Goodness of fit of a logistic regression model psychology essay

Logistic regression is a linear regression analysis to conduct when the dependent variable is dichotomous (binary). It is predictable and use to describe data and to explain relationship between one dependent binary variable and one or more metric independent variables. It assumes that dependent variable is a stochastic event (Dallag 2007, Field 2009, Gujarati, 2006, Sim 2009).

Logistic regression uses logistics function as illustrated below:

In (Ï€/1-Ï€) lies between - âˆž and âˆž (Dallag, 2007, Gujarati, 2006)

Dummy variables are used for ordinal variables or nominal variables with two levels. When exponentiated it expresses a probability between 0 and 1 it also linearizes the model over this range, so that predictors can be added together in the same way as in linear regression. (Dallag 2007, Field 2009, Gujarati 2006)

Binomial logistic regression is used when the predictor variables are dichotomous and the outcomes are of any type. Multinomial logistic regression is used for predictors with more classes than two and sometimes used for binary predictors. When multiple classes of the predictors' variable can be ranked, then ordinal logistic regression is preferred to multinomial logistic regression. (Field 2009, Gujarati, 2006, Menard 2002)

The predictors for logistic regression can be the same as for linear regression - either continuous or discrete and to determine the amount of variance of the predictor explained by the outcome; to rank the relative importance of outcome; to assess interaction effects; and to understand the impact of

covariate control variable. The impact of predictor variables is explained in terms of odds ratios. (Field 2009, Gujarati 2006)

Logistic regression uses maximum likelihood estimation after transforming of the predictor to a logit variable. By so doing , It estimates the odd of a certain event occurring. It calculates changes in log odds of the predictor variables not changes in the outcome (Field 2009, Garson 2008).

For logistic regression model to be fitted correctly there should be no occurrences of over fitting or under fitting. To achieve this, only the meaningful variable should be included and the entire meaningful variable must be included. The stepwise method is a good approach to estimate logistic regressions (Field 2009, Garson 2008, and Gujarati 2006).

Logistic regression requires error terms and each observation to be independent. The model should also have little or no multi-co linearity which independent variables to be independent from each other. Logistic regression requires that independent variables are linearly related to the log odds. Logistic regression requires large sample sizes because ordinary least square methods are more powerful than maximum likelihood (Dallag 2007, Field 2009, Garson 2008, and Gujarati 2006)

An important consideration for logistic regression analysis is the model fit . Independent variables that is added to a logistic regression model will always increase its statistical validity as it will always explain a bit more variance of the log odds (Field 2009). The model will be inefficient and over fitting will occur when more and more variables is added to the model because the

probability increases that the significance test finds the variables to be significant just by pure chance.

Goodness of fit test is a general test that assesses the fitted model's overall departure from the observed data (Hosmer and Lemeshow 2000). Goodness of fit should be examined by fitting the " design-based model" followed by estimating the corresponding probabilities using independently and " identically distributed- based test" and using any findings to " design- based model" (Hosmer and Lemeshow 2000).

Examining a goodness-of-fit in a logistic regression model can be problematic. The deviance (-2In Likelihood,(Sim 2009) or Pearson chi-square statistics does not have approximate chi-square distributions, if continuous covariates are modelled. Pearson chi-square and deviance test are used to assessing goodness-of-fit in ordinal response model when both categorical and continuous covariates are present (Agresti 2002). The test provides information where a model may not fit well (Garson 2008, Will 2009) The Deviance (-2 In Likelihood) chi-square test and the Hosmer-Lemeshow test shows if or not the model frequencies are better fit to the data than the null model , or worse fit than the " saturated model" (Sim 2009).

