The Korean Communications
in Statistics Vol. 9, No. 3, 2002
pp. 743-752

Optimal Design for Locally Weighted Quasi-Likelihood Response Curve Estimator

Dongryeon Park[1)]

Abstract

The estimation of the response curve is the important problem in the quantal bioassay. When we estimate the response curve, we determine the design points in advance of the experiment. Then naturally we have a question of which design would be optimal. As a response curve estimator, locally weighted quasi-likelihood estimator has several more appealing features than the traditional nonparametric estimators. The optimal design density for the locally weighted quasi-likelihood estimator is derived and its ability both in theoretical and in empirical point of view are investigated.

Keywords : Local quasi-likelihood, Nonparametric regression, Optimal design, Response curve.

1. Introduction

In many cases, the outcome of an experiment in bioassay is dichotomous - success or failure. The binary response Y_i of the ith subject at stimulus level x_i is assumed to be an independent random variables with mean $p(x_i)$, $i=1,\ldots,n$. Here p denotes response curve. Let assume $p \in C^2([0,1])$. The stimulus level x_i's are the design points fixed in advance. The statistical aim is the estimation of curve p without assuming a parametric model for p.

The traditional nonparametric regression methods can be used for the estimation of p. Muller and Schmitt (1988) defined the kernel response curve estimator in analogy to the nonparametric regression of Gasser and Muller (1984). It is known that the local polynomial regression has several more appealing features than the traditional nonparametric regression. The better performance near boundaries is one of them (Fan, 1992). Park (1999) considered the local linear regression as the response curve estimator and compared the finite sample performance with Muller and Schmitt's kernel response curve estimator.

However, these estimators ignore the binary nature of response, so they have some

1) Associate Professor, Department of Statistics, Hanshin University, Osan 447-791, KOREA.
E-mail : drpark@hanshin.ac.kr

problems as the estimator of $P(Y=1|X=x)$. The obvious one is that the fitted curve is not guaranteed to lie in the interval (0,1). To overcome these difficulties, a generalization of the weighting mechanism is needed. It is well known that generalized linear model (Nelder and Wedderburn, 1972) is the appropriate technique for binary response and can be applied to the nonparametric regression setting (Tibshirani and Hastie, 1987; Staniswalis, 1989). As a further extension, Wedderburn (1974) first considered a quasi-likelihood method, which requires only specification of a relationship between the mean and the variance of the response. Optimal properties of the quasi-likelihood methods have received considerable attention in the literature (Godambe and Heyde, 1987).

The kernel smoothing idea can be extended to the case where the quasi-likelihood is used. Fan, Heckman, and Wand (1995) proposed the locally weighted quasi-likelihood estimators in one-parameter exponential family, and we take their estimator as the response curve estimator in this paper.

We assume that the design points x_i's are chosen by

$$\int_0^{x_i} f(t)\,dt = \frac{i-1}{n-1} \tag{1}$$

where f is a strictly positive density satisfying $f \in \mathrm{Lip}([0,1])$. We refer to f as the design density which uniquely determines a sequence of designs. Muller (1984) first discussed the question of which design would be optimal with respect to asymptotic MISE for the nonparametric kernel regression. Muller and Schmitt (1988) and Park (1999) discussed the optimal design density for their response curve estimators.

In this paper, we will derive the optimal design density of the locally weighted quasi-likelihood response curve estimator with respect to asymptotic MISE criterion. In Section 2, the properties of the locally weighted quasi-likelihood estimator is summarized and its optimal design density is derived. In Section 3, the small sample properties of the optimal design is investigated by the simulation.

2. Response Curve Estimator and Its Optimal Design

2.1 Locally Weighted Quasi-Likelihood Estimator

Consider binary response variables with single covariate case. Let $Y_1, \ldots, Y_n$ be the independent binary random variables with success probability $p(x_i) = P(Y=1|X=x_i)$. We will assume that $p \in C^2([0,1])$. In parametric generalized linear model it is usual to model a transformation of the regression function $E(Y|X=x) = p(x)$ as linear. The model is given by

$$\eta(x) = \beta_0 + \beta_1 x + \cdots + \beta_d x^d = g(p(x)) \tag{2}$$

where g is the link function and the logit function is the canonical link.

In many applications, the full likelihood function is unknown and one can only specify the relationship between the mean and the variance. Suppose the conditional variance is modeled as $\mathrm{Var}(Y|X=x) = V(p(x))$ for some specific function V. In this case, the estimation of the mean can be achieved by replacing the conditional log-likelihood by the quasi-likelihood function $Q(p(x), y)$ which satisfies

$$\frac{\partial}{\partial w} Q(w, y) = \frac{y-w}{V(w)}, \tag{3}$$

and estimating $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_d)^T$ by maximizing the quasi-likelihood

$$\sum_{i=1}^{n} Q[g^{-1}(\beta_0 + \beta_1 x_i + \cdots + \beta_d x_i^d), Y_i]. \tag{4}$$

Since we deal with binary response, $V(p) = p(1-p)$ and in this case, the quasi-likelihood method coincides with the Bernoulli log-likelihood method.

