Binary Response Modeling and Validation of its Predictive Ability

HABSHAH MIDI[1], SOHEL RANA[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]habshahmidi@gmail.com, [2]srana_stat@yahoo.com , [3] sarojeu@yahoo.com

Abstract: – Assessment of the quality of the logistic regression model is central to the conclusion. 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. It is not unusual for statisticians to check fitted model with validation. Validation of predictions from logistic regression models is of paramount importance. 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. The present study have dealt with how to measure the quality of the fit of a given model and how to evaluate its performance regarding the predictive ability in order to avoid poorly fitted 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.

Key-Words: – Validation, training sample, deviance, prediction error rate, ROC curve.

1. Introduction

Logistic regression techniques have become an integral component of any data analysis. It describes the relationship between a response variable and one or more explanatory variables in which the outcome variable is often discrete and takes on two possible values. When the response variable is binary, the shape of the response function is usually sigmoidal. Over the last decade, binary logistic regression model has become, in many fields, the standard method of data analysis. Thus they are widely used in a number of different contexts. 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 goodness-of-fit analysis, performance analysis or model validation analysis for the model [12], [9], [11]. 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. In view of the fact that the principal aim of predictive modeling is generalization which implies the ability to predict the outcome on novel cases. In the prediction problem the statistician has available a set of cases, collectively called the training set and each case consists of two parts namely vector of predictors and a response vector. On the basis of training set, a prediction rule or model is constructed and use it to predict a future unobserved response on the basis of its predicted response vector [4].

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 [7] R^2 or Nagelkerke [17] 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 [15,19,21]. Hence validation is a useful and necessary part of the model-building process.

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. Graphical methods for model validation, such as residual analysis are also useful, but usually to a lesser degree than numerical methods due to binary outcomes. Numerical methods have an advantage over graphical methods for model validation because they readily illustrate a broad range of complex aspects of the relationship between the model and the data. Graphical methods for model validation tend to be narrowly focused on particular aspects of the relationship between the model and the data and often try to compress that information into two band of the graph in binary logistic regression. Thus the logistic regression with binary data is the area in which graphical residual analysis can be difficult to interpret as a model validation [3]. In addition, the binary regression residuals are not homosedastic and follow the property of heteroscedasticity, which is the another cause of interpret as a model validation [13, 14]

In some situations, it may be possible to exclude a sub sample of our observations, develop a model based on the remaining subjects, and then test the model in the originally excluded subjects. In other situations it may be possible to obtain a new sample of data to assess the goodness-of-fit of a previously developed model. This type of assessment is often called model validation, and may be especially important when the fitted model is used to predict outcome for future subjects. In some situations it may be possible to obtain a new sample of data from the same population or from a similar population. This new sample can then be used to assess the goodness-of-fit of a previously developed model by applying the model as it is to the new sample. This type of assessment is called external validation [11]. External validation is the most stringent and unbiased test for the model and for the entire data collection process. Nonetheless, most of the time it is not possible to obtain a new independent external sample from the same population or a similar one. It may then be possible to internally validate the model. The most accredited methods for obtaining a good internal validation of a model performance are data-splitting, repeated data-splitting, jackknife technique and bootstrapping. The core concept of these methods is similar in order to exclude a sub sample of observations, develop a model based on the remaining subjects, and then test the model in the originally excluded subjects. 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. With the data-splitting approach, model performance is determined on similar, but independent data. Common split is 50:50, 60:40 or 2/3:1/3. In order to check the internal validity of logistic regression model the study select 60% observations randomly as a training sample and the rest 40% of the observations as a validation sample [18], because the validation data set will need to be smaller than the model-building or training data set.