A pseudo-R2 measure quantifies the fit of the model to the data while Hosmer-Lemeshow provides a yes/no diagnosis of whether the model fits. If the frequency observed in the data are close to those predicted by the model, the fit is good, and the deviance will be small and statistically non-significant . The deviance of the model indicates how the model fits the data. The difference in deviance between two models shows how much better the

second model fit the data than the first (Sim 2009). For the fit to improve, the difference should be large and statistically significant for the analyst to achieve a significant gain in goodness of fit. (Archer and Lemeshow 2006)

Goodness of fit tests such as the likelihood ratio test is used as indicators of model appropriateness, as is the Wald statistics to test the significant of individual independent variables (Sim, 2009). The Hosmer and Lemeshow test, also called the chi-square test is not available in multinomial logistic regression (Field 2009). It is more robust than the traditional chi-square test, particularly if continuous covariates are in the model or sample is small (Dallag, 2007, Garson 2008, Gujarati, 2006). Hosmer and Lemeshow's goodness of fit test divides subjects into deciles based on predicted probabilities. Then calculate a chi-square from observed and expected frequencies. Then a (p) value is computed from the chi-square distribution with 5 degree of freedom to test the fit of the logistic model. If the H-L goodness-of-fit test is greater than . 05, as for well-fitting models and failure to reject the null hypothesis that there is no difference between observed and model-predicted values, this means that the model's estimates fit the data at an acceptable level. The well-fitting models show non significance on the H-L goodness-of-fit test, showing model prediction is not significantly different from observed values. This does not mean that the model necessarily explains much of the variance in the predictors, however much or little it does explain is significant. (Archer and Lemeshow 2006, Hosmer and Lemeshow 2000)

BIBILOGRAPHY

Archer, K. J and S . Lemeshow (2006) Goodness- of-fit for a logistics regression model fitted using

Survey sample data, The Stata Journal, Volume 6, Number 1, Pp. 97-105

Retrieved on 20-05-2009 Online from:

http://www. stat-journal. com/sjpdf. html? articlenum= st0099

Agresti, A., (2002). Categorical Data Analysis 2nd edition, New York, John Wiley

Clegg, F. (2007) Simple Statistics: A course for the social science, London, Cambridge University

Press

Dallag, G. E., (2007). The Little Handbook of Statistical Practice

Retrieved on 26-02-2010 online from:

http://www. tufts. edu/~gdallal/LHSP. HTM.

Field, A., (2009) Discovering Statistics Using SPSS, 3rd edition, London, Saga Publication

Garson, D. G., (2008) Logistic regression: Statnotes from North Carolina State University

Retrieved on 20-05-2009 Online from:

http://faculty. chass. ncsu. edu/garson/PA765/logistic. htm#concepts

Gujarati, D., (2006) Essential of Econometrics, 3rd edition, New York, McGraw- Hill

Hosmer, D. W., and S. Lemeshow (2000) Applied Logistics Regression, 2nd edition, New York, Wiley

Menard, S., (2002) Applied logistic regression analysis, 2nd edition, Thousand Oaks, Calif: Sage

Publications

Sim, J., (2009) Introduction to logistic regression, Lecture note from

Keele University for Quantitative Data Analysis 2

Will G Hopkins (2009). A new view of Statistics: AUT University Auckland

Retrieved on 26-02-2010 Online from:

http://www. sportsci. org/resource

(3) A statistical test is conducted and the obtained p value is smaller and than the stipulated value of alpha. Explain carefully the statistical conclusion that can be drawn from this.

P-value means probability value is the exact significant level of the test statistics. It is the lowest significant level which a null hypothesis can be rejected (Gujarati 2006). P- value represents the decreasing index of the reliability of the result (Will 2002, Will 2009). P -value shows the probability

of error that is involved in accepting our observed result as valid, that is, as 'representative of the population'. (Dallag 2007)

P-Value assesses how confident result from a sample is true of the population and the probability that result from the sample is not a reflection of the population from which the sample is taken. A way of using the p value is to avoid the arbitrariness in fixing alpha (Î±) at level 1, 5 or 10. Alpha (Î±) is an indicator of how extreme the result must be before we can reject the null hypothesis. The level of significance is same as alpha level (Î±) meaning the least probability value at which the result of a statistical test can be declared statistically significant while P- value illustrates how extreme data are. P value is computed with Î± to determine if observed data are statistically significant different from the null hypothesis (Dallag 2007, Field, 2009, Gujarati, 2006, Will 2002, Will 2009)

