1. Logistic Regression

In order to evaluate the probabilistic distribution of the level of consistency through a generalized multilinear model with binary dependent variable, defined by:

Where, , is a cumulative distribution function (*CDF*), is the residual term with and follows a Bernoulli distribution of parameter

Then,

The vector represents the independent variables for the observation. Finally, is the vector of coefficients to be estimated.

Therefore, the problem consists of estimating based on observations .

In our context of study, the dependent variable is qualitative with binary outcomes, the most commonly used *CDF* is that of the logistic distribution defined as:

That’s why, in this particular situation, our regression is called “Binary Logistic Model” (Cox 1958), where the probability, for observation , to be above a certain level of consistency will be estimated by:

However, equation (1) is not usable unless we estimate the parameters . To do this, we use the classical principle of maximum likelihood. In our case, the likelihood function to maximize is given by:

Where, and are the observed values of the variables , .

By applying the logarithmic function, we will obtain the log-likelihood function:

Thus, the estimation of is done by maximizing the log-likelihood through solving the system of partial derivatives:

Solution of system (3) is obtained by the iterative method of Newton-Raphson (Kendall 1989) and the estimated parameters will be denoted by .

In our application, we consider the following independent variables: “*time in minutes*”, “*temperature in degree Celsius*” and “*ratio*”. On the other hand, the level of consistency is considered as dependent variable.

Therefore, an explicit expression of the logistic model is:

Where, is the intercept coefficient.

Now, after estimating all the coefficients of the model and considering statistically significant ones, we can compute the probability that the consistency is above a certain apriori fixed threshold for a given time, temperature and ratio.

Analytically, these probabilities are given by:

Where, the index is added to indicate the observation with .

1. Application:

Before applying the logistic model to our set of data, we start by fixing a threshold for the consistency above which the binary dependent variable is considered to be equal to 1.

To do that, we will focus on approximating the increasing trend of the consistency around a certain consistency value fixed based on the previous graphical analysis.

Based on a second order Taylor expansion and using numerical approximation, one can approximate the derivative of a function at a point by:

Where is the length of the step between the points , and .

Then, by applying equation (4) we choose to approximate the increasing trend of consistency when the function reaches a level close to 40. The results, for different values of ratios, are presented in the below table:

|  |  |  |  |
| --- | --- | --- | --- |
| Ratio | 2.08 | 2.16 | 2.21 |
| Consistency level | 40.08 | 40.43 | 40.65 |
| approximation | 2.64 | 13.85 | 8.16 |

Based on the results of the above table, we remark that the trend around a consistency level 40 is large enough to consider 40 as apriori threshold to build the binary dependent variable of the logistic regression model.

Now, by applying the maximum likelihood principle and considering only the significant independent variables we get the following estimators for the different parameters:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coefficients | Estimate | Std. Error | z value | Pr(>|z|) |
| Intercept | 197.57181 | 26.06251 | 7.581 | 3.44e-14 \*\*\* |
|  | 0.10215 | 0.01252 | 8.159 | 3.36e-16 \*\*\* |
|  | -98.30019 | 12.75197 | -7.709 | 1.27e-14 \*\*\* |

Note that indicates statistically significant 0.1% level. The negative sign of indicates that when the ratio increases the probability that the consistency is above 40 decreases, which fits with logically expected results. In addition the positive sign of indicates that when the time increases the probability that the consistency is above 40 increases, which also fits with logically expected results. However, we remark that the variable  *“Temperature”* is not considered as a significant variable and this is mainly caused by the high correlation between the variables and which is given by .

1. Validation of results:

A good logistic regression model is the one who maximize the percentage of positives (i.e. the cases where ) that are successfully classified as positive (called specificity) and the percentage of negatives (i.e. the cases where ) that are successfully classified as negatives (called sensitivity). For example, in our application, if we consider a threshold (based on empirical observations) of (i.e. if we consider ) we obtain the following table that summarizes the performance of our model:

|  |  |  |  |
| --- | --- | --- | --- |
| Predicted  values of  Observed  values of | 0 | 1 | Total |
| 0 | 3304 | 637 | 3941 |
| 1 | 18 | 40 | 58 |
| Total | 3322 | 677 | 3999 |

Then, for this particular threshold, the sensitivity is and the specificity is . However, a more powerful criteria to evaluate the global performance of a logistic regression without fixing any threshold is the Receiver Operating Characteristic (ROC) curve. ROC Curve gives us an idea on the performance of the model under all possible values of threshold by plotting Sensitivity in terms of specificity. the model is as good as the curve is close to the optimal situation (i.e. the point of coordinates (1,1)). In other words when the area under the ROC curve is close to 1.

In our context, the ROC curve is given by figure (1). The area under the curve is (computed using R). Then we can say that our model has a global prediction accuracy of .

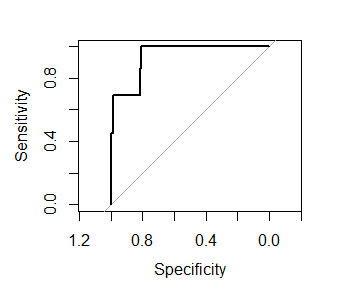


Figure 1: ROC curve of logistic model

Finally, note that we can now compute the probability of a consistency level greater than 40 for any given time and ratio by applying the following equation: