

Local Polynomial Regression for Binary Response

Aaron Aragaki and Naomi Altman *
 Biometrics Unit
 Cornell University

March 3, 1997

Abstract

Nonparametric regression methods can provide nonparametric modeling, guidance in selection of parametric models and diagnostic tools. This is particularly important for binary regression due to the lack of simple graphical tools for data exploration.

In this article, we discuss the application of local polynomial regression to the binary regression problem. We show that local polynomial regression is consistent, and is a simpler alternative to generalized smooth models. Bandwidth selection for good small sample performance remains problematic. We show by simulation that methods such as cross-validation and “plug-in” estimators, which perform well for continuous response, do poorly for binary data. However, bootstrap bandwidth selection, although very computer-intensive, appears to work well for binary response.

Keywords: smoothing; nonparametric regression; logistic regression; bandwidth selection; smoothed cross-validation; plug-in estimator; bootstrap bandwidth selection.

1 Introduction

Binary response is common in many problems. Often, the response is affected by a covariate, x (which may be multidimensional). If the data are independent, the response $Y(x)$ can be modeled by

$$Y(x) \sim \text{Bernoulli}(\mu(x))$$

where $\mu(x) = \text{Prob}(Y = 1|X = x)$ is the success probability. Since it is also true that

$$\mu(x) = E(Y|X = x)$$

regression methods can be used to estimate the unknown success probability from the data.

In extending methods developed for continuous data to binary regression, two problems arise: since $\mu(x)$ is the success probability, it is bounded between 0 and 1 and the response is heteroscedastic, with variance depending on $\mu(x)$.

A very useful parametric model for binary regression is the generalized linear model (GLM) (Wedderburn, 1974; McCullagh and Nelder, 1989). In GLMs, μ is transformed by a smooth

*Supported by Hatch Grant 151410 NYF and NSF Grant DMS-95-25350.

function g called the ‘link function’ which extends its range to the entire real line. A linear function is used to model the relationship between the transformed regression function and the covariates:

$$\begin{aligned} g(\mu(x)) &= x'\beta \\ \text{or} \quad \mu(x) &= g^{-1}(x'\beta) \end{aligned} \tag{1}$$

The distribution of the response is modeled and estimation is done by maximum likelihood. Alternatively, the dependence of the variance on the mean is modeled, and estimation is done by quasi-likelihood (defined in Section 2.1). Because g determines the curvature of the resulting regression function, appropriate choice of the link is critical to good fit of GLMs.

The main drawback of parametric modeling is lack of flexibility. Often the data analyst has no a priori idea of the functional form of $\mu(x)$, making model specification difficult. A number of nonparametric regression methods are available for continuous response with independent, homoscedastic errors and smooth regression function. Local polynomial regression (Cleveland, 1975; Cleveland and Devlin, 1988; Fan, 1992; Hastie and Loader, 1993) has become a popular method because it can adapt automatically to randomly spaced x and to x near a boundary, can readily be extended to multivariate x and has intuitive appeal as a local Taylor series expansion of the regression function.

If the regression function can be approximated by a $(p + 1)$ -term Taylor series

$$\mu(x) = \sum_{j=0}^p \frac{\mu^{(j)}(x_0)}{j!} (x - x_0)^j + o(|x - x_0|^p)$$

then a local polynomial regression estimator of degree p can be defined by

$$\hat{\mu}(x_0) = \hat{\beta}_0(x_0)$$

where $\hat{\mu}$ is the estimator of the regression function, and $\hat{\beta}_0(x_0)$ is the constant term in a polynomial regression of y on the set of x 's in neighborhood of x_0 . (In the remainder of the paper we will use the notation $\hat{\beta}_i$, suppressing the dependence of the regression coefficients on the neighborhood.) For example, the local least squares polynomial kernel regression (LPKR) estimator of degree p is defined via the weighted least squares estimator. $(\hat{\beta}_0, \dots, \hat{\beta}_p)$ minimize:

$$\frac{1}{h} \sum_{i=1}^n (Y_i - \beta_0 - \dots - \beta_p (X_i - x_0)^p)^2 K\left(\frac{X_i - x_0}{h}\right) \tag{2}$$

where $K(x)$ is a symmetric probability density, and h is a smoothing parameter, called the bandwidth, that controls the size of the neighborhood.

Choice of the degree of the polynomial depends to a large extent on the smoothness the investigator is willing to ascribe to μ . Fan et al (1996) have shown that the optimal kernel for polynomials of any degree is the Epanechnikov kernel, $K(x) = 3/4(1 - x^2)_+$. However, good performance of the estimator depends on appropriate choice of bandwidth. A number of techniques have been developed for adaptive choice of h for continuous response. These include cross-validation and its relatives (Allen, 1974; Craven and Wahba, 1979; Geisser, 1975), plug-in

estimators (Hall et al, 1991; Park and Marron, 1990; Ruppert, Sheather and Wand, 1995) and various types of bootstrap estimators (Faraway, 1990; Faraway and Jhun, 1990).

Fan, Heckman and Wand (1995) extended the generalized linear model to a generalized smooth model

$$g(\mu(x)) = s(x) \quad (3)$$

where $s(x)$ is a smooth function estimated by a local polynomial regression estimator based on local quasi-likelihood as discussed in Section 2.3. They show that their estimator is asymptotically normal and give closed form expressions for its asymptotic variance and bias. One drawback to their method is that it is very computationally intensive, requiring a numerical maximization at each design point. The Fan, Heckman and Wand (FHW) method is discussed in more detail in Section 2.3.

In this article we show that LPKR, a much less computationally intensive method than the FHW method, is suitable for nonparametric binary regression. This estimator is asymptotically Normal with the same asymptotic variance as the FHW estimator and asymptotic bias of the same order. Moreover, algorithms for LPKR are $O(n)$ and are available in many statistical packages. Asymptotic results for LPKR with binary response are discussed in Section 3.

In Section 4 the issue of bandwidth selection is addressed via simulation. Several bandwidth selection procedures which were developed for continuous data are compared for binary response data. Plug-in procedures, which appear to have very good performance for continuous response, perform very poorly. On the other hand, bootstrap procedures, while computationally intensive, appear to perform adequately.

In Section 6 LPKR is applied to the periparturient recumbency data (Clark et al, 1987) which were previously analyzed by kernel regression in Altman (1992) and Altman and McGibbon (1996).

2 Regression Estimators

In this section we discuss 3 regression estimators which are relevant to the application of LPKR to binary regression. These are GLMs, LPKR for continuous data and the FHW estimator for generalized smooth models. Notation and theoretical results needed for the comparison of LPKR and FHW for binary regression are introduced.

2.1 Generalized Linear Models

GLMs provide a flexible framework for parametric estimation of the regression function when the data are not normally distributed. Conditional on x , the data are assumed to come from an exponential family parametrized by its mean $\mu(x)$, which is transformed to linearity by (1). Although maximum likelihood may be used to compute the regression parameters, often quasi-likelihood methods are used. In quasi-likelihood, the conditional variance $Var(Y|X = x)$ is replaced by a known function of the regression $V(\mu(x))$. Then the quasi-likelihood function is

$$Q(\mu(x), y) = \int_y^{\mu(x)} \frac{y - t}{V(t)} dt \quad (4)$$

and

$$\frac{d}{d\mu(x)}Q(\mu(x), Y) = \frac{Y - \mu(x)}{V(\mu(x))}.$$

The quasi-likelihood estimator is the parameter set minimizing

$$\sum_{i=1}^n Q(\mu(x_i), Y_i).$$

Wedderburn (1974) discusses in detail the properties of quasi-likelihood functions, computational issues and the equivalence of quasi-likelihood and log-likelihood for exponential families when $Var(Y|X = x) = V(\mu(x))$.

For binary data, $Var(Y|X = x) = \mu(x)(1 - \mu(x))$. The link function g is chosen to be bounded between 0 and 1, which automatically forces $\mu(x)$ to be a valid probability. Common choices of g are the logistic and Gompertz functions.

2.2 LPKR for Continuous Response

Least squares local polynomial regression (2) was developed for the case of continuous response (Cleveland, 1979; Cleveland and Devlin, 1988). Interest in LPKR has grown since Fan (1992) demonstrated that the asymptotic bias and variance are smaller than those of local averaging methods, such as kernel regression, for designs which are not equally spaced and near boundaries. Because locally the solution to the estimating equation is a weighted least squares estimator, for fixed bandwidth the estimator is linear. Efficient $O(n)$ algorithms have been developed (Fan and Marron, 1994; Seifert et al, 1994).

Fan (1992) computed the asymptotic bias and variance of the least squares local linear regression estimator. These are:

$$Bias(\hat{\mu}(x)) = \frac{h^2}{2}\mu''(x) \int t^2 K(t) dt \quad (5)$$

$$\text{and } Variance(\hat{\mu}(x)) = \frac{Var(Y|X = x)}{f_x(x)nh} \int K^2(t) dt \quad (6)$$

where $f_x(x)$ is the density of the covariate.

