Gazi University Journal of Science
GU J Sci
24(4):753-762 (2011)

GUJS

XploRe Package for the Popular Parametric and the Semiparametric Single Index Models

Özge AKKUŞ♠

[1]*Muğla University, Department of Statistics, 48000, Muğla, TURKEY*

Received: 28.02.2011 Revised: 14.06.2011 Accepted: 16.06.2011

ABSTRACT

This study introduces and shows the applicability of the XploRe commands of the parametric and the semiparametric single index models, which are two most popular alternatives of each other. The commands required for the estimation in all stages of the semiparametric estimation and the parametric logistic, probit and complementary log log regression models are introduced in detail. An artificial data set is used to demonstrate the applicability of the commands in practice. The major contribution of this study is that it enables researchers to obtain additional outputs in easier way that are not so easy to have in the standard statistical packages especially for the semiparametric models. Additionally, users could extend and adapt these commands in conjunction with the new developments in this area.

Key Words: *Single index model, Semiparametric approach, Binary response modelling, XploRe.*

1. INTRODUCTION

When the dependent variable Y is binary, the conditional expectation function gives the probability of belonging to the category "1" in the dependent variable conditional on the explanatory variables X. The model is generally defined as,

$$\begin{aligned} E(Y/X=x) &= P[Y=1/X=x] \\ &= D\{\upsilon(x)\} = D(X^T\beta) \\ &= D(\beta_0 + X^T\beta) \end{aligned} \qquad (1)$$

where β is the unknown parameter vector and D is the distribution function related to the unobserved error term (also called link function between the probability $P[Y=1/X=x]$ and the index function $\upsilon(x)$). The linear form ($X^T\beta$) is usually determined for $\upsilon(x)$ by taking into consideration the simplicity of the model estimation under linearity. Since all explanatory variables are summed under only one linear index, these models are called "Single Index Models" (SIM).

♠Corresponding author, e-mail: ozge.akkus@mu.edu.tr

It is a well known fact that the parametric and the semi-parametric approaches are two most popular alternatives for the model estimation of binary responses. Binary Logit, Probit and Complementary log log models are widely used parametric models estimated based on the Maximum Likelihood Estimation (MLE) procedure whereas Density Weighted Average Derivative Estimator (DWADE) is one of the most popular estimators used in the semiparametric modelling.

In the current study, I mainly aimed to introduce the commands for the estimation of both parametric and semiparametric models according to the different estimators in the windows based version 4.8 of the XploRe package. It is evident that these commands will provide easier way to the estimation procedure by means of their some available commands in the form of libraries and quantlets (single XploRe programs). Additionally, they enable researchers to obtain additional outputs in easier way that are not at hand in the existing standard statistical packages.

In the application part of the study, all the commands are executed and the estimation of the model parameters is obtained over an artificial data set. In this way, the applicability of the commands in practice is supported.

2. THE METHODOLOGY

Because I mainly intended to introduce the XploRe commands I constructed for the parametric and the semiparametric single index models, a brief methodology of them is discussed in the following subsections.

2.1. The Parametric Single Index Model

In the standard parametric model, the function D, defined in Eq.(1), is assumed a known distribution function (denoted by $G_{\varepsilon/X} = G$) and ε is distributed independently of X. The model given by

$$\begin{aligned} E(Y/X=x) &= P[Y=1/X=x] = G\{\upsilon(x)\} \\ &= G(X^T\beta) = G(\beta_0 + X^T\beta) \end{aligned} \tag{2}$$

is called the Parametric Single Index Model (PSIM). If G is correctly specified, this approach satisfies the efficiency condition of the model parameters and allows for the extrapolation for the values x out of the support of X. However, G is rarely known in most applications and the results are highly misleading when G is misspecified. The name of the model changes in conjunction with the change in G. The parametric Logit and Probit models are obtained by using the link functions given in Eq.(3) and (4), respectively ([1], [10]).

$$\begin{aligned} E(Y/X=x) &= P[Y=1/X=x] \\ &= \frac{\exp(X^T\beta)}{1+\exp(X^T\beta)} \end{aligned} \tag{3}$$

$$E(Y/X=x) = P[Y=1/X=x] = \Phi(X^T\beta) \tag{4}$$

Here, Φ denotes the standard normal cumulative distribution function.

The complementary log-log model that is the third alternative to the binary logit and probit models is frequently used when the probability of an event is very small or very large. The major difference of the complementary log-log function given in Eq.(5) from the logit and probit is resulted

from its asymmetrical structure.

$$\begin{aligned} E(Y/X=x) &= P[Y=1/X=x] \\ &= 1-\exp[-\exp(X^T\beta)] \end{aligned} \tag{5}$$

2.1.1. The maximum likelihood estimator

In the PSIM, parameter estimates are obtained by the method of MLE. The likelihood and the logarithmic likelihood functions for the MLE of β are given by Eq.(6) and (7), respectively.

$$L(\beta/y,x) = \prod_{i=1}^{N} p_i^{Y_i}(1-p_i)^{1-Y_i} \tag{6}$$

The calculations become much easier when sums are used instead of products in Eq.(6). In this case, the general form of the logarithmic likelihood function given below should be identified.

$$\log L(\beta/y,x) = \sum_{i=1}^{N}\begin{bmatrix} Y_i \log p_i \\ +(1-Y_i)\log(1-p_i) \end{bmatrix} \tag{7}$$

We know that if $\hat{\beta}$ maximizes $L(\beta/y,x)$, it also maximizes $\log L(\beta/y,x)$. Hence, the first order derivatives are computed and set to "0" to obtain estimates maximizing the likelihood of observing the sample Y.

