$ are to understand intuitively. Notice that the bandwidth h governs the amount of smoothing being performed. For h tending to zero the estimated function will tend to interpolate the data and its variance will increase. On the other hand, for increasing h the estimate of η will tend towards the pth degree polynomial fit corresponding to (3.1) which, in general, corresponds to an increase in bias. This latter property is very appealing since it means that the proposed estimators can be viewed as extensions of the traditional parametric approach.

Local polynomial fitting also provides consistent estimates of higher order derivatives of η through the coefficients of the higher order terms in the polynomial fit. Thus, for $0 \leq r \leq p$ we can define the estimator of $\eta^{(r)}(x)$ to be

$$\hat{\eta}_r(x; p, h) = r!\hat{\beta}_r.$$

We will now investigate the asymptotic properties of $\hat{\eta}_r(x; p, h)$ and $\hat{\mu}(x; p, h)$. These are different for x lying in the interior of $\text{supp}(f)$ than for x lying near the boundary. Suppose that K is supported on $[-1, 1]$. Then the support of $K_h(x - \cdot)$ is $\mathcal{E}_{x,h} = \{z : |z - x| \leq h\}$. We will call x an interior point of $\text{supp}(f)$ if $\mathcal{E}_{x,h} \subseteq \text{supp}(f)$. Otherwise x will be called a boundary point. If $\text{supp}(f) = [a, b]$ then x is a boundary point if and

only if $x = a + \alpha h$ or $x = b - \alpha h$ for some $0 \leq \alpha < 1$. Finally, let $\mathcal{D}_{x,h} = \{z : x - hz \in \text{supp}(f)\} \cap [-1, 1]$. Then x is an interior point if and only if $\mathcal{D}_{x,h} = [-1, 1]$.

For any measurable set $\mathcal{A} \in \mathbb{R}$ define $\nu_\ell(\mathcal{A}) = \int_{\mathcal{A}} z^\ell K(z)\,dz$. Let $\mathbf{N}_p(\mathcal{A})$ be the $(p+1) \times (p+1)$ matrix having (i, j) entry equal to $\nu_{i+j-2}(\mathcal{A})$ and $\mathbf{M}_{r,p}(z; \mathcal{A})$ be the same as $\mathbf{N}_p(\mathcal{A})$, but with the $(r+1)$st column replaced by $(1, z, \ldots, z^p)^T$. Then for $|N_p(\mathcal{A})| \neq 0$ define

$$K_{r,p}(z; \mathcal{A}) = r!\{|\mathbf{M}_{r,p}(z; \mathcal{A})|/|\mathbf{N}_p(\mathcal{A})|\}K(z). \tag{3.4}$$

In the case where $\mathcal{A} \supseteq [-1, 1]$ we will suppress the $\mathcal{A}$ and simply write ν_ℓ, τ_ℓ, $\mathbf{N}_p$, $\mathbf{M}_{r,p}$ and $K_{r,p}$. One can show that $(-1)^r K_{r,p}$ is an order (r, s) kernel as defined by Gasser, Müller and Mammitzsch (1985), where $s = p+1$ if $p-r$ is odd and $s = p+2$ if $p-r$ is even. This family of kernels is useful for giving concise expressions for the asymptotic distribution of $\hat{\eta}_r(x; p, h)$ for x lying both in the interior of supp(f) and near its boundaries. Also, let

$$\rho(x) = \left\{g'(\mu(x))^2 V(\mu(x))\right\}^{-1}.$$

Note that when the model belongs to a one-parameter exponential family and the canonical link is used then $g'(\mu(x)) = 1/\text{var}(Y|X = x)$ and $\rho(x) = \text{var}(Y|X = x)$, if the variance V is correctly specified. The asymptotic variance of $\hat{\eta}_r(x; p, h)$ will depend on quantities of the form

$$\sigma^2_{r,p}(x; K, \mathcal{A}) = \text{var}(Y|X = x) g'(\mu(x))^2 f(x)^{-1} \int_{\mathcal{A}} K_{r,p}(z; \mathcal{A})^2\,dz.$$

Again, this will be abbreviated to $\sigma^2_{r,p}(x; K)$ when $\mathcal{A} \supseteq [-1, 1]$. The following theorems are proved in the Appendix.

Theorem 1a. Let $p - r > 0$ be odd and suppose that Conditions (1)–(5) stated in the Appendix are satisfied. Assume that $h = h_n \to 0$ and $nh^3 \to \infty$ as $n \to \infty$. If x is a fixed point in the interior of supp(f) then

$$\sqrt{nh^{2r+1}}\sigma_{r,p}(x; K)^{-1} \tag{3.5}$$

$$\times \left[\hat{\eta}_r(x; p, h) - \eta^{(r)}(x) - \left\{\int z^{p+1} K_{r,p}(z)\,dz\right\}\left\{\frac{\eta^{(p+1)}(x)}{(p+1)!}\right\} h^{p-r+1}\{1 + O(h)\}\right] \to_{\mathrm{D}} N(0, 1).$$

If $x = x_n$ is of the form $x = x_\partial + ch$ where x_∂ is a point on the boundary of supp(f) and $c \in [-1, 1]$ then (3.5) holds with $\sigma^2_{r,p}(x; K)$ and $\int z^{p+1} K_{r,p}(z)\,dz$ replaced by $\sigma^2_{r,p}(x; K, \mathcal{D}_{x,h})$ and $\int_{\mathcal{D}_{x,h}} z^{p+1} K_{r,p}(z)\,dz$ respectively.

Theorem 1b. Let $p > 0$ and $p - r \geq 0$ be even and suppose that Conditions (1)–(5) stated in the Appendix are satisfied. Assume that $h = h_n \to 0$ and $nh^3 \to \infty$ as $n \to \infty$. If x a fixed point in the interior of supp(f) then,

$$\sqrt{nh^{2r+1}}\sigma_{r,p}(x;K)^{-1}\left(\hat{\eta}_r(x;p,h) - \eta^{(r)}(x) - \left[\left\{\int z^{p+2}K_{r,p}(z)\,dz\right\}\left\{\frac{\eta^{(p+2)}(x)}{(p+2)!}\right\}\right.\right.$$
$$\left.\left. + \left\{\int z^{p+2}K_{r,p}(z)\,dz - r\int z^{p+1}K_{r-1,p}(z)\,dz\right\}\left\{\frac{\eta^{(p+1)}(x)(\rho f)'(x)}{(\rho f)(x)(p+1)!}\right\}\right]h^{p-r+2}\{1+O(h)\}\right)$$
$$\to_{\mathrm{D}} N(0,1).$$

If $x = x_n$ is of the form $x = x_\partial + ch$ where x_∂ is a point on the boundary of supp(f) for some $c \in [-1,1]$ then

$$\sqrt{nh^{2r+1}}\sigma_{r,p}(x;K,\mathcal{D}_{x,h})^{-1} \tag{3.6}$$
$$\times\left[\hat{\eta}_r(x;p,h) - \eta^{(r)}(x) - \left\{\int_{\mathcal{D}_{x,h}} z^{p+1}K_{r,p}(z)\,dz\right\}\left\{\frac{\eta^{(p+1)}(x)}{(p+1)!}\right\}h^{p-r+1}\{1+O(h)\}\right] \to_{\mathrm{D}} N(0,1).$$

The asymptotic distribution of $\hat{\eta}(x;0,h)$ admits a slightly more complicated form then for the for $p > 0$ cases given by Theorems 1a and 1b. However, observe that $\hat{\eta}(x;0,h) = g(\hat{\mu}(x;0,h))$ where $\hat{\mu}(x;0,h)$ has the explicit weighted average expression given by (3.3). Therefore, standard results for the conditional mean squared error of $\hat{\mu}(x;0,h)$ (see e.g. Ruppert and Wand 1992) lead to:

Theorem 1c. Let $p = 0$ and suppose that Conditions (1)–(5) stated in the Appendix are satisfied. Assume that $h = h_n \to 0$ and $nh \to \infty$ as $n \to \infty$. If x is a fixed point in the interior of supp(f) then

$$E\{\hat{\eta}(x;0,h) - \eta(x)|X_1,\ldots,X_n\}$$
$$= \left\{\int z^2K(z)\,dz\right\}\{\tfrac{1}{2}\eta''(x) + \eta'(x)f'(x)/f(x)\}g'(\mu(x))h^2\{1+o_P(1)\}$$

and $\mathrm{var}\{\hat{\eta}(x;0,h)|X_1,\ldots,X_n\} = \sigma^2_{0,0}(x;K)n^{-1}h^{-1}\{1+o_P(1)\}$. If $x = x_n$ is of the form $x = x_\partial + ch$ where x_∂ is a point on the boundary of supp(f) for some $c \in [-1,1]$ then

$$E\{\hat{\eta}(x;0,h) - \eta(x)|X_1,\ldots,X_n\} = \left\{\int_{\mathcal{D}_{x,h}} zK(z)\,dz\right\}\eta'(x)h\{1+o_P(1)\}$$

and $\mathrm{var}\{\hat{\eta}(x;0,h)|X_1,\ldots,X_n\} = \sigma^2_{0,0}(x;K,\mathcal{D}_{x,h})n^{-1}h^{-1}\{1+o_P(1)\}$.

