Logistic Regression: Concept and Application

*Ömay ÇOKLUK**

Abstract

The main focus of logistic regression analysis is classification of individuals in different groups. The aim of the present study is to explain basic concepts and processes of binary logistic regression analysis intended to determine the combination of independent variables which best explain the membership in certain groups called dichotomous dependent variable. Independent variables in this study are as follows: total score in the Scientific Thinking Skills Scale, total score in the Epistemological Belief Scale and total score in the Fatalism Scale. The dependent (predicted, criteria) variable is the level of critical thinking. While the three independent variables are constants, the dependent variable is defined as a categorical variable to include high and low critical thinking levels. The study group consists of 200 students from Ankara University, Faculty of Educational Sciences, Department of Guidance and Psychological Counseling during the 2006-2007 academic years. In the study, the following were used: the Epistemological Belief Scale adapted to Turkish by Deryakulu and Büyüköztürk (2002), the California Critical Thinking Disposition Scale adapted to Turkish by Kökdemir (2003), the Scientific Thinking Skills Scale" developed by Gündoğdu (2002) and the Fatalism Scale" developed by Şekercioğlu (2008). The study presents information about how to relatively assess each application step of logistic regression analysis. When the coefficient predictions of aimed model variables at the end of the study are considered, it is observed that a one-unit increase in the predictive variable of scientific thinking skills leads to an increase of 14.4% in high critical thinking odds. It is also seen that a one-unit increase in the predictive variable of epistemological belief leads to an increase of 4.9% in high critical thinking odds.

Key Words

Logistic Regression Analysis, Critical Thinking, Scientific Thinking Skills, Epistemological Belief, Fatalism.

**Correspondence:* Assist. Prof. Ömay Çokluk, Ankara University Faculty of Educational Sciences, 06590 Ankara/Turkey.
E-mail: cokluk@education.ankara.edu.tr

Kuram ve Uygulamada Eğitim Bilimleri / Educational Sciences: Theory & Practice
10 (3) • Summer 2010 • 1397-1407

Researchers almost from every discipline would like to define working principles of systems based on gathered data and they turn towards abstract structures to explain such systems. It is possible to define these abstractions as the term "model". A model is shaping information or concerns in a case, depending on certain principles (Tatlıdil, 1996). Intended use of logistic regression analysis is the same as those of other model structuring techniques in statistics. In such analysis, the main goal to form an acceptable model which could define the correlation between dependent (predicted) and independent (predictive) variables in best fit with the least variable (Atasoy, 2001).

The use of logistic regression model dates back to 1845. It first appeared during the mathematical studies for the population growth at that time (Gürcan, 1998). The term logistic regression analysis comes from logit transformation, which is applied to the dependent variable. This case, at the same time, causes certain differences both in estimation and interpretation (Hair, Black, Babin, Anderson and Tahtam, 2006). Logistic regression analysis is also called "Binary Logistic Regression Analysis", "Multinominal Logistic Regression Analysis" and "Ordinal Logistic Regression Analysis", depending on the scale type where the dependent variable is measured and the number of categories of the dependent variable. Logistic regression is divided into two: "univariate logistic regression" and "multivariate logistic regression" (Stephenson, 2008).

The main focus of logistic regression analysis is classification of individuals in different groups. The aim of the present study is to explain basic concepts and applications of binary logistic regression analysis intended to determine the combination of independent variables which best explain the membership in certain groups called dichotomous dependent variable.

