Research Article

THE IMPORTANCE OF LOGISTIC REGRESSION IMPLEMENTATIONS IN THE TURKISH LIVESTOCK SECTOR AND LOGISTIC REGRESSION IMPLEMENTATIONS/FIELDS

Murat KORKMAZ[1]*, Selami GÜNEY[2], Şule Yüksel YİĞÎTER[3]

ABSTRACT

Logistic regression analysis is one of the mostly preferred regression methods that can be implemented in modelling binary dependent variables. Logistic regression is a mathematical modelling approach used to define the relationship between such independent variables as X1, X2, …, Xn and Y binary dependent variable which is coded as 0 or 1 for two possible categories. The independent variables may be continuous, discrete, binary or a combination of them. In this paper, logistic regression models are researched. Maximum likelihood methods may be used to estimate the parameters of the logistic model. The interpretations of coefficients are made with odds rate values. In other words, in this paper, the logistic regression analysis has been reviewed that can define the relationship between the binary result variable and independent variables comprising of both continuous and discrete variables. Shortly, the applicability of logistic regression in the livestock has been researched.

Key Words: Logistic Regression Analysis, Binary variable, Stepwise (onward and backward) logistic regression, livestock

TÜRKİYEDE HAYVANCILIK SEKTÖRÜNDE LOJİSTİK REGRESYON UYGULAMALARININ ÖNEMİ LOJİSTİK REGRESYON UYGULAMALARI/ALANLARI

ÖZET

Lojistik regresyon analizi ikili bağımlı değişkenleri modellemek için uygulanabilen en çok tercih edilen regresyon metotlarından biridir. Lojistik regresyon X1, X2, …, Xn gibi bağımsız değişkenleri ile iki olası kategori için O veya 1 gibi kodlanmış Y ikili bağımlı değişkeni arasındaki ilişkiyi tanımlamak için kullanılan matematiksel modelleme yaklaşımıdır. Burada bağımsız değişkenler sürekli, kesikli, ikili veya bunların karışımı olabilir. Bu çalışmada lojistik regresyon modelleri araştırmaktadır. En çok olabilirlik metotları lojistik modelin parametrelerini tahmin etmek için kullanılır. Katsayıların yorumu odds oran değerleriyle yapılır. Bir başka deyişle bu araştırmada, ikili sonuç değişkeni ile hem sürekli hem de kesikli değişkenlerden oluşan bağımsız değişkenler kümesi arasındaki ilişkiyi tanımlayabilen lojistik regresyon analizi incelenmiştir. Kısaca, bu çalışmada lojistik regresyonun hayvancılıkta uygulanabilirliği ele alınmıştır.

Anahtar Kelimeler: Lojistik Regresyon Analizi, İkili değişken, adımsal (ileriye, geriye doğru) lojistik regresyon, Hayvancılık

[1]* Finance Manager of Güven Group, hakanmuratkorkmaz34@gmail.com
[2] Erzincan University İ.İ.B.F. (Faculty of Economics and Administrative Sciences)
[3] Department of Accounting and Financing

INTRODUCTION

Livestock has a significant place in the Turkish economy. As in many other sectors, livestock is now to be performed in the light of science and with technology. Developments in information technologies have given momentum to scientific research in the field of livestock. (Koçak H.)[1] Logistic Regression analysis system has begun to be used in almost all the studies in the cultivation and livestock sector of Turkey. Logistic regression is commonly used in such fields as clinical studies (Gardside, 1995), (Gibbon and fr., 1996), stock farming (Boer and fr., 1990), and agriculture, biology and environment (Wang, 1998).

The objective of this study is to implement the logistic regression analysis in data where variables are frequent, and to determine its frequency of usage in the field of zootechnics and livestock. Additionally, it aims at identifying the values obtained as a result of this usage and its contributions to zootechnics, studies observations of which have been completed and their contributions to the respective fields, and the contributions of Logistic Regression statistics implementations to the studies. The overall purpose of this study is to implement the logistic regression analysis in data where discrete variables are frequent, and to reach the best model that will appoint observations to one of the groups in the structure of data.

Among the goals of this study are to identify the importance of logistic regression analysis system in the livestock sector and zootechnics sector of Turkey, to ensure the use of (Logistic Regression) analysis system in studies carried out in these fields, to determine the its contributions to the sectors, and to define the importance of logistic regression analysis system. The most important aim of the usage of the Logistic Regression analysis is to ensure that it is the best analysis form which, in the event that the dependent variable in different fields of science contains two or more levels, and independent variables are both discrete and continuous, can appoint data to the groups (to which they belong) in the most proper way and determine the relevant risk factors, thereby has fundamental benefits to livestock and zootechnics.

[1] Prof. Dr. Hikmet Koçak, Atatürk University

ROAD MAP AND MATERIALS

In this paper, the importance of logistic regression will be introduced and emphasized; benefits will be identified that have been achieved by analyzing the logistic regression studies which were carried out with the Logistic Regression before. Additionally, data and results that have been obtained from the studies on the fields of livestock, zootechnics and so on, and the contributions of these data and results to the fields of livestock and zootechnics will be identified. It will also be ensured that a model is formed by examining the Logistic Regression analysis and the methods implemented and identifying its contributions to the studies carried out; and it is determined how the logistic regression analysis implementations in this model can be used in a more efficient way for livestock and zootechnics.

