Scientific Research Publishing

Journal of Data Analysis and Information Processing, 2019, 7, 190-207
https://www.scirp.org/journal/jdaip
ISSN Online: 2327-7203
ISSN Print: 2327-7211

A Review of the Logistic Regression Model with Emphasis on Medical Research

Ernest Yeboah Boateng, Daniel A. Abaye*

Department of Basic Sciences, School of Basic and Biomedical Sciences, University of Health and Allied Sciences,
Ho, Ghana
Email: eyboateng@uhas.edu.gh, *dabaye@uhas.edu.gh

How to cite this paper: Boateng, E.Y. and Abaye, D.A. (2019) A Review of the Logistic Regression Model with Emphasis on Medical Research. *Journal of Data Analysis and Information Processing*, 7, 190-207.
https://doi.org/10.4236/jdaip.2019.74012

Received: July 31, 2019
Accepted: October 9, 2019
Published: October 12, 2019

Abstract

This study explored and reviewed the logistic regression (LR) model, a multivariable method for modeling the relationship between multiple independent variables and a categorical dependent variable, with emphasis on medical research. Thirty seven research articles published between 2000 and 2018 which employed logistic regression as the main statistical tool as well as six text books on logistic regression were reviewed. Logistic regression concepts such as odds, odds ratio, logit transformation, logistic curve, assumption, selecting dependent and independent variables, model fitting, reporting and interpreting were presented. Upon perusing the literature, considerable deficiencies were found in both the use and reporting of LR. For many studies, the ratio of the number of outcome events to predictor variables (events per variable) was sufficiently small to call into question the accuracy of the regression model. Also, most studies did not report on validation analysis, regression diagnostics or goodness-of-fit measures; measures which authenticate the robustness of the LR model. Here, we demonstrate a good example of the application of the LR model using data obtained on a cohort of pregnant women and the factors that influence their decision to opt for caesarean delivery or vaginal birth. It is recommended that researchers should be more rigorous and pay greater attention to guidelines concerning the use and reporting of LR models.

Keywords

Logistic Regression Model, Validation Analysis, Goodness-of-Fit Measures, Odds Ratio, Likelihood Ratio Test, Hosmer-Lemeshow Test, Wald Statistic, Medical Research

1. Introduction

Logistic regression (LR) analysis has become an increasingly employed statistical tool in medical research, especially over the last two decades [1], although its origin can be dated back to the nineteenth century [2]. It is widely regarded as the statistic of choice for situations in which the occurrence of a binary (dichotomous) outcome is to be predicted from one or more independent (predicting) variables [3] [4] [5].

The logistic function was invented in the 19^{TH} century by Pierre François Verhulst a French mathematician for the description of growth of human populations, and the course of autocatalytic chemical reactions [6]. Verhulst published his suggestions which were edited by Quetelet between 1838 and 1847 [7]. The logistic model agreed very well with the actual course of the population of France, Belgium, Essex (UK), and Russia for the periods up to the early 1830's. The logistic function was discovered anew in 1920 by Pearl and Reed in a study of the population growth of the USA [8].

LR is used when the research method is focused on whether or not an event occurred, rather than when it occurred (time course information is not used). It is particularly appropriate for models involving disease state (diseased or healthy) and decision making (yes or no), and therefore is widely used in studies in the health sciences. There are more complex forms which can deal with situations where the predicted variable takes more than two categories, it is then referred to as polychotomous or multinomial logistic regression [9].

As in all models, certain assumptions are made in order to fit the model to the data. LR does not assume a linear relationship between the dependent and independent variables, but between the logit of the outcome and the predictor values [10]. The dependent variable must be categorical; the independent variables need not be interval; nor normally distributed, nor linearly related, nor of equal variance within each group, and lastly, the categories (groups) must be mutually exclusive and exhaustive. A case can only be in one group and every case must be a member of one of the groups. LR has the power to accommodate both categorical and continuous independent variables. Although the power of the analysis is increased if the independent variables are normally distributed and do have a linear relationship with the dependent variable [11]. Inspection of these assumptions shows that this technique can be employed somewhat more flexibly than traditional regression techniques, making it suitable for many clinically relevant situations. For any given case, LR computes the probability that a case with a particular set of values for the independent variables is a member of the modeled category. Larger samples are needed than for linear regression because maximum likelihood coefficients are large sample estimates [12].

Studies with small to moderate sample sizes employing LR overestimate the effect they measure [4] [13]. Thus, large sample sizes are required for LR to provide sufficient numbers in both categories of the outcome variable. Also, the more independent variables are included, the larger the sample size required. With small sample sizes, the Hosmer-Lemeshow test has low power and is unlikely to detect subtle deviations from the logistic model. Hosmer and Lemeshow recommend sample sizes greater than 400 and a minimum number of cases per independent variable is ten [4] [13].

In addition to its many uses for developing models that will predict events in the physical sciences [14], economics [15] [16] and political sciences [17], LR is increasingly being applied in medical research [18] [19] [20]. Examples of the use of logistic regression in medicine include a study of the factors that predict whether an improvement or no improvement will occur after an intervention [21] [22], the presence or absence of a disease in relation to a variety of factors [23], to explore the effects of and relationships between multiple predictors [24] [25], to determine which of a range of potential predictors actually are important [23] [26] and, to determine whether newly explored variables add to the predictive validity of already established models [27]. The other applications of LR are to develop novel statistical methods based on ranked-data [28].

