Mem. Fac. Sci. Eng. Shimane Univ.
Series B: Mathematical Science
44 (2011), pp. 33–53

KERNEL REGRESSION FOR BINARY RESPONSE DATA

HIDENORI OKUMURA

Communicated by Kanta Naito

(Received: February 7, 2011)

Abstract. This paper is based on the author's thesis, "Kernel Regression for Binary Response Data". We consider kernel-based estimators with additional weights of the regression functions in nonparametric binomial and binary regressions. Firstly, in the binomial regression with a single covariate and a fixed covariate design, we introduce a weighted Nadaraya-Watson estimator and its bias adjusted estimator discussed in Okumura and Naito [13]. Secondly, in the binary regression with multiple covariates and a random covariate design, we propose weighted local linear estimators.

1. Introduction

In various fields such as pharmacology, toxicology, econometrics etc, researchers are often interested in understanding the relationship between a binary response and a single covariate/covariate vector. The formal description of the relationship is as follows. Suppose that the response Y of a subject, which is encoded as $Y = 1$ if it exhibits interest and $Y = 0$ if it does not, is observed at each covariate/covariate vector x. The relationship can be expressed as

$$\Pr(Y = 1|x) = p(x),$$

where $p(x)$ is a function of x. The function $p(x)$ is referred to as the dose response curve in biostatistics. The aim of this paper is to estimate $p(x)$ with respect to x, which has various applications to practical situations.

This study focuses on the kernel smoothing methods used in nonparametric binary/binomial regression. The Nadaraya-Watson estimator (NWE) is one of the most popular and simplest estimators, which is referred to as the local constant estimator. Copas [2] discussed the application of NWE in the binary settings with a single covariate (Staniswalis and Cooper [17] and Müller and Schmitt [10]). The ordinary local linear estimator (OLLE) is superior to NWE in terms of the minimax efficiency and the boundary effect when the response variable is continuous in general (Fan [3] and Fan and Gijbels [4]). NWE is bounded on $[0, 1]$ while OLLE is

2000 *Mathematics Subject Classification.* 62G08, 62G20.
Key words and phrases. kernel regression, binary response.

not necessarily bounded. Aragaki and Altman [1] used the truncated LLE bounded in binary settings with a single covariate. To avoid the boundary problem, the local linear logistic estimator (LLLE) is considered, which is a special case of the local polynomial quasi-likelihood estimators discussed by Fan et al. [5]. Since LLLE is defined through the optimal solution of an optimization problem, an iterative calculation is required to derive the solution in practice. Moreover, in the case of single covariate settings, Signorini and Jones [16] proposed an NW-type estimator involving two bandwidths and Hazelton [7] proposed bias reduction for NWE.

We propose the kernel estimators of $p(x)$ with additional weights for binary and binomial regression data. In most studies on binomial settings, the dimension of the covariate is assumed to be one. On the other hand, in binary settings, it is not necessarily so. Hence, we consider the settings with multiple covariates, i.e., a covariate vector. This paper deals with the binomial regression problem with a single covariate and the binary regression problem with multiple covariates.

The remainder of this paper is organized as follows. In Section 2, the standard kernel smoothing methods are summarized from a general viewpoint, and NWE, OLLE and LLLE are given.

Sections 3 and 4 investigate the use of kernel estimators in nonparametric binomial regression with a single covariate. In Section 3, for the binomial data, a weighted NWE is derived by considering the variance of the proportion of responses at each covariate. Moreover, the efficient bias-adjusted estimator proposed in Okumura and Naito [13] is reviewed. In addition, an application of the bias-adjusted estimator to the quantal bioassay, discussed in Okumura and Naito [12], is also given. In Section 4, two methods for data-driven bandwidth selection for the bias-adjusted estimator discussed in Okumura and Naito [14] are described: the rule of thumb (ROT) method and the plug-in (PI) method. A scale adjustment and an efficient ROT method are discussed to improve the performance of the proposed method from practical viewpoints. The asymptotic properties of the proposed PI bandwidth selector are given.

Sections 5 and 6 are concerned with the kernel estimators in the nonparametric binary regression with multiple covariates. In particular, we will discuss the efficiency of weighting for the local linear type estimators (LLTEs). In Section 5, the LLTEs with weights are proposed, and their asymptotic properties are given. In Section 6, the ROT method and the PI method for data-driven bandwidth selection are discussed as in Section 3. Finally, we conclude the paper. Proofs of all theorems are omitted.

2. Regression with binary data

2.1. Binomial regression models. In bioassay, binomial responses are often observed at K levels designed through a single covariate, for example, a drug dose or a toxic dose. The sample can be expressed as independent random variables $Y_i (i = 1, \dots, K)$ having the binomial distribution $Bi(N_i, p(x_i))$ corresponding to covariates x_i, where p is unknown. Then, the binomial data is generally expressed as $\mathcal{B}_1 = \{(x_i, Y_i, N_i) : i = 1, \dots, K\}$. We consider the equispaced fixed design.

However, this assumption is not essential and can be relaxed by the manipulation described in Müller and Schmitt [10]. Without lack of generality, the covariates x_i are assumed to be

$$x_i = \frac{i-1}{K-1}, \qquad i = 1, ..., K.$$

Then, a regression model to estimate $p(x)$ is given as $\bar{Y}_i = p(x_i) + \bar{\epsilon}_i,\ i = 1, \ldots, K$, where $\bar{Y}_i = Y_i/N_i$ and $\bar{\epsilon}_i = \bar{Y}_i - p(x_i)$. The data set in this setting can be expressed as $\mathcal{B}_2 = \{(x_i, \bar{Y}_i) : i = 1, \ldots, K\}$. If Y_i can be expressed as $Y_i = \sum_{j=1}^{N_i} Y_{ij}$, where Y_{ij} has the Bernoulli distribution $Bi(1, p(x_i))$ for $j = 1, \ldots, N_i$, then another regression model can be given as $Y_{ij} = p(x_i) + \epsilon_i,\ i = 1, \ldots, K, j = 1, \ldots, N_i$, where $\epsilon_i = Y_{ij} - p(x_i)$. Then, $\mathcal{B}_3 = \{(x_i, Y_{ij}) : i = 1, \ldots, K, j = 1, \ldots, N_i\}$ is a general expression of the binary data.

2.2. **Binary regression models.** We consider the random design of multiple covariates. In this case, the outcome at each covariate vector is mostly observed as a binary scale. The binary data can be expressed as follows. Let $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ be independent random vectors, where $\mathbf{X}_i = (X_{i1}, \ldots, X_{id})^T$ is a d-dimensional random vector with a probability density function $\varphi(\mathbf{x})$ and Y_i is a binary random variable. In addition, assume that $\Pr(Y_i = 1|\mathbf{X}_i = \mathbf{x}) = p(\mathbf{x})$ for all i, where $p(\mathbf{x})$ is an unknown function of $\mathbf{x}$. The function $p(\mathbf{x})$ is to be estimated. A binary regression model is given as $Y_i = p(\mathbf{X}_i) + \varepsilon_i,\ i = 1, ..., n$, where $\varepsilon_i = Y_i - p(\mathbf{X}_i)$. Then, $\mathcal{B}_4 = \{(\mathbf{X}_i, Y_i) : i = 1, \ldots, n\}$ is a general expression of the binary data with multiple covariates.

2.3. **Kernel regression approaches.** We will introduce the standard kernel regression estimators. These can be derived from a general viewpoint. Assume that $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ are observed independently, where $\mathbf{X}_i$ is a d-dimensional random/nonrandom vector and Y_i has the distribution $Bi(1, p(\mathbf{X}_i))$. The kernel regression approaches are the methods to estimate $p(\mathbf{x})$ for a fixed $\mathbf{x}$ by fitting a parametric approximation to the observations $(\mathbf{X}_i, Y_i)$ using the weighted sum of divergence measures. In general, the standard kernel regression approaches can be described as follows. Let $D(s, t)$ be a measure that specifies the divergence of s and t and $\mathbf{Z}_i = (1, (\mathbf{X}_i - \mathbf{x})^T)^T$ for $i = 1, \ldots, n$. We consider the following criterion for estimation:

$$\ell(\boldsymbol{\beta}_r : g, D) = \sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{X}_i - \mathbf{x}) D(Y_i, g^{-1}(\boldsymbol{\beta}_r^T \mathbf{Z}_i)),$$

where $\mathbf{H}$ is a symmetric positive definite $d \times d$-matrix called the bandwidth matrix,

$$K_{\mathbf{H}}(\mathbf{x}) = |\mathbf{H}|^{-1/2} \psi(\mathbf{H}^{-1/2}\mathbf{x}),$$

$\psi(\mathbf{x})$ is the nonnegative d-variate kernel function with $\int \psi(\mathbf{x}) d\mathbf{x} = 1$, $\boldsymbol{\beta}_r = (\beta_0, r\beta_1, \ldots, r\beta_d)^T$ for $r = 0, 1$ and $g^{-1}(\boldsymbol{\beta}_r^T \mathbf{Z}_i)$ is a parametric approximation to $p(\mathbf{X}_i)$ in a neighborhood of $\mathbf{x}$. The kernel estimators are obtained by minimizing (2.1). Let $\ell_0(\beta_0; g, D) = \ell(\boldsymbol{\beta}_0; g, D)$ and $\ell_1(\boldsymbol{\beta}; g, D) = \ell(\boldsymbol{\beta}_1; g, D)$, where

$\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_d)^T$. The simplest kernel estimator is NWE that is defined as

$$\hat{p}_{NW}(\mathbf{x}; \mathbf{H}) = \frac{\sum_{i=1}^n w_i Y_i}{\sum_{i=1}^n w_i},$$

where $w_i = K_{\mathbf{H}}(\mathbf{X}_i - \mathbf{x})$. By setting $g_1(t) = t$ and $D_1(s,t) = (s-t)^2$, we see that

$$\ell_0(\beta_0; g_1, D_1) = \sum_{i=1}^n w_i (Y_i - \beta_0)^2.$$

Then, $\hat{p}_{NW}(\mathbf{x}; \mathbf{H})$ is given as $\hat{\beta}_0$ that minimizes $\ell_0(\beta_0; g_1, D_1)$. OLLE can be expressed as

$$\bar{p}_{OLL}(\mathbf{x}; \mathbf{H}) = \mathbf{e}_{1(d+1)}^T (\mathbf{X}_{\mathbf{x}}^T \mathbf{K}_{\mathbf{x}} \mathbf{X}_{\mathbf{x}})^{-1} \mathbf{X}_{\mathbf{x}}^T \mathbf{K}_{\mathbf{x}} \mathbf{Y},$$

where $\mathbf{e}_{k(d+1)}$ is the $(d+1)$ column vector having 1 in the kth entry and 0 in all other entries,

$$\mathbf{X}_{\mathbf{x}} = \begin{pmatrix} 1 & (\mathbf{X}_1 - \mathbf{x})^T \\ \vdots & \vdots \\ 1 & (\mathbf{X}_n - \mathbf{x})^T \end{pmatrix},$$