If p value is . 10 it means 10% probability that result from the sample is true of the population. A chance of being wrong 10% of time if you reject the true null hypothesis . If repeated samples are taken from the population, the analyst will find 90% of the time a value at least as great as the one found in the sample. " The smaller the p value, the stronger the evidence against the null hypothesis" (Gujarati 2006)

P value identifies if results are representative of the sampling distribution of the test statistic under the condition that the null hypothesis is true, that probability of sampling error alone is responsible for findings, helps make decision at the end of the study about which hypothesis is reasonable to retain, whether the findings obtained are members of sampling distribution for the condition when the null hypothesis is true; and If p value is less than alpha, then the observed results are members of some sampling distribution that represents a case when the alternative hypothesis is true (Field 2009 Gujarati, 2006)

P value demonstrates evidence for truthfulness or otherwise of hypothesis without proving anything because evidence is specific to conditions under which the study is conducted. Theoretical replication, validation, and judgments' of the findings are equally important to the scientific process when p value is small or large. P value does not identify a significant finding. The analysts are to use their own judgments to arrive at the correct interpretation of finding. All theoretical, practical significance of research finding are always present. Also, correlation is not necessarily causality, even when a p value is small. (Dallag 2007, Field, 2009, Gujarati, 2006, Will 2002, Will 2009).

A small p value is not an effect of a particular size. Large sample sizes can produce small P values related with very small effects, while studies with small sample sizes may produce large effects that are not statistically significant value. P value cannot identify that a statistical significance test was necessary. It does not show whether correct statistical procedure was used or assumption met, cost benefit or cost effectiveness issues related to the treatment nor make statement about side effects of the treatment. Also,

it does not make statement about the presence or absence of confounding variables in a study (Field 2009, Gujarati 2006, Hubbard and Armstrong, 2005).

If the critical p value is at level 5, the null hypothesis will be rejected if computed p value is . 00001 less that 5 percent. However, the null hypothesis will not be rejected if the computed p value is more than critical p value (Gujarati 2006).

The power of a test is its ability to reject the null hypothesis when it is false (Sim 2009). Hypothesis testing has gained much ground in modern research. The main idea behind this is to state the Null Hypothesis and the alternative hypothesis. This will enable the analyst to know if the data gathered will allow one to reject the null hypothesis (Hubbard and Armstrong, 2005). When the p-value deduced from the test statistic is less than the significance level, the null hypothesis will be rejected, meaning that the result is statistically significant. On the other hand, when the p-value is greater than the value of Î±, the null hypothesis is accepted and the alternative rejected, meaning the result of the research is non- significant. The analyst can conclude that, the lower the p-value, the less likely the result, assuming the null hypothesis, the more 'significant' the result, in terms of statistical significance. One often rejects the null hypothesis if the p-value is less than or equal to 0. 05 or 0. 01. Results that are significant at P <0. 005 or P <0. 001 are often called 'highly significant' (Hubbard and Armstrong, 2005).

If the probability of deriving no significant differences when null hypotheses are all true is greater than 0. 95, then overall P value is actually smaller than

the nominal 0. 05, by an amount based on dependence between the tests. The power of the test it ability to identify true differences in the population, is respectively diminished or the test is conservative. (Bland 2004)

If P-value is less than alpha level of 0. 05 it means that probability of incorrectly rejecting the null hypothesis is less than 5% statistically significant. If the p-value is small, then observed test statistic falls among a group of potential outcomes that the null hypothesis is not true. The result is statistically significant at the 5% level; because of some odds against the null hypothesis if it were true, then an unlikely event would have had to have occurred. If a fixed cut-off of 5% is used, then the probability of incorrectly rejecting a null hypothesis over the long run is 5%. Evidence or lack of that each particular set of data provides against the null, if P is more than 0. 05, then, it is a good indicator that data is normal. If the hypothesis for the p-value for a test of the null hypothesis of normality is less than 0. 05 then the conclusion in that case should be to reject the null. (Bland 2004, Field 2009, Gujarati, 2009, Hubbard and Armstrong, 2005, Sim 2009, Will 2002, Will 2009)

