ORIGINAL ARTICLE

GUJS

How to Test the Parametric Fit of the Complementary Log Log Model to the Data?

Özge AKKUŞ♠

Muğla University, Department of Statistics, 48000, Muğla, TURKEY

ABSTRACT

In this study, commands written on the basis of Uniform Confidence Bands (UCB) in the new windows version of the XploRe package for testing the validity of the link function assumed for binary parametric complementary log log model were introduced. Models including only continuous and both continuous and discrete (mixed) explanatory variables cases were discussed, separately. I intended here to present researchers an easier way for testing the accuracy of the assumed parametric complementary log log model for the data that could not be tested by the existing standard statistical packages. The applicability of the commands was shown over an artificial and a real data set on gastric cancer.

Key Words: *Complementary log log model; model validity test; binary dependent variable; semi-parametric approach; uniform confidence band.*

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

1. INTRODUCTION

Parametric modelling is based on the validity of the model assumptions related to the error term in binary response models. However, the validity tests of the assumptions are ignored in most applications. In such a case, the use of the semiparametric approach that requires less assumption compared with the parametric alternative is proposed.

Because the application and the interpretation of the results are easier in the parametric approach, testing the validity of the model assumptions is the most important point to be taken into consideration in the first step of the data modelling process. There is no user friendly statistical package testing the parametric model assumptions for binary responses. Therefore, I generated easily applicable and optional new commands here on the basis of the UCB limits in the statistical software package XploRe

This study focuses on testing the validity of the binary parametric complementary log log model assumptions.

Some related studies in the literature are briefly reviewed below.

Hardle (1988), Klinke et.al. (1997), Hardle et.al. (2003) and Hardle et.al. (2007) introduced the use of the XploRe package in detail.

Müller et.al. (1997) tested the validity of the generalized partially linear model against the parametric generalized linear model over a data set on East-West German Migration using the commands written in the XploRe package.

Proença and Werwatz (1994) introduced the commands written for testing the parametric logit model assumptions including only continuous explanatory variables in an old MsDos version of the XploRe package.

The current study may be considered as an updated and extended version of the study of Proença and Werwatz (1994) with four contributions.

The first one is that it updates the commands written in an old MsDos format to the windows based 4.8 version of the XploRe package. Therefore, testing the parametric model assumptions becomes easier.

The second one is related to the modification of the commands written for the binary logit model to the binary parametric complementary log log model.

The third one extends the commands restricted for the model including only continuous explanatory variables to the model with mixed (both discrete and continuous) explanatory variables.

The last one enables some parts of the commands to be optional and provides the user the chance to obtain alternative results according to the kernel function used (Gaussian, Epanechnikov etc.), the confidence level for the UCB limits (0.05, 0.10 etc.) and the bandwidth selection methods (Cross validation, AIC, etc.). Therefore, the commands become more flexible.

2. THE THEORETICAL BACKGROUND

In binary dependent variable modelling, the mean function that is conditional on the vector of the explanatory variables X given in Eq.(1) expresses the probability of belonging of observations to the category "1" coded in the dependent variable.

$$E(Y \,/\, X = x) = P[Y = 1 \,/\, X = x] \qquad (1)$$

The fully parametric and the semiparametric approaches are used for the estimation of the model parameters.

2.1. The Parametric Approach

There are finite number of estimates of the β parameter vector and the linear index restriction is accepted ($X\beta$) in the parametric approach. The model is defined as,

$$E(Y \,/\, X = x) = P[Y = 1 \,/\, X = x] = G(X\beta) \quad (2)$$

where G is a known function that represents the distribution of the error term. G or G^{-1} is called link function in the family of the generalized linear models. The name and the parameters of the distribution are also known. This approach is called the parametric approach due to the assumption of the linear functional form of the explanatory variables ($X\beta$).

The name of the models changes in conjunction with the change in G. The parametric complementary log log model is obtained by using the link function given in Eq. (3).

$$log\{-log[1-P(Y_i = 1)]\} = X\beta \qquad (3)$$

The probability expression of the model is defined as follows.

$$\begin{aligned} E(Y \,/\, X = x) &= P[Y = 1 \,/\, X = x] \\ &= 1 - exp[-exp(X\beta)] \end{aligned} \qquad (4)$$

Complementary log-log model is frequently used when the probability of an event is very small or very large. The major difference of the complementary log-log function from its alternatives such as logit and probit models is resulted from its asymmetrical structure.

The maximum likelihood estimation technique is used in the estimation of the model parameters (β's) (McCullagh and Nelder 1989; Wang and Dey 2008).

2.2. The Semiparametric Approach

The model estimation procedure is composed of two parts in the semiparametric modelling. In the first step, β is estimated and the linear index function $X\hat{\beta}$ is calculated. In the second step, an unknown function G (denoted by g) is estimated by the nonparametric regression of Y on the estimated linear index $X\hat{\beta}$.

The model is expressed as follows.

$$\begin{aligned} E(Y \,/\, X = x) &= P[Y = 1 \,/\, X = x] \\ &= g(X\hat{\beta}) \end{aligned} \qquad (5)$$

Various methods used according to the type of the explanatory variables have been developed for the estimation of β's. In this study, continuous and mixed explanatory variable cases of models are discussed, separately.

The XploRe commands in the estimation of β's are constructed according to the estimator of Powell et al. (1989) based on the Density Weighted Average Derivatives (DWADE). The command "dwade" is used in the estimation of the β's for the model with continuous explanatory variables whereas the "adedis" is used for the mixed model (Powell et al. 1989).