Data related to confronted and researched cases in applied social sciences are mostly categorical (nominal) data with discrete value or data obtained by an ordinal scale. For instance, a man either works or unemployed; he is either a member of a group or not; the party in power is either from the right wing or the left wing; a student is either a graduate or not (Arabacı, 2002; Kılıç, 2000; Mertler & Vannatta, 2005). In educational research, many problems relate with prediction of categorical results. For example, a student is either academically successful or not; he is either a slow learner or not; a teenager either has a tendency towards risky behavior or not (Peng, Lee & Ingersoll, 2002). For instance, in a study by Kayri and Okut (2008), the individuals in special ability exam of a university for the Department of Physical Education and Sports Teaching were modeled using mixed logistic regression analysis as those achieved or not (or those who get into the department or not) according to gender. Multivariable statistical analysis of categorical data is of importance for almost every discipline. Logistic regression analysis, with its advantage of being more eligible than other analysis and with its regression logic, has an important place in categorical data analysis (Kılıç, 2000).

Over the past years, logistic regression has been commonly used (Cook, 2008; Garson, 2008; Mertler & Vannatta, 2005; Seven, 1997; Tabachnick & Fidell, 1996). Logistic regression is similar to both multiple regression and discriminant analysis. Simple and multiple linear regression analysis are used to analyze the mathematical correlation between dependent (predicted or criteria) variables and independent (predictive or explanatory) variable or variables. In data sets where these methods could be used, it is essential that the dependent variable display a normal distribution, independent variables consist of a variable or variables with normal distributions and error terms variance display a normal distribution. Under no such circumstances, simple or multiple linear regression analysis cannot be used (Kılıç, 2000).

Although regression equation varies, basic terms of multiple linear regression analysis are the same as the terms of logistic regression (George and Mallery, 2000). A standard regression equation consists of true values of a few independent variables and weights produced by the model to predict the value of the dependent variable. On the other hand, in logistic regression, the estimated value ranges from 0 to 1. More clearly, logistic regression reveals the possibility of particular consequences for each subject (for example; “passed” or “failed”). The analysis produces a regression equation which enables us to make an accurate estimation for the possibility that an individual falls into one of the categories (“passed”) or (“failed”) (Tate, 1992). As a result, the main difference between the two techniques is that the value of the dependent variable is estimated in multiple linear regression analysis, while the possibility of occurrence of one of the values which the dependent variable might have is estimated in logistic regression analysis (Bircan, 2004).

Logistic regression is an analysis which enables us to estimate categorical results like group membership with the help of a group of variables.

Independent variables could be constant or categorical. Also, discriminant analysis aims at explaining and predicting group membership using a group of independent variables. When all these definitions are taken into account, it is clear that discriminant analysis and logistic regression analysis enable us to answer the same questions. In addition, logistic regression analysis and discriminant analysis are similar in that they both have a categorical dependent variable (Büyüköztürk & Çokluk-Bökeoğlu, 2008). However, logistic regression analysis differs from discriminant analysis and multiple regression analysis at certain points (Tabachnick & Fidell, 1996). Logistic regression analysis, unlike discriminant analysis and multiple regression analysis, does not require assumptions to meet concerning the distribution of independent variables. In other words, assumptions such as normal distribution of independent variables, linearity and equality of variance-covariance matrix do not have to be met. Therefore, it might be suggested that logistic regression analysis is much more flexible than the other two techniques. Also, it is sensible to state that it is easier to interpret the mathematical model obtained as a result of analysis by logistic regression analysis (Akkuş and Çelik, 2004; Grimm and Yarnold, 1995; Kalaycı, 2005; Leech, Barrett and Morgan, 2005; Poulsen and French, 2008; Tabachnick and Fidell, 1996; Tatlıdil, 1996). However, since maximum likelihood method, unlike least squares method is used in logistic regression analysis to obtain coefficients, it is important not to study with a low number of observations, because in estimations by low numbers of observations, reliability of the model decreases.

Although logistic regression could be used to predict a dependent variable with two or more categories, as it is mentioned above, the present study concerns only a dichotomous dependent variable, in other words, where the dependent variable is dichotomous. Therefore, it is aimed to introduce application processes of binary logistic regression analysis using real data. It is thought that introducing and illustrating logistic regression analysis will be useful for further use in applied social sciences such as education and psychology.

