Israel Nolazco

israelnolazco@lewis.edu

Overview of Logistic Regression

August 17, 2020

Table of Contents

# Background.

# Executive Summary.

# Analysis

* 1. Estimation of the Logistic Regression and Assessing Overall Fit
  2. Classification Results

## Summary

1. Background.

The purpose of this project is to perform various techniques of Logistic Regression Analysis in order to further understand the relationship between several independent variables compare to single nonmetric dependent variables. During the Logistic Regression analysis we will use 13 independent variables that percept variables from the database (x6 to x18) to discriminate between firms in each geographic area evaluated in the categorical variable in x4

The variables on the  HBAT \_with SPLITS\_ revised.xls and HBAT.xls dataset are presented in the table below:

* •  ‘x4 Region’
* •  ‘x6 Product Quality’
* •  ‘x7 E-commerce’
* •  ‘x8 Technical Support’
* •  ‘x9 Complaint Resolution’,’x10 Advertising’
* •  ‘x11 Product Line’
* •  ‘x12 Salesforce Image’
* •  ‘x13 Competitive Pricing’
* •  ‘x14 Warranty & Claimns’
* •  ‘x15 Packaging’
* •  ‘x16 Order & Billing’
* •  ‘x17 Price Fexibility’
* •  ‘x18 Delivery Speed’

The analysis will then use a logistic regression model that will weigh the probability for each variable and access those weight to determine where customer’s Region along with determining what variables provide the greatest factor in determining the category. The next section provides a high-level synopsis of details that are presented in the remainder of this document. Additionally, the information was analyze using the programming language SAS.

1. Executive Summary

This document presents an analysis of Logistic Regression Analysis that were obtained from the HBAT \_with SPLITS\_ revised.xls and HBAT.xls datasets. Within the aforementioned dataset we going to join both datasets to create a testing sample. It is important to ensure randomness in the selection of the holdout sample so that nay ordering of the observations does not affect the processes of estimation and validation. This measures how significant the dependent variable Region is truly a categorical variable and if actually plays a role in the overall analysis.

1. Analysis
   1. Estimation of the Logistic Regression Model and Assessing Overall Fit

**A screenshot of a cell phone

Description automatically generated**

**Figure1.** Number of variables used along with frequency

A Logistic Regression follows two key steps one is to set a training sample and the other is the validation sample. In this case Figure 1. Shows are training sample with only taking sixty out of the one hundred samples. As seen in the image, we have used and read all of the sixty variables. Additionally, Group 0 is corresponded to the USA/North America group whilst Group 1 is corresponded to the Outside North America group.

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

Figure 2. Logistic Regression Results

As seen in Figure 2 the Logistic Regression model provides an overall view as to the meaning behind the data we have chosen. Right away the likelihood ratio is 80.69. This means that the overall model fits the data at an 80% rate. Therefore, there are variables that could be eliminated to have a better fit. This is where the value of the Wald Chi-Square comes into play.

* 1. Classification Results for Two-Group Discriminant Analysis

The Wald Chi-Square value as seen in Figure 2. provides the level of significance the variable play in the overall fit. Variables such as x8, x10 have no significance in the over logistic regression model. However, variables such as x11 and x17 play a huge significance in the overall model. As explain in the summary these two variables correspond to Product Line(x11) and Price Flexibility (x17) which are crucial variables depending the region the item is being purchase from.

![A screenshot of a cell phone

Description automatically generated]()

![A screenshot of a social media post

Description automatically generated]()

Figure 3. Odds Ratio Estimates and Classification Table

An interesting piece of information gather by SAS is the odds ratio estimates. In essence Figure 3 demonstrates the odds the variables are going to be successful and failures based on the Wald Confidence. As we can see variables such as x14 and x15 shows a high Point of Estimate; therefore, stating that those two variables are more likely to create a failure or not create any data regarding our classification. Additionally, looking at what those variables are compose of X14(Warranty and Claims) and X15(New Products) we can visually see why they do not play a role in the region. Lastly, the classification Table then counts and provides a percentage rate of the accuracy the events were label. In this case the model only predicted about 86.7% of the variables with a high confidence. In theory by removing variables like x14 and x15 will increase our confidence and a far more accurate model.

1. Summary

Logistic Regression Analysis seem highly accurate in classifying and evaluating variables in a binary classification. Additionally, this model highlighted variables that are more likely to affect the model whilst additionally provide the accuracy in the classification table. In theory, the ability to add and remove variables could lead to a far accurate model. However, unlike Multiple Discriminant Analysis we are again at the mercy of a binary category; which depending on the type of problem could present its own challenges.