From the Expert: Logistic, Multinomial and Polynomial Regression

From the Expert: Logistic, Multinomial, and Polynomial Regression

Multiple linear regression is a flexible technique and can used to find the relationship between continuous dependent variables and independent variables, which may be continuous, categorical, or dichotomous. Nevertheless, there are many problems (e.g. social science, business, medicine, etc.) that require qualitative response such as classifying votes (e.g. yes, no), classify on-line transactions (e.g. fraud, not fraud), classify patients according to symptoms (e.g. Hepatitis A, Hepatitis B, Hepatitis C, Hepatitis D, Hepatitis E).

There are various types of regression that are appropriate to particular data types or data relationships such as logistic regression which is suitable when the dependent variable is dichotomous (e.g. yes, no), multinomial logistic regression is an extension of logistic regression when the dependent variable is categorical data with more than 2 categories, and polynomial regression is used when the relationship between independent variables (IV) and dependent variables (DV) is well expressed with polynomial terms (e.g. .

1. Logistic Regression

The outcome variable for logistic regression is dichotomous, usually coded as 1 or 0. When 1 refers to presence of a characteristic, and 0 is absence of a characteristic. The value produced from logistic regression is the probability value in the range of 0.0 to 1.0.
The assumptions about normality, linearity, and homogeneity of variance are not held in logistic regression.

In linear regression, coefficients of the independent variables are computed using the least squares method. But logistic regression uses *maximum likelihood* estimation to compute the coefficients. The maximum likelihood is an iterative process that tries to get closer and closer to the correct answer. The logistic slope coefficients are interpreted as one unit change in x affects the log of the odds when other variables in the model are held constant. How well the model fits is measured by likelihood value, which is similar to residual or error sum of squares in multiple linear regression. A small likelihood value indicates that a model fits the data well. Thus, the ideal or perfect model would have a likelihood value of zero.

The overall test of the relationship between independent and dependent is based on the difference in the likelihood values of the model that does not include independent variables, and the model that does include independent variables. This likelihood difference follows a chi-square distribution and is also known as the model chi-square. The relationship between the dependent variable and combination of independent variables is determined by the significance test for the model chi-square.

Examples of logistic regression usage include 1) model the probabilities as the function of explanatory variables 2) predict probabilities of cases 3) classify cases based on explanatory variables 4) social network analysis (e.g. is friend with, talks to, asks advice from). More examples in
http://courses.education.illinois.edu/EdPsy589/lectures/5logreg1_02.pdf

The outcome variable in linear regression is known as logit. Here, the natural log (base e) is used to convert logit to probability and vice versa.

1.1 Definition of logit

1.2 The logistic regression equation

The left hand side is called logit or log-odds. Logistic regression model has a *logit* that is linear in *x*. In addition, coefficients can be interpreted as the amount of change in the log- odds for one unit increase in predictor *x*.

1.3. Odds and Probability

Odds are just a different way to express the likelihood (probability) of an event. Odds are the likelihood of occurrence relative to the likelihood of non-occurrence. You can change back and forth from odds to probability.

The odds of an event are formulated as:

Odds (= = =

Therefore,

Or,

Example, suppose the probability of success is 0.9, so
Then probability of failure is
Therefore, odds (success) = (the odds of success is 9 to 1)
odds (failure) = (the odds of failure is 1 to 9)
odds Ratio (OR) = (odds of success are 82 times higher for failure)
and (odds of failure are twelve-thousandth odds of success)

logit is referred to log odds or natural log(e) of an odds. If odds = 1, then logit =0.

Therefore, odds, probability and logit are different ways to express the same thing.

Example from: http://www-hsc.usc.edu/~eckel/biostat2/notes/notes14.pdf
Assume that, an association between gender and smoking status is assessed in a study.
The primary predictor () is gender, and coded as 1 for men, 0 for women
The outcome () is smoking status, and coded as 1 for current smokers, 0 for non-smokers.

