Validation and Performance Analysis of Binary Logistic Regression Model

SOHEL RANA[1], HABSHAH MIDI[2], AND S. K. SARKAR [3]
[1,2,3]Laboratory of Applied and Computational Statistics,
Institute for Mathematical Research, University Putra Malaysia,
43400 Serdang, Selangor, MALAYSIA

E-mail: [1]srana_stat@yahoo.com, [2]habshahmidi@gmail.com, [3] sarojeu@yahoo.com

Abstract: Application of logistic regression modeling techniques without subsequent performance analysis regarding predictive ability of the fitted model can result in poorly fitting results that inaccurately predict outcomes on new subjects. Model validation is possibly the most important step in the model building sequence. Model validity refers to the stability and reasonableness of the logistic regression coefficients, the plausibility and usability of the fitted logistic regression function, and the ability to generalize inferences drawn from the analysis. The aim of this study is to evaluate and measure how effectively the fitted logistic regression model describes the outcome variable both in the sample and in the population. A straightforward and fairly popular split-sample approach has been used here to validate the model. Different summary measures of goodness-of-fit and other supplementary indices of predictive ability of the fitted model indicate that the fitted binary logistic regression model can be used to predict the new subjects.

Keywords: Validation, training sample, deviance, prediction error rate, ROC curve.

1 Introduction

Over the last decade, binary logistic regression model has become, in many fields, the standard method of data analysis. An important problem is whether results of the logistic regression analysis on the sample can be extended to the corresponding population. If this happens, then we say that the model has a good fit and we refer to this question as a model validation analysis [6].

Application of modeling techniques without subsequent performance analysis of the obtained models can result in poorly fitting results that inaccurately predict outcomes on new subjects. Model validation is possibly the most important step in the model building sequence. It is also one of the most overlooked sections. Model validity refers to the stability and reasonableness of the logistic regression coefficients, the plausibility and usability of the fitted logistic regression function, and the ability to generalize inferences drawn from the analysis. Often the validation of a model seems to consist of nothing more than quoting the Cox and Snell [4] R^2 or Nagelkerke [9] adjusted R^2 statistic as well as Correct Classification Rate (CCR) from the fit which measures the fraction of the total variability in the response that is accounted for by the model. Unfortunately, a high R^2 value and high percentage of CCR in logistic regression model do not guarantee that the model fits the data well. Use of a model that does not fit the data well cannot provide good answer to the underlying prediction or scientific questions under investigation. Hence validation is a useful and necessary part of the model-building process [7].

There are many statistical tools for model validation in binary logistic regression, but the primary tool for most process modeling applications is summary measures of goodness-of-fit analysis. Different types of summary measures of goodness-of-fit from a fitted model provide information on the adequacy of different aspects of the model. The logistic regression with binary data is the area in which graphical residual analysis can be difficult to interpret as a model validation [3].

The most accredited methods for obtaining a good internal validation of a model performance are data-splitting, repeated data-splitting, jackknife technique and bootstrapping. In order to validate the fitted model the study used the data-splitting technique. This is a straightforward and fairly popular approach in which the training data is randomly split into two parts; one to develop the model, and another to measure its performance.

The purpose of this study is to present a comprehensive approach to the internal validation of

logistic regression as a predictive model. Our focus is to measure the predictive performance of a model, i.e. its ability to accurately predict the outcome variable on new subjects. Thus the aim of this study is to assess the goodness-of-fit of a given model, and to determine whether the model can be used to predict the outcome of a new subject not included in the original or training sample.

2 Materials and Methods

The Bangladesh Demographic and Health Survey (BDHS-2004) is a part of the worldwide Demographic and Health Surveys program and a source of population and health data for policymakers and the research community. In the survey a total of 11,440 eligible women were furnished their responses. But in this analysis there are only 2,212 eligible women who have two living children and able to bear and desire more children are considered during the period of global two children campaign. The variable age of the respondent, fertility preference, place of residence, highest year of education, working status and expected number of children are considered in the analysis. The variable fertility preference involving responses corresponding to the question, would you like to have (a/another) child? The responses are coded 0 for 'no more' and 1 for 'have another' is considered the binary response variable (Y) in the analysis. The age of the respondent (X_1), place of residence (X_2) is coded 0 for 'urban' and 1 for 'rural', highest year of education (X_3), working status of respondent (X_4) is coded 0 for 'not working' and 1 for 'working' and expected number of children (X_5) is coded 0 for 'two or less' and 1 for 'more than two' are considered as covariates in the binary logistic regression model.