2.3. The Density Weighted Average Derivative Estimator of the Index Parameters

Since the DWADE is a direct estimator, the estimation of unknown parameters of the linear index function is easier compared with the other semiparametric estimators. It does not require iterative procedure but is considerably restrictive due to the fact that it can only be applied to the models including only continuous explanatory variables for satisfying the differentiability condition.

Assume that G is a differentiable function required for the identifiability of β, X is a continuously distributed random vector and W is a continuous function. The following two expressions are derived under these assumptions.

$$\frac{\partial E(Y/x)}{\partial x} = \beta G'(X\beta) \qquad (6)$$

$$E\left[W(X)\frac{\partial E(Y/X)}{\partial x}\right] = \beta E\left[W(X)\dot{G}(X\beta)\right] \qquad (7)$$

The left side of the Eq.(7) is called Weighted Average Derivative Estimator (WADE) with weight function W. The model given by Eq.(8) is called DWADE when the probability density function of X ($p(x)$) is determined as a weight function.

$$E\left[W(x)\frac{\partial E(Y/X)}{\partial x}\right] = E\left[p(X)\frac{\partial E(Y/X)}{\partial x}\right] \qquad (8)$$
$$= \int \frac{\partial E(Y/x)}{\partial x} p(x)^2 \, dx$$

It is evident in Eq.(8) that the DWADE of $E(Y/x)$ is proportional to the β. The scale normalization required in the semiparametric approach can be achieved by dividing each component on the left side of Eq.(8) by the first component of β ($\beta_1 = 1$).

The left side of Eq.(8) can be estimated by replacing the kernel estimator of $\dfrac{\partial E(Y/X)}{\partial x}$ and the sample mean for the population expected value [$E(.)$]. The resulting DWADE estimator weighted by $p(x)$ and denoted by δ is given as follows.

$$\delta = -\frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \left(\frac{1}{h_n}\right)^{k+1} K'\left(\frac{X_i - X_j}{h_n}\right) Y_i \qquad (9)$$

The weight function $p(x)$ is defined as,

$$p(x) = \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^{n} \left(\frac{1}{h_n}\right)^{k} K\left(\frac{x - X_j}{h_n}\right) \qquad (10)$$

where k denotes the dimension of X, K is a multivariate kernel function with k-dimensional component and h_n is the series of bandwidth parameters (Powell et al. 1989).

2.4. The Estimation Procedure of β's in the Model with Mixed Explanatory Variable

The conditional expectation is given as,

$$E(Y/X = x, Z = z) = g(X\beta + Z\alpha) \qquad (11)$$

where Z denotes the discrete and X denotes the continuous explanatory variables in the model. β and α are vectors of parameters to be estimated. The method DWADE explained in the Subsection 2.3 is used in the estimation of β. Horowitz and Hardle (1996) developed an estimator for α that is the parameter vector of the discrete explanatory variables. The horizontal distance between $g(v + z^{(i)}\alpha)$ and $g(v + z^{(1)}\alpha)$, $(i = 2,...,M)$ is used for the estimation. Here, $S_z \equiv \{z^{(i)} : i = 1,...,M\}$ defines the discrete random variable Z. They assumed that $g(v + z\alpha)$ satisfies a weak monotonicity condition. They also assumed that there are finite numbers v_0, v_1, c_0 and c_1 such that $v_0 < v_1$, $c_0 < c_1$, $g(v + z\alpha) < c_0$ for each $z \in S_z$ if $v < v_0$ and $g(v + z\alpha) > c_1$ for each $z \in S_z$ if $v > v_1$.

In the XploRe commands, the determination of the scalars c_0 and c_1 is required. In order to do this, the data is graphed on the each level of the discrete variable and the interval satisfying the monotonicity condition is determined (Horowitz 1998; Horowitz and Hardle 1996).

2.5. The Optimal Bandwidth Selection Problem in the Nonparametric Regression

After estimating the linear index function in the first step of the semiparametric approach, the nonparametric regression method is applied in the estimation of the link function g. The main problem here is to determine the method that gives the optimal bandwidth parameter value h.

Taking into consideration of its easier mathematical structure, the Least Square Cross-Validation (LSCV) method defined as,

$$LSCV(h)=\frac{1}{n}\sum_{i=1}^{n}\left[Y_i-\left(\frac{\sum_{j\neq i}^{n}Y_j K_h(X_i-X_j)}{\sum_{j\neq i}^{n}K_h(X_i-X_j)}\right)\right]^2 \quad (12)$$

is used here. The value h minimizing the LSCV function is the optimal one. In Eq.(12), K is a kernel function used in the nonparametric regression, Y is the vector of values of the observed dependent variable, X is the vector of the explanatory variables ($i\neq j=1,\ldots,n$) and n is the sample size (Hardle et al. 2004; Horowitz 1998).

3. THE UNIFORM CONFIDENCE BANDS

Because new XploRe commands testing the validity of the parametric complementary log log model assumptions are constructed based on the UCB limits, the three steps of the UCB procedure are discussed below.