Fan, Heckman, and Wand (1995) proposed the local quasi-likelihood using kernel weights, which is given by

$$\sum_{i=1}^{n} Q[g^{-1}(\beta_0 + \beta_1(x_i - x) + \cdots + \beta_d(x_i - x)^d), Y_i] K\left(\frac{x_i - x}{h}\right) \tag{5}$$

where h is the bandwidth and K is the kernel function. Maximizing (5) with respect to $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_d)^T$ leads to the maximum local quasi-likelihood estimate

$$\hat{\eta}(x; d, h) = \hat{\beta}_0 \tag{6}$$

and the local quasi-likelihood response curve estimate can be computed by applying the inverse link function

$$\hat{p}(x; d, h) = g^{-1}(\hat{\eta}(x; h)). \tag{7}$$

Suppose the order of local polynomial d is odd. Then, by the results of Fan, Heckman, and Wand (1995), asymptotic mean squared error of $\hat{p}(x; d, h)$ is given by

$$\text{AMSE}(\hat{p}(x;d,h)) = h^{2d+2}\cdot\left(\int z^{d+1}K_{0,d}(z)\,dz\right)^2\left(\frac{\eta^{(d+1)}(x)\cdot p(x)(1-p(x))}{(d+1)!}\right)^2 + \frac{1}{nh}\frac{p(x)(1-p(x))}{f(x)}\int K_{0,d}(z)^2dz \tag{8}$$

where $K_{0,d}$ is the equivalent kernel defined in Fan and Gijbels (1996), and $f(x)$ is the marginal density of X. We integrate (8) to obtain the asymptotic mean integrated error

$$\text{AMISE}(\hat{p}(x;d,h)) = h^{2d+2}\cdot\left(\int z^{d+1}K_{0,d}(z)dz\right)^2\cdot\int\left(\frac{\eta^{(d+1)}(x)\cdot p(x)(1-p(x))}{(d+1)!}\right)^2dx + \frac{1}{nh}\int K_{0,d}(z)^2dz\int\frac{p(x)(1-p(x))}{f(x)}dx \tag{9}$$

Then we can obtain the optimal bandwidth with respect to AMISE criterion by minimizing (9) with respect to h, which is given by

$$h_{\text{opt}} = \left[\frac{(d+1)!\int K_{0,d}(z)^2dz\int[p(x)(1-p(x))/f(x)]dx}{(2d+2)\left(\int z^{d+1}K_{0,d}(z)dz\right)^2\int[\eta^{(d+1)}(x)p(x)(1-p(x))]^2dx\cdot n}\right]^{1/(2d+3)}. \tag{10}$$

Substituting (10) into (9) leads to the minimal asymptotic MISE

$$\text{AMISE}_{\text{opt}} = c\cdot\left[\left(\int z^{d+1}K_{0,d}(z)dz\right)^2\cdot\int\left(\frac{\eta^{(d+1)}(x)\cdot p(x)(1-p(x))}{(d+1)!}\right)^2dx\right]^{1/(2d+3)} \times\left[\int K_{0,d}(z)^2dz\int\frac{p(x)(1-p(x))}{f(x)}dx\right]^{(2d+2)/(2d+3)}n^{-(2d+2)/(2d+3)} \tag{11}$$

where

$$c = (2d+2)^{-(2d+2)/(2d+3)} + (2d+2)^{1/(2d+3)}.$$

2.2 Optimal Design Density

Suppose that the design points, x_i, must be given in advance of the experiment. Then naturally we have a question of which design would be optimal and this question has been discussed by many authors. For the nonparametric response curve estimator, Muller and Schmitt (1988) derived the optimal design density of their estimator which is the Gasser and

Muller type kernel estimator. Park (1999) showed that the optimal design density of the local regression estimator is identical with that of Muller and Schmitt's estimator.

For the locally weighted quasi-likelihood estimator, we can derive the optimal design density with respect to the minimum AMISE by minimizing (11) with respect to $f(x)$.

Theorem 2.1 *Assume that the optimal bandwidth with respect to AMISE is used. Then the optimal design density* $f^*(x)$ *of* $\hat{p}(x;d,h)$ *with respect to minimal asymptotic MISE is given by*

$$f^*(x) = \frac{\sqrt{p(x)(1-p(x))}}{\int \sqrt{p(y)(1-p(y))}\,dy}. \tag{12}$$

To prove this theorem, we need to show that $f^*(x)$ is the minimizer of AMISE of (11) and this follows if we show that $f^*(x)$ is the solution of the variational problem

$$\min_f \int \frac{p(x)(1-p(x))}{f(x)}\,dx$$

with subject to $\int f(x)\,dx = 1$ and $f(x) > 0$, and this was done in Muller (1984). Since $f^*(x)$ is an obvious candidate solution for this problem, Muller (1984) compared it with $f^* + \delta f$, which represents another form of solution, where $\int_0^1 \delta f(x)\,dx = 0$ and $|\delta f(x)| < f^*(x)$, and showed that

$$\int_0^1 \frac{p(t)(1-p(t))}{f^*(t) + \delta f(t)}\,dt \geq \int_0^1 \frac{p(t)(1-p(t))}{f^*(t)}\,dt.$$

3. Simulation Study

A Monte Carlo study was carried out to investigate the small sample properties of the optimal design. We compared the small sample performance of the optimal design and the evenly spaced design. We want to compare the performance of the optimal design with as many designs as possible, but since we do not allow the sequential allocation of the design points, the only comparable design is the evenly spaced design. The optimal design points, x_i^* were chosen by

$$\int_0^{x_i^*} f^*(t)\,dt = \frac{i-1}{n-1} \tag{13}$$

where f^* is the optimal design density in (12), and the evenly spaced design points, x_i were chosen by $x_i = (i-1)/(n-1)$, $i = 1, \ldots, n$.

Since the optimal design density function contains the true response curve, p, we "cheat" by using the knowledge of p in choosing the optimal design points. However, without knowing p, there is no way to construct the optimal design. This unrealistic assumption can be avoided by using the two stage estimation method or the sequential design algorithm in Park and Faraway (1998). In this simulation study, we just want to investigate the ability of the optimal design. After getting the design points, all of the remaining procedures are data adaptive.

As the true response curve, we used the following five models:

1. The logit model, $p_1(x) = [1 + \exp(10 - 20x)]^{-1}$
2. The skewed logit model, $p_2(x) = [1 + \exp(10 - 20x)]^{-2}$
3. The complementary log-log model, $p_3(x) = 1 - \exp(-\exp(-8 + 12x))$
4. The normal mixture model, $p_4 = 0.5\,\Phi\left(\frac{x-0.4}{0.05}\right) + 0.5\,\Phi\left(\frac{x-0.6}{0.05}\right)$
5. The Weibull model, $p_5 = 1 - \exp(-6x)^2$

Parameters for each models were chosen such that $0 \le x_i^* \le 1$, $i = 1, \ldots, n$. $p_1(x)$ is a symmetric sigmoid curve, and $p_2(x)$ and $p_3(x)$ are a non-symmetric sigmoid curve, and $p_4(x)$ is a non-symmetric non-sigmoid curve with three inflection points, and $p_5(x)$ is a non-symmetric strictly concave. curve.

The sample size under consideration were n = 20, 25, 30, 35, 40, 45, and 50. To generate the responses for each designs, Uniform(0,1) pseudo random numbers were constructed and compared with $p(x_i)$ for the respective models.

The order of local polynomial was chosen by $d = 1$. An appropriate choice of the bandwidth is very important. There are some debates for the performance of the several bandwidth selectors in the literature. However, since we would apply the same bandwidth selector to both designs, we could choose the most straightforward selection method and chose the cross-validation method. The evaluation of the locally weighted quasi-likelihood estimator was done by S-Plus function *locfit* (Loader, 1999) with logit link function.

The performance of a design might be measured by the Monte Carlo MISE of the response curve estimator which is constructed by the design. The Monte Carlo MISE were computed

as the average of

$$\frac{1}{m}\sum_{i=1}^{m}(\hat{p}(t_i)-p(t_i))^2 \tag{14}$$

over 500 simulation samples, where $t_1,\ldots,t_m$ are the evenly spaced grid points in [0,1], and m was chosen by $m=1000$.

Simulation results are listed in Table 1 and Table 2. In these tables, MISE_1, MISE_2 denote the MISE of the evenly spaced design, and of the optimal design, respectively, and we put $R=\text{MISE}_2/\text{MISE}_1$. In each model, the optimal design outperforms the evenly spaced design. For the small sample case, the gain of the optimal design against the evenly spaced design can be estimated by the ratio of their Monte Carlo MISE. For the large sample case, the theoretical gain of the optimal design over the evenly spaced design can be computed by the ratio of AMISE in (11), which is given by

$$\frac{\text{AMISE(optimal design)}}{\text{AMIS(evenly spaced design)}}=\frac{\left(\int_0^1\sqrt{p(x)(1-p(x))}\,dx\right)^{8/5}}{\left(\int_0^1 p(x)(1-p(x))\,dx\right)^{4/5}}.$$

This ratio depends only on p and it is calculated in Table 3 for each model. For 5 models considered in the simulation, AMISE for the optimal design is at most 60 % of AMISE for the evenly spaced design.

