Research Paper Statistics

An Outline to Binary Logistic Regression Analysis

Kanika Grover | Student(M.Sc Statistics), Amity University

Shyama Gupta | Student(M.Sc Statistics), Amity University

ABSTRACT

The purpose of the article is to provide an outline of the building and applications of the Binary Logistic Regression model. The key highlights of the concept have been demonstrated well in the theory as well as illustration applied to the data set in testing a research hypothesis. The model strategies and the inferences of the model are all explained with the help of theoretical content, tables and graphs.

KEYWORDS | Binary data, regression, categories, logistic, strategies

Introduction

Researchers and educationists have usually resorted to regression models for prediction of quantitative variables. The question arises when statistical analysis has to be carried out for qualitative variables where data is classified into categories, where prediction is based on a dichotomous outcome. For example- whether a patient will die or not after being diagnosed of cancer at last stage, if the class topper is either a male or a female, if it will rain next week or not etc. In earlier times, OLS regression was opted to analyze such variables.

Owing to the statistical assumptions of such methods, linear regression analysis was not found suitable for dichotomous outcomes. Such outcomes or variables are either response or explanatory. Logistic Regression is thus a new technique where such variables can be represented in the form of ordered scores so as to run the analysis and find the predicted values. The variables under study can either be in the form of binary or multinomial data. In this paper we would lay emphasis on studying the building and interpreting the logistic models for binary data.

This paper aims at providing an outline to the structure of logistic model along with basic statistical inferences for analysis of categorical variables. A data has been sourced from *D.Collet, Encyclopedia of Biostatistics, Wiley, New York* to run a basic model and interpret its results.

Interpretation of the Logistic Regression Model

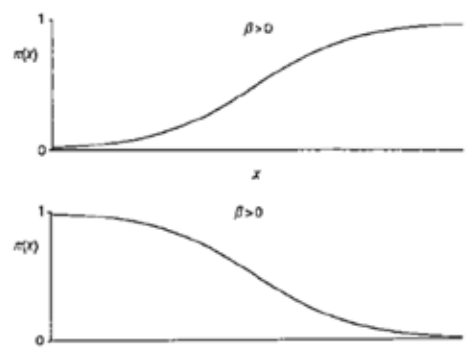
The distribution associated with studying a binary form of data is binomial distribution. In binomial distribution, the probability of success can be denoted by *p*. Let *Y* be the response variable and *X* be the continuous predictor variable, thus for the probability of success π at value *x*, linear form of the logistic regression model is given by the central mathematical concept of *logit* i.e. the natural logarithm of an odd ratio.

$$logit(Y) = natural\log(odds) = \ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta X \quad \ldots(1)$$

The probability of success using the above formula can be obtained as:

$$\pi(x) = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)} \quad \ldots(2)$$

Fig. 1- Graph for Logistic Regression Model



The graph and the formula show that π(x) increases or decreases as an S-shaped function of *x*.

The equation (1) implies that the logit of *Y* increases by β for every unit increase in *X*. It also determines the rate of increase or decrease of S-shaped curve of π(x). For β=0, π(x) is constant at all values of x, the curve ascends for β>0 whereas it descends for β<0.

Logistic Regression with Categorical Predictors

Like ordinary regression models, logistic regression models can also have multiple explanatory variables. These variables are categorical and thus they are qualitative in nature. Suppose for a binary response variable *Y,* there are two binary predictors given as *X* and *W*. The model is thus given as-

$$Y = \alpha + \beta_1 x + \beta_2 z \quad \ldots(3)$$

Statistical Inferences for Logistic Models

1) Confidence Interval for Effects- Confidence Intervals for parameter β is $\hat{\beta} + z_{\frac{\alpha}{2}}(S.E)$
2) Testing of Significance- The hypothesis can be stated as: H_0: β=0 i.e. probability of success is independent of X. Thus for large samples, z=β̂/SE

H_a: β≠0. $Z^2=(\hat{\beta}/SE)^2$. This can be approximated as large sample chi-square distribution with 1 degree of freedom.

