# Logistic Regression in Biomedical Study

**Kaizhi Lu**[*]

Statistics, Economics University of Washington Seattle, WA, United States

[*]Corresponding author: kaizhilu@uw.edu

**Keywords:** Logistic regression, variable selection, model building, model validation, output interpretation, biostatistics.

**Abstract:** Nowadays statistical tools have become an indispensable part in biomedical studies. Logistic regression, a model describing and estimating the relationship between one dependent variable and one or more independent variables, is one of the most widely used statistical analyses in multivariable models in medical research. Researchers need to be fully aware of the function of research designs applied, the applicability of statistical tests used, and the validity of the conclusions drawn. However, due to little time devoted in statistical training, researchers in epidemiology are short of the ability in system analysis and mathematical reasoning. This could result in generating avoidable statistical mistakes and compromising the final finding. So, a corresponding review and analysis of the logistic regression is in need. This article provides a walkthrough for creating logistic regression model within the context of medical study. It starts with the introduction of the model's definition and follows by the discussion of operation and caution in each step of its application including variable selection, model building, model validation, and output interpretation.

## 1. Introduction

The absence uses of statistical methods is nearly impossible in today's medical literature for reading a clinical study or other medical research report [1]. Statistical methods extract information from research data and make valid inferences in a wider population based on the occurrence of events in the small group. Improper statistical methods will result in erroneous conclusions leading to unethical practice and thus should be avoided [2]. However, the inappropriate application of statistical methods to analyze research data is a common error found in the medical literature [3]. So, it is crucial for the researchers to have a basic understanding of commonly used statistical methods in order to reach accurate conclusions. Over the last two decades, logistic regression analysis has become an increasingly employed statistical method in medical research [4]. It is widely regarded as the statistic of choice for situations in which the occurrence of a binary outcome is to be predicted from one or more independent variables [5].

Logistic function was firstly invented in the 19th century by Belgian mathematician Pierre François Verhulst for the description of the growth of populations, and the course of autocatalytic chemical reactions [6]. Verhulst used the logistic curve as a growth curve for exhibiting the course of a proportion P over time t as Prose monotonically between the bonds of 0 and 1. It agreed very well with the actual course of the population of France, Belgium, Essex and Russia for periods up to 1833 [6]. Nowadays, because of its ability to build a linear relationship between the binary response and predictors by using a link function, it has been widely used in medical research, being used to predict the risk of developing a given disease based on observed characteristics of the patient.

Logistic regression belongs to the family of generalized linear model (GLM). GLM is an advanced statistical modelling technique formulated by John Nelder and Robert Wedderburn in 1972 [7]. It allows researchers to build a linear relationship between the response and predictors by using a link function when their underlying relationship is not linear. Comparing with standard linear regression, GLM has relaxed the restrictions on independent variable and response variable for being able to be applied on a broader range of data: the random variable does not need to have the same probability distribution; the response variable does not need to be normally distributed; homoskedasticity (i.e., constant variances) needs not be satisfied. With less assumptions are made to the data and the facts that most clinical outcomes are defined in binary form, logistic regression is more attractive than other linear regressions.

In this paper, the components of and reporting requirements of the logistic regression model as applied in medical research are discussed and explained. It starts with a brief mathematical definition and transformation of the model; follows detailed discussion of the process of variable selection with special emphasis in sample size, dependent variable, and independent variable; then presents and evaluates the processes of model building, model validation and output interpretation in biomedical context.

## 2. Logistic Regression

Logistic regression is a model describing and estimating the relationship between one dependent variable and one or more independent variables. The dependent variable is normally binary taking only the value of 0 or 1; although, it can be extended to categorical variable with multiple classes in more complex models. The independent variables or predictors can be either binary or continuous (taken any real value). Log-odds, the log of the odds of being in one outcome category versus the other category, are introduced to resolve the discrepancy between the range of independent variables and the range of dependent variable. A linear relationship between the predictor variables and the log-odds is assumed, and the corresponding link function can be algebraically written as:

$$log - odds = log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \tag{1}$$

Where, $X_1, X_2, \dots, X_n$ are independent variables, $\beta_0$ is the intercept, $\beta_1 \dots \beta_n$ are corresponding estimate parameters, and $p$ is the probability that the event occurs. The odds can be recovered by exponentiating the log-odds:

$$odds = \frac{p}{1-p} = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n} \tag{2}$$

With introduction of binary variable Y and further algebraic manipulation, the probability of the event becomes:

$$p = P(Y = 1 |\, \mathrm{X}) = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n}} \tag{3}$$

This is the logistic function applied in the logistic regression. One can calculate the probability that the individual has the event of interest ($p$) from the natural exponent ($exp(X)$) of the sum of the product of each of the covariates ($X_i$) and their corresponding parameter estimates ($\beta_i$). The regression parameters produced by the logistic regression here,$\beta_1 \dots \beta_n$, represent log-odds ratios. They indicate the amount of change expected in the log-odds for one-unit change in one predictor variable with all other variables hold constant. As the model itself models probability of output in terms of input, it has also been used to make a binary classifier for machine learning application.

