Israel Nolazco

israelnolazco@lewis.edu

A Summary of Logistic Regression

August 16, 2020

Table of Contents

1. What is it?
2. *Even in a binary situation, why pick Logistic Regression over Discriminant?*
3. Best Suited Situation
4. Assumption of Logistic Regression
5. Relationships
6. Validations

**What is it?**

Logistic Regression is a model that specializes in a binary (two-group) categorical variable rather than a metric dependent measure.  Its variate is similar to regular regression and made up of metric independent variables.  It is less affected than discriminant analysis when the basic assumptions, particularly normality of the independent variables, are not met.  This is due to the binary nature of this type of analysis.

**Even in a binary situation, why pick Logistic Regression over Discriminant?**

The discriminant analysis assumes multivariate normality and equal variance-covariance matrices across groups, and these assumptions are often not met.  Logistic regression does not face these strict assumptions and is much more robust when these assumptions are not met, making its application appropriate in many situations. Even if the assumptions are met, some researchers prefer logistic regression because it is similar to multiple regression.  It has straightforward statistical tests, similar approaches to incorporating metric and nonmetric variables and nonlinear effects, and a wide range of diagnostics.

**Best Suited Situation**

Logistic Regression must first start by identifying the independent variables that impact group membership in the dependent variable.  Then, we must establish a classification system based on the logistic model for determining group membership. Keep in mind that the binary nature of the dependent variable (0 – 1) means the error term has a binomial distribution instead of a normal distribution, and it thus invalidates all testing based on the assumption of normality. Additionally, the variance of the dichotomous variable is not constant, creating instances of heteroscedasticity as well. Finally, neither of the above violations can be remedied through transformations of the dependent or independent variables.  Logistic regression was developed to specifically deal with these issues.

**Assumption of Logistic Regression**

The advantages of logistic regression are primarily the result of the general lack of assumptions. Logistic regression does not require any specific distributional form for the independent variables. Additionally, the heteroscedasticity of the independent variables is not required. Furthermore, linear relationships between the dependent and independent variables are not required.

**Relationships**

During the analytical research keep in mind that a positive relationship means an increase in the independent variable is associated with an increase in the predicted probability and vice versa. But the direction of the relationship is reflected differently for the original and exponentiated logistic coefficients. Original coefficient signs indicate the direction of the relationship. Exponentiated coefficients are interpreted differently since they are the logarithms of the original coefficients and do not have negative values.  Thus, exponentiated coefficients above 1.0 represent a positive relationship and values less than 1.0 represent negative relationships. The magnitude of metric independent variables is interpreted differently for original and exponentiated logistic coefficients:

•Original logistic coefficients – are less useful in determining the magnitude of the relationship since the reflect the change in the logit (logged odds) value.

•Exponentiated coefficients – directly reflect the magnitude of the change in the odds value.  But their impact is multiplicative and a coefficient of 1.0 denotes no change (1.0 times the independent variable = no change).

**Validations**

After making a logistic model, we must keep in mind one very important question; what does this mean?  In order to make proper validation it is important to ensure both the internal and external validity of the results. The most common form of estimating external validity is the creation of a holdout or validation sample and calculating the hit ratio. A second approach is cross-validation, typically achieved with a jackknife or “leave-one-out” process of calculating the hit ratio.