J Korean Acad Nurs Vol.43 No.2, 154-164
http://dx.doi.org/10.4040/jkan.2013.43.2.154

An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain

Park, Hyeoun-Ae

College of Nursing and System Biomedical Informatics National Core Research Center, Seoul National University, Seoul, Korea

Purpose: The purpose of this article is twofold: 1) introducing logistic regression (LR), a multivariable method for modeling the relationship between multiple independent variables and a categorical dependent variable, and 2) examining use and reporting of LR in the nursing literature. **Methods:** Text books on LR and research articles employing LR as main statistical analysis were reviewed. Twenty-three articles published between 2010 and 2011 in the Journal of Korean Academy of Nursing were analyzed for proper use and reporting of LR models. **Results:** Logistic regression from basic concepts such as odds, odds ratio, logit transformation and logistic curve, assumption, fitting, reporting and interpreting to cautions were presented. Substantial shortcomings were found in both use of LR and reporting of results. For many studies, sample size was not sufficiently large to call into question the accuracy of the regression model. Additionally, only one study reported validation analysis. **Conclusion:** Nursing researchers need to pay greater attention to guidelines concerning the use and reporting of LR models.

Key words: Logit function, Maximum likelihood estimation, Odds, Odds ratio, Wald test

INTRODUCTION

Multivariable methods of statistical analysis commonly appear in general health science literature (Bagley, White, & Golomb, 2001). The terms “multivariate analysis” and “multivariable analysis” are often used interchangeably in the literature. In the strict sense, multivariate analysis refers to simultaneously predicting multiple outcomes and multivariable analysis uses multiple variables to predict a single outcome (Katz, 1999).

The multivariable methods explore a relation between two or more predictor (independent) variables and one outcome (dependent) variable. The model describing the relationship expresses the predicted value of the outcome variable as a sum of products, each product formed by multiplying the value and coefficient of the independent variable. The coefficients are obtained as the best mathematical fit for the specified model. A coefficient indicates the impact of each independent variable on the outcome variable adjusting for all other independent variables. The model serves two purposes: (1) it can predict the value of the dependent variable for new values of the independent variables, and (2) it can help describe the relative contribution of each independent variable to the dependent variable, controlling for the influences of the other independent variables. The four main multivariable methods used in health science are linear regression, logistic regression, discriminant analysis, and proportional hazard regression.

The four multivariable methods have many mathematical similarities but differ in the expression and format of the outcome variable. In linear regression, the outcome variable is a continuous quantity, such as blood pressure. In logistic regression, the outcome variable is usually a binary event, such as alive versus dead, or case versus control. In discriminant analysis, the outcome variable is a category or group to which a subject belongs. For only two categories, discriminant analysis produces results similar to logistic regression. In proportional hazards regression, the outcome variable is the duration of time to the occurrence of a binary “fail-

Address reprint requests to : Park, Hyeoun-Ae
College of Nursing, Seoul National University, 103 Daehak-ro, Jongno-gu, Seoul 110-799, Korea
Tel: +82-10-3786-3284, Fax: +82-2-765-4103, E-mail: hapark@snu.ac.kr

Received: March 19, 2013 Accepted: April 2, 2013

ure" event (for example, death) during a follow-up period of observation.

The logistic regression is the most popular multivariable method used in health science (Tetrault, Sauler, Wells, & Concato, 2008). In this article logistic regression (LR) will be presented from basic concepts to interpretation. In addition, the use of LR in nursing literature will be examined by comparing the actual use of LR with published criteria for use and reporting.

CONCEPTS RELATED TO LOGISTIC REGRESSION

Logistic regression sometimes called the logistic model or logit model, analyzes the relationship between multiple independent variables and a categorical dependent variable, and estimates the probability of occurrence of an event by fitting data to a logistic curve. There are two models of logistic regression, binary logistic regression and multinomial logistic regression. Binary logistic regression is typically used when the dependent variable is dichotomous and the independent variables are either continuous or categorical. When the dependent variable is not dichotomous and is comprised of more than two categories, a multinomial logistic regression can be employed.

As an illustrative example, consider how coronary heart disease (CHD) can be predicted by the level of serum cholesterol. The probability of CHD increases with the serum cholesterol level. However, the relationship between CHD and serum cholesterol is nonlinear and the probability of CHD changes very little at the low or high extremes of serum cholesterol. This pattern is typical because probabilities cannot lie outside the range from 0 to 1. The relationship can be described as an 'S'-shaped curve. The logistic model is popular because the logistic function, on which the logistic regression model is based, provides estimates in the range 0 to 1 and an appealing S-shaped description of the combined effect of several risk factors on the risk for an event (Kleinbaum & Klein, 2010).

1. Odds

Odds of an event are the ratio of the probability that an event will occur to the probability that it will not occur. If the probability of an event occurring is p, the probability of the event not occurring is $(1-p)$. Then the corresponding odds is a value given by

$$\text{odds of \{Event\}} = \frac{p}{1-p}$$

Since logistic regression calculates the probability of an event occurring over the probability of an event not occurring, the impact of independent variables is usually explained in terms of odds. With 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 $p = \alpha + \beta x$.