## 3. Application

### 3.1 Variable Selection

### 3.1.1 Sample Size.

Sample size, which is usually represented by n, refers to the number of participants or observations included in a study. It influences the precision of estimates and the power of the study to draw conclusions. Larger sample sizes can present more accurate mean values, identify outliers that skew the data in a smaller sample, and provide a smaller margin of error. Margin of error, a statistic expressing the maximum deviation of the sample results from the real values, indicates how many percentage points the finding results will differ from the real population value and is inversely related to the sample size. Therefore, determination of sample size requirement is important and necessary before conducting the analysis.

The sample size requirement for logistic regression has been discussed in the literature. Peduzzi et al. propose that the minimum required sample size should be based on the rule of event per variable (EPV), and the concept of EPV of 10 is acceptable for logistic regression [8]. It suggests the logistic model should be used with a minimum of 10 events per predictor variable. For instance, the sample size of a model with 10 predictor variables should be at least 100 ($100 = 10 \times 10$). According to Bujang et al., observation studies that involve logistic regression in the analysis are recommended a minimum sample size of 500 to derive statistics that can represent the parameters in the targeted population [9]. Moreover, they propose a simpler formula for sample size estimation particularly for logistic regression in observational studies. The formula is $n = 100 + 50i$ where $i$ refers to number of independent variables in the model. For sample with 10 independent variables, the sufficient sample size being able to make inference on the targeted population is 600 ($600 = 100 + 50 \times 10$).

### 3.1.2 Dependent Sample.

In most cases, the outcome event or dependent variable is categorized into classes of having occurred or not having occurred. For example, diseased or disease free; dead or alive; positive or negative test are easily coded as either the outcome having happened or not having happened. In other cases, the dichotomous outcome may be derived from the censoring of continuous data. With a cut-off criterion been produced and settled, the raw data are recoded from continuous or multi-category to binary at the cut-off point. The cases with well-established cut-off points can make the translation relatively easy. For example, a hepatitis C virus (HCV) ribonucleic acid (RNA) result greater than 800,000 IU/L has been regarded as high viral load. However, the translation generally makes the situation of choosing the outcome variable more complicated and thus need detailed explanation and validation.

### 3.1.3 Independent Sample.

A major problem, also the key to success, when building a logistic regression model is to choose the correct independent variables to enter the model. A variable cannot feature in the final model if it is not selected for analysis. So, a detailed study of the literature related to the outcome variable is in need to include the full range of potential predictors. For being afraid of missing one significant variable, researchers may be tempted to include as many collected variables as possible in the model. However, too many independent variables in the model will lead to a mathematically unstable outcome and decreased generalizability beyond the study sample. The addition of unrelated variables has the tendency to inflate the apparent predictive validity of the final model. It can dilute true associations and display spurious associations between variables instead. So, the researchers must be very cautious with the selection of predictors' variables to be included in the model.

One important consideration during the variable selection process is the basic assumptions of logistic regression. The four basic assumptions must always be satisfied to conduct logistic regression. The first assumption is independence of observations and errors. The observations must be independent of each other, not coming from repeated or paired data. If one's data include repeated measures or other correlated outcomes, errors will be similarly correlated, and the assumption is violated. A special attention should be paid to when dealing with time-series data, where the correlation between sequential observations can be an issue.

The second assumption is linearity of independent variables and log-odds. The relationship between each continuous independent variable and respective logit-transformed outcome should be linear. The logit, also known as the log-odds, is the logarithm of the odds ratio discussed previously. One simple way for checking logit linearity is by visually inspecting the scatter plot between each predictor and the logit values. If the scatter plot shows a clear non-linear pattern of the variable related to its respective log-odds, then the assumption of log linearity is most likely violated. Another way for checking is Box-Tidwell test. It is effectively used by adding the non-linear transform of the original predictor as an interaction term to test if this addition made no better prediction. For example, for one continuous independent variable age, the new interaction term *age * ln (age)* will be added to check for its statistically significance. If the interaction term is not statistically significant (i.e., p < 0.05), then the independent variable is linearly related to the logit of the outcome variable and the assumption is satisfied.