Subject Matter: It is known that, as in other fields, the logistic regression analysis used in the field of livestock and zootechnics has great numerical benefits to the studies. Using Logistic Regression analysis has become an obligation for the evidence of the accuracy of data from the analysis of studies and for ensuring accuracy in the statistical works. What are the benefits and credibility of data from studies in which the logistic regression is used?

Other Matters:

Are researches in many fields on livestock and zootechnics and the logistic regression analysis data from these researches enough for further research?

Does today's research approach allow for estimating more credible results by granting numerical values in data?

It is important in that it is the most suitable method for the obtained data. Does data from researches provide credibility to the research?

Is it true that the logistic regression analysis, which is used in research on livestock and zootechnics provide numerical benefits to researches?

Hypothesis:

H1 Logistic Regression analysis data from the fields of livestock and zootechnics are enough.

H0 Logistic Regression analysis data from the fields of livestock and zootechnics are not enough.

H1 Numerical values from the research provide credible result.

H0 Numerical values from the research do not provide credible result.

H1 The model built upon the data from researches provides suitability to the research.

H0 The model built upon the data from researches does not provide suitability to the research.

H1 Logistic Regression analysis provides numerical credibility to the research on livestock and zootechnics.

H0 Logistic Regression analysis does not provide numerical credibility to the research on livestock and zootechnics.

METHODOLOGY

If there are medium-significant variables after beginning with variable selection using single variable logistic regression analysis, then multivariable logistic regression method should be selected. It should be determined whether these medium-significant variables will be included into the model as continuous or discrete, and interaction between variables should be examined. Statistically meaningful interaction terms are included in the model after tested by Odds rate test, thereby the model will be defined. The relationship between the logistic regression analysis and independent variables that affect the dependent variable has been reviewed, and the analyses have been reviewed as suitable for each hypothesis on the logistic regression method.

Linear Regression: One of the most important subjects of statistics is the regression analysis. Regression analysis provides us to make estimates through data from the past. According to Gujarati, regression analysis aims at predicting the dependency of a dependent variable to other expressive variables.

The most simple regression analysis is the binary variable regression analysis. The model is:

$Y=\beta_0 + \beta_1 X_i + u$

In this model, while Y represents the dependent variable, X shows the independent variable; β_0 the invariable; β_1 oblichity; and u the margin of error. Regression models can be solved through pretty much and long formulas. But, today, such programs as Excel, SPPS, SAS etc. can make these proceeds (Gujurati, 1995).

Regression analysis is one of the methods that are used in such disciplines as economy, mathematics, physics, biology and agriculture and is used in determining the relationship between two or more variables that have a cause and effect relation. The core of the regression analysis is to research which matters have an effect on an observed event while evaluating it. As these matters may be one or more, it may be directly or indirectly affected from them.

While making regression analysis, observation values and affected events should be represented with a mathematical demonstration that is a function. This established model is called regression model. In other words, when giving upon the pattern instead of groundmass, regression is the estimate of a dependent variable by expressive variables, and the measurement of errors made in this process.

While examining the regression analysis, the matters that affect it and generally constitute its subject are called variables, and the mathematical model including these variables are analyzed. Variables are patterns that include matters in a particular unit constituting a mass in a certain period. They should be measurable and countable. There should be a cause and effect relationship between the relevant events for the regression model to be used. In forming a regression model, cause and effect relationship are defined as dependent and independent variables. The model contains a dependent variable and one or more independent variables. If the dependent variables are considered as a variable in the model to be formed and simple, it constitutes a multiple regression model.

Model: Simple Linear Regression Model:

$Y = \beta_0 + \beta_1 X + e_i$

Multiple Regression Model:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + e_i$

Y: Dependent variable

X_1 , X_2 , X_3 , …. : Independent variables

β_0, β_1… : Coefficients

e_i : Error term (Average=0 and Variance=σ^2)

Regression analysis mainly aims at determining the quality of the relationship between variables. If a variable is used as an estimate variable, it is

simple regression; and if two or more variables are used as estimate variables, it is called multiple regression analysis. The point is to determine the contribution of each estimate variable to the total change in criteria variable, and thereby to estimate the criteria value based upon the value of a linear combination of estimate variables (Kurtuluş, 1985; Field, 2000). As different from the linear regression, the accrual possibility of one of the values that the result variable may get is estimated in logistic regression (Şahin 1999). The most distinct characteristic discriminating the logical regression from the linear is that the variable is binary. This difference between the logistic regression and linear regression reflects in the selection of the parametric model and assumptions. As in linear regression, assumptions are tried to be made based upon certain variable values in the logistic regression. However, there are three important differences between these two methods:

1. While the dependent variable to be predicted is continuous in the linear regression analysis, it should be a discrete value in the logistic regression.

2. In the linear regression analysis, the value of the dependent variable is estimated while one of the values that the dependent variable may get is estimated in the logistic regression.

