Thailand Statistician
July 2006; 4: 43-61
http://statassoc.or.th
Contributed paper

Goodness-of-fit Tests for Logit Models Based on Probability Levels of Response Categories

Veeranun Pongsapukdee *[a] and Thanittha Kumsri [b]

[a] Department of Statistics, Faculty of Science, Silpakorn University,
Nakhon Pathom 73000, Thailand.

[b] Siam Commercial Bank PCL, Head Office, Rutchadapisek Rd.,
Bangkok 10900, Thailand.

***Author for correspondence, email: veeranun@su.ac.th**

Received: 22 December 2005
Accepted: 30 March 2006

Abstract

For the basic logit models, the response Y takes the value 1 with the success probability P_1, and the value 0 with the failure probability $(1-P_1)$. Problems arise with several proposed statistics for assessing the fit of the models and often be questioned which one of them is more preferable. In this article, 1,000 computer simulation experiments in each condition of the probabilities of Y=1(P_1), the calculated parameters and X's distributions, were generated to evaluate the performance of various statistics, all of which were used for assessing the goodness-of-fit of the logit models. Ten statistics were computed for each combination of base rate levels and model conditions: the likelihood ratio statistics G_M, the indexes of predictive efficiency which consist of λ_P, τ_P and ϕ_P, the coefficients of determination or R^2 analogs which consist of R^2_C (the contingency coefficient R^2), R^2_L (the log likelihood ratio R^2), R^2_M (the geometric mean squared improvement per observation R^2), R^2_N (the adjusted geometric mean squared improvement R^2), and R^2_O (the ordinary least squares R^2). The correlation coefficients for determining their magnitude (absolute values) of the measures of independence from the base rate levels, the percentages of correct classification of the model (%correct) and the type II error rates, corresponding to the percentages of power of the tests (%accept) were also computed.

The research results show that, for hypothesis testing goodness-of-fit of models, both of the %correct and the %accept all are satisfied. The average of %correct, when X is Exponential is around 77% and when X's are Bernoulli and multinomial distributed, they are approximately equal to 99%. Similarly for the average of %accept which are all approximately equal to 95%. For X~ Exponential, the R^2_C, R^2_M, and R^2_O are preferable and for X~ Bernoulli R^2_C, R^2_M, R^2_O are still preferable but R^2_o outperforms. For (X_1, X_2)~ Multinomial, the results are similar but slightly superior to those of X~ Bernoulli. The indexes of predictive efficiency of the multinomial case, when the success probability P_1 is high, suggest that the λ_P , τ_P statistics may be used as the alternatives of the R^2_C, R^2_M and R^2_O. Some recommendations are made for logit models with the exponential explanatory variable, the statistics R^2_C, R^2_M, R^2_O, λ_P and ϕ_P probably be interesting to use. However, when P_1 is closed to 0.5 the %correct is low and the range is high. Therefore, further studies in more details for the exponential explanatory variable together with the increased sample sizes would be recommended. For the logit models with Bernoulli and multinomial explanatory variables are much improved. Then, the statistics R^2_C, R^2_M, R^2_O, λ_P and τ_P are probably appropriate, especially the R^2_O statistic.

Keywords: base rate levels, Bayes' theorem, indexes of predictive efficiency, likelihood ratio statistic, logit models.

1. Introduction

Dichotomous logit models for one or more than one explanatory variables have become the standard method of analysis for explaining the relationship between explanatory variables and a dichotomous response variable [5]. For multinomial or polytomous response, the logit models are also exist to handle the cases of response with more categories and can be called as many other names; such as, the cumulative logit models, the cumulative odds ratios [3]. In the usual case of the dichotomous logit model, it is now commonly used procedure in many disciplines; for example, in health-sciences research, particularly in medical sciences, engineering settings, and is becoming increasingly popular in the behavioral and social sciences. It is also an important endpoint in quality control and quality testing [11]. In this model the basic random variable Y is dichotomous response data taking the value 1 with the success probability P_1, and the value 0 with the failure probability $(1-P_1)$. The relationship between the response probability value P_j and the explanatory variable value x_j of the same individual is from the logit transformed function (1).

