Methodology

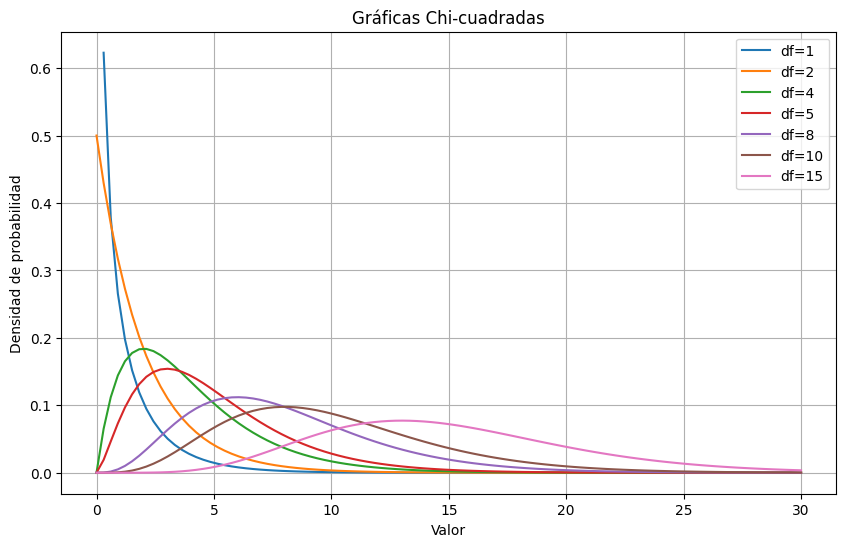
When considering categorical variables, we used the Chi-square test (χ2) as a statistical test to determine the existence of a significant relationship between the dependent variable and each of the independent variables that will be considered in the study. Since this is a non-parametric test, it is not necessary to verify that the variables have a normal distribution.

The hypotheses are:

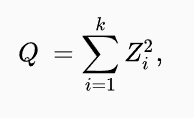
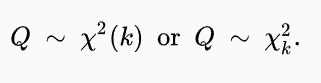
♦ Null hypothesis (Ho): The two variables are independent, which means there is no relationship between them.

♦ Alternative hypothesis (Ha): The two variables are not independent, that is, there is a relationship between them.

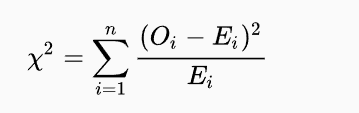
The shape of the χ2 distribution is positive and asymmetric.



This probability distribution represents that if *Z* 1 , ..., *Z k,* are independent variables and with Laplace-Gauss normality standards [[1]](#footnote-1)of random variables, then the sum of their squares is distributed according to the χ2 distribution, with a number of k degrees of freedom, designated as the number of random variables or independent responses in each of the questions compared, where by increasing the number of these with normality standards, their distribution is linked to a normal distribution [[2]](#footnote-2).



Now, basing the study on the search for correlation between variables using the χ2 test, the observed frequencies are compared with the expected frequencies under this hypothesis of independence (Ho), where the χ2 value is obtained from the sum of the quotients between the squares of the differences between the observed value and the expected value, with respect to the expected values for each category.



χ2 = Pearson cumulative test statistic, which asymptotically approximates the Chi-square distribution.

= Number of observations of type i

= Expected (theoretical) frequency of type i = (row total x column total) / grand total

= Number of cells in the table

This statistic follows a χ2 distribution with (rows – 1)( columns – 1) degrees of freedom. Referring to the rows and columns of the contingency table.

The assumptions that must be met for the value of this statistic to be statistically reliable in the study of independence between variables are:

1. That the variables are categorical. That is, that the values of the observed frequencies come from qualitative variables.
2. That each observation must be independent. That is, the same instance or individual must not belong to another record evaluated in the study.
3. That the expected frequency of each cell must be at least 5. In other words, that each crossing between the possible responses of both variables has an expected value (E) greater than or equal to a frequency of 5.

