Hacettepe Journal of Mathematics and Statistics
Volume 39 (4) (2010), 583 – 597

ON THE PERFORMANCE OF THE SEMIPARAMETRIC BINARY RESPONSE MODEL WHEN THE TRUE MODEL IS PARAMETRIC LOGISTIC

Özge Akkuş[*][†], Hüseyin Tatlıdil[‡] and Atilla Göktaş[*]

Received 18 : 11 : 2008 : Accepted 18 : 03 : 2010

Abstract

In this article, a simulation study is performed to reveal the deviations of the semiparametric binary response model from its parametric counterpart, based on various scenarios including different sample sizes, different bandwidth parameters containing the optimal ones, different forms of the linear index function and two and higher dimensional cases of the explanatory variables when the true model is logistic regression. The method of the Density Weighted Average Derivative Estimator (DWADE) is used in the semi-parametric estimation. A real data set on liquefaction is used to demonstrate the effectiveness of the simulation results with the results in practice. Additionally, new commands written for the estimation of both models in the Windows based 4.8 version of the XploRe package are introduced. This study may be seen as an updated form of the article by Proença and Werwatz (1994) which used XploRe commands written for both estimators in an old MS Dos format.

Keywords: Semi-parametric model, Density weighted average derivative estimator, XploRe, Logistic regression.

2000 AMS Classification: Primary: 62 J 99. Secondary: 62 J 12.

[*] Department of Statistics, Muğla University, Muğla, Turkey.
E-mail: (Ö. Akkuş) ozge.akkus@mu.edu.tr (A. Göktaş) gatilla@mu.edu.tr
[†] Corresponding Author.
[‡] Department of Statistics, Hacettepe University, 06800 Beytepe, Ankara, Turkey.
E-mail: tatlidil@hacettepe.edu.tr

1. Introduction

Most research fields of Applied Econometrics and Statistics focus on the estimation of the conditional expectation function $\mathrm{E}(Y/X = x)$ presented in Equation (1). If the dependent variable Y is binary, the conditional expectation function gives the probability of an observation having a positive response coded as $Y = 1$.

$$(1) \qquad \mathrm{E}(Y/X = x) = P(Y = 1/X = x).$$

Here X denotes the vector of explanatory variables. Possible approaches for the estimation of the model parameters are: the fully parametric approach, the fully nonparametric approach and the semi-parametric approach [3,8].

The fully parametric approach is given as,

$$(2) \qquad \mathrm{E}(Y/X = x) = P(Y = 1/X = x) = G[\upsilon(x)],$$

where G is a known distribution function related to the error term, and β is the vector of parameters. In the standard parametric models, the linear functional form of the explanatory variables defined as,

$$(3) \qquad \upsilon(x) = \beta_0 + x^T\beta,$$

are used. Because of the linear index assumption $X^T\beta$ and a known G, this approach is called the *fully parametric approach.* The binary logit model is a well known example for this kind of model. The logistic regression model given below is obtained when the logistic distribution is assumed for G [1,9,10,14]:

$$(4) \qquad \mathrm{E}(Y/X = x) = P(Y = 1/X = x) = \frac{exp[\upsilon(x)]}{1 + exp[\upsilon(x)]} = \frac{exp[\beta_0 + x^T\beta]}{1 + exp[\beta_0 + x^T\beta]}$$

If G is correctly specified, this approach achieves the property of the statistical efficiency of the parameters, and allows for the extrapolation of x values that are out of the support of x. However, it is a well known fact that G is rarely known in most applications and that when it is misspecified, the results will be highly misleading.

The fully nonparametric approach is given as follows.

$$(5) \qquad \mathrm{E}(Y/X = x) \equiv G(x).$$

In the approach, G is an unknown function that can be estimated using the nonparametric regression of Y on X. This approach minimizes the specification errors because no assumptions are required for the model. However, the main disadvantage of this approach is that the estimation and interpretation become gradually more difficult when the dimension of the vector X of explanatory variables increases [3,8,9].

The semiparametric approach, defined as,

$$(6) \qquad \mathrm{E}(Y/X = x) = P(Y = 1/X = x) = g\{\upsilon(x)\}$$

contains both an unknown finite-dimensional parameter $(\upsilon(x))$ and an unknown function $(g(\cdot))$. More assumptions are required for this model compared with the fully nonparametric, approach whereas less assumptions are needed than for the fully parametric one. The estimation procedure is composed of two steps.

In the first step, the parameter vector β is estimated using one of the semi-parametric estimation techniques appropriate for the data structure. In the second step, linear index values are computed using $\hat{\beta}$, which is the estimator of β. Then, an unknown distribution function "g" is estimated by applying one of the non-parametric regression methods of Y on $x^T\hat{\beta}$ [3,8].

It is also worthwhile to note that the fully parametric and semiparametric models are generally called "Single Index Models" due to the fact that all explanatory variables are summed under only one linear index function of the form $\upsilon(x) = x^T\beta$.

In this study, the parametric and the semiparametric approaches are focused. The principal aim of this study is to reveal deviations of the semiparametric binary response model from its parametric counterpart based on various scenarios mentioned in Section 3, when the true model is the parametric logistic regression. In addition, new XploRe commands generated in the Windows based version for the estimation of the semiparametric DWADE and the parametric logistic regression are introduced.

The comprehensive methodology for the estimation techniques of binary response data is given in Section 2. Particularly, the theory of the Maximum Likelihood Estimation (MLE) of the parametric logistic regression and the DWADE of the semiparametric estimation method are introduced here. Sections 3 and 4 contain the scenarios, and the design and some remarkable results of the study, respectively. Evidence for these is investigated in Section 5 using real data. New XploRe commands of the DWADE and logistic regression model are presented in Appendix 1 and 2, respectively, together with explanations.

2. Methodology of the estimation procedures

In the parametric approach, parameter estimates are obtained according to the MLE, whereas β is estimated by the DWADE in the semiparametric approach provided that all the explanatory variables are continuous.

2.1. The method of the maximum likelihood estimation of the parametric logistic regression. Under the assumption that the error term has a logistic distribution, the probability of having a $Y = 1$ could be expressed by,

$$\text{(7)} \qquad \mathrm{E}(Y/X = x) = P(Y = 1/X = x) = \frac{\exp(\sum \hat{\beta}_k x_{ik})}{1 + \exp(\sum \hat{\beta}_k x_{ik})},$$

where k denotes the number of explanatory variables. The likelihood and the logarithmic likelihood functions for the MLE of β are given by Equation (8) and (9), respectively.

$$\text{(8)} \qquad \mathrm{L}(Y/X, \hat{\beta}) = \prod_{i=1}^{N} \left[\frac{\exp\left(\sum \hat{\beta}_k x_{ik}\right)}{1 + \exp\left(\sum \hat{\beta}_k x_{ik}\right)}\right]^{Y_i} \left[\frac{1}{1 + \exp\left(\sum \hat{\beta}_k x_{ik}\right)}\right]^{1-Y_i},$$

$$\text{(9)} \qquad \log \mathrm{L}(Y/X, \hat{\beta}) = \sum_{i=1}^{N} \Big[Y_i \log P_i + (1 - Y_i)\log(1 - P_i)\Big].$$

If $\hat{\beta}$ maximizes $\mathrm{L}(Y/X, \hat{\beta})$, it also maximizes $\log \mathrm{L}(Y/X, \hat{\beta})$. Hence, the first order derivatives are performed and equalized to "0" in order to have the parameter estimates maximizing the likelihood of observing the sample Y. Clearly, K equations are obtained for the K parameters. The simultaneous solution of these equations gives the MLE estimates of β. The general form of the likelihood equations is given as follows [1,10,14].

$$\text{(10)} \qquad \frac{\partial \log \mathrm{L}}{\partial \hat{\beta}} = \sum_{i=1}^{N} \left[Y_i - \frac{\exp\left(\sum \hat{\beta}_k x_{ik}\right)}{1 + \exp\left(\sum \hat{\beta}_k x_{ik}\right)}\right] x_{ij} = 0;\ i = 1, 2, \ldots, N;\ k = 1, 2, \ldots, K.$$

2.2. The density weighted average derivative estimator. The DWADE estimator has two important advantages in terms of the distributional assumption and of the resulting estimator. That is, no distribution assumption is needed for the dependent variable Y, and the resulting estimator is a "direct estimator" which is not iterative. The main idea here is to estimate β using the average derivatives. A strict condition of the DWADE is that it can be applied directly only to the models containing continuous explanatory variables.

Assume that x is a continuously distributed random vector, and G is a differentiable function needed for the identifiability of β. Under these assumptions,

$$\frac{\partial \mathrm{E}(Y/x)}{\partial x} = \beta G(x^T\beta) \tag{11}$$

can be derived. Additionally, for any restricted and continuous function W, the following expression is obtained.

$$\mathrm{E}\left[W(x)\frac{\partial \mathrm{E}(Y/x)}{\partial x}\right] = \beta \mathrm{E}\left[W(x)G'(x^T\beta)\right]. \tag{12}$$

The left side of Equation (12) is called the Average Derivative Estimator (ADE) of the $\mathrm{E}(Y/x)$ with the weight function W. Equation (12) indicates that the weighted average derivative of $\mathrm{E}(Y/x)$ is proportional to $\boldsymbol{\beta}$ for each value of x. Because of the requirement of scale normalization, $\boldsymbol{\beta}$ is only defined according to the scale and any weighted average derivative of $\mathrm{E}(Y/x)$ is equal to $\boldsymbol{\beta}$ [15, Page 1404]. Therefore, it should be noticed that only estimating the left side of Equation (12) is adequate for the estimation of $\boldsymbol{\beta}$. Dividing each component on the left side of Equation (11) by the first component, the scale normalization of $\boldsymbol{\beta}_1 = 1$ can be achieved.

The left side of Equation (11) can be estimated by replacing the kernel estimator of $\frac{\partial \mathrm{E}(Y/x)}{\partial x}$ and the sample mean for the expected value of the population $\mathrm{E}(\,\cdot\,)$ [15].

2.1. Theorem. *Let $p(\,\cdot\,)$ be the probability density function of x and $W(x) = p(x)$. Then the left side of Equation (12) can be written as follows.*

$$\mathrm{E}\left[W(x)\frac{\partial \mathrm{E}(Y/x)}{\partial x}\right] = \mathrm{E}\left[p(x)\frac{\partial \mathrm{E}(Y/x)}{\partial x}\right] = \int \frac{\partial \mathrm{E}(Y/x)}{\partial x}[p(x)]^2\,dx. \tag{13}$$

Assuming that $p(x) = 0$ when x is on the boundary of the support of x, integration by parts gives:

$$\begin{aligned}\mathrm{E}\left[W(x)\frac{\partial \mathrm{E}(Y/x)}{\partial x}\right] &= -2\int \mathrm{E}(Y/x)\frac{\partial p(x)}{\partial x}p(x)\,dx \\ &= -2\mathrm{E}\left\{\mathrm{E}(Y/x)\frac{\partial p(x)}{\partial x}\right\} = -2\mathrm{E}\left[Y\frac{\partial p(x)}{\partial x}\right]\end{aligned} \tag{14}$$

The proof of the theorem is given in Appendix 3.