$$\boldsymbol{P(x_j)} = \frac{\mathbf{exp}(\beta_o + \sum_{i=1}^{k} \beta_i x_{ij})}{1 + \mathbf{exp}(\beta_o + \sum_{i=1}^{k} \beta_i x_{ij})}, \quad \text{i =1,…, k, j = 1,…, n.} \qquad (1)$$

where $\boldsymbol{P(x_j)}$ = P(Y=1|X=x) = E(Y=1|X=x) denote the expected probability value of Y given x, and x_{ij} = $(x_{0j}, x_{1j},\ldots, x_{kj})$ denote the j^{th} setting of values of k explanatory variables, i = 1,…, k, j = 1,…, n, for which x_{0j} = 1, k is a constant, n is the sample size, β_i, i = 1,…, k are the model parameters, and $y_j = P(x_j) + e_j$, whereas e is an random error which has a distribution with mean zero and variance equals to $\mathrm{P(x)}$ [1-$\mathrm{P(x)}$].

The values $\{y_j$, j =1,…, n} is assumed to follow a Bernoulli (P_1) or a binomial (1, P_1) distribution, so that y_j =1 represents a success and y_j = 0 represents a failure. The model in (1) is typically used with continuous explanatory variables and is often called as the logistic regression model; however, it is also appropriate when X's are categorical variables [13] and is usually called logit models, especially either when it is used with only categorical predictors or multinomial responses [1]. Since statistical methods and techniques for categorical data analyses have undergone development in the past 25 years, and several statistics for assessing and evaluating the goodness-of-fit of logit models have been proposed. It is probably concluded that there are two basic approaches to evaluating the association between the explanatory variables and the response variable in the logit model analysis. One approach discussed by Ryan, 1997 [12], Hosmer and Lemshow, 1989 [5], and Menard, 1995 [8] is to compare predicted and observed discrete values of the response variable, using the prediction table. Such measures are called indexes of predictive efficiency. Another approach is to use coefficients of determination, or R^2 analogs for logit models that compare the discrete observed values of the response with the continuous predicted values of the response (probabilities) [8]. As illustrated by DeMaris,1992 [4], Ryan 1997 [12] and Menard, 2000 [9], R^2 and its analogs are not necessarily consistent with measures of predictive efficiency. Moreover, base rate should also be considered when selecting an index of predictive efficiency, whereas the base rate refers to the relative frequency of occurrence or the ratio of successes to failures of events being studied in the population of interest. However, most measures of predictive accuracy are highly sensitive to changes in base rate [14]. Thus, for evaluating an logit model, no claim is made that any statistic is the best for the use of R^2 analogs and the indices of predictive efficiency. Only that the results illustrated the possible concerns in using some specific measures. Therefore, it probably be necessary and interesting to research the methodological development of assessing the logit model in more details. Both in the different logit models and their goodness-of-fit tests, and also investigating of the performance of R^2 analogs and indexes of predictive efficiency with their uses concerning to both the base rate issue and the inference tests for the overall model fit.

The objective of this article is to asses the adequacy-of-fit of logit models under the dichotomous response classified by its probability levels and the exponential, Bernoulli, and multinomial distributed explanatory variables. The different logit models used in this simulation studies depend on fixed conditions and various explanatory variables through the model parameters calculated using Bayes' theorem.

2. Methodology

Three separated sets of simulation studies were used to generate data in which the dichotomous outcome may depend on each of three different distributed explanatory variables. Three distributions of the explanatory variables consist of Exponential (λ, λ^2), Bernoulli (P), and multinomial (π_1, π_2, π_3, π_4). For the three simulation study sets, each of which four levels of the probability of Y = 1 **(**P_1**)**, 0.05, 0.20, 0.35, and 0.50 were taken to simulate data in each set.

The first set of simulation studies was carried out with a dichotomous response and one exponential distributed covariate. In this set, four simulation studies were performed accordingly to the levels of the probability of Y = 1 (P_1). Each of which was independently repeated 1,000 replicate data sets.

A second set of simulation study was performed with a dichotomous response and each of three different Bernoulli distributed explanatory covariates, namely Bernoulli (0.10), Bernoulli (0.30), and Bernoulli (0.50). Thus, according to each combination of the three distributions and four levels of the probability of Y = 1 (P_1) altogether, twelve simulation studies, were then performed in this set, with independently repeated 1,000 experiments in each simulation study.

