ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 10(36), DOI: 10.17485/ijst/2017/v10i36/118707, September 2017* ISSN (Online) : 0974-5645

Logistic Regression: For the Identification of Socio-Economic Variables that Influence on the Academic Performance of Students of Basic and Secondary Education in the Area of Mathematics

Liliana Margarita Vitola Garrido

Facultad de Educación y Ciencias, Universidad de Sucre cra 28 # 5-267 Sincelejo, Colombia; iliana.vitola@unisucre.edu.co

Abstract

Objective: To identify the socio-economic variables that affect the academic performance of students in Basic and Secondary Education in the Area of Mathematics, by classifying the students among the groups of the Students Who Approve the Area of Mathematics and the Students Who Do Not Approve the Area of Mathematics, using a Logistic Regression Model, as a Multivariate discriminate Technique. **Methodology/Statistical Analysis:** The data were obtained with the application of a questionnaire to a sample of students, among the 6th and 11th grades, of two Educational Institutions Simon Araujo and Pre-Universidad Estudiantilin the city of Sincelejo. Two groups were previously defined: Students Who Approve the Area of Mathematics and Students Who Do Not Approve the Area of Mathematics. We calculated the probabilities that a student has to belong to the group of "Students Who Approve the Area of Mathematics". With this result, the socio-economic characteristics of "Students Who Approve the Area of Mathematics" or "Students Who Do Not Approve the Area of Mathematics" were identified. **Findings:** The results showed that at the institution educative Simon Araujo, the greater the student's age, the probability of passing the Mathematics Area decreases. However, it increases according to the educational level reached by their parents. At the Institution Educative Pre-Universidad Estudiantil, the probability that a student has to pass the Mathematics Area is higher when he or she belongs to a complete or incomplete nuclear family. Meanwhile, if it belongs to an incomplete extended family, the lowest probability is recorded. **Applications:** Once the socioeconomic variables are found, for which the probability of approving the Mathematics Area is low, teachers should identify those students who have these socioeconomic characteristics, to design and execute an improvement plan aiming at these students, in a way that prevents them from repeating the subject.

Keywords: Academic Performance, Logistic Regression, Mathematics, Socioeconomic Variables

1. Introduction

Currently, the results of school tests applied in the Educational Institutions (I.E.) of Basic and Secondary Education, show the difficulties that many students in the area of Mathematics have. In view of this situation, the Educational Institutions (I.E.) plan, execute, and evaluate improvement plans that help counteract the poor performance of students in the different areas that make up the Curriculum, including the Mathematics Area. Most improvement plans are aimed at restructuring the pedagogical practices of the teaching-learning process, guided by teachers in the classroom, without taking into account other types of factors that could possibly be affecting academic performance of students, including the social and/or economic conditions in which students are.

*Author for correspondence

Therefore, a macro project was carried out in 10 Basic and Secondary Educational Institutions, which sought to identify the variables of the student's context that could influence the academic performance of said student in the Mathematics Area. In addition, to predict the type of performance that the student will have in the Area, in accordance with the socioeconomic factors that surround him/her.

In general, to try to solve a problem and make decisions, one of the first steps is to classify the problem or situation. Then, apply the corresponding methodology, which will depend, in great way, on the classification made[1]. In order to carry out such classification, this research took the approach in which the groups or categories are known and it is intended to locate the individuals within these categories, using the Logistic Regression, also known as Supervised Technique[2]. Such a technique, compared to the discriminate Analysis used for the same purpose, yields better results, according to studies conducted by[3-5].

Due to the above, two groups or categories were defined, in which the students were located. The first is the one in which the Students Who Approve the Area of Mathematics are located. The second is the one in which Students Who Do Not Approve the Area of Mathematics are located.

As the Logistic Regression Model estimates the probability of an event as a function of a set of explanatory variables, and that any observation belongs to one of the groups or categories previously defined[2], with the help of this one we could estimate the probability that a student belongs to one of the above groups (Students Who Approve the Area of Mathematics or Students Who Do not Approve the Area of Mathematics), in accordance with the socio-economic conditions that characterize him/her.

Therefore, we observe the probability that students have to belong to the category of Students Who Approve the Area of Mathematics, according to their socioeconomic conditions. If this probability is greater than 50%, they will belong to this group. Otherwise, if the probability is less than 50%, they will belong to the group of Students Who Do Not Approve the Area of Mathematics[6]. Later, we look at the socioeconomic traits that have the students that remained in the group of Students Who Do not Approve the Area of mathematics. In this way, it is possible to design improvement plans focused on students with these characteristics so that, in this way, the probability of not approving this subject is reduced.