1. In the first step, the linear index function $X\beta$ is estimated by using one of the semiparametric estimators.

2. In the second step, the nonparametric regression of Y on the estimated $X\hat{\beta}$ denoted by $m(x)$ is obtained.

3. In the last step, the UCB limits are constructed based on the nonparametric estimates.

If the parametric link function lies around the nonparametric estimates between the confidence limits, it is concluded that the parametric model assumptions are satisfied. The UCB limits for the nonparametric estimate ($m(x)$) at point x is defined as,

$$P\left\{\hat{m}_h(x)-z_{n,\alpha}\sqrt{\frac{\hat{\sigma}_h^2\|K\|_2^2}{nh\hat{f}_h(x)}}\leq m(x)\leq\hat{m}_h(x)+z_{n,\alpha}\sqrt{\frac{\hat{\sigma}_h^2\|K\|_2^2}{nh\hat{f}_h(x)}}\right\}\cong1-\alpha \quad (13)$$

where h is the optimal bandwidth parameter required for the nonparametric estimate, $\hat{\sigma}_h^2$ is the estimated variance of $m(x)$ given by Eq.(16), K is a kernel function, K' is the first order derivative of K and $\|K\|_2^2$ is the second order norm of K defined by Eq.(14).

$$\|K\|_2^2=\int\left[K(s)\right]^2 ds \quad (14)$$

$$z_{n,\alpha}=\left\{\frac{-log\left(-\frac{1}{2}log(1-\alpha)\right)}{(2\delta log n)^{1/2}}+d_n\right\}^{1/2}$$

Here, the open form of d_n is given as follows.

$$d_n=(2\delta log n)^{1/2}+(2\delta log n)^{-1/2}log\left(\frac{1}{2\pi}\frac{\|K'\|_2}{\|K\|_2}\right)^{1/2} \quad (15)$$

$$\hat{\sigma}_h^2(x)=\frac{\frac{1}{n}\sum_{i=1}^{n}K\left(\frac{x_i-x}{h}\right)\left\{y_i-\hat{m}_h(x)\right\}^2}{\sum_{i=1}^{n}K\left(\frac{x_i-x}{h}\right)} \quad (16)$$

Gaussian, Epanechnikov and Quadratic kernels are frequently used in practice. Any kernel function could be used in the estimation procedure by considering the support of the distributions and type of the data.

Another important point that worth mentioning here is that because different scale and normalization methods are used in the estimation of the semiparametric (g) and the parametric (G) link functions, they can not be shown in the same graph simultaneously. The following procedure is proposed for solving this problem (Horowitz 1998, Horowitz and Hardle 1994).

1. β is estimated using one of the semiparametric methods.

2. Index values are computed using $\hat{\beta}$ estimates. ($\upsilon_i=x_i\hat{\beta}$; $i=1,\ldots,n$)

3. The scale parameter s and the constant term c of the parametric model are estimated using y_i and υ_i.

4. A probability estimation for observation i is obtained from $\hat{y}_i=1-exp\{-exp[(\upsilon_i-c)/s]\}$ for the parametric complementary log log model.

5. The $\tilde{y}_i$'s are computed by applying the nonparametric regression of y_i on υ_i. Then the link function is estimated and the confidence limits are constructed.

6. $\hat{y}_i$, $\tilde{y}_i$ and the confidence limits are graphed against υ_i.

4. NEW XploRe COMMANDS FOR TESTING THE PARAMETRIC COMPLEMENTARY LOG LOG MODEL ASSUMPTIONS

In this section, new commands testing the parametric complementary log log model assumptions against its semiparametric alternative based on the UCB limits in the windows based version 4.8 of the XploRe package are introduced. Models including only continuous and mixed explanatory variables are discussed, separately.

The quantlet "dwade" in XploRe is used for the model with continuous explanatory variables whereas the "adedis" is used for the estimation of the mixed model. The explanations of the commands are given in detail below (Hardle 1988, Hardle et al. 2007, Hardle et al. 2003, Klinke et.al. 1997, Proença and Werwatz 1994).

4.1. Commands for the Model with Mixed Explanatory Variable (s)

proc(cb4)=procedure2()

dat=read("comploglog2");Reads the data set "comploglog2" written in ASCII format.

y=dat[,4]   ; 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.

z=dat[,3]   ; Describes the column number of the discrete explanatory variable(s) (*z*) in the data set.

x=x.-mean(x)   ; Centralizes *x* values to eliminate the possible correlation among them.

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

w=eigen.values   ; Expresses the eigenvalues using a matrix "*w*".

v=eigen.vectors   ; Expresses the eigenvectors using a matrix "*v*".

mah=v*(sqrt(1./w).*v') ;Applies the Mahalanobis transformation.

x=x*mah   ; Weights the 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 calculations.

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

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

{delt,alphahat,lim,hd,text}=adedis(z,x,y,h,1.5,0.2,0.8)

    ; Calls the command "adedis" for the estimation of the β's for the discrete and continuous explanatory variables, separately. The β estimations of the continuous variable(s) are assigned to the vector "delt" whereas β estimations of the discrete one(s) are assigned to the vector "alphahat". Additionally, *hfac=1.5; c_0= 0.2 and c_1=0.8* are determined here using the methods in Subsection 2.4.

b=mah*delt   ; Describes the transformations to the original values of the estimations of the continuous explanatory variables.

b=b./abs(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 complementary log log model and its semiparametric alternative.

υ_i = x*b+z*alphahat ; Computes the value of the linear index estimation of observation i.

x=matrix(rows(x))~υ_i ; Adds a column vector including only the elements "1" to the left side of the matrix υ_i.

; The estimation of the scale s and the constant c of the complementary log log model

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

g=glmest("bicll",x,y) ;Gives the estimations of the complementary log log model.

glmout("bicll",x,y,g.b,g.bv,g.stat) ; Displays the required outputs on the output screen.

c=g.b[1,]   ; Gives the first estimated coefficient of the model (b_0).

s=g.b[2,]   ; Gives the second estimated coefficient of the model (b_1).

yhat=$1-\exp\left(-\exp\left((\upsilon_i - c)/s\right)\right)$; Calculates the probability of belonging to the category "1" coded in the dependent variable for each observation using the *c* and *s* values of the model.

z=y~yhat   ; Adds the yhat column to the right side of *y*.

z1=υ_i ~yhat   ; Adds the yhat column to the right side of υ_i.

z1sorted=sort(z1)   ; Sorts *z1* values.

; Nonparametric regression of *y* on υ_i

data=υ_i ~y   ; Adds the column matrix *y* to the right side of υ_i.

h1=regxbwsel(data); Presents alternative bandwidth selection methods such as Cross-Validation, Shibata's Model Selector, Akaike's Information Criterion, Rice's T etc. The Cross-Validation, generally accepted method due to its simplicity in the estimation procedure, is used here.

{mh,clo,cup}=regxcb(data,h1,0.05,"gau") ; Calculates mh, the lower confidence band (clo) limit and the upper confidence band (cup) limit at the $\alpha = 0.05$ confidence level and with the "Gaussian" kernel function. This

command becomes the confidence level and the kernel function optional.

{mh,cli,cui}=regxci(data,h1,0.05,"gau"); Calculates mh and the pointwise confidence intervals at the $\alpha = 0.05$ level with the "Gaussian" kernel function.

; Graphical representation of mh, yhat and confidence bands

z1sorted=setmask(z1sorted,"circles","red")

; Describes the image of "z1 sorted" in the graph.

mh=setmask(mh,"line","black")

; Describes the image of "mh".

clo=setmask(clo,"line","blue","thin","dashed")

; Describes the image of "clo".

cup=setmask(cup,"line","blue","thin","dashed")

; Describes the image of "cup".

plot(z1sorted,mh,clo,cup)   ; Plots "z1sorted", "mh", "clo" and "cup"

endp

procedure2()

5. A NUMERICAL EXAMPLE

In this section, both an artificial and a real data set are used to show the applicability of all XploRe commands in practice. Some popular bandwidth selection methods such as Cross-Validation, Shibata's Model Selector, AIC for the estimation of mh are displayed in the screen when the commands are executed. The UCB confidence limits are calculated and graphed after doing the required selections.

5.1. The Results over an Artificial Data

The data used here has been taken from the original data library of the XploRe package. The name of the data is "dwade" with the sample size of 80.

Y is a binary outcome coded as 0 and 1; *X* is a nx2 matrix denoting the continuous explanatory variables and *Z* is a nx1 matrix that represents the discrete explanatory variable. *Z* has been added to the data "dwade" for illustrating the usage of the commands for the mixed explanatory variable model.

5.1.1. The Results of the model with only continuous explanatory variables

The optimal bandwidth value ($h_{opt} = 1.06988$) obtained by the method LSCV is shown in Figure1. The optimal range of *h* is (0.168495-2.69593). The quadratic kernel function (abbreviated by "qua") is used in the estimation.
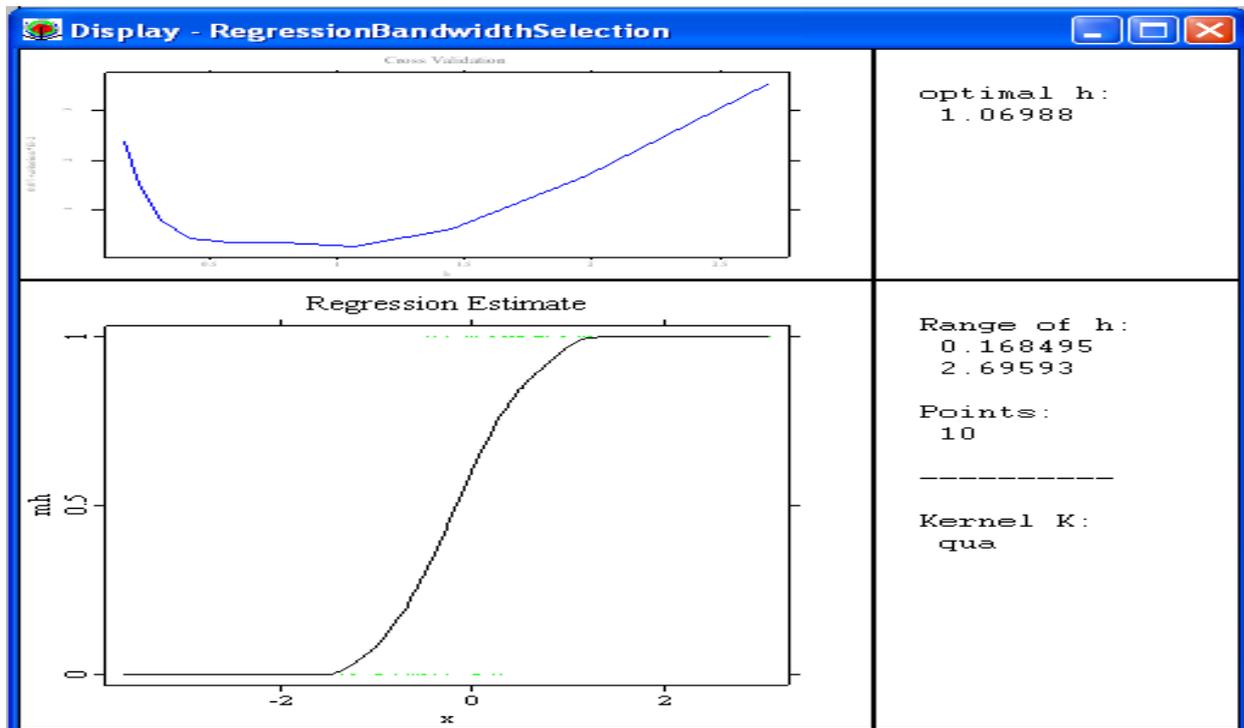

