LINEAR REGRESSION: AN ALTERNATIVE TO LOGISTIC REGRESSION THROUGH THE NON-PARAMETRIC REGRESSION

Ernesto P. Menéndez*, Julia A. Montano** Zoylo Morales** and Sergio Hernández**
*Faculty of Mathematics. Universidad Veracruzana. Veracruz. Mexico
**Faculty of Statistics. Universidad Veracruzana. Veracruz. Mexico

ABSTRACT
For applying the logistic regression, and any other type of parametric regression method, it is necessary to know the model that we want to adjust. In the case of a logistic regression, with only one independent variable, the use of a non-parametric regression is useful in order to obtain evidence of the possible relation between the success probability (π) and the independent variable (x). With this information is possible to determine the model to be adjusted. In this work, it is proposed to use a linear regression model, where the dependent variable is the probability ($\hat{\pi}$), and whose values are obtained from the estimated probabilities that resulting in the application of non-parametric regression. This proposal would avoid using logistic regression, whenever it is necessary to apply a non-parametric regression, for obtaining information of the type of model to be considered.

KEYWORDS: Linear regression, logistic regression, non-parametric regression.

MSC: 62G08; 62J12

RESUMEN
Para la aplicación de la regresión logística y cualquier otro tipo de regresión paramétrica, es necesario conocer el modelo que se desea ajustar. En el caso de una regresión logística, con una sola variable independiente, el uso de una regresión no paramétrica es de utilidad para obtener evidencia de la posible relación entre la probabilidad de éxito (π) y la variable independiente (x). Con esta información es posible determinar el modelo que debe ajustarse. En este trabajo se propone usar un modelo de regresión lineal, donde la variable dependiente es la probabilidad ($\hat{\pi}$), cuyos valores son obtenidos de las probabilidades estimadas que resultan de la aplicación de la regresión no paramétrica. Esta propuesta evitaría usar la regresión logística, siempre que sea necesario aplicar una regresión no paramétrica para obtener información del tipo de modelo a considerar.

1. INTRODUCTION

The aim of regression analysis is to construct mathematical models which describe or explain relationships that may exist among variables (Seber and Lee, 2003; Sheather, 2009). As it is known, from a parametric perspective, for applying linear regression is necessary to know the structure of the model which best expresses the relationship between the dependent variable and the independent variables. In the case of a single independent variable, if there is not knowledge about the possible relation between these variables, a scatter diagram can help to obtain evidence on this possible relation (Sheather, 2009). An alternative to the scatter diagram is the non-parametric regression (Menéndez, Gabriel and Hernández, 2015). The non-parametric regression does not impose any kind of model to data, and it allows determining graphically the existing relationship between the variables. With a non-parametric regression 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).
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. A linear regression model, in this case a multiple linear regression is model is given 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 $E(\varepsilon_i / x_{1i}, x_{2i}, \dots, x_{pi}) = 0$ and $Var(\varepsilon_i / x_{1i}, x_{2i}, \dots, x_{pi}) = \sigma^2$, 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

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

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, expressions (1.1) and (1.2) becomes in expressions (1.3) and (1.4) respectively

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \varepsilon_i \qquad (1.3)$$

or

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

On the other hand, the non-parametric regression is a set of smoothing techniques, that allows 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).A non-parametric regression does not assume a particular model. In this case the model is very general, and it is given by

$$E(y/x) = m(x)$$

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 information is, in some sense, equivalent to the information provided by a scatter diagram.

