

Application of Logistic Regression in Health and Allied Research: Part I – Basic Concepts, Purpose, Methods, Variable Requirements, and Assumptions

Mohammad Tajul Islam¹, Russell Kabir*², Sufia Nasrin Rita³, Ali Davod Parsa², SM Anwar Sadat⁴

¹ Department of Public Health, North South University, Dhaka, Bangladesh

² School of Allied Health and Social Care, Anglia Ruskin University, UK

³ Department of Orthodontics, Sapporo Dental College, Dhaka, Bangladesh

⁴ Department of Oral and Maxillofacial Surgery, Dhaka Dental College, Dhaka, Bangladesh

Corresponding author: Dr. Russell Kabir, School of Allied Health and Social Care, Anglia Ruskin University, Chelmsford, Essex, UK. Email: russell.kabir@aru.ac.uk

Abstract

Background: Logistic regression is a widely used multivariable statistical method in health and allied research for examining associations and predicting categorical outcomes. However, its appropriate application requires an understanding of the type of outcome variable, model specification, variable coding, and underlying assumptions. This article provides an accessible introduction to logistic regression for non-statisticians, clinical professionals, and early-career researchers.

Methods: This methodological article reviews the purpose, methods, mathematical concepts, variable requirements, sample size considerations, and assumptions of logistic regression. It distinguishes binary, multinomial, and ordinal logistic regression and explains odds, logit transformation, regression coefficients, odds ratios, predicted probabilities, reference category selection, and dummy variable coding.

Results: Binary logistic regression is appropriate for dichotomous outcomes, multinomial logistic regression for nominal outcomes with more than two categories, and ordinal logistic regression for ordered categorical outcomes. Appropriate coding and selection of reference categories are essential for interpretation. Although a minimum of 10 outcome events per estimated model parameter has traditionally been used as a rule of thumb, sample size requirements depend on the characteristics and purpose of the model. Logistic regression does not require normally distributed predictors but requires attention to model assumptions and diagnostics.

Conclusions: This article provides an accessible foundation for understanding and applying logistic regression in health and allied research and prepares readers for Part II, which focuses on binary logistic regression.

Keywords: Logistic regression; Odds ratio; Multivariable analysis; Health research

1. Introduction

Logistic regression is one of the most widely used multivariable statistical methods for data analysis in health and allied research, particularly in epidemiological studies (Domínguez-Almendros et al, 2011; Boateng and Abaye, 2019). This method can be applied to analyze data from cross-sectional, case-control, cohort studies, and experimental designs (Islam, Kabir and Nisha, 2022).

One of the aims of epidemiological studies is to determine the relationship between an exposure and an outcome and provide valid and precise estimates of associations (such as odds ratios) free from confounding effects. Epidemiological studies are also conducted to predict (or model) an outcome on a set of independent variables (also called predictors or explanatory variables) (Islam, Hawalder and Kabir, 2026).

Logistic regression provides odds ratios (ORs) and can be done with two different objectives: a) to obtain adjusted measures of association (such as adjusted ORs) after controlling for confounding factors (*adjusted model*) and b) to predict the outcome based on independent variables (*prediction model*) (Chan, 2004). For both these objectives, the underlying statistical technique is similar. However, when the goal is to predict the outcome (prediction model), the model's goodness of fit needs to be assessed.

This article, Application of Logistic Regression in Health and Allied Research, will be presented in several parts and is for non-technical audiences, such as non-statisticians, clinical professionals (doctors, nurses, and others), and young researchers for easy understanding. Part I of this article describes the basic concepts, purpose, methods, variable requirements, and assumptions for performing logistic regression. The subsequent parts discuss the specific techniques of the method, including how to perform the analysis using Stata.

2. Purpose of Performing Logistic Regression

Like other multivariable analyses, the purposes of multivariable logistic regression are to a) Adjust the estimate of risk (such as OR) for a number of independent variables (e.g., confounding factors) included in the model; b) Determine the relative contribution of independent variables to a single outcome; c) Predict the probability of an outcome for a number of independent variables in the model; and d) Assess the interaction of multiple independent variables on the outcome.

3. Logistic Regression Methods

Depending on the nature of the dependent variable (also called outcome or response variable), logistic regression analysis can be applied using several methods, such as binary, multinomial, and ordinal logistic regression. Table 1 presents the type of logistic regression method to use based on the nature of the dependent variable. Logistic regression can include both categorical and continuous variables as independent variables, and they are not the deciding factors for the selection of methods.