The reason for considering this type of assessment of model performance is that the fitted model always performs in an optimistic manner on the training data set. 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 part of the worldwide Demographic and Health Surveys program, which is designed to collect data on fertility, family planning, maternal and child health. The BDHS is a source of population and health data for policymakers and the research community. In this study women's data file is used. A total of 11,440 eligible women were furnished their responses. But in this analysis there are only 2,212 eligible women those are able to bear and desire more children are considered. The women under sterilization, declared in fecund, divorced, widowed, having more than and less than two living children are not involved in the analysis. Those women who have two living children and able to bear and desire more children are only considered here during the period of global two children campaign. In BDHS- 2004, there are three types of questionnaires, namely the household's, women's and men's. The information's obtained from the field are recorded in their respective data files. In this study, the information's corresponding to the women's data file is used. 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 as desire for children which is 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. In accordance with the principles of data-splitting we distinguish between training and validation samples. Due to random sampling, both samples from the same population, but are distinct and independent from one another. The size of the two sub-samples must be chosen in such a way as to have enough data in the training sample to fit the model and enough data in the validation sample. Since the sample size is large enough, the data are split into two sets. The first set or training sample consists of 1349 (approximately 60 percent of total sample) observations which were selected randomly from 2212 observations. The validate sample consists of the rest 863 (Approximately 40 percent) observations. 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 on it.

3. Fitting of the model for Training Sample

In order to fit the binary logistic regression model for the training sample, consider a collection of p explanatory variables be denoted by the vector X'=(X_1, X_2 …Xp). Let 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. Under usual assumptions, least square estimations have some desirable properties. However, the OLS method no longer has these properties applied to estimate a model with dichotomous outcome. In such a situation, the most commonly used method of estimating the parameters of a logistic regression model is the method of Maximum Likelihood (ML). In logistic regression, the likelihood equations are non-linear explicit function of unknown parameters. Therefore, we use a very effective and well known Newton-Raphson iterative method to solve the equations which is known as Iteratively Reweighted Least Square (IRLS) algorithm.

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 $f_i\left(Y_i\right)=\pi_i^{Y_i}\left(1-\pi_i\right)^{1-Y_i}$; $Y_i=0$ or 1 ; $i=1,2\cdots n$. As the Y's are assumed to be independent, the joint probability function or likelihood function is given by $g\left(Y_1,Y_2,\cdots Y_n\right)=\prod_{i=1}^{n}\pi_i^{Y_i}\left(1-\pi_i\right)^{1-Y_i}$, the log-likelihood function L (β_0, β1…β_p) =l_i (say),

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

The most effective and well known Newton-Raphson iterative method can be used to solve the equations. 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 $W=\frac{\hat{\beta}_i^2}{\left[S.E\left(\hat{\beta}_i\right)\right]^2}$ [5], [2], [6]. Under the null hypothesis $H_0:\beta_{\rm i}=0$, $\left(i=1,2,\cdots 5\right)$, 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.

Once, the particular multiple logistic regression model has been fitted, we begin the process of model assessment. The likelihood ratio test is performed to test the overall significance of all coefficients in the model on the basis of test statistic

$$G=\left[\left(-2\ln L_0\right)-\left(-2\ln L_1\right)\right] \tag{3}$$

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

where L_0 is the likelihood of the null model and $L_1$is the likelihood of the saturated model. Under the global 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, summary measure provides 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 [8] and Lemeshow and Hosmer [20] proposed grouping based on the values of the estimated probabilities. Hosmer-Lemeshow goodness-of-fit test divides subjects into deciles based on predicted probabilities and computes a chi-square from observed and expected frequencies (Table is not shown here). Using this grouping strategy, the Hosmer-Lemeshow goodness-of-fit statistic, $\hat{C}$ is obtained by calculating the Pearson chi-square statistic from the g×2 table of observed and estimated expected frequencies. A formula defining the calculation of $\hat{C}$ is as follows

$$\hat{C}=\sum_{k=1}^{g}\frac{\left(o_k-n_k'\bar{\pi}_k\right)^2}{n_k'\bar{\pi}_k\left(1-\bar{\pi}_k\right)} \qquad (4)$$

where g denotes the number of groups, n_k' is the number of observations in the kth group, o_k is the sum of the Y values for the kth group and $\bar{\pi}_k$ is the average of the ordered $\hat{\pi}$ for the kth group. Hosmer and Lemeshow [8] 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. A comparison of the observed and expected frequencies in each of the 20 cells indicates close agreement within each decile. Hosmer *et al.*, [10] examined the distributional properties of their test via simulations. 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

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

The use of validation data amounts to an assessment of goodness-of-fit where the fitted model is considered to theoretically known and no estimation is performed. The methods for assessment of fit in the validation sample parallel to the training sample can be done via summary measures of fit as well as logistic regression diagnostics. The major difference is that the values of the coefficients in the model are regarded as fixed constants rather than estimated values.

