NONPARAMETRIC REGRESSION: AN ALTERNATIVE TO THE SCATTER DIAGRAM

Ernesto Pedro Menéndez Acuña*, [1]Eliseo Gabriel Argüelles*, Sergio Hernández González**
*Faculty of Mathematics. Universidad Veracruzana. Mexico.
**Faculty of Statistics. Universidad Veracruzana. Mexico

ABSTRACT
In this paper is shown the possibility to use a nonparametric regression instead of a scatter diagram, when this last cannot be employed in order to obtain information about the shape of the model. In particular it is considered a logistic regression, and a single quantitative independent variable. The proposal is illustrated with two examples. The former considers a scatter diagram that shows a linear relationship between the values of the independent variable and the probabilities of success of an event. The other one presents a scatter diagram that evidences a nonlinear relation between them.

KEYWORDS: Generalized linear models;Linear regression model; Logistic regression;Regression analysis.

MSC: 62G08; 62J12

RESUMEN
En este trabajo se muestra la posibilidad de usar una regresión no paramétrica en lugar de un diagrama de dispersión, cuando éste último no pueda ser empleado para obtener información sobre la forma del modelo. En particular es considerada una regresión logística con sólo una variable independiente cuantitativa. La propuesta se ilustra con dos ejemplos. El primero considera un diagrama de dispersión que muestra una relación lineal entre la variable independiente y la probabilidad de éxito de un evento. El otro ejemplo presenta un diagrama de dispersión que evidencia una relación no lineal entre las variables.

1. INTRODUCTION

In the case of a logistic regression with one independent quantitative variable, a scatter diagram does not can give information about the linear structure of the systematic component of the logistic model. The objective of this work is to show as a non-parametric regression can be used to obtain information about the most appropriate model to be considered.

A major activity in statistics is the building of statistical models. In particular, the aim of regression analysisis to construct mathematical models which describe or explain relationships that may exist amongvariables (Seber and Lee, 2003; Sheather, 2009).Quite often, an experimental research work requires the empirical identification of the relationship among an observable response variable "y", and a set of associated variables, or factors, that is believed that they have an effect on "y". In general, if such relationship exists, it is unknown, but is usually assumed to be of a particular form, provided that it can adequately describe the dependence of "y" on the associated variables (or factors).This procedure leads to the so-called *postulated model,*which contains a number of unknown parameters, in addition to a random experimental error term (Khuri, 2009).

The simplest case is when there are just two variables, such as height and weight, income and intelligence quotient (IQ), the length and breadth of the leaves, temperature and pressure of a certain volume of gas, etc.If we have *n* pairs of observations $(x_i, y_i)$of the variables "x" and "y", we can plot these points, giving a scatter diagram, and endeavor to fit a smooth curve through the points, in such a way that the points are as close to the curve as possible. Clearly, we would not expect an exact fit, because at least one of the variables is subject to chance fluctuations due to factors outside our control. With two variables, the simplest regression model is the straight line. It is a particular case of the multiple regression models. The regression analysis is a statistical technique widely used, and it is employed in almost every field of application (Ryan, 1997). Many books and research papers have been published as an evidence

[1] emenendeza@gmail.com

of this fact. To mention only several of them, see Ryan (1997, pag. 1, 2).The linear regression analysis allows us to consider a dependent variable and one or more independent variables.

Suppose a set of n observations$(y_i, x_{1i}, x_{2i}, \dots, x_{pi})$from the variables *“y”* and "$x_1, x_2, \dots x_p$"where *“y”* represents the dependent variable and "$x_1, x_2, \dots x_p$"the independent variables. Amultiple linear regression is modelled by

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \varepsilon_i \tag{1.1}$$

In (1.1) the term ε_i represents a random error, where $E\left(\varepsilon_i / x_{1i}, x_{2i}, \dots, x_{pi}\right) = 0$ and $Var\left(\varepsilon_i / x_{1i}, x_{2i}, \dots, x_{pi}\right) = \sigma^2$, for $i = 1,2, \dots, n$. The role of the error term is to account for the extra variation in “y” that cannot be explained by the postulated model. Then, an equivalent form of the expression (1.1) is given by

$$E\left(y_i / x_{1i}, x_{2i}, \dots, x_{pi}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} \tag{1.2}$$

Additionally it can be assumed that the errors are normally distributed. With this assumption, it is possible to perform a broader inferential study (Draper and Smith, 1998). The objective is to estimate the parameters β from the data, applying the method of least squares, or likelihood estimation, if errors are normally distributed. Even with an independent variable, it is possible to study, not only linear relationships, but also non-linear relationships. In this case the expression (1.1) becomes in

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \varepsilon_i \tag{1.3}$$

or

$$E(y_i / x_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p \tag{1.4}$$

In this work the case where there exists only an independent variable is of interest.