To examine if commonly recommended assumptions for multivariable LR are addressed, Ottenbacher *et al.* [29] surveyed 99 articles from two journals; the *Journal of Clinical Epidemiology* and the *American Journal of Epidemiology*, under 10 criteria, six dealing with computation and four with reporting multivariable LR results. Their study revealed that three of the 10 criteria were addressed in 50% or more of the articles. Statistical significance testing or confidence intervals were reported in all articles. Methods for selecting independent variables were described in 82% and specific procedures used to generate the models were discussed in 65%. Fewer than 50% of the articles indicated if interactions were tested or met the recommended events per independent variable ratio of 10:1. Fewer than 20% of the articles described conformity to a linear gradient, examined collinearity, reported information on validation procedures, goodness-of-fit, discrimination statistics, or provided complete information on variable coding. There was no significant difference ($P > 0.05$) in the proportion of articles meeting the criteria across the two journals. They concluded that articles reviewed frequently did not report commonly recommended assumptions for using multivariable LR.

Bagley *et al.* [30] also identified 15 peer-reviewed articles and reported on substantial shortcomings in the use and reporting of LR results. Their study revealed that none of the articles reported any goodness-of-fit measures or regression diagnostics. The majority of the studies had events-per-variable ratios near or below 10, suggesting that those regression results themselves may be particularly unreliable, and finally, none of the studies reported any validation analysis.

In a review of four multivariate methods appearing in the literature from 1985 to 1989, Concato *et al.* [31] reported that LR was the most frequently used procedure comprising an average of 43% of the multivariate methods in the five-year period reviewed. Two reports ([21] [32]) described a significant increase in the use of LR in the public health, epidemiology, obstetrics and gynecology research literature. Bender [33] reviewed the statistical methods reported in a probability sample of 348 articles published between 1970 and 1998 in the American Journal of Public Health and the American Journal of Epidemiology.

The study revealed significant increases in the use of LR, proportional hazard regression and methods for the analysis of data from complex sample surveys.

Multivariable LR is a sophisticated statistical technique and concern has been expressed regarding its use and interpretation [29] [34] [35] [36]. The concerns have focused on assumptions associated with the appropriate use, correct interpretation and complete reporting of multivariable LR. The quality of the LR analysis depends heavily on researchers understanding the assumptions inherent in the method and following principles developed to ensure their sound application. Explicitness in modeling is also necessary for reporting the results to other researchers for verification and replication. It is against this back drop that this article aims to re-examine the components of and reporting requirements of the LR model as applied in medical research, and places emphasis on a more thorough and rigorous reporting for a wider audience.

2. Materials and Methods

2.1. The Logistic Regression Model

The LR gives each predictor a coefficient which measures its independent contribution to variation in the dependent variable. The dependent variable Y takes the value **1** if the response is "Yes" and takes a value **0** if the response is "No".

The model form for Predicted Probabilities is expressed as a natural logarithm (ln) of the odds ratio:

$$\ln\left[\frac{P(Y)}{1-P(Y)}\right]=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k \tag{1}$$

and,

$$\frac{P(Y)}{1-P(Y)}=\mathrm{e}^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k} \tag{2}$$

$$P(Y)=\mathrm{e}^{\beta_0+\beta_1X_1+\cdots+\beta_kX_k}-P(Y)\mathrm{e}^{\beta_0+\beta_1X_1+\cdots+\beta_kX_k} \tag{3}$$

$$P(Y)=\frac{\mathrm{e}^{\beta_0+\beta_1X_1+\cdots+\beta_kX_k}}{1+\mathrm{e}^{\beta_0+\beta_1X_1+\cdots+\beta_kX_k}} \tag{4}$$

where, $\ln\left[\frac{P(Y)}{1-P(Y)}\right]$ is the log (odds) of the outcomes, Y is the dichotomous outcome; $X_1,X_2,\cdots,X_k$ are the predictor variables, $\beta_0,\beta_1,\beta_2,\cdots,\beta_k$ are the regression (model) coefficients and β_0 is the intercept.

In Equation (4), the logistic regression model directly relates the probability of Y to the predictor variables. The goal of LR is to estimate the $k+1$ unknown parameters β in Equation (4). This is done with maximum likelihood estimation which entails finding the set of parameters for which the probability of the observed data is greatest. The regression coefficients indicate the degree of association between each independent variable and the outcome. Each coefficient represents the amount of change we would expect in the response variable if there was a one unit change in the predictor variable. The objective of LR is to correctly predict the category of outcome for individual cases using the best model. To accomplish this goal a model is created that include all predictor variables that are useful in predicting the response variable. LR calculates the probability of success over probability of failure. The results of the analysis are in the form of an odds ratio.

2.1.1. The Logistic Curve

The binary dependent variable has the values of **0** and **1** and the predicted value (probability) must be bounded to fall within the same range. To define a relationship bounded by **0** and **1**, LR uses the logistic curve to represent the relationship between the independent and dependent variable. At very low levels of the independent variable, the probability approaches 0, but never reaches **0**. Likewise, if the independent variable increases, the predicted values increase up the curve and approach **1** but never equal to 1.

2.1.2. Transforming a Probability into Odds and Logit Values

The logistic transformation ensures that estimated values do not fall outside the range of 0 and 1. This is achieved in two steps, firstly the probability is re-stated as odds which is defined as the ratio of the probability of the event occurring to the probability of it not occurring. For example, if a horse has a probability of 0.8 of winning a race, the odds of it winning are 0.8/(1 − 0.8) = 4:1. To constrain the predicted values to within 0 and 1, the odds value can be converted back into a probability; thus,

$$Probability\left(event\right)=\frac{odds\left(event\right)}{1+odds\left(event\right)} \tag{5}$$

It can therefore be shown that the corresponding probability is 4/(1 + 4) = 0.8. Also, to keep the odds values form going below 0, which is the lower limit (there is no upper limit), the logit value which is calculated by taking the logarithm of the odds, must be computed. Odds less than 1 have a negative logit value, odds ratio greater than 1.0 have positive logit values and the odds ratio of 1.0 (corresponding to a probability of 0.5) have a logit value of 0.