The third set of the simulation studies was generated with a dichotomous response and each of multiple explanatory variables which are in the forms of multinomial distributed explanatory variables, namely Multinomial (0.25,0.25,0.25,0.25), Multinomial (0.65,0.08,0.25,0.02), and Multinomial (0.10,0.35,0.45,0.10). Therefore, twelve simulation studies were then performed accordingly to each combination of different distributions and different levels of P_1 . Each of which was independently repeated 1,000 experiments. Detailed descriptions of the simulation and analyses are given for the first set of studies. Other descriptions, applied with the appropriate modification for the second and third sets of simulation studies, all of which are presented in section 2.1-2.2, respectively.

2.1 Simulation

The logit model was used as the form of model (1). The success probability of y = 1, or P(x_j) is a function of explanatory variable values x $_j$, j = 1,…, n. Data were simulated for the first set of simulation studies as an exponential distributed variable. In this first set, the parameters β_0 and β_1 were obtained using the Bayes' theorem approach, with the probability density function given by

$$f_{\lambda}(x) \;=\; \frac{1}{\lambda} e^{\frac{-x}{\lambda}}, \quad \lambda > 0, \quad x > 0$$

from which it readily follows that β_0 = log($P_1\lambda_2 / P_2\lambda_1$), P$_1$ and $\lambda = \lambda_1$, corresponding to the probability of Y = 1, P$_2$ and $\lambda = \lambda_2$, corresponding to the probability of Y = 0, and β_1 = ($\frac{1}{\lambda_2}$ - $\frac{1}{\lambda_1}$) which was set equal to log2. Taking λ_1= 2, we obtained $\frac{\lambda_2}{\lambda_1}$ = 0.419. The explanatory variable values were then generated from the selected distributions, X $_1$ ~Exp(2, 4), corresponding to the probability of Y=1and X $_2$ ~ Exp(0.838,0.702), corresponding to the probability of Y = 0, respectively. Thus, we combined both X$_1$ and X $_2$ based on P$_1$ to obtain a data set of 200 individuals.

To simulate 1,000 sets of the outcome y $_j$, with 200 individuals for each set, a pseudorandom realization, u $_j$, of a uniform (0,1) variable was then regenerated and compared with $P(x_j)$. If $P(x_j)$> u $_j$, then y $_j$ was set = 1, otherwise y $_j$ was set = 0, whereas $P(x_j)$ is the probability of Y=1 for the jth individual, j = 1,…, n which were based on the calculated p(x $_j$) values from the model (1). Then the proportion of cases for which y =1 or the base rate, is computed from each set of data of each logit model. The values of X's and the corresponding parameters which were determined for each probability of P$_1$, 0.05, 0.20, 0.35, and 0.50 and for each of the X's distribution conditions remain the same in 1,000 simulation. The sample size is n= 200 individuals for each of Y and X's. Each condition is performed for 1,000 independent data sets. Therefore, in this first set of study, the exponential distribution, there are all together four computer simulation studies. All details including the parameters and the explanatory variable distributions used in the simulation studies are summarized in Table 1.

Table 1. Parameter (β_0) for Simulation Studies.

Distributions Of X's	Simulation No. (P_1)			
	No.1 (0.05)	No.2 (0.20)	No. 3 (0.35)	No. 4 (0.50)
Exponential *	-1.2337	-0.5808	-0.2593	0.0000
Bernoulli *	-3.2910	-1.7328	-0.9656	-0.3465
Multinomial *	-3.6375	-2.0794	-1.3121	-0.6931

* Exponential distributions: X_1~Exp(2,4) for P_1, X_2 ~Exp(0.838,0.702) for P_2.

* Bernoulli distributions: Bernoulli (0.10), Bernoulli (0.30), and Bernoulli (0.45).

*Multinomial distributions: Multinomial (0.25,0.25,0.25,0.25), Multinomial (0.65,0.08,0.25,0.02), and Multinomial (0.10,0.35,0.45,0.10) .

In the second set of simulation studies, the distributions of the explanatory variables considered consist of Ber(0.10), Ber(0.30), and Ber(0.45). In each distribution, the logit model is of the form log[$\mathrm{P}(\mathrm{x})$/(1-$\mathrm{P}(\mathrm{x})$] = β_0 and log[$\mathrm{P}(\mathrm{x})$/(1-$\mathrm{P}(\mathrm{x})$] = β_1 X, for x = 0 and x = 1, respectively. Thus, from which it follows that β_0 **=** log [$\mathrm{P}(\mathrm{x})$/(1-$\mathrm{P}(\mathrm{x})$]- 0.5log2, corresponding to taking β_1 = log2. Once the data sets were generated from a selected distribution, each of which, according to the probability of Y= 1 levels or the conditions under the given parameters in Table1, the outcomes based on model (1) would then be computed similarly as those performed in the first set of simulation studies. Therefore, in the second set, it consists of twelve computer simulation studies, each with 1,000 data sets.

