Research Article

iMedPub Journals
www.imedpub.com

International Journal of Applied Science - Research and Review
ISSN 2394-9988

2021
Vol.8 No.3:10

Assessing Discrimination Power for Binary Logistic Regression Model Based on Parametric and Non-Parametric Methods

Md. Asadullah*

Department of Statistics, Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Gopalganj-8100, Bangladesh

***Corresponding author:** Md. Asadullah

asadullahstat@gmail.com

Department of Statistics, Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Gopalganj-8100, Bangladesh.

Citation: Md. Asadullah (2021) Assessing Discrimination Power for Binary Logistic Regression Model Based on Parametric and Non-Parametric Methods. Appl Sci Res Rev Vol.8 No.3:10

Abstract

The main focus of this paper is to measure the discrimination ability of the fitted binary logistic regression model after admission the patients in ICU (intensive care unit). In this paper we use parametric and non-parametric methods for measuring discrimination ability of the logistic regression classifier. The most important analysis in which the outcome variable is binary or dichotomous. It can be used to predict a binary dependent variable from a set of independent variables. Since our outcome variables have binary categories, so binary logistic regression prefers to estimate model parameters. This technique is preferred by many researchers in the analytical fields. It is also widely used in various clinical researches to predict the risk of a patient's future health status. Predictions based on these models have an important role in predicting the survival of patients in ICU. Concordance statistic (C-statistic), which is equivalent to the area under a receiver operating characteristic curve (AUC), is frequently used to quantify the discriminatory power of the logistic model because of its straightforward clinical interpretation. In this paper we assess the discrimination power in simulation and real data for binary logistic regression.

Keywords: Logistic regression; ICU; C-statistic; Simulation; Discriminatory power.

Received: March 13, 2021; **Accepted:** April 05, 2021; **Published:** April 23, 2021

Introduction

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). The binary logistic model is used to estimate the probability of a binary response based on one or more predictor (or independent) variables (features). As such it is not a classification method. It could be called a qualitative response/discrete choice model in the terminology of economics [1]. Thus, it treats the same set of problems as probability regression using similar techniques, with the latter using a cumulative normal distribution curve instead. Equivalently, in the latent variable interpretations of these two methods, the first assumes a standard logistic distribution of errors and the second a standard normal distribution of errors [2]. One situation in which logistic regression is applicable is in model clinical research, where the clinicians are concerned about predicting the patient's survival based upon predictor's [3,4]. For example, the Trauma and Injury Severity Score, which is widely used to predict mortality in injured patients, was originally developed by Boyd et al. using logistic regression [5]. Most regression models are described in terms of the way the outcome variable is modeled: in linear regression the outcome is continuous, logistic regression has a dichotomous outcome, and survival analysis involves a time to event outcome [6]. Logistic regression is the statistical technique used when we wish to estimate the probability of a dichotomous outcome such as the presence or absence of a disease or of death.

For modern clinical medicine, risk prediction procedures are valuable tools for disease prevention and management. Pioneered by the Framingham study, risk score systems have been established for assessing individual risks of developing cardiovascular diseases, cancer or many other conditions within a certain time period [7]. A key component in the assessment of risk algorithm performance is its ability to distinguish subjects who will develop an event ("cases") from those who will not ("controls"). This concept, known as discrimination, has been well studied and quantified for binary outcomes using measures such as the estimated area under the Receiver Operating Characteristics (ROC) curve (AUC), which is also referred to as a "C-statistic". Such a statistic is an estimated conditional probability that for any pair of "case" and "control," the predicted risk of an event is higher for the "case" [8]. To evaluate the adequacy of such a system, C-statistics are routinely used in the medical literature to quantify the capacity of the estimated risk score in discriminating among subjects with different event times. The C-statistic provides a global assessment of a fitted survival model for the continuous event time rather than focusing on the prediction of *t*-year survival for a fixed time. When the event time is possibly censored, however, the Population parameters corresponding to the commonly used C-statistics may depend on the study-specific censoring distribution [9]. We provide a large sample approximation to the distribution of this estimator for making inferences about the concordance measure. Results from numerical studies suggest that the new procedure performs well in finite samples.