Table 1: Logistic regression methods and nature of the dependent variable

Methods	Dependent variable
Binary logistic regression (LR):	
<i>Unconditional binary LR</i>	Dichotomous (or binary) categorical variable (e.g., disease present/absent) with unmatched study designs, e.g., unmatched case-control studies or cross-sectional studies.
<i>Conditional binary LR</i>	Dichotomous categorical variable (e.g., disease present/absent) with matched study designs, e.g., matched case-control studies.
Multinomial LR	Nominal categorical variable with more than two categories (or levels), e.g., health-seeking behavior (did not seek treatment/ received treatment from village doctors/ received treatment from pharmacists).
Ordinal LR	Ordinal categorical variable, e.g., severity of anemia (no/ mild/ moderate/ severe anemia).

3.1 Binary logistic regression:

This method is used when the dependent variable is a *dichotomous (or binary) categorical variable*, such as a disease (present/absent), vaccinated (yes/no), or the outcome of a patient (died/survived). The binary logistic regression can be applied as a) *Unconditional* binary logistic regression, where the dependent variable is a dichotomous categorical variable in an *unmatched* study design (e.g., unmatched case-control studies and cross-sectional studies), and b) *Conditional* binary logistic regression, where the dependent variable is a dichotomous categorical variable and the cases are *matched* with controls for one or more characteristics (e.g., matched case-control studies).

3.2 Multinomial logistic regression:

This method is used when the outcome variable is a nominal categorical variable with *more than two levels (or categories)*, such as health-seeking behavior (e.g., did not seek treatment/ received treatment from village doctors/ received treatment from pharmacists), type of cancer (stomach cancer/ lung cancer/ skin cancer), or others.

3.3 Ordinal logistic regression (or proportional odds regression):

This method is used when the outcome variable is an ordinal categorical variable, like severity of depression (mild/ moderate/ severe) or stage of a cancer (stage 1/ stage 2/ stage 3/ stage 4).

4. Mathematical Concept

Since this article is written for a non-technical audience, non-statisticians may skip this section without affecting their understanding of the data analysis using logistic regression.

An odds ratio is a commonly used measure of the strength of association between an exposure and an outcome. Odds (for success) are the probability of success of an event divided by the probability of failure of that event [i.e., (probability of success) ÷ (1 – probability of success)]. The odds ratio is the ratio of two odds, such as the ratio of odds of exposure among cases divided by the odds of exposure among controls.

In logistic regression analysis, odds are transformed into the natural logarithm (ln) of odds (i.e., ln odds). The natural log (ln) uses the mathematical constant “e” as its base ($\log_e$), where “e” is a constant with the value of approximately 2.718. When odds are transformed into ln odds, it is called the *logit transformation*. To reverse the transformation (i.e., converting the logit back to odds), we take the exponential of the logit value (i.e., e^x , where x is the logit value). The multivariable logistic regression model is expressed by the following equation (Domínguez-Almendros et al 2011; Boateng and Abaye, 2019; Sperandei, 2014):

$$\ln \left[\frac{P}{1-P} \right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$

Where P represents the probability of the outcome (success), β_0 is the intercept (constant), β_i is the regression coefficient for the i^{th} predictor variable ($i= 1, 2, \dots, n$), and X_i denotes the values of the predictor (or independent) variables in the model, $X_i = (X_1, X_2, \dots, X_n)$.

The values of regression coefficients (β) that we obtain from the logistic regression analysis are the ln odds of the outcome. The exponential (e^x) of the regression coefficient (e^x , where x is the regression coefficient) is the odds ratio (OR) for each categorical independent variable after controlling (adjusting) for other variables in the model. One can calculate the probability (P) of the outcome based on a number of independent variables by using the following formula:

$$P(Y) = \frac{1}{1 + e^{-Z}}$$

Where $P(Y)$ is the probability of the outcome variable (Y) and Z is $(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n)$. A detailed explanation of the model can be found in any standard biostatistics book.

5. Variables for Logistic Regression

Logistic regression is a multivariable analysis technique where the dependent variable must be a *categorical variable* (not a continuous variable). The categorical variable may be a dichotomous variable, a nominal variable with more than two categories, or an ordinal variable. The dichotomous dependent variable must be coded (for the use in Stata) as 0/1 with “0= disease absent and 1= disease present”, and the higher value is considered as the predicted outcome, while the lower value is the comparison group.

Logistic regression can include any number of independent variables (depending on the sample size), both quantitative and categorical. The categorical variables must be coded numerically (such as 0, 1, 2) to include in the analysis. It does not allow string variables (variables with alphabetic codes, such as m, f, n, y) for the analysis. If there are string variables to include, they must be recoded numerically.