The logarithmic likelihood equations with respect to the logistic, probit and complementary log log models are obtained by replacing p_i in Eq.(7) with the functions given by Eq.(3), (4) and (5), respectively.

$$\log L(\beta/y,x) = \sum_{i=1}^{N}\begin{Bmatrix} Y_i \log\left[\dfrac{\exp(X_i^T\beta)}{1+\exp(X_i^T\beta)}\right] \\ +(1-Y_i)\log(1-\dfrac{\exp(X_i^T\beta)}{1+\exp(X_i^T\beta)}) \end{Bmatrix} \tag{8}$$

$$\log L(\beta/y,x) = \sum_{i=1}^{N}\begin{Bmatrix} Y_i \log\Phi(X_i^T\beta) \\ +(1-Y_i)\log\left[1-\Phi(X_i^T\beta)\right] \end{Bmatrix} \tag{9}$$

$$\log L(\beta / y, x) = \sum_{i=1}^{N} \left\{ \begin{array}{l} Y_i \log\left[1-\exp(-\exp(X_i^T\beta))\right] \\ +(1-Y_i) \\ \log\left\{1-\left(1-\exp(-\exp(X_i^T\beta))\right)\right\} \end{array} \right\} \quad (10)$$

Unknown β parameters are estimated by maximizing the log likelihood functions given above ([1], [8], [10]).

2.2. The Semi-Parametric Single Index Model

Most estimation problems contain both an unknown finite-dimensional parameter (β) and an unknown link function. These kinds of models are called "Semi-Parametric".

In the Semiparametric Single Index Model (SSIM), the linearity assumption $\upsilon(x) = X^T\beta$ is still valid but no additional assumption is made related to the error term. In other words, a specific link function is not assumed in the model and represented by the term "g" instead of "G" in Eq.(2). SSIM is defined as follows.

$$\begin{aligned} E(Y / X = x) &= P[Y = 1 / X = x] \\ &= g\{\upsilon(x)\} = g(X^T\beta) \\ &= g(\beta_0 + X^T\beta) \end{aligned} \quad (11)$$

The estimation procedure of SSIM is composed of two steps. In the first step, the parameter vector β is estimated using one of the semi-parametric model estimation techniques according to the data structure such as DWADE introduced in detail in Subsection 2.2.1. In the second step, the values of the linear index function $X^T\hat{\beta}$ are computed. Finally, an unknown distribution function g is estimated and probabilities $P[Y = 1 / X = x]$ are obtained by the non-parametric regression of Y on $X^T\hat{\beta}$, which is introduced in Subsection 2.2.2. ([2], [7]).

2.2.1. The density weighted average derivative estimator

The DWADE has two important advantages except that it could only be applied to the data with continuous explanatory variables. The first one is that no distributional assumption is needed for the dependent variable Y and the second one is that the resulting estimator is a direct estimator. It is based on the average derivatives of the conditional expectation function expressed in Eq.(12) with respect to the continuously distributed random vector x.

$$\frac{\partial E(Y / x)}{\partial x} = \beta G'(x^T\beta) \quad (12)$$

The density weighted average derivatives are obtained by using the probability density function of x for any restricted and continuous function W.

$$\begin{aligned} &E\left[W(x)\frac{\partial E(Y / x)}{\partial x}\right] \\ &= \beta E\left[W(x)G'(x^T\beta)\right] \end{aligned} \quad (13)$$

Because of the scale normalization requirement of the semiparametric approach, β is only defined according to the scale and any weighted average derivative of $E(Y / x)$ is equal to the β ([9],[12]).

The scale normalization of $\beta_1 = 1$ can be achieved by dividing each component on the left side of Eq.(12) by the first component. The left side of Eq. (12) could be estimated by replacing the kernel estimator of $\frac{\partial E(Y / x)}{\partial x}$ and the sample mean for the expected value of the population $[E(.)]$. The resulting estimator proposed by [12] is defined as,

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

Here, $p(x)$ denotes the joint probability density function of the random vector x. When we accept the equality $W(x) = p(x)$; we conclude that $W(x) = p(x) = 0$ when x is on the boundary of the support of x. In such a case, we could easily obtain the following partial integration.

$$\begin{aligned} E\left[W(x)\frac{\partial E(Y / x)}{\partial x}\right] &= -\frac{2}{n(n-1)} \\ \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n}\left(\frac{1}{h_n}\right)^{k+1} &K'\left(\frac{X_i - X_j}{h_n}\right)Y_i \end{aligned} \quad (15)$$

In Eq.(15), K' is the first order derivatives of the kernel function K; n is the number of observations and h_n is the bandwidth parameter, dependent on n, required for the kernel estimation. (refer to [12] for detailed theoretical information and the proof of Eq.(14) and (15)).