Remark 1. While Theorem 1 describes the asymptotic behavior of $\hat{\eta}_r(x;p,h)$ for general r it is the case $r = 0$, which corresponds to the estimation of η itself, that is

of primary interest. The expressions involving $\int z^{j+1} K_{r,p}(z)\,dz$ and $\sigma^2_{r,p}(x;K)$ can be interpreted as the asymptotic bias and asymptotic variance of $\hat{\eta}(x;p,h)$ respectively. Notice that for odd p the asymptotic bias is of the same order of magnitude near the boundary as in the interior. This is a mathematical quantification of boundary bias not being a serious problem for odd degree polynomial fits. The asymptotic bias and variance can be combined to give the asymptotic mean squared error (AMSE) of $\hat{\eta}(x;p,h)$ for interior points as

$$\mathrm{AMSE}\{\hat{\eta}(x;p,h)\} = \left\{\int z^{p+1} K_{0,p}(z)\,dz\right\}^2 \left\{\frac{\eta^{(p+1)}(x)}{(p+1)!}\right\}^2 h^{2p+2} + \sigma^2_{0,p}(x;K) n^{-1} h^{-1}$$

for p odd and

$$\mathrm{AMSE}\{\hat{\eta}(x;p,h)\} = \left\{\int z^{p+2} K_{0,p}(z)\,dz\right\}^2 \left\{\frac{(\rho f)'(x)\eta^{(p+1)}(x)}{(\rho f)(x)(p+1)!} + \frac{\eta^{(p+2)}(x)}{(p+2)!}\right\}^2 h^{2p+4} + \sigma^2_{0,p}(x;K) n^{-1} h^{-1}$$

for p even and non-zero. Note that the bandwidth that minimizes $\mathrm{AMSE}\{\hat{\eta}(x;p,h)\}$ is of order $n^{-1/(2p+3)}$ for p odd and of order $n^{-1/(2p+5)}$ for p even. For a given sequence of bandwidths, the asymptotic variance is always $O(n^{-1}h^{-1})$, while the order of the bias tends to decrease as the degree of the polynomial increases. For instance, a linear fit gives bias of order h^2, a quadratic fit gives bias of order h^4, and a cubic gives bias of order h^4. Therefore, there is a significant reduction in bias for quadratic or cubic fits compared to linear fits, particularly when estimating η at peaks and valleys, where typically $\eta''(x) \neq 0$. However, it should also be realized that if globally optimal bandwidths are used then it is possible for local linear fits to outperform higher degree fits, even at peaks and valleys. This point is made visually by Marron (1992) in a closely related context.

With even degree polynomial fits, the boundary bias dominates the interior bias and boundary kernels would be required to make their asymptotic orders the same. This is not the case for odd degree polynomial fits. Furthermore, the expression for the interior bias for even degree fits is complicated, involving three derivatives of η instead of just one. Therefore, we do not recommend using even degree fitting. For further discussion on this matter see Fan and Gijbels (1992).

Remark 2. Staniswalis (1989) and Sevirini and Staniswalis (1992) have considered the case for the local constant fit (p=0) by using the explicit formula (3.3). However, significant gains can be made by using local linear or higher order fits, especially near the boundaries.

Remark 3. In the Appendix we actually derive the asymptotic joint distribution of the $\hat{\eta}_r(x;p,h)$ for all $r \leq p$. In addition, we are able to give expressions for the second order bias terms in (3.5) and (3.6).

Since $\hat{\mu}_r(x;p,h) = g^{-1}(\hat{\eta}(x;p,h))$ it is straightforward to derive:

Theorem 2. Under the conditions of Theorem 1 the error $\hat{\mu}(x;p,h) - \mu(x)$ has the same asymptotic behavior as $\hat{\eta}_0(x;p,h) - \eta(x)$ given in Theorem 1 with $r = 0$ except that the asymptotic bias is divided by $g'(\mu(x))$ and the asymptotic variance is divided by $g'(\mu(x))^2$. In addition to the conditions of Theorem 1, we require that $nh^{4p+5} \to 0$ for Theorems 1a and 1b, that $nh^3 \to \infty$ for the first statement of Theorem 1c, and that $nh^2 \to \infty$ for the second statement of Theorem 1c.

Remark 4. The additional conditions on n and h in Theorem 2 hold when the rate of convergence of h is chosen optimally. Notice that the interior variance is asymptotic to

$$n^{-1}h^{-1}\text{var}(Y|X = x)f(x)^{-1} \int K_{0,p}(z)^2\, dz.$$

The dependence of the asymptotic variance of $\hat{\mu}(x;p,h)$ on $\text{var}(Y|X = x)f(x)^{-1}$ reflects the intuitive notion of there being more variation in the estimate of the conditional mean for higher values of the conditional variance and regions of lower density of the covariate.

Example 1. Consider the case of a binary response variable with Bernoulli conditional likelihood. In this case $Q(w,y) = y\ln w + (1-y)\ln(1-w) = y\text{logit}(w) + \ln(1-w)$ and the canonical link is $g(u) = \text{logit}(u)$. For the canonical link with properly specified variance we have $\rho(x) = \text{var}(Y|X = x) = \mu(x)\{1-\mu(x)\}$ and $\sigma^2_{r,p}(x;K,\mathcal{A}) = [\mu(x)\{1-\mu(x)\}f(x)]^{-1} \int_{\mathcal{A}} K_{r,p}(z;\mathcal{A})^2\, dz$.

If the probit link $g(u) = \Phi^{-1}(u)$ is used instead then we obtain $\rho(x) = [\mu(x)\{1-\mu(x)\}]^{-1}\phi(\Phi^{-1}(\mu(x)))^2$ and $\sigma^2_{r,p}(x;K,\mathcal{A}) = \mu(x)\{1-\mu(x)\}\phi(\Phi^{-1}(\mu(x)))^{-2} \int_{\mathcal{A}} K_{r,p}(z;\mathcal{A})^2\, dz$ where ϕ is the standard normal density function.

Example 2. Consider the case of non-negative integer response with Poisson conditional likelihood. In this case $Q(w,y) = y\ln w - w$ and the canonical link is $g(u) = \ln(u)$. For the canonical link with properly specified variance we have $\rho(x) = \mu(x)$ and $\sigma^2_{r,p}(x;K,\mathcal{A}) = \{\mu(x)f(x)\}^{-1} \int_{\mathcal{A}} K_{r,p}(z;\mathcal{A})^2\, dz$.

If instead the conditional variance is modeled as being proportional to the square of the conditional mean, that is $V(\mu(x)) = \gamma^2\mu(x)^2$ where γ^2 is the coefficient of variation, then we have $Q(w,y) = (-y/w - \ln w)/\gamma^2$. If the logarithmic link is used then we have $\rho(x) = 1/\gamma^2$ and $\sigma^2_{r,p}(x;K,\mathcal{A}) = \gamma^2 f(x)^{-1} \int_{\mathcal{A}} K_{r,p}(Z;\mathcal{A})^2\, dz$, provided $V(\mu(x)) = \text{var}(Y|X = x)$.

3.2 Multiple Covariate Case

The extension of the local kernel estimator to the case of multiple covariates is straightforward for the important case of local linear fitting. The treatment of higher degree polynomial fits in the multivariate case requires careful notation to keep the expressions simple. Higher degree polynomial fits for multiple covariates can also be very computationally daunting, so we will concentrate on the local linear case in this section. For a flavor of asymptotic analyses of multivariate higher order polynomial fits see Ruppert and Wand (1992) where multivariate quadratic and cubic fits are considered for the ordinary local least squares regression estimator.