3. While the condition that independent values should demonstrate a multiple normal distribution is sought in the linear regression analysis, there is no precondition in the logistic regression regarding the distribution of independent variables (Bircan, 2004).

Today, it has also become an obligation to analyze qualitative variables in regression models. There is no natural measurement scale for these qualitative variables called "Dummy variable". So, to see the contribution of these variables to the model, such variables should fall into "a category". Dummy variables mostly get the value 0 and 1, and there is no numerical importance of these new variables. For, they just show which category the observation fall into. Most of these variables include two results. In this study, the relationship of variables in the relevant groups are tried to be explained through the Logistic Regression analysis.

Logistic Regression: The problems encountered in many fields and sectors are solved and interpreted based upon numerical values though statistical analysis. Logistic Regression estimates the effects of independent variables on the result variables as probability. The logistic regression ensuring the determination of the risk factors as probability is a method that investigates the relationship of the result variables with independent variables in binary or multiple phases.

In case of various assumption distortions (such as normality, common variances etc.), logistic regression studies and practices are used as an alternative to discriminant analysis and crosstabs. And if the dependent variable is binary like 0,1 or discrete containing more than two levels, as the normality assumption is distorted, it also is an alternative to the linear regression analysis (Ünal, 1996). The Logistic Regression analysis creates alternative solutions according to the data of emerging problems. There are certain reasons to use this method in many fields. They may be collected under many titles. Among them are parameter estimate methods (maximum likelihood method, weighted iterative least squares method, minimum logit chi-squared method). If it includes more variables than the model variables, then "Multiple Logistic Regression" model is used.

Linear probability function is one of the assumptions of the logistic regression that means the eligibility of the distribution of error terms to logistic distribution (http://www.deu.edu.tr/userweb/k.yaralioglu/dosyalar/ver_mad.doc). There are several methods to appoint the observations into possible groups included in the structure of data. These are:

- Clustering
- Discriminant
- Logistic Regression

The number of groups included in the structure of data is known in advance in Discriminant and Logistic Regression analyses, and discrimination model is obtained by using these data. The observations newly included in data set are appointed to groups through the discrimination model. While the logarithmic linear regression necessitates all the independent or regressor variables to be categorical, discriminant analysis provides for all the independent variables to be continuous. In case of the presence of categorical and numerical independent variables, the logistic regression analysis necessitates less assumption. Logistic regression is similar to discriminant analysis in terms of the aim of estimating a categorical dependent variable, and it necessitates less assumption. On the other hand, if the assumptions necessitated by the discriminant analysis are provided, the logistic

regression may also be implemented (Akgül and Çevik, 2003).

Logistic Regression is a regression method that helps in performing categorization and appointment process. It is a statistical method that makes categorization according to the rules of probability by estimating the value assumptions of dependent variables as probability. The point in the Logistic Regression is to offer a scientifically acceptable model that determines the relationship between dependent and independent variables to have the best suitability by using the least variable. It is a regression model that examines the relationship between discrete and continuous (independent) variables and those which have binary result variables (dependent variables). Logistic Regression is a method that is also used when dependent variables are binary, tertiary, ternary and quaternary (Bircan, 2004: Özdamar, 2002; BUİS, 2005). The Logistic Regression is a method that helps determining the cause and effect relationship between expressive variables when the regression response variable are observed in categorical, binary, ternary and multiple categories. It is a method in which the expected values of the result variable are estimated as probability according to expressive variables. Simple and multiple regression analyses are used to examine mathematical correlation between expressive variable(s) and dependent variable. To implement this method in data sets, the dependent variable should demonstrate a normal distribution; it should be singled from community(es) that demonstrate a normal distribution in independent variables; and the error variance should demonstrate a parametric normal distribution. In data sets in which such conditions are not met, simple or multiple regression analyses will not be implemented.

The Logistic Regression analysis is a method that helps in categorization and appointment process. There is no precondition such as a normal distribution variance or continuous variance. The effects of expressive variables on the dependent variables are estimated as probability thereby enabling the risk factors to be determined as probability. Logistic Regression aims at estimating parameters according to the logistic mode that is formed. It is possible to include common variables into the models in the Logistic Regression. Thus, Y values that are corrected according to the common variables may be estimated. The Logistic Regression is a statistical method that allows for categorization as appropriate to the rules of probability by estimating the values for the dependent variable. It analyzes tabulated or pure data sets. Depending upon the type of a dependent variable, there are three main methods of logistic regression analysis:

- Binary logistic regression (BLOGREG),
- Ordinal logistic regression (OLOGREG) and
- Nominal logistic regression (NLOGREG).

Binary Logistic Regression (BLOGREG) Analysis: It is a type of logistic regression analysis that are made through dependent variables including binary results (yes/no, does/do not, proves/does not provide etc.). It suggests the correlation between one or more expressive variables and binary result variable. Expressive variables are either expressive or common variables. Factor variables are nominal scaled and common variables should be continuous. Unless defined, Minitab and Spss programs consider expressive variables in a data set as common variables. As the process to define a model in Blogreg analysis may be performed according to direct user defining method (the enter method), it may be done through the progressive approach. In the progressive selection of model, an onward selection (according to conditional probability approach) or backward elimination methods may be implemented.