Once the criteria of the χ2 distribution and its link to the asymptotic approximation of the Pearson cumulative test statistic to a χ2 distribution have been described, we begin the search for independent variables that demonstrate contrasting evidence of dependence on our dependent variable.

However, there are variables that do not meet the expected value assumption with a frequency of at least 5 for each cell in the contingency table.

For these cases, the Fisher test [[3]](#footnote-3)provides a more accurate equivalent [[4]](#footnote-4)than the χ2 test when the number of events expected per level of each cell is small, that is, less than 5. Therefore, those variables that demonstrate a significant relationship by this means are evaluated through hypothesis contrast evidence with the Fisher test.

The hypotheses for this statistical test of independence are:

* Null hypothesis (Ho): The variables are independent so one variable does not change between different levels of the other variable.
* Alternative hypothesis (Ha): The variables are dependent, that is, one variable changes between different levels of the other variable.

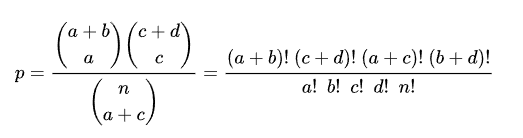
Where the assumptions that must be met for the correct implementation of this test are:

1. Independence among observations. As with Pearson's cumulative test statistic, each observation contributes only to a single level of the categorical variable, represented in the contingency table.
2. Fixed marginal frequencies. That is, there are no alterations in the total frequencies of the rows and columns, when obtaining different combinations between the levels of each variable.

Under these assumptions and the null hypothesis of independence between the variables, the calculation of the exact probability of obtaining the observed data from a contingency table follows a hypergeometric distribution. This distribution describes the probability of obtaining a given number of successes with a sample without replacement for a finite population.

For a general contingency table with two dichotomous variables

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature A** | | | |
| **Feature B** | Present | Absent | Total |
| Present | to | b | a + b |
| Absent | c | d | c + d |
| Total | a + c | b + d | n |

The exact probability of obtaining any specific combination with these conditions is given by the hypergeometric distribution:

Where is the binomial coefficient [[5]](#footnote-5)and the symbol “!” represents the factorial operator.

**Cramer's degree of association V**

Cramer's V is a metric used to measure the level of association between two qualitative variables. The function that calculates this coefficient is:

χ2 = Pearson cumulative test statistic, which asymptotically approximates the Chi-square distribution.

n = Total number of instances considered.

r = Number of rows in the contingency table.

c = Number of columns in the contingency table.

The assumption necessary to implement this relationship measure on the variables is:

* + - * 1. The χ2 statistic must be able to be calculated. This means that all the previous criteria necessary for the use of this test must be met [[6]](#footnote-6).

The range of values of this coefficient is [0, 1] [[7]](#footnote-7), where the effect size of the variables is classified as:

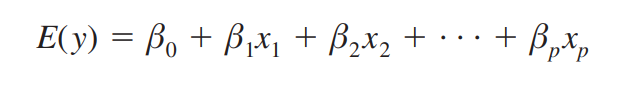
|  |  |
| --- | --- |
| **Effect size (ES)** | **Interpretation** |
| **ES ≤ 0.2** | The result is weak. Although the result is statistically significant, the fields are only weakly associated. |
| **0.2 < ES ≤ 0.6** | The result is moderate. The fields are moderately associated. |
| **EN > 0.6** | The result is strong. The fields are strongly associated. |

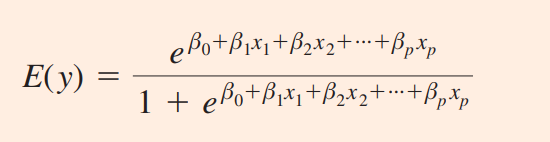
*Source: IBM Congos Analytics*

**Logistic regression**

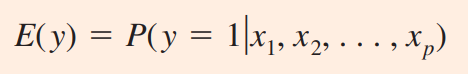
When the values that the dependent variable can assume reflect the occurrence of an event or its absence, given a set of chosen independent variables, logistic regression allows us to estimate the probability of this event occurring.