With regard to the use of statistical tools, applied in the identification of socioeconomic variables that influence the academic performance of a student, at the local level there are very few studies, but at the international level the work of[7] is highlighted, in his research on the curriculum of the home and the educational apprenticeships who, through a Factorial Experimental Design, found that the family-school iteration dynamically allowed to achieve high and equitable learning for children.

For its part, the research "Causal Model of Performance in Mathematics", fulfille[8], proposed a theoretical model that hierarchically collected the factors that significantly influence the performance in the Area of Mathematics.

This technique can be applied in other areas of knowledge. The present article has the results of the study applied in the educational institutions Simon Araujo and Pre-Universidad Estudiantil, of the city of Sincelejo, and was socialized in the XXVI International Symposium of Statistics, realized in the city of Sincelejo, in the year 2016.

2. Methodology

2.1 Type of Research, Study Population and Sample

An applied research with transversal design and quantitative approach was carried out, the population was made up of the students of Basic and Secondary Education, belonging to two Educational Institutions (I.E.) of the city of Sincelejo. While the sample was formed by 155 students, belonging to the I.E. Simon Araujo and 131 students, belonging to the I.E. Pre-Universidad Estudiantil, calculated with a margin of error of 5%. These Educational Institutions are part of a sample of 10 institutions that were selected from a population of 48 educational establishments in the city of Sincelejo, using the simple random sampling method. This is, because of all Educational Institutions were equally likely to be selected. On the other hand, the sample of students from each of these institutions was selected applying cluster sampling methods, proportional stratification and simple random, because each Educational Institution represents a conglomerate. Within each Educational Institution there are the levels Basic and Secondary Education. At the Basic level there are more students than at the Secondary level. Therefore,

each level represents a stratum, and within each level (stratum), students are equally likely to be selected.

2.2 Treatment of Information

In the Linear Models Theory, the Logistic Regression Model allows us to find the probability of occurrence of an event of interest, which is defined as[9]:

$$p = P(Y = 1 / X) = \frac{e^{g(x)}}{1 + e^{g(x)}} \quad (1)$$

Where (X) is:

$$\text{Logit}(p) = g(X) = \delta + \beta_1 X_1 + \ldots + \sum_{m=1}^{k_j - 1} \beta_{jm} D_{jm} + \beta_k X_k \quad (2)$$

The expression P (Y = 1 / X) is read as the "Probability that Y = 1, given the vector of independent variables" X "". Here Y is a dichotomous dependent variable, that is, it takes only two values "Y = 1 and Y = 0", according to the categories in which it is defined. For the purposes of this study, the dependent variable Y is defined as Academic Achievement in the Area of Mathematics, and has the following categories: "Y = 1: "Approve the Area of Mathematics" and Y = 0: "Do Not Approve the Area of Mathematics". The category "Y = 1: "Approve the Area of Mathematics" is the event of interest for which the probabilities P (Y = 1 / X) are calculated.

For its part, the vector of independent variables X is that which is composed of variables of socioeconomic type, among which are: Area where you live (Urban or Rural); Age; Sex; Type of Family (Nuclear Complete, Nuclear Incomplete, Extensive Complete, Extensive Incomplete, Composite, Recomposed); Stratum; You are displaced by violence (Yes or No); Number of Brothers; Maximum Academic Level achieved by their Parents (Primary, High school, Technical or Technological Courses, Professional University, Magister, Doctorate, Not Studying); Economic Activity of the Parents or Person in Charge (Employee, Unemployed, Independent); Housing Ownership (Own, Leased Family, Borrowed, Invaded); Housing Conditions (Good, Regulates, Poor); Basic Public Housing Services (Light, Water, Gas and Internet, Only Light, Water and Natural Gas, Only two of the above, Only one of the above, None of the above); In your house there were or have been problems such as: Family Fights; Sexual Abuse; Drugs; Alcohol; Abandonment by one or both parents (Yes or No).

The values **δ** and **βj** correspond to the constant and the coefficients of the Logit (p) = g (X) model, respectively, the value , represents each category in which a qualitative variable can be divided by means of a code that can be "1" or "0".