The study is correlational research. Correlational research includes studies where the correlation between two or more variables is examined, without any variable manipulation (Büyüköztürk, Kılıç Çakmak, Akgün, Karadeniz, & Demirel, 2008). The independent variables in this study are as follows: total score in the Scientific Thinking Skills Scale, total score in the Epistemological Belief Scale and total score in the Fatalism Scale. The dependent (predicted, criteria) variable is the level of critical thinking. While the three independent variables are constants, the dependent variable is defined as a categorical variable to include high and low critical thinking levels.

The study group consists of 200 students from Ankara University, Faculty of Educational Sciences, Department of Guidance and Psychological Counseling during the 2006-2007 academic years. In the study, the following were used: the Epistemological Belief Scale, developed by Schommer (1990) and adapted to Turkish by Deryakulu and Büyüköztürk (2002, 2005), the California Critical Thinking Disposition Scale adapted to Turkish by Kökdemir (2003), the Scientific Thinking Skills Scale developed by Gündoğdu (2002) and the Fatalism Scale developed by Şekercioğlu (2008).

In the study, the students were divided into two groups according to their scores from the California Critical Thinking Disposition Scale as high and low critical thinking level groups. The scores for this classification were almost the normal distributed arithmetic mean was used as cut-off point. The mean of the group according to their scores from the scale was approximately 220 and the students with the same or a lower score were assigned to "low", those with a higher score were assigned to "high" critical thinking level category. Thus, the dichotomous dependent variable for the analysis was obtained.

In binary logistic regression analysis, it is essential that the categories of dependent variable should be encoded as 0 and 1 in the analysis. In our case, 0 shows low critical thinking level and 1 means high critical thinking level. Accordingly, the coefficients obtained reflect the effects of the variables of scientific thinking, epistemological belief and fatalism on the possibility of having a high critical thinking level, since the category encoded as 1 shows "high" critical thinking level. For the study could be defined as exploratory by nature, a stepwise method was preferred. Although forward methods have some disadvantages, Todman and Dugard (2007) state that they present more reliable results when the number of parameters is low and emphasize similar results are obtained, although different selection criteria are used. Therefore, in this study, logistic regression analysis is carried out using "Forward Likelihood Ratio-Forward:LR".

In the study, before the application of logistic regression analysis, eigenvalues, condition indexes, variance ratios and standard errors in predictive variables, tolerance, VIF values and bivariate correlations between variables were examined to probe the problem of multicollinearity between variables. When all the findings obtained are considered, it is clear that there is no problem of multicollinearity between the predictive variables (Büyüköztürk, 2009; Menard, 1995; Myers, 1990).

The study presents information about how to relatively assess each application step of logistic regression analysis. In SPSS, when logistic regression analysis is used, first -2LL value of the baseline model or constant only model is presented. At further stages, with the predictive variable included in the model, there are changes in –2LL value. In other words, as the predictive variables are added to the model, there are developments in the model and model-data fit is improved.

The results of logistic regression analysis in SPSS first present a classification table. This case, at the same time, can be interpreted as the fact that in a constant only model, membership of only one group can be accurately classified (Kalaycı, 2005). In this study, as a result of the analysis, it was determined that all the students were classified in low critical thinking category and the percentage of accurate classification was 50.50 %. Following the classification table, the constant term which constituted the baseline model concerning Blok 0, standard error of the constant term, Wald statistics to test significance of variables (Agresti, 1996), degree of freedom of Wald statistics and significance level, and $Exp(B)$ or exponentiated logistic coefficients were assessed.

When residual chi-square or the initial chi-square value is considered, it is clear that the value is significant. This case shows the coefficients of predictive variables which are not included in the model significantly differ from 0, in other words, adding one or more variables to the model will increase the predictive power of the model. It is also understood that all the predictive variables in the study considerably contribute to the model and further selection of variables is decided.

Score statistics are Roa's efficient score statistics for each variable and decide whether each variable significantly contributes to the model or not (Field, 2005). Here, significant score statistics of all variables means all predictive variables will potentially contribute to the model.