BIBILOGRAPHY

Bland, J. M (2004) Multiple significant tests and Bonferroni Correction: An Introduction to Medical

Statistics, 3rd edition

Retrieved on 27/05/2009 online from: http://www-users. york. ac. uk/~mb55/intro/bonf. htm

Clegg, F. (2007) Simple Statistics: A course for the social science, London, Cambridge University

Press

Dallag, G. E., (2007) Historical background to the origin of p-values and the choice of 0. 05 as the cut

Off for significance

Retrieved on 26-02-2010 online from: http://en. wikipedia. org/wiki/P-value

Dallag, G. E., (2007). The Little Handbook of Statistical Practice

Retrieved on 26-02-2010 online from: http://www. tufts. edu/~gdallal/LHSP. HTM

Field, A., (2009) Discovering Statistics Using SPSS, 3rd edition, London, Saga

Gujarati, D., (2006) Essential of Econometrics, 3rd edition, New York, McGraw- Hill

Hubbarb, R and Armstrong, J. S., (2005) Historical background on the widespread confusion of the

P-values . Retrieved on 26-02-2010 online from: http://en. wikipedia. org/wiki/P-value

Sim, J., (2009) Statistical Power, Lecture note from Keele University for Quantitative Data Analysis 2

Will, G. H., (2002). A New view of statistics

. Retrieved on 26-02-2010 online from:

http://www. sportsci. org/resource/stats/effectmag. html.

Will, G. H., (2009). A new view of Statistics. Retrieve on 26-02-2010 online from

http://www. sportsci. org/resource/stats/

QUESTION 1

When constructing a multiple linear regression model, on the basis of what considerations can the analyst determine the most appropriate set of predictor variables?

A statistic or econometric, model is an example of a linear regression model where variable appearing on the left-hand side of the equation is called the dependant variable, and the variable on the right-hand side is called the independent, or explanatory variable. In linear regression analysis the main aim is to explain the behaviour of one or more other variable, allowing for relationship between them is inexact. Multiple regressions are ways of predicting an outcome variable from several predictors' variables. In multiple regressions, we include the error term u in the regression model (Gujarati 2006). The error term

A regression analysis is a Outcomei = (model) + errori

A multiple regression is expressed as illustrated in Gujarati (2006).

In non stochastic form as

âˆ‘(Yt) = B1 + B2 X2t + B3 X3t

And in the stochastic form as

Yt= B1 + B2 X2t + B3 X3t + ui

âˆ‘(Y) + ui

Where Y = the dependent variable

X2 and X3 = the explanatory variable

U = the stochastic disturbance term

t = the t observation

The individual demand will differ from the group mean by the factor u, in stochastic error term. In multiple regression we introduce the error term because we cannot take account all the forces that might affect the dependent variables (Bowerman 2003, Field 2009, and Gujarati 2006)

The distinctive features of the multiple regressions enable us not only to include more than one predictor's variable in the model but also to segregate the effect of each predictors variable on outcome from the other predictor variable included in the model (Field 2009, Garson 2008).

Bowerman and O'Connell (2003), Field (2009), Garson, 2008, Gujaraji (2006), Sim (2009), identified some assumptions of multiple regressions which recognise that outcome variables are at least interval, predictor variables are

at least interval or dichotomous, predictive relationship is linear and the model is correctly specified.