2.3 Local Polynomial Regression for Binary Response - The FHW Method

Fan, Heckman and Wand (1995) have extended local polynomial kernel regression to generalized smooth models in the spirit of GLMs.

They assume that the mean function is defined by (3) and estimate the smooth function $s(x_0)$ by weighted quasi-likelihood – that is, the local regression coefficients are estimated by maximizing

$$\frac{1}{h} \sum_{i=1}^n Q\left(g^{-1}(\beta_0 + \dots + \beta_p(X_i - x_0)^p), Y_i\right) K\left(\frac{X_i - x}{h}\right).$$

Then $\hat{\mu}(x_0) = g^{-1}(\hat{\beta}_0)$.

Fan, Heckman and Wand (1995) derive expressions for the asymptotic distribution of $\hat{s}(x)$ for local polynomials of degree p . For local linear regression they found that under regularity conditions on μ , s , g and $Var(Y|X = x)$, and the condition

$$\frac{d^2}{dx^2}Q(g^{-1}(x), y) < 0 \quad (7)$$

that $\hat{s}(x)$ is asymptotically Normal and asymptotically

$$\begin{aligned} Bias(\hat{s}(x)) &= \frac{s''(x)}{s} h^2 \int t^2 K(t) dt \\ \text{and } Variance(\hat{s}(x)) &= \frac{1}{f_x(x)nh} Var(Y|X = x) g'(\mu(x))^2 \int_D K^2(t, D) dt \end{aligned}$$

where $D = \{z : z - hz \in \text{supp}(K)\}$ and $K(t, D)$ is a weighted version of K where the weights depend on D and are 1 away from the boundary.

Based on these results, the asymptotic distribution of $\hat{\mu}(x)$ can be derived from the Taylor series expansion of $\mu(x) = g^{-1}(s(x))$ about $s(x)$. We find that $\hat{\mu}(x)$ is also asymptotically Normal with asymptotic variance (6) and asymptotic bias:

$$Bias(\hat{\mu}(x)) = \left(\frac{g''(\mu(x))\mu'(x)^2}{g'(\mu(x))} + \mu''(x) \right) \frac{h^2}{2} \int t^2 K(t) dt$$

Notice that the asymptotic variance does not depend on the link function, but the asymptotic bias does. Consider two link functions g_1 and g_2 . If $g_1(\mu(x)) = s_1(x)$ has sharper local minima and maxima than $g_2(\mu(x)) = s_2(x)$, then $g_1^{-1}(\hat{s}_1(x))$ will have more bias at local optima than $g_2^{-1}(\hat{s}_2(x))$, although the estimators have the same variance. This means that the performance of the FHW estimator can depend strongly on the choice of link g .

Use of the logit link function with local linear regression and binary data leads to what we shall refer to as $\hat{\mu}_{FHW}(x)$ with asymptotic variance (6) and bias:

$$Bias(\hat{\mu}_{FHW}(x)) = \left(\frac{2\mu(x) - 1}{\mu(x)(1 - \mu(x))} \mu'(x)^2 + \mu''(x) \right) \frac{h^2}{2} \int t^2 K(t) dt \quad (8)$$

3 LKPR for Binary Response

Nonparametric regression estimators are inherently computationally intensive. They are widely used as graphical tools to gain insight into the true nature of $\mu(x)$. As the local parameters of $\hat{\mu}_{FHW}(x)$ are obtained by numerically maximizing a quasi-likelihood function at each grid point, the method is computationally intensive relative to linear nonparametric regression estimators such as LKPR. Below we show that the LKPR estimator has similar asymptotic properties to $\hat{\mu}_{FHW}(x)$.

3.1 The Estimator

The local estimator of $s(x) = g(\mu(x))$ is robust to misspecification of the conditional variance. Fan, Heckman, and Wand (1995) point out that because of the localness of the fitting, Theorem 1 holds regardless of whether or not $V(\mu(x)) = Var(Y|X = x)$. The expressions for the asymptotic variance and asymptotic bias of their estimator do not involve $V(\mu(x))$. This coincides with earlier results given by Jones (1993) regarding heteroscedasticity and kernel nonparametric regression. He states, "... one can afford to ignore heteroscedastic errors when nonparametrically estimating the regression mean using a single overall smoothing parameter."

Since $V(\mu(x))$ determines the quasi-likelihood function through Equation (4) an appropriate misspecification of the conditional variance may lead to an explicit solution of the quasi-likelihood function. If we let $V(\mu(x)) = \sigma^2$, some constant, then the quasi-likelihood function is

$$Q(\mu(x), Y) = \frac{-(Y - \mu(x))^2}{2\sigma^2}$$

the least-squares function. Furthermore, if we choose the identity link function, $g(x) = x$, then the resulting estimator, $\hat{\mu}_{ilcv}(x)$ (identity link, constant variance), is the local least squares linear kernel estimator (2).

3.2 Asymptotic Properties

The computationally simpler estimator, $\hat{\mu}_{ilcv}(x)$, has the same asymptotic properties as $\hat{\mu}_{FW}(x)$. It can be shown that $\hat{\mu}_{ilcv}(x)$ is asymptotically normal with asymptotic variance and asymptotic bias given by (5) and (6) respectively. $\hat{\mu}_{ilcv}(x)$ has the same asymptotic variance as $\hat{\mu}_{FW}(x)$ and asymptotic bias of the same order.

3.3 The Identity and Other Link Functions

Since the identity link function is used when calculating $\hat{\mu}_{ilcv}(x)$, it is possible for $\hat{\mu}_{ilcv}(x)$ to lie outside the interval $[0, 1]$ although this seldom happens for reasonable bandwidths. If this is of concern, truncation of $\hat{\mu}_{ilcv}(x)$ is a simple way to guarantee that the estimator is truly a probability function. Under the assumptions giving the Fan, Heckman and Wand results, it can be shown that the truncated estimator is asymptotically Normal with the same asymptotic variance and asymptotic bias as $\hat{\mu}_{ilcv}(x)$. The truncated estimator will always lie closer to $\mu(x)$ than $\hat{\mu}_{ilcv}(x)$ and so will have a smaller risk than $\hat{\mu}_{ilcv}(x)$ with respect to squared error loss. The only drawback to truncation is that the truncated estimator will not have the same smoothness properties as $\mu(x)$.

The data analyst should be wary about the use of other link functions in conjunction with local least squares. The link functions that are commonly used in the parametric setting to guarantee that the estimator lie in the interval $[0, 1]$, may not work for local *least squares* linear kernel estimators. Application of these functions does guarantee that the resulting estimator is always a probability (i.e. $0 \leq \hat{\mu}(x) \leq 1$), but condition (7) need not hold, and so the local least squares estimator need not be consistent. It is easy to check that (7) does not hold for the logit, probit, log-log, and complementary log-log functions for $x \in R^1$ and $Y \in [0, 1]$.

3.4 Asymptotic Bias of $\hat{\mu}_{ilcv}(x)$ and $\hat{\mu}_{FHW}(x)$

The asymptotic biases of $\hat{\mu}_{ilcv}(x)$ and $\hat{\mu}_{FHW}(x)$ are given by (5) and (8) respectively. We see that the asymptotic bias of $\hat{\mu}_{ilcv}(x)$ is a much simpler expression, involving only the second derivative of $\mu(x)$. $\hat{\mu}_{ilcv}(x)$ will underestimate the peaks of $\mu(x)$, where $\mu''(x) < 0$, and fill in the valleys of $\mu(x)$, where $\mu''(x) > 0$. This makes intuitive sense because $\mu''(x)$ can be thought of as a measure of the curvature of $\mu(x)$; $\mu(x)$ will be harder to estimate in regions where $|\mu''(x)|$ is large. Also, the accuracy of this estimator does not depend on the value of $\mu(x)$, which agrees with intuition.

On the other hand, the asymptotic bias of $\hat{\mu}_{FHW}(x)$ is much more complicated. It involves $\mu'(x) > 0$ and more troubling, it involves $\mu(x)$. This conflicts with simple intuition. For instance, if $0 \leq \mu(x) \leq 1$, then FHW estimators of the parallel functions $\mu_1(x) = .5\mu(x)$ and $\mu_2(x) = \mu_1(x) + .5$ have different asymptotic biases.

The difference in asymptotic biases is due to the different “smoothing spaces”. $\hat{\mu}_{ilcv}(x)$ is obtained by smoothing the untransformed data, while $\hat{\mu}_{FHW}(x)$ is obtained by smoothing the data in the “logit” space, and then transforming back. The asymptotic bias of $\hat{s}(x)$ involves only the term $s''(x)$, but transformation back to the space of interest complicates things because

$$\begin{aligned} s''(x) &= \frac{d^2}{dx^2} \text{logit}(\mu(x)) \\ &= \left(\frac{-1 + 2\mu(x)}{\mu(x)^2(1 - \mu(x))^2} \right) \mu'(x)^2 + \frac{1}{\mu(x)(1 - \mu(x))} \mu''(x). \end{aligned}$$

The remainder of this paper will deal strictly with $\hat{\mu}_{ilcv}(x)$.