In simple linear regression:
is the mean outcome when gender is woman ()
is the mean outcome when gender is man ()
is the difference in the mean outcome when and

In Logistic regression model:

Suppose , and = 1

For women, (gender = 0) :

For men, (gender = 1) :
 is the log odds of smoking for women
 is the log odds of smoking for men
 is the difference in log odds of smoking comparing men to women

Transform from log odds to odds interpretation:
From then

For women, :
The odds of smoking for women is: = = 0.05

For men, :
The odds of smoking for men is: = 0.12

If the logit is linear in *x, then* refers to the change in log odds for a 1 unit increase in *x*.
 is the odds ratio corresponding to 1 unit increase in *x.*
Therefore, (corresponding to a 10 unit (e.g. year) increase in *x* (e.g. age).

One important property of odds ratio is that it is constant.

Suppose the logistic regression model is -1.2345 + 0.456 + 0.02
The effect of the odds of one unit increase in is exp(0.456) = 1.58. This means that, increases the odds by 58%, regardless of the value of

Odds ratio is a term for comparing the odds under one condition to the odds under another condition. For example,

Odds Ratio =

The odds of smoking is 2.4 times greater for men than women

As a summary:
To find the log odds:
For :
For :

To find odds:
For :
For :

To transform odds to probability:
For :

For :

Relative risk for men and women = =

Lastly, slope in a logistic regression is

1. A difference in log odds related to one unit change in *x* (fix other independent variables)

2. A log odds ratio is related to one unit change in *x* (fix other independent variables)

1.4 Assumptions

Logistic regression has different assumptions from traditional regression such as no homogeneity of variance, and no normality (errors are not distributed normally). Here are logistic regression assumptions:

- Independence of cases - Each case should be independent of other cases
- Non-Linearity - Outcome and predictors have no linear relationship but logit of outcome and any continuous predictors has a linear relationship
- No multi-colinearity - No predictor should be a linear function of other predictors, and predictors should not be too closely related to each other. Estimation is impossible for perfect multi-colinearity, whereas, strong multi-colinearity cause imprecise estimation.
- No complete separation - The value of one variable cannot be perfectly predicted by one variable or a set of variables. This usually happens when there are empty cells (or values are zero) on category variables in the model.

1.5.1 Overall Model fit

In logistic regression, there are several ways to evaluate model fit

1.5.1 Likelihood ratio test

Likelihood ratio test compare the difference in log likelihoods between full and reduced models. A likelihood ratio test can be written as:

For example,

Likelihood ratio test can be used for individual coefficient test and simultaneously several coefficients test.

As a side note, variability in logistic regression is determined by deviance residual not SS and MS residuals. In short, deviance is based on the differences between the response observed and predicted value from the likelihood function.

1.5.2 Wald Test

Test the null hypothesis for individual coefficients. This is comparable to t-tests for individual regression parameter in multiple regression.

There are several ways to compute this statistic. Here, chi-square test statistic is chosen:
Wald chi-squared: =

Wald's test is less reliable than the likelihood ratio test. When it is possible, use the likelihood ratio test.

1.5.3 Assessing model fit

This tells us how good the model predicts the outcomes or what percentage the model predicts correctly (e.g. classification accuracy)
In a case where 50% is used as a cutoff point for positive predictions, then if

Nevertheless, the cutoff point mainly depends on the applications:
To avoid false negatives (e.g. do not report cancer patient as normal), the cutoff should lower than 50%

To avoid false positives or false alarm (e.g. do not convict innocent people), the cutoff should be higher than 50%

1.6 Methods for including variables

- Simultaneous method – all independents are included at the same time
- Hierarchical method – control variables are entered in the analysis before predictors
- Stepwise method – variables are selected in the order to maximize the statistically significant of the model

1.7 Logistic Regression Process

- Identify the questions of interest
- Build models by parameter estimates and hypothesis testing
- Use likelihood ratio test (or Wald's test) to investigate single variable effect
- Use likelihood ratio to compare full and reduced models to assess variables contribution
- Determine the final model

2. Multinomial Logistic Regression

This is the extension of logistic regression, where the outcome has more than two categories. Outcome variable can be ordinal (e.g. excellent, very good, good, fair, poor) or nominal (by car, by bus, by train, by walk).