Unfortunately, this is not a good model because extreme values of x will give values of $\alpha + \beta x$ that does not fall between 0 and 1. The logistic regression solution to this problem is to transform the odds using the natural logarithm (Peng, Lee & Ingersoll, 2002). With logistic regression we model the natural log odds as a linear function of the explanatory variable:

$$\text{logit (y)} = \ln \text{(odds)} = \ln\left(\frac{p}{1-p}\right) = a + \beta\chi \qquad (1)$$

where p is the probability of interested outcome and x is the explanatory variable. The parameters of the logistic regression are α and β. This is the simple logistic model.

Taking the antilog of equation (1) on both sides, one can derive an equation for the prediction of the probability of the occurrence of interested outcome as

$$p = P\,(Y = \textit{interested outcome}/X = \chi, \textit{a specific vlaue})$$
$$= \frac{e^{a+\beta\chi}}{1 + e^{a+\beta\chi}} = \frac{1}{1 + e^{-(a+\beta\chi)}}$$

Extending the logic of the simple logistic regression to multiple predictors, one may construct a complex logistic regression as

$$\text{logit (y)} = \ln\left(\frac{p}{1-p}\right) = a + \beta_1\chi_1 + \ldots + \beta_k\chi_k$$

Therefore,

$$p = P\,(Y = \textit{interested outcome}/X_1 = \chi_1, \ldots\ X_k = \chi_k)$$
$$= \frac{e^{a+\beta_1\chi_1 + \ldots + \beta_k\chi_k}}{1 + e^{a+\beta_1\chi_1 + \ldots + \beta_k\chi_k}} = \frac{1}{1 + e^{-(a+\beta_1\chi_1 + \ldots + \beta_k\chi_k)}}$$

2. Odds ratio

The odds ratio (OR) is a comparative measure of two odds relative to different events. For two events A and B, the corresponding odds of A occurring relative to B occurring is

$$\text{odds ratio \{A vs.B\}} = \frac{\text{odds \{A\}}}{\text{odds \{B\}}} = \frac{p_A/(1-p_A)}{p_B/(1-p_B)}$$

An OR is a measure of association between an exposure and an outcome. The OR represents the odds that an outcome (e.g. disease or disorder) will occur given a particular exposure (e.g. health behavior, medical history), compared to the odds of the outcome occurring in the absence of that exposure.

When a logistic regression is calculated, the regression coefficient (b_1) is the estimated increase in the logged odds of the outcome per unit increase in the value of the independent variable. In other words, the exponential function of the regression coefficient (e^{b_1}) is the OR associated with a one-unit increase in the independent variable. The OR can also be used to determine whether a particular exposure is a risk factor for a particular outcome, and to compare the magnitude of various risk factors for that outcome. OR=1 indicates exposure does not affect odds of outcome. OR>1 indicates exposure associated with higher odds of outcome. OR<1 indicates exposure associated with lower odds of outcome. For example, the variable smoking is coded as 0 (=no smoking) and 1 (=smoking), and the odds ratio for this variable is 3.2. Then, the odds for a positive outcome in smoking cases are 3.2 times higher than in non-smoking cases.

Logistic regression is one way to generalize the OR beyond two binary variables (Peng & So, 2002). Suppose we have a binary response variable Y and a binary predictor variable *X*, and in addition we have other predictor variables $Z_1, ..., Z_k$ that may or may not be binary. If we use multiple logistic regression to regress *Y* on $X, Z_1, ..., Z_k$, then the estimated coefficient $\hat{\beta}_x$ for *X* is related to a conditional OR. Specifically, at the population level

$$e^{\hat{\beta}_x} = \frac{P\,(Y=1\,|\,X=1, Z_1, \ldots, Z_k)\,/\,P\,(Y=0\,|\,X=1, Z_1, \ldots, Z_k)}{P\,(Y=1\,|\,X=0, Z_1, \ldots, Z_k)\,/\,P\,(Y=0\,|\,X=0, Z_1, \ldots, Z_k)}$$

so $e^{\hat{\beta}_x}$ is an estimate of this conditional odds ratio. The interpretation of $e^{\hat{\beta}_x}$ is as an estimate of the OR between *Y* and *X* when the values of Z_1, ..., Z_k are held fixed.

3. The logistic curve

Logistic regression is a method for fitting a regression curve, $y = f(x)$, when *y* consists of binary coded (0, 1--failure, success) data. When the response is a binary (dichotomous) variable and *x* is numerical, logistic regression fits a logistic curve to the relationship between *x* and *y*. Logistic curve is an S-shaped or sigmoid curve, often used to model population growth (Eberhardt & Breiwick, 2012). A logistic curve starts with slow, linear growth, followed by exponential growth, which then slows again to a stable rate.

A simple logistic function is defined by the formula

$$y = \frac{e^x}{1 + e^x} = \frac{1}{1 + e^{-x}}$$

which is graphed in Figure 1.

To provide flexibility, the logistic function can be extended to the form

$$y = \frac{e^{a+\beta\chi}}{1 + e^{a+\beta\chi}} = \frac{1}{1 + e^{-(a+\beta\chi)}}$$

where α and β determine the logistic intercept and slope.