4 Bandwidth Selection

Appropriate selection of bandwidth is critical to the performance of local polynomial regression estimators. Too small a bandwidth unduly increases the variance of the estimator, and leads to an undersmoothed estimate with a number of spurious bumps. Too large a bandwidth introduces excessive bias, and leads to an oversmoothed estimate where important features of the true curve may be smoothed out.

Data-adaptive procedures for bandwidth selection are commonly based on minimizing mean integrated squared error, MISE:

$$E \int (\hat{\mu}(x; h) - \mu(x))^2 dx$$

The minimizing bandwidth, h_{MISE} is considered optimal. Alternatively, the target bandwidth is h_{ASE} , the bandwidth minimizing

$$ASE = \frac{1}{n} \sum (\hat{\mu}(x_i; h) - \mu(x_i))^2$$

However, because neither MISE nor ASE can be computed, a number of adaptive bandwidth selection procedures have been proposed.

4.1 Selection Procedures

In this section we describe several different bandwidth selection procedures for the estimator $\hat{\mu}_{ilcv}(x)$. Altman and MacGibbon (1996) discuss bandwidth selection for binary data with kernel regression estimators, but no articles have been published regarding LPKR and bandwidth selection for binary response data. Most of the procedures described here are adaptations of bandwidth selection procedures often used in kernel density estimation, kernel regression or LPKR with continuous response.

4.1.1 Plug-in Estimators

Plug-in estimators are based on the asymptotic mean integrated squared error (AMISE) which from (5) and (6) is readily computed:

$$AMISE = \frac{1}{nh} \int K^2(z) dz \int \mu(x) (1 - \mu(x)) f(x)^{-1} dx + \frac{h^4}{4} \left[\int z^2 K(z) dz \right]^2 \int \mu''(x)^2 dx.$$

Solving explicitly for the h that minimizes the above expression, yields

$$h_{AMISE} = \left(\frac{\int K^2(z) dz \int \mu(x) (1 - \mu(x)) f(x)^{-1} dx}{n \left[\int z^2 K(z) dz \right]^2 \int \mu''(x)^2 dx} \right)^{\frac{1}{5}}. \quad (9)$$

The *plug-in* estimator for h_{AMISE} is obtained by plugging nonparametric estimates of the unknown functionals into the right hand side of equation (9). For example, we might compute $\hat{\mu}_{ilcv}''(x) = 2\hat{\beta}_2$ (recalling the dependence of $\hat{\beta}_2$ on x) to estimate the corresponding functional $\int \mu''(x)^2 dx$. A pilot bandwidth is required to estimate the functionals.

For stability purposes, some statisticians (e.g. Fan, Heckman, Wand 1995) prefer to slightly perturb this by including the design density, $f(\cdot)$, and a weight function, $w(\cdot)$. $f(\cdot)$ is meant to give more weight to the estimator of $\mu''(x)$ where the data is expected to be more plentiful. $w(\cdot)$ is usually specified to be 0 near the boundaries of the covariate space and 1 elsewhere. This compensates for the high small sample variability of LPKR derivative estimators near the boundaries (Ruppert, Sheather, Wand 1994).

Plug-in estimators are very simple to understand and have received recent attention in the literature. In the contexts of density estimation and the usual regression setting, plug-in estimators have outstanding theoretical performance (some converge to the “optimal” at rate $O_p(n^{-\frac{1}{2}})$), and have performed well in simulations (Härdle, Hall, and Marron, 1988; Park and Marron, 1990; Sheather and Jones, 1991; Hall et al, 1991; Ruppert, Sheather, and Wand, 1994). A number of plug-in estimators have been proposed.

Plug-in estimator 1:

The first plug-in estimator is

$$\hat{h}_{plug1}(X) = \left(\frac{\bar{Y} (1 - \bar{Y}) (b - a) \int K^2(z) dz}{n \left(\int z^2 K(z) dz \right)^2 \left(\frac{1}{n} \sum_{i=1}^n \hat{\mu}_{ilcv}''(X_i; k)^2 1_{[(1-\alpha)a + \alpha b < X_i < \alpha a + (1-\alpha)b]} \right)} \right)^{\frac{1}{5}}.$$

where \bar{Y} is the sample mean, the notation $\hat{\mu}(X_i, k)$ refers to the LPKR estimator with bandwidth k , $[a, b]$ is the range of x , $1_{[c, d]}$ is the indicator function for the interval $[c, d]$ and α is a trimming percentage. The following should be noted

1. $\frac{1}{n} \sum_{i=1}^n \hat{\mu}_{ilcv}''(X_i; g)^2 1_{[(1-\alpha)a+\alpha b < X_i < \alpha a+(1-\alpha)b]}$ is a natural estimator of $\int \mu''(x)^2 f(x)w(x)dx$. The weight function trims the observations that lie within $\alpha\%$ of the boundaries. This type of estimator was suggested by Ruppert, Sheather and Wand (1994).
2. The bandwidth, k , for estimating $\mu''(x)$ needs to be specified. The method of Ruppert and Wand (1994), specified for continuous response data, is adapted here. This gives

$$\hat{k} = C_2(K) \left(\frac{\bar{Y}(1-\bar{Y})(b-a)}{n \left| \int \hat{\mu}_{para}''(x) \hat{\mu}_{para}^{(4)}(x) dx \right|} \right)^{\frac{1}{7}},$$

where $\hat{\mu}_{para}''(x), \hat{\mu}_{para}^{(4)}(x)$ are the parametric estimates obtained by fitting a cubic logit to the data, and

$$C_2(K) = \begin{cases} 315^{\frac{1}{7}} & \text{for } \int \hat{\mu}_{para}''(x) \hat{\mu}_{para}^{(4)}(x) dx < 0 \\ (1575/2)^{\frac{1}{7}} & \text{for } \int \hat{\mu}_{para}''(x) \hat{\mu}_{para}^{(4)}(x) dx > 0. \end{cases}$$

3. The variance functional $\int \mu(x)(1-\mu(x))dx$ has been replaced by $\bar{Y}(1-\bar{Y})(b-a)$. This approximation will be very good if a significant proportion of the true curve lies in the interval $[0.2, 0.8]$. This can be seen by the simple inequality

$$(b-a)0.8 \cdot 0.2 \leq \int_a^b \mu(x)(1-\mu(x))dx \leq (b-a)0.5^2$$

and the fact that the graph of $(v(1-v))^{\frac{1}{5}}$ is very flat for $0.2 \leq v \leq 0.8$

Plug-in estimator 2:

The second plug-in estimator is given by

$$\hat{h}_{plug2}(X) = \left(\frac{\bar{Y}(1-\bar{Y})(b-a) \int K^2(z) dz}{n \left(\int z^2 K(z) dz \right)^2 \left(\frac{1}{n} \sum_{i=1}^n \hat{\mu}_{ilcv}''(X_i; n^{-\frac{1}{7}}) \right)^2 1_{[(1-\alpha)a+\alpha b < X_i < \alpha a+(1-\alpha)b]}} \right)^{\frac{1}{5}}.$$

This is the same as \hat{h}_{plug1} except for the choice of pilot bandwidth, here $k = n^{-\frac{1}{7}}$. If the quantity $\frac{1}{n} \sum_{i=1}^n \hat{\mu}_{ilcv}''(X_i; \cdot)^2 1_{[(1-\alpha)a+\alpha b < X_i < \alpha a+(1-\alpha)b]}$, is robust to the choice of bandwidth, then $\hat{h}_{plug2}(X)$ and $\hat{h}_{plug1}(X)$ should be comparable.

Plug-in estimator 3:

The third plug-in estimator is given by

$$\hat{h}_{plug3}(X) = \left(\frac{\int K^2(z) dz \int \hat{\mu}_{ilcv}(x; \lambda) (1 - \hat{\mu}_{ilcv}(x; \lambda)) dx}{n \left(\int z^2 K(z) dz \right)^2 \left(\frac{1}{n} \sum_{i=1}^n \hat{\mu}_{ilcv}''(X_i; g) \right)^2 1_{[(1-\alpha)a+\alpha b < X_i < \alpha a+(1-\alpha)b]}} \right)^{\frac{1}{5}}.$$

$\hat{h}_{plug3}(X)$ has the following characteristics:

1. $\hat{h}_{plug3}(X)$ is just a high powered version of $\hat{h}_{plug1}(X)$ with the integrated variance estimated by $\int \hat{\mu}_{ilcv}(x; \lambda) (1 - \hat{\mu}_{ilcv}(x; \lambda)) dx$.
2. Since $\hat{h}_{plug3}(X)$ uses a consistent estimator of $\int \mu(x) (1 - \mu(x)) dx$, $\hat{h}_{plug3}(X)$ is consistent. $\hat{h}_{plug1}(X)$ and $\hat{h}_{plug2}(X)$ are not consistent.

