Cardiovasc Prev Pharmacother. 2020 Oct;2(4):142-146
https://doi.org/10.36011/cpp.2020.2.e15
eISSN 2671-700X

Cardiovascular Prevention and Pharmacotherapy

Special Article

Check for updates

Logistic Regression and Least Absolute Shrinkage and Selection Operator

Hyunyong Lee, MS[1], Hun-Sung Kim, MD, PhD[2,3]

[1]Clinical Research Coordinating Center, Catholic Medical Center, The Catholic University of Korea, Seoul, Korea
[2]Department of Medical Informatics, College of Medicine, The Catholic University of Korea, Seoul, Korea
[3]Department of Endocrinology and Metabolism, College of Medicine, The Catholic University of Korea, Seoul, Korea

OPEN ACCESS

Received: Aug 10, 2020
Accepted: Nov 11, 2020

Correspondence to
Hun-Sung Kim, MD, PhD
Department of Medical Informatics, College of Medicine, The Catholic University of Korea, 222 Banpo-daero, Seocho-gu, Seoul 06591, Korea.
E-mail: 01cadiz@hanmail.net

ORCID iDs
Hyunyong Lee
https://orcid.org/0000-0002-6132-9577
Hun-Sung Kim
https://orcid.org/0000-0002-7002-7300

Conflict of Interest
The author has no financial conflicts of interest.

Author Contributions
Conceptualization: Lee H, Kim HS; Formal analysis: Kim HS; Methodology: Lee H; Supervision: Kim HS; Writing - original draft: Lee H, Kim HS; Writing - review & editing: Lee H, Kim HS.

ABSTRACT

Logistic regression, a model that forms a binary dependent variable and one or more independent variable(s), is used especially in epidemiological studies. By understanding the logistic model and its applications, such as odds ratio (OR) and performance efficiency, the concept of logistic regression can be easily grasped. The purpose of this article is to 1) introduce logistic regression, including odds and OR, 2) present predictive efficiency, such as area under the curve, and 3) explain the caution of logistic regression analysis.

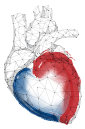
Keywords: Biostatistics; Logistic models; Models, statistical; Odds ratio; Regression analysis

INTRODUCTION

Regression analysis is a set of statistical processes used to estimate the relationship between a dependent variable (or outcome variable) and one or more independent variable(s). The most common form of regression analysis is linear regression, where the outcome variable is assumed to be continuous. In medical science, data are usually composed of two-level outcome variables (e.g., life or death, disease or no disease, pass or fail). Therefore, a linear statistical model would not be ideal, for several reasons. First, the predicted values for some covariate values are two-level outcomes (usually 0 or 1); therefore, using a standard linear regression for a bivariate outcome can produce very unsatisfactory results. Second, the assumption of constant variability of linear regression does not match the behavior of a bivariate outcome.[1)] To bridge this gap, logistic regression, which describes the relationship between a binary outcome and a set of independent variables, has been developed[2)] and explored because it reduces the potential bias resulting from differences in the groups being compared.[3)]

ODDS, ODDS RATIO (OR), AND LOGISTIC REGRESSION

The odds of an event are the ratio of probability of an event occurring (denoted by p) divided by the probability of the event not occurring (denoted by $1-p$). Since logistic regression calculates probability of an event occurring over the probability of the event not occurring, the impact of independent variables is usually explained in terms of odds. Using logistic



regression, the mean of the response variable p in terms of an explanatory variable x is modeled, relating p and x through the equation $p=\alpha+\beta x$.[4] However, this would not be an appropriate model because extreme values of x in the equation $\alpha+\beta x$ may lead to a value that does not fall between 0 and 1. To overcome this problem, we transform the odds and use the natural logarithm of the odds as a regression function,[5] as follows:

$$logit(y) = \ln(odds) = \ln(\frac{p}{1-p}) = \alpha + \beta x \quad \text{(Equation 1)},$$