Data splitting approach has been used to validate the fitted model. Since the sample size is large enough, the data are split into two sets. The study selected 1349 (60%) observations randomly as a training sample and the rest 863 (40%) observations as a validation sample [6], because the validation data set will need to be smaller than the model-building or training data set. Firstly, we use the training sample to fit the model. Then we take the fitted model as it is, apply it to the validation sample, and evaluate the model's performance by different summary measures of goodness-of-fit.

3 Fitting of the model for Training Sample

Consider a collection of p explanatory variables be denoted by the vector X'=(X_1, X_2 …Xp) and the conditional probability that the outcome is present be denoted by P(Y=1|X) =π. Then the logit of having Y=1 is modeled as a linear function of the explanatory variables as

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right)=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p\ ;\ 0\le\pi_i\le 1 \tag{1}$$

where the function

$$\pi_i=\frac{\exp\left(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p\right)}{1+\exp\left(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p\right)}$$ is known as logistic function. Suppose (y_1, y_2…y_n) be the n independent random observations corresponding to the random variables (Y_1, Y_2…Y_n). Since the Y_i is a Bernoulli random variable, the probability function of Y_i is $f_i(Y_i)=\pi_i^{Y_i}(1-\pi_i)^{1-Y_i}$; $Y_i=0$ or 1; $i=1,2\cdots n$. As the Y's are assumed to be independent, the likelihood function is given by $g(Y_1,Y_2,\cdots Y_n)=\prod_{i=1}^{n}\pi_i^{Y_i}(1-\pi_i)^{1-Y_i}$ and the log-likelihood function L (β_0, $\beta1$…β_p) =l_i (say)

$$=\sum_{i=1}^{n}Y_i\left(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p\right)$$
$$-\sum_{i=1}^{n}\ln\left\{1+\exp\left(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p\right)\right\} \tag{2}$$

Well known Newton-Raphson iterative method can be used to solve the equation (2) which is known as Iteratively Reweighted Least Square (IRLS) algorithm. Table 1 shows the coefficients β's, their standard errors, the Wald chi-square statistic, associated p-values, and odds ratio exp (β). In order to determine the worth of the individual regressor in logistic regression, the Wald statistic defined as $W=\frac{\hat{\beta}_i^2}{\left[S.E\left(\hat{\beta}_i\right)\right]^2}$ [2]. Under the null hypothesis $H_0:\beta_i=0,(i=1,2,\cdots 5)$, the statistic W is approximately distributed as chi-square with single degree of freedom. The Wald chi square statistics from Table 1 agree reasonably well with the assumption that all the individual predictors have significant contribution to predict the response variable.

Table 1 Analysis of maximum likelihood estimates

Variable	Coefficient β	Standard error	Wald chi-square statistics	df	p-value	Odds Ratio Exp(β)
X_1	-0.053	0.011	21.534	1	0.000	0.949
X_2	0.452	0.146	9.552	1	0.002	1.572
X_3	-0.085	0.018	21.690	1	0.000	0.919
X_4	-0.449	0.167	7.276	1	0.007	0.638
X_5	2.453	0.158	241.058	1	0.000	11.618
Intercept	0.389	0.343	1.290	1	0.256	1.476

The likelihood ratio test is performed to test the overall significance of all coefficients in the model on the basis of test statistic

$$G = [(-2\ln L_0) - (-2\ln L_1)] \quad (3)$$

where L_0 is the likelihood of the null model and $L_1$is the likelihood of the saturated model. Under the null hypothesis, $H_0 : \beta_1 = \beta_2 = \cdots = \beta_5 = 0$ the statistic G follows a chi-square distribution with 5 degrees of freedom and measure how well the independent variables affect the response variable. In the study, G=403.733 with $p < 0.001$, which indicate that as a whole the independent variables have significant contribution to predict the response variable.