4.1.2 Rule of Thumb Estimators

Rule of thumb estimators, ROT, are the parametric analog to plug-in estimators. Instead of estimating the unknown functionals by nonparametric procedures, the nonparametric functionals are estimated parametrically. For example, the functional $\int \mu''(x)^2 dx$, can be estimated by first fitting a cubic logistic regression curve to the data, and then substituting back the parameter estimates, $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$, into

$$\int \left(\frac{d^2}{dx^2} \frac{1}{1 + e^{\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3}} \right)^2 dx.$$

This expression can be computed explicitly and is a parametric estimator of $\int \mu''(x)^2 dx$. Since a cubic logistic regression curve is somewhat flexible, this is a reasonable estimator of $\int \mu''(x)^2 dx$; if necessary, higher order logistic curves can be fitted. It is important to notice that the dependence of h_{AMISE} on $\int \mu''(x)^2 dx$ is dampened by taking the $1/5^{\text{th}}$ power (e.g. $70^{\frac{1}{5}} - 20^{\frac{1}{5}} \cong 0.5$).

Rule of thumb estimator:

The rule of thumb estimator is

$$\hat{h}_{ROT}(X) = \left(\frac{\bar{Y} (1 - \bar{Y}) (b - a) \int K^2(z) dz}{n [\int z^2 K(z) dz]^2 \frac{1}{(b-a)}} \int \left(\frac{d^2}{dx^2} \frac{1}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 + \hat{\beta}_3 x^3}} \right)^2 dx \right)^{\frac{1}{5}},$$

Note that the weight function $w(\cdot)$ has been discarded since a parametric estimate of $\int \mu''(x)^2 dx$ is being used.

4.1.3 Smoothed Cross-Validation

Smoothed cross-validation (SCV) is a very computer intensive method of estimating h_{MISE} . SCV was first introduced by Hall, Marron, and Park (1992) in the context of density estimation as an alternative to least squares cross-validation (CV), which suffers from high sample variability. As with the plug-in and ROT estimators, its derivation begins with an approximation to the MISE of $\hat{\mu}_{ilcv}(x)$. Notice that

$$MISE \cong \frac{1}{nh} \int K^2(z) dz \int \mu(x) (1 - \mu(x)) f(x)^{-1} dx + \int (E\hat{\mu}_{ilcv}(x; h) - \mu(x))^2 dx.$$

The second term is the integrated squared bias.

Hall, Marron, and Park (1992) derive an approximation to the bias component by replacing $\mu(x)$ by a nonparametric estimator, $\hat{\mu}_{ilcv}(x; k)$. The resulting approximation is

$$\int (E^* \hat{\mu}_{ilcv}(x; h) - \hat{\mu}_{ilcv}(x; k))^2 dx,$$

where E^* indicates that expectation is taken over the binary responses generated by $\hat{\mu}_{ilcv}(x; k)$. $E^* \hat{\mu}_{ilcv}(x; h)$ can then be generated by first generating B bootstrap samples, $(Y_1^b, \dots, Y_n^b) \sim \text{Bernoulli}(\hat{\mu}_{ilcv}(\cdot; k))$, for $b = 1, \dots, B$, and then smoothing each sample to get $\hat{\mu}_{ilcv}^b(x; h)$. The resulting estimator is

$$\frac{1}{B} \sum_{b=1}^B \hat{\mu}_{ilcv}^b(x; h).$$

Notice that a plug-in estimator was used for the variance component and a bootstrap estimator for the bias component.

Hall, Marron, and Park (1992) suggest choosing k larger than h_{MISE} so that the estimator of the bias component is based on an oversmoothed version of the data.

SCV estimator:

The SCV estimator, \hat{h}_{SCV} is the minimizer of

$$\frac{(b-a)}{nh} \int K^2(z) dz \int \hat{\mu}_{ilcv}(x; \lambda) (1 - \hat{\mu}_{ilcv}(x; \lambda)) dx + \int \left(\frac{1}{B} \sum_{b=1}^B \hat{\mu}_{ilcv}^b(x; h) - \hat{\mu}_{ilcv}(x; k) \right)^2 dx$$

where $\lambda = \hat{h}_{plug1}(X)$.

4.1.4 Bootstrap

The last bandwidth estimator considered here, the bootstrap, is also based on an approximation to the MISE of $\hat{\mu}_{ilcv}(x; h)$. The bootstrap estimator of h_{MISE} was first introduced by Faraway (1990) and Faraway and Jhun (1990) in the contexts of nonparametric regression and density estimation, respectively. They point out that although the usual bootstrap estimate of the integrated variance,

$$\frac{1}{B} \sum_{b=1}^B \int \left(\hat{\mu}^b(x; h) - \bar{\mu}(x; h) \right)^2 dx,$$

should perform well, the usual bootstrap estimate of the integrated bias²,

$$\int \left(\hat{\mu}(x; h) - \bar{\mu}(x; h) \right)^2 dx, \tag{10}$$

does not. (Here $\hat{\mu}$ denotes a generic smooth of the data (i.e. nonparametric regression or density estimator), $\hat{\mu}^b(x; h)$ is the smooth of the b^{th} bootstrap sample, and $\bar{\mu}(x; h) = \frac{1}{B} \sum_{b=1}^B \hat{\mu}^b(x; h)$. Faraway shows that the difference, $\hat{\mu}(x; h) - \bar{\mu}(x; h)$, tends to decrease as h increases, incompatible with the true bias and bandwidth relationship. This allows (10) to dominate the approximation, resulting in a bandwidth that is too large.

Faraway (1990) and Faraway and Jhun (1990) propose instead that $\bar{\mu}$ be replaced by the same form of nonparametric smoother, with a different bandwidth k . For LPKR, this becomes

$$\frac{1}{B} \sum_{j=1}^B \int \left(\hat{\mu}_{ilcv}^b(x; h) - \hat{\mu}_{ilcv}(x; k) \right)^2 dx.$$

Faraway shows that provided $h = O\left(n^{-\frac{1}{5}}\right)$, and $ng^{-\frac{1}{5}} \rightarrow \infty$ as $n^{-1}, g \rightarrow 0$, then the bootstrapped difference $n^{\frac{2}{5}} \left(\hat{\mu}^b(x; h) - \hat{\mu}(x; k) \right)$ has the same asymptotic distribution as $n^{\frac{2}{5}} \left(\hat{\mu}(x; h) - \mu(x) \right)$. This indicates that an oversmoothed pilot curve is needed to handle the bias.

Bootstrap estimator:

The bootstrap estimator is

$$\hat{h}_{boot}(X) = \arg \min_h \left[\frac{1}{B} \sum_{j=1}^B \int \left(\hat{\mu}_{ilcv}^b(x; h) - \hat{\mu}_{ilcv}(x; k) \right)^2 dx \right].$$

5 Simulation

A simulation study was conducted to assess and compare the performance of the estimators proposed in Section 4.1.

Four mean functions were considered:

- 1) a linear logistic curve, $\mu_1(x) = \frac{1}{1+e^{-4x}}$, for $-2 \leq x \leq 2$
- 2) a complementary log-log curve, $\mu_2(x) = 1 - e^{-e^{4x}}$, for $-1 \leq x \leq 0.5$
- 3) a quadratic logistic curve, $\mu_3(x) = \frac{1}{1+e^{-(-1+12x-12x^2)}}$, for $0 \leq x \leq 1$
- 4) and a fifth degree polynomial,

$$\mu_4(x) = \frac{1}{2.77} \left(19.1 - 57.1x + 63x^2 - 31.9x^3 + 7.6x^4 - 0.69x^5 \right), \text{ for } 0.5 \leq x \leq 3.5.$$

Figure 1 contains plots of these mean functions. The X_i 's were generated from a uniform distribution on their respective intervals. Binary responses, Y_i , were then generated based on $\mu_j(X_i)$ for $j = 1, \dots, 4$. Sample sizes of $n = 50, 100$, and 500 were used for the ROT and plug-in methods. A sample size of $n = 50$ was used for the bootstrap and smoothed cross-validation methods since these methods are very computationally expensive. The number of replications for each of the twelve mean function/sample size scenarios was 100.

For the estimated bandwidths, all integrals of nonparametric smoothers were estimated by Simpson's method, with the number of partitions equal to the sample size. For the plug-in bandwidth selectors, α was set at 5%. The pilot bandwidth for the estimated integrated variance for $\hat{h}_{plug3}(X)$ was $\hat{h}_{plug1}(X)$. For SCV and the bootstrap bandwidth estimator, B , the number of bootstrap samples, was set at 100, and k was chosen to be $n^{-\frac{1}{7}}$.

Much of the discussion is based on the type of results displayed in Figure 2, which shows the logarithm of \hat{h}/h_{ASE} for each of the bandwidth estimators for sample size 50 and mean function μ_2 .