For the third set of simulation studies, the joint distribution of (X_1, X_2) is assumed to be multinomial with probabilities π_1, π_2, π_3, and π_4, corresponding to the (x_1, x_2) values of (0, 0), (0, 1), (1,0), and (1,1), respectively. Similarly to the second set of simulation studies, using β_1 = β_2 = log2 would then be leading to β_0 = log [$\mathrm{P(x)}$ /(1-$\mathrm{P(x)}$]- log2. Following all the next steps as performed in the previous simulations, the explanatory variables (X_1, X_2) data were then generated, and the corresponding outcomes were also obtained from model (1). In this last set, it consists of twelve computer simulation studies, each with 1,000 data sets.

In conclusion, for each combination of the probability of Y= 1 levels and distribution conditions together with the model parameters shown in Table1, 4,000 replicated data sets were independently generated for the first set, 12,000 replicated data sets for the second set, and also 12,000 data sets for the third set, by the computer simulations. In each replicated data set, each variable in the model (1) was simulated using 200 sample individuals.

2.2 Statistical Analyses

Several statistics were computed for each combination of probability of Y =1(P_1) levels and model conditions, the likelihood ratio statistics G_M, R^2 analogs, indexes of predictive efficiency ,and the type II error rates, all of which were used for assessing goodness-of-fit of the models. All the calculated coefficients of determination (R^2 analogs) are the followings: R^2_C (the contingency coefficient R^2; Aldrich and Nelson, 1984 [2]), R^2_L (the log likelihood ratio R^2; McFadden,1974 [7]; Menard,1995 [8]), R^2_M (the geometric mean squared improvement per observation R^2; Maddala,1983 [6]; Ryan,1997 [12]), R^2_N (the adjusted geometric mean squared improvement R^2; Nagelkerke,1991 [10]; Ryan,1997 [12]), and R^2_O (the ordinary least squares R^2). And also, the computed indexes of predictive efficiency consist of λ_P, τ_P and ϕ_P (Menard, 1995 [8]). Then the correlation coefficients were calculated for determining whether their magnitudes of the measures are independent of the base rate levels. In addition, the %correct of predictive efficiency and the %accept of the power of the tests (shown in Table 2). The average and range were consequently recorded. All the statistics were computed using the following formulae:-

G_M = -2 [ln(L_O)-ln(L_M)] (the model chi-square statistic) ……………(2)

$$R^2_C = \frac{G_M}{(G_M + n)} \qquad \text{……………(3)}$$

$$R^2_L = \frac{[\ln(L_O) - \ln(L_M)]}{\ln(L_O)} = 1 - \left[\frac{\ln(L_M)}{\ln(L_O)}\right] \qquad \text{……………..(4)}$$

$$R^2_M = 1 - \left[\frac{L_O}{L_M}\right]^{\frac{2}{n}} \qquad \text{……………..(5)}$$

$$R^2_N = \frac{\left[1 - (\frac{L_O}{L_M})^{\frac{2}{n}}\right]}{\left[1 - (L_O)^{\frac{2}{n}}\right]} \qquad \text{……………(6)}$$

$$R^2_O = 1 - \frac{\sum\left(y - \hat{P}(x)\right)^2}{\sum\left(y - \bar{Y}\right)^2} \qquad \text{……………..(7)}$$

$$\lambda_P = 1 - \frac{\left(n - \sum f_{ii}\right)}{n - n_{\text{mod}\,e}} \qquad \text{……………..(8)}$$

$$\tau_P = 1 - \frac{\left(n - \sum f_{ii}\right)}{\left[\sum f_i\left(n - f_i\right)/n\right]} \qquad \text{……………..(9)}$$

$$\phi_P = 1 - \frac{\left(n - \sum f_{ii}\right)}{\left[n - \sum E\left(f_{ii}\right)\right]}, \qquad \text{……………(10)}$$

P_1 = A parameter of probability of success in this simulation …………....(11)

Base rate = The proportion of cases for which y =1 from the models …...(12)

%correct = The average percentage correct classified of model (1)

from 1,000 data set ....(13)