Then, the model was assessed with p value of chi-square statistics. A significant value means a current correlation between the combination of predicted variables and the predictive variables. Significant model chi-square statistics means refusing the null hypothesis (H_0) stating "There is no difference between the baseline model with the constant only and the final model obtained when predictive variables are analyzed or between the former and the aimed model": It also means supporting the correlation between predictive variables and the predicted variable.

Next, changes in -2LL value of the baseline model were examined. This case shows improvements in the model, when one or more variables are added to the model. In our case, in the baseline or constant-only model, it was seen that the changes in the model fit were significant when the variable of epistemological belief with the highest score statistics was added at the first stage and the variable of scientific thinking at the second stage.

Cox & Snell R^2 and Nagelkerke R^2 values represent two different ways of prediction for the explained variance in the dependent variable by the model and are interpreted in a similar way to R^2 in multiple regression (Field, 2005). When Cox & Snell R^2 values are examined, it is observed that when the predictive variable of epistemological belief at the first stage is included in the analysis, it explains 8.1% of the variance of the predicted variable of critical thinking, and when the predictive variable of scientific thinking skills at the second stage is included in the analysis, it explains 10.1% of the variance of the predicted variable of critical thinking, together with the two predictive variables. Nagelkerke R^2 values are 10.8% for the first stage, and 13.5% for the second stage.

Hosmer and Lemeshow chi-squared goodness of fit test assess logistic regression model fit as a whole. It is particularly more powerful than the traditional chi-square test when predictive variables are constant variables or when there is a small sample. Here, the result of Hosmer and Lemeshow test is significant neither when only the predictive variable of epistemological belief at the first stage is included in the analysis, nor the predictive variable of scientific thinking skills is included in the analysis. This case shows a bivariate model is of acceptable fit, that is; the model-data fit is enough.

When classification results after logistic regression model were examined, it was determined that at first all the students were classified in low critical thinking category and the percentage of accurate classification was 50.50%. As a result of logistic regression model, the ratio in low critical thinking is 54.50% and 63.60% in high critical thinking group at the first stage, for the classification according to the predictive variable of epistemological belief. The total percentage of accurate classification at the first stage is 59.00%. When this case is compared to the original classification percentage, it might be suggested that the accurate classification percentage is increased when the predictive variable of epistemological belief is included in the analysis. The total percentage of accurate classification of the aimed model is increased to 63.50% at the second stage, when the variable of scientific thinking skills is included in the analysis, together with epistemological belief. This finding can be interpreted as an indicator of model-data fit.

When the predictions of the variables of the aimed model variables were considered, it was seen that a one-unit increase in the predictive variable of scientific thinking skills led to an increase of 14.4% in high critical thinking odds. A one-unit increase in the predictive variable of epistemological belief led to an increase of 4.9% in high critical thinking odds. When variable/variables were eliminated and whether that case affected the model was examined, it was determined that eliminating the variables of epistemological belief and scientific thinking skills was not a good idea because both significantly decreased the predictive power of the model.

As a result, logistic regression analysis is an alternative to discriminant analysis and cross level tables when assumptions such as normality and homogeneity of variance are not met, and it is the alternative of linear regression analysis when the dependent variable is a categorical variable including two (such as 0 and 1) or more levels, since the assumption of normality is invalid (Tatlıdil, 1996).

Intended use of logistic regression analysis is the same as those of other model structuring techniques in statistics. In such analysis, the main goal to form an acceptable model which could define the correlation between dependent (predicted) and independent (predictive) variables in best fit with the least variable (Atasoy, 2001). In other words, when the predicted variable (Y) is dichotomous or classified, the most eligible and economic model between the predicted variable and the predictive one or ones occurs (Seven, 1997). In the model, predictive variables generally referred as X are used to estimate the predicted variable Y.

