DOI: https://doi.org/10.24297/jam.v18i.8557

Implementation of The Logistic Regression Model and Its Applications

Gladiola Trushaj[1] Elmira Kushta[2]

[1]Teacher at Alidemi Vlore high School

[2]Department of Mathematics, Faculty of Technical Sciences University of Vlora

[1]gtrushaj@gmail.com [2]kushtamira@gmail.com

Abstract

The purpose of an analysis using this method is the same as that of any technique in constructing models in statistics, namely to find the best and most reasonable model to describe the relationship between a result variable and a set of variables independent. We are interested in how the costs affect them and if a customer has a travel card.

Credit card customers are shown to be 6 times more likely to use it regardless of the cost they make. It is also shown that a customer is more likely to use a travel card when costs increase Through logistic regression, which gives the probability that a result is an exponential function of the independent variables, we will see how through our data we will come to very important conclusions.

Keywords: Logistic Regression, Chances, Binary Data Analysis.

Introduction

Regression methods have been an integral part of any analysis of data related to the description of the relationship between a response variable and one or more explanatory variables. Logistic regression model is a form of regression which is used when the dependent variable (Y) is dichotomous (binary) and qualitative values i.e., 2 levels e.g., with or without customer card, etc., and the independent variable (X s) can be numerical, categorical or mixed.

This technique is applied in different research area, particularly in the medicine (e.g., Antonogeorgos G. et al., 2009), socio-logical sciences (e.g., HL Chuang, 1997) and in the education (state graduation results) (e.g., Sadri Alija, et al., 2011, Muca M., et al., 2013).

The purpose of an analysis that uses this method is the same as that of any technical building models in statistics, so to find the best and more reasonable model to describe the connection between a variable result and a set of variables independent. In logistic regression after adjusting the model, the emphasis is to evaluate and interpret is o the coefficients and their values, where conformity assessment methods are more of a technical nature.

The estimated coefficients for the independent variable present growth (i.e. the boundary change) a function of the dependent variable per unit change in the independent variable. Thus, interpretation involves two issues:

1 Determining the functional relationship between the dependent variable and independent.

2 Determining the unit change in the independent variable.

2 The logistic regression equation

The logistic regression model has the same form of the regression model, which is used when the dependent variable Y is qualitative (binary) with two values or categorical with more than two values and the independent variables can be quantitative, qualitative or mix. In multiple regression analysis, the mean or the expected value of y, E(Y), is calculated from equation (1)

$$E(Y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \qquad (1)$$

While, in the logistic regression is shown (Chao-Ying Joanne Peng and Tak-Shing Harry So, 2002) the connection between E(Y) and $x_1, x_2, \cdots, x_k$ that is described by nonlinear equation (2)

$$E(y) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)} \qquad (2)$$

If the dependent variable y is coded as 0 and 1, the value of E(y) in equation (2) provides the probability that y = 1 given a particular set of values for the independent variables $x_1, x_2, \cdots, x_k$ (see appendix). Because of the interpretation of E(y) as a probability p, the logistic regression equation is given in equation (3).

$$\text{E(Y)=P}(Y = 1 / \tilde{X} = \tilde{x}) = p(y) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)} \qquad (3)$$

Materials and Methods

3 Data. Linear regression analysis logistics LR

It is believed that the total costs in a supermarket and the fact that a customer has a credit card or not, are two important variables in predicting if a client who has client card will use it. Therefore, in this study we used a sample which consists of information for n = 100 clients (50 have travel card and a travel card 50 have not). The set of data consists of a dependent variables and forecast variables (predictor variables). For every customer the total costs marked 1000 ALL and {0,1} are listed whether or not the customer has credit card (see Table 1). Logistic regression model will be used to assess what kind of potential customer is to benefit from the promotion of the company.

Table 1. Description of features for the database community

Variables	**Description**
Use travel card	{0 $\to$ don't use and 1 $\to$ use}
Total costs	Are in "thousand ALL (Leke)"
Use credit card	{0 $\to$ don't use and 1 $\to$ use}

Figure 1, shows that customers can be grouped as follows:

• Customers who do not use travel cards (78%, 39 from 50) spend less (to 10 thousand) than those who buy with credit card.