From a parametric perspective, to apply linear regression is necessary to know the structure of the model which bestexpressesthe relationship between the dependent variable and the independent variables. In the case of asingle independent variable,if there is nota knowledge about the possible relation betweenthese variables, a scatter diagram can help to obtain evidence on this possible relation(Sheather, 2009).

For example, figures (1.a) and (1.b) represent two scatter diagrams.The first suggests that the existing relationship between the variables "y" and "x" is a straight line, and the second, a second degree polynomial.

But no always is possible to use a scatter diagram, even though it is considered an alone independent variable.For example, when the dependent variable takes only 0 and 1 values, since a scatter diagram displays ordered pairs (x, y) on two straight parallel lines, y = 1 and y = 0, which does not provide evidence on the possible relationship between these variables. In this situation, it is not appropriate to use a linear regression model to study the relationship betweenthe variables. However, through the use of a logistic regression, the study of the relationship between these variables can be done.

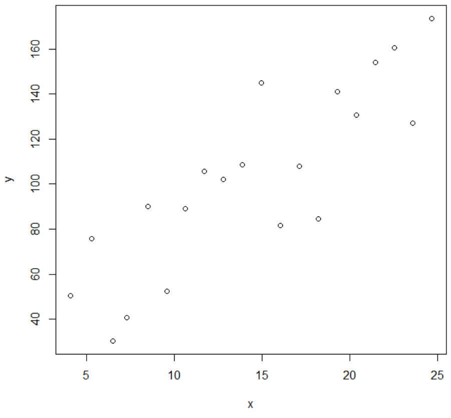


Figure1.a

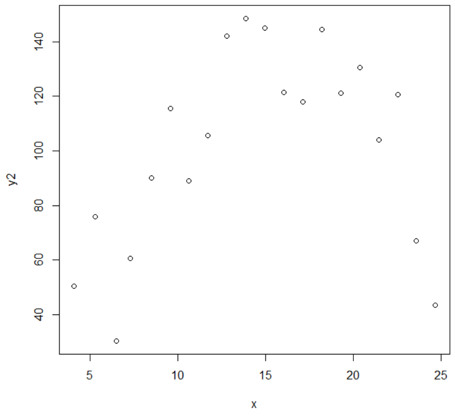


Figure1.b

On the other hand, when the functional form of the model which explains the relationship between the dependent variable and the independent variable is decided from a scatter diagram, what really is fixed is the second term of the expression (1.4). Therefore, once determined the values of the parameters$\beta_0, \beta_1, \dots, \beta_0$, estimates of$E\left(y_i / x_i, x_i^2, x_i^3, \dots, x_i^p\right)$, for each i = 1,2,..., n, are obtained.Then the figures (1.a) and (1.b) suggest respectively the linear models

$$E(y_i / x_i) = \beta_0 + \beta_1 x_i$$

and

$$E(y_i/x_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$$

2. NON-PARAMETRIC REGRESSION AND LOGISTIC REGRESSION

There is another way to perform a regression analysis without imposing any kind of model to data. The non-parametric regression (NPR) it allows determining graphically the existing relationship between the variables. With a NPR is not necessary to fix a model before to perform the regression analysis, but rather, the model is determined by the data (Takezawa, 2006; Eubank, 1999;Ruppert, Wand, and Carroll, 2003).

The NPR is a collection of techniques that allow estimating the functional form of the regression function from the data, and any assumption of linearity is replaced with a much weaker assumption of a smooth regression function; is therefore appropriate use it when do not exist prior knowledge of the relationship between the variables under study, or when the modelling using a parametric regression is very difficult, given the structure of the relationship between the dependent and independent variables.This characteristic makes very flexible the non-parametric regression (Eubank, 1999).The NPR does not assume a particular model. In this case the model is very general, and it is given by

$$E(y/x) = m(x) \qquad (2.1)$$

where m(x) is some unknown smoothed function and which expresses the functional form of the relationship between *"y"* and *"x"*. The objective is to estimate the functional form of $m(x)$ from the data (Keele, 2008). This estimate is achieved through some method of non-parametric estimation (Takezawa, 2006). Once estimated $m\ (x)$by $\widetilde{m}(x)$, an estimated of $E(y_i/x_i)$ for each value of x_i is obtained; and this informationis, in some sense, equivalent to the information provided by a scatter diagram.