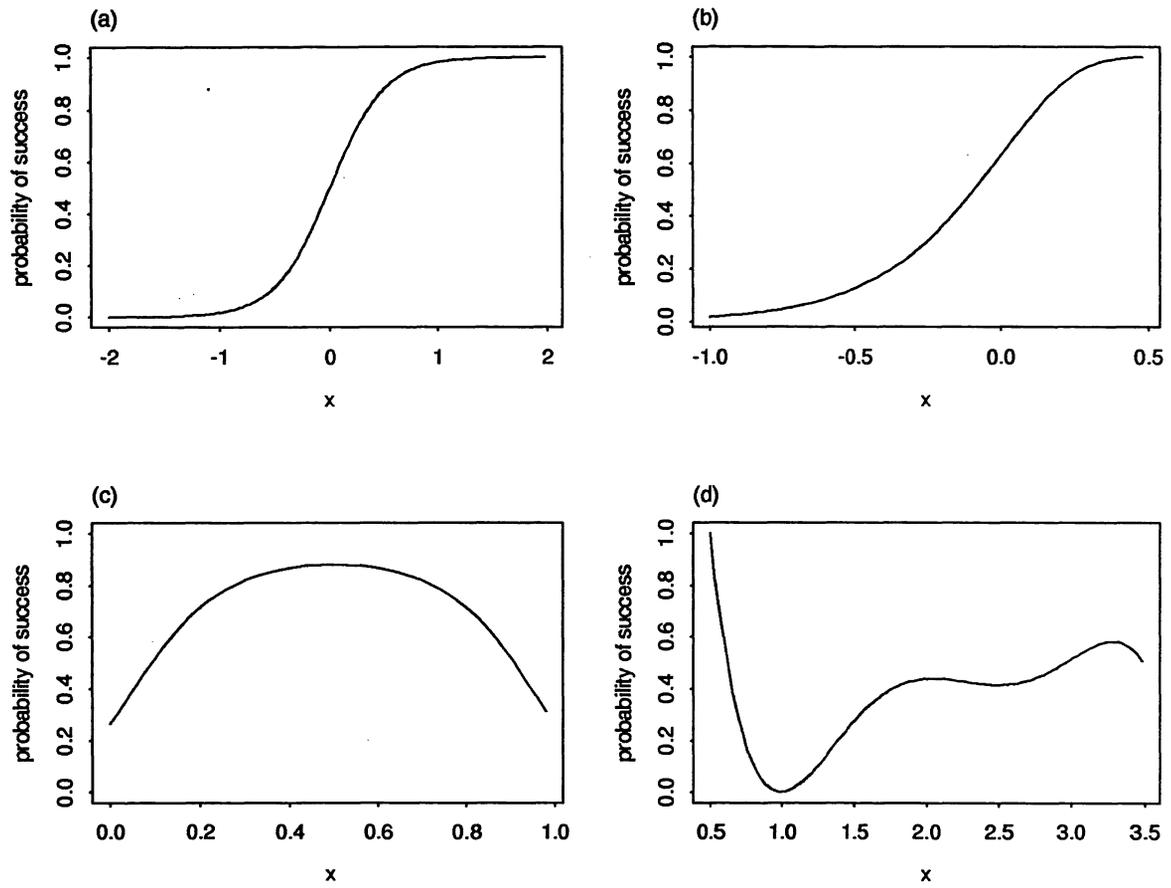


Figure 1: Mean functions: a) linear logit, b) complementary log-log, c) quadratic logit and d) 5th degree polynomial

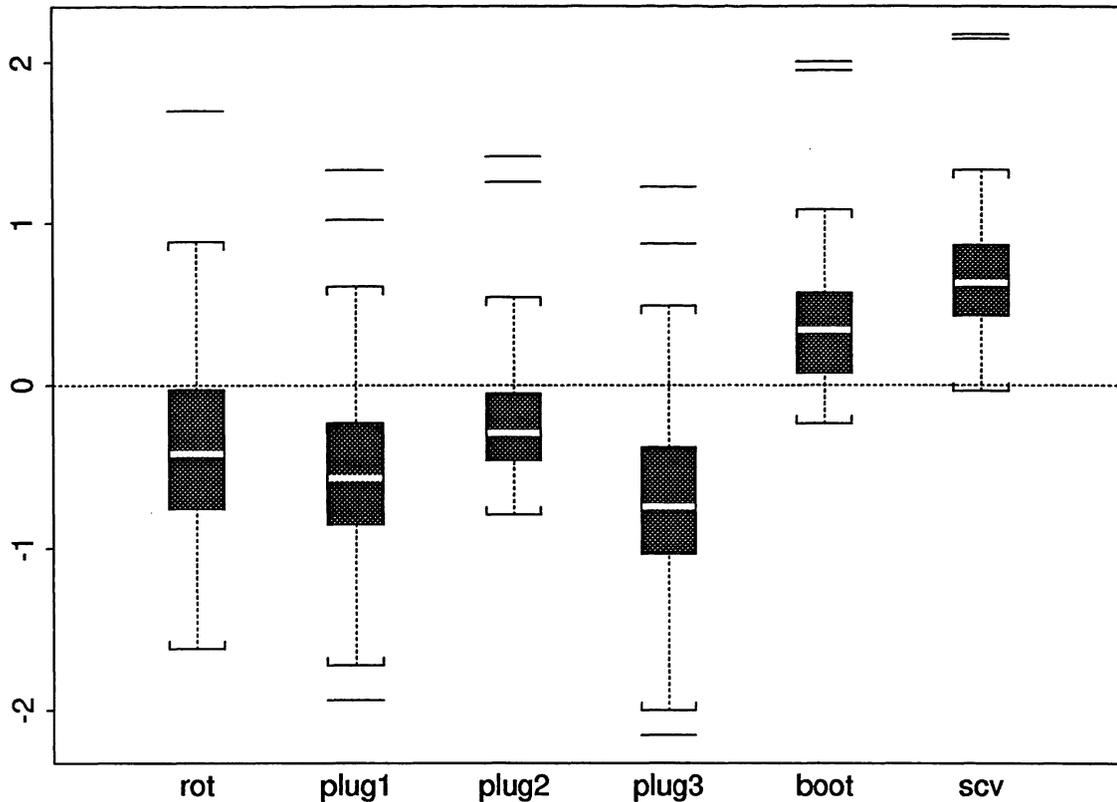


Figure 2: Boxplots of $\log(\hat{h}(X)/h_{ASE})$ for the complementary log-log regression function and sample size 50.

Section 5.1 discusses some of the important characteristics of the bandwidth estimators that can be seen graphically and offers explanations. Section 5.2 suggests which estimators are best by ranking them with respect to average squared error. Section 5.3 summarizes guidelines for choice of bandwidth selection method.

5.1 Performance of the Bandwidth Selection Methods

5.1.1 Plug-in Methods

Boxplots of estimated bandwidths and bandwidth ratios, showed that the plug-in methods tended to undersmooth the data. The undersmoothing was severe for $n = 50$, but tended to be much less of a problem for larger n . All the plug-in methods performed well for $n = 500$.

Plug-in methods tend to produce bandwidths that are too small because although the estimators of $\int \mu''(x)^2 dx$ are consistent, large samples sizes are required to obtain reasonable

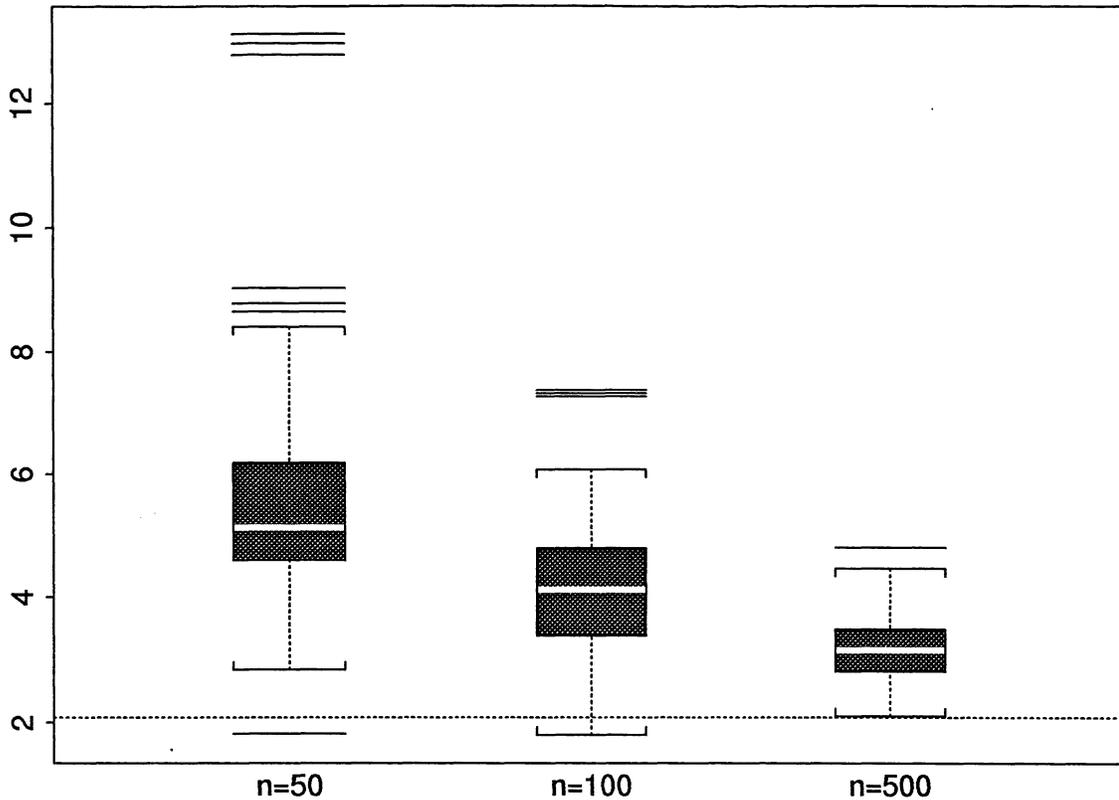


Figure 3: Estimates of $(\int \mu''(x)^2 dx)^{1/5}$ for the quadratic logit. The true value is indicated by the dotted line