$\mathbf{Y} = (Y_1, \dots, Y_n)^T$ and $\mathbf{K}_{\mathbf{x}} = \mathrm{diag}(w_1, \dots, w_n)$. Set

$$\ell_1(\boldsymbol{\beta}; g_1, D_1) = \sum_{i=1}^n w_i (Y_i - \boldsymbol{\beta}^T \mathbf{Z}_i)^2.$$

Then, $\bar{p}_{OLL}(\mathbf{x}; \mathbf{H})$ is given as $\hat{\beta}_0$ of $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \dots, \hat{\beta}_d)^T$ that minimizes $\ell_1(\boldsymbol{\beta}; g_1, D_1)$. Fan et al. [5] developed local quasi-likelihood estimators (LQLEs) for generalized linear models with one-parameter exponential families, which can be constituted naturally for binomial regression models. Their method estimates $\eta(\mathbf{x})$ locally under the assumption that $p(\mathbf{x}) = g^{-1}(\eta(\mathbf{x}))$. We assume that $g^{-1}(t)$ is strictly increasing for any t satisfying $0 < g^{-1}(t) < 1$ and $-\log g^{-1}(t)$ and $-\log(1 - g^{-1}(t))$ are convex. Because of its mathematical tractability and efficiency, we only consider the local linear estimator (LLE) derived in their theory, which corresponds to the case $p = 1$ and $r = 0$ in their notations. Set $D_2(s,t) = -\{s \log t + (1-s)\log(1-t)\}$, then,

$$\ell_1(\boldsymbol{\beta}; g, D_2) = -\sum_{i=1}^n w_i \{Y_i \log g^{-1}(\boldsymbol{\beta}^T \mathbf{Z}_i) + (1 - Y_i) \log(1 - g^{-1}(\boldsymbol{\beta}^T \mathbf{Z}_i))\}. \tag{2.1}$$

The LLE $\hat{\eta}_{LL}(\mathbf{x}; \mathbf{H})$ of $\eta(\mathbf{x})$ is given by $\hat{\beta}_0$ of $\hat{\boldsymbol{\beta}}$ that minimizes this $\ell_1(\boldsymbol{\beta}; g_2, D_2)$. The local maximum likelihood estimator (LMLE) of $p(\mathbf{x})$ is given as

$$\hat{p}_{LML}(\mathbf{x}; \mathbf{H}) = g^{-1}(\hat{\eta}_{LL}(\mathbf{x}; \mathbf{H})).$$

Note that $\hat{\boldsymbol{\beta}} = \hat{\eta}_{LL}(\mathbf{x}; \mathbf{H})$ cannot be expressed explicitly. This derivation requires iterative calculations. A method to identify the existence of $\hat{\boldsymbol{\beta}}$ is proposed in Okumura [11]. In particular, if $g^{-1}(t)$ is the logistic function $G(t) = e^t/(1+e^t)$, LMLE is referred to as LLLE, and is denoted as $\hat{p}_{LLL}(\mathbf{x}; \mathbf{H})$.

3. Nonparametric binomial regression

We will discuss estimators for the binomial regression models with a single covariate described in Section 2.1.

3.1. Nadaraya-Watson type estimators. The NWE of $p(x)$ based on data $\mathcal{B}_3$ can be written as

$$\hat{p}_{NW}(x;h) = \frac{\sum_{i=1}^{K} w_i Y_i}{\sum_{i=1}^{K} w_i N_i}$$

(cf. Lloyd [8]), where $w_i = \phi_h(x) = h^{-1}\phi(h^{-1}(x_i - x))$, $\phi(x)$ is a kernel function and h is the bandwidth. Note that $\hat{p}_{NW}(x;h)$ can be constructed based on the binomial data $\mathcal{B}_1$. Throughout this paper, we adopt as $\phi(x)$ a symmetric univariate probability density function whose support is $[-1,1]$ with $\int x\phi(x)dx = 0$ and $\int x^2\phi(x)dx > 0$.

In addition, the NWE of $p(x)$ based on data $\mathcal{B}_2$ is given as

$$\hat{p}_M(x;h) = \frac{\sum_{i=1}^{K} w_i \bar{Y}_i}{\sum_{i=1}^{K} w_i}.$$

This estimator was proposed in Müller [9], and utilized in Staniswalis and Cooper [17] in the multivariate covariate setting. Note that $\hat{p}_M(x;h)$ is equal to $\hat{p}_{NW}(x;h)$ if N_is are equal. From the viewpoint of local polynomial smoothing (Ruppert and Wand [15]), $\hat{p}_{NW}(x;h)$ based on $\mathcal{B}_3$ is characterized as the minimizer of

$$\ell_0(\beta_0; g_1, D_1) = \sum_{i=1}^{K} w_i \sum_{j=1}^{N_i} (Y_{ij} - \beta_0)^2 \tag{3.1}$$

with respect to β_0. Similarly, $\hat{p}_M(x;h)$ based on $\mathcal{B}_2$ is characterized as the minimizer of

$$\ell_0(\beta_0; g_1, D_1) = \sum_{i=1}^{K} w_i (\bar{Y}_i - \beta_0)^2 \tag{3.2}$$

with respect to β_0. In both (3.1) and (3.2), the differences in the variations of Y_{ij}s are not considered in the weights w_is. However, it seems natural that the weights should be related to the variations of the responses. Since $\bar{Y}_i$ has the variance $V[\bar{Y}_i] = v_i/N_i$ at each x_i, where $v_i = p_i(1-p_i)$, our approach for constructing the estimator of $p(x)$ is to use the weighted kernels defined by $N_i w_i / v_i$, $i = 1, \ldots, K$, which provides the criterion

$$\sum_{i=1}^{K} \frac{N_i w_i}{v_i} (\bar{Y}_i - \beta_0)^2$$

with respect to β_0. Through the minimization, we have the *ideal* estimator $p^*(x;h)$ as

$$p^*(x;h) = \frac{\sum_{i=1}^{K} Y_i w_i / v_i}{\sum_{i=1}^{K} N_i w_i / v_i}.$$

The phrase '*ideal*' means that $p^*(x;h)$ contains the v_is that are unknown. Therefore, v_i must be estimated and a reliable estimator should be built up by substituting v_is with the appropriate estimators $\hat{v}_i$s. To obtain the estimator of v_i, we adopt a pilot estimator of p_i defined by

$$\hat{p}_i = \frac{Y_i + \sqrt{N_i}/2}{N_i + \sqrt{N_i}},$$

which is known as the Bayes estimator with respect to the Beta distribution $Beta(s,t)$ with $s = t = \sqrt{N_i}/2$ that has the minimax property. Hence, the estimator of $p(x)$ is given by

$$\hat{p}(x;h) = \frac{\sum_{i=1}^{K} Y_i w_i/\hat{v}_i}{\sum_{i=1}^{K} N_i w_i/\hat{v}_i},$$

where $\hat{v}_i = \hat{p}_i(1-\hat{p}_i), i = 1,\ldots,K$. We call $\hat{p}(x;h)$ the weighted NWE (WNWE).

3.2. **Theoretical performance.** In this section, we develop a theory for the behavior of the proposed estimator and other estimators. In subsection 3.3.1, we consider their performance in the near exact situation. Subsection 3.3.2 considers the performance under a situation familiar in nonparametric smoothing.

3.2.1. *Near exact performance.* It is known that the consistent property of kernel estimators is guaranteed under the situation where h tends to zero as K and N_is go to infinity. However, their consistency is not held under such a near exact situation where h and K are fixed and N_is go to infinity. We assess the behavior of the estimators through their mean squared errors (MSEs) under the near exact situation. First, under the situation where h, K and N_is are fixed, the MSEs of basic estimators $\hat{p}_{NW}(x;h)$ and $\hat{p}_M(x;h)$ can be easily obtained as follows:

$$\text{MSE}[\hat{p}_{NW}(x;h)] = \left[\frac{\sum_{i=1}^{K} N_i p_i w_i}{\sum_{i=1}^{K} N_i w_i} - p(x)\right]^2 + \frac{\sum_{i=1}^{K} N_i v_i w_i^2}{(\sum_{i=1}^{K} N_i w_i)^2}, \tag{3.3}$$

$$\text{MSE}[\hat{p}_M(x;h)] = \left[\frac{\sum_{i=1}^{K} p_i w_i}{\sum_{i=1}^{K} w_i} - p(x)\right]^2 + \frac{\sum_{i=1}^{K} v_i w_i^2/N_i}{(\sum_{i=1}^{K} w_i)^2}. \tag{3.4}$$

Note that (3.4) and (3.5) can be derived without any asymptotic manipulations. In this sense, these are exact evaluations. On the other side, one assumption is needed to obtain the MSE of $\hat{p}(x;h)$.

Assumption 1 K and h are fixed, and $N = \sum_{i=1}^{K} N_i \to \infty$, $N_i/N \to 1/K$ for $i = 1,2,...,K$.

We explain the reason for making this assumption. In fact, there is often a case where the size of K is limited. In such a situation, one is interested in the behavior of the estimators when the number of subjects increases. Assumption 1 expresses the minimum situation wherein the MSE of $\hat{p}(x;h)$ is calculable. Thus, we can evaluate the performance of the estimators under the near exact situation. Direct calculations including Taylor expansion yield the following.

Theorem 3.1. *Under Assumption 1, we have*

$$\begin{aligned}\mathrm{MSE}[\hat{p}(x;h)] &= \left[\frac{\sum_{i=1}^K N_i w_i/(1-p_i)}{\sum_{i=1}^K N_i w_i/v_i} - p(x)\right]^2 \\ &\quad +2\left\{\frac{G_{21}}{\sqrt{N}}\frac{G_{22}}{N}\right\}\left[\frac{\sum_{i=1}^K N_i w_i/(1-p_i)}{\sum_{i=1}^K N_i w_i/v_i} - p(x)\right] \\ &\quad +\frac{G_{21}^2}{N} + \frac{\sum_{i=1}^K N_i w_i^2/v_i}{(\sum_{i=1}^K N_i w_i/v_i)^2} + \frac{H}{N} + o(\frac{1}{N}),\end{aligned}$$

where

$$\begin{aligned}G_1 &= \frac{1}{B^2}\sum_{i\neq j}\frac{N_j w_i w_j}{N v_i^2 v_j}\left\{p_i^3 - p_j(1-3v_i)\right\} - \frac{1}{B^3}\sum_{i,j}\frac{N_i N_j w_i^2 w_j}{N^2 v_i^3 v_j}(p_i-p_j)(1-2p_i)^2, \\ G_{21} &= -\frac{1}{2B^2}\sum_{i,j}\frac{N_j\sqrt{N_i}w_i w_j}{N\sqrt{N}v_i^2 v_j}(p_i-p_j)(1-2p_i)^2, \\ G_{22} &= G_1 + \frac{1}{4B^2}\sum_{i,j}\frac{N_i w_i w_j}{N v_i^3 v_j}(p_i-p_j)(1-2p_i)^2(1-v_i) \\ &\quad -\frac{1}{4B^3}\sum_{i,j,k}\frac{N_j\sqrt{N_i N_k}w_i w_j w_k}{N^2 v_i^2 v_j v_k^2}(p_i-p_j)(1-2p_i)^2(1-2p_k)^2. \\ H &= -\frac{2}{B^3}\sum_{i,j}\frac{N_i N_j w_i^2 w_j}{N^2 v_i^2 v_j}(p_i-p_j)(1-2p_i) \\ &\quad +\frac{1}{B^4}\sum_{i,j,k}\frac{N_i N_j N_k w_i^2 w_j w_k}{N^3 v_i^3 v_j v_k}(1-2p_i)^2(p_i-p_j)(p_i-p_k), \\ B &= \sum_{i=1}^K \frac{N_i w_i}{N v_i}.\end{aligned}$$

Theorem 3.1 evaluates the behavior of $\hat{p}(x;h)$ among the covariates, and plays a key role in deriving the asymptotic properties of the estimators discussed in subsection 3.3.2. The proof of Theorem 3.1 is given in Okumura and Naito [13].