2.2.2. The theory of the nonparametric regression

The estimation of the conditional expectation function given in Eq.(1) is defined as,

$$\hat{m}(x) = E(Y|x) = \int \frac{y\hat{f}(x,y)}{\hat{f}(x)}dy \quad (16)$$

where $\hat{f}(x,y)$ and $\hat{f}(x)$ represent the estimated joint probability density function of X and Y and the marginal density function of X, respectively. The estimators of the regression function proposed by [11]

and [14] are obtained when the kernel type estimators developed based on the bandwidth parameter h and the kernel functions are used in the estimation of $f(x,y)$ and $f(x)$. This requirement arises when no distributional assumption is made related to these two density functions.

$$\hat{f}(x) = \frac{1}{nh_X}\sum_{i=1}^{n} K(\frac{x - X_i}{h_X}) \qquad (17)$$

$$\hat{f}(x,y) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_X h_Y} \ddot{K}(\frac{x - X_i}{h_X}, \frac{y - Y_i}{h_Y}) \qquad (18)$$

Here, K is called a "kernel function". It is a symmetrical probability density function satisfying the following general assumptions.

1. $\int K(u)du = 1,$
2. $\int u\,K(u)du = 0,$
3. $\int u^2 K(u)du = \mu_2(K) \neq 0$

$\ddot{K}(...)$ is a bivariate kernel function, h_X is a fixed bandwidth of a random variable X, and h_Y is a fixed bandwidth of a random variable Y. A bivariate kernel function could be obtained by using the multiplicative kernel functions defined as follows ([3],[11],[14]).

$$\ddot{K}(\frac{x - X_i}{h_X}, \frac{y - Y_i}{h_Y}) = K_1(\frac{x - X_i}{h_X}) K_2(\frac{y - Y_i}{h_Y}) \qquad (19)$$

The kernel estimator of the bivariate probability density function of X and Y is obtained by using the same K_1=K_2=K kernel function in the estimation. The Nadaraya-Watson kernel estimator of the regression function ($\hat{m}_{NW}(x)$) is obtained by replacing the density functions $\hat{f}(x,y)$ and $\hat{f}(x)$ in Eq. (16) with the kernel estimators given by Eq.(17) and (18).

$$\hat{m}_{NW}(x) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{x - X_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)} \qquad (20)$$

Epanechnikov (K_1) and Gaussian (K_2) are two most popular kernel functions used in practice. The defination of them is given as follows.

$$K_1(u) = 3(1 - u^2)/4, \quad |u \le 1|$$

$$K_2(u) = \exp(-u^2/2)/\sqrt{2\pi}, \quad -\infty < u < \infty$$

The bandwidth parameter, also called the smoothing parameter, h of the Nadaraya-Watson kernel estimator controls for the smoothing level of the estimation. h plays very important role in the performance of the kernel estimators. Various methods such as the cross-validation, penalized functions, plug-in, bootstrap etc. have been developed to be able to obtain the optimal bandwidths. The cross-validation method is generally preferred due to its easier computable and applicable structure for any regression model.

The bandwidth value which minimizes the Cross-Validation (CV) function with a nonnegative weight function $w(X_i)$ given as,

$$CV(h) = n^{-1}\sum_{i=1}^{n}\left[Y_i - \hat{m}_i(X_i)\right]^2 w(X_i) \qquad (21)$$

is considered the optimal one. The CV function contains the leave-one-out kernel estimator defined as follows.

$$\hat{m}_i(X_i) = \frac{\sum_{j\neq i}^{n} Y_j K(\frac{X_i - X_j}{h})}{\sum_{j\neq i}^{n} K(\frac{X_i - X_j}{h})} \qquad (22)$$

The leave-one-out estimator is obtained by leaving out the observations i (the concerned observations X_i and Y_i) from the data each time for satisfying the unbiased estimate of the bandwidth parameter h. The procedure is replicated n times (for all observations). The final optimal bandwidth value required for the kernel estimation is the mean of all these values computed. The bandwidth that minimizes the cross-validation function also minimizes the mean square error which is a performance criterion of an estimator ([3],[11],[14]).

3. XploRe COMMANDS FOR THE SINGLE INDEX MODEL ESTIMATION

In this section, all XploRe commands for the popular parametric and the semiparametric models for binary response data are introduced in detail.

3.1. The Commands of the Parametric Single Index Models

The XploRe commands for the estimation of the parametric logistic, probit and complementary log log models are given below ([4],[5],[6]).