%accept = The average percentage of the power of the tests from 1,000 data sets ....(14)

whereas n is the total sample size, L_O is the likelihood function for the model containing only the intercept, L_M is the likelihood function for the model containing all of the predictors, $\hat{P}(x)$ is the predicted value of the dependent variable Y, obtained from the model, y is the observed value of the dependent variable Y, $\overline{Y}$ is the mean of Y, n_{mode} is the observed number of failures in the model category of the dependent variable, f_{ij} is the number of cases observed as having discrete value i and predicted as having discrete value j, f_{ii} is the number of cases for which the predicted value is equal to the observed value, f_i is the number of cases observed as having discrete value i (i.e., the row sum $\sum_j (f_{ij})$ where the rows represent observed values and the columns represent predicted values), $E(f_{ij})$ = $\left[\left(\sum_i f_{ij}\right)\left(\sum_j f_{ij}\right)\right]/n$ is the expected cell frequency for any cell, and $\sum E(f_{ii})$ is the expected number of correctly classified cases), calculated as the product of the row sum $\sum_i$ and the column sum $\sum_j$, divided by the total sample size. All tests of adequacy of logit models were performed under the null hypothesis H_0 : $\beta = 0$ using the model chi-square statistics, G_M and the computer works were programmed using the Minitab macro language and run by macros in MINITAB Release 11 for Windows^{TM} for all simulation studies.

3. Research Results

The results of the dichotomous response depend on each of three different distributed explanatory variables which were classified to three sets of simulations, the Exponential (λ, λ^2), the Bernoulli (P), and the Multinomial (π_1, π_2, π_3, π_4). Each condition 1,000 replicated data sets were carried out under each of four probability of Y =1 levels, 0.05, 0.20, 0.35, and 0.50.

The first set of simulation studies, four simulation studies were performed under a dichotomous response and the Exponential ($\lambda_{,}\lambda^2$) explanatory variable. It is found that the average percentages and the range of the %correct are equal 76.4727 % and 43.8891, respectively and the average percentages and the range of the power of the tests, corresponding to the %accept are equal 94.525% and 0.70, respectively (Table 2 Exponential). For investigating the correlation coefficients to determining whether their magnitude correlation of R^2 analogs and the predictive efficiency of λ_P, τ_P and ϕ_P are independent of the base rate levels. The lower the magnitude (absolute values) of correlation coefficients, the better the adequate of fit of the statistics. It is found in most situation that R^2_C, R^2_M, R^2_O are preferable than the others, especially when P_1 (the probability of Y=1) is low and that the statistics R^2_C, R^2_M , are better than the R^2_O. When P_1 is high all the three statistics R^2_M, R^2_C, R^2_O give approximately the same performance. For the Indexes of Predictive Efficiency : λ_P, τ_P and ϕ_P, it is found that the λ_P and the ϕ_P outperform the τ_P (Table 3).

The second set of simulation study was performed with a dichotomous response and each of three different Bernoulli distributed explanatory variables, namely Bernoulli (0.10), Bernoulli (0.30), and Bernoulli (0.50). The results of the %correct and the power of the test, corresponding to the %accept give the better results than those of the first set. The average percentages and the range of the %correct are approximately equal 99% and 0.4672, respectively and also those of the %accept are approximately equal 94.43% and 1.30, respectively (Table 2 Bernoulli). In assessing the correlation coefficients, it is found that R^2_O, R^2_M, R^2_C are still preferable but when P_1 is low the statistic R^2_O is better than the R^2_M and the R^2_C statistics. When P_1 is high all the three R^2_C, R^2_M, R^2_O give approximately the same performance. However for the Indexes of Predictive Efficiency : the λ_P and the τ_P outperform the ϕ_P (Table 4).

Table 2. The Average of Percentages of Correct Classification (%correct) and The Average of Percentages of the Powers of the tests (%accept) Classified by the probabilities of Y= 1 (P_1) and the X's Distributions.