3. The simulation study

As mentioned in Section 1, differences between the semiparametric binary response model and parametric logistic regression are examined here with a simulation study for the case when the latter is the true model. The simulation design is as follows.

Data simulated according to the sample sizes 25, 100, 250 and 500 are used for the model with two explanatory variables. The smallest sample size is determined as 50 for the case of higher dimensional explanatory variables due to the fact that the number of observations that are able to achieve consistency with the model decrease as the number of the variables increases.

The linear index function given in Equation (15) is determined by following the study of Proença and Silva [17], so that the minimal condition of the identifiability of the model parameters in the semiparametric estimation is satisfied by setting the first coefficient of a continuous variable (here X_1) to the value “1”.

$$\text{index}_{(i1)} = X^T\beta = 1 + X_{1(i1)} + X_{2(i1)};\ i = 1, 2, \ldots, n. \tag{15}$$

Here, n denotes the number of observations. X_1 and X_2 are assumed to follow a Standard Normal and Uniform (0,1) distribution, respectively. In addition to the study of [16], we also discuss the effect of a higher dimensional data structure on the results. To do this, we determine two additional continuous explanatory variables X_3 and X_4 that are drawn from the Exponential (2) and the Weibull (5,1) distributions, respectively.

It is a well known fact that the more variables a model has, the more observations it needs to obtain better estimates of the model parameters. Additionally, in such a case, the estimation gradually becomes hard, especially for the semiparametric approach. Therefore, we performed the second step of the simulation study based on four variables. The form of the true linear index function was taken as follows.

$$\text{index}_{(i2)} = X^T\beta = 1 + X_{1(i2)} + X_{2(i2)} + X_{3(i2)} - 2X_{4(i2)};\ i = 1, 2, \ldots, n. \tag{16}$$

The reason why we selected the true coefficients as given above arises from the fact that the supports of the variables X_2, X_3 and X_4 are strictly positive, and this leads to a problem in the data derivation process. That is, if we assume all the true coefficients are positive, the number of “1”’s in the dependent variable increases along with the increase in the index values in Equation (16), and accordingly in P_i in Equation (17). Large numbers of P_i give rise to the derivation of a larger number of the values “1” rather than “0” in the dependent variable. Therefore, we set the coefficient of X_4 to the negative value (-2) to obtain balanced data sets including a sufficient number of “0” values for the dependent variable. In this way, we also examined the effect of different functional forms of the linear index functions on the results.

The probabilities with respect to the index values are computed by following the ordinary logistic regression function. This means that all the simulated data are consistent with the parametric model. According to logistic regression, the probability of having the positive level coded as “1” in the dependent variable Y is computed as follows.

$$\text{E}(Y_i = 1/X_i) = P_i = \frac{\exp(\text{index}_i)}{1 + \exp(\text{index}_i)};\ i = 1, 2, \ldots, n. \tag{17}$$

The dependent variable Y_i is assumed to follow a Bernoulli distribution with probability P_i. Algorithms for simulating the data were constructed by following the procedures in [12]. The results are assessed in terms of the average estimates of the parameters β and the Averaged Mean Square Error (AMSE) of the estimates, defined as

$$\text{AMSE}_{\hat{\beta}} = \frac{\sum_{i=1}^{100}(\text{True coefficient} - \hat{\beta}_i)^2}{100}. \tag{18}$$

The optimal bandwidths are estimated by the method of Least Squares Cross Validation (for details, see [3, 4] and [8]). Hardle [4] presents the theory of the method well and in detail. The optimal bandwidth values of 0.5848, 0.4642, 0.3984 and 0.3550 were obtained for the case of two explanatory variables, whereas the values of 0.5210, 0.4642, 0.3984 and 0.3550 were computed for higher dimensional data for each sample size, respectively. That the estimated optimal bandwidth values are nearly the same for each sample size is an expected result because the samples are drawn from the same distribution.

In order to reveal the deviations of the semiparametric estimates from the logistic regression model as the bandwidth parameters change, five additional bandwidths were determined with 0.1 jumps in the support of the optimal ones. All these bandwidths are presented in Tables in Section 4.

100 replications were performed for each simulation scenario. Estimates of β for the parametric and the semiparametric models were carried using the programme XploRe 4.8. New XploRe commands for the estimation of β were generated and executed in the Windows based version of the XploRe package. This part of our study extends the study of Proença and Werwatz [16] in terms of the renewed and updated commands for two types of model [5,6,7].

3.1. The application steps.

(1) In the first step, logistic regression analysis was applied to the data sets and the unknown β parameter vector was estimated using the XploRe commands in Appendix 2.
(2) Five bandwidth parameters were determined in support of the optimal ones for each sample size and the DWADE estimates obtained using the XploRe commands in Appendix 1.
(3) The coefficient of the first variable was normalized to "1" to satisfy the identifiability conditions of the semiparametric DWADE.
(4) The coefficient normalized to "1" in (3) was also normalized to "1" in the logistic regression estimates to compare the DWADE estimates with the parametric alternatives.
(5) Bandwidth parameters having results very close to the true model (logit model) were determined.
(6) 100 replications were performed and the parameter estimates of β obtained when the true coefficients are assumed as given in Equation (15) and (16).
(7) In the last step, the average values of all the estimated parameters were calculated and the differences between the two models investigated by examining the averaged estimates of the β and AMSE values.

Note that all these steps are replicated both for the cases of two and higher dimensional explanatory variables.

4. Results and discussions

All the simulation results are summarized in Tables 1-4. Tables 1 and 3 present the estimated vector of coefficients $\hat{\beta}$, that is the average estimate of all results obtained from both the logistic regression and DWADE, whereas the averaged prediction errors of the normalized parameter estimates $\hat{\beta}$ are given in Tables 2 and 4 for each sample size and assumed linear index function.

