Journal of Modern Mathematics and Statistics 5 (2): 43-46, 2011
ISSN: 1994-5388

Alternative Goodness-of-Fit Test in Logistic Regression Models

[1]M.E. Nja, [2]E.I. Enang, [3]A.U. Chukwu and [3]C.A. Udomboso
[1]Department of Mathematics/Statistics, Cross River University of Technology, Calabar, Nigeria
[2]Department of Mathematics/Statistics/Computer Science, University of Calabar, Nigeria
[3]Department of Statistics, University of Ibadan, Nigeria

Abstract: The Deviance and the Pearson chi-square are two traditional goodness-of-fit tests in generalized linear models for which the logistic model is a special case. The effort involved in the computation of either the Deviance or Pearson chi-square statistic is enormous and this provides a reason for prospecting an alternative goodness-of-fit test in logistic regression models with discrete predictor variables. The Deviance is based on the log likelihood function while the Pearson chi-square derives from the discrepancies between observed and predicted counts. Replacing observed and predicted counts with observed proportions and predicted probabilities, respectively in a cross-classification data arrangement, the standard error of estimate is proposed as an alternative goodness-of-fit test in logistic regression models. The illustrative example returns favourable comparisons with Deviance and the Pearson chi-square statistics.

Key words: Deviance, Pearson chi-square, standard error, observed proportions, predicted probabilities, p value, Nigeria

INTRODUCTION

Goodness-of-fit of a model measures how well the model describes the response variable. Assessing goodness-of-fit involves investigating how close values predicted by the model are to the observed values. Goodness-of-fit test is a test of the explanatory power of a model. In general linear models, this power can be tested using the analysis of variance under the global null hypothesis. The coefficient of determination, R^2 and the standard error of estimate are also available for model evaluation in general linear models. In logistic regression models, a special case of the generalized linear model, the Deviance and the Pearson chi-square statistics are two traditional goodness-of-fit statistics. They are distributed as chi-square with k-m-I degrees of freedom where, k is the number of categories or subpopulations, m is the number of parameters to be estimated (Jennifer *et al*., 1996). The logistic regression tests are based on the assumption that the covariates involved in the model are all discrete.

In the presence of continuous covariates, Bewick conclude that the data is often too sparse to use the Deviance or Pearson chi-square. In that case, they proposed the Hosmer-Lemeshow goodness-of-fit test. Pulkstenis and Robinson (2002) also designed two goodness-of-fit tests for logistic regression models in the presence of continuous explanatory variables, using a methodology similar to that of Hosmer and Lemeshow goodness-of-fit test. Under the violation of the assumptions of independence and identical distribution, Acher *et al*. (2007) developed goodness-of-fit tests for logistic regression models when data are collected using a complex sampling design. Deng *et al*. (2009) designed an improved goodness-of-fit test for logistic regression models based on case-control data by random partition. The statistic has an asymptotic chi-square distribution.

This study proposes the use of the standard error of estimate as a goodness-of-fit test in logistic regression models. The standard error of the estimate Se is a measure of average amount by which the actual observations vary around the regression plane (Webster, 1992). It is a measure of the average amount of the error associated with the model:

$$S_e = \sqrt{\frac{\sum(y_i - \hat{y})^2}{m-k-1}}$$

Where, k is the number of parameter to be estimated and m is sample size. The smaller, the standard error of estimate, the better the fit. This statistic is closely similar to the Pearson chi-square statistic and is distributed as chi- square with m-k-1 degrees of freedom under the null hypothesis that the predicted probabilities closely approximate the observed proportions.

The test is justified against the premise that the values of the observed proportions and the predicted probabilities are continuous and follow a normal

Corresponding Author: M.E. Nja, Department of Mathematics/Statistics, Cross River University of Technology, Calabar, Nigeria

distribution as shown by the normal P-P Plot (plot of expected cumulative probability against observed cumulative probability). The cross-classification data on coronary artery disease (Koch *et al*., 1985) is used to demonstrate that the proposed alternative method returns favourable comparison with the traditional Deviance and Pearson chi-square statistics. In the illustrative example, sex and ECG status are independent variables while the occurrence of coronary artery disease (success) is the response variable The Newton-Raphson iterative scheme is employed in fitting, the logistic model using the SPSS software.