The coding of the categories of the qualitative variables was done as follows: If, for example, the qualitative independent variable is Sex, which has the Male and Female categories, its coding is:

Male = 0 Female = 1

Taking into account that there are variables with more than two categories or levels, such as the variable Economic activity of parents, which has three categories that are Employee, Unemployed and Independent, its coding was as follows:

Independent = 0 Employee = 1 Unemployed = 2

And, in this way, the codification of the qualitative variables was performed, according to the number of categories they had.

The categories of the dependent variable Y were defined in consultation with the directives of the institutions that were part of the study, the final grade in the Area of Mathematics of the students who answered the survey.

Considering that both Educational Institutions use the assessment scale from 1.0 to 5.0 points, to qualify the performance of their students, the coding was as follows:

Grades between:

- 2.9: 0 Y = 0: Not Approved 3.0 - 5.0: 1 Y = 1: Approved

In order to classify an observation or a student who has certain attributes or socioeconomic characteristics, in the group of Students Who Approves the Mathematics Area or in the group that does not approve the Mathematics Area, the probability of belonging to each one of them is calculated among these groups. Once this is done, it is assigned to the group in which it is most likely to be[9]. Or, if the probability of passing the Mathematics Area exceeds 50%, they will belong to that group. Otherwise, if it is less than 50% then it will belong to the group of thos e who do not approve the Area of Mathematics.

3. Results

3.1 Results of the Logistic Regression Model, for the I. E. Simon Araujo.

Table 1 shows the independent variables, which are significant for the Logistic Regression Model, with the

estimated values of the constant δ, the coefficients βj, of each independent variable and its corresponding estimation errors. Table 2 shows the Reliability Analysis of the Model. With the P-Value, the hypothesis test of goodness of fit of the model found, with a confidence of 95% is made.

Table 1. Estimated Regression Model (Maximum Likelihood)

Parameter	Estimated	Estimation Errors
Constant	= 21,5200	504,156
Age	-0,4430	000,102
Academic Levels= Primary	-14,0203	504,155
Academic Levels= High school	-14,7459	504,154
Academic Levels= Technician or Technologist	-13,7821	504,155
Academic Levels= Professional	-13,3593	504,155
Academic Levels= Magister	29,5544	1540,240

Table 2. Deviation Analysis

Source	Deviation	DF	P-Value
Model	39,8956	6	0,0000
Residue	134,9750	148	0,7709
Total (corr.)	174,8710	154	

Hypothesis to contrast:

H_0: The logistic model does not fit the data.
H_i:The logistic model fits the data.

Table 3 shows the significance test of each of the independent variables that make up the model. With the P-Value, we test the hypothesis of this significance, with a confidence of 95%.

Hypothesis to contrast:

The variable is not significant for the model.
The variable is significant for the model.

In accordance with the values shown in Table 1, 5 equations of Logit (p) = g (X) are generated. Each of them, corresponding to one of the following Academic Levels: Primary; High school; Technician or Technologist; Professional and Magister; and the age of the student. It should be noted that the reference variable in this case is the one corresponding to Academic Level None, that is, the one in which the parents did not have the opportunity to study. The equations are shown below:

$$\begin{aligned} Logit\,(p) &= g(x) = 21{,}52 - 0{,}44304 * X_1 - 14{,}0203 \quad \text{(Primary)} \\ Logit\,(p) &= g(x) = 21{,}52 - 0{,}44304 * X_1 - 14{,}7459 \quad \text{(High school)} \\ Logit\,(p) &= g(x) = 21{,}52 - 0{,}44304 * X_1 - 13{,}7821 \quad \text{(Technician or Technologist)} \\ Logit\,(p) &= g(x) = 21{,}52 - 0{,}44304 * X_1 - 13{,}3593 \quad \text{(Professional)} \end{aligned} \tag{2}$$

Table 3. Probability Ratio Tests

Factor	Chi- Square	DF	P-Value
Age	22,7883	1	0,0000
Academic Levels	12,5876	5	0,0276

In the variable of equations (2), arbitrary age values were assigned to each of them, for example, 12 and 17 years. By replacing the results in the equation, thereby obtaining the probabilities , which has a Student Who Approves the Area of Mathematics, according to their age and the maximum academic level reached by their parents, results shown in Table 4.

3.2 Results of the Logistic Regression Model, for the I. E. Pre-Universidad Estudiantil

Table 5 shows the independent variables that are significant for the Logistic Regression Model, with the estimated values of the constant δ, the coefficients βj of each independent variable and their corresponding estimation errors.