3.2.2. *Asymptotic performance.* We focus on the estimators discussed above from the viewpoint of the estimation of $p(x)$. Then, it is natural to consider a nonparametric smoothing situation such as the smoothing parameter tends to zero as the sample size grows. Indeed, no estimator including the bandwidth h is endowed with consistency without such assumptions. The necessary assumptions are as follows.

Assumption 2 $h \to 0$ as $K \to \infty$ and $N_i = N_1 \to \infty$ for $i = 2, \ldots, K$ in such a manner that $Kh^{3+\varepsilon} = O(1)$ and $N_1 h^{2-\varepsilon} = O(1)$ for some $0 < \varepsilon < 1$.

Assumption 3 The support of the kernel $\phi(x)$ is $[-1,1]$ and $\phi(x)$ has continuous and bounded derivatives of order m for any x in $[-1,1]$ with

$$(-1)^k \int_{-1}^{1} x^\ell \phi^{(k)}(x)dx = \begin{cases} 0, & \ell < k \text{ or } \ell = k+1, \\ \ell!, & \ell = k, \\ c_{\ell,k}, & \text{otherwise}, \end{cases}$$

where $0 \le k \le m$ and the $c_{\ell,k}$s are non-positive constants.

Note that the symmetric beta densities on the interval $[-1,1]$ satisfy Assumption 3 (cf. Wand and Jones [18]):

$$\phi(x) = \phi_B(x;r) = \{2^{2r+1}B(r+1,r+1)\}^{-1}(1-x^2)^r 1_{\{|x|<1\}}(x), \quad r = 0,1,\ldots,$$

where $B(\cdot,\cdot)$ is the beta function. Particular cases are the uniform, Epanechnikov, biweight and triweight kernels for $r = 0,1,2$ and 3, respectively.

Assumption 4 The curve $p(x)$ has continuous and bounded derivatives of order $m+2$ for any x in [0,1], and satisfies $0 < p(x) < 1$ for any x in $[0,1]$.

Note again that $\hat{p}_{NW}(x;h)$ is equal to $\hat{p}_M(x;h)$ under these assumptions, since $N_i = N_1$ for $i = 2,\ldots,K$. Under Assumptions 2–4, the MSE of $\hat{p}_{NW}(x;h)$ is given by

$$\mathrm{MSE}[\hat{p}_{NW}(x;h)] = \mathrm{AMSE}[\hat{p}_{NW}(x;h)] + O\left(\frac{h}{K} + h^6\right),$$

where

$$\mathrm{AMSE}[\hat{p}_{NW}(x;h)] = \frac{h^4\mu_2(\phi)^2}{4}p^{(2)}(x)^2 + \frac{v(x)R(\phi)}{N_1Kh},$$

$v(x) = p(x)(1-p(x))$, $R(\phi) = \int_{-1}^{1}\phi(z)^2dz$ and $\mu_r(\phi) = \int_{-1}^{1} z^r\phi(z)dz$. Using the Taylor expansion, we obtain the MSE expression of the ideal estimator $p^*(x;h)$ as

$$\mathrm{MSE}[p^*(x;h)] = \mathrm{AMSE}[p^*(x;h)] + O\left(\frac{h}{K} + h^6\right),$$

where

$$\mathrm{AMSE}[p^*(x;h)] = h^4\mu_2(\phi)^2 f(x)^2 + \frac{v(x)R(\phi)}{N_1Kh} \tag{3.5}$$

and

$$f(x) = \frac{1}{2}\left\{p^{(2)}(x) - \frac{2(1-2p(x))p^{(1)}(x)^2}{v(x)}\right\}.$$

Next, we focus on the proposed estimator $\hat{p}(x;h)$. We have the following.

Theorem 3.2. *Under Assumptions 2–4, we have*

$$\mathrm{MSE}[\hat{p}(x;h)] = \mathrm{AMSE}[\hat{p}(x;h)] + O\left(\frac{h}{K} + h^6\right),$$

$$\mathrm{AMSE}[\hat{p}(x;h)] \quad = \quad \mathrm{AMSE}[p^*(x;h)] - \frac{h^2}{N_1}(1-2p(x))\mu_2 f(x).$$

The proof of Theorems 3.2 is also given in Okumura and Naito [13]. The term $v(x)R(\phi)\{N_1Kh\}^{-1}$ appears in all MSE expressions above, which is the leading term of the variances of $\hat{p}_{NW}(x;h)$, $p^*(x;h)$ and $\hat{p}(x;h)$. Thus, we note that the essential difference in these estimators appears in the bias. We can immediately obtain the following.

Corollary 3.3. *If* $(1-2p(x))\left\{v(x)p^{(2)}(x)-(1-2p(x))p^{(1)}(x)^2\right\}\geq 0$ *for* $x\in[0,1]$, *then*

$$\text{MSE}[p^*(x;h)]\leq \text{MSE}[\hat{p}_{NW}(x;h)]$$

by neglecting the terms of $O(hK^{-1}+h^6)$ *and smaller.*

The proposed WNWE $\hat{p}(x;h)$ is constructed by plugging the variance estimators into $p^*(x;h)$. However, Theorem 3.2 reveals that there is a difference of $O(h^2N_1^{-1})$ between the AMSE expressions of $\hat{p}(x;h)$ and $p^*(x;h)$. We also understand that the effect of the plug-in variance estimators appears as $O(h^2N_1^{-1})$ in the sense of MSE. The order of $O(h^2N_1^{-1})$ is important since it dominates the order in the MSE of $\hat{p}(x;h)$ under Assumption 2. Here, we discuss a manipulation that makes the MSE of $\hat{p}(x;h)$ close to that of $p^*(x;h)$. We have from the calculations presented in the Appendix of Okumura and Naito [13] that

$$\text{Bias}[\hat{p}(x;h)]=h^2\mu_2 f(x)-\frac{1-2p(x)}{N_1}+o(h^2).$$

Hence, the bias-adjusted version of $\hat{p}(x;h)$ can be obtained as

$$\tilde{p}(x;h)=\frac{1}{N_1}+\left(1-\frac{2}{N_1}\right)\hat{p}(x;h).$$

Then, we have the following.

Theorem 3.4. *Under Assumptions 2–4, we have*

$$\text{MSE}[\tilde{p}(x;h)]=\text{AMSE}[p^*(x;h)]+O\left(\frac{h}{K}+h^6\right).$$

Further, we have the following asymptotic normality of $\tilde{p}(x;h)$:

Theorem 3.5. *Assume that Assumptions 2–4 hold and also that there exists a constant* $\rho\geq 0$ *such that* $N_1Kh^5\to\rho^2$. *Then,*

$$\sqrt{N_1Kh}\,\{\tilde{p}(x;h)-p(x)\}\to_D N(\rho f(x),v(x)R(\phi)).$$

It can be easily verified that $\hat{p}_{NW}(x;h)$ also has asymptotic normality and that its asymptotic variance equals $v(x)R$. The same argument for $\hat{p}_{LML}(x;h)$ can been seen in Theorem 2 of Fan et al. [5]. These facts along with Theorem 3.5 reveal that $\tilde{p}(x;h)$ has asymptotically the same precision for estimating $p(x)$ as $\hat{p}_{NW}(x;h)$ and $\hat{p}_{LML}(x;h)$, and we again note that $\hat{p}_{LML}(x;h)$ does not always exist. Hence, the use of $\tilde{p}(x;h)$ is justified since it does not have a fault on existence and it has the same asymptotic precision. Using Theorem 3.5, we obtain an asymptotic $100(1-\beta)\%$ confidence interval for $p(x)$ as

$$\left[\tilde{p}(x;h) - \Phi^{-1}(1-\frac{\beta}{2})V(x),\ \tilde{p}(x;h) + \Phi^{-1}(1-\frac{\beta}{2})V(x)\right], \tag{3.6}$$

where $V(x) = \sqrt{\tilde{p}(x;h)(1-\tilde{p}(x;h))R(\phi)/(N_1Kh)}$ and $\Phi(z)$ is the standard normal distribution function. The bias adjusted estimator $\tilde{p}(x;h)$ has a nonzero asymptotic bias $\rho f(x)\mu_2$ in Theorem 3.5 if $f(x) \neq 0$ and the bandwidth is optimally selected as $h \sim (N_1K)^{-1/5}$. However, this bias will be small for $\alpha = 0.5$ since $p^{(2)}(x)$ is close to 0, that is, $f(x)$ is close to 0, if $p(x)$ can be well approximated by symmetric and sigmoid models such as the probit models or the logistic models. On the other hand, the bias may also varnish since $\rho = 0$ for a small h. Hence, we neglect the bias term when the confidence interval is constituted as in Müller and Schmitt [10].

3.3. **Quantal Bioassay.** Okumura and Naito [12] proposed quantal bioassay using $\tilde{p}(x;h)$. We assume that $p(x)$ is strictly monotone, and here discuss the estimation for $\Theta_\alpha = p^{-1}(\alpha)(0 < \alpha < 1)$. The estimator of Θ_α is similarly defined as the way of Müller and Schmitt [10]. Put $M_\alpha = \{x \in [0,1] : \tilde{p}(x;h) = \alpha,\ \tilde{p}^{(1)}(x;h) > 0\}$, then the estimator of Θ_α is defined by $\tilde{\Theta}_\alpha = (\inf M_\alpha + \sup M_\alpha)/2$. Note that if α is increasing, then $\tilde{\Theta}_\alpha$ is also increasing. The proposed estimator $\tilde{\Theta}_\alpha$ has consistent property.

Theorem 3.6. *Let $p(x)$ be strictly monotone. Under Assumptions 2–4,*

$$\sup_{\alpha:\Theta_\alpha\in[0,1]} |\tilde{\Theta}_\alpha - \Theta_\alpha| \to 0 \quad \textit{in probability}.$$

The asymptotic normality of $\tilde{\Theta}_\alpha$ is given as follows.