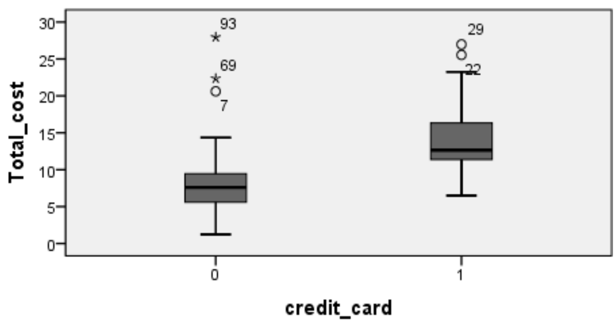

Figure 1. Graphical presentation according to the type of payment.

4 Logistic Regression Analysis

The linear logistic regression model was adapted into the set of the data presented in Table 1. In Table 2 are represented the results of RL method performed in SPSS.

$$P(Y=1/\underset{\%}{X}=\underset{\%}{x})=\frac{\exp(-1.965+0.035\text{Total\_Cost}+1.794\text{Credit\_}Cart)}{1+\exp(-1.965+0.035\text{Total\_Cost}+1.794\text{Cr}edit\_Cart)} \quad (4)$$

According to the model, the natural logarithm of odds that a client use client card is positively associated with both variables (cost and Credit card) with p = .467 and p = .001 respectively. In other words, more money a client has spent, the greater will be the chance of being a user (that a client used client cart). Exp(B) for the total cost (expenditure) is equal to 1.036, which means when the total costs are added to a unit (1000 ALL), the client is 1.036 times more likely to use the card customer than not, or if the difference in domestic currency (ALL) is 20 thousand then a client is 2.03 = 1.036^{20} times more likely to use the travel card. Customers who use credit cards are more likely to use client card compares to they that not have travel card because credit card customers are codified 1 while customers who have no credit card are codified with 0. Actually, the odds ratio for clients with a travel card (odds) who use the customers card is 6.015 (=$e^{1.794}$ Table 2) times greater than the odds ratio for customers who have no credit card. Also based on the 95% confidence interval for the expenditure we see that it contains the value 1 and in such cases it is said that change in expenditures is statistically insignificant, and for other intervals that do not contain the value 1 the opposite is true.

Table 2. Results of the logistic regression analysis

Classification Table[a,b]

			Predicted		
			kartë_udhëtimi		Percentage Correct
Observed			.00	1.00	
Step 0	kartë_udhëtimi	.00	15	0	100.0
		1.00	9	0	.0
	Overall Percentage				62.5

a. Constant is included in the model.

b. The cut value is .500

We can now use equation (4) to estimate the probability of using the customer card for a particular type of customer. Thus, for this group of customers, the probabilities of using or not using the customer card when his total expenses are equality as some numerical indicators (minimum, maximum, mode, and the average costs) are shown in Table 3.

As you can see the tables display in SPSS works by steps, where in the zero interation we see that the tables are simplified because they do not yet take into account the data I modify in the respective windows, and the first interation includes all details as a result of tacing into account data from windows that SPSS extracts.

Table 3.Assessing the probability according to some numerical characteristics

		Minimum =1.235	Average costs =11.366	Mode =12.640	Maximum =27.957
Credit card	1	46.81%	55.65%	56.74%	69.16%
	0	12.77%	17.26%	17.91%	27.16%

Results and Discussion

Binary logistic regression interpretation (binary logistic regression)

In Table 4 are shown the ratios of two conditional probability (equation 8) for customers using the customer card when the total of her expenses as some numerical indicator (minimum, maximum, mode, and the average cost).

Table 4. Assessment report two conditional probability by some numerical characteristics

Classification Table[a,b]

Observed			Predicted: kartë_udhëtimi .00	Predicted: kartë_udhëtimi 1.00	Percentage Correct
Step 0	kartë_udhëtimi	.00	15	0	100.0
		1.00	9	0	.0
	Overall Percentage				62.5

a. Constant is included in the model.

b. The cut value is .500

The multiple logistic regression model has the form (1)

$$Ln(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k \quad (1)$$

Therefore

$$P = prob(Y = 1 / X_1 = x_1, \dots, X_k = x_k) = \frac{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)},$$

$\beta_0, \beta_1, \dots \beta_k$ they are the regression coefficients which are estimated by the maximum likelihood method, X-s are a set of predictors, p is the probability of occurrence of the eventA={Y=1/X=x}.

$Y = X\beta + \varepsilon$ it is a linear model with $Y \sim B(p)$. If $E(\varepsilon) = 0$ then $E(Y) = 1 * p + 0 * (1 - p) = p$ (in general) and $E(y_i) = x_i^{'}\beta = p_i$ (in particular). From the last equation shows that the mean of the response to the linear function $Y = X\beta$ is equal to the probability of the event A={Y=1/X=x}.