2.1.3. Interpreting the Odds Ratio (OR)

When an independent variable X_i increases by one unit (X_{i+1}), with all other factors remaining constant, the odds of the dependent variable increase by a factor exp(β_i) which is called the odds ratio (OR) and ranges from zero (0) to positive infinity. It indicates the relative amount by which the odds of the dependent variable increase (OR > 1) or decrease (OR < 1) when the value of the corresponding independent variable increases by one (1) unit.

2.1.4. Selecting the Dependent Variables

In many cases, that outcome event is easily categorized into classes of having occurred, or not having occurred. For example, the occurrence of a heart attack or not; or delivering through caesarean or not, are relatively easily discerned and coded as either having happened, or not having happened. Once this categorization has been achieved, the predictors of that outcome can be studied [37]. In other cases, the outcome may be treated as dichotomous, but, in fact, it derives from the censoring of continuous data; that is, a cutoff criterion has been produced and the data recoded from continuous to categorical at the cutoff point. In these cases, the situation in choosing the outcome variable may be more complicated [16]. In some instances, continuous outcomes translate relatively easily into a dichotomous event. These cases are most often concerned with measures for which well-established cutoff points for the presence of an event have been developed. The presence or absence of high blood pressure is one such example, where a systolic pressure of greater than 140 mm/Hg is considered to be high [32]. It is worth noting that many multi-category or even continuous variables can be reduced to dichotomous ones. For example, if the health condition of patients is expressed on, say a seven-category scale, from "completely healthy" to "terminal condition", this could be reduced to two categories such as "healthy" and "unhealthy" [9].

2.1.5. Selecting Potential Predictors

Another aspect to consider in the development of a LR study concerns the selection of which variables to analyse as potential predictors of the outcome. This can only be achieved by a careful study of the literature in relation to the outcome, in order to ensure that the full range of potential predictors is included [20]. However, there are a number of drawbacks in selecting predictor variables that can lead to the presented logistic model appearing to explain greater or lesser amounts of variance than it actually may explain in reality [38]. The results of any LR will depend on the variables selected as potential predictors, put simply, if a variable is not selected for analysis, then it cannot feature in the final model. However, the choice regarding whether or not to include factors in the initial data set can impact on the results [37].

Further, if interaction terms between the variables are to be considered, then the omission of some variables could potentially have major impacts of the results. Unfortunately, the solution is not simply to include as many variables as possible, as the inclusion of variables that are unrelated to the outcome in question, this (the addition of unrelated variables) has the tendency to inflate the apparent predictive validity of the final model [33]. There is no one best way to tell that the set of predictors that have been chosen are appropriate, but a number of rules-of-thumb can show that the choice is reasonable. For example, if the specificity (the degree to which the predictors correctly identify individuals not showing the particular outcome, true negatives), and sensitivity (the degree to which the predictors correctly identify individuals showing the outcome, true positives), of the model are both above 80%, then it is likely that the chosen predictors have validity [1].

There may well be constraints acting on any particular study that lead to bias in the selection of the data used for the analysis. One potential constraint is the sample size, which limits the number of variables that can be studied. There is some debate as to the number of participants per variable that are needed, however, Agresti [39] suggests that a minimum of 10 participants are needed for every variable studied; a suggestion that is based on some statistical evidence confirming the reliability of logistic regressions performed on different numbers of events per variable [40]. This obviously places some constraints on the number of variables that can be employed in a study, although it should be noted that most studies of medical outcome using LR do follow this rule.

Another source of selection bias in the variables that are studied is that of missing data, where the presence of missing data in the sample can drive down the sample size if those participants with missing data are excluded, or can lead to the exclusion of certain variables from the analysis if large amounts of data are missing. Unfortunately, both of these outcomes can lead to bias in the variables selected that may be highly important as it will leave the sample as self-selected—that is, comprising only those individuals who chose to supply certain data, or only that data which is readily supplied by the sample, as well as other reasons why the other data are missing [19].

Finally, in addition to selection bias effects from these sources, the selection of variables is also constrained by the properties of the data that are collected. For example, predictor variables that are related to one another (that show colinearity or multi-colinearity) or predictor variables that have excessively influential observations (outliers), will impact adversely on the results of a LR. Particularly, in small or moderate samples, colinearity can result in overall levels of significance from the LR when individual predictors are not in themselves predictive of the outcome, or in the degree of relationship between a predictor and the outcome being incorrectly established [22]. Although LR is particularly useful in providing a parsimonious combination of the best predictor variables, such a procedure has the tendency to capitalize on chance sample characteristics [17]. The set of predictors yielded by one sample may not hold for another sample. It is therefore considered desirable when employing this procedure to correct for capitalizing on chance by cross-replicating to a new sample.

2.2. Evaluation of the LR Model

The goodness-of-fit for the LR model can be assessed in several ways. First, is to assess the overall model (relationship between all of the independent variables and dependent variable). Second, the significance of each of the independent variables needs to be assessed. Thirdly, the predictive accuracy or discriminating ability of the model needs to be evaluated, and finally, the model needs to be validated.