where p is the probability of the outcome variable and x is the independent variable. Suppose Y=1 when the event occurs, and Y=0 when the event does not occur. α is the intercept term, β represents the regression coefficient; the change in the logarithm of the odds of the event a 1-unit change in the predictor x, in other words, a 1-unit change in x is the ratio of the odds (OR). For example, the variable smoking relative was coded as 0 (no relative who smoked) or 1 (at least one relative who smoked) and the OR of acute myocardial infarction (AMI) was 1.76. The meaning of odds in cases with at least one relative who smoked is 1.76 times higher than that in cases with non-smoking relatives. A useful way to remember the OR is "100 times the OR minus 1". In this case, the OR of 1.76 is calculated as 100×(1.76–1)=76%; this means that the passive smoking group, where the participants have a relative who smokes, is at 76% higher risk of AMI than non-passive smoking group. The OR is sometimes confused with the relative risk (RR), which is the ratio of probability of an outcome in an exposed group to the probability of the outcome in an unexposed group. If the OR, without ignoring its meaning as a ratio of odds, is interpreted as RR, then this inaccuracy entails potentially serious problems.[6] Using a two-by-two contingency table (**Table 1**), the formula of OR and RR have been proposed.

$$RR = \frac{I_1}{I_0} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}} \approx \frac{\frac{a}{b}}{\frac{c}{d}} = OR \quad \text{(Equation 2)}$$

Mathematical formulas deduce the OR always overestimates the RR. Assuming that a disease is "rare" (below 10%), then OR is similar to the RR and can be considered a good approximation to the RR.[3]

APPLICATION AS A PREDICTIVE EFFICIENCY

A logistic model can be used to distinguish the observed cases from the observed non-cases. One approach for assessing such discriminatory performance involves using a fitted model to predict accuracy. If the fitted logistic model is made and the cut-off value (Cp) determined, then the predicted cases can be classified as *the cut-off value exceeds* and *the non-cases as the cut-off value does not exceed*. From the classification table (**Table 2**), sensitivity and specificity can be calculated as A/(A+C) and B/(B+D), respectively. Higher sensitivity and specificity indicate a better fit for the model.

Table 1. Introduction to contingency (2×2) table

Exposed	Disease	
	Yes	No
Yes	A	B
No	C	D

https://doi.org/10.36011/cpp.2020.2.e15

Table 2. Classification/diagnostic table

Predicted	Observed	
	Yes	No
Yes	A	B
No	C	D

Another approach involves plotting a receiver operating characteristic (ROC) curve for the fitted model and computing the area under the curve (AUC) as a measure of discriminatory performance. The ROC curve is defined as one minus specificity and the sensitivity as x- and y-axes, respectively. Each point on the ROC curve represents a sensitivity/(1-specificity) pair corresponding to a particular decision threshold. The AUC is an integrated ROC curve and can indicate how much better the predictability of the logistic model is. A rough guide for grading AUC is as follows: excellent discrimination for 0.9–1.0, good discrimination for 0.8–0.9, fair discrimination for 0.7–0.8, poor discrimination for 0.6–0.7, and failed discrimination for 0.5–0.6.

CONSIDERATION OF SAMPLE SIZE AND SPARSE DATA

The simple linear regression model is represented as follows:

$$Yi = \beta_0 + \beta_1 X \quad \text{(Equation 3)},$$

where β_0 is the intercept and β_i is the coefficient. The goal of the linear regression is to find the best-fit line that can accurately predict the output for the continuous variable and whether independent variables are single or multiple. Before making a regression model, the assumptions of the regression model (outlined in **Table 3**) will be considered.

In logistic regression, no assumption is needed, unlike in linear regression. However, longitudinal data should be considered for analysis in the logistic form, otherwise it would lead to problems with estimation. One consideration for analyzing logistic regression is sample size. The recommended sample size for employing logistic regression can be greater than 400[2] because a small sample size overestimates the effect measure. In practice, small sample size is possible given the following "one in ten rule. One predictive variable can be studied for every ten events.[7] Another consideration is that a small sample or sparse event dataset, which estimates odds ratios, may be biased.[8] To address the problem, several methods are available to improve accuracy, including exact logistic regression,[9] Firth's penalized likelihood approach,[10] Cauchy priors,[11] data augmentation priors,[12] and Ridge and Least Absolute Shrinkage and Selection Operator method.[13][14]

Table 3. Assumptions of regression model to consider

Assumptions
1. Assumption of linearity: the relationship between mean value of outcome variable and independent variable is linear.
2. Assumptions of normality: normality means that the test is normally distributed (or bell-shaped) with 0 mean, 1 standard deviation, and a symmetric bell-shaped curve.
3. Assumption of homoscedasticity: homoscedasticity means the error term (or residuals) is the same across all values of the independent variables.