4.1. The case of two explanatory variables. Before interpreting the results, it is evident that the assumed linear index function given by Equation (15) is satisfied as the sample size increases. Additionally, as emphasized above, it should be noted that the first coefficient of a continuous explanatory variable (here, the coefficient of X_1) is normalized to the value "1", and the estimation of the intercept term is not required to be able to achieve the identifiability conditions of the model parameters in the semiparametric approach (for details, see [11] and [13]).

Different normalizations are required in the estimation of the parametric and the semiparametric models. Therefore, we also normalized the coefficient of X_1 to the value "1" so that we could compare the results of the parametric logistic regression and its semiparametric alternative.

Table 1. The averaged estimates of β obtained by DWADE and logistic regression for the case of two explanatory variables

	LOGIT True Model	DWADE				
		$n = 25$				
		$h_1 = 0.3848$	$h_2 = 0.4848$	$h_3 = 0.5848$ **(optimal)**	$h_4 = 0.6848$	$h_5 = 0.7848$
Intercept	1.22738	-	-	-	-	-
$\bar{\hat{\beta}}_1$ (fixed)	1	1	1	1	1	1
$\bar{\hat{\beta}}_2$	**2.74084**	-12.35920	9.72429	**0.17758**	0.89666	0.44658
		$n = 100$				
		$h_1 = 0.2642$	$h_2 = 0.3642$	$h_3 = 0.4642$ **(optimal)**	$h_4 = 0.5642$	$h_5 = 0.6642$
Intercept	1.13577	-	-	-	-	-
$\bar{\hat{\beta}}_1$ (fixed)	1	1	1	1	1	1
$\bar{\hat{\beta}}_2$	**0.88059**	3.16893	4.65912	**1.03031**	1.11832	1.02002
		$n = 250$				
		$h_1 = 0.1984$	$h_2 = 0.2984$	$h_3 = 0.3984$ **(optimal)**	$h_4 = 0.4984$	$h_5 = 0.5984$
Intercept	1.07989	-	-	-	-	-
$\bar{\hat{\beta}}_1$ (fixed)	1	1	1	1	1	1
$\bar{\hat{\beta}}_2$	**0.99135**	1.07161	1.11725	**1.06907**	1.05005	1.04229
		$n = 500$				
		$h_1 = 0.1550$	$h_2 = 0.2550$	$h_3 = 0.3550$ **(optimal)**	$h_4 = 0.4550$	$h_5 = 0.5550$
Intercept	1.00605	-	-	-	-	-
$\bar{\hat{\beta}}_1$ (fixed)	1	1	1	1	1	1
$\bar{\hat{\beta}}_2$	**1.09475**	0.82830	1.02378	**1.02057**	1.03566	1.05449

We see from Table 1 that the averaged DWADE estimates of the coefficient β are about: 0.17758, 1.03031, 1.06907 and 1.02057 for sample sizes of 25, 100, 250 and 500, respectively. The corresponding averaged true estimates obtained from the parametric logistic regression model are about: 2.74084, 0.88059, 0.99135 and 1.09475.

In the light of all these findings, the most important points to be emphasized related to the estimates of β are summarized below.

(1) The semiparametric DWADE estimates that are very close to the optimal h approximate the true coefficients as the sample size increases. This is evident especially for the sample sizes of 250 and 500. That is, the true coefficient for $n = 250$ is is 0.99135, whereas the DWADE estimate related to the optimal h is 1.06907. Similarly, the true coefficient is 1.09475 and the corresponding DWADE estimate is 1.02057 for $n = 500$.

(2) All the DWADE estimates are closely related to the optimal bandwidth parameter h. This result indicates that the optimal bandwidth values should be carefully determined.

(3) We could say that if we suspect whether the true model is a parametric model or not, one of the best solutions to the problem is to use its semiparametric alternative by conditioning that the optimal bandwidths are correctly specified.

Table 2 gives the prediction errors (AMSE values) with respect to the estimated parameters $\hat{\beta}$ for each model.

Table 2. Prediction errors of the normalized $\hat{\beta}$ for the case of two explanatory variables

	LOGIT True Model	DWADE				
	$n = 25$					
		$h_1 = 0.3848$	$h_2 = 0.4848$	$h_3 = 0.5848$ **(optimal)**	$h_4 = 0.6848$	$h_5 = 0.7848$
Intercept	14.11394	-	-	-	-	-
$\bar{\hat{\beta}}_1$ (fixed)	0	0	0	0	0	0
$\bar{\hat{\beta}}_2$	**139.23772**	7146.55026	8251.83774	**547.20053**	27.23247	49.06606
	$n = 100$					
		$h_1 = 0.2642$	$h_2 = 0.3642$	$h_3 = 0.4642$ **(optimal)**	$h_4 = 0.5642$	$h_5 = 0.6642$
Intercept	0.40984	-	-	-	-	-
$\bar{\hat{\beta}}_1$ (fixed)	0	0	0	0	0	0
$\bar{\hat{\beta}}_2$	**1.09134**	221.56722	507.69305	**5.57699**	2.45342	1.76881
	$n = 250$					
		$h_1 = 0.1984$	$h_2 = 0.2984$	$h_3 = 0.3984$ **(optimal)**	$h_4 = 0.4984$	$h_5 = 0.5984$
Intercept	0.17439	-	-	-	-	-
$\bar{\hat{\beta}}_1$ (fixed)	0	0	0	0	0	0
$\bar{\hat{\beta}}_2$	**0.56702**	18.58513	1.57858	**0.98616**	0.81485	0.75076
	$n = 500$					
		$h_1 = 0.1550$	$h_2 = 0.2550$	$h_3 = 0.3550$ **(optimal)**	$h_4 = 0.4550$	$h_5 = 0.5550$
Intercept	0.06042	-	-	-	-	-
$\bar{\hat{\beta}}_1$ (fixed)	0	0	0	0	0	0
$\bar{\hat{\beta}}_2$	**0.20687**	13.51334	0.57463	**0.30976**	0.25344	0.23922