The product binomial distribution: There are two independent and identically distributed explanatory variables X_1 and X_2, representing sex and ECG, respectively. For this reason, the product binomial distribution (Stokes *et al*., 1979) $P_r(n_{kij})$ is given as:

$$P_r(n_{kij}) = \prod_{k-1}^{2}\prod_{1-1}^{2} \frac{n_{ki+}!}{n_{ki1}!n_{ki2}!}\theta_{ki}^{n_{ki1}}(1-\theta_{ki})^{n_{ki2}}$$

Where, n_{ki1} is the number of Persons of the kth sex and ith ECG with coronary artery disease, n_{ki2} is the number of persons of the kth sex and ith ECG without coronary artery disease, k = 1 for females, k = 2 for males; i = 1 for ECG<0.1, i = 2 for ECG ≥0.1, $n_{ki+} = n_{ki1} + n_{ki2}$. θ_{ki} is the probability that a person of the kth sex with an ith ECG status has coronary artery disease:

$$\theta_{ki} = \frac{\exp\left\{\beta_0 + \sum_{j=1}^{n}\beta_j x_{ij}\right\}}{1+\exp\left\{\beta_0 + \sum_{j=1}^{n}\beta_j x_{ij}\right\}} \text{ or } \frac{1}{1+\exp\left\{-\left(\beta_0 + \sum_{j=1}^{n}\beta_j x_{ij}\right)\right\}}$$

$\frac{\theta_{ki}}{1-\theta_{ki}}$ = odds of coronary artery disease for the kth group

$$\frac{\theta_{ki}}{1-\theta_{ki}} = \exp\left\{\beta_0 + \sum_{j=1}^{n}\beta_j x_{ij}\right\}$$

$$\hat{\theta}_{ki}(\text{estimate of } \theta_{ki}) = \frac{\exp\left\{\beta_0 + \sum_{j=1}^{n}\hat{\beta}_j x_{ij}\right\}}{1+\exp\left\{\beta_0 + \sum_{j=1}^{n}\hat{\beta}_j x_{ij}\right\}}$$

Let m_{kij} = Model-predicted counts defined as:

$$m_{kij} = \begin{cases} n_{ki+}\hat{\theta}_{ki} \text{ for } j = 1 \text{ for desease} \\ n_{ki+}(1-\hat{\theta}_{ki}) \text{ for } j = 2, \text{ no disease} \end{cases}$$

$$\text{Deviance} = \sum_{k=1}^{2}\sum_{l=1}^{2}\sum_{j=1}^{2} 2n_{kij}\log\left(\frac{n_{kij}}{m_{kij}}\right)$$

Where, n_{kij} = number of persons of the kth sex and ith ECG with jth disease status.

The Deviance: The Deviance is used to compare two models in generalized linear models and is similar to residual variance in ANOVA (McCullagh and Nelder, 1989). Let:

$$\text{Log}\left[P_r\left(n_{kij}/\hat{\theta}_{ki}\right)\right]$$

be the maximum achievable log likelihood and;

$$\text{Log}\left[P_r\left(n_{kij}/\bar{\theta}_{ki}\right)\right]$$

be the log likelihood under consideration. The deviance is defined as:

$$2\log\left[P_r\left(n_{kij}/\hat{\theta}_{ki}\right)\right] - 2\log\left[P_r\left(n_{kij}/\bar{\theta}_{ki}\right)\right]$$

The Pearson chi-square: The Pearson chi-square (χ_p^2) is the other traditional goodness-of-fit test in logistic regression models. It tests the null hypothesis that the frequency distribution of certain events observed in a sample is consistent with a particular theoretical distribution (Chernoff and Lehmann, 1954). The null distribution of the Pearson statistic with j rows and k columns is approximated by the chi-square distribution with (k-1) (j-1) degrees of freedom (Plackett, 1983):

$$\chi_p^2 = \sum_{k=1}^{2}\sum_{l=1}^{2}\sum_{j=1}^{2}\left(n_{kij} - m_{kij}\right)^2 \Big/ m_{kij}$$

Where, n_{kij} and m_{kij} are as defined earlier.

THE ALTERNATIVE GOODNESS-OF-FIT TEST

Let $\hat{\theta}_{ki}$ be as defined before, P_{ki} be the proportion of success in the kth sex level and ith ECG level:

$$P_{ki} = \frac{n_{kij}}{n_{ki+}}$$

Table 1: Coronary artery disease data

Sex	ECG	Disease	No disease	Total
Female	<0.1 ST Seg. dep.	4	11	15
Female	≥0.1 ST Seg. dep.	8	10	18
Male	<0.1 ST Seg. dep.	9	9	18
Male	≥0.1 ST Seg. dep.	21	6	27

Let $d_i = (\hat{\theta}_{ki} - P_{ki})$ = discrepancy between the ith model-predicted response probability and the ith proportion of success. D = $\{d_1, \ldots, d_n\}$ is a sequence of discrepancies. The proposed alternative goodness-of-fit test is the standard error of estimated, Se. This is given as:

$$S_e = \sqrt{\frac{\sum(\hat{\theta}_{ki} - p_{ki})^2}{m-k-1}}$$

Where:

m = Number of categories or sub-populations

k = Number of parameters to be estimated

ILLUSTRATIVE EXAMPLE

It is required to assess the goodness-of-fit associated with the Newton-Raphson method applied in the logistic regression modeling of the data is shown in Table 1.

