Logistic Regression

A B M Abir Mahboob
Department of Electronic Engineering
Hochshule Hamm Lippstadt
Lippstadt, Germany
a-b-m-abir.mahboob@stud.hshl.de

***Abstract*—In this paper we will be looking at concepts related to logistic regression, as well as the statistical significance of individual regression coefficients. In this process we always get a binary outcome, which can have two values such as yes/no or true/false. When looking at modeling approaches we can use logistic regression to define a relationship between independent and dependent variables. We could use logistic regression to figure out which new sample fits where best therefore it can be a useful method for classification problems. We will also be looking at how robust logistic regression is as compared to linear regression and how it has helped solved problems that linear regression was not able to until now.**

***Index Terms*—**

I. Introduction

When the dependent variable is dichotomous, logistic regression is the best regression strategy to use. The logistic regression, like other regression studies, is a predictive analysis. To describe data and explain the connection between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables, logistic regression is utilized.

Logistic regression, like contingency table analyses and x2 tests, provides for the investigation of binary or dichotomous outcomes with two mutually exclusive levels. However, logistic regression allows for the inclusion of both continuous and categorical factors, as well as the adjustment of numerous predictors. This makes logistic regression particularly effective when analyzing observational data and adjusting for any bias caused by variations between the groups being compared.

For a two-level outcome, ordinary linear regression might provide unacceptable results. Some covariate values are likely to have predicted values that are either above the upper level (typically 1) or below the lower level (usually 0). (usually 0).

Furthermore, the validity of linear regression is contingent on the outcome's variability being the same for all predictor values. The behavior of a 2-level result does not fit this premise of continuous variability. As a result, linear regression is insufficient for such data, and logistic regression was created to fill the void. A recent example of the use of logistic regression in Circulation is the assessment of gender as a predictor of operative mortality after coronary artery bypass grafting surgery, a meta-analysis of the relationship between the TaqlB genotype and the risk of cardiovascular disease, and an investigation of the relationship between lipoprotein abnormalities and the incidence of diabetes.

A. Model of Logistic Regression

The probability of a 2-level outcome of interest form the basis of the logistic regression model. For the sake of simplicity, I'll assume that one of the result levels has been selected as the event of interest, and refer to it as the event in the following text. The ratio of the chance of the event occurring divided by the probability of the event not occurring is the event's odds. Odds are frequently employed in gambling, and ”even odds” (odds=1) indicate that the event will occur half of the time. When rolling an even number on a single die, this is the case. Because rolling a number 5 is twice as often as rolling a 5 or 6, the odds for rolling a number 5 are 2. The reciprocal is used to find symmetry in the chances, and the odds of rolling at least a 5 are 0.5 (=1/2).

The natural logarithm of the odds is used as a regression function of the predictors in the logistic regression model. With only one predictor, X, this is written as ln[odds(Y=1)]=0+1X, where ln stands for natural logarithm, Y is the outcome and Y=1 when the event occurs (versus Y=0 when it does not), 0 is the intercept term, and 1 is the regression coefficient, or the change in the logarithm of the event's odds with a 1-unit change in the predictor X. Because the difference in the logarithms of two values equals the logarithm of the ratio of the two values, we can get the odds ratio corresponding to a 1-unit change in X by calculating the exponential of 1.

Odds ratios often are used in the analysis of 2-by-2 contingency tables6 and case-control studies. The odds ratio is sometimes confused with the relative risk, which is the ratio of probabilities rather than odds. Only when the probability of the event is very low can the odds ratio be considered a good approximation to the relative risk. The odds ratio is more extreme than the relative risk, which leads to exaggeration of the effect of a predictor when it is misinterpreted as a relative risk. In many settings, the relative risk is preferred over the odds ratio because it addresses the more readily understood probability of the event rather than its odds. However, logistic regression results are typically presented by odds ratios because these are the natural estimates from the model and attempts to transform these to relative risks can distort the results.10 A useful way to think of the odds ratio is that 100 times the odds ratio minus 1, ie, 100×(odds ratio1), gives the percent change in the odds of the event corresponding to a 1-unit increase in X. If this value is negative, then the odds of the event decrease with increasing values of X; if positive, the odds increase. This percentage change is the same for any 1-unit increase in X because of the assumed linearity between X and the logarithm of the odds in the regression model above. For some continuous predictors, this assumption may not match the data,11 in which case careful checking of the model results is required. For example, if the logarithm of the odds against the predictor X has a U shape (both low and high values have large odds of the outcome relative to the intermediate values) and the model assumes a linear (straight line) pattern, then goodness-of-fit checking should show that the model and the data are not compatible. In such a case, splitting the predictor values into categories and using dummy variables to code for the categories may improve the fit. Other methods such as splines also may be used to lessen the assumption of linearity. When adjusted values are needed, more predictors can be added to the right side of the regression equation above, along with corresponding regression coefficients (). In this case, the odds ratio value for X would be adjusted for the other predictors in the model. The equation above, 100×(odds ratio1), would then be interpreted as the percent change in the odds corresponding to a 1-unit increase in X while holding all other predictors fixed. The selection of appropriate predictors to reduce confounding and to improve the precision of estimates is done similarly for logistic regression and for linear regression; guidelines can be found in many statistical textbook. Unlike linear regression, there is no formula for the estimates of for logistic regression. Finding the best estimates requires repeatedly improving approximate estimates until stability is reached. This is done easily on a computer, and there are many statistical software packages that perform logistic regression, but it makes logistic regression less understandable and more of a “black box” approach for many researchers.

II. Prepare Your Paper Before Styling

Reference

Peng, J. (2002, September). (PDF) an introduction to logistic regression analysis and reporting. ResearchGate. Retrieved April 6, 2022, from An Introduction to Logistic Regression Analysis and Reporting