Figure1.The optimal bandwidth value of the nonparametric regression of Y on X in the model with continuous explanatory variables.
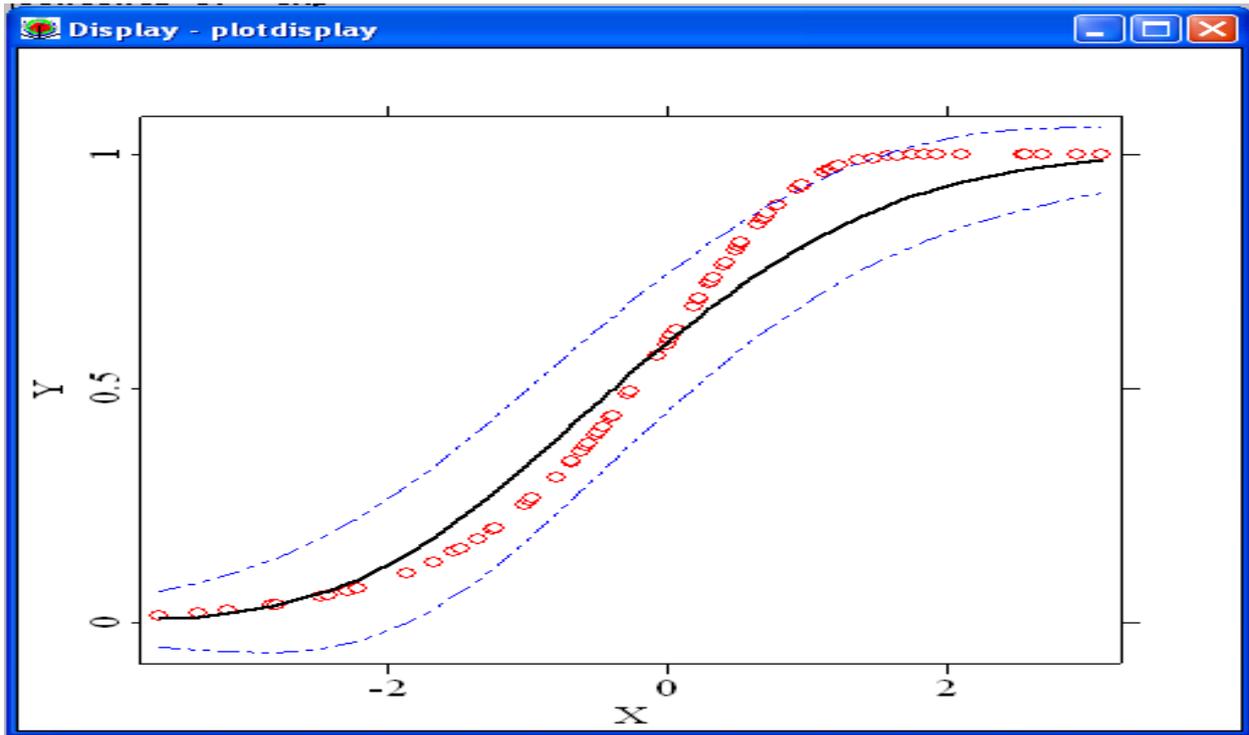
Figure 2. The estimated parametric curve, the nonparametric curve and the UCB limits at the $1-\alpha = 0.95$ confidence level in the model with continuous explanatory variables.

The estimated parametric curve, the nonparametric curve and the UCB limits are graphed in Figure 2. at the $1-\alpha = 0.95$ confidence level with the Gaussian kernel.

The circles represent the parametric link function, the line represents the estimated nonparametric curve and the broken line represents the lower and the upper UCB limits.

Because some parts of the circles lie out of the UCB upper limit in Figure 2, it is concluded that the use of the parametric complementary log log model is not appropriate for modelling the data.