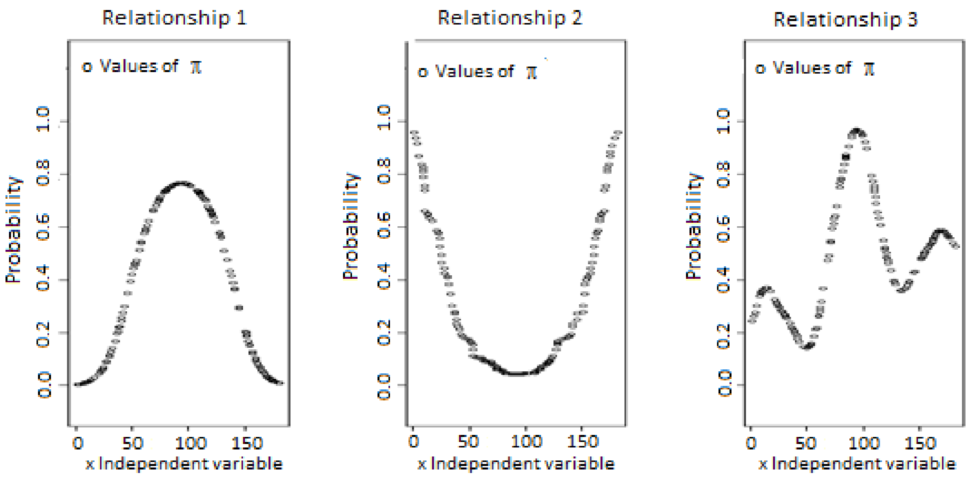
However, a linear regression cannot be used to study the relationship between a dichotomous dependent variable "y", with Bernoulli distribution, and a quantitative independent variable "x", because the least squares method for estimating the parameters of the linear model, does not guarantee estimated values of the dependent variable inside the interval [0,1]. When the dependent variable is dichotomic, (takes values 0 or 1), the logistic regression is a good option for studying the dependency relationship of this variable with respect to one, or more independent variables. Additionally, if there is only an independent variable in the logistic regression, a scatter diagram is not informative of the possible relationship between the variables under study. With the implementation of a non-parametric regression, it is possible to obtain this information (Menendez et al., 2015).

The goal of this work consists in to analyze the possibility of substituting the logistic regression by a linear regression, when a non-parametric regression is applied in order to obtain evidence on the relationship between the dependent dichotomic variable *"y"* and the independent variable "x", avoiding to use of logistic regression. How do apply linear regression instead of the logistic regression? Our proposal consists in applying a non-parametric regression for obtaining evidence on the structure of the possible relationship between the variables. Once this information has been obtained, it is possible determining the structure of the linear model to be adjusted. For example, $\beta_0 + \sum_{j=1}^{p} \beta_j x^j$, for p≥1, where the dependent variable is not the original variable "y", but the link function, and whose values are obtained when this function is evaluated in the estimated probabilities resulting from the application of non-parametric regression.

2. SIMULATION STUDY AND RESULTS

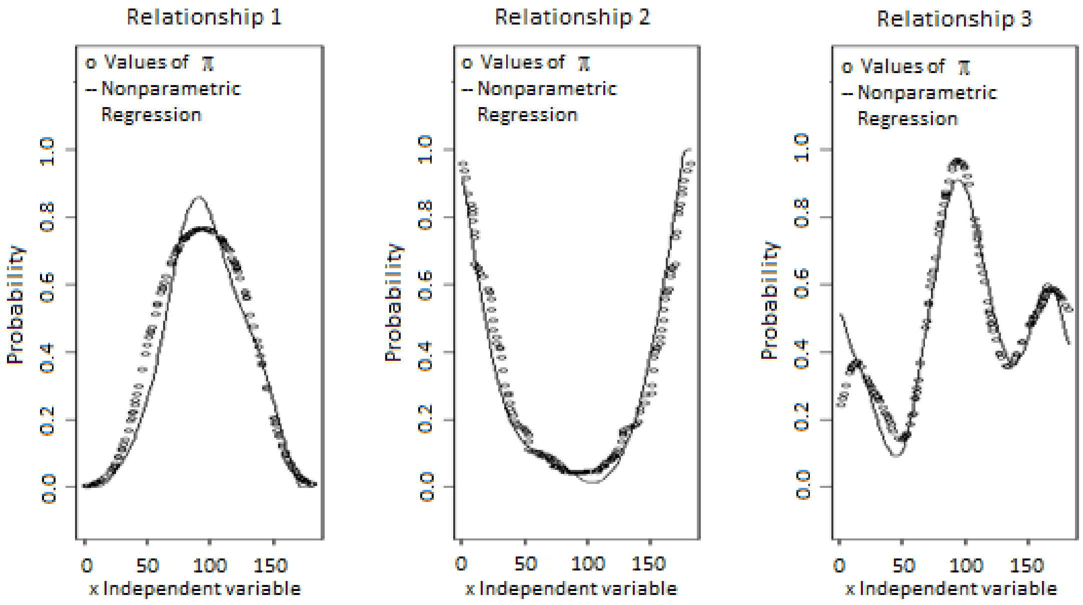
In order to assess the proposal, three different relationships between the success probability (π) and the independent variable "x" were considered. Figure 1 shows the different relationships. It was considered a set of pairs (π_i, x_i) of size 150 for each relationship. Then, from each success probability, a random number with a Bernoulli distribution was generated, obtaining a sample of 150 pairs (y_i, x_i), where y_i and x_i represent the values of the dependent variable and the independent variable respectively. This procedure was repeated 100 times for each relationship. Therefore, from each relationship were obtained 100 samples.