In order to find the overall goodness-of-fit, Hosmer and Lemeshow [5] and Lemeshow and Hosmer [10] proposed grouping based on the values of the estimated probabilities. Using this grouping strategy, the Hosmer-Lemeshow goodness-of-fit statistic under usual notations, $\hat{C}$ is as follows

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_k - n_k'\overline{\pi}_k)^2}{n_k'\overline{\pi}_k(1-\overline{\pi}_k)} \quad (4)$$

Hosmer and Lemeshow [5] demonstrated that under the null hypothesis that the fitted logistic regression model is the correct model, the distribution of the statistic $\hat{C}$ is well approximated by the chi-square distribution with g-2 degrees of freedom. This test is more reliable and robust than the traditional chi-square test [1]. The value of the Hosmer-Lemeshow goodness-of-fit statistic computed from the frequencies is $\hat{C}$=5.209 and the corresponding p-value computed from the chi-square distribution with 8 degrees of freedom is 0.74. The large p-value signifies that there is no significant difference between the observed and the predicted values of the outcome. This indicates that the model seems to fit quite reasonable. The other supplementary summary measures of goodness-of-fit like Cox and Snell R^2 is 0.26, Nagelkerke adjusted R^2 is 0.35, predicted correct classification rate is 77.4% indicate that the model fit the data at an acceptable level. Thus the fitted binary logistic response function from the training sample is

$$\hat{\pi} = [1 + \exp(-0.389 + 0.053X_1 - 0.452X_2 + 0.085X_3 + 0.449X_4 - 2.453X_5)]^{-1} \quad (5)$$

Suppose that the validation sample consists of n_v observations $(y_i, \mathbf{x_i})$, i=1, 2…n_v, which may be grouped into J_v covariate patterns. If some subjects have the same value of $\mathbf{x}$, then $J_v < n_v$. We denote the number of subjects with $\mathbf{x}=\mathbf{x_j}$ by m_j, j=1, 2…J_v. It follows that $\sum m_j = n_v$. Let y_j denote the number of positive responses among the m_j subjects with covariate pattern $\mathbf{x}=\mathbf{x_j}$ for j=1, 2…J_v. For the validation sample under study, the number of covariate patterns J_v=626. The logistic probability for the jth covariate pattern is π_j, the value of the previously estimated logistic model obtained in equation (5) using the covariate pattern $\mathbf{x_j}$, from the validation sample. These quantities become the basis for the computation of the summary measures of fit like Hosmer-Lemeshow goodness-of-fit, prediction error rate, area under Receiver Operating Characteristic (ROC) curve. Each of these summary measures of goodness-of-fit is considered in turn in the following.

3.1 Hosmer-Lemeshow Goodness-of-fit Test

Hosmer-Lemeshow goodness-of-fit test may be used to obtain the summary measure of test statistic for the validation sample. Let n_j denote approximately n_v/g or $n_v/10$ subjects in the jth decile. Let $O_j = \sum y_j$ be the number of positive responses among the covariate patterns falling in the jth decile. The estimate of the expected value of O_j under the assumption that the fitted model is correct is $E_j = \sum m_j \pi_j$. Thus the Hosmer-Lemeshow test statistic is obtained as the Pearson chi-

Table 2 Hosmer-Lemeshow goodness-of-fit chi-square statistic

Decile (j)	Mean predicted Prob.	Total observation (n_j)	Observed positive response (O_j)	Expected positive response (E_j)	χ^2	p-value
1	.0777134	63	7	4.89594		
2	.1378498	63	11	8.68454		
3	.2033223	63	16	12.80931		
4	.2317175	62	11	14.36649		
5	.3439117	63	20	21.66644	5.57	0.85
6	.5372998	62	35	33.31259		
7	.8483141	63	49	51.44379		
8	.7502874	63	43	45.26811		
9	.9219760	62	51	52.16251		
10	1.298966	62	76	75.53588		

Table 3 Predicted classification table based on Training sample and Validation sample taking 0.5 as cutoff

Training Sample				Validation Sample			
Observed (Y)	Expected (Y)			Observed (Y)	Expected (Y)		
	0	1	Total		0	1	Total
No more (0)	785	66	851	No more(0)	307	111	418
Have another (1)	239	259	498	Have another (1)	58	150	208

square statistic computed from the observed and expected frequencies as

$$C_v = \sum_{j=1}^{g} \frac{(O_j - E_j)^2}{n_j \bar{\pi}_j (1 - \bar{\pi}_j)} \quad (8)$$

where $\bar{\pi}_j = \sum m_j \hat{\pi}_j / n_j$. The subscript v has been added to C to emphasize that the statistic has been calculated from a validation sample. Under the hypothesis that the model is correct, and the assumption that each E_j is sufficiently large for each term in C_v to be distributed as χ^2 (1), it follows that C_v is distributed as χ^2 (10). Results presented in Table 2 indicate that the model seems to fit quite well.