3) Confidence Interval for Probabilities:

$$\hat{\pi}(x) = \exp(\hat{\alpha} + \hat{\beta}x)/[1 + \exp(\hat{\alpha} + \hat{\beta}x)]$$

Strategies for Model Selection
The selection process for a particular model becomes more challenging as the number of explanatory variables increases. Two competing goals arise: simpler models are easy to interpret and complex model are difficult to fit. Searching many models may provide a clue to select predictor variable associated with response.

a) Number of Predictors to be used in the model:
In binary data Y can take the value either 1 or 0 which limits the number of predictors. One guideline suggests that there should be at least 10 outcomes of each type for every predictor. Even if it does not satisfies, software still fits the model which may leads to biased ML estimates and poor standard errors. Models of two or more predictor variables may suffer from *Multicollinearity*- correlation among the predictors. A variable may linearly predicted from other with a major degree of accuracy. Deleting such variable may be helpful to get better estimates.

b) Stepwise selection of Variables
Forward selection- add term sequentially until further additions do not improve the fitted model. *Backward elimination*- eliminating terms sequentially. Eliminate the term in the model that has largest P-value in the test. The process stops when further elimination leads to significantly poorer fit. Statistically significant should not be the sole criteria in selection of particular variable. If a variable is important for studies but not statistically significant, one should report the estimates.

c) AIC and correct model
Before fitting a model one should look that a simple model has the advantage of model parsimony and if a model is little bias but describing reality well, provides good estimates of outcomes probabilities and of odd ratios that describe the effects of predictors. Another criterion to select a good model is Akaike information criterion. It weights model fits against parsimony-

AIC = -2(log Likelihood – number of parameters in model)

The model with lowest value of AIC should be selected.

d) Diagnostic- R^2
R is the correlation between the n binary $\{y_j\}$ observation and estimated probabilities . It is useful to compare the fitted model for the same data.

Ways of checking model fit
a) Model for Likelihood Comparison
It helps in detecting lack of fit by comparing the model with more complex ones. A more complex model might contain non linear effect. Usually, model with multiple predictors might have interaction terms. If the complex model do not fit better than the chosen model is better.

b) Deviance and Goodness of fit
This test compares the model fit with the data. Let us define the model as M. In testing the fit of M, we test whether all the parameters are in saturated model but not in M equal to 0. Saturated model is the most complex model possible, which has a separate parameter for each observation. When the predictors are in categorical form, the data are summarized by counts in contingency table. The deviance statistics can be calculated for all the cells in the table as:

$G^2(M) = 2 \sum$ observed [log (observed/fitted)]

Pearson Statistics is:
$X^2(M) = \sum$ (observed - fitted) ^{2} / fitted

When the fitted values are at least 5, Deviance statistics and Pearson statistics have approximate chi-square null distribution. The degree of freedom is called residual degrees of freedom for the model. This can be calculated by subtracting the number of parameters in the model from the number of parameters in the saturated model. Large values of any test provide an evidence of lack of fit.

Application of Logistic Regression
The data has been sourced from *D.Collet, Encyclopedia of Biostatistics, Wiley, New York*. The aim is to study whether a patient undergoing a surgery after being exposed to general anesthesia experiences a sore throat or not. The factors taken into consideration are the duration of the surgery (in minutes) and the type of device used to secure the airway: laryngeal mask airway and tracheal tube.
Hypotheses

Null Hypothesis: Patient having the surgery with general anesthesia experiences sore throat on waking, it does not depend on duration of the surgery and the types of devices used.

Alternative Hypothesis: Patient having the surgery with general anesthesia experiences sore throat on waking, it depends on duration of the surgery and the types of devices used.

Statistically,
H_0: β_i's = 0; (i=0,1,2,3)