The assumed true coefficients are: $\beta_0 = 1$ (constant term), $\beta_1 = 1$ and $\beta_2 = 1$. As mentioned above, the constant term is not predictable in the semiparametric approach. Therefore, no comparison could be made related to this term. Additionally, another important point to be stressed is that the prediction error is equal to "0" for the fixed parameters ($\hat{\beta}_1' s$) due to the coincidence of the true (1) and the fixed (1) parameter values. The following findings are derived from Table 2 and are worth mentioning.

(1) The AMSE values for each model strictly decrease in conjunction with the increase in the sample size. This is an expected result because the estimated linear index function gradually approximates to the assumed linear index function as the sample size increases. Similarly, the estimated coefficients get closer to the true coefficients.

(2) The closest prediction errors of the logit model and its semiparametric alternative arise in the support of the optimal bandwidth values.

(3) Although the errors are not precisely overlapped in the optimal h, the magnitudes of deviations between two methods in this level could not be assessed extremely significant. That is, it is a well known fact that the semiparametric

approach is more flexible than the parametric counterpart in that it does not require restrictive assumptions related to the error term. In other words, minimal deviations from the true values could be ignored at the expense of obtaining better results without testing the parametric model assumptions.

4.2. The case of more than two explanatory variables. We see from Table 3 that the assumed linear index function given by Equation (16) is also satisfied here as the sample size increases, especially for the sample sizes of 100, 250 and 500, similar to the case of two explanatory variables.

Table 3. The averaged estimates of β obtained by DWADE and logistic regression analysis for the case of high dimensional explanatory variables

	LOGIT True Model	DWADE				
	$n = 50$					
		$h_1 = 0.3210$	$h_2 = 0.4210$	$h_3 = 0.5210$ **(optimal)**	$h_4 = 0.6210$	$h_5 = 0.7210$
Intercept	-0.40961	-	-	-	-	-
$\bar{\hat{\beta}}_1$ (fixed)	1	1	1	1	1	1
$\bar{\hat{\beta}}_2$	**0.59176**	4.15083	0.64357	**0.74963**	1.68570	1.60373
$\bar{\hat{\beta}}_3$	**1.03166**	0.42493	0.71214	**0.98518**	0.89218	0.66020
$\bar{\hat{\beta}}_4$	**-1.5586**	10.10531	-1.62747	**2.78875**	-1.30269	-0.33948
	$n = 100$					
		$h_1 = 0.2642$	$h_2 = 0.3642$	$h_3 = 0.4642$ **(optimal)**	$h_4 = 0.5642$	$h_5 = 0.6642$
Intercept	-0.07196	-	-	-	-	-
$\bar{\hat{\beta}}_1$ (fixed)	1	1	1	1	1	1
$\bar{\hat{\beta}}_2$	**0.97056**	-0.27443	3.43967	**1.76146**	3.26923	1.03277
$\bar{\hat{\beta}}_3$	**1.01266**	0.98159	0.57208	**1.02931**	-0.42772	0.85534
$\bar{\hat{\beta}}_4$	**-1.92629**	-17.87740	-3.15632	**-2.04997**	-1.59315	-1.95999
	$n = 250$					
		$h_1 = 0.1984$	$h_2 = 0.2984$	$h_3 = 0.3984$ **(optimal)**	$h_4 = 0.4984$	$h_5 = 0.5984$
Intercept	-0.06173	-	-	-	-	-
$\bar{\hat{\beta}}_1$ (fixed)	1	1	1	1	1	1
$\bar{\hat{\beta}}_2$	**1.01639**	0.57392	-0.19480	**1.10901**	0.99633	0.99667
$\bar{\hat{\beta}}_3$	**1.07099**	0.51764	1.32687	**1.00623**	0.88124	0.82745
$\bar{\hat{\beta}}_4$	**-2.06540**	-0.23570	-6.59050	**-2.44444**	-2.13307	-2.0447
	$n = 500$					
		$h_1 = 0.1550$	$h_2 = 0.2550$	$h_3 = 0.3550$ **(optimal)**	$h_4 = 0.4550$	$h_5 = 0.5550$
Intercept	-0.09479	-	-	-	-	-
$\bar{\hat{\beta}}_1$ (fixed)	1	1	1	1	1	1
$\bar{\hat{\beta}}_2$	**0.98830**	-1.12294	1.71249	**1.16324**	1.07162	1.04316
$\bar{\hat{\beta}}_3$	**1.01627**	-0.07824	1.26020	**0.88526**	0.81939	0.77778
$\bar{\hat{\beta}}_4$	**-1.92381**	8.00478	-1.53216	**-2.00846**	-1.98448	-1.95670

Another remarkable point to be mentioned is the poor fitting between the true and the semiparametric models for small sample size, even though the smallest sample size is taken as 50, as opposed to 25 for the case of two variables. This is not surprising because we know that additional observations are required in conjunction with the increasing number of explanatory variables included in the model. Additionally, that the DWADE estimates are closely related to the optimal h once again emphasizes the importance of the determination of an accurate optimal bandwidth parameter h for obtaining more realistic results.