Some problems encountered in linear model:

1. If the response is binary, then the error terms can take on two values, namely,

$$\varepsilon_i = \begin{cases} 1 - x_i'\beta, & y_i = 1 \\ -x_i'\beta, & y_i = 0 \end{cases}$$

Because the errors is dicreate, normaly assuption is violated.

2. The error variance is not constant and it can be seen that it is a function of probabilities, because

$$D(y_i) = E(y_i - E(y_i))^2 = (1 - p_i)^2 * P(y_i = 1) + (0 - p_i)^2 P(y_i = 0)$$

$$D(y_i) = (1 - p_i)^2 * p_i + p_i^2 (1 - p_i) \; D(y_i) = (1 - p_i) * p_i = E(y_i)(1 - E(y_i))$$

where $E(y_i) = x_i'\beta = p_i$

Therefore the assuption of homoscadasticity does not hold.

From the last equation shows that the variance response (that It is equal to with the error variance because $\varepsilon_i = y_i - p_i$, where p_i an constant) is a function of waiting Mathematical and $0 \le E(y_i) = p_i \le 1$. Therefore the assuption of homoscadasticity does not hold.

From the last equation shows that the variance response (that It is equal to with the error variance because $\varepsilon_i = y_i - p_i$, where p_i an constant) is a function of waiting Mathematical and $0 \le E(y_i) = p_i \le 1$. Last inequality It causes problems if elected linear model.

The interpretation of β s is rendered using either the odds ratio (for categorical predictors) or the delta-*p* (for continuous predictors). The null hypothesis states that all β s equal zero. A rejection of this null hypothesis implies that at least one of the parameters does not equal zero in the popullation, which means that the logistic regression model predicts the probability of the answer better than the mean of the answer Y (see Item 2 to some problems encountered in linear model) that determined by draw (2).

$$\mathrm{P}(Y = 1 / \underset{\%}{X} = \underset{\%}{x}) = p(y) = \frac{\exp(\beta_0 + \beta_1 x_1 + \mathrm{L} + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \mathrm{L} + \beta_k x_k)} \quad (2)$$

Results and Discussion

Conclusions

The analysis showed that customers who use credit cards are more favorable to use customers card compared to those who don't have credit card and the chances increase when total cost rise. Also, it is seen that a credit card customer is 6 times more favorable to use the customer card compared to those who don't have customer card.

References

1. Bewick V, Cheek L., Ball J: Statistics review 8: Qualitative data–tests of association. Crit Care 2003, 8:46-53

2. Chao-Ying Joanne Peng and Tak-Shing Harry So, (2002). Logistic regression Analysis and Reporting: A Primer. Understanding Statistics, l(1), 31-70. Lawrence ErlbaumAssociates, Inc

3. D.W. Hosmer, S. Lemeshow (1989). Applied Logistic Regression. Willey, New York, .

4. George Antonogeorgos**,** Demosthenos B. Panagiotakos, Kostas N. Priftis and Anastasia Tzonou, (2009). Logistic Regression and Linear Discriminat Analysis in Evaluating Factors Associated ëith Asthma Prevalence among 10- to 12-Years –Old Children: Divergence and Similaritity of the Two Statistical Methods**.** Hindawi Publishing Corporation, International Journal of Pediatrics.

5. Muça M., PUKA Ll., BANI K, SHAKAJ F. (2013) "Logistic Regression Analysis: A Model to Predict the Entrance Probability in Higher Education"**.** "1st International Western Balkans Conference of Mathematical Sciences – IWBCMS-2013", në Elbasan/ALBANIA PROCEEDINGS

6. Sadri Alija, Lazim Kamberi and Llukan Puka, (2011). Logistic regressions and an application in teaching practice assessment. In AKTET, Journal of Institute Alb-Shkenca. Volume IV, 8.

7. 'Multivariate data analysis'---Joseph F. Hair, Barry J. Babin, Rolph E. Anderson, William C. Black, ISBN: 0-6952-2541-4

8. 'Applied multivariate statistical analysis (sixth edition)'---Richard A. Johnson & Dean W. Wichern, 0-13-187715-1

9. 'Logistic regression and discriminant analysis'---Gillez Stoltz, 2002, ISBN: 0-2156-2155-1

10. 'Statistics_Understanding regression analysis'---Apelt Stome, 2000, ISBN: 0-1689-7485-3