Theorem 3.7. *Let $p(x)$ be strictly monotone. Assume that there exists a constant $\rho \geq 0$ such that $KN_1h^5 \to \rho^2$. Then under Assumptions 2–4, for any $\Theta_\alpha \in [h, 1-h]$,*

$$\sqrt{KN_1h}(\tilde{\Theta}_\alpha - \Theta_\alpha) \xrightarrow[d]{} N\left(\frac{\rho f(\Theta_\alpha)\mu_2}{p^{(1)}(\Theta_\alpha)}, \frac{\alpha(1-\alpha)R(\phi)}{p^{(1)}(\Theta_\alpha)^2}\right).$$

We also neglect the bias term from the same reasons in the construction of the confidence interval of $p(x)$ given by (3.6). Hence an asymptotic $100(1-\beta)\%$ confidence interval for Θ_α can be obtained as

$$\left[\tilde{\Theta}_\alpha - \frac{\Phi^{-1}(1-\frac{\beta}{2})}{\tilde{p}^{(1)}(\tilde{\Theta}_\alpha)}\sqrt{\frac{\alpha(1-\alpha)R(\phi)}{KN_1h}},\ \tilde{\Theta}_\alpha + \frac{\Phi^{-1}(1-\frac{\beta}{2})}{\tilde{p}^{(1)}(\tilde{\Theta}_\alpha)}\sqrt{\frac{\alpha(1-\alpha)R(\phi)}{KN_1h}}\right].$$

This confidence interval is incalculable for all α such that $M_\alpha = \emptyset$.

4. Bandwidth selection for WNWE

4.1. **Bandwidth selection.** From (3.5), the integrated mean squared error (IMSE) with a weight function $\kappa(x)$ of the bias-adjusted estimator $\tilde{p}(x;h)$ can be approximated under Assumptions 1–3 with $m \geq 0$ as follows:

$$\begin{aligned} \mathrm{IMSE}[\tilde{p}(\cdot;h)] &= \int E[(\tilde{p}(x;h)-p(x))^2]\kappa(x)dx \\ &\approx h^4\mu_2(\phi)^2\int\kappa(x)f(x)^2dx+\frac{R(\phi)}{N_1Kh}\int\kappa(x)v(x)dx. \end{aligned} \tag{4.1}$$

In Section 4, we employ $\kappa(x)=1_{[\delta_1,1-\delta_2]}(x)$, where $1_{[\delta_1,1-\delta_2]}(x)$ is the indicator function of the interval $[\delta_1,\delta_2]$ for some small positive constants δ_1 and δ_2. Therefore, the optimal bandwidth h_{opt} that minimizes the right hand side of (4.1) is given as

$$h_{\mathrm{opt}}=C(\phi)\left(\frac{\theta_2}{\theta_1}\right)^{1/5}(N_1K)^{-1/5}, \tag{4.2}$$

where $C(\phi)=\{R(\phi)/(4\mu_2(\phi)^2)\}^{1/5}$, $\theta_1=\int\kappa(x)f(x)^2dx$ and $\theta_2=\int\kappa(x)v(x)dx$. The unknown amounts θ_1 and θ_2 that are functionals of p must be estimated. First, in order to construct an estimator of θ_1, we adopt

$$\widehat{p^{(i)}}(x;g)=\frac{\sum_{i=1}^K Y_i\phi_g^{(i)}(x_i-x)}{\sum_{i=1}^K N_i\phi_g(x_i-x)}$$

as a simple convenient estimator of $p^{(i)}(x)$ having the bandwidth g. The NWE for $\mathcal{B}_1$ is $\widehat{p^{(0)}}(x;g)$. Then, a consistent estimator of $f(x)$ is given by

$$\bar{f}(x;g)=\frac{\widehat{p^{(2)}}(x;g)}{2}-\frac{(1-2\widehat{p^{(0)}}(x;g))\widehat{p^{(1)}}(x;g)^2}{\tau_{K^{-r}}(\widehat{p^{(0)}}(x;g))(1-\tau_{K^{-r}}(\widehat{p^{(0)}}(x;g)))},$$

where r is a positive integer and

$$\tau_c(t)=\begin{cases} c, & t\leq c, \\ t, & c<t<1-c, \\ 1-c, & t\geq 1-c. \end{cases}$$

Note that it can be easily shown to hold that $\tau_{K^{-r}}(\widehat{p^{(0)}}(x;g))-\widehat{p^{(0)}}(x;g)=o_p(K^{-r})$.

Hence, our proposed estimator of θ_1 is given by

$$\bar{\theta}_1(g) = \int\kappa(x)\bar{f}(x;g)^2dx.$$

To select the optimal bandwidth g, we make the following assumption instead of Assumption 2.

Assumption 5 $g\to 0$ as $K\to\infty$ and $N_i=N_1\to\infty$ for $i=2,\ldots,K$ in such a manner that $Kg^{2m+2+\varepsilon}=O(1)$ and $N_1g^{1-\varepsilon}=O(1)$ for some $0<\varepsilon<1$.

Then, we have the following theorem:

Theorem 4.1. *Under Assumptions 3–5 with $m = 2$, we have*

$$\mathrm{MSE}[\bar{\theta}_1(g)] \simeq \left[g^2\Delta_1(p, p^{(1)}, p^{(2)}, p^{(3)}, p^{(4)}) + \frac{\Delta_2(p)}{N_1Kg^5}\right]^2 + \frac{\Delta_3(p)}{N_1K^2g^9}, \tag{4.3}$$

where

$$\begin{aligned}
\Delta_1 &= \Delta_1(p, p^{(1)}, p^{(2)}, p^{(3)}, p^{(4)}) \\
&= \int \kappa(x)\left[\frac{\mu_4(\phi^{(2)})p^{(2)}(x)p^{(4)}(x)}{48} - \frac{\mu_2(\phi)(1-2p(x))(1-2v(x))p^{(1)}(x)^4p^{(2)}(x)}{v(x)^3}\right. \\
&\quad -\frac{2\mu_2(\phi^{(1)})(1-2p(x))^2p^{(1)}(x)^3p^{(3)}(x)}{3v(x)^2} + \frac{\mu_2(\phi)(1-2v(x))p^{(1)}(x)^2p^{(2)}(x)^2}{2v(x)^2} \\
&\quad \left.-\frac{(1-2p(x))p^{(1)}(x)}{v(x)}\left(\frac{\mu_4(\phi^{(2)})p^{(1)}(x)p^{(4)}(x)}{24} - \frac{\mu_3(\phi^{(1)})p^{(2)}(x)p^{(3)}(x)}{3}\right)\right]dx, \\
\Delta_2 &= \Delta_2(p) = \frac{R(\phi^{(2)})}{4}\int \kappa(x)v(x)dx, \\
\Delta_3 &= \Delta_3(p) = \frac{R(\phi^{(2)} * \phi^{(2)})}{4}\int \kappa(x)p(x)^2v(x)dx, \\
&\phi * \phi(x) = \int \phi(t)\phi(x-t)dt.
\end{aligned}$$

Futhermore, it holds that

$$K\sqrt{N_1g^9}\{\bar{\theta}_1(g) - \theta_1 - g^2\Delta_1 - (N_1Kg^5)^{-1}\Delta_2\} \to_d N(0, \Delta_3).$$

In the right hand side of (4.3), the order of the first term, which is the asymptotic squared bias (ASB) of $\bar{\theta}_1(g)$, is larger than that of the second term, which is the asymptotic variance. Therefore, the optimal bandwidth g_{opt} that minimizes the ASB is given by

$$g_{\mathrm{opt}} = g_{\mathrm{opt}}(\Delta_1, \Delta_2) = C_1(\Delta_1)\left(\frac{\Delta_2}{|\Delta_1|}\right)^{1/7}(N_1K)^{-1/7},$$

where

$$C_1(\Delta_1) = \begin{cases} 1 & , \quad \Delta_1 < 0, \\ (5/2)^{1/7} & , \quad \Delta_1 > 0. \end{cases}$$

We adopt

$$\bar{\theta}_2 = \frac{1}{K^*}\sum_i{}^* \frac{N_i}{N_i - 1}(\bar{Y}_i - \bar{Y}_i^2)$$

as an estimator of θ_2, where K^* is the number of x_is which belongs to $[\delta_1, 1-\delta_2]$ and $\sum_i{}^*$ designates the summation for those x_is. Then, under Assumptions 3 and 4 with $n \geq 0$, it can be shown that $\bar{\theta}_2 - \theta_2 = O_p((N_1K)^{-1/2})$. Hence, from (4.2), a selection of h_{opt} can be performed on the basis of the following:

$$\bar{h}_{\mathrm{PI}} = \bar{h}_{\mathrm{PI}}(g_{\mathrm{opt}}) = C(\phi)\left(\frac{\bar{\theta}_2}{\bar{\theta}_1(g_{\mathrm{opt}})}\right)^{1/5}(N_1K)^{-1/5}.$$

The relative error of $\bar{h}_{\rm PI}$ has the following property.

Theorem 4.2. *Under Assumptions 3–5 with $m = 2$,*

$$N_1^{-1/7}K^{5/14}\left\{\frac{\bar{h}_{\rm PI}}{h_{\rm opt}} - 1 + \frac{1}{5(N_1K)^{2/7}}\theta_1^{-1}(c^2\Delta_1 + c^{-5}\Delta_2)\right\} \to_D N\left(0, \frac{\Delta_3}{25c^9\theta_1^2}\right),$$

where $c = (N_1K)^{1/7}g_{\rm opt} = C_1(\Delta_1)(\Delta_2/|\Delta_1|)^{1/7}$.

From Theorem 4.2, we immediately have

$$\frac{\bar{h}_{\rm PI}}{h_{\rm opt}} - 1 = O_P\left(\frac{1}{(N_1K)^{2/7}}\right).$$

4.2. **Implementation.** The subsection briefly illustrates the proposed practical algoritheorem to select the bandwidth from the data. Our devices in the algoritheorem include a scale adjustment pointed out by Yang and Tschernig [19] and the GSK approach discussed in Grizzle et al. [6] for the ROT step.

4.2.1. *Scale adjustment.* As pointed out in Yang and Tschernig [19], the bandwidth that minimizes an ASB as described in Section 4.1 is smaller than the optimal bandwidth that minimizes an asymptotic MSE (AMSE) in finite samples. Hence, they recommended using a bandwidth slightly larger than the bandwidth that minimizes the ASB.