Table 4. Prediction errors of the normalized parameter estimates $\hat{\beta}$ for the case of high dimensional explanatory variables

	LOGIT True Model	DWADE				
	$n = 50$					
		$h_1 = 0.3210$	$h_2 = 0.4210$	$h_3 = 0.5210$ **(optimal)**	$h_4 = 0.6210$	$h_5 = 0.7210$
Intercept	10.88737	-	-	-	-	-
$\hat{\beta}_1$ (fixed)	0	0	0	0	0	0
$\hat{\beta}_2$	**4.88908**	2392.67232	123.65149	**82.46214**	74.84145	19.51970
$\hat{\beta}_3$	**6.73068**	21.00283	26.26624	**49.05042**	10.08207	4.82660
$\hat{\beta}_4$	**10.48524**	7861.05452	312.53586	**1044.54455**	1002.98701	276.66716
	$n = 100$					
		$h_1 = 0.2642$	$h_2 = 0.3642$	$h_3 = 0.4642$ **(optimal)**	$h_4 = 0.5642$	$h_5 = 0.6642$
Intercept	3.39374	-	-	-	-	-
$\hat{\beta}_1$ (fixed)	0	0	0	0	0	0
$\hat{\beta}_2$	**1.25318**	1692.86610	258.83485	**13.82980**	417.97227	4.191967
$\hat{\beta}_3$	**0.17097**	161.59666	46.16513	**1.52196**	163.51049	0.744236
$\bar{\hat{\beta}}_4$	**1.88015**	31306.44008	157.11766	**27.39113**	11.79027	3.61303
	$n = 250$					
		$h_1 = 0.1984$	$h_2 = 0.2984$	$h_3 = 0.3984$ **(optimal)**	$h_4 = 0.4984$	$h_5 = 0.5984$
Intercept	2.04739	-	-	-	-	-
$\hat{\beta}_1$ (fixed)	0	0	0	0	0	0
$\hat{\beta}_2$	**0.36112**	83.29981	278.32245	**2.49383**	0.91080	0.63160
$\hat{\beta}_3$	**0.08002**	7.22085	21.87966	**0.57970**	0.20723	0.14728
$\hat{\beta}_4$	**0.98941**	255.53889	328.66594	**6.10585**	2.87019	2.04672
	$n = 500$					
		$h_1 = 0.1550$	$h_2 = 0.2550$	$h_3 = 0.3550$ **(optimal)**	$h_4 = 0.4550$	$h_5 = 0.5550$
Intercept	1.58511	-	-	-	-	-
$\hat{\beta}_1$ (fixed)	0	0	0	0	0	0
$\hat{\beta}_2$	**0.17467**	486.83613	17.42593	**0.81280**	0.43399	0.36464
$\hat{\beta}_3$	**0.02224**	28.54659	4.92197	**0.10593**	0.07556	0.07902
$\hat{\beta}_4$	**0.42384**	2981.29233	41.01429	**1.79859**	1.16975	1.00468

The averaged deviations of the predicted parameter values given in Table 3 from the true coefficients are summarized below. The assumed true coefficients are: $\beta_0 = 1$ (constant term), $\beta_1 = 1$, $\beta_2 = 1$, $\beta_3 = 1$ and $\beta_4 = -2$ for the variables X_1, X_2, X_3, and X_4

respectively. As emphasized in Subsection 4.2, it should be noted that more observations are needed as the number of variables increases in order to obtain the best model fit. Therefore, high level prediction errors could be expected, especially for the sample size of 50.

It may be easily concluded from Table 4 that the results support the findings obtained in the case of two explanatory variables. That is, the prediction errors sharply decrease in parallel with the increase in the sample sizes corresponding to the optimal bandwidth parameters where the estimated coefficients approximate to the true coefficients assumed.

In accordance with all the results, we could say that all findings and interpretations are valid both for models including two and higher dimensional explanatory variables and different forms of the linear index function.

In all, since the prediction errors of the normalized parameter estimates $\hat{\beta}$ for logit and DWADE are close enough and the differences may be ignored, DWADE proves promising and gives very much hope for h optimal.

5. Evidence from real data

In this section, a real data set taken from the registry of the General Directorate of Highways in Turkey in 2009 on the potential liquefaction of ground in Ízmir, one of the largest cities in Turkey, is used [2]. Findings from 314 well-bores are examined to determine whether the simulation results given in the previous sections are supported by the results of a real data set in practice.

Y is a binary variable coded as follows.

$$Y_i = \begin{cases} 0, & \text{if liquefaction exits in wellbore } i, \; i = 1, 2, \ldots, 314 \\ 1, & \text{otherwise.} \end{cases}$$

Three important factors that may affect the liquefaction are determined. These are, the Corrected Standard Penetration Test (CSPT) computed based on the correction factors according to the energy rates, well-hole size and covering stress for the stroke number obtained from the SPT experiments; the Cyclic Stress Ratio (CSR) that is a proportional expression related to the active and the total stresses resulted from the earthquake and the Factor of Safety (FS) that is a value obtaining by dividing the CSR values required for liquefying the soil to the CSR values resulting from the earthquake. These continuous variables are represented by X_1, X_2 and X_3, respectively.

We do not intend to focus on the interpretations of the results of the liquefaction data here. We only aim to give some remarkable results indicating the model quality and their interpretations. Additionally, the general discussion on the simulation results and the validity and the reliability of them in practice are discussed here.

5.1. The results of the liquefaction data. The variable CSPT is determined as a fixed parameter. The optimal h is 0.38357. The logistic regression model is significant at the $\alpha = 0.05$ level. The chi-square value is 133.432 and the corresponding probability is $p = 0.00$. The parameter estimates of the liquefaction data are given in Table 5.

It is evident from Table 5 that the parameter estimates of the logistic regression gradually approximate to the semiparametric estimates, and the closest results arise in the event that the optimal bandwidth value is used in the semiparametric approach. This once again emphasizes the importance of the determination of the optimal h.