2.2.1. Overall Model Evaluation

1) The likelihood ratio test

The overall fit of a model shows how strong a relationship between all of the independent variables, taken together, and dependent variable is. It can be assessed by comparing the fit of the two models with and without the independent variables. A LR model with the k independent variables is said to provide a better fit to the data if it demonstrates an improvement over the model with no independent variables (the null model). The overall fit of the model with k coefficients can be examined through a likelihood ratio test, which tests the null hypothesis:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \tag{6}$$

To do this, the deviance with just the intercept (−2 log likelihood of the null model) is compared with the deviance when the k independent variables have been added (−2 log likelihood of the given model). The difference between the two yields a goodness of fit index G, χ^2 statistic with k degrees of freedom (DoF) [41]. This is a measure of how well all of the independent variables affect the outcome or dependent variable.

$$G = \chi^2 = (-2 \text{ log likelihood of null model}) - (-2 \text{ log likelihood of given model}) \tag{7}$$

An equivalent formula sometimes presented in the literature is,

$$G = \chi^2 = -2 \log\left(\frac{\textit{likelihood of the null model}}{\textit{likelihood of the given model}}\right) \tag{8}$$

where, the ratio of the maximum likelihood is calculated before taking the natural logarithm (ln) and multiplying by −2. The term “likelihood ratio test” is used to describe this test. If the p-value for the overall model fit statistic is less than the significance level of the test, conventionally 0.05 ($P < 0.05$), then H_0 is rejected, with the conclusion that there is evidence that at least one of the independent variables contributes to the prediction of the outcome.

2) Hosmer-Lemeshow test

The Hosmer-Lemeshow test is used to examine whether the observed proportions of events are similar to the predicted probabilities of occurrence in subgroups of the model population. The Hosmer-Lemeshow test is performed by dividing the predicted probabilities into deciles (10 groups based on percentile ranks) and then computing a Pearson’s *Chi*-square (χ^2) that compares the predicted to the observed frequencies in a 2-by-10 table. The value of the test statistics is expressed as,

$$H = \sum_{g=1}^{10} \frac{O_g - E_g}{E_g} \tag{9}$$

where, O_g and E_g denote the observed and expected events for the g^{th} risk decile group. The test statistic asymptotically follows a χ^2 distribution with 8 (number of groups minus 2) DoF. Small values (with large P-value closer to 1) indicate a good fit to the data, therefore, good overall model fit. Large values (with $P < 0.05$) indicate a poor fit to the data. Hosmer and Lemeshow [4] do not recommend the use of this test when n is small (*i.e.* $n < 400$).

2.2.2. Statistical Significance of Individual Regression Coefficients

If the overall model works well, the next question is how important each of the independent variables is. The LR coefficient for the i^{th} independent variable shows the change in the predicted log odds of having an outcome for one unit change in the i^{th} independent variable, all other things being equal. That is, if the i^{th} independent variable, with regression coefficient b, is changed by 1 unit while all of the other predictors are held constant, log odds of outcome is expected to change b units. There are a couple of different tests designed to assess the significance of an independent variable in logistic regression including the likelihood ratio test and the Wald statistic [40].

1) Wald statistic

Statistical tests of significance can be applied to each variable's coefficients. For each coefficient, the null hypothesis that the coefficient is zero is tested against the alternative that the coefficient is not zero using a Wald test, W_j. A Wald test can also be used to compare a full model containing all the predictor variables with a reduced model with some coefficients set to zero. The Wald statistic can be used to assess the contribution of individual predictors or the significance of individual coefficients in a given model [41]. The Wald statistic is the ratio of the square of the regression coefficient to the square of the standard error of the coefficient. The Wald statistic is asymptotically distributed as a χ^2 distribution:

$$W_j = \frac{\beta_j^2}{SE_{\beta j}^2} \tag{10}$$

Each Wald statistic is compared with a χ^2 critical value with 1 DoF.

2) Likelihood ratio test

The likelihood-ratio test used to assess overall model fit can also be used to assess the contribution of individual predictors to a given model. The likelihood ratio test for a particular parameter compares the likelihood of obtaining the data when the parameter is zero, L_0 with the likelihood L_1 of obtaining the data evaluated at the maximum likelihood estimation of the parameter.

The test statistic is calculated as follows:

$$G = -2\ln\frac{L_0}{L_1} = -2\ln\left(L_0 - L_1\right) \tag{11}$$

This statistic is compared with a χ^2 distribution with 1 DoF. To assess the contribution of individual predictors one can enter the predictors hierarchically, then compare each new model with the previous model to determine the contribution of each predictor.

2.3. Predictive Accuracy and Discrimination

2.3.1. Classification Table

The classification table (**Table 1**) is a method to evaluate the predictive accuracy of the logistic regression model [42]. In this table, the observed values for the dependent outcome and the predicted values (at a user defined cutoff value) are cross-classified. For example, if a cutoff value is 0.5, all predicted values above 0.5 can be classified as predicting an event, and all below 0.5 as not predicting the event. Then a two-by-two table of data can be constructed with dichotomous

Table 1. Sample classification table.

Observed	Predicted	
	1	0
1	*a*	*b*
0	*c*	*d*

observed outcomes and dichotomous predicted outcomes. The table has the following form.

Where, *a, b, c* and *d* are the number of observations in the corresponding cells.

If the logistic regression model has a good fit, we expect to see many counts in the *a* and *d* cells, and few in the *b* and *c* cells. In an analogy with medical diagnostic testing, we can consider,

$$\text{Sensitivity} = a/(a+b) \quad \text{and} \quad \text{Specificity} = d/(c+d) \tag{12}$$

where, higher sensitivity and specificity indicate a better fit of the model.