Table 6 shows the Model Reliability Analysis. With the P-Value, the hypothesis test of goodness of fit of the model found, with a confidence of 95%, is made.

Hypothesis to contrast:

H_0: The logistic model does not fit the data.
H_i: The logistic model fits the data.

Table 7 shows the significance test of each of the independent variables that make up the model. With the P-Value, we test the hypothesis of this significance with a confidence of 95%.

Hypothesis to contrast:

H_0: The variable is not significant for the model.

Table 4. Probabilities to Approve the Mathematics Area I.E. Simon Araujo

Age: X1=12; Academic Levels Primary: P(Y=1/ X)= 0,89	Age: X1=17; Academic Levels Primary: P(Y=1/ X)= 0,49
Age: X1=12; Academic Levels High school : P(Y=1/ X)=0,81	Age: X1=17; Academic Levels High school : P(Y=1/ X)= 0,32
Age: X1=12; Academic Levels Technician or Technologist: P(Y=1/ X)= 0,92	Age: X1=17; Academic Levels Technician or Technologist: P(Y=1/ X)= 0,55
Age: X1=12; Academic Levels Professional: P(Y=1/ X)= 0,95	Age: X1=17; Academic Levels Professional : P(Y=1/ X)= 0,65
Age: X1=12; Academic Levels Magister: P(Y=1/ X)= 1	Age: X1=17; Academic Levels Magister: P(Y=1/ X)=1

Table 5. Estimated Regressión Model (Maximum Likelihood)

Parameter	Estimated	Estimation Errors
Constant	δ= 1,70475	0,543557
Family Types= Complete Nuclear	β_{11}= 1,77649	0,900466
Family Types= Incomplete Nuclear	β_{12}= 14,86130	565,578000
Family Types= Extensive Complete	β_{13}= 14,86130	1199,770000
Family Types= Extensive Incomplete	β_{14}= -0,60614	1,276240
Family Types= Composite	β_{15}= 0,60614	0,860174

H_i:: The variable is significant for the model.

According to the values shown in Table 5, 5 equations of logit (p) = g (X) are generated. Each one corresponding to one of the following Family Types: Complete Nuclear; Nuclear Incomplete; Extensive Complete; Extensive Incomplete or Composite. It should be noted that the reference variable in this case is the one that corresponds to the Type of Family Recomposed, that is, the one that is made up of one of the parents and the stepmother or stepfather, as appropriate. The equations are shown below.

$$Logit\ (p) = g(x) = 1{,}70475 + 1{,}77649 \quad \text{(Complete Nuclear)}$$
$$Logit\ (p) = g(x) = 1{,}70475 + 14{,}8613 \quad \text{(Incomplete Nuclear)}$$
$$Logit\ (p) = g(x) = 1{,}70475 + 14{,}8613 \quad \text{(Extensive Complete)}$$
$$Logit\ (p) = g(x) = 1{,}70475 - 0{,}606136 \quad \text{(Extensive Incomplete)}$$
$$Logit\ (p) = g(x) = 1{,}70475 + 0{,}606136 \quad \text{(Composite)}$$
$$(2)$$

In a similar way to the process performed with the data from the first institution, the results were replaced in the equation, $p = P(Y = 1/X) = \frac{e^{g(x)}}{1+e^{g(x)}}$ thus obtaining the probabilities P (Y = 1 / X), which a student has to pass the Mathematics Area, according to the type of family to which he belongs. Results are shown in Table 8.

Table 6. Deviation Analysis

Source	Deviation	DF	P-Value
Model	12,3633	5	0,0301
Residue	58,3054	125	1,0000
Total (corr.)	70,6687	130	

Table 7. Probability Ratio Tests

Factor	Chi-Square	DF	P-Value
Family Types	12,3633	5	0,0301

Table 8. Probabilities to Approve the Mathematics Area I.E. Pre – Universidad Estudiantil

Family Complete Nuclear	Family Incomplete Nuclear	Family Extensive Complete	Family Extensive Incomplete	Family Composite
P(Y=1/ =0,97	P(Y=1/ = 0,99	P(Y=1/ = 0,99	P(Y=1/ = 0,75	P(Y=1/ =0,91

4. Discussion

The results found in the I.E. Pre-Universidad Estudiantil match those found[10] in the I.E. Santa Rosa de Lima of the city of Sincelejo, who, applying the concept of Logistical Regression, also found that the type of family to which a student belongs influences his academic performance in the Area of Mathematics. Only in that study, it was found that when a student belongs to a family, whether complete nuclear, incomplete nuclear, or reconstituted, the student's probability of approving the Mathematics Area is very small (0.00046%), placing the students with this characteristic, in the group of those Who Do Not Approve the Area of Mathematics. While in the I.E. Pre-Universidad Estudiantil, students who belong to families, whether complete nuclear, incomplete nuclear, extensive complete, extensive incomplete or compound, have high probabilities of Approving the Area of Mathematics, (between 75% and 99%), locating the students with any of these characteristics in the group of those Who Would Approve the Area of Mathematics.