Ordinal Logistic Regression (OLOGREG) Analysis: It is a method implemented when the result variable is ordinal. The ordinal scaled-result variable should include values that are observed at least in three categories. In coding ordinal scaled data or determining the categories nominally, the results should have a structure of natural order (such as light/medium/strong or didn't like/like/like much). OLOGREG analysis operates through code values rather than nominal categories.

Parameter estimates should be refreshed in OLOGREG analysis-it makes maximum resemblance parameter estimates according to the weighted least squares method. The assumption that the categories are parallel to each other is used in OLOGREG. In identifying the most appropriate logit models in OLOGREG, models as many as the binary combinations of the category number ((c-1)/2) are determined, and the analogy of these sub-models is analyzed, or the analysis is made based on the result that has the highest value and by forming logit models according to this reference. The factors

that are included into the model as expressive variable may be categorical or continuous. If a common variable is included to the model, the common variable should be continuous.

Nominal Logistic Regression (NLOGREG) Analysis: It is a method implemented when the result variable is nominal. The nominal scaled-result variable should include values that are observed at least in three categories. In coding the values observed, these categories do not have to be in order. For instance, the category of profession names or sportive activities may be determined nominally.

In NLOGREG, parameter estimates are made according to the repetitive-weighted least squares method. They are the maximum resemblance estimates. In identifying the parameter estimates in NLOGREG, the assumption of determining the most appropriate logit models is used. In identifying logit models, models as many as the binary combinations of the category number ((c-1)/2) are determined, or the analysis is made based on one of the categories and by forming binary logit models according to this reference. If the reference value is not stated, the first result will be taken as reference. To ensure that the values that dependent variable may get are between 0-1, the model providing a curvilinear relationship between the dependent variable and independent variable should be used (http://epidemiyoloji.org/moodle/mod/wiki/view.php?id=741&page=Lojistik+regresyon&MoodleSession=16b88071cfe0a1c9581788013d2eb068).

When the logistic regression is in the qualitative data form of the dependent variable, it is used to define the relationship between the dependent variable and one or more expressive (independent) variables. The Logistic Regression analysis may be extended from the case of two categorical dependent variables to the one with more than two categories. In literature, the analysis of the cases where the dependent value is observed to be more than two is called multiple categorical or multi-nominal logistic regression analysis. For example, when a dependent variable is observed in at least three categories like light/medium/strong, multiple categorical logical regression analysis should be implemented.

Shift from the two categorical logistic regression analyses to multiple categorical logistic regression analysis is mathematically possible. A value of the dependent variable (generally the first or the last value) is selected as reference. Then, the probability of the selected category is compared to the probabilities of other categories. It is pretty simple to make this comparison for the dependent variable that is measured in the ordinal scale. For the dependent categories comprising of M category, in explaining the relationship between the dependent variable and independent variables, M-1 equation with which the reference category and each category are examined respectively should be calculated (http://www.sayisalyontemler.com/?q=content/cok-kategorili-lojistik-regresyon-analizi).

Forming the Model: The point in forming the model is to, as much as possible, explain the change in the dependent variable through the most independent variable. If many variables are included in the model, standard error estimates will increase. At the same time, it will become more complicated to form and develop a model with many independent variables. There are many different methods in selecting variables in the logistic regression model. They are performed according to two main analyses as single variable analysis and multivariable analysis. Multivariable analysis includes two methods. These are stepwise method and the best sub-sets method. The best sub-sets method is not often used in the logistic regression analysis (Costanza at all,1992). The stepwise method is divided into onward selection and backward screening (Lee and Koval, 1997).

Different methods for model selection may be used in forming logistic regression model. While there is no variable in the regression equation in the onward variable selection method, it is based on, beginning from the most related variable, adding significant variables one by one to the equation in each phase. The backward variable screening method is implemented by screening insignificant variables one by one in each phase from the regression equation in which all the variables are included. In other words, the processes of selecting and screening variables are performed according to a statistical proceed that control the significance of variables. The significance of a variable is defined with demonstrating the statistical significance of coefficient for the variable. Onward variable selection and backward selection screening may be performed in three different ways-Wald, probability rate and conditional-in package software such as SPSS for Windows.

In the enter method, there is no stepwise procedure, and the significance of coefficients of all the variables are assessed statistically in one step. Model forming methods are similar to those in the linear regression. The three differences between them are:

1. In the logistic regression, dependent variable is in the form of qualitative data like there is/there isn't, ill/healthy. Independent variables may be numerical continuous-discrete or in the form of qualitative data as in the multiple linear regression.

2. While a value for the dependent variable is predicted for each observation in the linear regression, the probability of a risk to emerge for each observation is obtained in the logistic regression. That is, the result is a value between 0-1. In other words, in the logistic regression, the probability (to happen) of one of two values that the dependent variable may get is obtained.

3. Many of the assumptions in the linear regression are not present in the logistic regression. For instance, while the normal distribution of errors is rechi-red with zero average and certain variance in the multiple linear regressions, there is no such precondition in the logistic regression.