We describe an algoritheorem to ditermine the constant $\rho^* > 1$ such that $g_{\rm opt}^\dagger = \rho^* g_{\rm opt}$, where $g_{\rm opt}^\dagger$ is the optimal bandwidth minimizing the AMSE of $\bar{\theta}_1(g)$ which is the right hand side of (4.3). Denote the AMSE$[\bar{\theta}_1(\rho g)]$ as $\alpha(\rho : g, \Delta_1, \Delta_2, \Delta_3)$, which can be regarded as a function of ρ. We apply the Newton–Raphson method to the derivation of the solution minimizing $\alpha(\rho : g_{\rm opt}, \Delta_1, \Delta_2, \Delta_3)$, where the existence of the minimizer is guaranteed from the form of the function. The m-th approximation of ρ is given by

$$\rho^{[m]} = \rho^{[m-1]} - \frac{\alpha'(\rho^{[m-1]} : g_{\rm opt}, \Delta_1, \Delta_2, \Delta_3)}{\alpha''(\rho^{[m-1]} : g_{\rm opt}, \Delta_1, \Delta_2, \Delta_3)},$$

where the derivatives are calculated with respect to ρ and $\rho^{[0]} = 1$. From the fundamental result of the Newton–Raphson method, it follows that $\rho^* = \lim_{m\to\infty} \rho^{[m]}$.

In practice, we adopt $\hat{g}_{\rm opt}^\dagger = \hat{\rho}^*\hat{g}_{\rm ROT}$, where $\hat{g}_{\rm ROT}$ and $\hat{\Delta}_i (i = 1, 2, 3)$ are quantities given by using a ROT method in the sequent discussion and $\hat{\rho}^*$ is the minimizer of $\alpha(\rho : \hat{g}_{\rm ROT}, \hat{\Delta}_1, \hat{\Delta}_2, \hat{\Delta}_3)$. Finally, we have the data-driven bandwidth

$$\tilde{h}_{\rm PI}^\dagger = \bar{h}_{\rm PI}(\hat{g}_{\rm opt}^\dagger).$$

4.2.2. *GSK approach for ROT.* The quantities $g_{\rm opt}$ and $\Delta_i (i = 1, 2, 3)$ are determined by the ROT method using certain parametric estimation. In the parametric estimation for $p(x)$, it will be desirable that estimators of $p(x)$ take values in $(0, 1)$ for any x and can be derived without iterative calculations. We present here a method of ROT endowed with such desirable properties. The method we employ is called the GSK approach given in Grizzle et al. [6], which is based on a generalized least squares method. The method is as follows. We consider the following polynomial logistic regression model: $p(x;\boldsymbol{\beta}) = G(\mathbf{z}^T\boldsymbol{\beta})$, where $\mathbf{z} = (1, x, \ldots, x^r)^T$ and unknown parameter $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_r)^T$. The estimator $\hat{\boldsymbol{\beta}}$ is obtained by minimizing the weighted sum of squared errors $\sum_{i=1}^K N_i\hat{v}_i(G^{-1}(\hat{p}_i) - \mathbf{z}_i^T\boldsymbol{\beta})^2$, where $\hat{p}_i$s are the Bayes estimators given in Section 3.1 and $\mathbf{x}_i = (1, x_i, \ldots, x_i^r)^T$. We can explicitly express $\hat{\boldsymbol{\beta}}$ as $\hat{\boldsymbol{\beta}} = (X^T\hat{\Omega}^{-1}X)X^T\hat{\Omega}^{-1}\boldsymbol{\eta}$, where $Z = (\mathbf{z}_1, \ldots, \mathbf{z}_K)^T$, $\hat{\Omega}^{-1} = \mathrm{diag}[N_1\hat{v}_1, \ldots, N_K\hat{v}_K]$ and $\boldsymbol{\eta} = (G^{-1}(\hat{p}_1), \ldots, G^{-1}(\hat{p}_K))^T$. Note that the empirical logit $G^{-1}(\bar{Y}_i)$ is asymptotically normally distributed with mean $\mathbf{z}_i^T\boldsymbol{\beta}$ and variance $(N_iv_i)^{-1}$ and is not well defined if $\bar{Y}_i$ is 0 or 1 in practice. The ROT estimator of $p(x)$ and its ith derivative $p^{(i)}(x)$ are obtained by $p(x;\hat{\boldsymbol{\beta}})$ and $p^{(i)}(x;\hat{\boldsymbol{\beta}})$, respectively. We can obtain the estimators $\hat{\Delta}_i (i = 1, 2, 3)$ of $\Delta_i (i = 1, 2, 3)$, respectively, by substituting the ROT estimators defined above to the corresponding parts in the definitions of $\Delta_i (i = 1, 2, 3)$. Hence, we have $\hat{g}_{\mathrm{ROT}} = g_{\mathrm{opt}}(\hat{\Delta}_1, \hat{\Delta}_2)$.

4.2.3. *Fully ROT method.* As a simpler alternative method, the ROT method for h_{opt} can be directly applied. This bandwidth selector is written as

$$h_{\mathrm{ROT}} = C(\phi)\left(\frac{\int \kappa(x)\hat{v}(x;\hat{\beta})dx}{\int \kappa(x)\hat{f}(x;\hat{\beta})^2dx}\right)^{1/7}(N_1K)^{-1/7},$$

where $\hat{v}(x;\hat{\beta})$ and $\hat{f}(x;\hat{\beta})$ are the ROT estimators of $v(x)$ and $f(x)$, respectively. Note that h_{ROT} is not consistent in general.

5. Nonparametric binary regression

5.1. **Kernel estimators.** We consider the binary regression problem for the binary data $\mathcal{B}_4$. Let us introduce the standard kernel-based estimators of the regression function $p(\mathbf{x})$. Set $w_i = K_{\mathbf{h}}(\mathbf{X}_i - \mathbf{x}) = \Pi_{j=1}^d h_j^{-1}\phi(h_j^{-1}(X_{ij} - x_j))$, where $\mathbf{h} = (h_1, \ldots, h_d)$. Then, the NWE can be expressed as

$$\hat{p}_{NW}(\mathbf{x};\mathbf{h}) = \frac{\sum_{i=1}^n K_{\mathbf{h}}(\mathbf{X}_i - \mathbf{x})Y_i}{\sum_{i=1}^n K_{\mathbf{h}}(\mathbf{X}_i - \mathbf{x})}.$$

Table 1. Asymptotic conditional biases of three kernel-based estimators.

Estimator	Asymptotic bias
$\hat{p}_{NW}(\mathbf{x};\mathbf{h})$	$\frac{1}{2}\mu_2(\phi)\sum_{j=1}^d h_j^2\left\{p_{jj}(\mathbf{x}) + 2\frac{p_j(\mathbf{x})\varphi_j(\mathbf{x})}{\varphi(\mathbf{x})}\right\}$
$\breve{p}_{LLT}(\mathbf{x};\mathbf{h})$	$\frac{1}{2}\mu_2(\phi)\sum_{j=1}^d h_j^2 p_{jj}(\mathbf{x})$
$\hat{p}_{LLL}(\mathbf{x};\mathbf{h})$	$\frac{1}{2}\mu_2(\phi)\sum_{j=1}^d h_j^2\left\{p_{jj}(\mathbf{x}) - \frac{(1-2p(\mathbf{x}))p_j(\mathbf{x})^2}{p(\mathbf{x})(1-p(\mathbf{x}))}\right\}$

In order to derive local linear type estimators (LLTEs), we employ the following criterion:

$$q_\omega(\boldsymbol{\beta};\mathbf{x}) = \sum_{i=1}^{n} \omega(\mathbf{X}_i) K_{\mathbf{h}}(\mathbf{X}_i - \mathbf{x})(Y_i - \boldsymbol{\beta}^T \mathbf{Z}_i)^2, \tag{5.1}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_d)^T$ and $\omega(\mathbf{t})$ is a weight function of $\mathbf{t}$. Let LLTE be defined as β_0 of $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_d)^T$ that minimizes $q_\omega(\boldsymbol{\beta};\mathbf{x})$. LLTE can be expressed as the following matrix formula:

$$\bar{p}_{LLT}(\mathbf{x};\mathbf{h}) = \mathbf{e}_{1(d+1)}^T (\mathbf{X}_{\mathbf{x}}^T \mathbf{W}_{\mathbf{x}} \mathbf{X}_{\mathbf{x}})^{-1} \mathbf{X}_{\mathbf{x}}^T \mathbf{W}_{\mathbf{x}} \mathbf{Y},$$

where $\mathbf{W}_{\mathbf{x}} = \Omega \mathbf{K}_{\mathbf{x}}$, in which $\Omega = \mathrm{diag}(\omega(\mathbf{X}_1), \ldots, \omega(\mathbf{X}_n))$ and $\mathbf{K}_{\mathbf{x}} = \mathrm{diag}(K_{\mathbf{h}}(\mathbf{X}_1 - \mathbf{x}), \ldots, K_{\mathbf{h}}(\mathbf{X}_n - \mathbf{x}))$. Note that $\bar{p}_{LLT}(\mathbf{x};\mathbf{h})$ depends on ω. The truncated version of LLTE proposed by Aragaki and Altman [1] can be expressed as $\breve{p}_{LLT}(\mathbf{x};\mathbf{h}) = \tau_0(\bar{p}_{LLT}(\mathbf{x};\mathbf{h}))$. In particular, OLLE can be defined by $\breve{p}_{LLT}(\mathbf{x};\mathbf{h})$ with $\omega(\mathbf{t}) = 1$ for any $\mathbf{t}$, which is expressed as $\hat{p}_{OLL}(\mathbf{x};\mathbf{h}) = \tau_0(\bar{p}_{OLL}(\mathbf{x};\mathbf{h}))$, where $\bar{p}_{OLL}(\mathbf{x};\mathbf{h}) = \mathbf{e}_{1(d+1)}^T (\mathbf{X}_{\mathbf{x}}^T \mathbf{K}_{\mathbf{x}} \mathbf{X}_{\mathbf{x}})^{-1} \mathbf{X}_{\mathbf{x}}^T \mathbf{K}_{\mathbf{x}} \mathbf{Y}$.

The LLLE of $p(\mathbf{x})$ is defined by $\hat{p}_{LLL}(\mathbf{x};\mathbf{h}) = G(\mathbf{e}_{1(d+1)}^T \hat{\boldsymbol{\beta}})$ if the following solution exists uniquely:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} K_{\mathbf{h}}(\mathbf{X}_i - \mathbf{x}) \left\{ Y_i \log(G(\boldsymbol{\beta}^T \mathbf{Z}_i)) + (1 - Y_i) \log(1 - G(\boldsymbol{\beta}^T \mathbf{Z}_i)) \right\}^2.$$

The following asymptotic properties of $\hat{p}_{NW}(\mathbf{x};\mathbf{h}), \breve{p}_{LLT}(\mathbf{x};\mathbf{h})$ and $\hat{p}_{LLL}(\mathbf{x};\mathbf{h})$ can be obtained through the standard calculations used in kernel smoothing. The asymptotic conditional variances of the three estimators given $\dot{\mathbf{X}} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ are equal and are given as

$$\frac{R(\phi)^d}{nh_1 \cdots h_d} \frac{v(\mathbf{x})}{\varphi(\mathbf{x})}, \tag{5.2}$$

where $v(\mathbf{x}) = V[Y_i|\mathbf{X}_i = \mathbf{x}] = p(\mathbf{x})(1 - p(\mathbf{x}))$. The asymptotic conditional biases of the three estimators given $\dot{\mathbf{X}}$ are shown in Table 1, where $p_j(\mathbf{x})$ and $p_{jj}(\mathbf{x})$ denote the first and second order partial derivatives of $p(\mathbf{x})$ with respect to the jth variable x_j, respectively. The asymptotic conditional biases of LLTEs given $\dot{\mathbf{X}}$ depending on ω obtained using (5.1) are found to be the same. If $\varphi(\mathbf{x})$ is not the uniform density, then the bias of NWE only includes functionals of φ. The bias of LLLE is more complex than that of LLTEs. If $p(\mathbf{x})$ is a logistic linear model, then the asymptotic bias of LLLE becomes 0.