The analyst also consider that for any given combination of values of x1, x2, ..., xz, the population of potential error term values has a mean equal to 0. Second, the different populations of potential error term values corresponding to different combinations of values of x1, x2, ..., xz have equal variances. We symbolize the constant variances as Ï2. Third, at any given combination of values of x1, x2... xz, the population of potential error term values has a normal distribution, fourth, any one value of the error term Ïµ is statistically independent of any other value of Ïµ, fifth, residual analysis are uncorrelated with the predictor variables and last, predictor variable are fixed and measured without error. (Bowerman and O'Connell 2003, Garson 2008, Gujarati, 2006, Field, 2009, Sim 2009, Thorne and Giesen 2000)

Multiple regression can established that a set of outcome variables at a significant level and can determined the relative predictive importance of the independent variables. Power terms are added as predictors' variable to explore curvilinear effect. Cross-product term is also added as predictors' variable to investigate interaction effect. One can test the significance of the difference of the two R2's to determine whether adding a predictor variable to the model helps significantly, Using hierarchical regression, the analyst can see how most variance in the predictor variable can be explained by one or a set of new predictor variables (Field, 2009, Sim 2009).

A predictor whose inclusion in the model leads to increase in R2 value would be considered. R-squared is the percentage of the variance in the predictor

variables explained distinctively or jointly by the outcome. Residual are the difference between the observed values and those predicted by the regression equation. Residual is used for three main purposes of detecting heteroscredasticity, outliers and identify pattern of error in a regression. (Bowerman and O'Connell 2003, Garson 2008, Gujarati, 2006, Field, 2009, Sim 2009, Thorne and Giesen 2000)

The method of regression and procedure of selection of the predictors' variable which could be the forward selection method, backward elimination method, Forced entry, hierarchical (block wise entry) and the stepwise regression were the predictors are entered based on purely mathematical results or selection are very important in helping the analyst. Variables that do not contribute to the prediction will be deleted. The entry order will also impact on which variables will be selected . The variable entered in the earlier stages have a better chance of being retained (Field 2009, Garson 2008).

A method is used to minimise the Residual Sum of Square (RSS) which is the sum of the squared difference between the observation and their predictor values by regression. It is equivalent to maximizing the multiple correlation coefficients R. If the RSS were the sole method, then analyst would use all of the variables. An additional criterion must be used if one wishes to reduce the number of variable. The degree to which these criteria are weighted is arbitrary. The use of the Mean Square Error of Prediction (MSEP) as a method of selecting variables, which takes into account the value of the predictor variable associated with the future observation and remove the

arbitrariesness with the RSS. Since the values associated with the future observation cannot generally be given definitely. (Field 2009, Garson 2008)

Examination of the goodness of fit: Goodness of fit test is a general test that assesses the fitted model's overall departure from the observed data. It is suggested that goodness of fit be examined by first fitting the design-based model then estimating the corresponding probabilities and subsequently using independently and identically distributed- based test and applying any findings to design- based model (Bowerman and O'Connell 2003, Garson 2008, Gujarati, 2006, Field, 2009, Hosmer and Lemeshow 2000, Sim 2009, Thorne and Giesen 2000).

The predictor variable found to have been used in previous work can be used. This include literature search among other because the viability of such predictor variable could have been tested. These predictors' variables should be seen to have real life importance. The effect size predictor variables should also be measured . Also, adding and subtracting from the model can cause the b and beta weights to change distinctly, making the analyst to conclude that a predictor variable initially perceived as unimportant is actually an important variable . Also, importance of ratio of beta weights be emphasised by examining the correlation and semi-partial correlation of the given predictor with dependent. Standardisation is needed before comparison because only standardized b-coefficient can be compared to judge relative predictive power of predictive variable (Field 2009, Garson, 2008).

Correlation with outcome variable- A predictor with a high correlation with outcome variable is likely going to be included in the model. The analyst should also consider problem of multicolinearity which is a problem typical to multiple regression when a strong correlation exist between several predictors variables in a multiple regression model. Predictors' variables in the model should be uncorrelated with external variables that are not included in the regression model to influences the outcome variable. If this happens it can lead to unrealistic conclusion from the model because other variable exist that predict the outcome just as well. The analyst will also consider residual terms of the predictors' variables are uncorrelated or independent. This is described as a lack of autocorrelation (Bowerman and O'Connell 2003, Garson 2008, Gujarati, 2006, Field, 2009, Hosmer and Lemeshow 2000, Sim 2009, Thorne and Giesen 2000).