5.1.2. The results of the model with mixed explanatory variables

As shown in Figure 3, the optimal bandwidth value obtained by the method LSCV and using the quadratic kernel function is $h_{opt} = 0.386358$. The optimal range of h is (0.283922-4.54275)
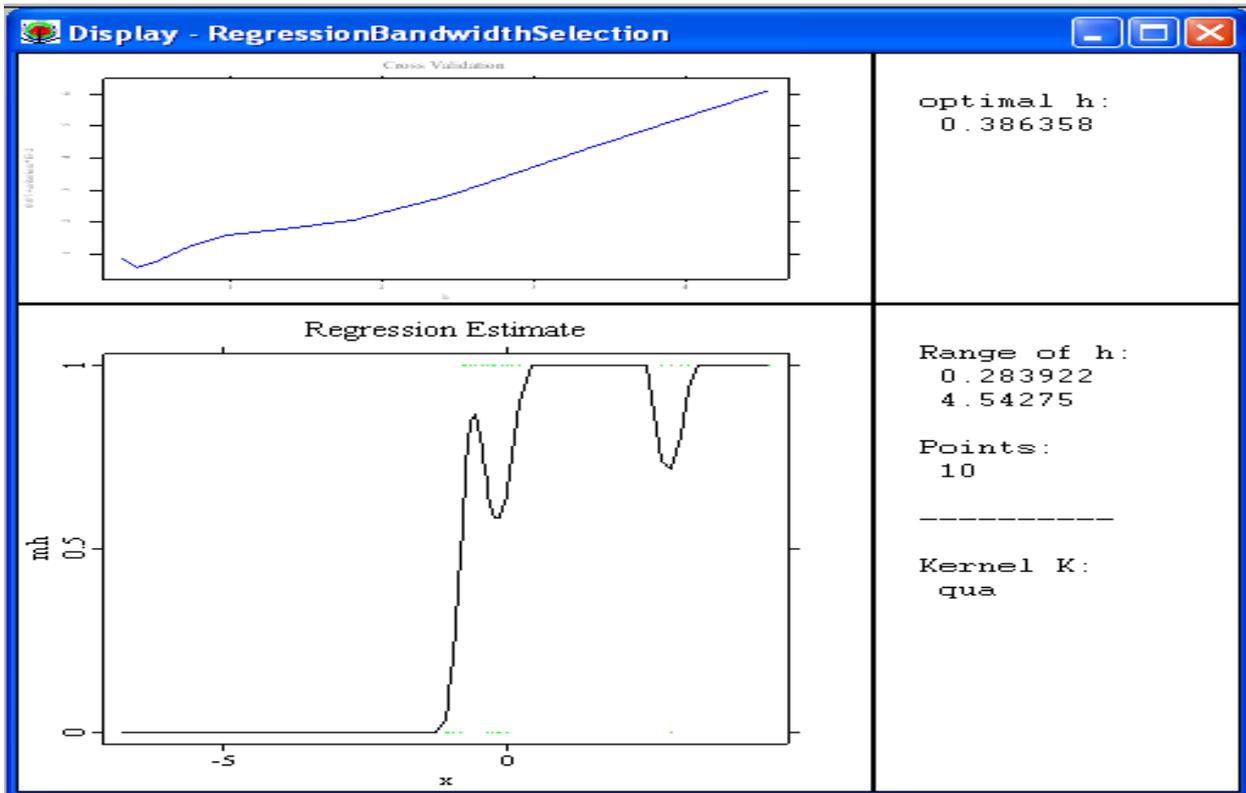
Figure 3. The optimal bandwidth value of the nonparametric regression of Y on X in the model with mixed explanatory variables.
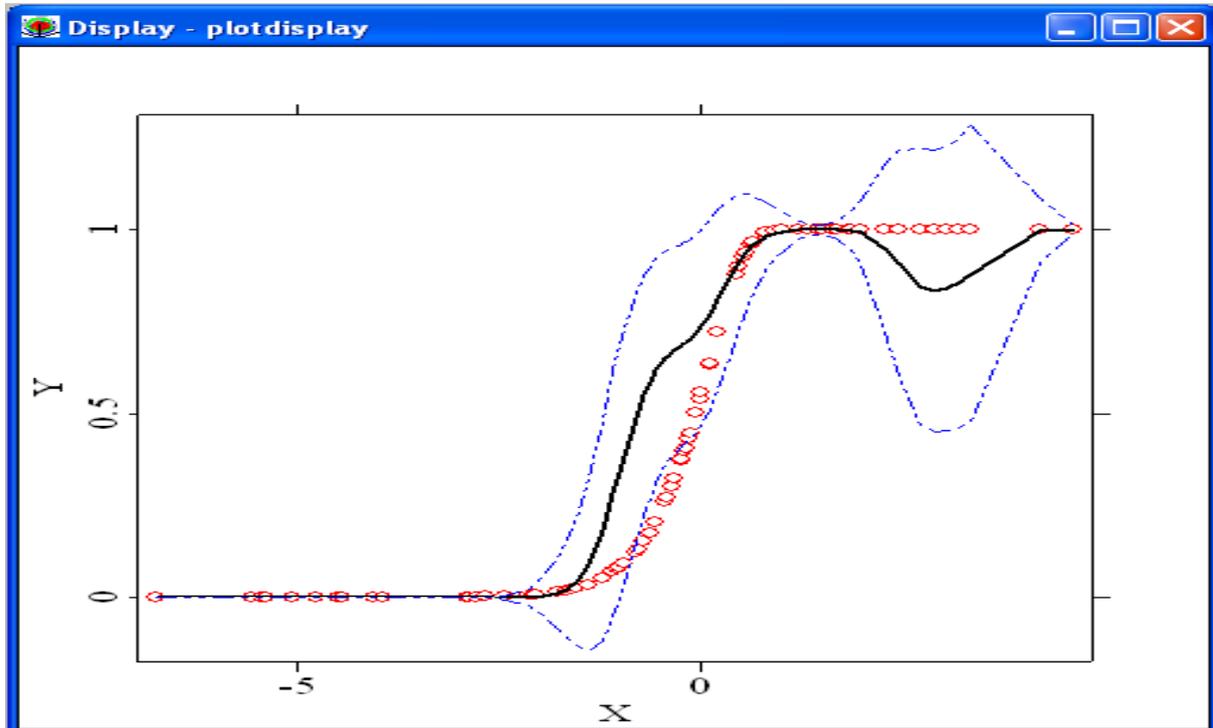

Figure 4. The estimated parametric curve, the nonparametric curve and the UCB limits at the $1 - \alpha = 0.95$ confidence level in the model with mixed explanatory variables.

Figure 4 indicates that the use of the parametric complementary log log model with mixed explanatory variables is also rejected.

5.2. A Real Data Results

Here, the commands given in Subsection 4.2 are run over a real data set on gastric cancer including male patients of 65 to illustrate the usage of them in practice. The gastric cancer

data is obtained from the project report of Akkaya (2008) in the department of Biostatistics of Hacettepe University of Turkey.