5.2. **Weighted local linear estimators.** We will focus on LLTE because it has several advantages. In ordinary parametric linear regression with heteroskedasticity, the advantage of weighting by inverse variance is given by the Gauss-Markov theorem. To improve the inverse variance weighting in local linear regression, we consider $\breve{p}_{LLT}(\mathbf{x};\mathbf{h})$ weighted through $\omega(\mathbf{x}) = \omega_\alpha(\mathbf{x})$, where

$$\omega_\alpha(\mathbf{t}) = \eta_\alpha(p(\mathbf{t})) = \{p(\mathbf{t})^\alpha (1 - p(\mathbf{t}))^{1-\alpha}\}^{-2}$$

for some $\alpha (0 < \alpha < 1)$, which we refer to as the weighted local linear estimator (WLLE) for $100\alpha\%$. The corresponding estimator is given by $\breve{p}_{\omega_\alpha}(\mathbf{x};\mathbf{h})$. If $\alpha = 0.5$, then $\omega(\mathbf{t}) = v(\mathbf{t})^{-1}$. The difference between the variances of $\hat{p}_{OLL}(\mathbf{x};\mathbf{h})$ and $\breve{p}_{\omega_\alpha}(\mathbf{x};\mathbf{h})$ is asymptotically given by the following theorem.

Theorem 5.1. *Under Assumptions 6–8 in Appendix, we have*

$$V[\hat{p}_{OLL}(\mathbf{x};\mathbf{h})|\dot{\mathbf{X}}] = V[\breve{p}_{\omega_\alpha}(\mathbf{x};\mathbf{h})|\dot{\mathbf{X}}] + \frac{\nu(\phi)}{nh_1\cdots h_d}\sum_{j=1}^{d} h_j^2 \delta_{\alpha j}(\mathbf{x}) + o_p\left(\frac{\|\mathbf{h}\|^2}{nh_1\cdots h_d}\right),$$

where $\|\mathbf{h}\| = \sqrt{\mathbf{h}^T\mathbf{h}}$, $\nu(\phi) = R(\phi)^{d-1}\{\mu_2(\phi^2) - R(\phi)\mu_2(\phi)\}$ *and*

$$\begin{aligned}\delta_{\alpha j}(\mathbf{x}) &= \frac{2}{\varphi(\mathbf{x})^2 p(\mathbf{x})^{2(1-\alpha)}(1-p(\mathbf{x}))^{2\alpha}} \\ &\quad \times \left[(\alpha - p(\mathbf{x}))p(\mathbf{x})(1-p(\mathbf{x}))\left\{-2\varphi_j(\mathbf{x})p_j(\mathbf{x}) + \varphi(\mathbf{x})p_{jj}(\mathbf{x})\right\}\right. \\ &\quad\quad \left. + \left\{-\alpha(1-\alpha) + (\alpha - p(\mathbf{x}))^2\right\}\varphi(\mathbf{x})p_j(\mathbf{x})^2\right].\end{aligned}$$

It can be shown that $\mu_2(\phi^2) - R(\phi)\mu_2(\phi) < 0$ when $\phi(x) = \phi_B(x,r) \quad r = 1,2,\ldots$. Hence, $\nu(\phi) < 0$. Let $A_\alpha = \bigcap_{j=1}^{d}\{\mathbf{x} : \delta_{\alpha j}(\mathbf{x}) \le 0\}$. For $\mathbf{x} \in A_\alpha$, it can be expected that the variance of $\breve{p}_{\omega_\alpha}(\mathbf{x};\mathbf{h})$ is smaller than that of $\hat{p}_{OLL}(\mathbf{x};\mathbf{h})$. At least, for $\mathbf{x}$ with $\alpha = p(\mathbf{x})$, $\delta_{\alpha j}(\mathbf{x}) \le 0$ for any j.

Moreover, if φ is uniform on $[0,1]^d$ and the value of $p(\mathbf{x})$ is close to α and $p_{jj}(\mathbf{x})$ is rather small, that is, $p(\mathbf{x})$ is rather flat, $\breve{p}_{\omega_\alpha}(\mathbf{x};\mathbf{h})$ can be superior to $\hat{p}_{OLL}(\mathbf{x};\mathbf{h})$ in terms of variance.

Furthermore, if $\phi(x) = \phi_B(x;0)$, then $\mu_2(\phi^2) - R(\phi)\mu_2(\phi) = 0$. Note that $\phi_B(x;0)$ is a uniform kernel. In addition, if $\alpha = 0.5$, then we obtain the following exact result:

$$V[\hat{p}_{OLL}(\mathbf{x};\mathbf{h})|\dot{\mathbf{X}}] \ge V[\breve{p}_{\omega_{0.5}}(\mathbf{x};\mathbf{h})|\dot{\mathbf{X}}].$$

In practice, $\breve{p}_{\omega_\alpha}(\mathbf{x};\mathbf{h})$ requires the estimation of $\Omega = \Omega_\alpha$, where $\Omega_\alpha = \mathrm{diag}(\omega_\alpha(\mathbf{X}_1), \ldots, \omega_\alpha(\mathbf{X}_1))$. Let $\hat{p}_{LLT}(\mathbf{x};\mathbf{h})$ denote $\breve{p}_{LLT}(\mathbf{x};\mathbf{h})$ in which Ω is replaced by $\hat{\Omega}$, where $\hat{\Omega} = \mathrm{diag}(\hat{\omega}(\mathbf{X}_1), \ldots, \hat{\omega}(\mathbf{X}_n))$ is an estimator of Ω. Then, we obtain the conditional AMSE of $\hat{p}_{LLT}(\mathbf{x};\mathbf{h})$ given $\dot{\mathbf{X}}$ in general as follows.

Theorem 5.2. *Under Assumptions 6–9 in Appendix, if* $E[n^{-1}\mathrm{tr}\{\mathbf{K}_{\mathbf{x}}(\hat{\Omega}-\Omega)^2\}|\dot{\mathbf{X}}] = O_p(||\mathbf{h}||^4)$, *we have*

$$\mathrm{AMSE}[\hat{p}_{LLT}(\mathbf{x};\mathbf{h})|\dot{\mathbf{X}}] = \frac{\mu_2(\phi)^2}{4}\left\{\sum_{j=1}^{d} h_j^2 p_{jj}(\mathbf{x})\right\}^2 + \frac{R(\phi)^d}{nh_1\cdots h_d}\frac{v(\mathbf{x})}{\varphi(\mathbf{x})}.$$

Let $\hat{\Omega}_\alpha = \mathrm{diag}(\hat{\omega}_\alpha(\mathbf{X}_1), \ldots, \hat{\omega}_\alpha(\mathbf{X}_n))$, where $\hat{\omega}_\alpha(\mathbf{X}_i) = \eta_\alpha(\tau_{n^{-1}}(\hat{p}_{OLL}(\mathbf{X}_i;\mathbf{h})))$. Then, it follows that $E[n^{-1}\mathrm{tr}\{\mathbf{K}_{\mathbf{x}}(\hat{\Omega}_\alpha - \Omega_\alpha)^2\}|\dot{\mathbf{X}}] = O_p(||\mathbf{h}||^4)$. Practical WLLE for $100\alpha\%$ can be expressed as $\hat{p}_{\omega_\alpha}(\mathbf{x};\mathbf{h}) = \tau_0(\bar{p}_{\omega_\alpha}(\mathbf{x};\mathbf{h}))$, where

$$\bar{p}_{\omega_\alpha}(\mathbf{x};\mathbf{h}) = \mathbf{e}_1^T(\mathbf{X}_{\mathbf{x}}^T\hat{\mathbf{W}}_{\mathbf{x}}\mathbf{X}_{\mathbf{x}})^{-1}\mathbf{X}_{\mathbf{x}}^T\hat{\mathbf{W}}_{\mathbf{x}}\mathbf{Y}$$

and $\hat{\mathbf{W}}_{\mathbf{x}} = \hat{\Omega}_\alpha \mathbf{K}_{\mathbf{x}}$. On the other hand, assume that $\tilde{p}(\mathbf{x})$ is a parametric estimator of $p(\mathbf{x})$ and $\tilde{p}^*(\mathbf{x})$ is the best parametric estimator of $\tilde{p}(\mathbf{x})$. Let $\tilde{\Omega}_\alpha = \mathrm{diag}(\tilde{\omega}_\alpha(\mathbf{X}_1), \ldots .\tilde{\omega}_\alpha(\mathbf{X}_n))$ and $\tilde{\Omega}^*_\alpha = \mathrm{diag}(\tilde{\omega}^*_\alpha(\mathbf{X}_1), \ldots .\tilde{\omega}^*_\alpha(\mathbf{X}_n))$, where $\tilde{\omega}_\alpha(\mathbf{X}_i) = \eta_\alpha(\tilde{p}(\mathbf{X}_i))$ and $\tilde{\omega}^*_\alpha(\mathbf{X}_i) = \eta_\alpha(\tilde{p}^*(\mathbf{X}_i))$. If the true function $p(\mathbf{x})$ is not included in a family of parametric models, then $E[n^{-1}\mathrm{tr}\{\mathbf{K}_{\mathbf{x}}(\tilde{\Omega}_\alpha - \Omega_\alpha)^2\}|\dot{\mathbf{X}}]$ does not converge to 0 in general but it will follow that $E[n^{-1}\mathrm{tr}\{\mathbf{K}_{\mathbf{x}}(\tilde{\Omega}_\alpha - \tilde{\Omega}^*_\alpha)^2\}|\dot{\mathbf{X}}] = O_p(n^{-1}) = o_p(||\mathbf{h}||^4)$ under certain regularity conditions. Hence, if the parametric fitting is possible, the parametric estimator $\tilde{\Omega}_\alpha$ can be used as an alternative for $\hat{\Omega}_\alpha$. We obtain an alternative estimator for $\hat{p}_{\omega_\alpha}(\mathbf{x};\mathbf{h})$ by replacing $\hat{\Omega}_\alpha$ with $\tilde{\Omega}_\alpha$, and denote it as $\breve{p}_{\omega_\alpha}(\mathbf{x};\mathbf{h})$. Moreover, we adopt $\hat{p}_\omega(\mathbf{x};\mathbf{h})$ and $\tilde{p}_\omega(\mathbf{x};\mathbf{h})$ as the estimators of $p(\mathbf{x})$ for any $\mathbf{x}$, where $\hat{p}_\omega(\mathbf{x};\mathbf{h})$ is $\hat{p}_{\omega_\alpha}(\mathbf{x};\mathbf{h})$ with $\alpha = \tau_{n^{-1}}(\hat{p}_{OLL}(\mathbf{x};\mathbf{h}))$ and $\tilde{p}_\omega(\mathbf{x};\mathbf{h})$ is $\tilde{p}_{\omega_\alpha}(\mathbf{x};\mathbf{h})$ with $\alpha = \tilde{p}(\mathbf{x})$. It can also be shown that the condition of Theorem 5.2 is satisfied for $\hat{p}_\omega(\mathbf{x};\mathbf{h})$ and $\tilde{p}_\omega(\mathbf{x};\mathbf{h})$.