estimates. Due to the high variance of binary data, $\hat{\mu}''(x)^2$ tends to be too wiggly resulting in very large estimates of $\int \mu''(x)^2 dx$. Figure 3 shows boxplots of the estimated value of $(\int \mu''(x)^2 dx)^{1/5}$ used in $\hat{h}_{plug3}(X)$. Notice that these estimated values are extremely poor for $n = 50$, but get better as n increases.

Ironically, the plug-in estimator that provides the only consistent estimate of the numerator, $\hat{h}_{plug3}(X)$, does the worst of all three plug-in methods. The simple estimator, $(b - a)\bar{Y}(1 - \bar{Y})$, tends to over-estimate $\int_b^a \mu(x)(1 - \mu(x)) dx$, which helps compensate for the larger than optimal denominator. As displayed in Figure 4, the estimator of the integrated variance is reasonably good. Plots of the regression estimators show that undersmoothing is less of a problem for estimates using $\hat{h}_{plug2}(X)$ instead of $\hat{h}_{plug1}(X)$, especially for $n = 50$.

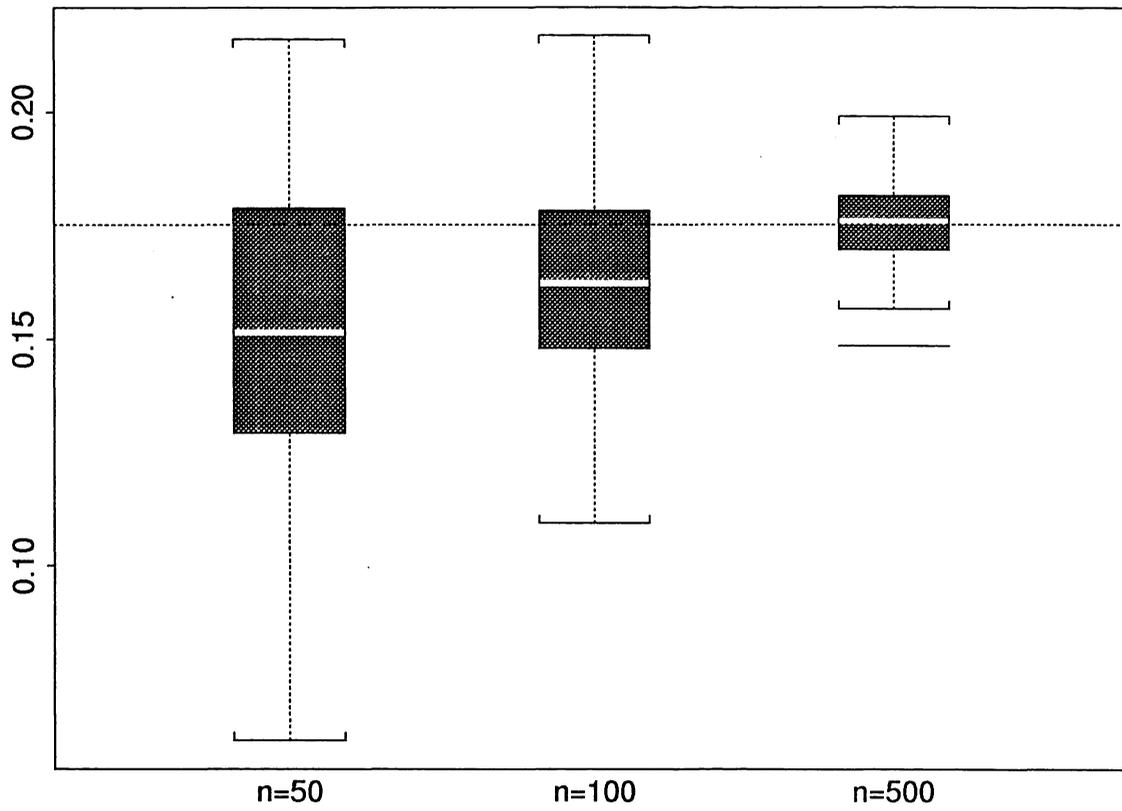


Figure 4: Estimates of $\int \mu(x)(1 - \mu(x))dx$ for the quadratic logit. The true value is indicated by the dotted line.

5.1.2 Rule of Thumb

For the first three types mean functions, $\hat{h}_{ROT}(X)$ tended to perform very similarly to $\hat{h}_{plug2}(X)$; and for small n , were of noticeably higher quality than estimates based on $\hat{h}_{plug1}(X)$ or $\hat{h}_{plug3}(X)$. This is to be expected because $\mu_1(x)$ and $\mu_3(x)$ are logistic curves, and $\mu_2(x)$ and its second derivative are shaped very similarly to a linear logistic curve and its second derivative.

Since $\mu_4(x)$ does not resemble a polynomial logistic curve, $\hat{h}_{ROT}(X)$ did not perform as well. Estimates using $\hat{h}_{ROT}(X)$ tend to oversmooth $\mu_4(x)$ for $n = 100$, and especially for $n = 500$. This highlights the lack of flexibility of the parametric rule of thumb estimator.

5.1.3 SCV and the Bootstrap

Both the bootstrap and SCV methods did very well, although they tend to slightly oversmooth the data generated from the smoother curves, $\mu_2(x)$ and $\mu_3(x)$. In these cases, $\mu(x)$ has very little structure, so the oversmoothed pilot estimator used in estimating the bias component has even less structure leading to underestimating the bias.

The oversmoothing due to the bootstrap and SCV estimators is not nearly as distressing as the undersmoothing due to the plug-in estimators for $n = 50$. In general, AMISE increases quite rapidly as h approaches zero, but increases slowly as h moves past h_{AMISE} . In fact, MISE and ASE as a function of h also have this general shape (see Härdle 1990). Therefore, both the bootstrap and SCV produce reasonable estimates.

5.2 Rankings

In hopes of determining which bandwidth selection method resulted in the best estimate of $\mu(x)$, the ASEs were compared via a nonparametric variant of Fisher's Protected LSD. For each of the twelve probability function/sample size scenarios, Friedman's test (see Conover 1980) was used to determine whether there were differences among the bandwidth selectors. Here a block is a sample, the treatments are the selection procedures, and $\alpha = 0.05$. If a difference was detected, paired Wilcoxon tests were used to rank the selection methods based on their median ASE's (Ruppert, Sheather and Wand 1994). These tests were done at $\alpha = 0.05/c$ level, where c is the number of pairwise comparisons. The results are given in Table 1. A ranking of "1" indicated the best method, and the *'s in indicate "ordered" ties. For example, for $n = 100$, $\hat{h}_{plug2}(X)$ was significantly better than all other estimators. On the other hand, $\hat{h}_{plug1}(X)$ was significantly better than $\hat{h}_{plug3}(X)$ but not $\hat{h}_{ROT}(X)$, and $\hat{h}_{plug2}(X)$ was not significantly different from $\hat{h}_{ROT}(X)$.

Mean func.	rot	plug1	plug2	plug3	boot	scv
$n = 50$						
$\mu_1(x)$	4	5	3	6	1	1
$\mu_2(x)$	3	5	1	6	1	3
$\mu_3(x)$	1	5	1	6	1	4
$\mu_4(x)$	1	4	4	6	1	1
mean rank	2.25	4.75	2.25	6	1	2.25
$n = 100$						
$\mu_1(x)$	3*	2*	1	4		
$\mu_2(x)$	2	2	1	4		
$\mu_3(x)$	1	3	1	4		
$\mu_4(x)$	3*	1	4*	2*		
mean rank	2.25	2	1.75	3.5		
$n = 500$						
$\mu_1(x)$	3	3	1	2		
$\mu_2(x)$	1	1	1	4		
$\mu_3(x)$	1	3	1	4		
$\mu_4(x)$	1	1	2	2		
mean rank	2.25	2	1.25	3		

Table 1: Rankings of the bandwidth selection methods based on a Friedman test of the equality of the average squared error. 1 indicates the best method (lowest average squared error). An asterisk (*) indicates that the method did not differ significantly from a method with adjacent rank.