Y is a binary variable coded as follows.

$$Y_i = \begin{cases} 0 & \text{if patient } i \text{ is alive.} \\ 1 & \text{if patient } i \text{ is dead.} \end{cases}$$

Three important factors that may affect the disease are determined. The age of the patient and the tumour size are continuous variables represented by X_1 and X_2, respectively. X_3 is a discrete variable coded as,

$$X_3 = \begin{cases} 0 & \text{if the metastasiz is not stated.} \\ 1 & \text{if stated.} \end{cases}$$

No other change in the commands is required beyond rearrangement of the column numbers of the new variables.
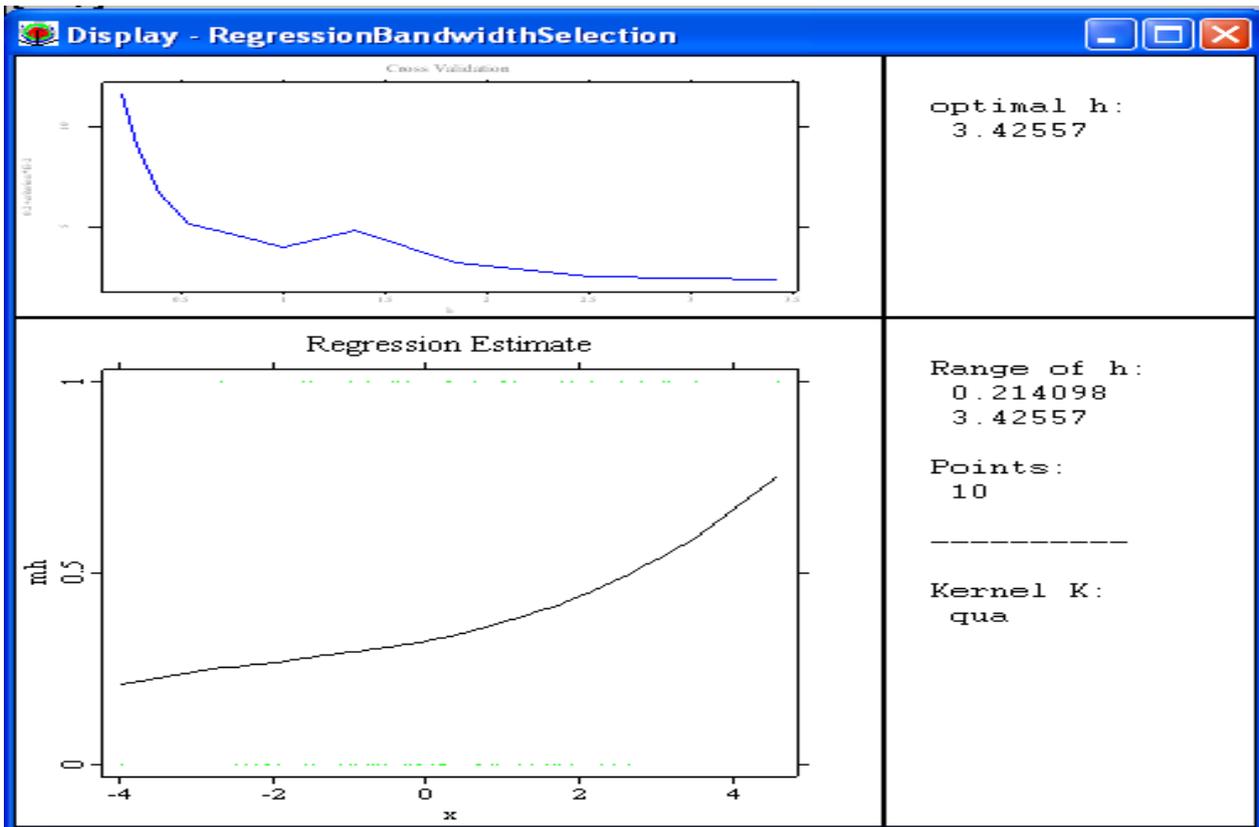

Figure 5. The optimal bandwidth value of the nonparametric regression of *Y* on *X* for the gastric cancer data.

Figure 5 shows that the optimal bandwidth value obtained based on the method LSCV and the quadratic kernel function is $h_{opt} = 3.42557$. The optimal range of *h* is (0.214098-3.42557).

Figure 6 displays that the important factors affecting the probability of being dead of patients could not be determined by using the parametric complementary log