Because the probabilities of Approving the Area of Mathematics in the I.E. Simón Araujo are higher when a student's parents have reached the academic level of masters (see Table 4.), this result is consistent, because a large percentage of parents who reach these academic levels, belong to socioeconomic groups between middle and high levels

5. Conclusions

With respect to the results of the probabilities that has a student who belong to the I.E. Simon Araujo, of Approving the Area of Mathematics, we can conclude the following:

- Setting the age, the highest probability, of Approving the Area of Mathematics, is recorded when a student's parents have reached, at most, the academic level of master, which is 100%. In addition, the probabilities of approving the Area increase when a student's parents move from one educational level to another, except when the parents have reached the high school level, where the lowest possibilities are recorded.
- Setting the academic level attained by parents, as a student is older, the probability of approving the Area of Mathematics decreases. This is because the relation between the variable age and academic performance in the Mathematics Area is the reverse.
- The results in that when a student has, for example, 17 years and their parents have reached the maximum level of primary or secondary school, the probability of Approving the Area of Mathematics is 49% and 32%, respectively. This indicates that, because they are values lower than 50%, the students of this institution with these socio demographic characteristics, are located in the group of Students Who Do Not Approve the Area of Mathematics[6]. Therefore, it would be advisable to design an improvement plan for these students, aimed at counteracting these poor results. In all other cases, the probabilities exceed 50%, which indicates that the students of this institution that have these other social characteristics, are in the group that Would Approve the Area ofmathematics. It should be noted that this does not imply that the teaching-learning process of students is neglected, so it is classified in the group that would approve the Area of Mathematics.

With respect to the results of the probabilities that a Student Would Approve the Area of Mathematics, obtained in the I.E. Pre-Universidad Estudiantil, you can conclude that:

- The probability of approving the Area of Mathematics is high, regardless of the type of family to which a student belongs, ranging from 75% to 99%. However, the lowest probability of Approving the Area of Mathematics is recorded when a student belongs to an incomplete extended family of 75%.

- A student's probability of passing the Area of Mathematics is higher when he/she belongs to a complete extended or incomplete nuclear family (99%).
- The results in that the probability of approving the Mathematics Area exceeds 50%, regardless of the type of family the student belongs to. This could indicate that, if there are any students Who Do Not Approve the Mathematics Area, then it would be due to another type of factor that would be worth identifying.

6. References

1. Pineda SCS. Comparacion de arboles de regresion y clasificacion y regresion logistica. Universidad nacional. 2009; 1-60.
2. Webb AR, Copsey KD. Statistical pattern recognition. Wiley. 2011 November; p. 1- 642.
3. Bull SB, Donner A. The Efficiency of multinomial mogistic megression mompared With multiple group discriminant analysis. Journal of the american statistical association. 1987 December; 82(400):1118-22. Crossref
4. Press SJ, Wilson S. Choosing between logistic regression and discriminant analysis. Journal of the american statistical association. 1978 December; 73(364):699-705. Crossref
5. Barajas HF, Correa Morales JC. Comparacion entre tres tecnicas de clasificacion comparison for three classification techniques. Revista colomb estadistica diciembre. 2009; 32(2):247-65.
6. Clark LS, Pizarro SR. Curriculum del hogar y aprendizajes educativos: Interacción versus status. Revista de psicologia universidad de chile. 1998; 7:25-34. 8. Illarramendi AM. Modelo causal del rendimiento en mate maticas. Ensenanza de las ciencias. 1998; 16(1):11-12.
7. Hosmer DW, Lemeshow S, Sturdivant RX. Applied logistic regression. Wiley authenticity guarantee. 2013 March; p. 1-528.
8. Garrido LV. Regresion logistica: una aplicacion en la identificación de variables que inciden en el rendimiento académico, en el area de matemáticas. Revista educación y desarrollo social. 2015 December; 9 (1):118-131.