PROPOSED METHOD

The following results were obtained:

$\beta_0 = 1.1567765$, $\beta_1 = -1.276955$ and $\beta_2 = -1.0545$

$$\hat{\theta}_{11}(k=1, i=1) = \frac{\exp(\beta_0)}{1+\exp(\beta_0)} = 0.760742$$

$$\hat{\theta}_{12}(k=1, i=1) = \frac{\exp(\beta_0+\beta_2)}{1+\exp(\beta_0+\beta_2)} = 0.525547$$

$$\hat{\theta}_{21}(k=2, i=1) = \frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)} = 0.559802$$

$$\hat{\theta}_{22}(k=2, i=2) = \frac{\exp(\beta_0+\beta_1+\beta_2)}{1+\exp(\beta_0+\beta_1+\beta_2)} = 0.236010$$

$$P_{11} = \frac{4}{15} = 0.266666, \quad P_{12} = \frac{8}{18} = 0.444444$$

$$P_{21} = \frac{9}{18} = 0.500000, \quad P_{22} = \frac{21}{27} = 0.777777$$

D = {0.4940816, 0.0811069, 0.0598016, 0.054176}

With n = 4, k = 2, S_e = 0.2928 having a p value ($P_r > \chi^2$) = 0.6908. Tested against chi-square (χ^2) with 1 degree of freedom at 0.05 level of significance, the null hypothesis of a good fit is accepted. The deviance = 0.2141 with a p value of 0.6436. The Pearson chi-square (χ^2) statistic = 0.2151 with a p value of 0.6425.

DISCUSSION

One of the goodness-of-fit tests in general linear models is the standard error of estimate which assesses the overall goodness-of-fit in a model. It is a measure of the average amount by which the actual observations vary around the regression line. When the assumption of normality apply to a set of observed values, it becomes reliable to apply this measure for model evaluation. Applying the P-P Plot; a plot of the expected cumulative probability against observed cumulative probability, it is shown that the observed proportions are approximately normally distributed.

Several advantages are associated with the use of the standard error of estimate. Like the Pearson chi-square, the standard error of estimate is distributed as chi-square so, the measure can be tested for significance at m-k-1 chi-square degrees of freedom.

The test can be supported by the coefficient of determination (R^2) which measures the explanatory power of the model. From the illustrative example, the standard error of estimate is given as:

S_e = 0.2928 with a p value ($P_r > \chi^2$) of 0.6908

This compares favourably with the Deviance which has a value of 0.2141 and a p value of 0.6436. The Pearson chi-square statistic is 0.2151 with a p value of 0.6425. Tested against chi-square distribution with 1 (m-k-1) degrees of freedom at 0.05 level of significance, the null hypothesis of a good fit is accepted. m-k-1 = 4-2-1 = 1 degree of freedom. The Deviance, Pearson chi-square values and parameter estimates were obtained using the SPSS software.

CONCLUSION

The proposed alternative goodness-of-fit test observes the assumptions of generalized linear models in general. It therefore, extends beyond the scope of the logistic regression models. Its computational ease renders it more user-friendly than the Deviance or the Pearson chi-square. It is an attempt to unify the general linear model and the generalized linear model with respect to goodness-of-fit test.

REFERENCES

Acher, K.J., S. Lemeshow and D.W. Hosmer, 2007. Computational Statistics and Data Analysis. Vol. 51, Elsevier Science Publishers, Amstadam, Netherlands.

Chernoff, H. and E.L. Lehmann, 1954. The use of maximum likelihood estimate in C2 tests for goodness-of-fit. Ann. Math. Stat., 25: 579-586.

Deng, X., S. Wan and B. Zhang, 2009. An improved goodness-of-fit test for logistic regression models based on case-control data by random partition. Commun. Stat. Simul. Comput., 38: 233-343.

Jennifer, L.K., S.W. Alice, T. Douglas and S.E. Alfred, 1996. Methods in Observational Epidemiology. Oxford University Press Inc. New York.

Koch, G.G., P.B. Imrey, J.M. Singer, S.S. Atkinson and M.E. Stokes, 1985. Analysis of Categorical Data. Montreal, Canada.

McCullagh, P. and J.A. Nelder, 1989. Generalized Linear Models. 2nd Edn., Chapmann and Hall, London, ISBN: 0-412-31760-5, pp: 536.

Plackett, R.L., 1983. Karl pearson and the chi-squared test. Int. Stat. Rev., 51: 59-72.

Pulkstenis, E. and T.J. Robinson, 2002. Two goodness-of-fit tests for logistic regression models with continuous covariates. Stat. Med., 21: 79-93.

Stokes, M.E., C.S. Davis and G.G. Koch, 1979. Categorical data analysis using the SAS System. SAS Institute Inc., Carry, NC, USA.

Webster, A., 1992. Applied Statistics for Business and Economics. IRWIN, Boston.