3.1.1. The logistic regression model

proc(b)=main1()

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

y=dat[,1] ; Describes the column number of the dependent variable (y) in the data set.

x=dat[,2:5] ; Describes the column numbers of the explanatory variable(s) (x) in the data set.

x = matrix (rows (x)) ~ x[,1:4] ; Adds column vector "1" to the left side of the matrix x.

library ("glm") ; Calls the library "glm" (**g**eneralized **l**inear **m**odel) for the estimation of β.

g=glmest ("bilo",x,y) ; Applies the logistic regression analysis to the data using the option "**bi**nomial **lo**git" abbreviated by "bilo" by calling the command "glmest" (**g**eneralized **l**inear **m**odel **e**stimation) in the library "glm".

glmout ("bilo",x,y,g.b,g.bv,g.stat) ; Creates outputs of the logistic regression model by the command "glmout" (**g**eneralized **l**inear **m**odel **out**put). "g.b"; "g.bv" and "g.stat" include the estimated parameter vector b; variance-covariance matrix of b and some basic statistics such as the logarithmic likelihood value, degrees of freedom, residuals, some information criterion etc., respectively.

index1=x*g.b ;Computes the linear index values $\upsilon(x) = x^T(g.b)$.

index 1 ; Displays the index values on the output screen.

prob1=exp (index1) / (1+exp (index1)

;Computes the probabilities of belonging to the category "1" coded in the dependent variable obtained from the logistic regression analysis in connection with the linear index values.

prob1; Displays the calculated probabilities related to the each observation on the output screen.

write1 (prob1, "output1.xls") ;Writes the probabilities obtained from the logistic regression analysis to the file "output 1" in the xls format.

endp

main1()

3.1.2. The probit regression model

proc(b)=main2()

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

y=dat[,1] ; Describes the column number of the dependent variable (y) in the data set.

x=dat[,2:5] ; Describes the column numbers of the explanatory variable(s) (x) in the data set.

x = matrix (rows (x)) ~ x[,1:4] ; Adds column vector "1" to the left side of the matrix x.

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

g=glmest ("bipro",x,y) ; Applies the probit regression analysis to the data using the option "**bi**nomial **pro**bit" abbreviated by "bipro" by calling the command "glmest" in the library "glm".

glmout ("bipro",x,y,g.b,g.bv,g.stat) ; Creates outputs of the probit regression by the command "glmout".

index2=x*g.b ; Computes the index values $\upsilon(x) = x^T(g.b)$.

index 2; Displays the index values on the output screen.

prob2=cdfn(index2) ; Calculates the probability of belonging to the category "1" coded in the dependent variable. The command "cdfn" stands for "**c**umulative **d**istribution **f**unction of **n**ormal distribution", which is the link function of the probit model.

prob2; Displays the calculated probabilities related to the each observation on the output screen.

write2 (prob2, "output2.xls"); Writes the probabilities obtained from the probit regression analysis to the file "output 2" in the xls format.

endp

main2()

3.1.3. The complementary log log model

proc(b)=main3()

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

y=dat[,1] ; Describes the column number of the dependent variable (y) in the data set.

x=dat[,2:5] ; Describes the column numbers of the explanatory variable(s) (x) in the data set.

x = matrix (rows (x)) ~ x[,1:4] ; Adds column vector "1" to the left side of the matrix x.

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

g=glmest ("bicll",x,y) ; Applies the complementary log log regression analysis to the data using the option "**bi**nomial **c**omplementary **l**og **l**og" abbreviated by "bicll" by calling the command "glmest" in the library "glm".

glmout ("bicll",x,y,g.b,g.bv,g.stat); Creates outputs of the complementary log log model by the command "glmout".

index3=x*g.b ; Computes the index values $\upsilon(x) = x^T(g.b)$.

index3; Displays the index values on the output screen.

prob3= $1-\exp(-\exp(\text{index3}))$; Calculates the probability of belonging to the category "1" coded in the dependent variable.

prob3; Displays the calculated probabilities related to the each observation on the output screen

write3 (prob3, "output3.xls") ; Writes the probabilities obtained from the probit regression analysis to the file "output 3" in the xls format.

endp

main3()

3.2. The Commands of the Semiparametric Single Index Model

In this subsection, the XploRe commands related to the DWADE estimator used in the first step of the semiparametric modelling and the nonparametric regression that constitutes the second step of the estimation procedure are introduced.

The commands related to the DWADE in an old version of XploRe package are written by [13], however, these commands could not be used in the current windows based version of the XploRe.

proc(b) = main4()

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

y=dat[,1] ; Describes the column number of the dependent variable (y) in the data set.

x=dat[,2:5] ; Describes the column numbers of the explanatory variable(s) (x) in the data set.

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

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 library "smoother" for the estimation of β .

library ("metrics") ; Calls the library "metrics" for the mathematical computations.

library("plot") ; Calls the library "plots" for the graphical representation.

h=0.2*(max(x).-min(x))' ; Describes the optimal bandwidth values required for the estimation of β .

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

b=mah*b ; Computes the original values of the b estimates.

b=b./abs(b[1,]) ; Normalizes all estimated b s' dividing by the first estimated coefficient.

index4=x*b ; Gives the estimated linear index values.

index4; Displays all the calculated index values on the output screen.

write4 (index4, "output4.xls") ; Writes the linear index values obtained from the dwade to the file "output 4" in the xls format.