Logistic regression fits α and β, the regression coefficients. Figure 1 shows logistic function when α and β are 0 and 1, respectively. The logistic or logit function is used to transform an 'S'-shaped curve into an approximately straight line and to change the range of the proportion from 0 – 1 to -∞ – +∞ as

$$\text{logit}\,(y) = \ln\,(odds) = \ln\,\left(\frac{p}{1-p}\right) = \alpha + \beta\chi$$

where *p* is the probability of interested outcome, α is the intercept parameter, β is a regression coefficient, and χ is a predictor.

ASSUMPTIONS OF LOGISTIC REGRESSION

Logistic regression does not require many of the principle assumptions of linear regression models that are based on ordinary least squares method–particularly regarding linearity of relationship between the dependent and independent variables, normality of the error distribution, homoscedasticity of the errors, and measurement level of the independent variables. Logistic regression can handle non-linear relationships between the dependent and independent variables, because it applies a non-linear log transformation of the linear regression. The error terms

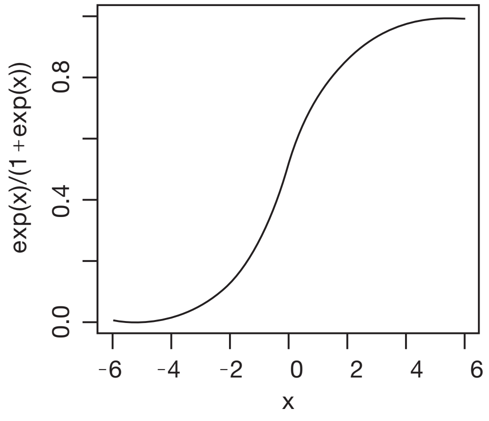

Figure 1. Graph of logistic curve where α = 0 and β = 1.

http://dx.doi.org/10.4040/jkan.2013.43.2.154

(the residuals) do not need to be multivariate normally distributed–although multivariate normality yields a more stable solution. The variance of errors can be heteroscedastic for each level of the independent variables. Logistic regression can handle not only continuous data but also discrete data as independent variables.

However, some other assumptions still apply (Bewick, Cheek, & Ball, 2005; Peng & So, 2002): First, logistic regression requires the dependent variable to be discrete mostly dichotomous. Second, since logistic regression estimates the probability of the event occurring (*P(Y=1)*), it is necessary to code the dependent variable accordingly. That is the desired outcome should be coded to be 1. Third, the model should be fitted correctly. It should not be over fitted with the meaningless variables included. Also it should not be under fitted with meaningful variable not included. Fourth, logistic regression requires each observation to be independent. Also the model should have little or no multicollinearity. That is, independent variables are not linear functions of each other. Fifth, whilst logistic regression does not require a linear relationship between the dependent and independent variables, it requires that the independent variables are linearly related to the log odds of an event. Lastly, logistic regression requires large sample sizes because maximum likelihood estimates are less powerful than ordinary least squares used for estimating the unknown parameters in a linear regression model.

STUDY DESIGN OF LOGISTIC REGRESSION

Logistic regression model corresponds to data from either a cross-sectional, prospective, or retrospective case-control study (Hsieh, Bloch & Larsen, 1998). In the cross-sectional studies a random sample is taken from a population, and outcome and explanatory variables are collected simultaneously. The fitted probabilities from a logistic regression model are then estimates of proportions of an outcome in the underlying population.

In the prospective studies, a set of subjects are selected and the explanatory variables are observed. Subjects are then followed over some standard period (e.g. a month or a year) or episode (hospital stay) to determine the response outcome. In this case, the fitted probabilities are estimates of the probability of the response outcomes occurring.

In the retrospective case-control studies, separate samples of case and control groups are first assembled and potential explanatory variables are collected later often through their recollections. In this case the fitted probabilities do not have a direct interpretation since they are determined by the relative sample sizes for case and control groups. However, odds ratios can be estimated based on logistic regression.

SAMPLE SIZE FOR LOGISTIC REGRESSION

Sample size calculation for logistic regression is a complicated problem, because there are so many factors involved in determining sample size such as statistical power, number of parameters to estimate, percentage of 1's, effect size, and standard error. There are many researchers suggesting different methods to calculate the required sample size for logistic regression (Hosmer & Lemeshow, 2000; Hsieh et al., 1998).

Hsieh et al. (1998) proposed a sample size formula for a simple logistic regression with a continuous variable with a normal distribution:

$$n=\frac{(Z_{1-\alpha/2}+Z_{1-\beta})^2}{P_1\,(1-P_1)\,\beta^{*2}}$$

where n is the required total sample size, β^* is the effect size to be tested the null hypothesis $H_0 : \beta_1 = 0$ against the alternative $H_1 : \beta_1 = \beta^*$, where $\beta^* \neq 0$, P_l is the overall event rate at the mean of X, and Z_u is the upper uth percentiles of the standard normal distribution.