Figure 1. Different relationships between the success probability (π) and the independent variable.



With these samples, a non-parametric regression was computed, in order to obtain evidence of the relationship between the success probability and the independent variable. The result of the application of non-parametric regression checks that the non-parametric regression recovers the relationship that it is considered. The model suggested from the non-parametric regression for the two first relationships is a polynomial of second degree and for the third relationship one of sixth degree. The figure 2 shows, for each of the different relationships considered, the estimated probabilities ($\hat{\pi}_i$) by the non-parametric regression. Only it is shown the result of a sample of all possible.

Figure 2. Original probabilities and estimated probabilities ($\hat{\pi}_i$) by the non-parametric regression.



Subsequently, logistic regression is used in each of the different relationships, to check how well the probabilities are estimated. The models employed are given by

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

$$ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_6 x^6$$

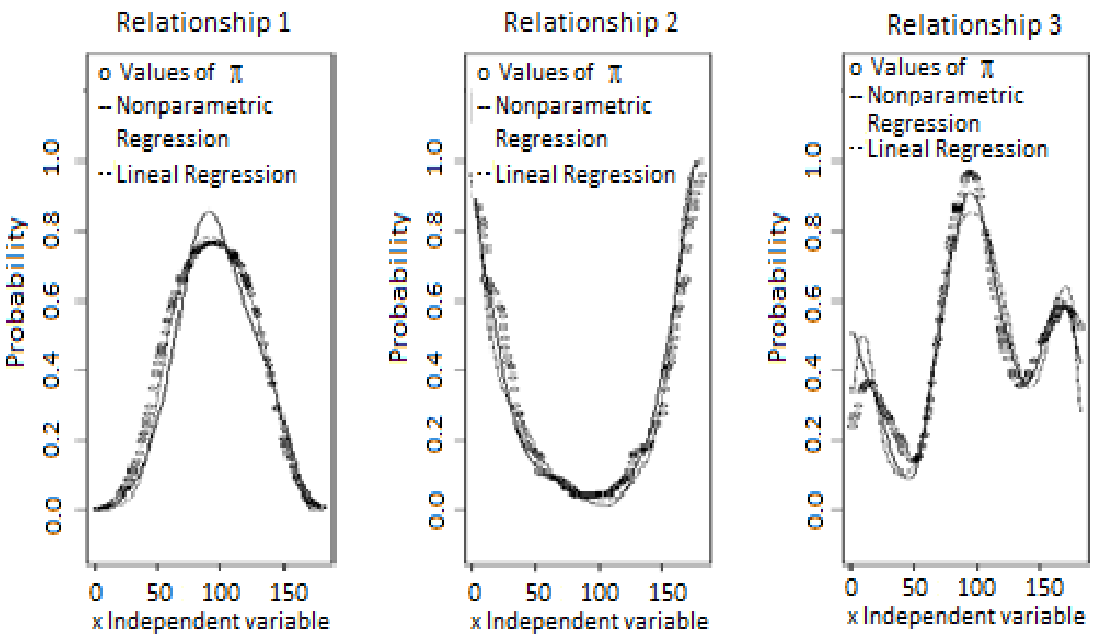
Now, with the estimated probabilities ($\hat{\pi}_i$) by the employ of a non-parametric regression, a linear regression is applied. The considered models are given by

$$E(y'|x) = \hat{\pi} = \beta_0 + \beta_1 x + \beta_2 x^2$$

$$E(y'|x) = \hat{\pi} = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_6 x^6$$

The use of a linear regression model shows the estimated probabilities ($\tilde{\pi}_i$) which are not very different from the one obtained with the use of logistic regression. Figure 3 graphically displays, for each of the different relationships considered, the estimated probabilities ($\tilde{\pi}_i$). Only it is shown the result of a sample of all possible.

Figure 3. Original probabilities, estimated probabilities ($\hat{\pi}_i$) by the non-parametric regression and the linear regression.