In summary, we conclude that the simulation results are largely consistent with the results in practice. This indicates that the semiparametric approach could be directly applied to the data set if we are not certain about the validity of the parametric logistic regression model on condition that the optimal bandwidths are correctly specified.

Table 5. Parameter estimates of the liquefaction data obtained by logistic regression and DWADE

	LOGIT True Model	DWADE				
	$n = 50$					
		$h_1 = 0.18357$	$h_2 = 0.28357$	$h_3 = 0.38357$ **(optimal)**	$h_4 = 0.48357$	$h_5 = 0.58357$
Intercept	-3.7107	-	-	-	-	-
$\hat{\beta}_1$ (fixed)	1	1	1	1	1	1
$\hat{\beta}_2$	**10.145**	-9.0514	0.30208	**9.0853**	17.045	23.432
$\hat{\beta}_3$	**-18.12**	-15.743	-13.995	**-13.596**	-13.034	-12.303

Appendix 1. The XploRe commands for the DWADE method and their descriptions

`proc(b) = main1()`

`dat=read ("dwade1");` Reads the data set labeled "dwade1" written in ASCII format.

`y=dat[,3];` Describes the column number of the dependent variable y in the data set

`x=dat[,1:2];` Describes the column numbers of the continuous explanatory variable(s) x in the data set.

`x=x.-mean (x);` Centralizes x values for eliminating the high correlation.

`ozdeg=eigsm (cov (x));` Calculates the eigenvalues and eigenvectors of the covariance matrix of x.

`v=ozdeg.vectors;` Expresses the eigenvectors by the matrix "v".

`w=ozdeg.values;` Expresses the eigenvalues by the matrix "w".

`mah=v*(sqrt (1./w).*v');` Applies the Mahalanobis transformation to the values of the explanatory variables for eliminating the possible high correlation among them.

`x=x*mah;` Weights raw data matrix x by the transformation matrix "mah".

`library ("smoother");` Calls the "smoother" library for the estimation of β.

`library ("metrics"):` Calls the "metrics" library for the mathematical computations.

`h= 0.3848*matrix(cols(x));` Describes the optimal bandwidth values estimated by the method of the least square cross-validation required for the estimation of β.

`b=dwade (x,y,h);` Gives the semiparametric estimation of β by the method DWADE .

`b=mah*b:` Computes the original values of the estimations.

`b=b./(b [1,]);` Normalizes all estimated **b** s' by dividing them to the first estimated coefficient. This normalization is required for the comparison of the estimated parameters of the parametric logistic regression model and its semiparametric alternative.

`indexdw=x*b;` Gives the linear index estimation of observation i.

`write(indexdw, "output1.xls");` Writes the linear index values obtained from the method dwade to the file "output 1" in the xls format.

`endp`

`main 1()`

Appendix 2. The XploRe commands for the logistic regression model and their descriptions

`proc(b)=main2()`

`dat=read ("logistic1");` Reads the data set “logistic” written in ASCII format.

`y=dat[,3];` Describes the column number of the dependent variable y in the data set.

`x=dat[,1:2];` Describes the column numbers of the continuous explanatory variable(s) x in the data set.

`x = matrix (rows (x))` $\sim$ `x[,1:2];` Adds column vector “1” to the left side of the matrix x.

`library ("glm");` Calls the “glm” library for the estimation of β.

`g=glmest (‘‘bilo",x,y);` Applies the logistic regression analysis to the data using the option “binomial logit” abbreviated by “bilo” by calling the command “glmest” in the library “glm”.

`glmout (‘‘bilo",x,y,g.b,g.bv,g.stat);` Describes the basic statistics in the output.

`g.b=g.b/(g.b[2,]);` Normalizes all estimated g.bs$'$ by dividing them to the first estimated coefficient as in the case of the “dwade”. Here, line 2 represents the coefficient of the first explanatory variable whereas line 1 denotes the constant term.

`index=x*g.b;` Computes the index values xb.

`prob=exp (index) / (1+exp (index));` Computes the probabilities of belonging to the category “1” coded in the dependent variable obtained from the logistic regression analysis.

`write (prob, "output2.xls");` Writes the probabilities obtained from the logistic regression analysis to the file “output 2” in the xls format.

`endp`

`main2()`

Appendix 3. Proof of Theorem 2.1

.

We obtain the following expressions by applying integration by parts.

$$u = [p(x)]^2; \quad dv = \frac{\partial \mathrm{E}(Y/x)}{\partial x}; \quad du = 2[p(x)]\frac{\partial p(x)}{\partial x} \text{ and } \int dv = \int \frac{\partial \mathrm{E}(Y/x)}{\partial x}$$
$$\Longrightarrow \; v = \mathrm{E}(Y/x)\, \underbrace{[p(x)]^2 \mathrm{E}(Y/x)\Big|}_{\text{“0”}} - \int \mathrm{E}(Y/x) 2p(x)\frac{\partial p(x)}{\partial x} \qquad (i)$$

Because we are assuming $p(x) = 0$ on the boundary areas of the support of x, the first term in (i) is zero and the following expression is obtained.

$$\begin{aligned} \mathrm{E}\left[W(x)\frac{\partial \mathrm{E}(Y/x)}{\partial x}\right] &= -2\int \mathrm{E}(Y/x)\frac{\partial p(x)}{\partial x}p(x)\,dx \\ &= -2\mathrm{E}\left\{\mathrm{E}(Y/x)\frac{\partial p(x)}{\partial x}\right\} = -2\mathrm{E}\left[Y\frac{\partial p(x)}{\partial x}\right] \end{aligned} \qquad (ii)$$