H_a: At least one β is not equal to zero

Data Analysis of the model
Table 1: Binary Logistic Regression Model for Interaction Effect of Predictors

```
> library(ISLR)
>
glm.out=glm(sore.
throat~device*duration,family=binomial(logit),data=logistic)
> summary(glm.out)
```

```
Call:
glm(formula = sore.throat ~ device * duration, family =
binomial(logit),
    data = logistic)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7360  -1.2530   0.7988   0.9248   1.3015

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.74997    1.19302  -0.629    0.530
device           1.81087    1.53359   1.181    0.238
duration         0.03087    0.02733   1.129    0.259
device:duration -0.03635    0.03152  -1.153    0.249
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 45.004  on 34  degrees of freedom
Residual deviance: 43.276  on 31  degrees of freedom
AIC: 51.276
Number of Fisher Scoring iterations: 4
```

```
> anova(glm.out,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: sore.throat

Terms added sequentially (first to last)


                Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                            34     45.004
device           1  0.12093        33     44.883   0.7280
duration         1  0.15073        32     44.732   0.6978
device:duration  1  1.45604        31     43.276   0.2276
```

```
> confint(glm.out)
                    2.5 %     97.5 %
(Intercept)     -3.32100931 1.52607914
device          -1.10711851 5.02245001
duration        -0.01736861 0.09460498
device:duration -0.10622585 0.02183569
```

Table 2: Binary Logistic Regression model for Linear Effect of Predictors

```
> glm=glm(sore.
throat~device+duration,family=binomial(logit),data=logistic)
> summary(glm)
Call:
glm(formula = sore.throat ~ device + duration, family =
binomial(logit),
    data = logistic)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7305  -1.3820   0.8589   0.9281   1.0165

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.312954   0.754470   0.415    0.678
device      0.228390   0.722054   0.316    0.752
duration    0.005205   0.013572   0.383    0.701

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 45.004  on 34  degrees of freedom
Residual deviance: 44.732  on 32  degrees of freedom
AIC: 50.732

Number of Fisher Scoring iterations: 4
> anova(glm,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: sore.throat
Terms added sequentially (first to last)


         Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                        34     45.004
device    1  0.12093        33     44.883   0.7280
duration  1  0.15073        32     44.732   0.6978
> confint(glm)
                 2.5 %     97.5 %
(Intercept) -1.20695118 1.81151431
device      -1.18532493 1.68644574
duration    -0.02055247 0.03439477
```

After analyzing the two outputs, it can be inferred that the AIC value for the binary logistic regression model for linear effect of predictors (Table 2) is low as compared to the binary logistic regression model for interaction effect of predictors (Table 1). The value of AIC is dependent on the number of parameters in the model. The model with linear effect of predictors is thus considered to be simpler compared to the other one hence considered to be the best fit to the data.

Coefficients:

a) Device: For every one unit change in the device used, there is an increased chance of i.e 0.06% chance of having sore throat of the patient.
b) Duration: : For every one unit change in the duration of the surgery, there is an increased chance of i.e 0.01% chance of having sore throat of the patient.

Deviance:

In order to analyse the overall performance of the model, look at the values of null deviance and residual deviance in the model (Table 2). The value of Null deviance will show how well the response is predicted by the model. Adding the two predictors-Device and Duration in this case, the deviance of the model gets decreased by 1.728 points on 3 degrees of freedom. The residual deviance is 43.276 at 31 degrees of freedom.

Confidence Interval:

The confidence interval in Linear Effects is greater as compared to the Interaction Effects. Thus, the model defined in Linear Effect approaches more to the Null Hypothesis as it increases the Acceptance Region.

P-value:

```
> 1-pchisq(1.728,df=3)
[1] 0.6307268
```

The p-value above results in acceptance of null hypothesis i.e the Patient having the surgery with general anesthesia experiences sore throat on waking; it does not depend on duration of the surgery and the types of devices used.

Summary

The paper shows the building and application of the Binary Logistic Model as a powerful technique for the analysis of categorical variables. The model over a past few decades has been gaining significant importance and can be easily accessed with the help of statistical softwares available in the market. The application used in the paper has been performed with the help of R. The paper has aimed to answer all the basic questions regarding to the logistic model of regression and aims to provide guidance to the masses.

REFERENCES

i) An introduction to Categorical Data Analysis, Second Edition. By Alan Agresti Copywright 2007 John Wiley and Sons, Inc. | ii) An Introduction to Logistic Regression Analysis and Reporting, Chao-Ying Joanne Peng, Kuk Lida Lee, Gary M. Ingersoll; Indiana University-Bloomingston | iii) Applied Logistic Regression, Second Edition, By David W. Hosmer, Stanley Lemeshow, Copyright 2000 John Wiley & Sons, Inc. | iv) Statistics for Biology and Health, By Klienbaum, David G., Klien, Mitchel Copyright 2010