Similar to logistic regression, no assumptions regarding to normality, linearity, and homogeneity of variance are made in multinomial logistic regression. The overall test relationship between independent and dependent is based on the difference in the likelihood value of the model with independent and the model without independent variables. This difference in likelihood follows chi-square distribution, and known as model chi-square. The significance test of the final model chi-square (after independent variables are added) is an evidence-based for the relationship between dependent and independent variables. Classification accuracy is a useful measure to assess the utility of a multinomial logistic regression model.

To simplify the model, only 3 outcome categories are assumed (Y = 0,1,2 and 0 is arbitrary reference)

- ***Multinomial probability in each category***

- ***Relative Risk Ratio for comparing x=b to x=a (of outcome j to outcome 0)***

3. Polynomial Regression

This model is useful to approximate the complex nonlinear relationships such as when the curvilinear effects appear in the true response function. To deal with nonlinearity, a sequence of power transformation of the same variables is added to the model.

The polynomial regression model is:

The quadratic model (has linear, and squared term for independent variables):

The coefficient is linear effect parameter
 is quadratic effect parameter

The cubic model (has linear, squared, and cubic term):

In general, a number of extreme points equal to the highest order in the polynomial term such as quadratic model has one maximum(or minimum) point, cubic model has both a (relative) maximum and minimum.
Comparable to standard multiple linear regression:

- measures the strength of the relationship between dependent variable and a set of independent variables.
- -test is used to evaluate the relationship between each independent and dependent variable
- -test is used to determine the overall relationship (and examine whether sample can be generalized to the population)

For example, in the Quadratic model, if we want to test whether the quadratic term is necessary, then the hypothesis:

Using -test, if the test is significant then we conclude that there is significant additional explained by the quadratic term, given that the linear term (including intercept) are presented.

3.1 Model building

3.1.1 Forward selection procedure – this approach fits the model in increasing order, as well as tests the significance of the coefficients at each step. Here, the order is increasing until there is no significance on the highest order term.

3.1.2 Backward elimination procedure – this approach fits the model with appropriate highest order, and then gets rid of one term at a time. Begin with significance test for the highest order, and the remaining term until it is significant.
In general, the polynomial order should be as low as possible.

Examples of logistic regression and multinomial using R:

http://www.columbia.edu/~so33/SusDev/Lecture_10.pdf

https://courses.washington.edu/b536/Archive/handouts/Lecture18.pdf
Examples of polynomial regression using R:
http://wiener.math.csi.cuny.edu/Statistics/R/simpleR/stat015.html

References:

Anderson, C.J. (2015). Applied Categorical Data Analysis. EdPsych/Psych/Soc 589.
Retrieved from: http://courses.education.illinois.edu/EdPsy589/lectures/5logreg1_02.pdf

Boslaugh, S. (2013), Statistics in a nutshell. 2nd Edition, O'Reilly Media, Inc., CA.

Eckel, S. (2008). Biostatistic II: Biostatistical Modelling. (slide)
Retrieved from: http://www-hsc.usc.edu/~eckel/biostat2/notes/notes13.pdf
http://www-hsc.usc.edu/~eckel/biostat2/notes/notes14.pdf
http://www-hsc.usc.edu/~eckel/biostat2/notes/notes17.pdf

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) An Introduction to Statistical Learning with Applications in R. Springer New Science+Business Media New York. (4th printing). Retrieved from: http://ckwri.tamuk.edu/fileadmin/user_upload/A-Litt/Regression_Sp10_-_Lecture_17.pdf

O'Halloran, S. Logistic Regression II-Multinomial Data (slide). Sustainable Development U9611 Econometrics II. Retrieved from: http://www.columbia.edu/~so33/SusDev/Lecture_9.pdf http://www.columbia.edu/~so33/SusDev/Lecture_10.pdf

UW course web server. BIOST 536 (slide), University of Washington. Retrieved from: https://courses.washington.edu/b536/Archive/handouts/Lecture18.pdf https://courses.washington.edu/b536/Archive/handouts/Lecture4.pdf

Reflect in ePortfolio Download Print

Activity Details

Task: View this topic