;the nonparametric regression of Y on the estimated index values and the graphical representation

yindex4=index4~y ; Adds the dependent variable column y to the right side of the variable "index4". This is required for the nonparametric regression method used for the second step of the semiparametric modelling.

h1=regxbwsel (yindex4) ; Selects the optimal bandwidth value for the nonparametric regression by the command "regxbwsel" (**reg**ression **b**and**w**idth **sel**ection).

mh=regxest(yindex4,h1,"qua") ; Computes the nonparametric estimates mh by the command "regxest" (**reg**ression **est**imate) dependent on the values of y and index4, optimal bandwidth value h1 and the kernel function used (here the quadratic kernel function abbreviated by "qua" is preferred).

mh ;Displays the nonparametric regression estimates on the output screen.

mh=setmask(mh,"line","blue"); Describes some diagrammating characteristics such as the color and the shape.

xy=setmask(yindex4,"cross","small")

; Describes the image of the "yindex4".

plot(xy,mh) ; Plots "xy" and "mh".

endp

main4()

4. A NUMERICAL EXAMPLE

The applicability of all the XploRe commands I constructed above in practice are supported over an artificial data derived according to sample size of 100 in the linear index functional form of

$$(X\beta)_{(i)} = \text{index}_{(i)} = 1 + X_{1(i)} + X_{2(i)} + X_{3(i)} - 3X_{4(i)} \quad ;i = 1,2,\ldots,n \tag{23}$$

where n denotes the number of observations and X represents the vector of the explanatory variables. It should be taken into consideration that the estimation of the constant term is not required and the first coefficient

of a continuous explanatory variable of the linear index function should set to "1" to satisfy the identifiability condition of the parameters in the semiparametric modelling (for details, see [9]).

X_1 X_2, X_3 and X_4 are assumed to follow a Standard Normal, Uniform (0,1), Exponential (3) and Weibull (6,2) distributions, respectively. The dependent variable Y_i is assumed to follow a Bernoulli distribution with the parameter $p_{(i)}$, which is the probability of belonging to the category "1" coded in the dependent variable of observation i. $p_{(i)}$ is assumed to be calculated based on the logistic regression function given below.

$$p_{(i)} = \frac{\exp\left[\text{index}_{(i)}\right]}{1+\exp\left[\text{index}_{(i)}\right]} \qquad (24)$$

The outputs obtained by executing all the XploRe commands given above are discussed in the following subsections.

It should be again noted that I do not focus on to the interpretations of the model results here. I mainly intended to introduce the XploRe commands of the most popular SIM that enable to display the required additional outputs on the screen, which are not so easy in the standard statistical packages.

Researchers interesting in the interpretations to the model results could refer to [1],[2],[7],[8] and [10].

4.1. The Results of the Logistic Regression Model

The results related to the logistic regression model are expressed by Figure 1 and 2.

Estimated parameters (b), standard error of the estimate (s.e) and related t values, other statistics measuring the quality of the model such as the log-likelihood value and some residual types such as Pearson and Deviance are presented on the left side of Figure1. The graphical representation of the link function "logit" is given on the right side of the figure.

Figure 2 shows the output screen related to the computed index values (xb) and the probabilities of belonging to the category "1" coded in the dependent variable.

Further details could be obtained by extending and adapting the existing commands.

```
====================================
 GLM fit, 'bilo', n=100
====================================

Estimates (b, s.e., t-value)

 b[1]     10.5047 4.191       2.51
 b[2]     1.40636 0.6891      2.04
 b[3]    -2.46795 2.444      -1.01
 b[4]      2.2705 0.8546      2.66
 b[5]    -10.5247 3.627      -2.90

 * constant variable: b[1]

Statistics

 df                        95
 Deviance             23.6591
 Log-Likelihood      -11.8295
 Pearson              27.0558
 R^2                   0.8113
 adj. R^2              0.8033
 AIC                  33.6591
 BIC                  46.6849
 iterations                8
 distinct obs.           100
```