Measuring predictive variables in order to estimate the predicted variable over a probation time and using the obtained regression equation to estimate the predicted variable in the future are common practices. In both cases, when the nature of correlation between variables X and Y is not fully known, it is essential to use data from the selected variables and define the correlation which could show the nature of the correlation to be used for prediction (Miller, 1990).

The higher the number of variables to be included in a regression equation designed to explain variance of the predicted variable, the less error rate in the equation will be. However, problems and possible errors caused by workload for observation of each predictive variable and time limit for such observations would entail a decrease in the number of predictive variables. Therefore, accuracy of predictions should be high as much as possible and it is suggested to work with a reasonable number of predictive variables to lower systematic errors caused by data gathering using too many variables (Önder & Cebeci, 2001).

References/Kaynakça

Agresti, A. (1996). *An introduction to categorical data analysis.* New York: John Wiley and Sons.

Akkuş, Z., ve Çelik, M. Y. (2004, Eylül-Ekim). *Lojistik regresyon ve diskriminant analizi yöntemlerinde önemli ölçütler.* VII. Ulusal Biyoistatistik Kongresinde sunulan bildiri. Mersin Üniversitesi, Tıp Fakültesi, Biyoistatistik Anabilim Dalı, Mersin.

Arabacı, Ö. (2002). *Lojistik regresyon analizi ve bir uygulama denemesi.* Yayınlanmamış yüksek lisans tezi, Uludağ Üniversitesi, Sosyal Bilimler Enstitüsü, Bursa.

Atasoy, D. (2001). *Lojistik regresyon analizinin incelenmesi ve bir uygulaması.* Yayınlanmamış yüksek lisans tezi, Cumhuriyet Üniversitesi, Sosyal Bilimler Enstitüsü, Sivas.

Bircan, H. (2004). Lojistik regresyon analizi: Tıp verileri üzerine bir uygulama. *Kocaeli Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 2*, 185-208.

Büyüköztürk, Ş. (2009). *Sosyal bilimler için veri analizi el kitabı: İstatistik, araştırma deseni, SPSS uygulamaları ve yorum* (9. bs). Ankara: PegemA Yayıncılık.

Büyüköztürk, Ş., & Çokluk-Bökeoğlu, Ö. (2008). Discriminant analysis: Concept and application. *Eurasian Journal of Educational Research, 33*, 73-92.

Büyüköztürk, Ş., Kılıç-Çakmak E., Akgün, Ö. E., Karadeniz, Ş. ve Demirel, F. (2008). *Bilimsel araştırma yöntemleri (2. baskı).* Ankara: PegemA Yayıncılık.

Cook, D. (2008). *Binary response and logistic regression analysis*. www.public.iastate.edu/~stat415/ stephenson/stat415_chapter3.pdf adresinden 22 Kasım 2008 tarihinde edinilmiştir.

Deryakulu, D. ve Büyüköztürk, Ş. (2002). Epistemolojik İnanç Ölçeği'nin geçerlik ve güvenirlik çalışması. *Eğitim Araştırmaları, 18*, 111-125.

Deryakulu, D. ve Büyüköztürk, Ş. (2005). Epistemolojik inanç ölçeğinin faktör yapısının yeniden incelenmesi: Cinsiyet ve öğrenim görülen program türüne göre epistemolojik inançların karşılaştırılması. *Eğitim Araştırmaları, 5* (18), 57-70.

Field, A. (2005*). Discovering statistics using SPSS* (2nd ed.). London: Sage.

Garson, G. D. (2008). *Logistic regression.* http://www2.chass.ncsu.edu/garson/ PA765/ logistic.htm adresinden 22 Kasım 2008 tarihinde edinilmiştir.

George, D., & Mallery, P. (2000). *SPSS for Windows step-by-step: A simple guide and reference* (2nd ed). Boston: Allyn and Bacon.

Grimm, L.G., & Yarnold, P. R. (Eds.). (1995). *Reading and understanding multivariate statistics.* Washington D.C.: American Psychological Association.