The third assumption is lack of strongly influential outliers. With an unexpectedly large impact on model results, these problematic values can distort the outcome and accuracy of the model. Cook's Distance has normally been used to determine the influence of a data point. With Cook's Distance greater than *4/n* where *n* is the number of observations, the observations are deemed as influential. Standardized residuals are used to determine the outlier. Data points with absolute standardized residual values greater than 3 represent possible extreme outliers. Observations that are identified as influential outliers should be removed or transformed.

The final assumption is the absence of multicollinearity among independent variables. Collinearity, or multicollinearity, refers to the phenomenon in which one predictor variable in a multiple regression model is highly correlated with the other predictor and can be linearly predicted from it. Collinearity is a threat problem because it can reduce the precision of the estimated coefficients and weaken the statistical power of the regression model. When independent variables are correlated that changes in one variable are associated with shifts in another, it becomes difficult for the model to estimate the relationship between each independent variable and the dependent variable independently. It can lead to highly unstable and biased estimate of exposure effect and erroneous conclusions about significant or no significant individual predictors. Signs of collinearity include large standard errors with wide confidence intervals and large changes in estimated parameters of the affected predictors when small changes in data happened. Variable Inflation Factor (VIF), the ratio of the overall model variance to the variance of a model that includes only that single independent variable, has been used to measure the degree of multicollinearity. A VIF value that exceeds 5 indicates a problematic variable. Clinical and medical expertise can also be used to detect collinearity. For example, the body mass index (BMI) and waist circumference (WC) are some of the widely known risk factors for obesity related health outcomes existing significant correlation that may cause collinearity [10].

### 3.2 Model Building

A model building strategy is closely linked to the selection of independent variables. The conventional technique is to first run the univariate analyses (the simplest form of analysing data with only one explanatory variable at a time) on all predictors and then run the multivariate model with variables that meet a pre-set cut-off for significance [11]. The pre-set cut-off for significance is often more liberal and relaxed than the conventional cut-off for significance, i.e., $P \leq 0.25$ instead of the usual $P < 0.05$, since it is a pre-selection strategy aiming to identify potential predictor variables and no inference will be derived from this step. So, there is no need to worry about a rigorous p-value criterion at this stage. Indeed, this relaxed P-value criterion can help reduce the initial number of variables in the model while at the same time reduce the risk of missing important variables. While applying this conventional technique, researchers need to consider the scientific plausibility and the clinical meaningfulness of the association between independent variables. For instance, variables such as white hair and baldness may show significant result in association with the risk of occurrence of myocardial infarction using univariate analyses. However, these associations are due to the association with older age and male sex and hence must not be entered into the regression model.

There are multiple statistical tools that can be used to evaluate the goodness-of-fit for the logistic regression model, such as likelihood ratio test, pseudo$R^2$, and the Hosmer-Lemeshow test. The likelihood ratio test is performed by estimating two models and comparing the fit of one model to the fit of the other. It compares the likelihood of the data under the full model against the likelihood of the data under a model with fewer predictors. The null hypothesis holds the smaller model provides as good a fit for the data as the larger model. A p-value with less than 0.05 would provide evidence against the reduced model in favour of the current model. The most notable pseudo $R^2$ in logistic regression is McFadden's $R^2$, which is defined as $1 - ln(L_m)/ln(L_0)$ with $L_m$ referring to the log likelihood value for the fitted model and $L_0$ being as the log likelihood for the null model with only an intercept as a predictor. Values closer to zero indicate the model has no predictive power and thus should not be used. The Hosmer-Lemeshow test is performed by dividing the predicted probabilities into subgroups (commonly 10) and then computing a Pearson's Chi-square ($\chi^2$) that compares the predicted to the observed frequencies. It is used to examine whether the observed proportions of events are like the predicted probabilities of occurrence in subgroups of the model population. This suggests that suppose the observations in the first group (10 in total) have a predicted probability of 0.1, then if the model is correctly specified, the observed proportion who have Y=1 would be expected to be 10%. A significant p-value would indicate the poor fit. And it is not recommended to use this test when the sample size is small (i.e., n < 400) [12]. For example, Ozdemir et al. used Hosmer-Lemeshow test to evaluate and determine the logistic regression model and coefficients aiming for assess the effects of physical activity on quality of life, depression and anxiety levels during the COVID-19 outbreak [13].