2.3.2. Discrimination with Receiver Operating Characteristic Curves

Extending the above two-by-two idea (Table 1), rather than selecting a single cutoff, the full range of cutoff values from 0 to 1 can be examined. For each possible cutoff value, a two-by-two table can be formed. Plotting the pairs of sensitivity and one minus specificity on a scatter plot provides a Receiver Operating Characteristic (ROC) curve. The area under this curve (AUC) provides an overall measure of fit of the model [43]. The AUC varies from 0.5 (no predictive ability) to 1.0 (perfect predictive ability). Larger AUC indicates better predictability of the model. Points above the diagonal dividing the ROC space represent good classification results (better than random), while points below represent the poor results (worse than random).

2.4. Validation of the LR Model

Validation is an important test of the regression's internal validity, a crucial step in the argument that the regression model is not an idiosyncratic artifact but instead that it has captured essential relationships in the domain of study. An important question is whether results of the LR analysis on the sample can be extended to the population the sample has been chosen from. This question is referred as model validation. In practice, a model can be validated by deriving a model and estimating its coefficients in one data set, and then using this model to predict the outcome variable from the second data set, then check the residuals, and so on. When a model is validated using the data on which the model was developed, it is likely to be over-estimated. Thus, the validity of model should be assessed by carrying out tests of goodness of fit and discrimination on a different data set [44].

If the model is developed with a sub-sample of observations and validated with the remaining sample, it is called internal validation. The most widely used methods for obtaining a good internal validation are data-splitting, repeated data-splitting, jackknife technique and bootstrapping [45]. If the validity is tested with a new independent data set from the same population or from a similar population, it is called external validation. Obtaining a new data set allows us to check the model in a different context. If the first model fits the second data set, there is some assurance of generalizability of the model. However, if the model does not fit the second data, the lack of fit can be either due to the different contexts of the two data sets or true lack of fit of the first model [25].

2.4.1. Pseudo R^2 Measures

If, however, the model does not fit the data set exactly, some indication of how well it does fit should be given. A summary of goodness-of-fit measures describe how well the entire model matches the observed values; in addition, regression diagnostics (including residual, leverage, and influence measures) are important in revealing the effect of individual subjects on the estimated model. A perfect fit has −2 LL value of 0 and R^2_{LOGIT} of 1. The Cox and Snell R^2 measure and Nagelkerke R^2 measure are common in most statistical software packages [38].

2.4.2. Determining the Number of Significant Variables to Retain

Since the estimates of the included variable may be sensitive to changes in the variable(s) omitted, some researchers have chosen to retain all the variables representing the same factor if at least one of them is statistically significant. They refer to such a model as the full model [46] [47] while others chose to eliminate all insignificant variables from the model to increase efficiency of estimation and refer to such a model as the reduced model [48]. To increase the efficiency in medical research, the reduced model with only the statistically significant variables retained is mostly used. In the reduced model, variables with P-value less than or equal to α-value are treated as statistically significant [49].

3. Reporting and Interpreting LR Results

The following four types of information should be included when presenting the LR results; 1) An overall evaluation of the logistic model; 2) statistical tests of individual predictors; 3) goodness-of-fit statistics; and 4) an assessment of the predicted probabilities. We demonstrate this from recent work on variables informing expectant mothers to opt for caesarean delivery or vaginal birth [49]. Tables 2-5 are examples to illustrate the presentation of these four types of information. Table 2 presents the logistic regression model with statistical significance of individual regression coefficients (β) tested using the Wald χ^2 statistic.

From Table 2 baby's birth weight has a significant effect on the event ($P <$ 0.05). Compared with babies with birth weight above 3.5 kg, babies with birth weight less than 3.5 kg were found to have a decreased probability on the event. The negative sign of the estimated coefficients and the sign of the odds ratio being less than 1 ($\beta = -1.5381$, $P < 0.001$ and OR = 0.2148) for babies with birth

Table 2. Example of LR output: statistical tests of individual predictors.

Explanatory Variable	Co-Efficient β	Standard Error	*P*-Value	Wald Test W_j	Odds Ratio OR
Baby's Birth Weight (3.5 kg and above as Reference)					
2.5 - 3.5 kg	−1.5381	0.3988	*0.00012*	−3.857	0.2148
Less than 2.5 kg	−1.6042	0.5148	*0.00183*	−3.116	0.2010
Parity (None as Reference)					
One	1.1588	0.5700	*0.04205*	2.033	3.1861
Two	1.0248	0.5063	*0.04296*	2.024	2.7865
Three	1.1322	0.5273	*0.03178*	2.147	3.1025
Above Three	1.6898	0.6047	*0.0052*	2.794	5.4184

Figures in *italics* are significant ($P < 0.05$). See Reference [50] for full description.

Table 3. Example of output from LR: overall model evaluation and goodness-of-fit statistics.

Test	Categories	χ^2	DoF	*P*-value
Overall Model Evaluation	Likelihood Ratio Test	12.02	2	0.002
	Wald Test	11.06	2	0.004
Goodness of Fit Test	Hosmer and Lemeshow Test	5.975	8	0.65

Table 4. Example output from LR: model summary.

	Likelihood	Cox & Snell R^2 Square	Nagelkerke R^2
1	273.175	0.576	0.723

Table 5. Example output from LR: a classification table.

			Predicted		
			Caesarean Delivery		Percentage % Correct
Observed			Yes	No	
Step 1	Caesarean Delivery	Yes	158	31	83.6
		No	32	148	82.2
	Overall Percentage %				82.9

[a]The cut off value is 0.500.

weight from 2.5 kg to 3.5 kg and ($\beta = -1.6042$, $P < 0.001$ and OR = 0.2010) for babies with birth weight below 2.5 kg show that the probability of caesarean delivery is higher for babies with birth weight above 3.5 kg than babies with birth weights below 3.5 kg. That is, the relative probability of caesarean delivery decreases by 78.52% for babies with birth weight from 2.5 kg to 3.5 kg and 79.9% for babies with birth weight below 2.5 kg.