When the covariate is a binary variable, the sample size formula for the total sample size required for comparing two independent event rates has the following form

$$n=\frac{\left(Z_{1-\alpha/2}\sqrt{\frac{P(1-P)}{B}}+Z_{1-\beta}\sqrt{P_1(1-P_1)+P_2(1-P_2)\frac{(1-B)}{B}}\right)^2}{(P_1-P_2)^2\,(1-B)}$$

where $P= (1-B)P_1+BP_2$ is the overall event rate; B is the proportion of the sample with $X=1$; P_1 and P_2 are the event rates at $X=0$ and $X=1$, respectively.

For multiple logistic regression, Peduzzi , Concato, Kemper, Holford, & Feinstein (1996) suggested a very simple guideline for a minimum number of cases for logistic regression study. Let p be the smallest of the proportions of negative or positive cases in the population and k the number of independent variables, then the minimum number of cases to include is:

$$N=10\,k\,/\,p$$

For example, if there are 5 explanatory variables to include in the model and the proportion of positive cases in the population is 0.25 (25%). The minimum number of cases required is

$$N=10 \text{ x } 5\,/\,0.25=200$$

Long (1997) suggested that if the resulting number is less than 100 it should be increased to 100.

FITTING THE LOGISTIC REGRESSION MODEL

Although logistic regression model, logit (y) $=\alpha+\beta\chi$ looks similar to a simple linear regression model, the underlying distribution is binomial and the parameters, α and β cannot be estimated in the same way as for simple linear regression. Instead, the parameters are usually estimated using the method of maximum likelihood of observing the sample values (Menard, 2001). Maximum likelihood will provide values of α and β which maximize the probability of obtaining the data set. It requires iterative computing with computer software.

The likelihood function is used to estimate the probability of observing the data, given the unknown parameters (α and β). A "likelihood" is a probability that the observed values of the dependent variable may be predicted from the observed values of the independent variables. The likelihood varies from 0 to 1 like any other probabilities. Practically, it is easier to work with the logarithm of the likelihood function. This function is known as the log-likelihood. Log-likelihood will be used for inference testing when comparing several models. The log likelihood varies from 0 to $-\infty$ (it is negative because the natural log of any number less than 1 is negative).

In logistic regression, we observe binary outcome and predictors, and we wish to draw inferences about the probability of an event in the population. Suppose in a population from which we are sampling, each individual has the same probability p, that an event occurs. For each individual in our sample of size n, $Y_i=1$ indicates that an event occurs for the ith subject, otherwise, $Y_i=0$. The observed data are $Y_1, \ldots, Y_n$ and $X_1, \ldots, X_n$

The joint probability of the data (the likelihood) is given by

$$L=\prod_{i=1}^{n} p\,(y/x)^{Y_i}\,(1-p\,(y/x))^{1-Y_i}=p\,(y/x)^{\sum_{i=1}^{n} Y_i}\,(1-p\,(y/x))^{n-\sum_{i=1}^{n} Y_i}$$

Natural logarithm of the likelihood is

$$l=\log(L)=\sum_{i=1}^{n} Y_i \log\,[p\,(y/x)]+\left(n-\sum_{i=1}^{n} Y_i\right)\log\,[1-p\,(y/x)]$$

In which

$$p\,(y/x)=\frac{e^{\alpha+\beta\chi}}{1+e^{\alpha+\beta\chi}}$$

Estimating the parameters α and β is done using the first derivatives of log-likelihood, and solving them for α and β. For this, iterative computing is used. An arbitrary value for the coefficients (usually 0) is first chosen. Then log-likelihood is computed and variation of coefficients values observed. Reiteration is then performed until maximization of l (equivalent to maximizing L). The results are the maximum likelihood estimates of α and β.

EVALUATION OF A LOGISTIC REGRESSION MODEL

There are several parts involved in the evaluation of the logistic regression model. First, the overall model (relationship between all of the independent variables and dependent variable) needs to be assessed. Second, the importance of each of the independent variables needs to be assessed. Third, predictive accuracy or discriminating ability of the model needs to be evaluated. Finally, the model needs to be validated.

1. Overall model evaluation

1) The likelihood ratio test

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 logistic regression model with the k independent variables (the given model) 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 via a likelihood ratio test which tests the null hypothesis

$$H_0: \beta_1=\beta_2=\ldots=\beta_k=0.$$

To do this, the deviance with just the intercept (-2 log likelihood of the null model) is compared to the deviance when the k independent variables have been added (-2 log likelihood of the given model). Likelihood of the null model is the likelihood of obtaining the observation if the independent variables had no effect on the outcome. Likelihood of the given model is the likelihood of obtaining the observations with all independent variables incorporated in the model.

The difference of these two yields a goodness of fit index G, χ^2 statistic with k degrees of freedom (Bewick, Cheek, & Ball, 2005). This is a measure of how well all of the independent variables affect the outcome or dependent variable.

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

An equivalent formula sometimes presented in the literature is

$$=-2\log\frac{\text{likelihood of the null model}}{\text{likelihood of the given model}}$$

where the ratio of the maximum likelihood is calculated before taking the natural logarithm 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 conventional 0.05, then reject H_0 with the conclusion that there is evidence that at least one of the independent variables contributes to the prediction of the outcome.