When δ is defined as $\delta = \mathrm{E}\left[W(x)\frac{\partial \mathrm{E}(Y/x)}{\partial x}\right]$, an efficient estimator of δ can be obtained by replacing p with its nonparametric estimator and replacing the expectation operator

E with the sample mean. The estimator of δ is given as,

$$\delta \to -\frac{2}{n}\sum_{i=1}^{n} Y_i \frac{\partial p_{ni}(x_i)}{\partial x} \text{ a.s. if } \mathrm{E}\left(\left|\frac{\partial p(x)}{\partial x}\right|\right) < \infty, \qquad (iii)$$

where $\{Y_i, X_i;\ i = 1, \dots, n\}$ represent the sample values of observation "i", and $\mathbf{p}_{ni}(x_i)$ is the estimator of the joint probability density function $p(x_i)$. Since the joint probability density function of X is used as a weight function, the resulting δ_n estimator is called the "Density Weighted Average Derivative Estimator" (DWADE).

In order to complete the estimation process in Equation (iii), the estimator of p should be defined. Kernel type estimators are widely used on account of their easier estimation procedures. The following estimator of $\mathbf{p}(x_i)$ is obtained by using the "leave-one-out" kernel density estimation method.

$$\mathbf{p}_{ni}(x) = \frac{1}{n-1}\sum_{j=1, j\neq i}\left(\frac{1}{h_n}\right)^{k} K\left(\frac{x - X_j}{h_n}\right). \qquad (iv)$$

Here, k denotes the dimension of X, K is a multivariate kernel function with k-dimensional component, and $\{h_n\}$ the series of bandwidth parameters. The function $\mathbf{p}_{ni}(x)$ possesses the standard properties of a kernel density estimator, such as being an efficient estimator of $\mathbf{p}(x)$. Moreover, $\frac{\partial \mathbf{p}(x)}{\partial x}$ can be efficiently estimated by $\frac{\partial \mathbf{p}_{ni}(x)}{\partial x}$. The formulation of $\frac{\partial \mathbf{p}_{ni}(x)}{\partial x}$ is given by

$$\begin{aligned}\frac{\partial \mathbf{p}_{ni}(x)}{\partial x} &= \frac{1}{n-1}\sum_{j=1, j\neq i}^{n}\left(\frac{1}{h_n}\right)^{k} K'\left(\frac{x - X_j}{h_n}\right)\left(\frac{1}{h_n}\right) \\ &= \frac{1}{n-1}\sum_{j=1, j\neq i}^{n}\left(\frac{1}{h_n}\right)^{k+1} K'\left(\frac{x - X_j}{h_n}\right),\end{aligned} \qquad (v)$$

where K' is the first order derivatives of K (gradient vector). The resulting DWADE estimator δ_n is obtained by substituting Equation (v) in Equation (iii):

$$\delta_n = -\frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n}\left(\frac{1}{h_n}\right)^{k+1} K'\left(\frac{X_i - X_j}{h_n}\right) Y_i. \qquad (vi)$$

It should be noted that by taking $W(x) = p(x)$, the right side of the denominator of Equation (vi) does not contain a density estimator or a random variable. The absence of randomness in the denominator constitutes the basis of the applicability and interpretability of δ_n [15].

References

[1] Aldrich, J. H. and Nelson, F. D. *Linear Probability, Logit and Probit Models* (Sage Publications, London, 1984).

[2] *Geological Research Report* (The General Directorate of Highways, Turkey, 2009).

[3] Hardle, W., Müller, M., Sperlich, S. and Werwatz, A. *Nonparametric and Semiparametric Models* (Springer-Verlag, New York, 2004).

[4] Hardle, W. *Applied Nonparametric Regression* (Cambridge University Press, Cambridge, 1990).

[5] Hardle, W., Klinke, S. and Müller, M. *XploRe Learning Guide* (MDtech, Springer-Verlag, New York, 1999).

[6] Hardle, W., Hlavka, Z. and Klinke, S. *XploRe Application Guide*, (e-book, MD Tech, Springer-Verlag, New York, 2003).

[7] Hardle, W., Klinke, S. and Turlach, B. A. *XploRe: An Interactive Statistical Computing Environment: Statistics and Computing* (Springer-Verlag, New York, 2007).

[8] Horowitz, J. L. *Semiparametric Methods in Econometrics* (Springer-Verlag, New York, 1998).

[9] Horowitz, J. L. and Hardle, W. *Testing a parametric model against a semiparametric alternative*, Econometric Theory **10**, 821–848, 1994.

[10] Hosmer, D. W. and Lemeshow, S. *Applied Logistic Regression* (John Wiley and Sons, New York, 1989).

[11] Ichimura, H. *Semiparametric least squares (sls) and weighted sls estimation of single-index models*, Journal of Econometrics **58**, 71–120, 1993.

[12] Law, A. M. and Kelton, W. D. *Simulation Modeling and Analysis* (McGraw-Hill International Series, Singapore, 2000).

[13] Manski, C. F. *Identification of binary response models*, Journal of the American Statistical Association **83**, 729–738, 1988.

[14] McCullagh, P. and Nelder, J. A. *Generalized Linear Models: Monographs on Statistics and Applied Probability* **37** (Chapman and Hall, London, 1989).

[15] Powell, J. L., Stock, J. H. and Stoker, T. M. *Semiparametric estimation of index coefficients*, Econometrica **57** (6), 1403–1430, 1989.

[16] Proença, I. and Werwatz, A. *Comparing Parametric and Semiparametric Binary Response Models* (Humboldt University, Berlin, 1994).

[17] Proença, I. and Silva, S. *Parametric and semiparametric specification tests for binary choice models. A comparative simulation study*, Econometrics, Econ WPA, No: 0508008, 2005.