As in multiple linear regressions, appropriate variables should also be included in the model in logistic regression, and those that are not casually appropriate should not be included in the model. As a general approach, minimum 10 observations should be made for each independent variable in the model. If there is one independent variable, Binary Logistic model:

$$P(Y=1)=\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}=\frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$$

Multiple Binary Logistic Regression model(http://www.biyoistatistik.hacettepe.edu.tr/Donem_III/Turkce/coklu_dogrusalolmayan_lojistik.pps#364,).:

$$P(Y=1)=\frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_p X_p}}$$

$$=\frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_p X_p)}}$$

Conditional and Limited Logistic Regression: A limited binary data is a data structure in which observation may be made in groups, there are at least two units in each group and at least one positive reply for each group on the event examined is provided. The logistic regression model in which limited binary data is used as dependent value is conditional and limited logistic regression models. If independent variables whose effects are researched are at unit level, either method may be used. But if the effects of independent variables at group level as well as unit level are to be researched, since the independent variables at group level cannot be included in the conditional logistic regression model, the limited logistic regression model should be used. Grouped data sets are often found in the studies in health. It is of very importance in grouped data to include group effects into the analysis for achieving correct results. Limited Logistic Regression or Conditional Logistic Regression models are used in examining the relationship between the categorical results from grouped data sets and the factors affecting the formation of these results through statistical methods. While both models may be implemented limited or grouped data sets, group effects can be examined in the limited logistic regression model.

Comparing Conditional and Limited Logistic Regressions: Limited and conditional logistic regression models are two different logistic regression models that are used in the analysis of data in the same form and for the same purpose. Independent variables to be used in the conditional logistic regression model should be necessarily at individual level. In the limited logistic regression model, independent variables may be at either unit or group level. So, the limited logistic regression model is more effective than the conditional logistic regression model.

Intended Purpose of the Logistic Regression: The most important intended purpose of the logistic regression analysis is to form a model that, when dependent variable includes two or more levels and independent variable are discrete and continuous, can appoint data to groups they belong to in the most proper way, and that can determine the risk factors regarding the research. There are two main reasons for selecting the logistic regression:

- It can be implemented with mathematically ease and flexibility.
- It provides biologically meaningful interpretations.

If Y is defined in the logistic regression and linear regression as indicator variable that get two values (0 and 1), the expected value of error terms related them (e_i) is zero; the assumption that $E(e_i)=0$ and its variances are stable, $Var(e_i)=\sigma_e^2$ does not occur. As a result, the estimates obtained in case of deviation from assumptions will not be the best linear and unbiased predictors. This failure prevents the use of linear regression in categorization analyses.

Thus, the logistic regression is one of the methods commonly used in categorization analyses. As it does not rechi-re multivariable normal distribution assumption, it prevails in such studies. Additionally, it has a feature of identifying the possibilities on the category membership.

Importance of the Logistic Regression: The point in using logistic Regression is that it is same as the other model building methods used in statistics. It aims at building a biologically acceptable model that can identify the relationship between dependent and independent variables in a way that will have the best suitability with use of the least variable. The logistic regression has begun to be used commonly especially in recent years. The method is an alternative to the linear regression as the normality assumption fails in case of binary categorical or multi-categorical discrete variable. As from its flexible use due to not having any assumption limitation, the fact that the model from analysis is mathematically flexible and it can be easily interpreted has increased the interest in the method (Özdamar,2002). The logistic regression does not rechi-re multivariable normality assumption. It plays an important role in the tendency to the method that the model from implementations is mathematically flexible and can be easily interpreted, and results in meaningful implementations (Ünal, 1996).

Reasons for the Logistic Regression Analysis Being Up-to-date:

1. It may be implemented if the dependent variable is discrete, but independent variables are both discrete and continuous.

2. It is the same as the discriminant function and linear regression model that corresponds to the number of the parameters in the logistic model.

3. As parameters of the logistic model are similar to measurements in the up-to-date life, it may be interpreted easily.

4. There are many computer programs for logistic model-based analyses.

5. It is a stronger model against assumption distortions.

6. It is a function whose usage is mathematically easy.

Research on the Logistic Regression: It is known that there are a large number of researches in literature. Use of the logistic regression for the analysis of biological experiments was firstly suggested by Berkson (1944), and Cox (1970) reviewed this model and developed various implementations. The summing-up advances were firstly granted by Anderson (1979-1983). There have been also works on the suitability of data with the logistic model. Among them are works by Aranda-Ordaz (1981) and Johnson (1985) are the most important. Pregibon (1981) examined influential and outlier observations and diagnostics in binary group logistic model, and Lesaffre (1986) and Lesaffre and Albert (1989) examined influential and outlier observations and diagnostics in multiple group logistic models.

Gardside and Glueck (1995) examined the effects of such factors as diet, cigarette smoke and alcohol use and physical activity on heart diseases in people (Gardside and Glueck, 1995). Kloiber *et al.* (1996), Peoples *et al.* (1991), Buescher *et al.* (1993) examined the risk factors affecting the low birth weight in women; Santos *et al.* (1998) the relationship between caffeine consumption and low birth weight; Sable and Herman (1997) the relationship between preterm labour and low birth weight (Kloiber et al,1996: Peoples et al,1991: Buescher, 1993: Santos, 1998).