“Logistic regression is similar in many ways to regular regression. It requires a dependent variable y, and one or more independent variables. In multiple regression analysis, the mean or expected value of y is referred to as the multiple regression equation” (Anderson et al., 2012)

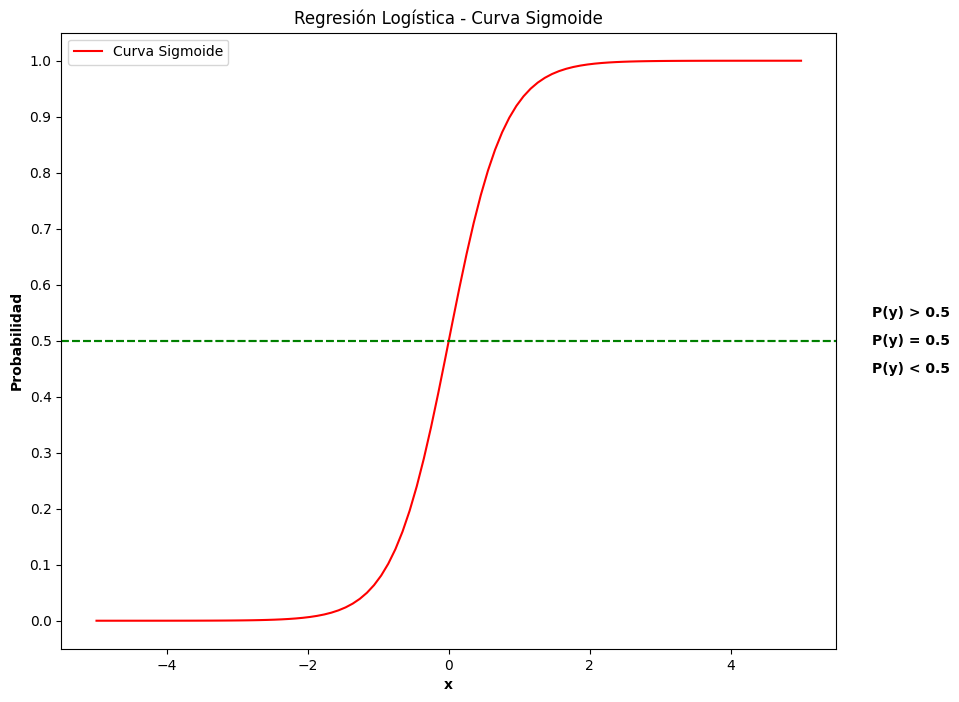


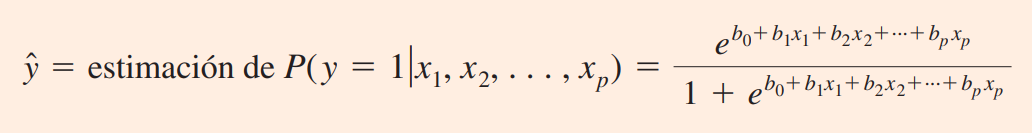
However, this relationship between the expected value of the dependent variable, with respect to the selected set of independent variables, for the logistic regression, shows a non-linear relationship, both in practice and in theory.

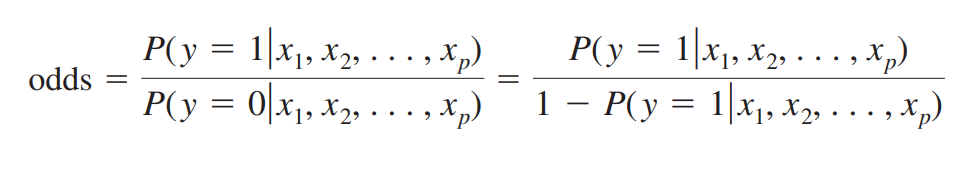
In addition, the values of occurrence and absence of the event are coded, respectively, as 1 and 0, for the value obtained by this function. This indicates that the expected value, which is obtained with the non-linear relationship, obtained for the logistic regression, provides the probability that the dependent variable is coded as an occurrence event (1), based on the values of the independent variables of the logistic model.