2) Chi–Square Goodness of Fit Tests

With logistic regression, instead of R^2 as the statistics for overall fit of the linear regression model, deviance between observed values from the expected values is used. In linear regression, residuals can be defined as $y_i - \hat{y}_i$, where y_i is the observed dependent variable for the *i*th subject, and $\hat{y}_i$ the corresponding prediction from the model. The same concept applies to logistic regression, where y_i is equal to either 1 or 0, and the corresponding prediction from the model is as

$$\hat{y}_i = \frac{\exp(\alpha + \beta_1 x_{i1} + \ldots + \beta_k x_{ik})}{1 + \exp(\alpha + \beta_1 x_{i1} + \ldots + \beta_k x_{ik})}$$

Chi-square test can be based on the residuals, $y_i - \hat{y}_i$ (Peng & So, 2002). A standardized residual can be defined as

$$r_i = \frac{y_i - \hat{y}_i}{\sqrt{\hat{y}_i(1-\hat{y}_i)}}$$

where the standard deviation of the residuals is $\hat{y}_i(1-\hat{y}_i)$. One can then form a χ^2 statistic as

$$\chi^2 = \sum_{i=1}^{n} r_i^2$$

This statistic follows a χ^2 distribution with $n-(k+1)$ degrees of freedom, so that *p*-values can be calculated.

3) Hosmer-Lemeshow test

The Hosmer–Lemeshow test is 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 Chi-square that compares the predicted to the observed frequencies in a 2-by-10 table.

The value of the test statistics is

$$H = \sum_{g=1}^{10} \frac{(O_g - E_g)^2}{E_g}$$

where O_g and E_g denote the observed events, and expected events for the *g*th risk decile group. The test statistic asymptotically follows a χ^2 distribution with *8* (number of groups -2) degrees of freedom. Small values (with large *p*-value closer to 1) indicate a good fit to the data, therefore, good overall model fit. Large values (with $p < .05$) indicate a poor fit to the data. Hosmer and Lemeshow do not recommend the use of this test when there is a small *n* less than 400 (Hosmer & Lemeshow, 2000).

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 logistic regression 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 is changed 1 unit while all of the other predictors are held constant, log odds of outcome is expected to change b_i units. There are a couple of different tests designed to assess the significance of an independent variable in logistic regression, the likelihood ratio test and the Wald statistic (Menard, 2001).

1) 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 MLE of the parameter. The test statistic is calculated as follows:

$$G = -2\ln\frac{L_0}{L_1} = -2(\ln L_0 - \ln L_1)$$

This statistics is compared with a χ^2 distribution with 1 degree of freedom. 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) Wald statistic

The Wald statistic can be used to assess the contribution of individual predictors or the significance of individual coefficients in a given model (Bewick et al., 2005). 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 Chi-square distribution.

$$W_j = \frac{\beta_j^2}{SE_{\beta_j}^2}$$

Each Wald statistic is compared with a Chi-square with 1 degree of freedom. Wald statistics are easy to calculate but their reliability is questionable.

3) Odds ratios with 95% CI

Odds ratio with 95% confidence interval (CI) can be used to assess the contribution of individual predictors (Katz, 1999). It is important to note however, that unlike the p value, the 95% CI does not report a measure's statistical significance. It is used as a proxy for the presence of statistical significance if it does not overlap the null value (e.g. OR=1).

The 95% CI is used to estimate the precision of the OR. A large CI indicates a low level of precision of the OR, whereas a small CI indicates a higher precision of the OR. An approximate confidence interval for the population log odds ratio is

$$95\%\ \text{CI for the In (OR)} = \text{In (OR)} \pm 1.96 \times \{\text{SE In (OR)}\}$$

where ln(OR) is the sample log odds ratio, and SE ln(OR) is the standard error of the log odds ratio(Morris & Gardner, 1988). Taking the antilog, we get the 95% confidence interval for the odds ratio:

$$95\%\ \text{CI for OR} = e^{\text{In (OR)} \pm 1.96 \times \{\text{SE In (OR)}\}}$$

3. Predictive Accuracy and Discrimination

1) Classification table

The classification table is a method to evaluate the predictive accuracy of the logistic regression model (Peng & So, 2002). In this table the observed values for the dependent outcome and the predicted values (at a user defined cut-off 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 observed outcomes, and dichotomous predicted outcomes.

The table has following form.

Table 1. Sample Classification Table

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

a, *b*, *c*, and *d* are 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 sensitivity $= a/(a+b)$ and specificity $= d/(c+d)$. Higher sensitivity and specificity indicate a better fit of the model.

2) Discrimination with ROC curves

Extending the above two-by-two table idea, 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 an ROC (Receiver Operating Characteristic) curve. The area under this curve (AUC) provides an overall measure of fit of the model (Bewick, Cheek, & Ball, 2004). 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).

4. Validation of the logistic regression

Logistic regression models are frequently used to predict a dependent variable from a set of independent variables. An important question is whether results of the logistic regression 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 cab 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 (Giancristofaro & Salmaso, 2003). 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 (Harrell, Lee, & Mark, 1996).

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.

REPORTING AND INTERPRETING LOGISTIC REGRESSION RESULTS

In presenting the logistic regression results, following four types of information should be included: a) an overall evaluation of the logistic model; b) statistical tests of individual predictors; c) goodness-of-fit statistics; and d) an assessment of the predicted probabilities. Table 2, 3, and 4 are examples to illustrate the presentation of these four types of information.