From Table 1, we see that the bootstrap method does well for $n = 50$, but was too computationally intensive to assess for larger sample sizes. The SCV method also does well, although it is prone to oversmoothing when $\mu(x)$ has little structure as indicated by its rankings for $\mu_2(x)$ and $\mu_3(x)$. $\hat{h}_{ROT}(X)$ and $\hat{h}_{plug2}(X)$ are also quite good but $\hat{h}_{plug3}(X)$ cannot be recommended for small samples.

We see that $\hat{h}_{plug2}(X)$ does the best in many of the scenarios. $\hat{h}_{ROT}(X)$ and $\hat{h}_{plug1}(X)$ seemed to do equally well, while $\hat{h}_{plug3}(X)$ was ranked last more often than not.

5.3 Summary

For small sample size, the bootstrap method is recommended. For samples too large to bootstrap, $\hat{h}_{plug2}(X)$ appears to work well. (Note that bootstrapping is practical for much larger samples than used here - in a simulation study, the need to bootstrap numerous samples is a limiting factor.) Neither method requires estimation of the design density.

6 Analysis of Periparturient Recumbency Data

In New Zealand, 3% to 5% of dairy cows will suffer from periparturient recumbency (Clark et al, 1987), a potentially fatal syndrome following calving. For humanitarian and economic reasons, it is important to predict which cows are likely to recover. The data analysed in this section were collected by Clark et al (1987). A sample of 110 cows diagnosed with the syndrome were examined, and blood and urine samples were taken from each animal. Recovery is considered a “success”. Animals which did not recover either died or were euthanized due to poor prognosis. Here we will consider the binary regression on serum urea, which was found by the investigators to be one of the predictors of recovery. Both high and low levels of serum urea are associated with poor health in dairy cows.

These data were previously analyzed in Altman (1992) using kernel regression and an ad hoc bandwidth selector, and in Altman and McGibbon (1996) using kernel regression with bandwidth selected by CV and generalized cross-validation.

Figure 5 shows the LPKR estimates of recovery based on the 2 most promising bandwidth selection methods, plug-in estimator 2, and the bootstrap. Although the bootstrap estimate is much smoother, both estimates show a peak in the recovery probability at about $\log(\text{Serum Urea})=1.7$ and then a slow decline in the probability with increasing urea. There is also a hint of a small bump near $\log(\text{Serum Urea})=3$.

The plug-in estimator actually picked the smallest bandwidth for these data, while SCV picked the largest, and showed evidence of severe over-smoothing.

A parametric model was also fitted to the data for comparison. A suitable parametric model was difficult to find. Since use of logit, probit and complementary log-log link functions led to similar results, the discussion will be limited to the logit link.

The analysis began by fitting a fifth degree logit polynomial curve, and proceeded by backwards selection, removing the highest order term when the change of deviance was not significant at level $\alpha = .05$.

This process led to a linear logistic model, which seems very unrealistic. Figure 6a shows the linear logistic fit, the 95% pointwise confidence interval for the linear logistic, and $\hat{\mu}_{ilcv}(x)$. Notice that the linear logistic does not account for the mode or the flat region at high serum urea. The failure of the parametric model to identify the mode is particularly distressing because it is very evident in the data, and it is known that both high and low serum urea are indicators of poor health. The cubic logistic fit (Figure 6b) is much closer to the nonparametric fit, and picks up the features of interest.

7 Conclusions

In this article we have shown that LPKR is a consistent estimator of a smooth mean function for binary response data. Although the estimate is not constrained to lie in the interval $[0, 1]$, its asymptotic variance and bias are of the same order of magnitude, and the bias is of simpler form, than those of the generalized smooth estimator proposed by Fan, Heckman and Wand (1995). LKPR offers computational advantages over the FHW estimator, as it is linear and can be computed in $O(n)$ operations.

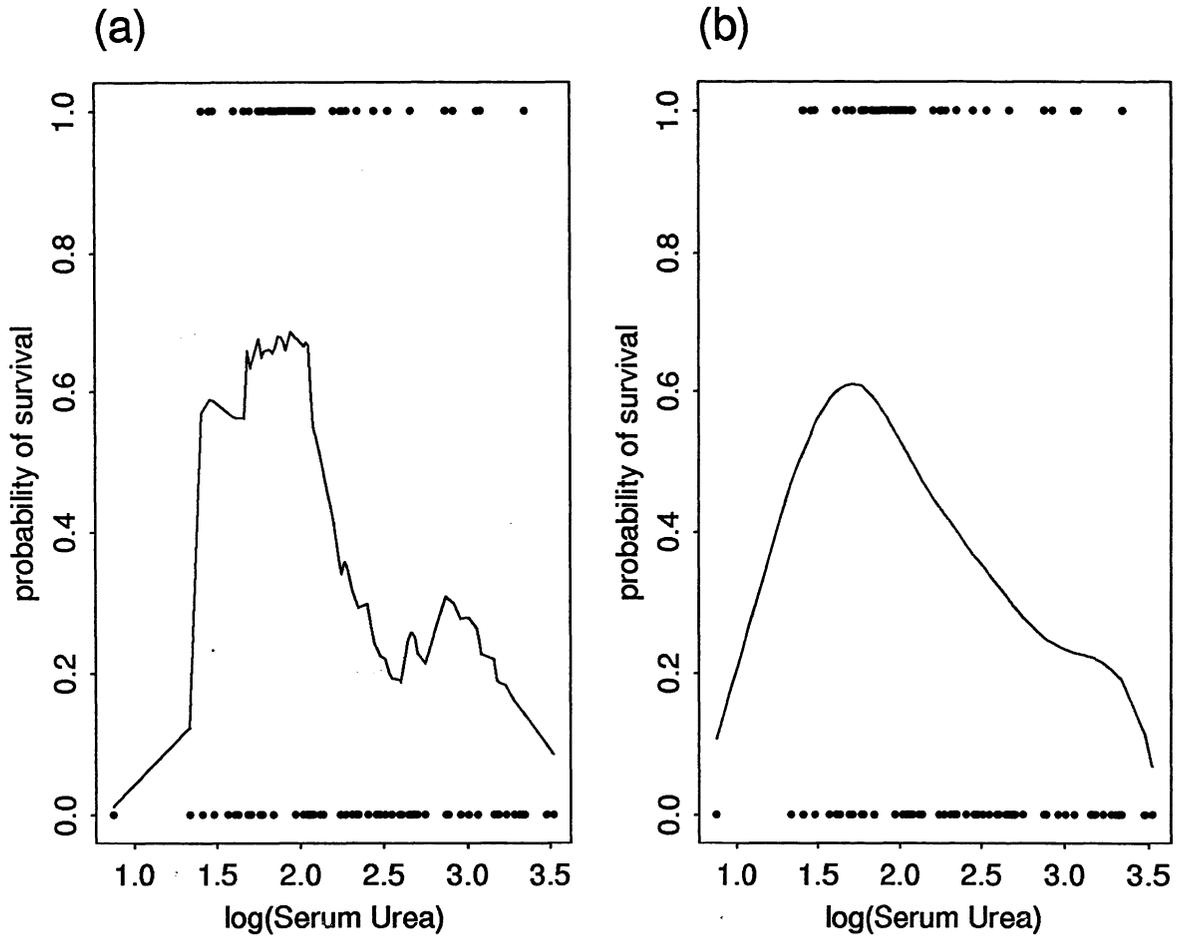


Figure 5: Local linear estimator of survival probability of periparturient recumbent cows as a function of log(Serum Urea) with bandwidth selected by (a) plug2 and (b) bootstrap.