A problem that must be faced when confronted with multiple covariates is the well known "curse of dimensionality" – the fact that the performance of nonparametric smoothing techniques deteriorate as the dimensionality increases. One way of overcoming this problem is to assume that the model is additive in the sense that

$$\eta(\mathbf{x}) = \eta_1(x_1) + \ldots + \eta_d(x_d)$$

where each η_i is a univariate function corresponding to the ith coordinate direction. This is the *generalized additive model* as described by Hastie and Tibshirani (1990) and it is recommended that each η_i be estimated by an appropriate scatterplot smoother. There are, however, many situations where the additivity assumption is not valid, in which case multivariate smoothing of the type presented in this section is appropriate.

Throughout this section we will take K to be a d-variate kernel with the properties that $\int K(\mathbf{z})\,d\mathbf{z} = 1$, $\int \mathbf{z}\mathbf{z}^T K(\mathbf{z})\,d\mathbf{z} = \nu_2 \mathbf{I}$ where $\nu_2 = \int_{\mathbb{R}^d} z_i^2 K(\mathbf{z})\,d\mathbf{z}$ is non-zero and independent of i. In the multivariate case we define $K_{\mathbf{H}}(\mathbf{z}) = |\mathbf{H}|^{-1/2} K(\mathbf{H}^{-1/2}\mathbf{z})$ where $\mathbf{H}$ is a positive definite matrix of bandwidths. The d-variate local linear kernel estimator of $\eta(\mathbf{x})$ is $\hat{\eta}(\mathbf{x};\mathbf{H}) = \hat{\beta}_0$ where

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\boldsymbol{\beta}}_1 \end{bmatrix} = \arg\max_{(\beta_0,\boldsymbol{\beta}_1)^T} \sum_{i=1}^{n} Q(g^{-1}(\beta_0 + \boldsymbol{\beta}_1^T(\mathbf{X}_i - \mathbf{x})), Y_i) K_{\mathbf{H}}(\mathbf{X}_i - \mathbf{x}).$$

The corresponding estimator for $\mu(\mathbf{x})$ is $\hat{\mu}(\mathbf{x};\mathbf{H}) = g^{-1}(\hat{\eta}(\mathbf{x};\mathbf{H}))$.

Before giving the asymptotic distributions of these estimators we need to extend some definitions to the multivariate setting. A point $\mathbf{x} \in \mathbb{R}^d$ will be called an interior point of $\text{supp}(f)$ if and only if $\{\mathbf{z} : \mathbf{H}^{-1/2}(\mathbf{x} - \mathbf{z}) \in \text{supp}(K)\} \subseteq \text{supp}(f)$. The Hessian matrix of a η at a point $\mathbf{x}$ will be denoted by $\mathcal{H}_\eta(\mathbf{x})$.

Theorem 3. Suppose that Conditions (1) and (3) in the Appendix are satisfied and f and all entries of $\mathcal{H}_\eta$ are continuous at $\mathbf{x}$. Also suppose the $n^{-1}|\mathbf{H}|^{-3/2}$ and all entries of $\mathbf{H}$ tend to zero in such a way that $\mathbf{H}$ remains positive definite and the ratio of the largest

to smallest eigenvalues of $\mathbf{H}$ remain bounded. If x is an interior point of supp(f) then

$$\sqrt{n|\mathbf{H}|^{1/2}}\Big\{\mathrm{var}(Y|\mathbf{X}=\mathbf{x})g'(\mu(\mathbf{x}))^2 f(\mathbf{x})^{-1}\int K(z)^2\,dz\Big\}^{-1/2}$$

$$\times\left[\hat{\eta}(\mathbf{x};\mathbf{H})-\eta(\mathbf{x})-\tfrac{1}{2}\nu_2\mathrm{tr}\{\mathbf{H}\mathcal{H}_\eta(\mathbf{x})\}\{1+o(1)\}\right]\to_{\mathrm{D}} N(0,1).$$

Theorem 4. Under the conditions of Theorem 3 and g^{-1} being differentiable, if $\mathbf{x}$ is an interior point of supp(f) then the error $\hat{\mu}(\mathbf{x};\mathbf{H})-\mu(\mathbf{x})$ has the same asymptotic normal distribution as $\hat{\eta}(\mathbf{x};\mathbf{H})-\eta(\mathbf{x})$ given in Theorem 3 except that the bias is divided by $g'(\mu(\mathbf{x}))$ and the variance is divided by $g'(\mu(\mathbf{x}))^2$.

4. BANDWIDTH SELECTION

As is the case for all nonparametric curve estimators the smoothing parameter, in this case the bandwidth, plays a very important role in the trade-off between reducing bias and variance. Often the user will be able to choose the bandwidth satisfactorily by eye. However it is also desirable to have a reliable data-driven rule for selecting the value of h.

In simpler curve estimation settings such as kernel density estimation there has been considerable recent progress in the objective choice of the bandwidth. See Park and Marron (1990) for a discussion and comparison of several of these developments.

A straightforward bandwidth selection idea is that of cross-validation. The extension of this idea to the generalized linear model setting is trivial. However, cross-validation performs poorly in simpler settings, exhibiting a large amount of sample variation, and this behavior carries over to this setting. Therefore, we do not view cross-validation as a sensible bandwidth selection rule for practice.

An alternative bandwidth selection idea that makes use of the results derived in Sections 2 and 3 is that based on "plugging-in" estimates of unknown quantities to a formula for the asymptotically optimal bandwidth. This has been shown to be more stable in both theoretical and practical performance (see e.g. Park and Marron 1990, Sheather and Jones 1991). Indeed, Fan and Marron (1992) show that, in the density estimation case, the plug-in selector is an asymptotically efficient method from a semiparametric point of view. For odd degree polynomial fits the simplicity of the bias and variance expressions indicates that it would be reasonably straightforward to extend the plug-in ideas used in kernel density estimation to estimation of η and hence μ. In the following suppose that p

is odd. A convenient approximate error criterion for $\hat{\eta}(\cdot; p, h)$ is

$$\text{AMISE}\{\hat{\eta}(\cdot; p, h)\} = h^{2p+2} \left\{ \int z^{p+1} K_{0,p}(z)\, dz \right\}^2 \int \left\{ \frac{\eta^{(p+1)}(x)}{(p+1)!} \right\}^2 f(x) w(x)\, dx + n^{-1} h^{-1} \int \sigma^2_{0,p}(x; K) f(x) w(x)\, dx .$$

We will call this the asymptotic mean integrated squared error since it is obtained from the above AMSE expression by integrating over $\text{supp}(f)$. The design density f and weight function w are included for stability purposes. With respect to this criterion the optimal bandwidth is

$$h_{\text{AMISE}} = \left[\frac{\{(p+1)!\}^2 \int \sigma^2_{0,p}(x, K) f(x) w(x)\, dx}{(2p+2) \left\{ \int z^{p+1} K_{0,p}(z)\, dz \right\}^2 \left\{ \int \eta^{(p+1)}(x)^2 f(x) w(x)\, dx \right\} n} \right]^{1/(2p+3)} . \quad (4.1)$$

A plug-in bandwidth selection rule that replaces the unknown quantities by other local polynomial kernel estimators would be a worthwhile future project. However, it is also possible to use (4.1) to motivate "rough-and-ready" bandwidth selectors as well. For example, one could define $\hat{h}_q$ to be the bandwidth that is obtained by replacing η in (4.1) by a qth degree polynomial parametric fit, where $q \geq p+1$. While such a selector would not have any asymptotic optimality properties, it should return a bandwidth that is reasonable for a wide range of functions that arise in practice.

5. APPLICATION TO BURNS DATA

We applied the local linear estimate to a data set consisting of dispositions of burns victims and several covariates. The data were collected at the University of Southern California General Hospital Burn Center. The binary response variable Y is 1 for those victims who survive their burns and zero otherwise. As a predictor we used

$$X = \ln(\text{area of third degree burn} + 1).$$

Since children have significantly smaller body areas only the 435 adults (older than 17 years) who suffered third degree burns were considered in our study.

We applied the local linear kernel estimator with the Bernoulli likelihood and logit link function to these data. The kernel was the standard normal density. The maximization was performed using Newton-Raphson iteration. We used the data-driven bandwidth $\hat{h}_2 = 1.242$ as defined in the previous section. Figure 2 shows (a) the kernel estimate of $\eta(x) = \text{logit}\{P(Y = 1|X = x)\}$ and (b) $\mu(x) = P(Y = 1|X = x)$, the survival probability

for a given third degree burn area (measured in square centimeters). The plus signs represent the data.