Table 2 presents the statistical significance of individual regression coefficients (βs) tested using the Wald Chi-square statistic. According to Table 2, Cholesterol was a significant predictor for event ($p<.05$). The slope coefficient 1.48 represents the change in the log odds for a one unit increase in cholesterol. The test of the intercept ($p<.05$) was significant suggesting that the intercept should be included in the model. Odd ratio 4.04 indicates that the odds for a event increase 4.04 times when the value of the cholesterol is increased by 1 unit.

Table 3 presents three inferential statistical tests for overall model evaluation: the likelihood ratio, score, and Wald tests. All three tests yield similar conclusions for the given data set, namely that given logistic model with independent variables was more effective than the null model. Table 3 also presents an inferential goodness-of-fit test, the Hosmer-Lemeshow test. Hosmer-Lemeshow test statistics 7.76 was insignificant ($p>.05$), suggesting that the model was fit to the data well.

Table 4 presents the degree to which predicted probabilities agree with actual outcomes in a classification table. The overall correct prediction, 66.84% 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.

Table 3. Example Output from Logistic Regression: Overall Model Evaluation and Goodness-of-Fit Statistics

Test	Categories	χ^2	df	p
Overall model evaluation	Likelihood ratio test	12.02	2	.002
	Score test	11.52	2	.003
	Wald test	11.06	2	.004
Goodness-of-fit test	Hosmer & Lemeshow	7.76	8	.457

Table 4. Example Output from Logistic Regression: A Classification Table

Observed	Predicted		% Correct
	Yes	No	
Yes	3	57	5.00
No	6	124	95.48
Overall % correct			66.84

Sensitivity=3/(3+57)=5.00%; Specificity=124/(6+124)=95.48%;
False positive=6/(6+124)=4.62%; False negative=57/(3+57)=95.00%.

CAUTIONS AND CONSIDERATIONS

In logistic regression no assumptions are made about the distributions of the independent variables. However, the independent variables should not be highly correlated with one another because this could cause problems with estimation.

Studies with small to moderate samples size employing logistic regression overestimate the effect measure. Thus, large sample sizes are required for logistic regression to provide sufficient numbers in both categories of the outcome variable. Also, the more independent variables were included, the larger the sample size is 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 (Hosmer & Lemeshow, 2000).

Table 2. Example Output from Logistic Regression: Statistical Tests of Individual Predictors

Predictor	ß	SE (ß)	Wald's χ^2	df	p	e^{β} (OR)	95% CI for OR	
							Lower	Upper
Cholesterol	1.48	0.45	10.98	1	<.001	4.04	1.83	10.58
Constant	-12.78	1.98	44.82	1	<.001			

CI=Confidence interval; df=Degrees of freedom; OR=Odds ratio; SE=Standard error.

ANALYSIS OF USE OF LOGISTIC REGRESSION IN NURSING

Original research articles containing explicit mention of LR were searched in the Journal of Korean Academy of Nursing published between 2010 and 2012. In total 23 articles were identified. There are 11 articles performed logistic regression with secondary data from large surveys, six articles from the cross-sectional survey, four articles from the prospective studies and two from the retrospective case-control studies. To evaluate the research reports, a list of criteria from Bagley et al.'s (2001) study was used. None of the studies reported any interaction and only one study reported a validation of the model with threefold cross validation technique, so those columns have been omitted.

Table 5 shows the adherence to the guidelines for using and reporting LR for each of the 23 articles. Although it is important the logistic regression model includes all relevant variables, it is also important the model not start with more variables than are justified for the given number of observations (Peduzzi et al., 1996). Peduzzi et al. suggested that the number of the less common of the two possible outcomes divided by the number of independent variables should be at least 10 as a useful rule of thumb. For analysis of 23 articles, the number of events-per-variable ranged from 0.7 to 16409. Sixteen of 23 of the analyses had an events-per-variable ratio above 10. Sample size calculation was mentioned in 11 articles. Five out of 11 used G*Power to calculate the sample size.

For best results from the use of logistic regression, any given change in a continuous independent variable should have the same effect on the log-odds of a positive outcome. However, in all the studies with continuous or ordinal independent variables, none tested the conformity with the linear gradient for continuous variables. As noted before, no interactions were reported in any of 23 studies. However, it is unclear whether interactions were considered but not found to be significant or whether interactions were not considered. If two highly correlated variables are included in the model, then their estimated contributions may be imprecise. Thus collinearity should be tested. However, only 5 of the 23 studies mentioned collinearity with the details of testing. Again it is not clear whether collinearities was considered but not found to be significant or whether collinearities were not considered. As noted before, only one study reported the model validation.

All of the studies reported measures of statistical significance, typically confidence intervals and P-values for each of the independent variables. Sometimes, these statistics were reported only for those variables found to be significant. The statistical significance for the entire model was reported for 13 of the 23 analyses. Goodness of fit measures describing how well the entire model matches the observed values were reported in 9 articles.