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. That is our fitted model contains p independent variables, $\mathbf{x}'=(x_1, x_2\ldots x_p)$, and let J_v denote the number of distinct values of **x** observed. If some subjects have the same value of **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 Pearson's χ^2 goodness-of-fit, deviance goodness-of-fit D, and Hosmer-Lemeshow goodness-of-fit C. Each of these summary measures of goodness-of-fit is considered in turn in the following.

3.1 Pearson Chi-Square Goodness-of-fit Test

The Pearson chi-square goodness-of-fit test assumes only that the y_{ij} observations are independent and that the replicated data of reasonable sample size are available. The test can detect major departure from a logistic response function, but is not sensitive to small departures from a logistic response function. Here our objective is to test the hypothesis $H_0 : E\{y\} = [1+\exp(-X'\beta)]^{-1}$ against $H_1 : E\{y\} \neq [1+\exp(-X'\beta)]^{-1}$. Here the number of distinct combinations of the predictor variable be denoted by c, the ith binary response at predictor combination x_j by y_{ij}, and the number of cases in the jth class will be denoted by n_j. The number of cases in the jth class with outcome 1 will be denoted O_{j1} and the number of cases in the jth class with outcome 0 will be denoted by O_{j0}. Because the response variable y_{ij} is a Bernoulli variable whose outcomes are 1 and 0, the number of cases O_{j1} andO_{j0} can be easily obtained. Also suppose $\overline{\pi}_j$ is the average of the predicted probability for the jth class. If the logistic response function is appropriate, the expected value of y_{ij} is given by $E\{y_{ij}\} = \pi_j = [1+\exp(-X_j'\beta)]^{-1}$ and is estimated by the fitted value $\hat{\pi}_j = \left[1+\exp\left(-X_j'\hat{\beta}\right)\right]^{-1}$. Consequently, if the logistic response function is appropriate, the expected number of cases with y_{ij}=1 and y_{ij}=0 for the jth class are estimated as $E_{j1} = n_j\overline{\pi}_j$ and $E_{j0} = n_j(1-\overline{\pi}_j)$ respectively, where E_{j1} denotes the estimated expected number of 1s in the jth class, and E_{j0} denotes the estimated expected number of 0s in the jth class. The following Pearson chi-square test statistic is used to test the null hypothesis H_0.

$$\chi_v^{\ 2} = \sum_{j=1}^{c}\sum_{k=0}^{1}\frac{(O_{jk}-E_{jk})^2}{E_{jk}} \tag{6}$$

If the logistic response function is appropriate and n_j is large enough with p<c, χ^2_v follows approximately a chi-square distribution with c-p degrees of freedom. As with other chi-square goodness-of-fit tests, it is advisable that most expected frequencies E_{jk} be moderately large, that is 5 or more. The value of the Pearson chi-square statistic from Table 2 is 7.18 with p-value 0.21. Smaller value of the test statistic χ^2_v equivalently the large p-value indicates that the fitted logistic response function is appropriate for prediction.

Table 2: Pearson chi-square and Deviance chi-square statistic as summary measures of goodness-of-fit tests

Number of observations (n_j)	$\overline{\pi}_j$	O_{j0}	E_{j0}	O_{j1}	E_{j1}	χ^2	df	p-value	p_j	D	df	p-value
63	.090	56	57.29	7	5.70				.111			
63	.150	52	53.52	11	9.47				.174			
63	.179	47	51.66	16	11.33				.253			
62	.190	51	50.17	11	11.82				.177			
63	.251	45	47.16	18	15.83	7.18	5	0.21	.285	6.73	5	0.24
62	.405	35	36.83	27	25.16				.435			
63	.453	31	34.40	32	28.59				.507			
63	.726	21	17.23	42	45.76				.666			
62	.764	16	14.57	46	47.42				.741			
62	.873	11	7.826	51	54.17				.822			

3.2 Deviance Goodness-of-fit Test

The deviance goodness-of-fit test for logistic regression models is completely analogous to the F test for lack of fit for multiple linear regression models. Suppose there are c unique combinations of predictors and the number of observations in the jth class is n_j, and the ith binary response at jth class be denoted as y_{ij}. The lack of fit test for standard regression was based on the general linear test of the reduced model E $\{y_{ij}\}$ =$X_j'\beta$ against the full model E $\{y_{ij}\}$ =μ_i. In similar fashion, the deviance goodness-of-fit test is based on a likelihood ratio test of the reduced model E $\{y_{ij}\}$ = $[1+\exp(-X_j'\beta)]^{-1}$ against the full model E $\{y_{ij}\}$ =π_j. This full model in the logistic regression case is usually referred to as the saturated model. To carry out the likelihood ratio test, we must obtain the values of the maximized likelihoods for the full and reduced models, namely L (F) and L (R). L(R) is obtained by fitting the reduced model, and the maximum likelihood estimates of the parameters extracted from the training sample and hence the sample proportion $p_j=y_j/n_j$. Let $\bar{\pi}_j$ is the average predicted probability for the jth class and to test the same hypothesis as in the case of Pearson chi-square goodness-of-fit the test statistic

$$D_v = -2\left[Log_e L(R) - Log_e L(F)\right]$$

$$= -2\sum_{j=1}^{c} y_j \left[Log_e\left(\frac{\bar{\pi}_j}{p_j}\right) + \left(n_j - y_j\right) Log_e\left(\frac{1-\bar{\pi}_j}{1-p_j}\right)\right] \quad (7)$$

The likelihood ratio test given in equation (7) is called the deviance. If the fitted logistic response function is correct one and the class sizes n_j are large enough, then the deviance will follow a chi-square distribution with c-p degrees of freedom, where p is the number of predictors. Small value of the deviance given in Table 2 as D=6.73 with 5 degrees of freedom and p-value 0.24 indicates that the fitted logistic response function is correct one.

3.3 Hosmer-Lemeshow Goodness-of-fit Test

Hosmer-Lemeshow goodness-of-fit test may be used to obtain an equivalent summary measure of test statistic for the validation sample. Assume that we wish to use g or10 groups composed of deciles based on ordered predicted probabilities. Any other grouping strategy could be used with obvious modifications in the calculations. 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 $E_j=\sum m_j\pi_j$, where sum is over the covariate patterns in the jth decile. Thus the Hosmer-Lemeshow test statistic is obtained as the Pearson chi-square statistic computed from the observed and expected frequencies as

Table 3: 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		

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

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). In general, if we use g groups then the distribution is χ^2 (g). In addition to calculating a p-value to assess overall fit, it is recommended that each term in C_v be examined to assess the fit within each decile. The value of the Hosmer-Lemeshow goodness-of-fit statistic computed from the frequencies in Table 3 for validation sample is C_v=5.57 and the corresponding p-value computed from the chi-square distribution with 10 degrees of freedom is 0.85. This indicates that the model seems to fit quite well. A comparison of the observed and expected frequencies in Table 3 shows close agreement within each decile. The appropriateness of the p-value depends on the validity of the assumption that the estimated expected frequencies must be greater than 5.

It can be observed from table 3 that only one of the estimated expected frequencies is less 5. However, its value is fairly closed to 5. In the present case, there is reason to believe that the calculation of the p-value is accurate enough to support the hypothesis that the model fits well. If one is concerned about the magnitude of the expected frequencies, selected adjacent rows of the table may be combined to increase the size of the expected frequencies while, at the same time, reducing the number of degrees-of-freedom. The advantage of the Hosmer-Lemeshow chi-square test is that it provides analysis with a single, easily interpretable value that can be used to measure the calibration of a model.

3.4 Validation of Prediction Error Rate

The classification table is the remaining summary statistic that we are likely to use with the validation sample and then only in instances where classification is an important use of the model. The classification table is constructed for validation sample with the modification that probabilities are obtained from the fitted response function in (5). The resulting table may then be used to compute statistic such as prediction error rate, area under the Receiver Operating Characteristic 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 [16].

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 4 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 and the chosen prediction rule.

Table 4: Predicted classification table based on Training sample and Validation sample taking 0.5 as cutoff.

Training Sample				Validation Sample			
	Expected (Y)				Expected (Y)		
Observed (Y)	0	1	Total	Observed (Y)	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

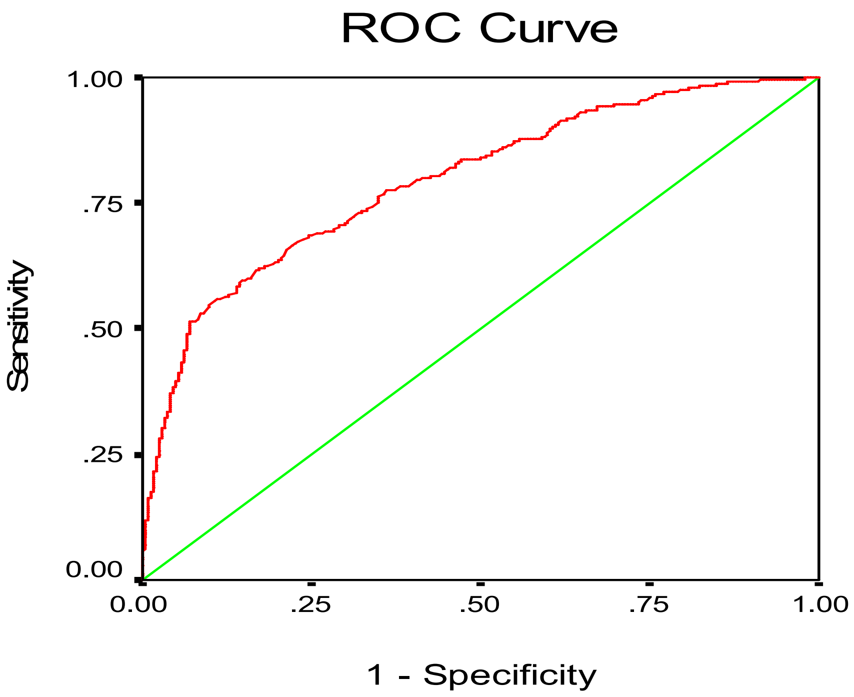

Figure 1.1: Area under the ROC curve for the training sample.

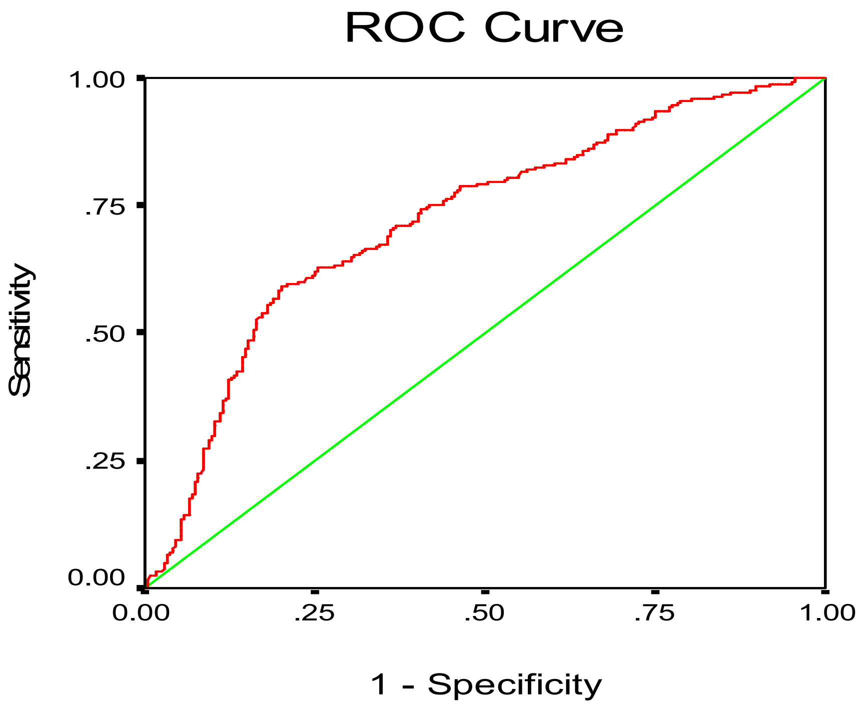

Figure 1.2: Area under the ROC curve for the validation sample.