Distributions Statistics	Probability of Y= 1	X ~ Exponential	X ~ Ber(0.1)	X ~ Ber(0.3)	X ~ Ber(0.45)	(X_1, X_2) ~ Multinomial (0.25, 0.25, 0.25, 0.25)	(X_1, X_2) ~ Multinomial (0.65, 0.08, 0.25, 0.02)	(X_1, X_2) ~ Multinomial (0.10, 0.35, 0.45, 0.10)
% Correct	P_1 =0.05	95.1925	99.2820	99.2005	99.2535	98.4915	98.7050	98.5025
% Correct	P_1 =0.20	87.6315	99.8265	99.4010	99.4570	99.4385	98.7390	99.3560
% Correct	P_1 = 0.35	71.7635	99.8875	99.4870	99.5350	99.4885	99.0010	99.4790
% Correct	P_1 = 0.50	51.3034	99.8845	99.6990	99.5510	99.5150	99.1840	99.4930
Average		76.4727	99.7201	99.4468	99.4491	99.23338	98.9072	99.2076
Range		43.8891	0.6055	0.4985	0.2975	1.0235	0.4790	0.9905
% Accept	P_1 =0.05	94.90	93..70	94.70	94.70	94.40	95.40	95.10
% Accept	P_1 =0.20	94.70	93.20	95.80	94.30	93.20	97.00	95.30
% Accept	P_1 = 0.35	94.20	92.60	94.80	95.00	95.20	94.80	95.00
% Accept	P_1 = 0.50	94.30	93.80	94.70	95.90	93.80	94.60	95.00
Average		94.525	93.325	95	94.975	94.15	95.45	95.1
Range		0.70	1.20	1.10	1.60	2.00	2.40	0.30

Table 3. Coefficients of Correlation between $\boldsymbol{R}^2$ Analogs and Base rates, Indexes of predictive efficiency and Base rates, Classified by the probabilities of Y= 1 (P_1) under X ~Exponential .

X ~Exponential	P_1 =0.05	P_1 =0.20	P_1 = 0.35	P_1 = 0.50
R_0^2	-0.037	0.031	-0.027	0.024
R_L^2	-0.122	-0.192	-0.085	0.024
R_M^2	-0.002	0.048	-0.033	0.023
R_N^2	-0.085	-0.159	-0061	0.024
R_N^2	-0.002	0.048	-0.034	0.024
λ_P	-0.021	0.017	0.054	0.445
τ_P	-0.985	-0.905	-0.992	-0.790
$\phi_{\mathbf{P}}$	-0.025	0.018	0.050	0.221

The last set of the simulation studies was generated with a dichotomous response and each of multiple explanatory variables, which are in the forms of multinomial distributed explanatory variables, namely Multinomial(0.25,0.25,0.25,0.25), Multinomial (0.65,0.08,0.25,0.02), and Multinomial(0.10,0.35,0.45,0.10), it is found that the average percentages and the range of the %correct are approximately equal 99% and 0.8310, respectively. The power of the test, corresponding to the %accept provides the approximately the same results as those of the second set and the average percentages and the range are approximately equal 94.90% and 1.5666, respectively (Table 2 Multinomial). In evaluating the correlation coefficients, it is found that statistics R^2_C, R^2_M, R^2_O are still preferable and all of them R^2_C, R^2_M, R^2_O give approximately close performance, except for the R^2_O that has a little prominent performance than others. However, for all the indexes of predictive efficiency, their results are more better than those of the first set and both λ_P and τ_P outperform the ϕ_P (Table 5).

Table 4. Coefficients of Correlation between $\boldsymbol{R}^2$ Analogs and Base rates, Indexes of predictive efficiency and Base rates, Classified by the probabilities of Y= 1 (P_1) under X ~ Bernoulli (P), P=0.1, 0.3, 0.45.

X ~ Bernoulli (0.1)	**P_1 = 0.05**	**P_1 = 0.20**	**P_1 = 0.35**	**P_1 = 0.50**
R_0^2	-0.009	0.002	-0.033	-0.036
R_L^2	-0.200	-0.190	-0.155	-0.201
R_M^2	-0.032	-0.026	0.005	-0.035
R_N^2	-0.156	-0.148	-0.112	-0.158
R_N^2	-0.033	-0.027	0.004	-0.035
λ_P	-0.310	-0.297	-0.210	-0.161
τ_P	-0.398	-0.219	-0.216	-0.142
$\phi_{\mathbf{P}}$	-0.421	-0.316	-0.199	-0.159
X ~ Bernoulli (0.3)	**P_1 =0.05**	**P_1 =0.20**	**P_1 =0.35**	**P_1 =0.50**
R_0^2	-0.032	-0.009	-0.060	-0.044
R_L^2	-0.103	-0.038	-0.097	-0.031
R_M^2	-0.025	0.014	-0.049	0.017
R_N^2	-0.072	-0.013	-0.074	-0.008
R_N^2	-0.025	0.014	-0.049	0.017
λ_P	0.156	-0.221	-0.309	-0.159
τ_P	-0.112	-0.352	-0.344	-0.143
$\phi_{\mathbf{P}}$	-0.126	-0.617	-0.630	-0.156
X ~ Bernoulli (0.45)	**P_1 =0.05**	**P_1 =0.20**	**P_1 =0.35**	**P_1 =0.50**
R_0^2	-0.059	0.015	0.023	-0.048
R_L^2	-0.083	-0.036	0.048	-0.003
R_M^2	-0.026	-0.026	0.059	0.008
R_N^2	-0.058	-0.031	0.054	0.003
R_N^2	-0.026	-0.026	0.059	0.008
λ_P	0.214	0.110	-0.047	-0.101
τ_P	0.062	-0.088	-0.223	-0.272
$\phi_{\mathbf{P}}$	0.051	-0.094	-0.533	-0.502