As a numerical evaluation to check how the probability (π_i) are estimated by the logistic regression and linear regression, we calculated the difference in absolute value from these probabilities ($\hat{\pi}_i$), ($\tilde{\pi}_i$) and those originals probabilities, by assigning the value 1 (success) if the difference was lower (0.01, 0.03 and 0.05 respectively) and 0 otherwise. Subsequently, in each sample, for each value of the probability was determined the total the differences with value 1 of the 100 samples. Tables 1, 2.a, 2.b and 2.c show the averages per sample as well as the confidence intervals for the mean. These results show similar behaviors of the logistic regression and the linear regression.

Table 1. Difference in absolute value from these probabilities ($\hat{\pi}_i$), ($\tilde{\pi}_i$) and those originals probabilities.

	Difference < 0.01		**Difference < 0.03**		**Difference < 0.05**	
Relationship	**Logistic regression**	**Linear regression**	**Logistic regression**	**Linear regression**	**Logistic regression**	**Linear regression**
1	3.30	3.52	6.80	7.02	10.32	10.56
2	3.33	3.42	10.52	9.66	17.55	15.98
3	9.25	9.60	28.09	27.83	47.52	45.60

Table 2.a. Average per sample as well as the confidence intervals for the mean.

	Confidence interval 95% (Relationship 1)			
	Logistic regression		**Linear regression**	
Difference	**Lower limit**	**Upper limit**	**Lower limit**	**Upper limit**
< 0.01	3.085	3.515	3.269	3.771
< 0.03	6.473	7.127	6.699	7.341
< 0.05	9.941	10.699	10.128	10.992

However, the non-parametric regression can produce negative values or values greater than 1. When this occurs the negative values are substituted by 0.001 and when they are greater than 1 by 0.999. All calculations were performed using the code R (Core Team, 2008) and the smoothing technique, which is named cubic spline, for applying the non-parametric regression.

Table 2.b. Average per sample as well as the confidence intervals for the mean.

	Confidence interval 95% (Relationship 2)			
	Logistic regression		**Linear regression**	
Difference	**Lower limit**	**Upper limit**	**Lower limit**	**Upper limit**
< **0.01**	2.945	3.715	3.029	3.811
< **0.03**	9.896	11.144	8.989	10.331
< **0.05**	16.780	18.320	15.056	16.904

Table 2.c. Average per sample as well as the confidence intervals for the mean.

	Confidence interval 95% (Relationship 3)			
	Logistic regression		**Linear regression**	
Difference	**Lower limit**	**Upper limit**	**Lower limit**	**Upper limit**
< **0.01**	8.536	9.964	8.856	10.344
< **0.03**	26.502	29.678	26.249	29.411
< **0.05**	45.395	49.645	43.595	47.605

3. CONCLUSIONS

The results of the study show that the proposal works properly. Users of the statistics tend to be more familiar with the linear regression and not with the logistic regression. Our proposal can facilitate the analysis of the relationship between a dichotomous dependent variable and a continuous independent variable, applying a linear regression instead a logistic regression.

Acknowledgements: The authors are grateful to the referees for their comments leading to the improvement of the paper.

RECEIVED: MAY, 2016
REVISED: NOVEMBER, 2016

REFERENCES

[1] DRAPER, N. and SMITH, H. (1998): **Applied Regression Analysis**. 3rd. ed. Wiley. N. YORK.

[2] EUBANK, R. (1999**): Nonparametric Regression and Spline Smoothing**. Marcel Dekker. New York.

[3] KEELE, L. (2008): **Semiparametric Regression for the social sciences**. Wiley. N. York.

[4] MENÉNDEZ, E. P., GABRIEL, E. and HERNÁNDEZ, S. (2015): Nonparametric regression: an alternative to the scatter diagram. **Revista Investigación Operacional**, 36, 146-150.

[5] R CORE TEAM (2008**): R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Austria.

[6] RUPPERT D., WAND P. and CARROLL R. (2003): **Semiparametric Regression**. Cambridge University Press. N. York.

[7] SEBER, G. A. F. and LEE A. (2003**): Linear Regression Analysis**. Wiley. N. York.

[8] SHEATHER, S. (2009): **A Modern Approach to Regression with R**. Springer. New York.

[9] TAKEZAWA, K. (2006): **Introduction to Nonparametric Regression**. Wiley. N. York.