With commonly use of the logistic regression models, methods of coefficient estimate have been developed and the logistic regression models have been reviewed in a more detailed way. Discriminant function approach was firstly used in the proceeds of coefficient estimate in logistic regression and made popular by Cornfield (1962). Lee (1984) focused on linear logistic models for cross-over testing plans. Bonney (1987) worked on the use and development of the logistic regression model (Bonney, 1987). Robert et al. (1987) worked on chi- square, probability rate (G2), "pseudo" maximum likelihood estimates, excellence of suitability and hypothesis tests. Duffy (1990) examined the proximity of the distribution of error terms and the parameter values in the logistic regression to actual values. Başarır (1990) worked on the multivariable logistic regression analysis in clinical data and the discrimination issue. Hsu and Leonard (1995) focused on obtaining Bayes estimates in the logistic regression functions and shows that Monte Carlo transformation could be used in logistic regression. Akkaya and Pazarlıoğlu (1998) examined the use of logistic regression models in economy using examples. Cox *et al.* (1998) worked on the relationship between cardiovascular diseases and hypertensive diseases.

Logistic Regression Analysis System: Assumptions on logistic regression are briefly as follows:

• Y1,……,Yn Y1 is statistically independent.

• Independent variables (X_k) are independent from each other.

• Yi Є (0,1) i = 1,2,……,n

• $P(Y_i=1/X_i)=P_i$ i = 1,2,……,n

The linear probability function, which is one of the assumptions on logistic regression, is the suitability of error terms with the logistic distribution

$$P = \frac{e^{\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_pX_p}}{1+e^{\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_pX_p}}$$

$$= \frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_pX_p)}}$$

P: Observing probability of an analyzed event

β_0: Another statement for the value of dependent variable when independent variables get a value of zero

β_0, β_1: Regression coefficients of independent variables

X_1 X_2... X_p: *Independent* variables

p: The number of independent variables

e: Error Term

P represents the observing probability of an event analyzed in the logistic regression equation. The proportion of probability of an analyzed event to probability of the other events is *Odds Value* (Çolak, and Özdamar, 2004). The proportion of the Odds values of two different analyzed events to each other is *Odds Rate*. Odds Rate is defined as exp(β) in the logistic regression. As Odds is the proportion of probability of an event happening to the probability of not happening (Gujarati, 1999), exp (β_p) shows how many more folds in what percentage Y variable has observing probability with the effect of X_p variable (Girginer, 2008).

Features of the Logistic Regression: The dependent variable gets a value of 0 or 1.

Therefore, Pr(Y = 1 | x) = π (*x*)

Y is interpreted as the odds rate. That is, probability is built by the contrary of an event. Natural logarithm of the odds rate is logit transformation. Another feature of a logistic function is that it can be linearized easily.

Binary Group Logistic Regression Analysis System: If dependent variable has two levels like 0,1, with the probability of $P(Y_i=1)$ to get a value of i-fold event, the expected value will be:

$E(Y_i) = 1xp(y_{i=1})+0xp(y_i=0)=P(y_i=1)$.

If the result is to be shown as a regression equation:

$$E(y_i)=p(y_i=1)=\sum_{k=0}^{p}\beta_k\ x_{ik}$$

In the linear probability model with equity, left side of the equation gets a value between 0-1. The regression model in which Y_i values of dependent variable are binary is the linear probability model. Expected equity is not always reached when these probability values are linked with independent values that may get infinite values except the said values. In this case, the probability values as dependent variable become defined in (-∞,+∞) range by modifying them. Some of these transformations are probit and logistic (Ünal, 1996).

Binary Group Logistic Regression

1) y_i ε (0,1) i=1,2,…,n

2) $P(y_i=1/x_i) = P_i$

$$P_i = \frac{e^{\sum_{k=0}^{p}\beta_k X_{ik}}}{1+e^{\sum_{k=0}^{p}\beta_k X_{ik}}} \qquad (2.3.2.)$$

3) $y_1,y_2,\ldots,y_n$ are statistically independent.

4) Expressive variables are independent from each other. They are defined as follows according to being discrete, continuous or both discrete and continuous:

a. If all the expressive variables are discrete, the logistic model is:

$$\ln (P_i / (1-P_i)) = \sum_{k=0}^{p}\beta_k X_{ik}$$

b. If all the expressive variables are continuous, as $P(x_1,\ldots,x_p)$ is conditional probability of success on the p expressive variables, the logistic model is defined as:

$$\ln(\frac{P(x_1,...,x_p)}{1-P(x_1,...,x_p)}) = \beta_0 + \sum_{k=1}^{p}\beta_k X_{ik}$$

c. If some of the expressive variables are discrete, and some are continuous, multivariable frequency distribution is $f_1(x_1,\ldots,x_p)$ for success and $f_0(x_1,\ldots,x_p)$ for failure, and the logistic model is.

$$\ln\left(\frac{Pf_1(x_1,...,x_p)}{(1-P)f_0(x_1,...,x_p)}\right)=\beta_0+\sum_{k=1}^{p}\beta_k X_{ik}$$

P: Pre-probability of the reply variable to get a value of 1 (Başarır, 1990). Methods to estimate the coefficients of a logistic model in Binary Group Logistic models are maximum likelihood, reweighted iterative least squares method and minimum logit chi- square method.