log model. In other words, the gastric cancer data could not be modeled by the method of the parametric complementary log log model due to the violation of the model assumptions. Instead, the semiparametric alternative of the model could be dealt with.
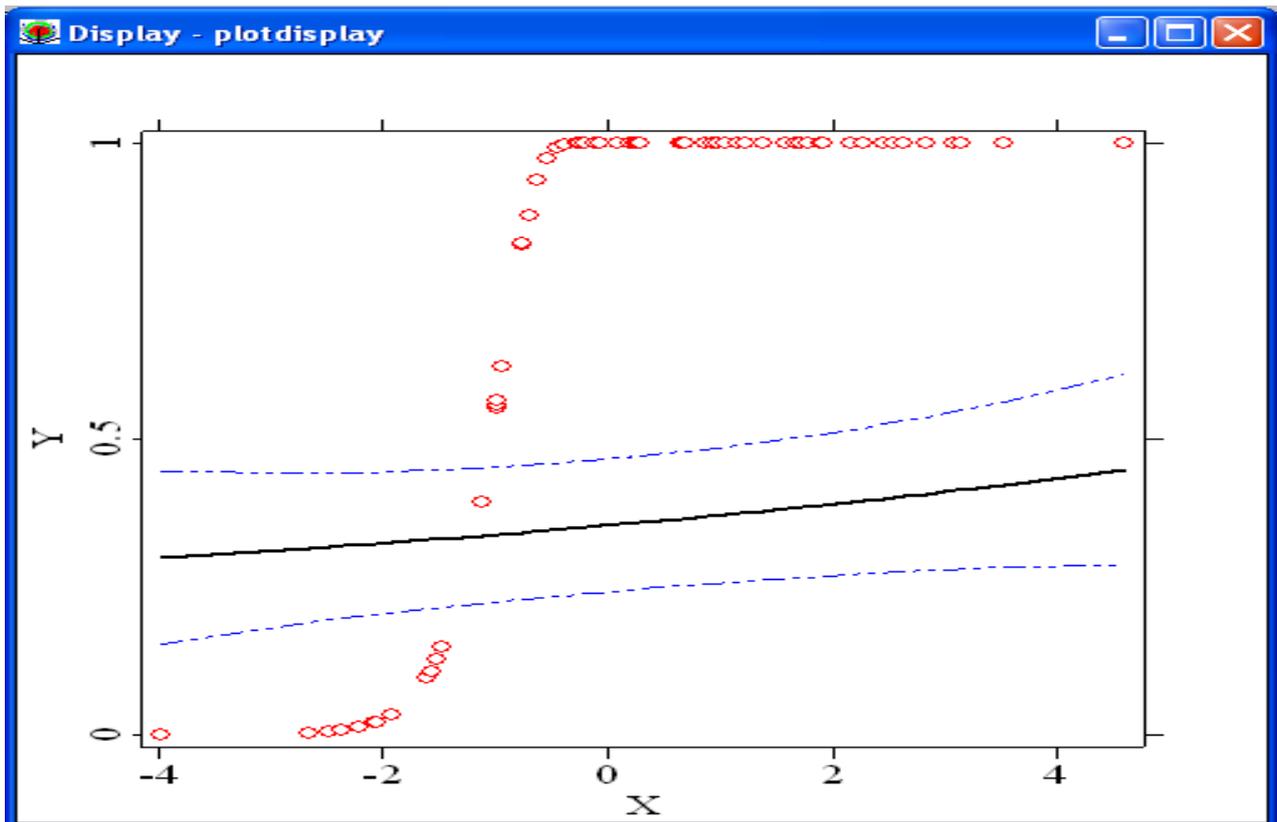
.

Figure 6. The estimated parametric curve, the nonparametric curve and the UCB limits at the $1 - \alpha = 0.95$ confidence level for the gastric cancer data.

6. CONCLUDING REMARKS

In binary dependent variable modelling, the use of the parametric approach is common in most applications because of its simplicity in interpretation and application. However, the validity of these types of models is all based on the assumptions related to the error term. The main problem in the parametric modelling is how to test the validity of the model assumptions. At this point, a statistical testing criterion is required.

In this study, the validity test of the model assumptions related to the one of the most popular binary parametric models, the complementary log log model, is focused. The UCB limits are used as testing criteria in the XploRe commands written in the windows based version 4.8 for models with continuous and mixed explanatory variables. This study also modifies and updates the commands of Proença and Werwatz (1994) to the case of the parametric complementary log log model. It also extends the commands restricted for the model including only continuous explanatory variable to the model with mixed explanatory variables and enables some commands to be optional. The explanations of the commands are given in Section 4.

The applicability of all XploRe commands is shown over an artificial and a real data set. They are executed for testing the validity of the parametric complementary log log model assumptions. In conclusion, the parametric model is rejected against the semiparametric alternative in both situations. These commands are considered to provide an easy way for testing the validity of the parametric complementary log log model assumptions by

taking into consideration the lack of a user friendly standard statistical package in existing literature.

REFERENCES

[1] Hardle, W., Klinke, S., Turlach, B.A., "XploRe: An interactive statistical computing environment and state-of-the-art density and regression smoothing", *Stat Comput*, 3:23-26 (2007).

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

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

[4] Hardle, W., "XploRe-A computing environment for exploratory regression and density smoothing", *Stat Softw Newsl*, 14:113-119 (1988).

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

[6] Horowitz, J.L, Hardle, W., "Direct semiparametric estimation of single-index models with discrete covariates", *J Am Stat Assoc*, 91:1632-1640 (1996).

[7] Horowitz, J.L, Hardle, W., "Testing a parametric model against a semiparametric alternative", *Econ Theory*, 10:821-848 (1994).

[8]   Klinke, S., Golubev, Y., Hardle, W., Neuman, M., "Teaching wavelets in XploRe", ***Comput Stat***, 13:141-151 (1997).

[9]   McCullagh, P., Nelder, J.A., "Generalized Linear Models", Monogr Stat Appl Probab 37, ***Chapman and Hall***, London (1989).

[10] Müller, M., Ronz, B., Hardle, W.K., "Computerassisted semiparametric generalized linear models", ***Comput Stat***, 12:153-172 (1997).

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

[12] Proença, I., Werwatz, A., "Comparing parametric and semiparametric binary response models", Sonderforschungsbereich 373 2000-20. Humboldt Universitaet. Berlin, Available via DIALOG. ftp://amadeus.wiwi.hu-berlin.de/pub/papers/sfb373 (1994).

[13] Wang, X., Dey, D.K., "A flexible skewed link function for binary response data", Technical Report, Department of Statistics. University of Connecticut, Available via DIALOG. http://www.samsi.info/TR/tr2008-05.pdf (2008).