A graphical display through the Receiver Operating Characteristic (ROC) curve is an effective way to exhibit the classification information. For possible points $\hat{\pi}$, the ROC plots $P(\hat{Y}=1|\text{Y=1})$ against $\{1-P(\hat{Y}=0|\text{Y=0})\}$ which are called sensitivity and specificity, respectively. The area under the ROC curve is another summary measure of the model's predictive power and is identical to the concordance index. Suppose any pair of observations (i, j) such that Y_i=1 and Y_j=0. Since Y_i>Y_j, this pair is said to be concordant if $\hat{\pi}_i > \hat{\pi}_j$. The concordant index estimates the probability that the predictions and the outcomes are concordant. The area under ROC curve having the value 0.5 means that the predictions were no better than random guessing. In the present study the area under the ROC curve for the training sample is 0.80 (Figure 1.1) while the area for the validation sample is 0.72 (Figure 1.2) for all possible cut points between 0 and 1. 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

Logistic regression is a technique for fitting a regression curve to the data in which the dependent variable is dichotomous. An interesting and useful property of the logistic response function is that it can be linearized easily. The principal aim of predictive modeling is generalization and determination of its ability to predict the outcome on new subjects. In contrast, the principal aim of traditional statistical analysis is inference. Confidence intervals, hypothesis test, and p-values are the common inferential tools. Similar methods used by predictive modelers may be used to infer how input variables affect the response variable. The validity of the inference relies on understanding the statistical properties of methods and applying them correctly. Understanding the relationships between random variables can be important in predictive modeling as well. However, many of the methods, used are adhoc with poorly understood statistical properties. Consequently, the discovery of structure in predictive modeling is informal and exploratory. Some predictive modeling methods are inscrutable yet successful because they generalize well. The validity of predictive modeling methods is assessed empirically. If a model generalizes well, then the method is useful, regardless of its statistical properties.

Model validation is done to ascertain whether predicted values from the model are likely to accurately predict responses on future subjects or subjects not used to develop the model. There are two major modes of model validation, that is the external and the internal validation. Even though the external validation is frequently favored by non-statisticians, it is often problematic and more stringent than internal validation. 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 Pearson's chi-square, Deviance and 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] B. Efron, Estimating the Error Rate of a Prediction Rule Improvement on Cross-Validation, *Journal of the American Statistical Association*, Vol.78, No.382, 1983, pp. 316-331.

[5] C. R. Rao, *Linear Statistical Inference and Its Applications*, 2nd edition, John Wiley & Sons, Inc. 1973.

[6] D. E. Jennings, Judging inference adequacy in logistic regression. *Journal of the American Statistical Association*, Vol.81, 1986a, pp. 471-476.

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

[8] 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.

[9] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, 2nd edition, Wiley InterScience, New York, 2000.

[10] D. W. Hosmer, T. Le Cessie and S. Lemeshow, A comparison of goodness-of-fit tests for the logistic regression model, *Statistics in Medicine*, Vol.16, 1997, pp. 965-980.

[11] 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.

[12] F. E. Harrell, *Regression Modeling Strategies with Applications to Linear Models, Logistic Regression, and Survival Analysis*, Springer Series in Statistics, Springer-Verlag, New York, Inc, 2001.

[13] H. Midi, S. Rana and A. H. M. R. Imon, The performance of robust weighted least squares in the presence of outliers and heteroscedastic errors, *WSEAS TRANSACTIONS on MATHEMATICS,*Vol. 8, 2009, pp. 351-361.

[14] I. Juutilainen and J. Roning, Heteroscedastic Linear Models for Analysing Process Data, *WSEAS TRANSACTIONS on MATHEMATICS,*Vol. 2, 2003, pp. 179-187.

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

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

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

[18] R. A. Giancristofaro and L. Salmaso, Model Performance Analysis and Model Validation in Logistic Regression, *Statistica*, Vol.LXIII, No.2, 2003, pp. 375-396.

[19] S. Al-Hajjar, Improving Probability Education Through Statistical Experiments, *WSEAS TRANSACTIONS on MATHEMATICS,*Vol. 7, 2008, pp. 382- 390.

[20] 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.

[21] T. C. Yang, C. M. Kao, T. Y. Yeh, C. E. Lin, Y. C. Lai , Evaluation of NPS Pollution in Drinking Water Protection Area of Kaoping River Watershed, *WSEAS TRANSACTIONS on MATHEMATICS,*Vol. 5, 2006, pp. 1131-1137.