6. Bandwidth selection for WLLEs

6.1. **Preliminaries.** Using a known weight function $\kappa(\mathbf{x})$, the weighted IMSE of $\hat{p}_{LLT}(\cdot;\mathbf{h})$ is defined as

$$\mathrm{IMSE}[\hat{p}_{LLT}(\cdot;\mathbf{h})] = E[\int \{(\hat{p}_{LLT}(\mathbf{x};\mathbf{h}) - p(\mathbf{x})\}^2 \kappa(\mathbf{x})\varphi(\mathbf{x})d\mathbf{x}].$$

From the result of Theorem 5.2, the asymptotic IMSE of $\hat{p}_{LLT}(\cdot;\mathbf{h})$ is given as

$$\mathrm{AIMSE}[\hat{p}_{LLT}(\cdot;\mathbf{h})] = \frac{\mu_2(\phi)^2}{4}\sum_{\alpha=1}^{d}\sum_{\lambda=1}^{d} h_\alpha^2 h_\lambda^2 C_{\alpha\lambda} + \frac{R(\phi)^d}{nh_1\cdots h_d}B, \tag{6.1}$$

where $C_{\alpha\lambda} = \int \kappa(\mathbf{x})p_{\alpha\alpha}(\mathbf{x})p_{\lambda\lambda}(\mathbf{x})d\mathbf{x}$ and $B = \int \kappa(\mathbf{x})v(\mathbf{x})d\mathbf{x}$. Under some regularity assumptions, given in the Appendix of Yang and Tschernig [19], it is confirmed that the optimal bandwidth exists uniquely. In such a case, the solution can be calculated numerically. We will define the optimal bandwidth as follows:

$$\mathbf{h}_{\mathrm{opt}} = \arg\min_{\mathbf{h}} \mathrm{AIMSE}[\hat{p}_{LLT}(\cdot;\mathbf{h})].$$

We will present the ROT and PI methods for bandwidth selection using (6.1), which are constructed as in Yang and Tschernig [19]. In practice, $C_{\alpha\lambda}$ and B are to be estimated. For each $\alpha, \lambda = 1, \ldots, d$, let

$$\begin{aligned}\bar{C}_{\alpha\lambda} &= \bar{C}_{\alpha\lambda}(p_{\alpha\alpha}, p_{\lambda\lambda}) = \frac{1}{n}\sum_{i=1}^{n} p_{\alpha\alpha}(\mathbf{X}_i)p_{\lambda\lambda}(\mathbf{X}_i)\kappa(\mathbf{X}_i),\\ \bar{B} &= \bar{B}(p,\varphi) = \frac{1}{n}\sum_{i=1}^{n}\frac{(Y_i - p(\mathbf{X}_i))^2\kappa(\mathbf{X}_i)}{\varphi(\mathbf{X}_i)}\end{aligned}$$

and

$$AIMSE[\mathbf{h}; \{\bar{C}_{\alpha\lambda}(p_{\alpha\alpha}, p_{\lambda\lambda})\}, \bar{B}(p,\varphi)] = \frac{\mu_2(\phi)^2}{4}\sum_{\alpha=1}^{d}\sum_{\lambda=1}^{d} h_\alpha^2 h_\lambda^2 \bar{C}_{\alpha\lambda} + \frac{R(\phi)^d}{nh_1\cdots h_d}\bar{B}.$$

To derive the ROT and PI bandwidth selectors, $\bar{C}_{\alpha\lambda}$ and $\bar{B}$ are estimated in parametric and nonparametric approaches, respectively.

6.2. **ROT method.** To estimate $p(\mathbf{x})$ and its higher order partial derivatives, Yang and Tschernig [19] have proposed blocked polynomial fitting. A polynomial model is fitted to each block constructed by separating the covariate space. Mallows's C_p is used to select the optimal separation for the various separated blocks. Since $p(\mathbf{x})$ is known to be bounded on $(0,1)$, we will carry out the blocked fitting of the following logistic quadratic model:

$$p(\mathbf{x};\boldsymbol{\gamma}) = G\left(\gamma_0 + \sum_{j=1}^{d}\sum_{k=1}^{2}\gamma_{2(j-1)+k}x_j^k\right).$$

The maximum likelihood estimation of the model in each block is required. We will select the optimal separation of blocks by the Bayesian information criterion. The maximum likelihood estimator (MLE) of $p(\mathbf{x})$ obtained by this approach is denoted as $\tilde{p}(\mathbf{x})$. As an estimator of $\varphi(\mathbf{x})$, we adopt the uniform density function, which is denoted as $\dot{\varphi}(\mathbf{x})$. Therefore, the ROT bandwidth of $\mathbf{h}$ is given by

$$\hat{\mathbf{h}}_{\mathrm{ROT}} = \arg\min_{\mathbf{h}} \mathrm{AIMSE}[\mathbf{h}; \{\bar{C}_{\alpha\lambda}(\tilde{p}_{\alpha\alpha}, \tilde{p}_{\lambda\lambda})\}, \bar{B}(\tilde{p}, \dot{\varphi})].$$

6.3. **PI method.** We will use the estimator with a scalar bandwidth proposed by Yang and Tschernig [19] as an estimator of $p_{\alpha\alpha}(\mathbf{x})$. The estimator of $p_{\alpha\alpha}(\mathbf{x})$ can be expressed as $\hat{p}_{\alpha\alpha}(\mathbf{x};g) = 2\mathbf{e}^T_{(\alpha+1)(5d-1)}(\mathbf{X}_\alpha^T\Xi\mathbf{X}_\alpha)^{-1}\mathbf{X}_\alpha^T\Xi\mathbf{Y}$, where $\mathbf{g} = (g,...,g)$, $\Xi = \mathrm{diag}(K_{\mathbf{g}}(\mathbf{X}_1 - \mathbf{x}),\ldots,K_{\mathbf{g}}(\mathbf{X}_n - \mathbf{x}))$, $\mathbf{X}_\alpha$ is the $n\times(5d-1)$ matrix defined by

$$\begin{aligned}\mathbf{X}_\alpha &= [1, \{(X_{ij}-x_j)^2\}_{1\le j\le d}, \{(X_{ij}-x_j)(X_{i\alpha}-x_\alpha)\}_{j\neq\alpha}, \{(X_{ij}-x_j)\}_{1\le j\le d},\\ &\quad \{(X_{ij}-x_j)(X_{i\alpha}-x_\alpha)^2\}_{j\neq\alpha}, \{(X_{ij}-x_j)^2(X_{i\alpha}-x_\alpha)\}_{j\neq\alpha}, (X_{i\alpha}-x_\alpha)^3]_{i=1}^n\end{aligned}$$

and $[(a_{i1},\ldots,a_{im})]_{i=1}^n = (a_{ij})_{n\times m}$. Put $\hat{C}_{\alpha\lambda}(g_{\alpha\lambda}) = n^{-1}\sum_{i=1}^n \hat{p}_{\alpha\alpha}(\mathbf{X}_i;g_{\alpha\lambda})\hat{p}_{\lambda\lambda}(\mathbf{X}_i;g_{\alpha\lambda})$ for each $\alpha,\lambda = 1,\ldots,d$. Then, from the result of Yang and Tschernig [19], an expression of the asymptotic MSE of $\hat{C}_{\alpha\lambda}(g_{\alpha\lambda})$ can be expressed as

$$\mathrm{AMSE}[\hat{C}_{\alpha\lambda}(g_{\alpha\lambda})] = \left\{g_{\alpha\lambda}D_{\alpha\lambda} + \frac{BV_{\alpha\lambda}}{ng_{\alpha\lambda}^{d+4}}\right\}^2 + \frac{\sigma^2_{\alpha\lambda}}{n^2g^{d+2}}, \tag{6.2}$$

in which

$$\begin{aligned}D_{\alpha\lambda} &= \int\{p_{\alpha\alpha}(\mathbf{x})b_{\lambda\lambda}(\mathbf{x}) + p_{\lambda\lambda}(\mathbf{x})b_{\alpha\alpha}(\mathbf{x})\}\varphi(\mathbf{x})\kappa(\mathbf{x})d\mathbf{x},\\ V_{\alpha\lambda} &= \frac{4R(\phi)^{d-2}}{(\mu_4(\phi)-\mu_2(\phi)^2)^2}\{\delta_{\alpha\lambda}\mu_4(\phi^2)R(\phi) + (1-\delta_{\alpha\lambda})\mu_2(\phi^2)^2\\ &\quad -2\mu_2(\phi)\mu_2(\phi^2)R(\phi) + \mu_2(\phi)^2R(\phi)^2\},\\ \sigma^2_{\alpha\lambda} &= c_{\alpha\lambda}(\phi)\int p(\mathbf{x})^2(1-p(\mathbf{x}))^2\kappa(\mathbf{x})^2d\mathbf{x},\end{aligned}$$

where $\delta_{\alpha\lambda} = 1$ if $\alpha = \lambda$, $\delta_{\alpha\lambda} = 0$ if $\alpha \neq \lambda$,

$$\begin{aligned} b_{\alpha\alpha}(\mathbf{x}) &= \frac{\mu_6(\phi) - \mu_2(\phi)\mu_4(\phi)}{12(\mu_4(\phi) - \mu_2(\phi)^2)} p_{\alpha\alpha\alpha\alpha}(\mathbf{x}) + \frac{\mu_2(\phi)}{2} \sum_{j \neq \alpha} p_{jj\alpha\alpha}(\mathbf{x}), \\ c_{\alpha\lambda}(\phi) &= \frac{32}{(\mu_4(\phi) - \mu_2(\phi)^2)^4} \int F_{\alpha\alpha}(\mathbf{x}) F_{\lambda\lambda}(\mathbf{x}) d\mathbf{x}, \\ F_{\alpha\lambda}(\mathbf{x}) &= (1 - \delta_{\alpha\lambda})(\phi^* * \phi)(x_\alpha)(\phi^* * \phi)(x_\lambda)\Pi_{j \neq \alpha,\lambda}\phi^{(2)}(x_j) \\ &\quad + \delta_{\alpha\lambda}\phi^{*(2)}(x_\alpha)\Pi_{j\neq\alpha}\phi^{(2)}(x_j), \\ \phi^{(2)}(t) &= \int \phi(u - t)\phi(u) du \quad \text{and} \quad \phi^*(u) = \phi(u)(u^2 - \mu_2(\phi)). \end{aligned}$$