Materials and Methods

Data

The ICU data is a type of secondary data. This data was taken from Hosmer and Lemeshow [10]. The ICU study dataset consists of a sample of 200 subjects who were part of a much larger study on survival of patients following admission to an adult intensive care unit (ICU). Vital status (lived/died) of patients after the admission in the ICU are depend mostly in some variable such as age of the patients, service at ICU admission, history of chronic renal failure, sex, systolic blood pressure at ICU admission etc. Data were collected on 200 patients, 40 of which had died and 160 of which had vital status. The predictors of interest were AGE, SEX (male/female), RACE (white/black/other), service at ICU admission (SER (medical/surgical)), cancer part of present problem (no/yes), history of chronic renal failure (CRN) (no/yes), infection probable at ICU admission (INF(no/yes)), CPR prior to ICU admission(CPR (no/yes)), systolic blood pressure at ICU admission (SYS), heart rate at ICU admission (HRA), previous admission to an ICU within 6 months (PRE (no/yes)), type of admission (TYP (elective/emergency)), long bone, multiple, neck, single area, or hip fracture (FRA (no/yes)), p^{O_2} from initial blood gases (P$_{O_2}$(bg>60,bg≤60)), ph from initial blood gases (PH (ph ≥7.25, ph<7.25)), pc$_{o_2}$ from initial blood gases (PCO (pco ≤45, pco>45)), bicarbonate from initial blood gases (BIC (bic ≥18, bic<18)), creatinine from initial blood gases (CRE(cre ≤2.0, cre >2.0)) and level of consciousness at ICU admission (LOC (no/deep stupor/coma)). The main focus here is to illustrate the validation measures for simple binary data.

Estimation of C-statistic for logistic regression model

The C- statistics is numerically identical to the area under the receiver operating characteristic curve (AUC). It equals the proportion of pairs in which the predicted event probability is higher for the subject who experienced the event of interest than that of the subject who did not experience the event. For a pair of subjects (i, j), where i and j correspond to those who experienced the event and those who did not respectively, with event probabilities $\{\pi(\beta|x_i),\pi(\beta|x_j)\}$ $\left\{\pi\left(\beta\,|\,x_i\right),\pi\left(\beta\,|\,x_i\right)\right\}$ the C-statistics can be defined as

$$C = P[\pi(\beta|x_i) > \pi(\beta|x_j)|Y_i = 1\,\&Y_j = 0] \qquad (1)$$

$$C = P\left[\pi\left(\beta\,|\,x_i\right) > \pi\left(\beta\,|\,x_i\right)|\,Y_i = 1\,\&\,Y_j = 0\right]$$

Since there exists a one-to-one transformation between π and $\beta^T x$, the above probability expression can be written as

$$C = Pr[\beta^T x_i > \beta^T x_j|Y_i = 1\,\&Y_j = 0] \qquad (2)$$

$$C = \Pr\left[\beta^T x_i > \beta^T x_i\,|\,Y_i = 1\,\&\,Y_j = 0\right]$$

The value of C statistic lies between 0.5 and 1. A value of 0.5 indicates that the model has no ability to discriminate between low and high-risk subjects, whereas a value of 1 indicates that the model can perfectly discriminate between these two groups. The C-statistic for the logistic regression models can be estimated using both parametric and nonparametric approaches. Under the assumption of normal distribution, the method of maximum likelihood may be used to estimate the C-statistic.

Non-parametric (C-statistic) estimation: Mann-Whitney U statistic (C$_U$)

The widely used non-parametric approach to estimate the C-statistic is based on the Mann-Whitney U statistic and does not require any distributional assumptions regarding the prognostic index. The C-statistic or AUC has been shown to be equal to the $U-statistic$. Let $\eta^{(1)} = \beta^T x_i\,|\,Y_i = 1\,and\,\eta_j = \beta^T x_i\,|\,Y_j = 0$ be PI derived by the model for subject i with event and for subject j without event, respectively. Further, let η_0 and η_1 be the number of events and non-events, respectively. Considering all pairs (i, j), the C-index can be estimated by analogy to the U statistic formulations [11] as:

$$C_U = \frac{1}{n_1 n_0}\sum_{i=1}^{n_1}\sum_{j=1}^{n_0} I(\eta_i^{(1)},\eta_j^{(0)}) \qquad (3)$$

$$C_U = \frac{1}{n_1 n_0}\sum_{i=1}^{n_1}\sum_{j=1}^{n_0} I\left(n_i^{(1)},n_j^{(0)}\right)$$

Where

$$I(\eta_i^{(1)},\eta_j^{(0)}) =$$

$$\{1 \quad if\ \ \eta^1 > \eta^{(0)}\ \tfrac{1}{2} \quad if\ \ \eta^{(1)} = \eta^{(0)}\ 0 \quad if\ \ \eta^{(1)} < \eta^{(0)}$$

$$I\left(\eta_i^{(1)},\eta_j^{0}\right) = \{1\ \ if\,\eta^1 > \eta^{(0)}\frac{1}{2}\ \ if\,\eta^{(1)} = \eta^{(0)}0\ if\,\eta^{(1)} < \eta^{(0)}$$

A concordance pair can be defined (as above using the indicator function) as a pair in which the subject who experienced the event had a higher predicted probability of experiencing the event than the subject who did not experience the event. The total number of pars is the product of number of subjects with event of interest and the number of subjects without an event.