Figure 2. (a) Local linear kernel estimate of logit P(Y=1|X=x) for the burns data as described in the text. The number of observations is n=435 and the bandwidth is $\hat{h}_2$=1.24. (b) Estimate of P(Y=1|X=x) obtained from the estimate in (a) by application of the inverse logit transformation. The plus signs indicate the observed data.

As expected, the survival probability decreases although Figure 2(a) suggests that this decrease is slightly non-linear in the logit space.

6. DISCUSSION

We have shown that, in addition to their simplicity, the class of polynomial kernel estimators have many attractive mathematical properties. This is especially the case for low odd degree polynomial fits such as lines and cubics. While we have shown the applicability of these estimators through a real data example, there is still room for further study of their practical implementation and performance. Two important questions that need to be addressed are the fast computation of the estimator and automatic selection of the bandwidth. The solution to the first of these problems could involve a version of the method of scoring for the minimization of (3.2) and application of discretization ideas for computation of the kernel-type estimators required for each iteration (e.g. Härdle and Scott, 1992). Bandwidth selection rules of the type considered by Park and Marron (1990) and Sheather and Jones (1991) could also be developed. The theory for estimating higher-order derivatives of η derived in this article would be important for this problem. Each of these would be fruitful topics for future research.

APPENDIX: PROOFS OF THEOREMS

Since $\hat{\boldsymbol{\beta}}$ is calculated using X_i near x, we would expect that

$$\hat{\beta}_0 + \hat{\beta}_1(X_i - x) + \cdots + \hat{\beta}_p(X_i - x)^p \sim \eta(x) + \eta'(x)(X_i - x) + \cdots + \eta^{(p)}(x)(X_i - x)^p/p!.$$

Therefore, one would expect that $r!\hat{\beta}_r \to \eta^{(r)}(x)$. We thus study the asymptotics of

$$\hat{\boldsymbol{\beta}}^* \equiv (nh)^{1/2} \begin{bmatrix} \hat{\beta}_0 - \eta(x) \\ h\{\hat{\beta}_1 - \eta'(x)\} \\ \vdots \\ h^p\{p!\hat{\beta}_p - \eta^{(p)}(x)\} \end{bmatrix}$$

so that each component has the same rate of convergence. Let $\mathbf{Q}_p(\mathcal{A})$ and $\mathbf{T}_p(\mathcal{A})$ be the $(p+1)\times(p+1)$ matrices having (i,j)th entry equal to $\nu_{i+j-1}(\mathcal{A})$ and $\int_{\mathcal{A}} z^{i+j-2}K(z)^2\,dz$ respectively. Also define

$$\mathbf{D} = \mathrm{diag}(1, 1/1!, \cdots, 1/p!), \qquad \mathbf{\Sigma}_x(\mathcal{A}) = \rho(x)f(x)\mathbf{D}\mathbf{N}_p(\mathcal{A})\mathbf{D},$$

$$\mathbf{\Gamma}_x(\mathcal{A}) = \frac{f(x)\mathrm{var}(Y|X=x)}{\{V(\mu(x))g'(\mu(x))\}^2}\mathbf{D}\mathbf{T}_p(\mathcal{A})\mathbf{D}, \qquad \mathbf{\Lambda}_x(\mathcal{A}) = (\rho f)'(x)\mathbf{D}\mathbf{Q}_p(\mathcal{A})\mathbf{D},$$

$$a_{1,i}(\mathcal{A}) = \frac{\eta^{(p+1)}(x)}{(p+1)!}\int_{\mathcal{A}} z^{p+1}K_{i-1,p}(z;\mathcal{A})\,dz,$$

and

$$\begin{aligned} a_{2,i} =& \frac{\eta^{(p+2)}(x)}{(p+2)!}\int_{\mathcal{A}} z^{p+2}K_{i-1,p}(z;\mathcal{A})\,dz \\ &+ \frac{\eta^{p+1}(x)}{(p+1)!}\frac{(\rho f)'(x)}{(\rho f)(x)}\Big\{\int_{\mathcal{A}} z^{p+2}K_{i-1,p}(z;\mathcal{A})\,dz \\ &- (i-1)\int_{\mathcal{A}} z^{p+1}K_{i-2,p}(z;\mathcal{A})\,dz - \frac{1}{p!}\int_{\mathcal{A}} z^{p+1}K_{i-1,p}(z;\mathcal{A})\,dz\int_{\mathcal{A}} z^{p+1}K_{p,p}(z;\mathcal{A})\,dz\Big\}. \end{aligned}$$

Let $\mathbf{b}_x(\mathcal{A})$ be the $(p+1)\times 1$ vector having ith entry equal to

$$nh^{2p+3}a_{1,i} + nh^{2p+5}a_{2,i}.$$

Lastly, let $q_i(x,y) = (\partial^i/\partial x^i)Q(g^{-1}(x),y)$. Note that that q_i is linear in y for fixed x and that

$$q_1(\eta(x),\mu(x)) = 0 \quad \text{and} \quad q_2(\eta(x),\mu(x)) = -\rho(x). \tag{A.1}$$

Conditions.

(1) The function $q_2(x,y) < 0$ for $x \in \mathbb{R}$ and y in the range of the response variable.

(2) The functions f', $\eta^{(p+2)}$, $\mathrm{var}(Y|X=\cdot)$, V'' and g''' are continuous.

(3) For each $x \in \mathrm{supp}(f)$, $\rho(x)$, $\mathrm{var}(Y|X=x)$, and $g'(\mu(x))$ are non-zero.

(4) The kernel K is a symmetric probability density with support $[-1,1]$.

(5) For each point x_∂ on the boundary of $\mathrm{supp}(f)$ there exists an interval $\mathcal{C}$ having nonnull interior such that $\inf_{x\in\mathcal{C}} f(x) > 0$.

Note that Condition (2) implies that q_1, q_2, q_3, ρ', and μ' are continuous, and that Conditions (1) and (3) imply that ρ is strictly positive over $\mathrm{supp}(f)$.

Theorems 1a and 1b are simple consequences of the following Main Theorem.

Main Theorem. Suppose that the above conditions hold and that $h = h_n \to 0$, $nh^3 \to \infty$ as $n \to \infty$. If x is an interior point of $\text{supp}(f)$ and $p > 0$ then

$$\left\{\boldsymbol{\Sigma}_x([-1,1])^{-1}\boldsymbol{\Gamma}_x([-1,1])\boldsymbol{\Sigma}_x([-1,1])^{-1}\right\}^{-1/2}\{\hat{\boldsymbol{\beta}}^* - \mathbf{b}([-1,1]) + o\{(nh^{2p+5})^{1/2}\}\} \to_{\mathrm{D}} N(\mathbf{0}, \mathbf{I}_{p+1}).$$

If $x = x_n$ is of the form $x = x_\partial + hc$ where $c \in [-1,1]$ is fixed and x_∂ is a fixed point on the boundary of $\text{supp}(f)$ then

$$\left\{\boldsymbol{\Sigma}_x(\mathcal{D}_{x,h})^{-1}\boldsymbol{\Gamma}_x(\mathcal{D}_{x,h})\boldsymbol{\Sigma}_x(\mathcal{D}_{x,h})^{-1}\right\}^{-1/2}\{\hat{\boldsymbol{\beta}}^* - \mathbf{b}(\mathcal{D}_{x,h}) + o\{(nh^{2p+5})^{1/2}\}\} \to_{\mathrm{D}} N(\mathbf{0}, \mathbf{I}_{p+1}).$$

The proof of the main theorem follows directly from Lemmas 1 and 2 below. The following lemma outlines the key idea of our proof.