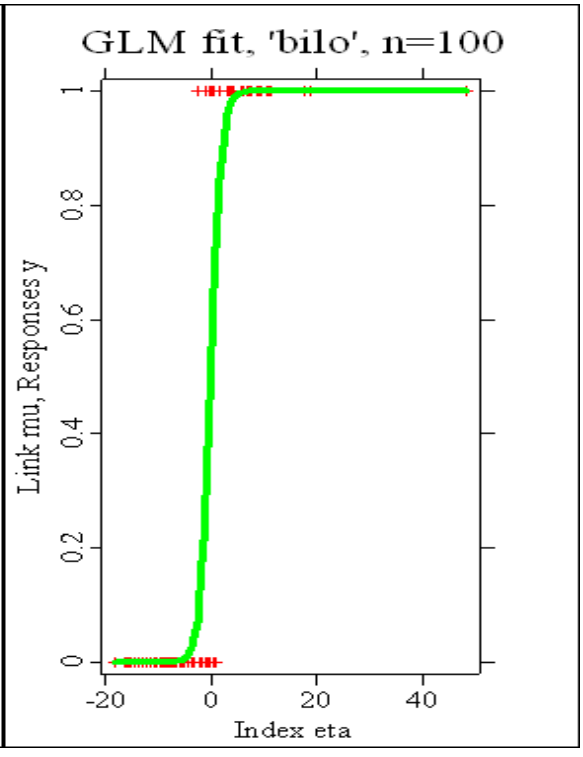
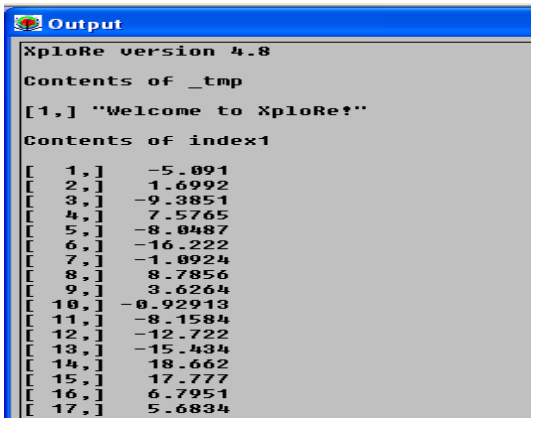


Figure 1. Estimated parameters and some basic statistics related to the logistic regression model.



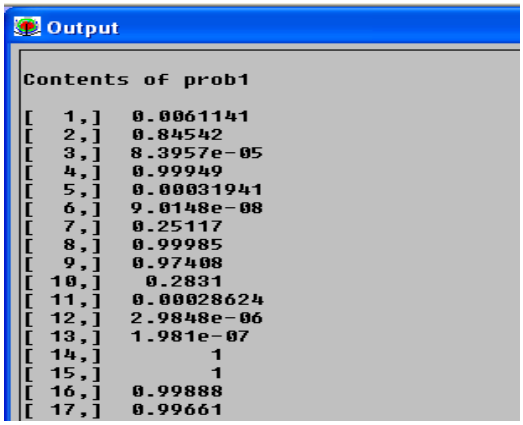


Figure 2. Estimated index values and probabilities in the logistic regression model.

4.2. The Results of the Probit Regression Model

The similar outputs with respect to the probit regression model as in the case of the logistic regression could be obtained. Therefore I only gave the basic results by Figure 3. It is evident that the results obtained from the logistic and probit regression models are parallel with each other due to the fact that they only differ from the link functions used in the terminology.

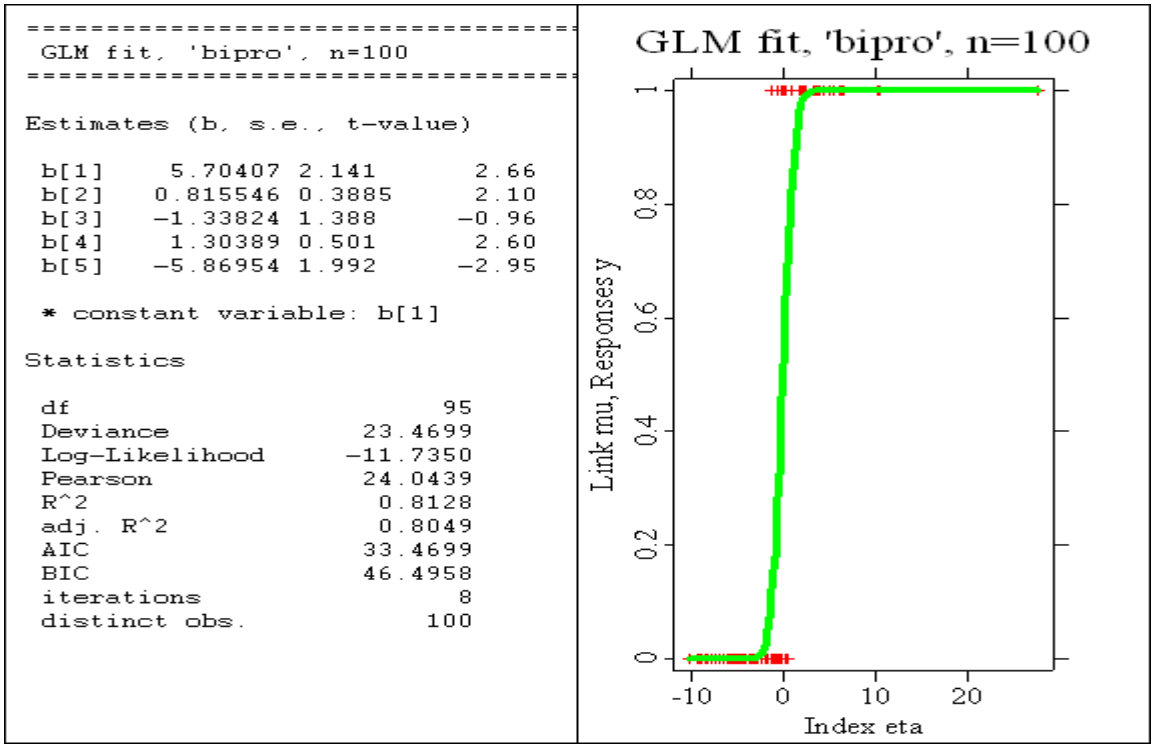


Figure 3. Estimated parameters and some basic statistics related to the probit regression model.