This is why the expected value of the logistic regression is expressed as an equation of probability of occurrence

Which is linked to the non-linear function, expressed above, of the expected value of the logistic regression, where the exponential relationship with the independent variables and their beta coefficients is shown [[8]](#footnote-8), as well as the estimated coefficient for the constant term .



Based on this, the probability estimate will represent the estimated logistic regression equation, on the occurrence of the event, as the representation of the coefficients with the notation of estimated coefficients beta .

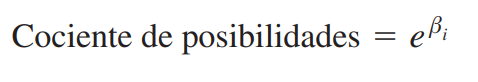
For the interpretation of the estimated coefficients, since it is a non-linear relationship between the dependent variable and the combination of this set of independent variables, the individual relationship of each coefficient is investigated through its impact in favor of its occurrence on the desired event (y = 1), caused by a unitary change in one of the independent variables.[[9]](#footnote-9)

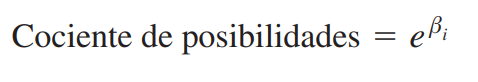
In addition to the possibility function in favor of the event of the study occurring, the possibility quotient is obtained [[10]](#footnote-10), which provides the relationship between the possibility of the event occurring, due to the unitary change in one of the independent variables ( ) and the possibility of the event occurring, without the unitary change in said independent variable ( ).



This provides information about the direct effect of a one-unit increase in one of the independent variables. That is, the degree of possibility that the event will occur due to the individual, unitary increase in that independent variable.

These functions are linked to the beta coefficients of the independent variables of the model, because for every independent variable of the logistic regression, “there is a unique relationship between the odds ratio of a variable and its corresponding regression coefficient” (Anderson et al., 2012).





**Instance Classifier Model Performance Metrics**

The performance metric “ precision ” indicates the proportion of correctly classified instances, for each class, with respect to the total number of instances estimated as that class.

Instances correctly classified as class i

Instances incorrectly classified as class i.

recall ” performance metric indicates the percentage of correctly classified instances, relative to the total number of real instances of that class. This formula is identical to the True Positive Rate (TPR).

Instances correctly classified as class i

Instances incorrectly classified as class i.

Now, the performance metric “f1-score”, this is calculated as the harmonic mean of the “ precision ” and “ recall ” metrics for each of the classes.

Finally, the “ accuracy ” metric measures the accuracy of the model, that is, the proportion of instances of both classes that have been correctly classified, with respect to the total number of instances in the study.

Instances correctly classified as class i

Instances correctly classified as class j

Instances incorrectly classified as class i

Instances incorrectly classified as class j

1. Referring to a continuous probability distribution, where the values belong to real numbers, with a mean of 0 and a standard deviation of 1. [↑](#footnote-ref-1)
2. The fulfillment of these assumptions denotes the relationship of the Chi-square distribution with the normal distribution, as the number of random variables that meet these conditions increases. [↑](#footnote-ref-2)
3. Statistical method that allows evaluating the relationships between two qualitative variables when the assumption of expected values >= 5 in each of the cells of the contingency table is not met. [↑](#footnote-ref-3)
4. It is an exact calculation of the deviation, rather than an asymptotic (infinite) approximation. [↑](#footnote-ref-4)
5. Combinations that correspond to the number of possible subsets, from the given set. [↑](#footnote-ref-5)
6. Assumptions linked to the χ2 statistic. [↑](#footnote-ref-6)
7. This allows it to be an alternative to the C contingency test, which only approximates the degree of association to 1, but never reaches this value. [↑](#footnote-ref-7)
8. Magnitude of the perceived impact of the dependent variable on the unit change of an independent variable, keeping the other independent variables constant. [↑](#footnote-ref-8)
9. The probability in favor of an event occurring is defined as the “probability in favor of the event occurring , relative to the probability of this event not occurring ” (Anderson et al., 2012) [↑](#footnote-ref-9)
10. Also known as odds ratio. [↑](#footnote-ref-10)