Table 5. Coefficients of Correlation between R^2 Analogs and Base rates, Indexes of predictive efficiency and Base rates, Classified by the probabilities of Y=1 (P_1) under (X_1, X_2)~Multinomial (0.25, 0.25, 0.25, 0.25) (X_1, X_2)~Multinomial (0.65, 0.25, 0.08, 0.02) and (X_1, X_2)~Multinomial (0.10, 0.35, 0.45, 0.10).

(X_1, X_2) ~ Multinomial (0.25,0.25,0.25,0.25)	**P_1 = 0.05**	**P_1 = 0.20**	**P_1 = 0.35**	**P_1 = 0.50**
R_0^2	-0.048	-0.052	0.004	-0.015
R_L^2	-0.103	-0.102	-0.091	-0.074
R_M^2	-0.011	-0.018	-0.006	0.013
R_N^2	-0.062	-0.065	-0.054	-0.035
R_N^2	-0.010	-0.018	-0.005	0.013
λ_P	0.213	-0.350	-0.202	-0.198
τ_P	-0.070	-0.226	-0.183	-0.168
ϕ_P	-0.152	-0.428	-0.212	-0.190
(X_1, X_2) ~ Multinomial (0.65,0.25,0.08,0.02)	**P_1 = 0.05**	**P_1 = 0.20**	**P_1 = 0.35**	**P_1 = 0.50**
R_0^2	0.011	-0.012	0.016	-0.003
R_L^2	-0.050	-0.106	-0.040	-0.040
R_M^2	-0.003	-0.055	0.011	0.006
R_N^2	-0.027	-0.081	-0.014	-0.017
R_N^2	-0.003	-0.056	0.011	0.006
λ_P	0.177	0.125	-0.222	-0.287
τ_P	-0.014	-0.203	-0.234	-0.191
ϕ_P	-0.056	-0.233	-0.534	-0.548

(X_1, X_2) ~ Multinomial (0.10,0.35,0.45,0.10)	**P_1 = 0.05**	**P_1 = 0.20**	**P_1 = 0.35**	**P_1 = 0.50**
R_0^2	-0.014	0.014	0.021	0.049
R_L^2	-0.122	-0.130	-0.048	-0.104
R_M^2	-0.045	-0.046	0.043	-0.013
R_N^2	-0.088	-0.093	-0.007	-0.064
R_N^2	-0.044	-0.046	0.044	-0.012
λ_P	0.208	-0.429	-0.176	-0.0169
τ_P	-0.022	-0.283	-0.149	-0.143
ϕ_P	-0.129	-0.563	-0.186	-0.168

4. Conclusion

This research is performed to assessing the adequacy of fit of logit models, using likelihood ratio statistic or GM, R^2 analogs, indexes of predictive efficiency, and to determining the magnitude values of the correlation coefficients between R^2 analogs, indices of predictive efficiency and the base rate levels and also to evaluating the performance of the inferential tests. The logit models depend on the explanatory variables which corresponding to the relationship between the dichotomous response variable and the explanatory variables, namely Exponential, Bernoulli, and Multinomial distributed variables through the model parameters calculated using Bayes' theorem. Each condition for n=200 is repeated for 1,000 simulations, corresponding to the random Y. It is shown that the %correct of model prediction and the %accept corresponding to the percentage of the power of the tests are all probably satisfied. The average %correct, when X's is exponential distributed, is approximately equal 77% (Table 2). The average %correct, for both when X's are Bernoulli and multinomial distributed, are closed together and are approximately equal to 99%. Meanwhile, all of these results are also consistent with the results among the average of %accept when X's are exponential, Bernoulli, and multinomial of which they are approximately equal to 94.52%, 94.43%, and 94.90%, respectively.