Generally in forming a regression model, either the least squares method or maximum likelihood is used. Parameters are estimated by, if there is an assumption that the error term demonstrates a normal distribution, using maximum likelihood or, if there is no assumption on the distribution of error term, using the least squares method (http://fikretgultekin.com/yukseklisans/Regresyon%20Analizi.pdf). In this study, maximum likelihood, which is one of the parameter estimating methods, will be reviewed.

Maximum Likelihood Estimate Method: Maximum likelihood is one of the point estimate methods used in statistics and econometrics. In general, the maximum likelihood method gives the values of unknown parameters that make maximum the probability of obtaining a data set observed. Maximum likelihood function is rechi-red to use this method. It explains the probability of data observed as a function of unknown parameters. Maximum estimators of these parameters are selected to find the values making the function maximum. Thus, the estimators that are obtained in the end have highly similar values to the observed data (Bircan,2004). This method was developed by the English statistician, Sir Ronald A. Fisher (1890-1962) in 1920s.

The maximum likelihood method tries to find groundmass parameters that make maximum the probability of certain sampling values happening(http://www.yildiz.edu.tr/~tastan/teachi-ng/tahminyont_slides.pdf). The result of maximum likelihood estimate of a multiple group logistic model is dependent upon structure of the function. Type of the sampling plan and data is important in forming the function (Ünal, 1996). The goal of the maximum likelihood method is to find β estimates of p expressive variable as to make maximum the probability of Y variable observing.

It is necessary to estimate β parameters in order to make (2.3.2) ideal position. Estimate operations in linear regression are made using the Least Squares Method. It is the Maximum Likelihood method that makes estimations below the linear regression model when error terms demonstrate a normal distribution. The first thing to do in implementing the model is to form the maximum likelihood function. It shows the probability of data observing. β_0 and β_1 values in the logistic regression model are selected fort he observed values of Y. While the probability of an event happening is $p_i=p(y_i=1/x)$, the probability of it not happening is $1-p_i$. (i=1,…,n)

It can be made as $P(y_i/x_i)=p_i^{yi}\,(1-p_i)^{1-yi}$

If we generalize the probability for n probability:

$$L(Y/X)=\prod_{i=1}^{n}p_i^{\,y_i}\left(1-p_i\right)^{1-y_i}$$

p_i : probability of the event happening

$1-p_i$: probability of the event not happening

$$L(Y/X)=\prod_{i=1}^{n}\left(\frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}\right)^{y_i}\left(\frac{1}{1+\exp(\beta_0+\beta_1 x)}\right)^{1-y_i}$$

β_0 and β_1 are sought in maximum likelihood function(http://oak.cats.ohiou.edu/~milesd/logistic.ppt#283,1). As L(Y/X, β) is the probability function, the method selects the estimate value of β to make maximum this function. Generally, it maximizes ln L(Y/X) rather than L(Y/X) to find the maximum likelihood estimator.

$$\ln L(Y/X,\beta)=\sum_{i=1}^{n}\left(y_i\ln p_i+(1-y_i)\ln(1-p_i)\right)$$

To find the β vale that makes ln L(Y/X, β) maximum, derivative of ln L(Y/X,β) to β is calculated, and it is made equal to zero. The equities obtained are called probability equations (Bircan,2004).

$$\sum_{i=1}^{n}(y_i-p_i)x_{ij}=0 \qquad j=1,\ldots,p$$

It is impossible to reach a result since P_i is exponential. It rechi-res iterative solution methods (Seven,1997).

Parameter Significance Test: The actual matter we should consider is to define the best suitability model with the least parameters. The next rational step is to make a new analysis adding significant variables into the model and to compare it with a full model. When independent variables that are scaled categorically are

excluded from (or included into) the model, all the design variables of that variables should also be excluded (or included) (Atakurt, 1999). Comparison of observed and expected values in logistic regression is made with the log likelihood function.

D= -2 ln (Probability of the actual model/ Probability of the saturated model)

The statement above given in parentheses is called "probability ratio". When multiplied by (–2ln), it gives a mathematical value whose distribution is known. This value is used for hypothesis testing. Such tests are called probability rate test. If the statements in parentheses are placed, the following equation is obtained:

$$D = -2\sum_{i=1}^{n}\{y_i \ln\left(\frac{\hat{\pi}_i}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\hat{\pi}_i}{1-y_i}\right)$$

D values of the cases when there is an independent variable in the equation and when there is not are compared to determine the significance of a variable. Change in D for having an independent value is:

G = D (for non-variable model) – D (for variable model)

The statistics above has the same role as share side in F test that is used in linear regression. Since probabilities of the saturated model are the same for both D values that will be subtracted from each other to calculate G, the G statistics is:

G= -2 ln (Probability of the non-variable model/Probability of the variable model)

In $\beta_1=0$, G statistics shows χ^2 distribution with 1 degree of freedom. The validity of the model, which is built upon the idea that measures based on the difference of the odds rate values of the model that includes all the variables and the estimated model are chi- square distributed, is tested according to χ^2 table (Elhan, 1997). If the calculated G statistics is lower than the value in chi- square table with 1 degree of freedom ($G<\chi^2$), it is determined that the analyzed variable should be removed from the model. And it is concluded that this variable makes no contributions to the model. The most suitable variable logistic model is reached, by realizing the same process for the other insignificant variables (Çolak and Özdamar 2004).