3.2 Validation of Prediction Error Rate

The classification table may then be used to compute statistic such as prediction error rate, area under the ROC curve, positive and negative predictive power. The reliability of the prediction error rate observed in the training data set is examined by applying the chosen prediction rule to a validation data set. If the new prediction error rate is about the same as that for the training data set, then the latter gives a reliable indication of the predictive ability of the fitted binary logistic regression model and the chosen prediction rule. If the new data lead to a considerably higher prediction error rate, then the fitted binary logistic regression and the chosen prediction rule do not predict new observations as well as originally indicated [8].

In the current study, the fitted logistic response function based on the training sample given in (5) was used to calculate the estimated probabilities for the 626 cases of validation data set. The chosen prediction rule is applied to the estimated probabilities as predict 1 if $\hat{\pi}_j \geq 0.5$ and predict 0 if $\hat{\pi}_j < 0.5$. The percent prediction error rate for the validation sample given in Table 3 is 26.9 while the rate for the training sample was 22.6. Thus the total prediction error rate for the validation sample is not considerably higher than the training sample and we may conclude that it is a reliable indicator of the predictive capability of the fitted logistic regression model.

The area under the ROC curve is another summary measure of the model's predictive power. In the present study the area under the ROC curve for the training sample was 0.80 while the area for the validation sample is 0.72. The area under ROC curve

for the validation sample is smaller than the training sample and it may be considered that the predictive ability of the fitted logistic response function for the new subject is acceptable.

4 Discussion and Conclusion

Model validation is done to ascertain whether predicted values from the model are likely to accurately predict responses on future subjects. Internal validation involves fitting and validating the model by carefully splitting one series of subjects into training set and validating set. The study evaluated the model performance on the validating data set based on the model developed in the training set. Comprehensive approaches to the validation of the predictive logistic regression model have been introduced in the study. Different summary measures of goodness-of-fit and indices have been used to calibrate the model. The summary measures like Hosmer-Lemeshow goodness-of-fit test suggest that the fitted logistic regression model has significant predictive ability for future subjects. Prediction error rate for validation of the model is not so high. The area under the ROC curve for the training sample was 0.80 and it was decreased by 0.08 to 0.72 for the validation sample which indicates that the predictive ability of the fitted model is good. Thus different summary measures of goodness-of-fit and others supplementary indices of predictive ability of the fitted model indicate that the fitted binary logistic regression model can be used to predict the future subjects.

References

[1] A. Agrest, *Categorical data analysis*, Wiley InterScience, New York, 2002.

[2] A. Wald, Test of statistical hypotheses concerning several parameters when the number of observations is large, *Transactions of the American Mathematical Society*, Vol.54, 1943, pp. 426-482.

[3] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, Chapman and Hall/CRC, 1983.

[4] D. R. Cox and E. J. Snell, *The Analysis of Binary Data*, 2nd edition, Chapman and Hall, London, 1989.

[5] D. W. Hosmer and S. Lemeshow, A goodness-of-fit test for the multiple logistic regression models, *Communications in Statistics*, Vol.A10, 1980, pp. 1043-1069.

[6] F. E. Harrell, K. L. Lee and D. B. Mark, Tutorial in Biostatistics: Multivariable prognostic models: Issues in developing models, evaluating assumptions and measuring and reducing errors, *Statistics in Medicine*, Vol.15, 1996, pp. 361-387.

[7] J. Shao, Linear Model Selection by Cross-Validation, *Journal of the American Statistical Association*, Vol.80, No.422, 1993, pp. 486-494.

[8] M. H. Kutner C. J. Nachtsheim, J. Neter and W. Li, *Applied Linear Statistical Models*, Fifth Edition, McGraw-Hill, Irwin, 2005.

[9] N. J. D. Nagelkerke, A note on the general definition of the coefficient of determination. *Biometrika*, Vol.78, 1991, pp. 691-692.

[10] S. Lemeshow and D. W. Hosmer, The use of goodness-of-fit statistics in the development of logistic regression models, *American Journal of Epidemiology*, Vol.115, 1982, pp. 92-106.