Quadratic Approximation Lemma. Let $\{c_n(\boldsymbol{\theta})\colon\ \boldsymbol{\theta} \in \Theta\}$ be a sequence of random concave functions defined on a convex open subset Θ of $\mathbb{R}^k$. Let $\mathbf{F}$ and $\mathbf{G}$ be non-random matrices, with $\mathbf{F}$ positive definite and $\mathbf{U}_n$ a sequence of random vectors that is stochastically bounded. Lastly, let α_n be a sequence of constants tending to zero. Write $c_n(\boldsymbol{\theta}) = \mathbf{U}_n^T\boldsymbol{\theta} - \frac{1}{2}\boldsymbol{\theta}^T(\mathbf{F} + \alpha_n\mathbf{G})\boldsymbol{\theta} + f_n(\boldsymbol{\theta})$. If, for each $\boldsymbol{\theta} \in \Theta$, $f_n(\boldsymbol{\theta}) = o_P(1)$, then

$$\hat{\boldsymbol{\theta}}_n = \mathbf{F}^{-1}\mathbf{U}_n + o_P(1),$$

where $\hat{\boldsymbol{\theta}}_n$ (assumed to exist) maximizes c_n. If, in addition, $f'(\boldsymbol{\theta}) = o_P(\alpha_n)$ and $f_n''(\boldsymbol{\theta}) = o_P(\alpha_n)$ uniformly in $\boldsymbol{\theta}$ in a neighborhood of $\hat{\boldsymbol{\theta}}$, then

$$\hat{\boldsymbol{\theta}}_n = \mathbf{F}^{-1}\mathbf{U}_n - \alpha_n\mathbf{F}^{-1}\mathbf{G}\mathbf{F}^{-1}\mathbf{U}_n + o_p(\alpha_n).$$

Proof. The first statement follows from the Convexity Lemma (Pollard 1991). Denote $\boldsymbol{\theta}_0 = \mathbf{F}^{-1}\mathbf{U}_n$. Let $c_n'(\boldsymbol{\theta})$ and $c_n''(\boldsymbol{\theta})$ be respectively the gradient and the Hessian matrix of $c_n(\boldsymbol{\theta})$. Taylor's expansion leads to

$$\mathbf{0} = c_n'(\hat{\boldsymbol{\theta}}_n) = c_n'(\boldsymbol{\theta}_0) + c_n''(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\{1 + o_P(1)\}$$

so that

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 =& -\{c_n''(\boldsymbol{\theta}_0)\}^{-1}c_n'(\boldsymbol{\theta}_0)\{1 + o_P(1)\} \\
=&(\mathbf{F} + \alpha_n\mathbf{G})^{-1}\{\mathbf{U}_n - (\mathbf{F} + \alpha_n\mathbf{G})\boldsymbol{\theta}_0 + o_P(\alpha_n)\}\{1 + o_P(1)\} \\
=& -\alpha_n(\mathbf{F} + \alpha_n\mathbf{G})^{-1}\mathbf{G}\boldsymbol{\theta}_0 + o_P(\alpha_n) \\
=& -\alpha_n\mathbf{F}^{-1}\mathbf{G}\mathbf{F}^{-1}\mathbf{U}_n + o_P(\alpha_n)
\end{aligned}$$

The conclusion follows from the last display.

In the proofs of Lemmas 1 and 2 we will suppress the region of integration $\mathcal{A}$. If x is in the interior of supp(f) then $\mathcal{A}$ can be taken to be $[-1,1]$. If $x = x_n$ is a boundary point then $\mathcal{A} = \mathcal{D}_{x,h}$. For example, for x an interior point, $\int K_{r,p}(z)^2\,dz$ will be shorthand for $\int_{-1}^{1} K_{r,p}(z)^2\,dz$ while for x a boundary point it will be shorthand for $\int_{\mathcal{D}_{x,h}} K_{r,p}(z;\mathcal{D}_{x,h})^2\,dz$.

Lemma 1. Let $\bar{\eta}(x,u) = \eta(x) + \eta'(x)(u-x) + \cdots + \eta^{(p)}(x)(u-x)^p/p!$,

$$\mathbf{Y}_i^* = q_1(\bar{\eta}(x,X_i),Y_i)K\{(X_i-x)/h\}\begin{bmatrix} 1 \\ (X_i-x)/h \\ \vdots \\ (X_i-x)^p/(h^p p!) \end{bmatrix} \quad \text{and} \quad \mathbf{W}_n = (nh)^{-1/2}\sum_{i=1}^{n}\mathbf{Y}_i^*.$$

Then, under the above conditions,

$$\hat{\boldsymbol{\beta}}^* = \boldsymbol{\Sigma}_x^{-1}\mathbf{W}_n - h\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\Lambda}_x\boldsymbol{\Sigma}_x^{-1}\mathbf{W}_n + o_P(h).$$

Proof. Recall that $\hat{\boldsymbol{\beta}}$ maximizes (3.2). Let

$$\boldsymbol{\beta}^* = (nh)^{1/2}\begin{bmatrix} \beta_0 - \eta(x) \\ h\{\beta_1 - \eta'(x)\} \\ \vdots \\ h^p\{p!\beta_p - \eta^{(p)}(x)\} \end{bmatrix} \quad \text{and} \quad \mathbf{Z}_i = \begin{bmatrix} 1 \\ (X_i-x)/h \\ \vdots \\ (X_i-x)^p/(h^p p!) \end{bmatrix}.$$

Then $\beta_0 + \beta_1(X_i-x) + \cdots + \beta_p(X_i-x)^p = \bar{\eta}(x,X_i) + a_n\boldsymbol{\beta}^{*T}\mathbf{Z}_i$ where $a_n = (nh)^{-1/2}$. If $\hat{\boldsymbol{\beta}}$ maximizes (3.2), $\hat{\boldsymbol{\beta}}^*$ maximizes $\sum_{i=1}^{n} Q(g^{-1}(\bar{\eta}(x,X_i) + a_n\boldsymbol{\beta}^{*T}\mathbf{Z}_i),Y_i)K\{(X_i-x)/h\}$ as a function of $\boldsymbol{\beta}^*$. To study the asymptotic properties of $\hat{\boldsymbol{\beta}}^*$ we use the Quadratic Approximation Lemma applied to the maximization of the normalized function

$$\ell_n(\boldsymbol{\beta}^*) = \sum_{i=1}^{n}\Big\{Q(g^{-1}(\bar{\eta}(x,X_i) + a_n\boldsymbol{\beta}^{*T}\mathbf{Z}_i),Y_i) - Q(g^{-1}(\bar{\eta}(x,X_i)),Y_i)\Big\}K\{(X_i-x)/h\}.$$

Then $\hat{\boldsymbol{\beta}}^*$ maximizes ℓ_n. We remark that Condition (1) implies that ℓ_n is concave in $\boldsymbol{\beta}^*$. Using a Taylor series expansion of $Q(g^{-1}(\cdot),Y_i)$

$$\begin{aligned}\ell_n(\boldsymbol{\beta}^*) =& a_n\sum_{i=1}^{n} q_1(\bar{\eta}(x,X_i),Y_i)\boldsymbol{\beta}^{*T}\mathbf{Z}_i K\{(X_i-x)/h\} \\ &+ \frac{a_n^2}{2}\sum_{i=1}^{n} q_2(\bar{\eta}(x,X_i),Y_i)(\boldsymbol{\beta}^{*T}\mathbf{Z}_i)^2 K\{(X_i-x)/h\} \qquad (A.2) \\ &+ \frac{a_n^3}{6}\sum_{i=1}^{n} q_3(\eta_i,Y_i)(\boldsymbol{\beta}^{*T}\mathbf{Z}_i)^3 K\{(X_i-x)/h\},\end{aligned}$$

where η_i is between $\bar{\eta}(x, X_i)$ and $\bar{\eta}(x, X_i) + a_n \beta^{*T} \mathbf{Z}_i$.