The results from the coefficients of correlation between R^2 analogs and base rate levels of exponential distributed X's, show that the R^2_C, R^2_M, and R^2_O statistics are most useful in term of smaller values of their correlation coefficients with the base rate levels. When P_1 is low, R^2_M is preferable. However, when P_1 is moderate to high, these three statistics give similar results (Table 3). For X's are Bernoulli and multinomial, it is found that when P_1 is low, R^2_O is preferable; otherwise, the correlation of coefficients are approximately the same and most values are tend to be more independent than those from the exponential distribution (Tables 4-5). The results of the correlation coefficients between the indexes of predictive efficiency and base rate levels of exponential distribution, show that the statistics λ_P and ϕ_P have better performance than the τ_P statistic does. However, when X's are Bernoulli and multinomial the statistics λ_P and τ_P dominate the statistic ϕ_P. Thus, in general conclusion for the Bernoulli and multinomial distributions, they give lower magnitude of correlation coefficients than those of the exponential distribution. Therefore, we prefer using λ_P and τ_P to ϕ_P.

5. Recommendations

From the results of this research it is found that for the logit model with dichotomous response and exponential explanatory variable, the statistics R^2_C, R^2_M, R^2_O, λ_P and ϕ_P probably be interesting to use; however, when P_1 is closed to 0.5 the %correct is low and the range is high. Thus further studies for more details in the exponential explanatory distribution together with the increased sample sizes would be recommended. The logit models with dichotomous response and Bernoulli and multinomial exponential explanatory variables are much improved, the statistics R^2_C, R^2_M, R^2_O, λ_P and τ_P probably be appropriated. The average of %correct, for both when X's are Bernoulli and multinomial distributed, are approximately equal to 99% with only low range. Similarly for the average %accepted which are also approximately equal to 94%. Therefore, the R^2_o statistics outperforms the others and is also recommended.

Acknowledgements

This research was partially supported by the Institute of Research and Development of Silpakorn University. We would also like to thank the Editor and the referees for their helpful comments.

References

[1] Agresti, A. Categorical Data Analysis. New York: Wiley, 2002.

[2] Aldrich, J. H., and F. D. Nelson. Linear Probability Logit and Probit Models. Beverly Hills: Sage, 1984.

[3] Cole S.R., P.D. Allison, and C.V. Ananth . Estimation of Cumulative Odds Ratios. Copyright Elsevier Inc. 2003. AEP, 14(3), 2004: 172-178.

[4] DeMaris, A. Logit Modelling. Practical Applications, Newbury Park, CA: Sage, 1992.

[5] Hosmer, D.W. and S. Lemeshow. Applied Logistic Regression. New York: Wiley, 1989.

[6] Maddala, G. S. Limited-Dependent and Qualitative Variables in Econometrics. n.p.: Cambridge University Press, 1983.

[7] McFadden, D. The Measurement of Urban Travel Demand. Journal of Public Economics 3, 1974: 303-328.

[8] Menard, S. Applied Logistic Regression Analysis. A Sage University Papers series, CA: Sage, 1995.

[9] Menard, S. Coefficients of Determination for Multiple Logistic Regression Analysis. The American Statistician 54, 2000:17-24.

[10] Nagelkerke, N. J. D. A Note on a General Definition of the Coefficient of Determination. Biometrika 78, 1991: 691-692.

[11] Piegorsch, W. W. Introduction to Binary Response Regression and Associated Trend Analysis. Journal of Quality Technology, 30(3), 1998:269-281.

[12] Ryan, T. P. Modern Regression Methods. New York : Wiley, 1997.

[13] Simonoff, J. S. Logistic Regression, Categorical Predictors, and Goodness-of-Fit: It Depends on Who You Ask. The American Statistician, February, 52(1), 1998: 10-14.

[14] Soderstrom, I., and D., Leitner. The Effects of Base Rate, Selection Ratio, Sample Size, and Reliability of Predictors on Predictive Efficiency Indices Associated with Logistic Regression Models. Paper Presented at the Annual Meeting of the Mid-Western Educational Research Association. Chicago, October 1997: 1-20.