It could also be seen from Table 2 that parity was estimated to be a significant predictor for the event. Compared with pregnant women with no parity, expectant mothers with one parity (OR = 3.1861, $P = 0.05$), two parities (OR = 2.7865, $P = 0.05$) and three parities (OR = 3.1025, $P = 0.05$) are characterized by significantly higher probability of not undergoing caesarean deliveries. However, expectant mothers with more than three parities, that is, four or more parities (OR = 5.4184, $P < 0.005$) are associated with a very higher probability of not undergoing caesarean delivery. That is, compared with a pregnant woman with no parity and, all other variables held constant, expectant mothers with four or more parities are five times more as likely not to undergo caesarean delivery. The relative probability of not undergoing caesarean delivery increases by 441.84% for expectant mothers with more than four parities and approximately 210.25% for expectant mothers with one to three parities.

From Table 3 two inferential statistical tests for overall model evaluation: the likelihood ratio and Wald tests, are shown. All two tests yield similar conclusions for the given data set. It could be noticed from the results of the likelihood ratio test and the Wald test presented in Table 3 that the logistic model with independent variables was more effective than the null model. Table 3 also presents the Hosmer-Lemeshow goodness-of-fit test. This statistical test measures the correspondence of the actual and predicted (expected) values of the dependent variable (caesarean delivery). A better model fit is characterized by insignificant differences between the actual and expected values. It tests the hypothesis H_0, there is no difference between the predicted and actual values against H_1, there is difference between the predicted and actual values. At p-value of 0.650 the null hypothesis is accepted and we conclude that insignificant differences remain between the actual and expected values, suggesting that the model fitted the data well.

A model summary of the logistic model is presented in Table 4. It could be observed that the model has a relatively larger pseudo R^2 of 0.723 for the Nagelkerke R^2 and 0.576 for the Cox and Snell R^2 That is, the fitted model can explain or account for 72.3% of the variation in the dependent variable. This is an indication of a good model.

Table 5 presents the degree to which predicted probabilities agree with actual outcomes in a classification table. The overall correct prediction, 82.9% shows an improvement over the chance level which is 50%. With the classification table, sensitivity, specificity, false positive and false negative can be measured. Sensitivity measures the proportion of correctly classified events, whereas specificity measures the proportion of correctly classified nonevents. The false positive measures the proportion of observations misclassified as events over all of those classified as events. The false negative therefore measures the proportion of observations misclassified as nonevents over all of those classified as nonevents.

4. Conclusion

This study explored the components LR model, a type of multivariable method used frequently for modeling the relationship between multiple independent variables and a categorical dependent variable, with emphasis on medical research. Six text books on logistic regression and 37 research articles published between 2000 and 2018 which employed logistic regression as the main statistical tool were reviewed. Logistic regression concepts such as odds, odds ratio, logit transformation, logistic curve, assumption, selecting dependent and independent variables, fitting, reporting and interpreting were presented. Upon perusing literature, considerable deficiencies were found in both the use and reporting of LR. For many studies, the ratio of the number of outcome events to predictor variables (events per variable) was sufficiently small to call into question the accuracy of the regression model. Also, most studies did not report validation analysis, regression diagnostics or goodness-of-fit measures. Proper use of this powerful and sophisticated modeling technique requires considerable care both in the specification of the form of the model, in the calculation and interpretation of the model's coefficients. We presented an example of how the LR should be applied. It is recommended that researchers be more thorough and pay greater attention to these guidelines concerning the use and reporting of LR models. In future, researchers could compare LR with other emerging classification algorithms to enable better or more rigorous evaluations of such data.

Authors' Contributions

The idea was developed by EYB. Literature was reviewed by both authors. Both authors contributed to manuscript writing and approved the final manuscript.

Acknowledgements

We thank the anonymous reviewers whose comments made this manuscript more robust.

Funding

This study attracted no funding.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this manuscript.

References

[1] Oommen, T., Baise, L.G. and Vogel, R.M. (2011) Sampling Bias and Class Imbalance in Maximum-Likelihood Logistic Regression. *Mathematical Geosciences*, **43**, 99-120. https://doi.org/10.1007/s11004-010-9311-8

[2] Cramer, J.S. (2002) The Origins of Logistic Regression. Tinbergen Institute Working Paper.

[3] Tu, J.V. (1996) Advantages and Disadvantages of Using Artificial Neural Networks versus Logistic Regression for Predicting Medical Outcomes. *Journal of Clinical Epidemiology*, **49**, 1225-1231. https://doi.org/10.1016/S0895-4356(96)00002-9

[4] Hosmer D.W. and Lemeshow, S. (2000) Applied Logistic Regression. 2nd Edition, Wiley, New York. https://doi.org/10.1002/0471722146
https://onlinelibrary.wiley.com/doi/book/10.1002/0471722146

[5] King, G. and Zeng, L. (2001) Logistic Regression in Rare Events Data. *Political Analysis*, **9**, 137-163. https://doi.org/10.1093/oxfordjournals.pan.a004868

[6] Hosmer, D.W. and Lemeshow, S. (1989) Applied Logistic Regression. John Wiley & Sons, New York.