Non-parametric estimation (C-statistic): Kernel statistic (C_K)

To obtain C-statistic or AUC from a smooth ROC curve an alternative to the above estimator suggested by Lloyd [12] is based on standard normal Kernel smoothing. The resulting Kernel estimate of the C-statistic can be written as

$$C_K = \frac{1}{n_1 n_0} \sum_{i=1}^{n} \sum_{i=}^{n} \Phi\left(\frac{\eta_i^{(1)}-\eta_i^{(0)}}{\sqrt{h_1^2+h_0^2}}\right) \qquad (4)$$

$$C_K = \frac{1}{n_1 n_0} \sum\nolimits_{i=1}^{n} \sum\nolimits_{i=}^{n} \phi\left(\frac{\eta_i^{(1)}-\eta_i^{(0)}}{\sqrt{h_1^2+h_0^2}}\right)$$

With the bandwidth $h_1 = 0.9\min\left(s_1, iqr_1/1.34\right)n_1^{-1.5}$, where s_1 and iqr are the standard deviation and inter quartile range of risk score $\eta_i^{(1)}$ and $\phi(\cdot)$ is the standard normal cumulative distribution function. Similarly, one can defined h_0 for risk score $\eta_i^{(0)}$.

Parametric Estimation (C-statistic) (CP)

Based on the central limit theorem, the prognostic index is likely to follow normal distribution as the dimension of the parameter vector β increases [13]. The estimation of the parametric C-index is as follows:

Let $\eta_i^{(1)} = \left(\beta^T x_i \mid Y_j = 1\right) \sim N\left(\mu_0, \sigma_0^2\right)$ and $\eta_i^{(1)} - \eta_j^{(0)} \sim N\left(\mu_1 - \mu_0, \sigma_1^2 + \sigma_0^2\right)$.

The parametric concordance statistic is:

$$C_p = \Pr\left[n_i^{(1)} > n_j^{(0)}\right]$$

After standardizing the term $\eta_i^{(1)} - \eta_j^{(0)}, C_p$ can be obtained as:

$$C_p = \Pr\left[z < \frac{\mu_1 - \mu_0}{\sqrt{\sigma_1^2 + \sigma_0^2}}\right]$$

$$= \phi\left(\frac{\mu_1 - \mu_0}{\sqrt{\sigma_1^1 + \sigma_0^2}}\right) = \Phi\left(\frac{\mu_1 - \mu_0}{\sqrt{\sigma_1^1 + \sigma_0^2}}\right) \qquad (5)$$

Where $Z \sim (0,1)$ and $\phi(\cdot)$ is the standard normal CDF. The estimate of C_p can be obtained by replacing μ_1, μ_0 and, σ_1^2, σ_0^2 by their sample estimates and $\underline{x}_1, \underline{x}_0$ and s_1^2, s_0^2 respectively.

Results and Analysis

Simulation study

In this section, we check the properties of the C-statistics by the standard error for simulation studies. The properties of the C-statistics were investigated in a range of scenarios, created by varying sample size of simulated data. The C-statistics are also evaluated by varying the distributional assumptions of risk score derived from the model. The aim was to identify scenarios where the C-statistics did not perform adequately, for example, whether C- statistics were affected by decreasing sample size. The section begins by describing the simulation design which is followed by describing the strategies for evaluating the C-statistics and the results. Only binary responses were generated from Bernoulli distribution with probability derived from a true logistic model based on ICU data. The covariates of the data were the same. The probability form logistic model were calculated using:

$$\pi\left(\beta \mid x_i\right) = \frac{1}{1 + \exp\left[-\eta_i\right]}$$

$$\pi(\beta|x_i) = \frac{1}{1 + exp[-\eta_i]} \qquad (6)$$

Where $\eta_i = \hat{\beta}^T x$, That is, the estimated coefficients from the fitted model for ICU data were used as the true value of the parameter of the true logistic model.

Simulations with normal distributions

The performance of the different C-statistics was investigated under various simulations Scenarios. As distributional assumption of parametric approach of C-statistic is required, the simulation was conducted varying the distribution of prognostic index. Based on normality assumptions, we try to show the evaluation and comparison of validation measures for the estimator of concordance statistics for different approaches. According to the simulation procedure, we were conducted the simulation study and different validation measures of our estimates are obtained. **Table 1** provides the estimates of C-statistic for methods and its validation properties in simulation scenarios.