Let $\mathbf{A}_n = a_n^2 \sum_{i=1}^n q_2(\bar{\eta}(x, X_i), Y_i) K\{(X_i - x)/h\} \mathbf{Z}_i \mathbf{Z}_i^T$. Then the second term in (A.2) equals $\frac{1}{2} \beta^{*T} \mathbf{A}_n \beta^*$. Now $(\mathbf{A}_n)_{ij} = (E\mathbf{A}_n)_{ij} + O_P[\{\mathrm{var}(\mathbf{A}_n)_{ij}\}^{1/2}]$ and $E\mathbf{A}_n = h^{-1} E\{q_2(\bar{\eta}(x, X_1)), \mu(X_1)) K\{(X_1 - x)/h\} \mathbf{Z}_1 \mathbf{Z}_1^T\}$. We will use a Taylor expansion of q_2 about $(\eta(X_1), \mu(X_1))$. Since $\mathrm{supp}(K) = [-1, 1]$ we only need consider $|X_1 - x| \leq h$, and thus

$$\bar{\eta}(x, X_1) - \eta(X_1) = -\frac{\eta^{(p+1)}(x)}{(p+1)!}(X_1 - x)^{p+1} - \frac{\eta^{(p+2)}(x)}{(p+2)!}(X_1 - x)^{p+2} + o(h^{p+2}). \quad (A.3)$$

Using the second result of (A.1), we obtain

$$(i-1)!(j-1)!(E\mathbf{A}_n)_{ij} = -(\rho f)(x)\ \nu_{i+j-2} - h(\rho f)'(x)\ \nu_{i+j-1} + o(h).$$

Similar arguments show that $\mathrm{var}\{(\mathbf{A}_n)_{ij}\} = O\{(nh)^{-1}\}$ and that the last term in (A.2) is $O_P\{(nh)^{-1/2}\}$. Therefore

$$\begin{aligned} \ell_n(\beta^*) &= \mathbf{W}_n^T \beta^* - \tfrac{1}{2} \beta^{*T} (\mathbf{\Sigma}_x + h\mathbf{\Lambda}_x) \beta^* + O_P\{(nh)^{-1/2}\} + o_P(h) \\ &= \mathbf{W}_n^T \beta^* - \tfrac{1}{2} \beta^{*T} (\mathbf{\Sigma}_x + h\mathbf{\Lambda}_x) \beta^* + o_P(h) \end{aligned}$$

since $nh^3 \to \infty$ and $h \to 0$. Similar arguments show that

$$\ell_n'(\beta^*) = \mathbf{W}_n - (\mathbf{\Sigma}_x + h\mathbf{\Lambda}_x)\beta^* + o_P(h)$$

and

$$\ell_n''(\beta^*) = -(\mathbf{\Sigma}_x + h\mathbf{\Lambda}_x) + o_P(h).$$

The result follows directly from the Quadratic Approximation Lemma.

Lemma 2. Suppose that the conditions of Theorem 1 hold. For $\mathbf{W}_n$ as defined in Lemma 1,

$$\left\{\mathbf{\Sigma}_x^{-1} - h\mathbf{\Sigma}_x^{-1}\mathbf{\Lambda}_x\mathbf{\Sigma}_x^{-1}\right\} E(\mathbf{W}_n) = \mathbf{b}_x + o\{(nh^{2p+5})^{1/2}\},$$

$$\mathbf{\Gamma}_x^{-1/2} \mathrm{cov}(\mathbf{W}_n) \to \mathbf{I}_{p+1} \quad \text{and} \quad \mathbf{\Gamma}_x^{-1/2}(\mathbf{W}_n - E\mathbf{W}_n) \to_{\mathrm{D}} N(\mathbf{0}, \mathbf{I}_{p+1}).$$

Proof. We compute the mean and covariance matrix of the random vector $\mathbf{W}_n$ by studying $\mathbf{Y}_1^*$, as defined in Lemma 1. The mean of the ith component of $\mathbf{Y}_1^*$ is easily shown to be

$$(E\mathbf{Y}_1^*)_i = \frac{h}{(i-1)!} \int q_1(\bar{\eta}(x, x + hz), \mu(x + hz)) z^{i-1} K(z) f(x + hz)\, dz.$$

Now by Taylor expansion, (A.1), and (A.3)

$$\begin{aligned} q_1(\bar{\eta}(x,x+hz),\mu(x+hz)) &= q_1(\eta(x+hz),\mu(x+hz)) \\ &\quad + \{\bar{\eta}(x,x+hz)-\eta(x+hz)\}q_2(\eta(x+hz),\mu(x+hz)\} \\ &\quad + \tfrac{1}{2}\{\bar{\eta}(x,x+hz)-\eta(x+hz)\}^2\{q_3(\eta(x),\mu(x))+o(1)\} \\ &= \left\{(hz)^{p+1}\frac{\eta^{(p+1)}(x)}{(p+1)!}+(hz)^{p+2}\frac{\eta^{(p+2)}(x)}{(p+2)!}\right\}\rho(x+hz)+o(h^{p+2}). \end{aligned}$$

Thus

$$(E\mathbf{Y}_1^*)_i = h^{p+2}\frac{\eta^{(p+1)}(x)}{(p+1)!}\frac{(\rho f)(x)}{(i-1)!}\nu_{p+i} + h^{p+3}\frac{\zeta_p(x)(\rho f)(x)}{(i-1)!}\nu_{p+i+1}+o(h^{p+3}) \qquad (A.4)$$

where

$$\zeta_p(x)=\frac{\eta^{(p+2)}(x)}{(p+2)!}+\frac{\eta^{(p+1)}(x)}{(p+1)!}\frac{(\rho f)'(x)}{(\rho f)(x)}.$$

The ith component of $\boldsymbol{\Sigma}_x^{-1}E\mathbf{W}_n$ is

$$\begin{aligned} (\boldsymbol{\Sigma}_x^{-1}E\mathbf{W}_n)_i &= (nh^{2p+3})^{1/2}\frac{\eta^{(p+1)}(x)}{(p+1)!}(i-1)!\sum_{j=1}^{p+1}(\mathbf{N}_p^{-1})_{ij}\nu_{p+j} \\ &\quad + (nh^{2p+5})^{1/2}\zeta_p(x)(i-1)!\sum_{j=1}^{p+1}(\mathbf{N}_p^{-1})_{ij}\nu_{p+j+1} \\ + o(nh^{2p+5})^{1/2} &= (nh^{2p+3})^{1/2}\frac{\eta^{(p+1)}(x)}{(p+1)!}\int z^{p+1}K_{i-1,p}(z)dz \\ &\quad + (nh^{2p+5})^{1/2}\zeta_p(x)\int z^{p+2}K_{i-1,p}(z)dz \;+\; o(nh^{2p+5})^{1/2} \end{aligned}$$

by Lemma 3 below. Next consider the second term in the expectation.

$$\begin{aligned} &h(\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\Lambda}_x\boldsymbol{\Sigma}_x^{-1}E\mathbf{W}_n)_i \\ &\quad = (nh^{2p+5})^{1/2}\frac{\eta^{(p+1)}(x)}{(p+1)!}\frac{(\rho f)'(x)}{(\rho f)(x)}(i-1)!\sum_{j=1}^{p+1}(\mathbf{N}_p^{-1}\mathbf{Q}_p\mathbf{N}_p^{-1})_{ij}\nu_{p+j}+O\{(nh^{2p+7})^{1/2}\}. \end{aligned}$$

Using the fact that $(\mathbf{Q}_p)_{kl}=(\mathbf{N}_p)_{k,l+1}$ for $l<p+1$, it can be shown that

$$(\mathbf{N}_p^{-1}\mathbf{Q}_p\mathbf{N}_p^{-1})_{ij}=(\mathbf{N}_p^{-1})_{i-1,j}+\left\{\sum_{k=1}^{p+1}(\mathbf{N}_p^{-1})_{i,k}\nu_{p+k}\right\}(\mathbf{N}_p^{-1})_{p+1,j}$$

for $i=2,\ldots,p+1$ and that

$$(\mathbf{N}_p^{-1}\mathbf{Q}_p\mathbf{N}_p^{-1})_{1j}=\left\{\sum_{k=1}^{p+1}(\mathbf{N}_p^{-1})_{1,k}\nu_{p+k}\right\}(\mathbf{N}_p^{-1})_{p+1,j}.$$

So by Lemma 3

$$(i-1)!\sum_{j=1}^{p+1}(\mathbf{N}_p^{-1}\mathbf{Q}_p\mathbf{N}_p^{-1})_{ij}\nu_{p+j} = (i-1)\int z^{p+1}K_{i-2,p}(z)\,dz + \frac{1}{p!}\int z^{p+1}K_{p,p}(z)\,dz\int z^{p+1}K_{i-1,p}(z)\,dz.$$

The statement concerning the asymptotic mean follows immediately.

By (A.4), the covariance between the ith and jth component of $\mathbf{Y}_1^*$ is $E((\mathbf{Y}_1^*)_i(\mathbf{Y}_1^*)_j)$ $+O(h^{2p+4})$. By a Taylor series expansion

$$E((\mathbf{Y}_1^*)_i(\mathbf{Y}_1^*)_j) = \frac{1}{(i-1)!(j-1)!}E\left[q_1^2(\eta(X_1),Y_1)K^2\{(X_1-x)/h\}\{(X_1-x)/h\}^{i+j-2}\right]+O(h^2)$$

and one easily calculates

$$\{\mathrm{cov}(\mathbf{Y}_1^*)\}_{i,j} = \frac{hf(x)\mathrm{var}(Y|X=x)}{[V(\mu(x))g'(\mu(x))]^2}\int\frac{z^{i+j-2}}{(i-1)!(j-1)!}K(z)^z\,dz + o(h).$$

Therefore $\mathbf{\Gamma}_x^{-1/2}\mathrm{cov}(\mathbf{W}_n) \to \mathbf{I}_{p+1}$.

We now use the Cramer-Wold device to derive the asymptotic normality of $\mathbf{W}_n$. For any unit vector $\mathbf{u} \in \mathbb{R}^{p+1}$, if

$$(na_n^2)^{-1/2}\mathbf{u}^T\mathrm{cov}(\mathbf{Y}_1^*)^{-1/2}(\mathbf{W}_n - E\mathbf{W}_n) \to_{\mathrm{D}} N(0,1), \tag{A.5}$$

then $h^{1/2}\mathrm{cov}(\mathbf{Y}_1^*)^{-1/2}(\mathbf{W}_n - E\mathbf{W}_n) \to_{\mathrm{D}} N(\mathbf{0},\mathbf{I}_{p+1})$ and so $\mathbf{\Gamma}_x^{-1/2}(\mathbf{W}_n - E\mathbf{W}_n) \to_{\mathrm{D}}$ $N(\mathbf{0},\mathbf{I}_{p+1})$. To prove (A.5), we need only check Lyapounov's condition for that sequence, which can easily be verified.

Lemma 3. For $\ell = 0,1,\ldots,$

$$\int_{\mathcal{A}} z^{p+\ell+1}K_{r,p}(z;\mathcal{A})dz = r!\sum_{i=1}^{p+1}\{\mathbf{N}_p(\mathcal{A})^{-1}\}_{r+1,i}\nu_{p+i+\ell}(\mathcal{A}).$$

Proof. Let c_{ij} denote the cofactor of $\{\mathbf{N}_p(\mathcal{A})\}_{ij}$. By expanding the determinant of $\mathbf{M}_{r,p}(z;\mathcal{A})$ along the $(r+1)$st column, we see that

$$\int z^{p+\ell+1}K_{r,p}(z)dz = \frac{r!}{|\mathbf{N}_p|}\int\sum_{i=1}^{p+1}z^{p+i+\ell}c_{i,r+1}K(z)dz = r!\sum_{i=1}^{p+1}\frac{c_{i,r+1}}{|\mathbf{N}_p|}\nu_{p+i+\ell}.$$

The lemma follows, since $(\mathbf{N}_p^{-1})_{ij} = c_{ij}/|\mathbf{N}_p|$ from the symmetry of $\mathbf{N}_p$ and a standard result concerning cofactors.

Lemma 4. Let $K_{r,p}(z;\mathcal{A})$ be as defined by (3.4) where K satisfies Condition (4). Then for $p \geq r$, $p-r$ even, $\int_{-1}^{1} z^{p+1} K_{r,p}(z)\,dz = 0$.

Proof. Suppose that both p and r are odd. The case when both p and r are even is handled similarly. Then by writing $\int_{-1}^{1} z^{p+1} K_{r,p}(z)\,dz$ in terms of the defining determinants, interchanging integral and determinant signs, and interchanging rows and columns of the determinant we can obtain a determinant of the form

$$\begin{vmatrix} \mathbf{M}_1 & \mathbf{0}_{(p+1)/2,(p-1)/2} \\ \mathbf{0}_{(p+1)/2,(p+3)/2} & \mathbf{M}_2 \end{vmatrix}$$

where $\mathbf{0}_{l,k}$ is an $l \times k$ matrix of zeroes. Since $\mathbf{M}_1$ is $\frac{1}{2}(p+1) \times \frac{1}{2}(p+3)$, there exists a non-zero vector $\mathbf{x}$ in $\mathbb{R}^{(p+3)/2}$ such that $\mathbf{M}_1\mathbf{x} = \mathbf{0}$. Thus the above determinant is zero.

Proof of Theorem 1. Theorem 1 follows from the Main Theorem by reading off the marginal distributions of the components of $\hat{\beta}^*$. To calculate the asymptotic variance, we calculate the $(r+1, r+1)$ entry of $(r!)^2 \mathbf{N}_p(\mathcal{A})^{-1}\, \mathbf{T}_p(\mathcal{A}) \mathbf{N}_p(\mathcal{A})^{-1}$ as

$$(r!)^2 \sum_{k=1}^{p+1} \sum_{\ell=1}^{p+1} \frac{c_{r+1,k} c_{r+1,\ell}}{|N_p(\mathcal{A})|^2} \{\mathbf{T}_p(\mathcal{A})\}_{k\ell} = \int_{\mathcal{A}} K_{r,p}(z;\mathcal{A})^2\,dz$$

where c_{ij} is the cofactor of $\{\mathbf{N}_p(\mathcal{A})\}_{ij}$.

Proof of Theorem 3. The proof of this theorem can be accomplished using exactly the same arguments as the univariate case with $p = 1$ and $r = 0$ and using multivariate approximations analogous to those used in Ruppert and Wand (1992).

REFERENCES

Cox, D.R. (1983), "Some Remarks on Over-dispersion," *Biometrika*, 70, 269-274.

Cox, D. D. and O'Sullivan, F. (1990), "Asymptotic Analysis of Penalized Likelihood and Related Estimators," *Annals of Statistics*, 18, 1676–1695.

Eubank, R. (1988), *Spline Smoothing and Nonparametric Regression*, New York: Dekker.

Fan, J. (1992a), "Local Linear Regression Smoothers and their Minimax Efficiency," *Annals of Statistics*, 20, in press.

Fan, J. (1992b), "Design-adaptive Nonparametric Regression," *Journal of the American Statistical Association*, 87, 998-1004.

Fan, J. and Gijbels, I. (1992), "Spatial and Design Adaptation: Variable Order Approximation in Function Estimation." *Institute of Statistics Mimeo Series # 2080*, University of North Carolina at Chapel Hill.

Fan, J. and Marron, J. S. (1992), "Best Possible Constant for Bandwidth Selection," *Annals of Statistics*, 20, Annals of Statistics, in press.

Gasser, T., Müller, H-G. and Mammitzsch, V. (1984), "Kernels for Nonparametric Curve Estimation," *Journal of the Royal Statistical Society, Series B*, 47, 238–252.

Godambe, V.P. and Heyde, C.C. (1987), "Quasi-likelihood and Optimal Estimation," *Inter. Statist. Rev.*, 55, 231-244.

Green, P. J. and Yandell, B. (1985), "Semiparametric Generalized Linear Models," in *Proceedings of the 2nd International GLIM Conference* (Lecture Notes Statistics 32), Berlin: Springer-Verlag.

Härdle, W. (1990), *Applied Nonparametric Regression*, New York: Cambridge University Press.

Härdle, W. and Scott, D. W. (1992), "Smoothing by Weighted Averaging of Rounded Points," *Computational Statistics*, 7, 97–128.

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

Marron, J. S. (1992), "Graphical Understanding of Higher Order Kernels," unpublished manuscript.

McCullagh, P. and Nelder, J. A. (1988), *Generalized Linear Models, Second Edition*, London: Chapman and Hall.

Nadaraya, E. A. (1964), "On Estimating Regression," *Theory of Probability and its Applications*, 10, 186–190.

Nelder, J. A. and Wedderburn, R. W. M. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society, Series A*, 135, 370–384.

O'Sullivan, F., Yandell, B., and Raynor, W. (1986), "Automatic Smoothing of Regression Functions in Generalized Linear Models," *Journal of the American Statistical Association*, 81, 96–103.

Park, B. U. and Marron, J. S. (1990), "Comparison of Data-driven Bandwidth Selectors," *Journal of the American Statistical Association*, 85, 66–72.

Pollard, D. (1991), "Asymptotics for Least Absolute Deviation Regression Estimators," *Econometric Theory*, 7, 186–199.

Ruppert, D. and Wand, M. P. (1992), "Multivariate Locally Weighted Least Squares Regression," unpublished manuscript.

Sevirini, T. A. and Staniswalis, J. G. (1992), "Quasi-likelihood Estimation in Semiparametric Models," unpublished manuscript.

Sheather, S. J. and Jones, M. C. (1991), "A Reliable Data-based Bandwidth Selection Method for Kernel Density Estimation," *Journal of the Royal Statistical Society, Series B*, 53, 683–690.