### 3.3 Model Validation

Model validation, referring to the process of confirming that the logistic regression model can be extended to its intended population, is an important step. It argues the regression model does capture essential relationships in the domain of study rather than serves as an artifact. The process can be divided into internal and external validation based on the choice of its validation data set. If the model is developed with a sub-sample of observations and validated by the remaining sample, it is called internal validation. The commonly used methods for obtaining a good internal validation include the holdout method, K-fold cross-validation, and bootstrapping. External validation refers to the situation that the validity is tested with a new independent data set from the same population or from a similar population. If the model fits the new data set in different context, then there is some assurance of generalizability of the model. However, if the results of either internal or external validation fail, it is advisable to adjust the model as needed, or to explicitly define any restrictions for the model's future use.

Common statistical validation indexes include classification accuracy, specificity, sensitivity, and the area under receiver operating characteristic (AUC). Classification accuracy, equalling to the number of correct predictions divided the number of all predictions, measures the percentage of the correct prediction. Specificity, also known as the true negative rate (TNR), measures the degree to which the predictors correctly identify individuals not showing the particular outcome. Sensitivity, also known as the true positive rate (TPR), measures the degree to which the predictors correctly identify individuals showing the outcome. If the specificity and sensitivity of the model are both above 80%, then it is likely that the tested model has validity. The receiver operating characteristic (ROC) curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold varies. It is plotted with TPR against the FPR. AUC represents the degree of separability and its capability of distinguishing between classes. Its values range from 0.50 to 1.00, with 0.5 reflecting random forecasts and 1.0 implying perfect forecasts. For example, to provide support for the validity of the nine important risk factors for lupus nephritis (LN) patients with hypothyroidism, Huang et al. calculated their model's AUC value finding it to be 0.885 [14].

### 3.4 Output Interpretation

Independent variables are usually presented as odds ratio (OR) in the output of the logistic regression model. It indicates how much the odds of a particular outcome change for a 1-unit increase in the independent variable. When a logistic regression is calculated, the regression coefficient ($\beta_i$) is the estimated increase in the log-odds of the outcome per unit increase in the value of the exposure; and its exponentiated form ($e^{\beta_i}$) is interpreted as OR.

Odds are the ratio between probabilities: the probability of an event favourable to an outcome and the probability of an event against the same outcome. And odds ratio is the ratio of odds: the odds of an event in the treatment group to the odds of an event in the control group. It effectively represents the constant effect of an exposure on the likelihood that one outcome will occur, revealing the strength of the independent variable's contribution to the outcome. Crude or unadjusted OR refers to the odds ratio for a logistic regression model with only one independent variable. For example, in Bari et al.'s study of hepatitis C virus (HCV) infection, an OR of 2.6 for the variable armpit shave from barber in univariate analysis represents the odds of adult males with armpit shaving from barber been exposed to HCV are 2.6 times greater than the odds of those without (reference group) [15]. It is also equivalent to express that there is a $(2.6 - 1.0) \times 100\% = 160\%$ increase in the odds of being infected for adult males with armpit shaving versus those without. Adjusted OR refers to the OR that controls for the other predictor variables in the model, when the model includes multiple independent variables. It is aimed to understand how a predictor variable affects the odds of an event occurring after adjusting for the effect of other predictor variables. For example, with predictor variables age, therapeutic injections received in past 10 years and frequency of facial shave from barber included in the regression model, the adjusted OR of 2.9 represents the unique contribution of the variable armpit shave from barber to HCV infection with the other three variables holding at constant values.

Interpretation of odds ratio is mistakable for mixing up with relative ratio (RR), another commonly used statistic for quantifying the relationship between variables. It is the ratio of the probability of an outcome in an exposed group to the probability of an outcome in an unexposed group. For instance, it is incorrect to state that the risk of infection for armpit shaving is 2.6 times greater. This is because OR tends to exaggerate the estimate of relationship between exposure and response than RR [16]. While the definition of RR is straightforward and intuitive making it preferable than OR, the failure of the calculation of risk in certain studies makes it be less used than OR. Specifically, the calculation of risk requires the use of "people at risk" as the denominator, but it is not always available in study as retrospective case-control study. Researchers should be careful in explaining the result to avoid stating false conclusion from correct mathematical results.

Finally, the 95% confidence interval (CI) is routinely reported with OR to provide an estimate of its precision. The level of precision is inversely related with the width of CI. It can be used as a proxy for the presence of statistical significance by observing if it overlaps the null value (OR=1). For example, as the OR of 1.7 for past dental treatment in Bari's study has a 95% CI of 0.9 to 3.1 spans 1.0, researchers cannot state there exists definitive evidence that past dental treatment is a significant contributor to the infection of HCV. Further detailed research is needed for studying the effect of this independent variable on the outcome.