Nearly all the articles explained how variables were selected for inclusion in the model. Most of the articles selected variables based on the literature review. However, 15 of 23 reported performing the statistical tests (such as bivariate analyses) before considering the variable for the models. None of the articles provided complete details on the coding for all the variables. However, it was possible to infer the coding from the textual description in all cases. Eight studies explicitly stated the fitting procedure. In one study the variables included in the model was determined in hierarchically grouped subsets.

CONCLUSION

Logistic regression is a type of multivariable analyses used with increasing frequency in the nursing domain because of its ability to model the relationship between dichotomous dependent variable and one or more independent variables. In this paper, logistic regression from basic concepts to interpretation of analysis results was introduced. In addition, twenty-three articles published between 2010 and 2011 in the Journal of Korean Academy of Nursing were examined to see if logistic regressions were properly used and reported.

It was found that there were substantial shortcomings in the use and reporting of logistical regression results. Most notably, one-thirds of the studies had not enough sample size with events-per-variables ratios below 10, suggesting that the regression results may not be reliable. Only one study reported internal validation analysis. Validation is a crucial step to test the regression model captured essential relationships in the domain of study. Another problem is that only five studies reported tests for collinearity. If there is high correlation between two independent variables, variance of the coefficients of these variables will be increased, with a consequent loss of statistical significance. In addition, none reported tests for any conformity with the linear gradient for continuous independent variables, and tests for interactions. These problems represent failures to perform important aspect of the analysis or failures to report details of the analyses. It is hard to distinguish these two possibilities by reviewing published articles.

Thus, proper use and report of this powerful and sophisticated modeling technique requires considerable care both in the specification of the model and in the estimation and interpretation of the model's coefficients.

Table 5. Adherence to Guidelines for Using and Reporting Logistic Regression

Author	Event per variable	Collinearity	Statistical test	Goodness of fit	Variable selection	Coding of variables	Fitting procedure	Number of observations	Sample size calculation	Type of study
Jung & Lee	262/7 = 29.1	Mentioned, tested by VIF	Variables: OR, CI, p Model: Hosmer-Lemeshow	Hosmer-Lemeshow	Yes, $p<.05$	Yes	Stepwise	1,458	No	Secondary analysis of 4th Korean National Health and Nutrition Examination Survey
Cho & Chung	205/17 = 12.1	NR	Variables: B, CI, p Model: Wald Chi-square, AIC, R^2, Hosmer-Lemeshow Goodness of fit test	Hosmer-Lemeshow	Yes, $p<.05$	No	Forward stepwise	3,348	No	Retrospective cohort study
Lee, Jung, Yun, Um, & Jee	142/4 = 35.5	NR	Variables: B, SE, Wald, OR, CI, p Model: Hosmer-Lemeshow Goodness of fit test, Chi-Square, ROC curve, Correct prediction, Nagelkerke R^2	Hosmer-Lemeshow	Yes, $p<.05$	No	Forward stepwise	401	Mentioned (Tabachnick & Fidell)	Secondary Analysis of Survey
Kim & Kim	114863/7 = 16409	NR	Variables: B, SE, OR, CI, p Model: NR	NR	Informal	No	Generalized estimating equation logistic regression	254,414	No	National Health Insurance Data
Yi, Yi, & Jung	2700/20 = 135	NR	Variables: OR, CI Model: NR	NR	Yes, $p<.001$	Yes	NR	17783	No	Korea Youth Health Risk Behavior Web-based Survey
Yeoum & Lee	181/4 = 43	Mentioned, tested by Collinearity statistics Tolereance, VIF	Variables: OR, CI, p Model: -2LL, Chi-Square, AUC	NR	Yes, $p<.05$	Yes	NR	732	No	Survey
Cha, Cho, & Yoo	26/3 = 8.7	NR	Variables: OR, CI, p Model: NR	NR	Yes, $p<.05$		Stepwise	103	Mentioned (Korinek)	Retrospective case-control study
Choi & Lee	123/20 = 6.1	NR	Variables: OR, CI, p Model: Hosmer-Lemeshow Goodness of fit test, Nagelkerke R^2, Correct prediction	Hosmer-Lemeshow	Informal	Yes	NR	246	Mentioned (Biderman, Fried & Galinsky)	Case-control study Interview and survey with secondary data from public health center records
Choi, Jung, Kim, & Park	318/15 = 21.2	Mentioned, tested by VIF	Variables: OR, CI, p Model: Hosmer- Lemeshow Goodness of fit test	Hosmer-Lemeshow	Informal		Hierarchical, in 3 blocks	9094	NR	Secondary analysis of Korean Working Condition Survey
Park, & Jung	60/4 = 15	Mentioned, tested by Collinearity statistics Tolereance, VIF	Variables: OR, CI Model: NR		Yes, $p<.05$		NR	804	NR	Survey
Kim & Park	756/2 = 378	NR	Variables: OR, CI, p Model: NR		Informal		NR	6,521	NR	Secondary Analysis of Health Interview and Health Behavior Surveys
Cho & Yoo Yang &	88/3 = 29.3	NR	Variables: OR, CI, p Model: -2LL, AIC, Score		Yes, $p<.05$		NR	276	G*Power	Survey
Kim	103/8 = 12.9	NR	Variables: OR, CI, p Model: NR		Yes, $p<.05$		NR	324	G*Power	Survey
Choi, Park, & Lee	249/9 = 27.7	NR	Variables: B, SE, Wald, OR, CI, p Model: NR		Yes, $p<.05$		NR	3,024	Mentioned	Secondary Analysis of Survey
Yeon et al.	451/11 = 41	NR	Variables: OR, CI Model: NR		Yes, $p<.05$		NR	2,639	NR	Secondary Analysis of Community Health Survey