On the other hand, the study of the relationship between a dichotomous dependent variable "y", with Bernoulli distribution and a quantitative independent variable "x", it is possible using a generalized linear model (GLM) (Agresti, 2002; Dobson, 2002; Faraway, 2006; McCulloch and Searle, 2001).If $\pi = P\{y = 1\,/x\} = E(y/x)$, the simplest model is given by the expression (2.1), in this case

$$E(y/x) = \pi = \beta_0 + \beta_1 x \qquad (2.2)$$

But the expression (2.2) makes no sense, because once estimated the parameters, the adjusted model $\hat{\pi} = \widehat{\beta_0} + \widehat{\beta_1}x$cannot guarantee values inside the interval (0, 1). An alternative is to formulate the relationship in terms of a function g (.), this is

$$\pi = g(\beta_0 + \beta_1\, x)$$

But when the function g (.) is a distribution function, their values will be in the interval (0, 1). If it is assumed that

$$\pi = g(\beta_0 + \beta_1 x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \qquad (2.3)$$

g (.) is the distribution function of the logistic distribution with location parameter μ = 0 and scale parameter s =1, evaluated in $(\beta_0 + \beta_1\, x)$. From the expression (2.3),

$$ln\left[\frac{\pi}{1-\pi}\right] = \beta_0 + \beta_1 x \qquad (2.4)$$

is obtained.

The expressions (2.3) and (2.4) can be generalized for more than one independent variable, resulting

$$\pi = g\left(\beta_0 + \beta_1 x_1 + \cdots + x_p\right) = \frac{e^{\beta_0+\sum_{j=1}^{p}\beta_j x_j}}{1+e^{\beta_0+\sum_{j=1}^{p}\beta_j x_j}} \qquad (2.5)$$

and

$$ln\left[\frac{\pi}{1-\pi}\right] = \beta_0 + \sum_{j=1}^{p}\beta_j x_j\ . \qquad (2.6)$$

In particular if $x_j = x^j$ for j=1,2,…,p,

$$\pi = g\left(\beta_0 + \sum_{j=1}^{p}\beta_j x^j\right) = \frac{e^{\beta_0+\sum_{j=1}^{p}\beta_j x^j}}{1+e^{\beta_0+\sum_{j=1}^{p}\beta_j x^j}} \qquad (2.7)$$

and

$$ln\left[\frac{\pi}{1-\pi}\right] = \beta_0 + \sum_{j=1}^{p}\beta_j x^j \qquad (2.8)$$

The expressions (2.4), (2.6) and (2.8) represent logistic regression models, as special cases of a GLM, where the link function is$n\left[\frac{\pi}{1-\pi}\right]$, known as logit function. The systematic component in each case is a linear predictor, given by $\beta_0+\beta_1 x, \beta_0+\sum_{j=1}^{p}\beta_j x_j$ and $\beta_0+\sum_{j=1}^{p}\beta_j x^j$ respectively, and the random component is a Bernoulli random variable.

3. NUMERICAL ILLUSTRATION

Two different scenarios were considered. In each one it is fixed a set of 20 pairs of values (x_i, π_i) ,The variable "x"represents the independent variable, and "π" the dependent variable. The variable "π" represents the probability of success of an event, that is, it only takes valuesbetween 1 and 0. With each value"π", it was generated a random variable Bernoulli with probability of successπ_i, resulting a sample of size 20, of pairs of values(x_i, y_i). In the tables 1.a and 1.b, appear the values of the variables "x", "π" and "y", whose scatter diagrams (figures 2.a and 2.b) suggest a linear relationship and a nonlinear between the variables "x" and "π", respectively. In practice the values of the variable "π"$(\pi_i)$are unknown, and neither is possible to make a scatter diagram, nor obtain evidence of the possible model of the systematic component in a logistic regression. In this work it is assumed that the $\pi_i$values are known, but just to illustrate how the use of a scatter diagram can be substituted by a nonparametric regression, in order to obtain information on the model that to be considered in the logistic regression.

On the other hand, on each figure (2.a and 2b), it is shown the scatter diagram of pairs$(x_i,\ y_i)$, the π_i values, estimated by the application of a nonparametric regression using spline smoothing, and a logistic regression with the model

$$ln\left[\frac{\pi}{1-\pi}\right]=\beta_0+\beta_1 x$$

on figure 2.a, and

$$ln\left[\frac{\pi}{1-\pi}\right]=\beta_0+\beta_1 x+\beta_2 x^2$$

on figure 2.b.