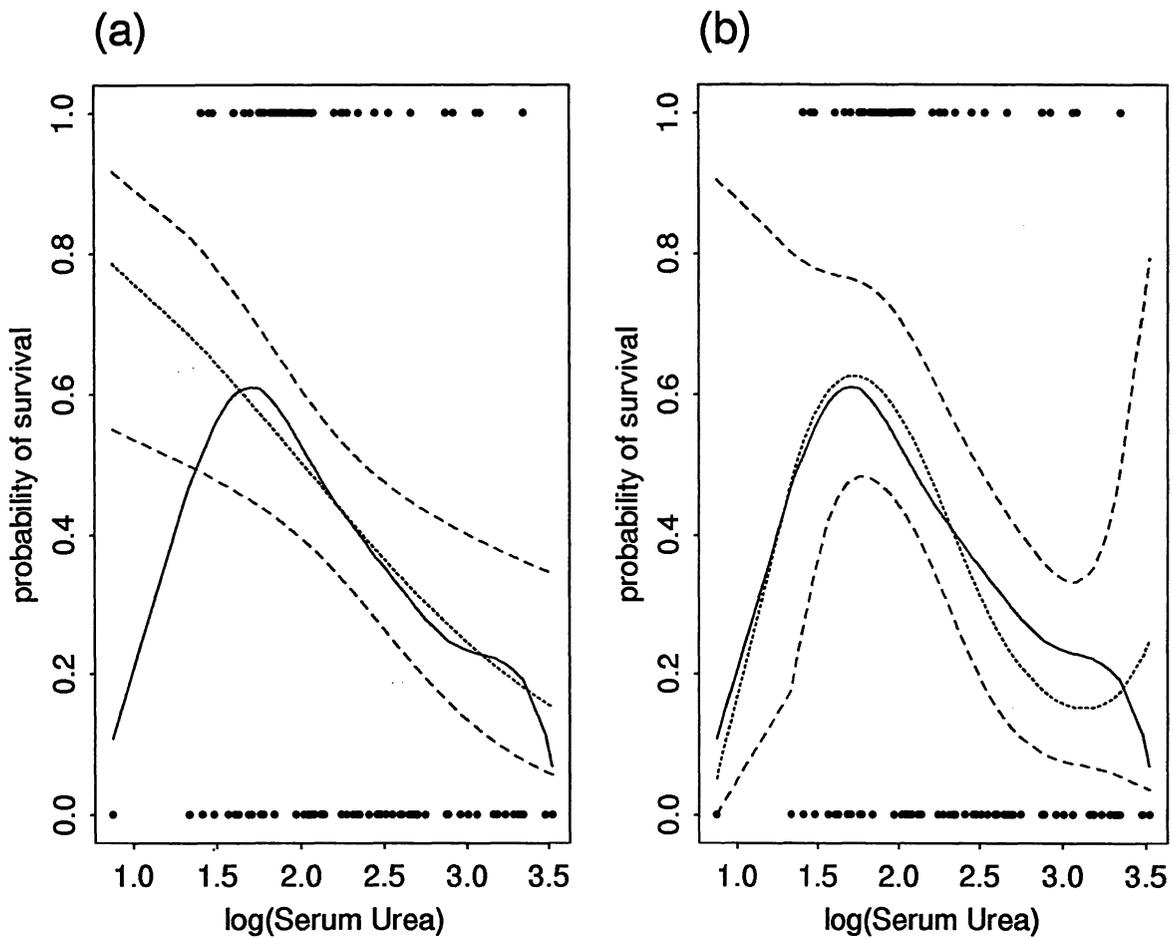


Figure 6: Polynomial logit estimates of survival probability as a function of $\log(\text{Serum Urea})$. (a) Linear logit (b) Cubic logit. The solid line is the local linear fit. The dotted line is the parametric fit. The dashed line is the parametric 95% confidence interval for the parametric fit.

A number of bandwidth selectors proposed for nonparametric regression with continuous response were tested for binary response data. Because binary response data have a low signal to noise ratio, plug-in bandwidth estimators, which require estimation of integrated derivatives, perform poorly for small sample sizes. Bootstrap bandwidth selection performed best in the simulations and provided a reasonable picture for the example analyzed here.

Nonparametric regression has been used with continuous response to provide guidance about the goodness-of-fit of a parametric model (Cox and Koh, 1989; Cox et al, 1988; Eubank and Speigelman, 1990). In Section 6 we used the nonparametric regression estimate to supplement the parametric analysis. Although the parametric analysis of deviance did not show a lack of fit of a linear logistic model, the biological theory indicated that both low and high values of the predictor were detrimental to survival. The LKPR fit indicated this as well, and informal visual inspection showed that a cubic logistic model provides a reasonable parametric fit to the data. Azzalini, Bowman and Härdle (1989) suggest an informal method for using nonparametric regression to test goodness-of-fit for parametric binary regression. Formal testing methods for goodness-of-fit of binary regression models, such as those provided by the references above for continuous response, would be a useful addition to the data analysis toolkit.

References

- Allen, D. M. (1974) "The relationship between variable selection and data augmentation and a method for prediction," *Technometrics*, **16**, 1307-1325.
- Altman, N.S. and MacGibbon, B. (1996), "Consistent Bandwidth Selection for Kernel Binary Regression," BU-1126-M.
- Altman, N.S. (1992), "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression," *The American Statistician*, **46**175-185.
- Azzalini, A., Bowman, A.W. and Härdle, W. (1989), "On the Use of Nonparametric Regression for Model Checking," *Biometrika*, **76**1-11.
- Clark, R.G., Henderson, H.V., Hoggard, G.K., Ellison, R.S. and Young, B.J. (1987), "The Ability of Biochemical and Haematological Tests to Predict Recovery in Periparturient Recumbent Cows," *New Zealand Veterinary Journal*, **35**126-133.
- Cleveland, W. S. (1979) "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, **74**, 829-836.
- Cleveland, W.S. and Devlin, S.J. (1988), "Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting," *Journal of the American Statistical Association*, **83**596-610.
- Conover, W.J. (1980), *Practical Nonparametric Statistics 2nd Edition*, New York: John Wiley and Sons.
- Cox, D. D., and Koh E. (1989), "A smoothing spline based test of model adequacy in polynomial regression," *Annals of the Institute of Statistical Mathematics*, **41**, 383-400.
- Cox, D. D., Koh, E., Wahba, G. and Yandell, B. (1988) "Testing the (parametric) null model hypothesis in (semiparametric) partial and generalized spline models," *Annals of Statistics*, **16**, 113-119.

- Craven, P. and Wahba, G. (1979) "Smoothing Noisy Data with Spline Functions". *Numerische Mathematik* **31**, 377-403.
- Eubank, R. L. and Spiegelman, C. H. (1990) "Testing the Goodness of Fit of a Linear Model Via Nonparametric Regression Techniques," *Journal of the American Statistical Association*, **85**, 387-392.
- Fan, J. (1992), "Design-Adaptive Nonparametric Regression," *Journal of the American Statistical Society* **87**998-1004.
- Fan, J., Gasser, T., Gijbels, I., Brockmann, M., and Engel, J. (1996) "Local polynomial regression: optimal kernels and asymptotic minimax efficiency". (to appear: *Annals of the Institut. of Math. Statist.*).
- Fan, J., Heckman, N.E. and Wand, M.P. (1995), "Local Polynomial Kernel Regression for Generalized Linear Models and Quasi-Likelihood Functions," *Journal of the American Statistical Association* **90**141-150.
- Fan, J. and Marron, J.S. (1994) "Fast implementations of nonparametric curve estimators," *Journal of Computational and Graphical Statistics*, **3**, 35-56.
- Faraway, J. (1990), "Bootstrap Selection of Bandwidth and Confidence Bands for Nonparametric Regression," *Journal of Statistical Computation and Simulation* **37**37-44.
- Faraway, J. and Jhun, M. (1990), "Bootstrap Choice of Bandwidth for Density Estimation," *Journal of the American Statistical Association*, **85**1119-1122.
- Geisser, S. (1975) "The Predictive Sample Reuse Method with Applications," *Journal of the American Statistical Association*, **70**, 320-328.
- Hall, P., Sheather, S. J., Jones, M. C., Marron, J. S. (1991) "On optimal data-based bandwidth selection in kernel density estimation," *Biometrika*, **78**, 263-269.
- Hall, P., Marron, J.S. and Park, B.U. (1992), "Smoothed Cross-Validation," *Probability Theory and Related Fields*, **92**1-20.
- Härdle, W., Hall, P. and Marron, J.S. (1988), "How Far Are Automatically Chosen Regression Smoothing Parameters From Their Optimum?," *Journal of the American Statistical Association*, **83**86-95.
- Hastie, T. and Loader, C. (1993), "Local Regression: Automatic Kernel Carpentry," *Statistical Science*, **8**120-143.
- Jones, M.C. (1993), "Do Not Weight for Heteroscedasticity In Nonparametric Regression," *Austral. J. Statist.*, **35**89-92.
- McCullagh, P. and Nelder, J.A. (1989), *Generalized Linear Models*, London: Chapman and Hall.
- Park, B.U. and Marron, J.S. (1990), "Comparison of Data-Driven Bandwidth Selectors," *Journal of the American Statistical Association*, **85**66-72.
- Ruppert, D.R. and Wand, M.P. (1994), "Multivariate Locally Weighted Least Squares Regression," *Annals of Statistics* **22**,1346-1370.
- Ruppert, D.R., Sheather, S.J. and Wand, M.P. (1995), "An Effective Bandwidth Selector for Local Least Squares Regression," *Journal of the American Statistical Association*, **90**, To appear.
- Sheather, S.J. and Jones, M.C. (1991), "A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation," *JRSS B*, **53**683-690.

- Seifert, B., Brockmann, M., Engel, J and Gasser, T. (1994) "Fast algorithms for nonparametric curve estimation." *Journal of Computational and Graphical Statistics*, **3**, 192-213.
- Wedderburn, R.W.M. (1974), "Quasi-likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method," *Biometrika* **61** 439-447.