Name – Gaurav

Roll No. – 40821046

Logistic Regression

Purpose: Introducing logistic regression , a multivariable method for modeling the relationship between multiple independent variables and a categorical dependent variable

Methods : LR and research articles employing LR as main statistical analysis were reviewed from the book

Results: Logistic regression such as odds, odds ratio, logit transformation and logistic curve, assumption, fitting, reporting and interpreting to cautions were presented

Introduction

Multivariate analysis refers to simultaneously predicting multiple outcomes andmultivariable analysis uses multiple variablesto predict a single outcome

CONCEPTS RELATED TO LOGISTIC REGRESSION

Logistic regression analyse the relationship between multiple independent variables and a categorical dependent variable, and estimates the probability of occurence of an event by fitting data to a logistic curve.There are two models of logistic regression, binary logistic regression and multinomial logistic regression. Binary logistic regression is typically used when the dependent variable is dichotomous and the independent variables are either continuous orcategorical. When the dependent variableis not dichotomous and is comprised of more than two categories, a multinomial logisticregression can beemployed

Odds

Odds of an event aretheratio of the probability that an event will occur to the probability that it will not occur. If the probability of an event occurring is p, the probability of the event not occurring is (1-p). Then the corresponding odds is a value given by odds = p 1-p

Odds=P/1-P

The impact of independent variable is explained in term of odd

p=P (Y=interested outcome/X= χ, a specific vlaue) =

e^a+βχ/ 1 + e^a+βχ =1/ 1 + e^-(a+βχ)

. Odds ratio

The odds ratio (OR) is a comparative measure of two odds relativeto different events.

Odd ratio = odds(A)/odds(B) = P(A)/odds(B)

The Logestic Curve

A simple logistic function is defined by the formula =

1/1+e^-(a+bx)

where a and b determinethelogistic intercept and slope. Logistic regression fits a and b, the regression coefficients. Figure 1 showslogisticfunction when α and β are 0 and 1,respectively.Thelogistic orlogit function is used to transform an ‘S’-shaped curveinto an approximately straight line and to changetherange of the proportion from 0 – 1 to -∞ - +∞ as

logit(y)= ln (P/1-P) =a + bx

Assumptions In LR

The dependent variable for logistic regression must be discrete and primarily dichotomous. Second, since a logistic regression assesses the likelihood that an event will occur (P(Y=1)), the dependant variable must be coded appropriately. That is the desired result, which ought to be coded as 1. Third, the model needs to be well fitted. With the added useless factors, it shouldn't be overfit. Additionally, it shouldn't be underfitted with omitting relevant variables. Fourth, each observation must be independent for logistic regression. Furthermore, there should be little to no multicollinearity in the model. In other words, independent variables do not relate to one another linearly.

Fifth, while a linear link between the dependent and independent variables is not necessary for logistic regression, it does require that the independent variables have a linear relationship with the log probability of an occurrence. Finally, because maximum likelihood demands large sample sizes, logistic regression

FITTING THE LOGISTIC REGRESSION MODE

The likelihood function is used to estimatethe probability of observing thedata, giventhe unknown parameters(a andb).A “likelihood” is a probability that the observed values of the dependent variable may be predicted fromthe observed values of the independent variables.The likelihood varies from 0 to 1 like any other probabilities

EVALUATION OF A LOGISTIC REGRESSION MODEL

1. Overall model evaluation
2. The likelihood ratio test

A logisticregression model with the k independent variables (the given model) is said to provide a better fit to the data if it demonstrates an improvement over the model with no independent variables

(the null model). The overall fit of the model with k coefficientscan beexamined via a likelihood ratio testwhich teststhe null hypothesis

= -2 log (likelihood of the null model /likelihood of the given model)

1. Chi-Square Goodness of Fit Tests

With logisticregression, instead of R2 as thestatistics for overall fit of thelinear regression model, deviance between observed values from the expected values is used.

1. Hosmer-Lemeshow test

The Hosmer–Lemeshow test is to examine whether the observed proportions ofevents aresimilar to the predicted probabilities of occurrence in subgroups of the model population.

1. Statistical significance of individual regression coefficients
2. Likelihood ratio test

The likelihood-ratio test used to assess overall model fit can also be used to assess the contribution of individual predictorsto a given model. The likelihood ratio test for a particular parameter compares thelikelihood of obtaining the datawhen the parameter is zero (L0)with the likelihood (L1) of obtaining the data evaluated at the MLE of the parameter. Thetest statisticiscal culated as follows:

G = -2 (ln L0 -ln L1)

1. Wald statistic

The Wald statisticcan be used to assess the contribution of individual predictors or the significance of individual coefficients in a given The wald statistic is the ratio of the square of the regression coefficient to the square of the standard error of the coefficient. The Wald statistic is asymptotically distributed as a Chi-square distribution.

1. Odds ratios with 95%

CI Odds ratio with 95%confidence interval (CI) can be used to assess the contribution of individual predictors It isimportant to note however, that unlikethe p value, the 95% CI does not report a measure’s statistical significance. It is used as a proxy for the presence of statistical significanceif it does not overlap the null value(e.g. OR=1). The 95% CI is used to estimatethe precision of the OR. A large CI indicates a low level of precision of the OR, whereas a small CI indicates a higher precision of the OR

1. Validation of the logistic regression

The validity of model should be assessed by carrying out tests of goodness of fit and discrimination on a different data set . If the model is developed with a sub sample of observations and validatedwith the remaining sample, it is called internal validation. The most widely used methods for obtaining a good internal validation are data-splitting, repeated data-splitting.

CAUTIONS AND CONSIDERATIONS

In logistic regression no assumptions are made about the distributions of the independent variables. However, the independent variables should not be highly correlated with one another because this could cause problems with estimation. Studies with small to moderate samples size employing logisticregression overestimate the effect measure. Thus, large sample sizes are required for logistic regression to provide sufficient numbers in both categories of the outcome variable. Also, the more independent variables were included, the larger the sample sizeis required. With small sample sizes.