[7] Bacaër, N. (2011) Verhulst and the Logistic Equation. In: *A Short History of Mathematical Population Dynamics*, Springer, London, 35-39.
https://doi.org/10.1007/978-0-85729-115-8_6

[8] Pearl, R. and Reed, L.J. (1920) On the Rate of Growth of the Population of the United States since 1790 and Its Mathematical Representation. *Proceedings of the National Academy of Sciences of the United States of America*, **6**, 275-288.
https://doi.org/10.1073/pnas.6.6.275

[9] Boateng, E.Y. and Oduro, F.T. (2018) Predicting Microfinance Credit Default: A Study of Nsoatreman Rural Bank Ghana. *Journal of Advances in Mathematics and Computer Science*, **26**, 1-9. https://doi.org/10.9734/JAMCS/2018/33569

[10] Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X. (1989) The Multiple Logistic Regression Model. *Applied Logistic Regression*, **1**, 25-37.

[11] Glantz, S.A. and Slinker, B.K. (1990) Primer of Applied Regression and Analysis of Variance.
https://pdfs.semanticscholar.org/112d/1cdf27a5e3ac74971b7197a1007d00da8271.pdf

[12] Burns, R. P. and Burns, R. (2008) Cluster Analysis. In: *Business Research Methods and Statistics Using SPSS*, Sage, London, 178-211.

[13] Hosmer, D.W., Jovanovic, B. and Lemeshow, S. (1989) Best Subsets Logistic Regression. *Biometrics*, **45**, 1265-1270. https://doi.org/10.2307/2531779

[14] López, L. and Sánchez, J.L. (2009) Discriminant Methods for Radar Detection of Hail. *Atmospheric Research*, **93**, 358-368.
https://doi.org/10.1016/j.atmosres.2008.09.028

[15] Boyacioglu, M.A., Kara, Y. and Baykan, Ö.K. (2009) Predicting Bank Financial Failures Using Neural Networks, Support Vector Machines and Multivariate Statistical Methods: A Comparative Analysis in the Sample of Savings Deposit Insurance Fund (SDIF) Transferred Banks in Turkey. *Expert Systems with Applications*, **36**, 3355-3366. https://doi.org/10.1016/j.eswa.2008.01.003

[16] Karp, S. (2009) The Problem of Media Economics: Value Equations Have Radically Changed. The (r) Evolution of Media, Publishing 2.0.
https://publishing2.com/2009/01/07/

[17] King, G., Tomz, M. and Wittenberg, J. (2000) Making the Most of Statistical Analyses: Improving Interpretation and Presentation. *American Journal of Political Science*, **44**, 341-355. https://doi.org/10.2307/2669316
https://web.stanford.edu/~tomz/pubs/ajps00.pdf

[18] Srivastava, N. (2005) A Logistic Regression Model for Predicting the Occurrence of Intense Geomagnetic Storms. *Annales Geophysicae*, **23**, 2969-2974.
https://doi.org/10.5194/angeo-23-2969-2005

[19] Jiang, X., El-Kareh, R. and Ohno-Machado, L. (2011) Improving Predictions in Imbalanced Data Using Pairwise Expanded Logistic Regression. *AMIA Annual Symposium Proceedings*, **2011**, 625-634.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3243279/

[20] Reed, P. and Wu, Y. (2013) Logistic Regression for Risk Factor Modelling in Stuttering Research. *Journal of Fluency Disorders*, **38**, 88-101. https://doi.org/10.1016/j.jfludis.2012.09.003

[21] Khan, K.S., Chien, P.F. and Dwarakanath, L.S. (1999) Logistic Regression Models in Obstetrics and Gynecology Literature. *Obstetrics & Gynecology*, **93**, 1014-1020. https://doi.org/10.1097/00006250-199906000-00024 https://www.ncbi.nlm.nih.gov/pubmed/10362173

[22] Kim, Y., Kwon, S. and Song, S.H. (2006) Multiclass Sparse Logistic Regression for Classification of Multiple Cancer Types Using Gene Expression Data. *Computational Statistics & Data Analysis*, **51**, 1643-1655. https://doi.org/10.1016/j.csda.2006.06.007

[23] Howell, P. and Davis, S. (2011) Predicting Persistence of and Recovery from Stuttering by the Teenage Years Based on Information Gathered at Age 8 Years. *Journal of Developmental & Behavioral Pediatrics*, **32**, 196-205. https://doi.org/10.1097/DBP.0b013e31820fd4a9

[24] Jones, S.R. and McEwen, M.K. (2000) A Conceptual Model of Multiple Dimensions of Identity. *Journal of College Student Development*, **41**, 405-414. https://www.researchgate.net/publication/292759031_A_conceptual_model_of_multiple_dimensions_of_identity

[25] Vollmer, R.T. (1996) Multivariate Statistical Analysis for Pathologists: *Part I, The Logistic Model. American Journal of Clinical Pathology*, **105**, 115-126. https://doi.org/10.1093/ajcp/105.1.115

[26] Holland, A.L., Greenhouse, J.B., Fromm, D. and Swindell, C.S. (1989) Predictors of Language Restitution Following Stroke: A Multivariate Analysis. *Journal of Speech, Language, and Hearing Research*, **32**, 232-238. https://www.ncbi.nlm.nih.gov/pubmed/2739374

[27] Fleck, M.P.D.A., Simon, G., Herrman, H., Bushnell, D., Martin, M. and Patrick, D. (2005) Major Depression and Its Correlates in Primary Care Settings in Six Countries: 9-Month Follow-up Study. *The British Journal of Psychiatry*, **186**, 41-47. https://doi.org/10.1192/bjp.186.1.41

[28] Mahdizah, M. and Zamanzade, E. (2019) Efficient Body Fat Estimation Using Multistage Pair Ranked Set Sampling. *Statistical Methods in Medical Research*, **28**, 223-234. https://doi.org/10.1177/0962280217720473