Note that the first and second terms in the right hand side of (6.2) are the leading terms of the ASB and the asymptotic variance of $\hat{C}_{\alpha\lambda}(g_{\alpha\lambda})$, respectively. We substitute

$$\bar{D}_{\alpha\lambda} = \frac{1}{n} \sum_{i=1}^{n} \{p_{\alpha\alpha}(\mathbf{X}_i) b_{\lambda\lambda}(\mathbf{X}_i) + p_{\lambda\lambda}(\mathbf{X}_i) b_{\alpha\alpha}(\mathbf{X}_i)\} \kappa(\mathbf{X}_i)$$

and

$$\bar{\sigma}^2_{\alpha\lambda} = \frac{1}{n} \sum_{i=1}^{n} \frac{p(\mathbf{X}_i)^2 (1 - p(\mathbf{X}_i))^2 \kappa(\mathbf{X}_i)^2}{\varphi(\mathbf{X}_i)}$$

for $D_{\alpha\lambda}$ and $\sigma_{\alpha\lambda}$ in (6.2), respectively. Instead of (6.2), we will consider

$$\text{AMSE}[g_{\alpha\lambda}; p, p_{\alpha\alpha}, p_{\lambda\lambda}, \{p_{jj\alpha\alpha}, p_{jj\lambda\lambda}\}, \varphi = \left\{ g_{\alpha\lambda}\bar{D}_{\alpha\lambda} + \frac{\bar{B}V_{\alpha\lambda}}{n g_{\alpha\lambda}^{d+4}} \right\}^2 + \frac{\bar{\sigma}^2_{\alpha\lambda}}{n^2 g_{\alpha\lambda}^{d+2}}. \tag{6.3}$$

Yang and Tschernig [19] adopted for each $\alpha, \lambda = 1, \ldots, d$,

$$\hat{g}_{\alpha\lambda} = \text{argmin}_{g_{\alpha\lambda}} \text{ASB}[g_{\alpha\lambda}; \tilde{p}, \tilde{p}_{\alpha\alpha}, \tilde{p}_{\lambda\lambda}, \{\tilde{p}_{jj\alpha\alpha}, \tilde{p}_{jj\lambda\lambda}\}, \dot{\varphi}],$$

where

$$\text{ASB}[g_{\alpha\lambda}; p, p_{\alpha\alpha}, p_{\lambda\lambda}, \{p_{jj\alpha\alpha}, p_{jj\lambda\lambda}\}, \varphi] = \left\{ g_{\alpha\lambda}\bar{D}_{\alpha\lambda} + \frac{\bar{B}V_{\alpha\lambda}}{n g_{\alpha\lambda}^{d+4}} \right\}^2.$$

We obtain $\hat{C}_{\alpha\lambda}(\hat{g}_{\alpha\lambda})$ as an estimator of $\bar{C}_{\alpha\lambda}$. Therefore, the PI bandwidth selector of $\mathbf{h}$ can be obtained as follows:

$$\hat{\mathbf{h}}_{\text{PI}} = \arg\min_{\mathbf{h}} \text{AIMSE}[\mathbf{h}; \{\hat{C}_{\alpha\lambda}(\hat{g}_{\alpha\lambda})\}, B(\bar{p}, \bar{\varphi})],$$

where $\bar{p}(\mathbf{x}) = \hat{p}_{OLL}(\mathbf{x}; \hat{\mathbf{h}}_{\text{ROT}})$ and $\bar{\varphi}(\mathbf{x}) = n^{-1}\sum_{i=1}^n K_{\hat{\mathbf{h}}_{\text{ROT}}}(\mathbf{X}_i - \mathbf{x}) + n^{-2}$. Since $B(\bar{p}, \bar{\varphi})/B = 1 + O_p(n^{-2/(d+4)})$, under the regularity assumptions given in Yang and Tschernig [19], we have

$$\hat{\mathbf{h}}_{\text{PI}} = \mathbf{h}_{\text{opt}}\{1 + O_p(n^{-2/(d+6)})\}.$$

The bandwidth that minimizes the AMSE given by (6.3) can be selected through the method proposed in Okumura and Naito [14]. The proposed bandwidth selector of $g_{\alpha\lambda}$ is given as

$$\hat{g}_{\alpha\lambda} = \arg\min_{g_{\alpha\lambda}} \text{AMSE}[g_{\alpha\lambda}; \tilde{p}, \tilde{p}_{\alpha\alpha}, \tilde{p}_{\lambda\lambda}, \{\tilde{p}_{jj\alpha\alpha}, \tilde{p}_{jj\lambda\lambda}\}, \dot{\varphi}],$$

which was used in the simulation study in the thesis.

In practice, the $d \times d$ matrix whose (α, λ)-element is $\hat{C}_{\alpha\lambda}$ is not always a positive definite. Therefore, we replace $\hat{C}_{\alpha\lambda}$ by $n^{-1}\sum_{i=1}^{n}\hat{p}_{\alpha\alpha}(\mathbf{X}_i; g_{\alpha\alpha})\hat{p}_{\lambda\lambda}(\mathbf{X}_i; g_{\lambda\lambda})$. The ROT and PI bandwidth selectors for NWE and LLLE can also be derived similarly based on each conditional bias in Table 1 and the same conditional variance (5.2).

7. Conclusion

We proposed the weighted kernel estimators of the regression function in the binomial and binary regression problems.

For the binomial regression problem with a single covariate, WNWE was proposed in Section 3. This estimator is NWE with a kernel weighted by the inverse of the variance estimator at each covariate. The bias-adjusted version of WNWE was also obtained through the asymptotic properties. The MSE of the bias-adjusted WNWE is asymptotically smaller than that of NWE under some conditions. The simulation study in Okumura and Naito [13] also demonstrated the good performance of the bias-adjusted WNWE. Two efficient data-driven bandwidth selectors were also proposed: the PI bandwidth selector and the ROT bandwidth selector. Their efficiencies are described in Okumura and Naito [14].

For the binary regression problem with multiple covariates, two practical WLLEs, which are LLTEs with a kernel weighted at each observed covariate, were proposed in Section 5. The behavior of the estimators and the bandwidth selectors was discussed. WLLE with weights constructed by using the true regression function is asymptotically shown in Theorem 5.1 to have a variance smaller than that of OLLE. With regard to practical applications, the WLLEs with weights constructed by using the parametric and nonparametric estimators were proposed. The practical WLLEs cannot necessarily guarantee the variance reduction asymptotically as given in Theorem 5.1. A simulation study to compare estimators described was done in the thesis. From the simulation study, we inferred that the proposed WLLEs reduce the variance and have good performance in terms of IMSE. Further, it was shown that it is effective to use the ROT bandwidth selector in comparatively small samples if the parametric model fits well. The proposed WLLEs can be recommended as a nonparametric estimator of the regression function, considering the disadvantages of the boundary bias problem inherent in NWE and the troubles involved in the calculation of the LLLE. The WLLEs are also useful for the estimation of effective covariates corresponding to a response probability.

Appendix

The following assumptions are needed in Theorems 5.1 and 5.2.

Assumption 6. At $\mathbf{x} \in \mathrm{supp}(\varphi)$, $\varphi(\mathbf{x})$ is second continuously differentiable.

Assumption 7. At $\mathbf{x} \in \mathrm{supp}(\varphi)$, $p(\mathbf{x})$ is second continuously differentiable.

Assumption 8. As $n \to \infty$, $h_j \to 0 (j = 1, \dots, d)$ in such a manner that $h_k/h_j = O(1) (j \neq k)$ and $||\mathbf{h}||^4/(nh_1 \dots h_d) = O(1)$.

Assumption 9. At $\mathbf{x} \in \mathrm{supp}(\varphi)$, $\omega(\mathbf{x})$ is second continuously differentiable.

References

[1] A. Aragaki and N. S. Altman, *Local polynomial regression for binary response*, Computer Science and Statistics: Proceedings of the 29th Symposium on the Interface (1997), 467-472.

[2] J. B. Copas, *Plotting p against x*, Applied Statistics **32** (1983), 25-31.

[3] J. Fan, *Local linear regression smoothers and their minimax efficiencies*, Ann. Statist. **21** (1993), 196-216.

[4] J. Fan and I. Gijbels, *Local Polynomial Modeling and its Applications*, Chapman and Hall, London, 1996.

[5] J. Q. Fan, N. E. Heckman and M. P. Wand, *Local polynomial kernel regression for generalized linear models and quasi-likelihood functions*, J. Amer. Statist. Assoc. **90** (1995), 141-150.

[6] J. E. Grizzel, C. F. Starmer and G. G. Koch, *Analysis of growth and dose-response curves*, Biometrics **25** (1969), 489-504.

[7] M. L. Hazelton, *Bias reduction in kernel binary regression*, Comput. Stat. Data Anal. **51** (2007), 4393-4402.

[8] C. J. Lloyd, *Statistical Analysis of Categorical Data*, John Wiley and Sons, New York, 1999.

[9] H.-G. Müller, *Kernel estimators of zeros and of location and size of extrema of regression functions*, Scand. J. Statist. **12** (1985), 221-232.

[10] H.-G. Müller and T. Schmitt, *Kernel and probit estimates in quantal bioassay*, J. Amer. Statist. Assoc. **83** (1988), 750-759.

[11] H. Okumura, *A note on local likelihood regression for binary response data*, Commun. Statist.-Simul. Comp. **38** (2009), 1019-1025.

[12] H. Okumura and K. Naito, *Kernel smoothing in quantal bioassay*, Japanese Journal of Applied Statistics **32** (2003), 127-144.

[13] H. Okumura and K. Naito, *The weighted estimators in nonparametric binomial regression*, J. Nonparam. Statist. **16** (2004), 39-62.

[14] H. Okumura and K. Naito, *Bandwidth selection for kernel binomial regression*, J. Nonparam. Statist. **18** (2006), 343-356.

[15] D. Ruppert and M. P. Wand, *Multivariate locally weighted least squares regression*, Ann. Statist. **22** (1994), 1346-1370.

[16] D. F. Signorini and M. C. Jones, *Kernel estimators for univariate binary regression*, J. Amer. Statist. Assoc. **99** (2004), 119-126.

[17] J. G. Staniswalis and V. Cooper, *Kernel estimates of dose response*, Biometrics **44** (1988), 1103-1119.

[18] M. P. Wand and M. C. Jones, *Kernel Smoothing* Chapman and Hall, London, 1995.

[19] L. Yang and R. Tschernig, *Multivariate bandwidth selection for local linear regression*, J. R. Statist. Soc. B **61** (1999), 793-815.

Hidenori Okumura: Department of Information Science and Business Management, Chugoku Junior College, Niwase, Okayama, 701-0197, Japan

E-mail address: `okumura@cjc.ac.jp`