Gündoğdu, M. (2002). Üniversite öğrencilerinin bilimsel düşünme becerilerinin yordanması. *Türk Psikolojik Danışma ve Rehberlik Dergisi, II* (17), 11-18.

Gürcan, M. (1998). *Lojistik regresyon analizi ve bir uygulama.* Yayınlanmamış yüksek lisans tezi, Ondokuz Mayıs Üniversitesi, Fen Bilimleri Enstitüsü, Samsun.

Hair, J. F., Black, W. C., Babin, B., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate data analysis* (6th ed). Upper Saddle River, NJ: Prentice-Hall.

Kalaycı, Ş. (Ed.). (2005). *SPSS uygulamalı çok değişkenli istatistik teknikleri.* Ankara: Asil Yayın Dağıtım.

Kayri, M. ve Okut, H. (2008). Özel yetenek sınavındaki başarıya ilişkin risk analizinin karışımlı lojistik regresyon modeli ile incelenmesi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 35,* 227-239.

Kılıç, S. (2000). *Lojistik regresyon analizi ve pazarlama araştırmalarında bir uygulama.* Yayınlanmamış yüksek lisans tezi, İstanbul Teknk Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul.

Kökdemir, D. (2003). *Belirsizlik durumlarında karar verme ve problem çözme.*Yayımlanmamıs doktora tezi, Ankara Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Leech, N. L., Barrett, K. C., & Morgan, G. A. (2005). *SPSS for intermediate statistics: Use and interpretation* (2nd ed). Mahwah, NJ: Lawrence Erlbaum Associates.

Menard, S. (1995). *Applied logistic regression analysis*. Thousand Oaks, CA: Sage.

Mertler, C. A., & Vannatta, R. A. (2005). *Advanced and multivariate statistical methods: Practical application and interpretation* (3rd ed.). Glendale, CA: Pyrczak Publishing.

Miller, A. J. (1990). *Subset selection in regression*. London: Chapman and Hall.

Myers, R. (1990). *Classical and modern regression with applications* (2nd ed). Boston, MA: Duxbury.

Önder, H. ve Cebeci, Z. (2001). Lojistik regresyonlarda değişken seçimi. *Çukurova Üniversitesi Ziraat Fakültesi Dergisi, 17*(2), 105-114.

Peng, C. Y. 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.

Poulsen, J., & French, A. (2008). *Discriminant function analysis.* http://userwww.sfsu.edu/~efc/classes /biol710/discrim/discrim.pdf adresinden 22 Kasım 2008 tarihinde edinilmiştir.

Schommer, M. (1990). Effects of beliefs about the nature of knowledge on comprehension. *Journal of Educational Psychology, 82*(3), 498-504.

Seven, Z. (1997). *Değişken seçimi yöntemi olarak adimsal lojistik regresyon ile adımsal diskiriminant analizinin karşılaştırılması.* Yayınlanmamış yüksek lisans tezi, Gazi Üniversitesi, Fen Bilimleri Enstitüsü, Ankara.

Stephenson, B. (2008). *Binary response and logistic regression analysis.* www.public.iastate.edu / ~stat415/stephenson/stat415_chapter3.pdf. adresinden 22 Kasım 2008 tarihinde edinilmiştir.

Şekercioğlu, G. (2008). *Fatalizm Ölçeği'nin geliştirilmesi: Geçerlik ve güvenirlik çalışması.* Yayınlanmamış makale taslağı.

Tabachnick, B. G., Fidell, L. S. (1996). *Using multivariate statistics* (3rd ed.). New York, USA: HarperCollins College Publishers.

Tatlıdil, H. (1996). *Uygulamalı çok değişkenli istatistiksel analiz.* Ankara: Engin Yayınları.

Tate, R. (1992). *General linear model applications*. Unpublished manuscript, Florida State University.

Todman, J., & Dugard, P. (2007). *Approaching multivariate analysis: An introduction for psychology.* New York: Taylor & Francis Group.