From the above table in standard error approaches, we were found that the estimated value of C_U and C_K were close to the true value but C_P has more deviation than others for each scenarios of simulation that means standard error for non-parametric estimators are smaller than parametric estimator. Allowing the effect of sample size on the validation measures, it may be noticed that standard error for all the approaches were increased when the sample size was decreased for all scenarios but it was more for parametric estimators than non-parametric. From above discussion, we can draw an approximate conclusion; standard error was affected by sample size. From the above all discussion, we say that non-parametric estimators C_U provide better result among all estimators.

Table 1 Empirical comparison of concordance statistics for normal distribution.

True Value = 0.746			
Sample Size	C-Statistic	Estimate	SE
189	C_U	0.74639	0.00108
	C_K	0.74639	0.00108
	C_P	0.74702	0.00105
100	C_U	0.74744	0.00216
	C_K	0.74744	0.00215
	C_P	0.74747	0.00192
75	C_U	0.76083	0.00332
	C_K	0.76084	0.00332
	C_P	0.76199	0.00294
50	C_U	0.72038	0.00698
	C_K	0.72033	0.00699
	C_P	0.64758	0.01005

Table 2 Concordance statistics for ICU data.

Concordance Statistics	Estimate	SE	95% C.I.
C_u	0.7995	0.0418	[0.7306, 0.8684]
C_K	0.7996	0.0419	[0.7307, 0.8685]
C_p	0.7933	0.1381	[0.7227, 0.8520]

Assessing the risk model for ICU data

The logistic model was fitted to the ICU data and its discriminatory power assessed using C-statistic. We have to fit the model to derive the prognostic index as well as C-statistic and its properties by using binary response variable STA (vital status) against independent variables AGE, CAN, CPR, SYS, TYP. After constructing the model, estimate of coefficients, its standard error we have to assess the estimated risk model using the C-statistic. Discrimination measure for ICU data is very close for parametric and non-parametric approaches but standard error larger for parametric estimator than the non-parametric and hence provides larger confidence. The estimated C-statistic, its standard error (SE) and confidence interval are showed in **Table 2.**

Discussion and Conclusion

The Concordance statistic is frequently used to assess the discriminatory ability of the for-risk model for binary data. Several approaches including parametric and non-parametric of estimating C-statistic has been proposed in the literature but it is still unclear to the practitioner which approach should generally be used. The results from ICU datasets suggest that “non-parametric” and “Kernel smoothing” estimators provided approximately similar results but “parametric” estimators provided different results particularly produced larger standard error than the others. If the sample size is large under normality of the prognostic index all the approaches produced comparable results. However, when sample size is small, the non-parametric Mann-Whitney U estimator performed better than the non-parametric Kernel smoothing estimators and parametric estimator. Parametric estimator performed well when the sample size was large. Sample size is less depending on the distribution of prognostic index. Above discussion provides the conclusion that nonparametric estimators may be used generally in practice rather than the other estimators. In summary, before evaluating the predictive performance (discriminatory power) of the risk models for binary data using C-statistics, it is essential to check sample size and distribution of the log-odds derived from the model. In the both real data and simulation data we conclude that nonparametric estimator C_U having more discriminatory power than other methods.

References

1 Vella M, Cardozo L, Duckett J (2012) Prognostic research and its potential role in modern gynaecology: A call for more prognostic research in urogynaecology. J Obstet Gynaecol 32: 730-732.

2 Steyerberg EW, Vergouwe Y (2014) Towards better clinical prediction models: Seven steps for development and an ABCD for validation. Eur Heart J 35: 1925-1931.

3 Schisterman EF, Faraggi D, Reiser B (2004) Adjusting the generalized ROC curve for covariates. Stat Med 23: 3319-3331.

4 Choodari-Oskooei B, Royston P, Parmar MK (2012). A simulation study of predictive ability measures in a survival model I: Explained variation measures. Stat Med 31: 2627-2643.

5 Koh K, Kim SJ, Boyd S (2007) An interior-point method for large-scale L1-regularized logistic regression. J Mach Learn Res 8: 1519-1555.

6 Dobson AJ, Barnett G (2018) An introduction to generalized linear models. CRC Press, USA.

7 Altman DG, Royston P (2000) What do we mean by validating a prognostic model? Stat Med 19: 453-473.

8 Austin PC, Steyerberg EW (2012) Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. BMC Med Res Methodol 12: 1-8.

9 DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: A non-parametric approach. Biometrics 837-845.

10 Hosmer DW, Lemeshow S (2000) Applied Logistic Regression (2nd edn) John Willey & Sons, New York, USA.

11 Royston P, Altman DG (2010) Visualizing and assessing discrimination in the logistic regression model. Stat Med 29: 2508-2520.

12 Lloyd CJ (1998) Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. J Am Stat Assoc 93: 1356-1364.

13 Wyatt JC, Altman DG (1995) Commentary on prognostic models: Clinically useful or quickly forgotten? BMJ Case Reports 311: 1539-1541.