https://doi.org/10.36011/cpp.2020.2.e15

CONCLUSION

The logistic regression model is a powerful tool for epidemiologic studies, although with some substantial shortcomings in the use and reporting of logistic regression results noted. Therefore, careful model selection and interpretation is warranted. The extended logistic regression model is widely used. In this paper, I have discussed only a binary logistic regression model; nevertheless, there are extensions to the logistic regression models that allow analysis of outcomes with more than three-ordered levels (for instance, no, moderate, or severe pain). A multinomial logistic regression model can be applied with the proportional odds logistic regression method,[15] although some other models are equally applicable.[16][17] The standard logistic regression model presumes that observations are independent. In longitudinal or clustered data, observations are not independent, and assuming they are independent can lead to misleading conclusions.[18] Therefore, it is appropriate to use generalized estimating equations[19] and random effects model in these cases.[20]

REFERENCES

1. LaValley MP. Logistic regression. Circulation 2008;117:2395-9.
PUBMED | CROSSREF
2. Hosmer DW, Lemeshow S. Applied Logistic Regression. 2nd ed. New York, NY: John Wiley & Sons, Inc.; 2000.
3. Kirkwood BR, Sterne JA. Essential Medical Statistics. Oxford: Blackwell Science Ltd; 2003.
4. Park HA. An introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain. J Korean Acad Nurs 2013;43:154-64.
PUBMED | CROSSREF
5. Peng CJ, Lee KL, Ingersoll GM. An introduction to logistic regression analysis and reporting. J Educ Res 2002;96:3-14.
CROSSREF
6. Schmidt CO, Kohlmann T. When to use the odds ratio or the relative risk? Int J Public Health 2008;53:165-7.
PUBMED | CROSSREF
7. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996;49:1373-9.
PUBMED | CROSSREF
8. Mansournia MA, Geroldinger A, Greenland S, Heinze G. Separation in logistic regression: causes, consequence, and control. Am J Epidemiol 2018;187:864-70.
PUBMED | CROSSREF
9. Agresti A. Categorical Data Analysis. 3rd ed. Hoboken, NJ: John Wiley & Sons; 2013.
10. Firth D. Bias reduction of maximum likelihood estimates. Biometrika 1993;80:27-38.
CROSSREF
11. Gelman A, Jakulin A, Pittau MG, Su YS. A weakly informative default prior distribution for logistic and other regression models. Ann Appl Stat 2008;2:1360-83.
CROSSREF
12. Greenland S, Mansournia MA. Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. Stat Med 2015;34:3133-43.
PUBMED | CROSSREF
13. Le Cessie S, Van Houwelingen J. Ridge estimators in logistic regression. J R Stat Soc Ser C Appl Stat 1992;41:191-201.
14. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol 1996;58:267-88.
15. Bender R, Grouven U. Ordinal logistic regression in medical research. J R Coll Physicians Lond 1997;31:546-51.
PUBMED
16. Harrell FE Jr, Margolis PA, Gove S, Mason KE, Mulholland EK, Lehmann D, Muhe L, Gatchalian S, Eichenwald HF; WHO/ARI Young Infant Multicentre Study Group. Development of a clinical prediction

https://doi.org/10.36011/cpp.2020.2.e15

model for an ordinal outcome: the World Health Organization Multicentre Study of Clinical Signs and Etiological agents of Pneumonia, Sepsis and Meningitis in Young Infants. Stat Med 1998;17:909-44.
PUBMED | CROSSREF

17. Scott SC, Goldberg MS, Mayo NE. Statistical assessment of ordinal outcomes in comparative studies. J Clin Epidemiol 1997;50:45-55.
PUBMED | CROSSREF

18. Cannon MJ, Warner L, Taddei JA, Kleinbaum DG. What can go wrong when you assume that correlated data are independent: an illustration from the evaluation of a childhood data are independent: an illustration from the evaluation of a childhood health intervention in Brazil. Stat Med 2001;20:1461-7.
PUBMED | CROSSREF

19. Lipsitz SR, Kim K, Zhao L. Analysis of repeated categorical data using generalized estimating equations. Stat Med 1994;13:1149-63.
PUBMED | CROSSREF

20. Twisk JW. Applied Longitudinal Data Analysis for Epidemiology. Cambridge: Cambridge University Press; 2003.