4.3. The Results of the Complementary Log Log Model

The results of the third alternative, the complementary log log model, to the logistic and the probit regression models are given below in the same form as the other two parametric SIM models.

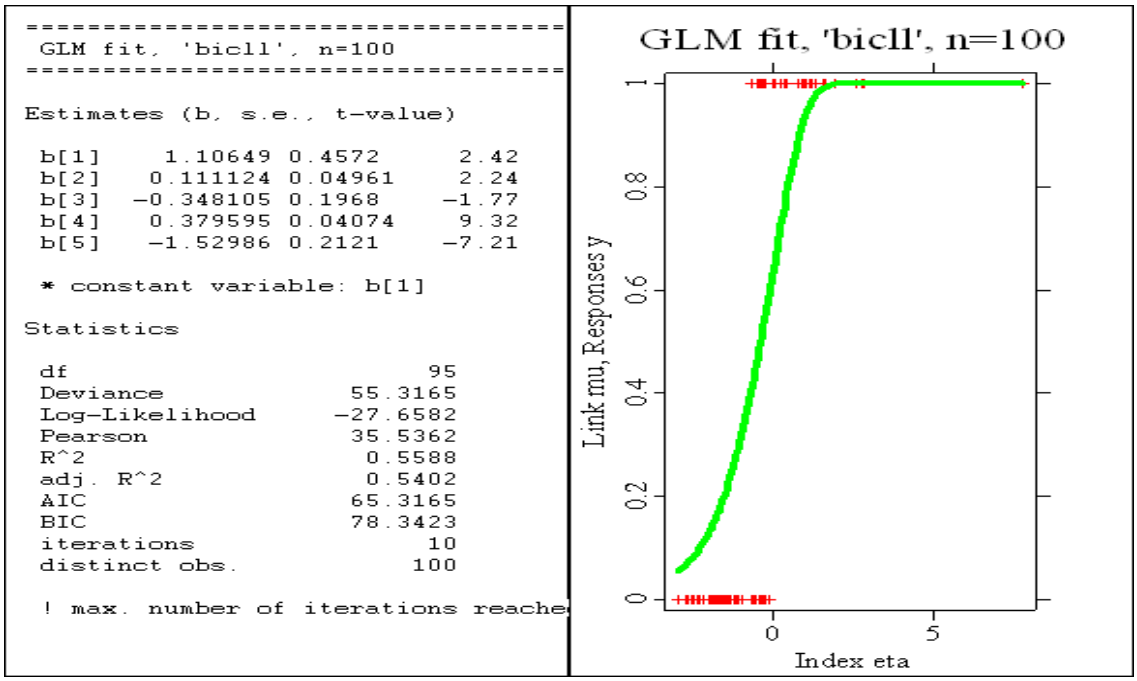


Figure 4. Estimated parameters and some basic statistics related to the complementary log log model.

4.4. The Results of the Semiparametric SIM

The output screen on the left side of Figure5 appears by executing the first part of the commands given in Subsection 3.2. It shows the optimal bandwidth values (h), the estimated parameters (b) and the related index values (index4) of the method DWADE in the semiparametric approach. The window on the right side of Figure5 appears when we run the command "regxbwsel" and enables users to choose one of the estimation methods of the parameter h1 such as Cross Validation, Shibata's Model Selector etc. required in the nonparametric estimation.

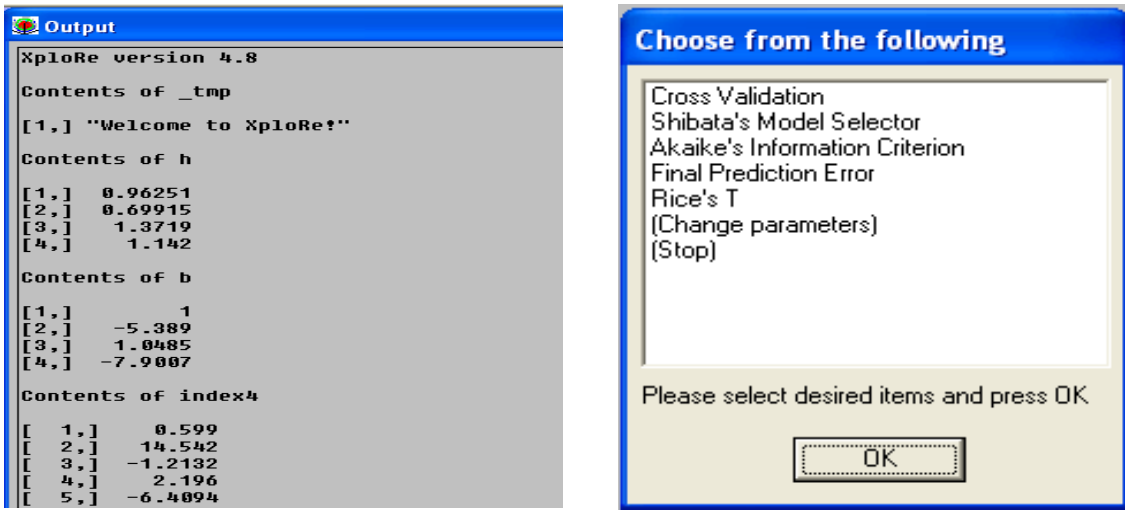


Figure 5. The results obtained by the method DWADE.