For the computation was used the Rlanguage(R Code Team,2008).

Table 1.a

x	*π*	*y*
4.1	0.25	0
5.3	0.38	0
6.5	0.15	0
7.3	0.20	0
8.5	0.45	1
9.58	0.26	1
10.66	0.45	0
11.74	0.53	0
12.82	0.51	1
13.9	0.54	1
14.98	0.73	0
16.06	0.41	0
17.14	0.54	1
18.22	0.42	0
19.3	0.71	0
20.38	0.65	1
21.46	0.77	1
22.54	0.80	1
23.62	0.64	1
24.7	0.87	1

Table 1.b

x	*π*	*y*
4.1	π	0
5.3	0.25	0
6.5	0.38	0
7.3	0.15	0
8.5	0.30	0
9.58	0.45	1
10.66	0.58	1
11.74	0.45	0
12.82	0.53	1
13.9	0.71	1
14.98	0.74	1
16.06	0.73	1
17.14	0.61	0
18.22	0.59	0
19.3	0.72	0
20.38	0.61	1
21.46	0.65	1
22.54	0.52	0
23.62	0.60	0
24.7	0.34	0

Figure 2.a

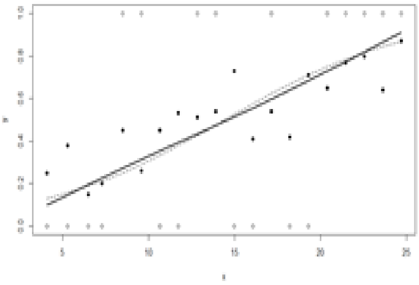
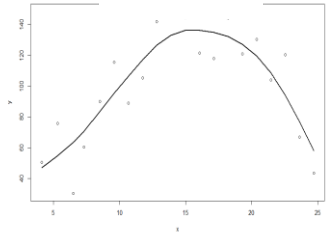


Figure 2.a



***Symbols used: •, °, ---, and –, represent the probabilities*π_i*,, the values* y_i*, the estimated probabilitiesvia the applicationof a logistic regression and NPR, respectively*.

4. CONCLUSIONS

From the obtained results, it can be concluded that the proposal formulated, is reasonable.A logistic regression was considered in this work, but there are other cases where a scatter diagram is not informative.For example, when one tries to adjust growth models, and as it is known, there are many of them. Therefore, different models offer different settings, which can not be assessed from the result of a scatter diagram. In this situation the applicationof a non-parametric regression can help in the selection of the most suitable model.

RECEIVED JULY 2014
REVISED DECEMBER 2014

REFERENCES

[1] AGRESTI, A. (2002): **Categorical Data Analysis**. Wiley. New York.
[2] DOBSON, A. J. (2002): **An Introduction to Generalized Linear Models**. 2nd. Ed. Chapman and Hall. USA.
[3] DRAPER, N. and SMITH, H. (1998): **Applied Regression Analysis**. 3rd. ed. Wiley. USA.
[4] EUBANK, R. (1999): **Nonparametric Regression and Spline Smoothing**. Marcel Dekker. New York.
[5] FARAWAY J. (2006): **Extending the Linear Model with R, Generalized Linear, Mixed Effects and Nonparametric Regression Models**. Chapman and Hall. USA.
[6] KEELE, L. (2008): **Semiparametric Regression for the social sciences**. Wiley. USA.
[7] KHURI, A. I. (2009): **Linear Model Methodology**. Chapman and Hall/CRC. London.
[8] MCCULLOCH, C. E and SEARLE, S.R. (2001): **Generalized Linear and Mixed Models.** Wiley. USA.
[9] R CORE TEAM (2008): R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL **http://www.R-project.org/.**
[10] RUPPERT D., WAND P. and CARROLL R. (2003): **Semiparametric Regression**. Cambridge University Press. USA.
[11] RYAN, T. (1997): **Modern Regression Methods**. Wiley. New York.
[12] SEBER, G. A. F. and LEE A. (2003): **Linear Regression Analysis.** Wiley. USA.
[13] SHEATHER, S. (2009): **A Modern Approach to Regression with R**. Springer. New York.
[14] TAKEZAWA, K. (2006): **Introduction to Nonparametric Regression**. Wiley. USA.