4. Discussion

The estimation of the response curve is one of the major topic in the quantal bioassay. When we estimate the response curve, we determine the design points in advance of the experiment. Then naturally we have a question of which design would be optimal. We have driven the optimal design density of the locally weighted quasi-likelihood estimator with respect to AMISE and investigated its ability both in theoretical and in empirical points of view.

We would need some prior knowledge of p to construct the optimal design in practice. Failing that, it is often possible to observe the results of the measurements sequentially so that we may decide on the position of the next design point on the basis of the previous observation. In this case, we could adapt the sequential design algorithm in Park and Faraway (1998).

Table 1 : Monte Carlo MISE for each design.

Model	n	MISE $_1$	MISE $_2$	R
p_1	20	0.0209	0.0072	0.3444
	25	0.0162	0.0058	0.3580
	30	0.0133	0.0043	0.3233
	35	0.0103	0.0039	0.3786
	40	0.0084	0.0036	0.4285
	45	0.0072	0.0029	0.4027
	50	0.0067	0.0025	0.3731
p_2	20	0.0194	0.0063	0.3247
	25	0.0160	0.0046	0.2875
	30	0.0121	0.0040	0.3305
	35	0.0104	0.0035	0.3365
	40	0.0090	0.0032	0.3555
	45	0.0079	0.0023	0.2911
	50	0.0072	0.0020	0.2777
p_3	20	0.0208	0.0102	0.4903
	25	0.0151	0.0081	0.5364
	30	0.0132	0.0061	0.4621
	35	0.0109	0.0052	0.4770
	40	0.0094	0.0044	0.4680
	45	0.0075	0.0041	0.5466
	50	0.0067	0.0036	0.5373

Table 2 : Monte Carlo MISE for each design.

Model	n	$MISE_1$	$MISE_2$	R
p_4	20	0.0222	0.0127	0.5720
	25	0.0159	0.0108	0.6792
	30	0.0131	0.0078	0.5954
	35	0.0108	0.0065	0.6018
	40	0.0088	0.0061	0.6931
	45	0.0081	0.0049	0.6049
	50	0.0072	0.0048	0.6666
p_5	20	0.0198	0.0071	0.3585
	25	0.0145	0.0058	0.4000
	30	0.0138	0.0041	0.2971
	35	0.0073	0.0038	0.5205
	40	0.0063	0.0033	0.5238
	45	0.0062	0.0030	0.4838
	50	0.0061	0.0027	0.4426

Table 3 : Ratio of AMISE of the optimal design to the evenly spaced design

Model	p_1	p_2	p_3	p_4	p_5
Ratio	0.5606	0.4715	0.5882	0.4717	0.4887

References

[1] Collett, D. (1991). *Modelling Binary Data*, London, Chapman & Hall

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

[3] Fan, J., Heckman, N., and Wand, M. (1995). Local Polynomial Kernel Regression for Generalized Linear Models and Quasi-Likelihood Functions. *Journal of the American Statistical Association*. **90**. 141-150

[4] Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*, London, Chapman & Hall

[5] Gasser, T. and Muller, H.G. (1984). Estimating Regression Functions and Their Derivatives by the Kernel Method. *Scandinavian Journal of Statistics*, **11**, 171-185

[6] Godambe, V.P., and Heyde, C.C. (1987). Quasi-Likelihood and Optimal Estimation. *International Statistical Review*, **55**, 231-244

[7] Loader, C. (1999). *Local Regression and Likelihood*, New York, Springer-Verlag

[8] Morgan, B.J.T. (1992). *Analysis of Quantal Response Data*, London, Chapman & Hall

[9] Muller, H. (1984). Optimal Designs for Nonparametric Kernel Regression. *Statistics & Probability Letters*, **2**, 285-290

[10] Muller, H. and Schmitt, T. (1988). Kernel and Probit Estimates in Quantal Bioassay. *Journal of the American Statistical Association*, **83**, 750-759

[11] Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized Linear Models. *Journal of Royal Statistical Society*, Series A, **135**, 370-384

[12] Park, D. (1999). Comparison of Two Response Curve Estimators. *Journal of Statistical Computation and Simulation*, **62**, 259-269

[13] Park, D. and Faraway, J.J. (1998). Sequential Design for Response Curve Estimation. *Journal of Nonparametric Statistics*. **9**, 155-164

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

[15] Tibshirani, R. and Hastie, T. (1987). Local Likelihood Estimation. *Journal of the American Statistical Association*, **82**, 559-568

[16] Wedderburn, R.W.M. (1974). Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method. *Biometrika*, **61**, 439-447

[2002년 7월 접수, 2002년 12월 채택]