For an independent categorical variable (whether dichotomous or multinomial), the analysis needs to consider a comparison group (also called the reference category or base). When the independent variable is dichotomous, Stata, by default, considers the first category as the reference group. However, one can change the reference category, as desired (see Part II).

When an independent categorical variable is multinomial (has more than two categories), we need to generate dummy variables before including it in the analysis (Sperandei, 2014). Dummy variables are the dichotomous indicator variables (coded as 0/1) representing the categories of a multinomial

categorical variable. The number of dummy variables generated for a categorical variable with more than two levels is equal to the number of levels minus one. For example, if we want to include the variable "religion" in the analysis, which has three levels (Muslim, Hindu, and Christian), we need to generate two dummy variables.

Dummy variables can be generated separately before entering them in the analysis. However, Stata can generate dummy variables automatically during the analysis and provide the ORs for each category compared to the reference category (see Part II of the publication). Like a dichotomous variable, we also need to select a reference category of the multinomial variable for the analysis, or Stata will select the first category by default.

The selection of the reference category is straightforward when there is clear prior knowledge about the exposure-outcome relationship. For example, to analyze the relationship between smoking and oral cancer, obviously non-smokers should be the choice for the reference category.

When there is no clearly established reference category, the reference group should be selected primarily on the basis of clinical or scientific relevance and ease of interpretation. A common or clinically meaningful category is often preferred because it facilitates stable and interpretable comparisons. For example, when examining the association between smoking status and oral cancer, non-smokers may be selected as the reference category because they represent the clinically meaningful unexposed group. Similarly, for categorical variables with no natural reference group, researchers should select a category that provides the most meaningful comparison for the research question. The choice of reference category changes the interpretation of the estimated odds ratios but does not change the overall fit of the logistic regression model.

6. Sample Size

Sample size is always a concern for the analysis of data. The sample size needed for a logistic regression analysis depends on the effect size (odds ratio) that the researcher is trying to demonstrate and the variability of data. It is always preferable to calculate the sample size in the planning phase of the study by using an appropriate formula and selecting relevant parameters. A commonly cited traditional rule of thumb is to have at least 10 outcome events per estimated model parameter (Bergmire-Sweat et al, 2007 and Katz, 2011). However, this should not be regarded as a universal requirement, as adequate sample size depends on factors such as outcome prevalence, number of model parameters, effect sizes, and the intended purpose of the model.

7. Assumptions

Logistic regression does not require the independent variables to be normally distributed. However, several assumptions and considerations should be addressed. Observations should be independent, problematic multicollinearity among independent variables should be absent, and continuous predictors should have an approximately linear relationship with the logit of the outcome. Influential observations and outliers should also be examined because they may substantially affect model estimates. Adequate numbers of outcome events relative to the number of estimated parameters are important for obtaining stable estimates. For ordinal logistic regression, the proportional odds assumption should additionally be assessed.

The next article of the series (Part II) will be on binary logistic regression.

References

- Bergmire-Sweat, D., Nelson, A. and FOCUS Workgroup (2007) 'Advanced data analysis: methods to control for confounding (matching and logistic regression)', *Focus on Field Epidemiology*, 4(1). Chapel Hill, NC: UNC School of Public Health, North Carolina Center for Public Health Preparedness.
- 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, pp. 190–207. doi: 10.4236/jdaip.2019.74012.
- Chan, Y.H. (2004) 'Biostatistics 202: logistic regression analysis', *Singapore Medical Journal*, 45(4), p. 149.
- Domínguez-Almendros, S., Benítez-Parejo, N. and Gonzalez-Ramirez, A.R. (2011) 'Logistic regression models', *Allergologia et Immunopathologia*, 39(5), pp. 295–305. doi: 10.1016/j.aller.2011.05.002.
- Islam, M.T., Hawlader, M.D.H. and Kabir, R. (2026) *Basic principles of epidemiology*. 2nd edn. New York: CRC Press.
- Islam, M.T., Kabir, R. and Nisha, M. (2022) *Data analysis with Stata: a comprehensive guide for data analysis and interpretation of outputs*. Dhaka, Bangladesh: Altaf Publications.
- Katz, M.H. (2011) *Multivariable analysis: a practical guide for clinicians and public health researchers*. 3rd edn. Cambridge: Cambridge University Press.
- Sperandei, S. (2014) 'Understanding logistic regression analysis', *Biochemia Medica*, 24(1), pp. 12–18. doi: 10.11613/BM.2014.003.