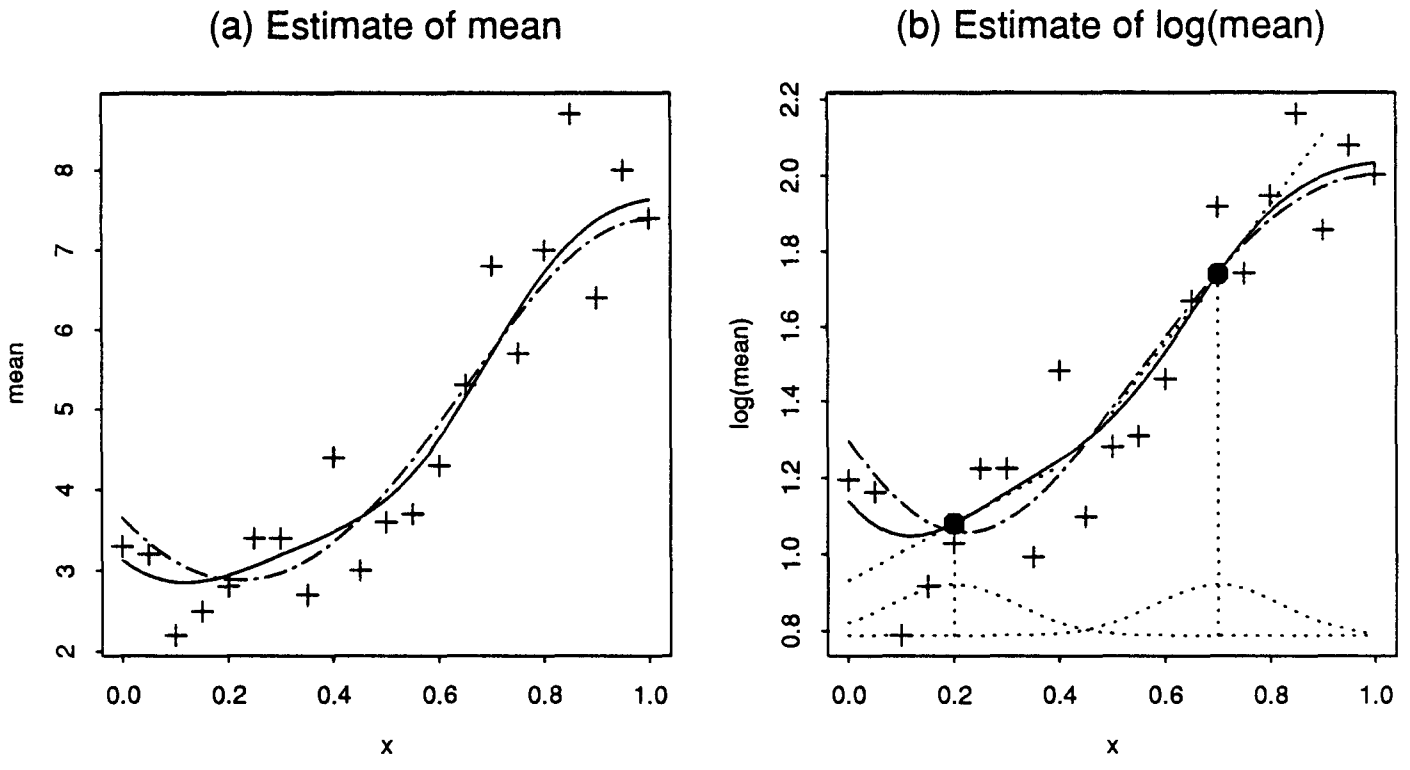
Staniswalis, J. G. (1989), "The Kernel Estimate of a Regression Function in Likelihood-based Models," *Journal of the American Statistical Association*, 84, 276–283.

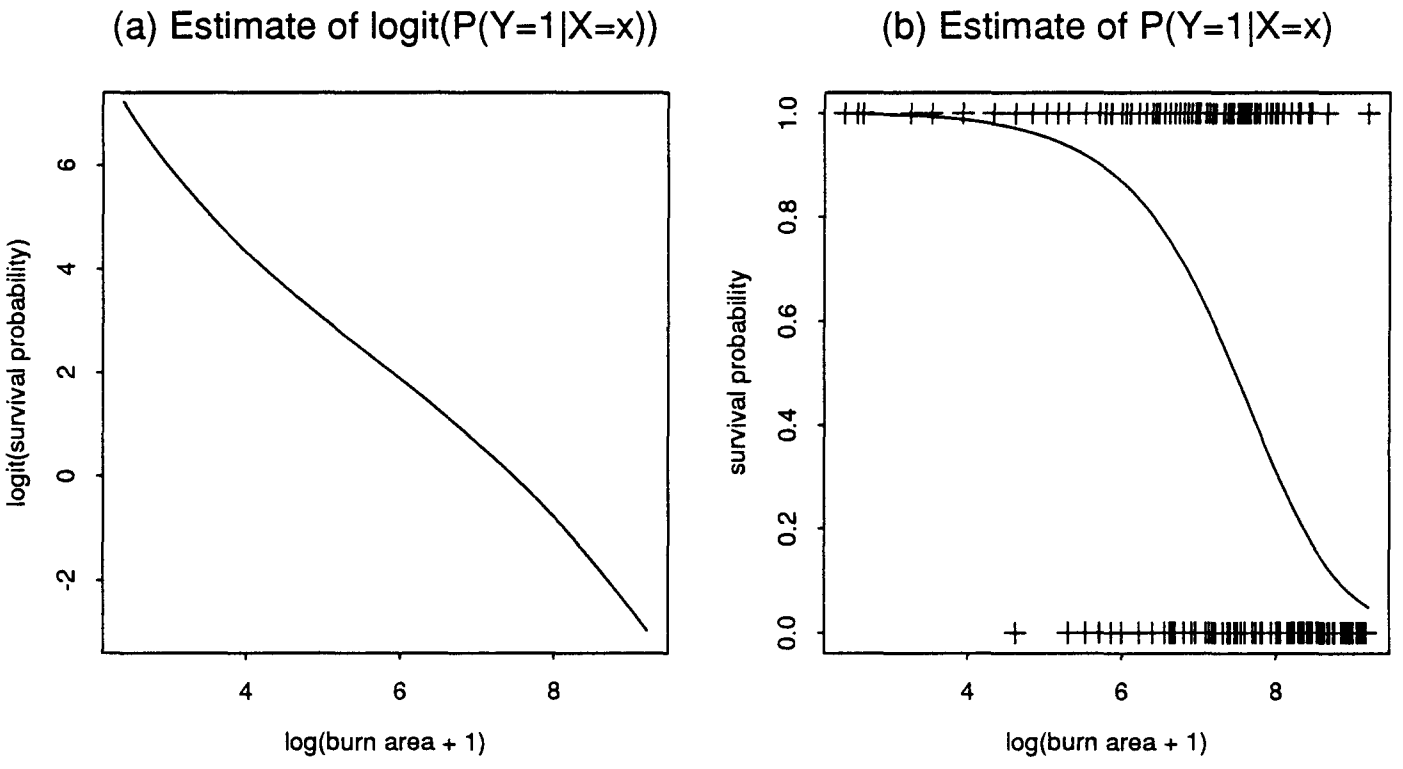
Tibshirani, R. and Hastie, T. (1987), "Local Likelihood Estimation," *Journal of the American Statistical Association*, 82, 559–568.

Wahba, G. (1990), *Spline Models for Observational Data*, Philadelphia: SIAM.

Watson, G. S. (1964), "Smooth Regression Analysis," *Sankhyā, Series A*, 26, 101–116.

Wedderburn, R. W. M. (1974), "Quasi-likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method," *Biometrika*, **61**, 439–447.


(a) Estimate of mean
mean
x
0.0
0.2
0.4
0.6
0.8
1.0
2
3
4
5
6
7
8
(b) Estimate of log(mean)
log(mean)
x
0.0
0.2
0.4
0.6
0.8
1.0
0.8
1.0
1.2
1.4
1.6
1.8
2.0
2.2


(a) Estimate of logit(P(Y=1|X=x))
logit(survival probability)
-2
0
2
4
6
4
6
8
log(burn area + 1)
(b) Estimate of P(Y=1|X=x)
survival probability
0.0
0.2
0.4
0.6
0.8
1.0
4
6
8
log(burn area + 1)