As mentioned above, the use of the method CV is commonly preferred in most studies due to its easier mathematical structure and the power in applications. The results obtained after selecting one of the methods are given by Figure6.

The bandwidths (h1's) giving the best results range between the values 1.23 and 19.68 and the optimal h1 is determined as 2.27765.

The left side of Figure6 graphs the values of h1 and the nonparametric regression (mh) of Y on the estimated index values for the optimal bandwidth parameter with the quadratic kernel function, which is optional in the commands.

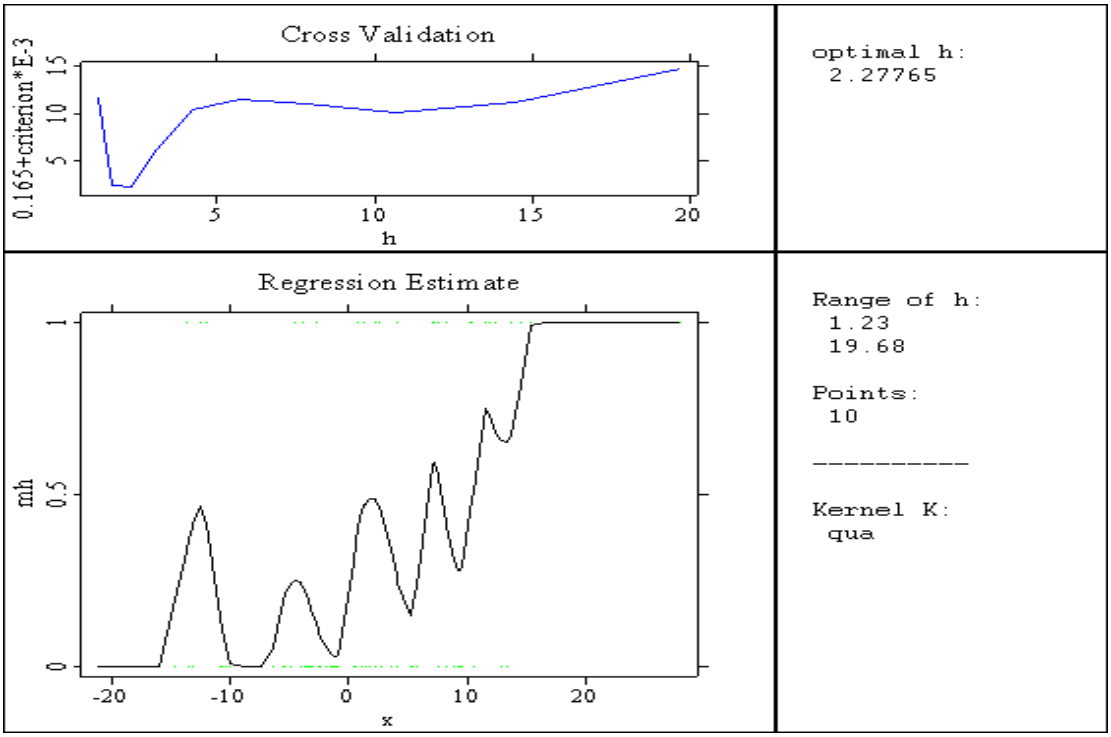


Figure 6. The results of the nonparametric regression applied in the second part of the semiparametric approach.

5. THE CONCLUDING REMARKS

Because the main aim of this study is to introduce the commands constructed for the estimation of the parametric and the semiparametric SIM in the new windows based version of the XploRe package, researchers studying in this area could refer to related references listed below for comprehensive interpretations of the results.

As emphasized above, the major contribution of this study is that it simplifies the estimation procedures of PSIM and SSIM and provides users to create their own outputs easily by some additions to the commands.

REFERENCES

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

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

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

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

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

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

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

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

[9] Manski, C. F., "Identification of binary response models", ***Journal of the American Statistical Association***, 83: 729-738 (1988).

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

[11] Nadaraya, E.A., "On estimating regression", ***Theory of Probability and its Applications***, 10: 186-190 (1964).

[12] Powell, J.L., Stock, J.H., Stoker, T.M., "Semiparametric estimation of index coefficients", ***Econometrica***, **57**(6): 1403-1430 (1989).

[13] Proença, I., Werwatz, A., "Comparing parametric and semiparametric binary response models", Humboldt University, Series: Sonderforschungsbereich 373, Number: 1995-36, Berlin, (1994).

[14] Watson, G.S., "Smooth regression analysis", Sankhya, 26(A): 359-72 (1964).