## 4. Conclusion

Logistic regression is an efficient and powerful tool allowing assessment of the relationship between one or more independent variables and a binary outcome. However, deficiencies such as sufficiently small ratio of the number of outcome events to predictor variables, lack of regression diagnostics or goodness-of-fit measures, and confusion of odds ratio and relative risk in the application of the model can compromise the accuracy the results. Therefore, researchers must pay considerable attention to the proper use of this powerful and sophisticated modelling technique and avoid simply putting raw data into the computer and jumping straightforward to conclusion. A basic understanding of the model's background and mathematical definition; special attention to the selection process of sample size, dependent variable and independent variable; and the operation of model building, model validation and output interpretation in biomedical context are in need for a more thorough research study. In future, a more direct and rigorous evaluation model of the existed clinical data is expected.

## References

[1] Elenbaas, R. M., Elenbaas, J. K., & Cuddy, P. G. (1983). Evaluating the medical literat ure. Part II: Statistical analysis. Annals of emergency medicine, 12(10), 610–620. https://doi.org/10.1016/s0196-0644(83)80205-4

[2] Sprent P. (2003). Statistics in medical research. Swiss medical weekly, 133(39-40), 522–529.

[3] Glantz SA: Primer of Biostatistics. New York, McGraw-Hill Book Co, 1981.

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

[5] Boateng, E.Y., Abaye, D.A.: A review of the logistic regression model with emphasis on medical research. J. Data Anal. Inf. Process. 07(04), 190–207 (2019). https://10.4236/jdaip.2019.74012

[6] Cramer, J.S. (2002) The Origins of Logistic Regression, TI 2002-119/4. Tinbergen Institute Discussion Paper. https://doi.org/10.2139/ssrn.360300

[7] Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized Linear Models. Journal of the Royal Statistical Society. Series A (General), 135(3), 370–384. https://doi.org/10.2307/2344614

[8] Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. Journal of clinical epidemiology, 49(12), 1373–1379. https://doi.org/10.1016/s0895-4356(96)00236-3

[9] Bujang, M. A., Sa'at, N., Sidik, T., & Joo, L. C. (2018). Sample Size Guidelines for Logistic Regression from Observational Studies with Large Population: Emphasis on the Accuracy Between Statistics and Parameters Based on Real Life Clinical Data. The Malaysian journal of medical sciences: MJMS, 25(4), 122–130. https://doi.org/10.21315/mjms2018.25.4.12

[10] Tran, N., Blizzard, C. L., Luong, K. N., Truong, N., Tran, B. Q., Otahal, P., Nelson, M., Magnussen, C., Gall, S., Bui, T. V., Srikanth, V., Au, T. B., Ha, S. T., Phung, H. N., Tran, M. H., & Callisaya, M. (2018). The importance of waist circumference and body mass index in cross-sectional relationships with risk of cardiovascular disease in Vietnam. PloS one, 13(5), e0198202. https://doi.org/10.1371/journal.pone.0198202

[11] Ranganathan, P., Pramesh, C. S., & Aggarwal, R. (2017). Common pitfalls in statistical analysis: Logistic regression. Perspectives in clinical research, 8(3), 148–151. https://doi.org/1 0.4103/picr.PICR_87_17

[12] Hosmer, D. W., Jovanovic, B., & Lemeshow, S. (1989). Best Subsets Logistic Regression. Biometrics, 45(4), 1265–1270. https://doi.org/10.2307/2531779

[13] Ozdemir, F., Cansel, N., Kizilay, F., Guldogan, E., Ucuz, I., Sinanoglu, B., Colak, C., & Cumurcu, H. B. (2020). The role of physical activity on mental health and quality of life during COVID-19 outbreak: A cross-sectional study. European journal of integrative medicine, 40, 101248. https://doi.org/10.1016/j.eujim.2020.101248

[14] Huang, T., Li, J., & Zhang, W. (2020). Application of principal component analysis and logistic regression model in lupus nephritis patients with clinical hypothyroidism. BMC medical research methodology, 20(1), 99. https://doi.org/10.1186/s12874-020-00989-x

[15] Bari, A., Akhtar, S., Rahbar, M. H., & Luby, S. P. (2001). Risk factors for hepatitis C virus infection in male adults in Rawalpindi-Islamabad, Pakistan. Tropical medicine & international health: TM & IH, 6(9), 732–738. https://doi.org/10.1046/j.1365-3156.2001.00779.x

[16] Davies, H. T., Crombie, I. K., & Tavakoli, M. (1998). When can odds ratios mislead? BMJ (Clinical research ed.), 316(7136), 989–991. https://doi.org/10.1136/bmj.316.7136.989