[29] Ottenbacher, K.J., Ottenbacher, H.R., Tooth, L. and Ostir, G.V. (2004) A Review of Two Journals Found that Articles Using Multivariable Logistic Regression Frequently Did Not Report Commonly Recommended Assumptions. *Journal of Clinical Epidemiology*, **57**, 1147-1152. https://doi.org/10.1016/j.jclinepi.2003.05.003

[30] Bagley, S.C., White, H. and Golomb, B.A. (2001) Logistic Regression in the Medical Literature: Standards for Use and Reporting, with Particular Attention to One Medical Domain. *Journal of Clinical Epidemiology*, **54**, 979-985. https://doi.org/10.1016/S0895-4356(01)00372-9 https://www.ncbi.nlm.nih.gov/pubmed/11576808

[31] Concato, J., Feinstein, A.R. and Holford, T.R. (1993) The Risk of Determining Risk with Multivariable Models. *Annals of Internal Medicine*, **118**, 201-210. https://doi.org/10.7326/0003-4819-118-3-199302010-00009

[32] Chien, K., Cai, T., Hsu, H., Su, T., Chang, W., Chen, M., Lee Y. and Hu, F.B. (2009) A Prediction Model for Type 2 Diabetes Risk among Chinese People. *Diabetologia*, **52**, 443-450. https://doi.org/10.1007/s00125-008-1232-4

[33] Bender, R. (2009) Introduction to the Use of Regression Models in Epidemiology. In: Verma, M., Ed., *Cancer Epidemiology. Methods in Molecular Biology*, **471**, 179-195. https://doi.org/10.1007/978-1-59745-416-2_9

[34] Hall, G.H. and Round, A.P. (1994) Logistic Regression-Explanation and Use. *Journal of the Royal College of Physicians of London*, **28**, 242-246. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5400976/

[35] Sun, G.W., Shook, T.L. and Kay, G.L. (1996) Inappropriate Use of Bivariable Analysis to Screen Risk Factors for Use in Multivariable Analysis. *Journal of Clinical Epidemiology*, **49**, 907-916. https://doi.org/10.1016/0895-4356(96)00025-X

[36] Bender, R. and Grouven, U. (1996) Logistic Regression Models Used in Medical Research Are Poorly Presented. *British Medical Journal*, **313**, 628. https://doi.org/10.1136/bmj.313.7057.628

[37] Levy, P.S. and Stolte, K. (2000) Statistical Methods in Public Health and Epidemiology: A Look at the Recent Past and Projections for the Next Decade. *Statistical Methods in Medical Research*, **9**, 41-55. https://doi.org/10.1177/096228020000900106

[38] Park, H.A. (2013) An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain. *Journal of Korean Academy of Nursing*, **43**, 154-164. https://doi.org/10.4040/jkan.2013.43.2.154

[39] Agresti, A. (2007) Building and Applying Logistic Regression Models. In: *An Introduction to Categorical Data Analysis*, 2nd Edition, John Wiley & Sons, Hoboken, NJ, 137-172. https://doi.org/10.1002/9780470114759.ch5

[40] Menard, S. (2002) Applied Logistic Regression Analysis. 2nd Edition, Volume 106, Sage, New York.

[41] Bewick, V., Cheek, L. and Ball, J. (2005) Statistics Review 14: Logistic Regression. *Critical Care*, **9**, 112-118. https://doi.org/10.1186/cc3045

[42] Peng, C.Y.J. and So, T.S.H. (2002) Logistic Regression Analysis and Reporting: A Primer. *Understanding Statistics*: *Statistical Issues in Psychology*, *Education*, *and the Social Sciences*, **1**, 31-70. https://doi.org/10.1207/S15328031US0101_04 https://www.researchgate.net/profile/Joanne_Peng2

[43] Bewick, V., Cheek, L. and Ball, J. (2004) Statistics Review 13: Receiver Operating Characteristic Curves. *Critical Care*, **8**, 508-512. https://doi.org/10.1186/cc3000

[44] Giancristofaro, R.A. and Salmaso, L. (2007) Model Performance Analysis and Model Validation in Logistic Regression. *Statistica*, **63**, 375-396.

[45] Harrell, F.E., Lee, K.L. and Mark, D.B. (1996) Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. *Statistics in Medicine*, **15**, 361-387. https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4

[46] Lee, J. and Mannering, F. (2002) Impact of Roadside Features on the Frequency and Severity of Run-off-Roadway Accidents: An Empirical Analysis. *Accident Analysis & Prevention*, **34**, 149-161. https://doi.org/10.1016/S0001-4575(01)00009-4

[47] Tay, R., Choi, J., Kattan, L. and Khan, A. (2011) A Multinomial Logit Model of Pedestrian-Vehicle Crash Severity. *International Journal of Sustainable Transportation*, **5**, 233-249. https://doi.org/10.1080/15568318.2010.497547

[48] Wang, X. and Abdel-Aty, M. (2008) Analysis of Left-Turn Crash Injury Severity by Conflicting Pattern Using Partial Proportional Odds Models. *Accident Analysis & Prevention*, **40**, 1674-1682. https://doi.org/10.1016/j.aap.2008.06.001

[49] Campillo, C. (1993) Standardizing Criteria for Logistic Regression Models. *Annals of Internal Medicine*, **119**, 540-541.
https://doi.org/10.7326/0003-4819-119-6-199309150-00036

[50] Boateng, E.Y., Bosson-Amedenu, S., Nortey, E.N.N. and Abaye, D.A. (2019) Non-Medical Determinants of Caesarean Deliveries in Ghana: A Logistic Regression Approach. *Open Journal of Applied Sciences*, **9**, 492-505.
https://doi.org/10.4236/ojapps.2019.96039