Table 5. Adherence to Guidelines for Using and Reporting Logistic Regression (Continued)

Author	Event per variable	Collinearity	Statistical test	Goodness of fit	Variable selection	Coding of variables	Fitting procedure	Number of observations	Sample size calculation	Type of study
Kim, Cho, Cho, & Kim	17/2 = 8.5	NR	Variables: b, OR, CI, p Model: Hosmer-Lemeshow Goodness of fit test, Correct classification, Nagelkerke R^2		Yes, $p < .05$		NR	175	NR	Prospective descriptive study
Sung et al.	5/5 = 1	NR	Variables: b, OR, CI, p Model: -2LL, Chi-Square, Hosmer-Lemeshow Goodness of fit test, Correct prediction		Informal		NR	145	Mentioned	Randomized Prospective study
Kim & Jung	11/6 = 1.8	NR	Variables: RR, CI, p Model: Hosmer-Lemeshow Goodness of fit test, Correct prediction		Yes, $p < .05$		NR	197	G*Power	Prospective cohort study
Hyun & Cho	122/7 = 17.4	NR	Variables: OR, CI, p Model: -2LL, Max-rescaled R^2		Yes, $p < .05$		NR	508	G*Power	Survey
Kim et al.	1053/12 = 87.8	NR	Variables: OR, CI, p Model: NR		Informal		Generalized estimating equation Logistic regression	111,491	NR	National Health Insurance Data
Jang & Park	13/19 = 0.7	Mentioned, tested by VIF	Variables: OR, CI, p Model: Hosmer-Lemeshow Goodness of fit test. Nagelkerke R^2		Informal		Stepwise	416	NR	Survey
Oak & Lee	8494/12 = 797.8	NR	Variables: b, SE, Wald, OR, CI, p Model: -2LL, Chi-Square, Correct prediction, Negelkerke R^2				NR	37,570	NR	Secondary analysis
Cho, Lee, Mark, & Lee	192/7 = 2.7	NR	Variables: OR, CI, p Model: NR		Informal		NR	507	NR	Public Survey Data

NR = Not reported; VIF = Variance inflation factor; AUC = Area under curve; ROC = Receiver operationg characteristic; AIC = Akaike's information criterion; LL = Log-likelihood CI = Confidence interval; df = Degrees of freedom; OR = Odds ratio; SE = Standard error.

REFERENCES

Bagley, S. C., White, H., & 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*(10), 979-985.

Bewick, V., Cheek, L., & Ball, J. (2004). Statistics review 13: Receiver operating characteristic curves. *Critical Care (London, England), 8*(6), 508-512. http://dx.doi.org/10.1186/cc3000

Bewick, V., Cheek, L., & Ball, J. (2005). Statistics review 14: Logistic regression. *Critical Care (London, England), 9*(1), 112-118. http://dx.doi.org/10.1186/cc3045

Eberhardt, L. L., & Breiwick, J. M. (2012). Models for population growth curves. *ISRN Ecology, 2012, 1-7*. http://dx.doi.org/doi:10.5402/2012/ 815016

Giancristofaro, R. A., & Salmaso, L. (2003). Model performance analysis and model validation in logistic regression. *Statistica, 63*(2), 375-396.

Harrell, F, E., LEE, K. L., & 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.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York, NY: John Wiley & Sons Inc.

Hsieh, F. Y., Bloch, D. A., & Larsen, M. D. (1998). A simple method of sample size calculation for linear and logistic regression. *Statistics in Medicine, 17*(14), 1623-1634.

Katz, M. H. (1999). *Multivariable analysis: A practical guide for clinicians*. Cambridge: Cambridge University Press.

Kleinbaum, D. G., & Klein, M. (2010). *Logistic regression(statistics for biology and health)* (3rd ed.). New York, NY: Springer-Verlag New York Inc.

Long, J. S. (1997). *Regression models for categorical and limited dependent vriables*. Thousand Oaks, CA: Sage Publications.

Menard, S. W. (2001). *Applied logistic regression analysis (quantitative applications in the social sciences)* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Morris, J. A., & Gardner, M. J. (1988). Calculating confidence intervals for relative risks (odds ratios) and standardised ratios and rates. *British Medical Journal (Clinical Research Ed.), 296*(6632), 1313-1316.

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.

Peng, C. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research, 96*(1), 3-14.

Peng, C. J., & So, T. H. (2002). Logistic regression analysis and reporting: A primer. *Understanding Statistics, 1*(1), 31-70.

Tetrault, J. M., Sauler, M., Wells, C. K., & Concato, J. (2008). Reporting of multivariable methods in the medical literature. *Journal of Investigative Medicine, 56*(7), 954-957. http://dx.doi.org/10.231/JIM.0b013e31818914ff