Usage areas of the Logistic Regression: Logistic regression is used in analyzing the data that show binomial distribution. Parameter estimates are obtained through logistic models in the logistic regression (Bonney, 1987: Wang and Putterman, 1998). Logistic regression analysis, or shortly logit models are commonly used in social and biological sciences. In most of the research in social sciences, it is assumed that the dependent variable may have one of two possible values. In recent years, logistic regression has become a model that is used more in social sciences (http://en.wikipedia.org/wiki/Logistic_regression). Moreover, it has been found that it is also commonly used in epidemiology, medicine, meteorology and economy (Ünal, 1996).

CONCLUSION

In this research, the logistic regression has been researched in literature in general terms, and binary result dependent variables and independent variables have been reviewed. Logistic regression analysis methods have been used to find whether independent variables are important or not. It has also been found that the logistic regression, which is only one of the multivariable statistical models, may be used in zootechnics and livestock. Besides, it has been reviewed whether a statistically insignificant variable in the single variable logistic regression is biologically meaningful before excluding from the model. If it is found that a statistically insignificant variable is biologically meaningful (significant), variables should be included to the multivariable logistic regression model.

REFERENCES

Akgül A. and Çevik O., 2003, Statistical Analysis Methods "Management Implementation is SPSS", Emek Offset, Ankara.

Alpar, R., Regression Overview Presentation, http://www.biyoistatistik.hacettepe.edu.tr/Donem_III/Turkce/coklu_dogrusalolmayan_lojistik.pps#364,30, Slide 30

Atakurt, Y., 1999, Logistic Regression Analysis and an Implementation in Its Use in Medicine, Ankara University Faculty of Medicine Journal, C.52, Issue 4, P.195, Ankara

Başarır, G.,1990, Discrimination Issue in Multivariable Data and Logistic Regression Analysis (Applied statistics doctoral thesis) Hacettepe. U., 1-36, Ankara

Bircan H., Logistic Regression Analysis: Practice in Medical Data, Kocaeli University Social Sciences Institute Journal, 2004 / 2: 185-208

Bonney, G. E., 1987, Logistic Regression for dependent binary observations. Biometrics, 43(4): 951-973.

Buescher, P.A., Larson, L.C., Nelson, M.D., Lenihan, A.J., 1993, Prenatal WIC Participation Can Reduce Low Birth Weight and Newborn Medical Costs: A Cost Benefit Analysis of Wic Participation in North Carolina, Journal of the American Dietetic Association, 93:163-166.

BUİS, 2005

Costanza M.E., Staddat A.M., Gaw V. and Zaplea J.G., 1992, "The Risk Factors of Age and Family History and Relationship To Screening Mammography Utilization", Journal of The American Statistical Association, 40, 776.

Çolak, E., Özdamar K., 2004, Review of Conditional and Limited Regression Models by the Risk Factors in Fatal Traffic Accidents, OGÜ Faculty of Medicine Journal, Volume 26 P.1 Eskişehir

Elhan, A.H, 1997, Review of Logistic Regression Analysis and Implementation in Medicine. (PhD thesis in biostatistics) A.U., 4-29, Ankara

Field, A., 2000, Discovering Statistics, Sage Publications

Gardside, P.S., Glueck, C.J., 1995, The Important Role of Modifiable Dietary and Behaviour Characteristic in the Causation and Prevention of Coronary Heart Disease Hospitalization and Mortality. Journal of American College of Nutrition, 14: 71-79.

Girginer, N, 2008, Measuring the Satisfaction of Tramway Passengers with Logistic Regression Analysis: Estram Pattern, Celal Bayar University FEAC Management and Economics Journal, Manisa

Gujurati, D. N., 1995, "Basic Econometrics", McGraw-Hill, Inc., New York

Kloiber, L.L., Winn, N.J., Shaffer, S.G., Hassanein, R.S., 1996, Late Hyponatremia in very Low Birth Weight Infants: Incidence and Associated Risk Factors. Journal of the American Dietetic Association, 96: 880-884.

Kurtuluş, K., 1985, Marketing Research, Economics and Management Institute, 3. Edition

Lee K. and Koval J.J., 1997, "Determination of The Best Significance Level in Forward Stepwise Logistic Regression", Communication in Statistics, 26(B), 566.

Peoples, M.D., Siegel, E., Suchi-ndran, C.M., Origasa, H., Ware, A., Barakat, A., 1991, Characteristics of Maternal Employment during Pregnancy: Effects on Low Birth weight. American Journal of Public Health. 81: 1007-1012.

Santos, I.S., Victoria, C.G., Huttly, S., Carvalhal, J.B., 1998, Caffeine Intake and Low Birth Weight: A Population Based Case Control Study. American Journal of M., 1988, The Retreat From Class: A New True Socialism, London: Verso.

Seven, Z., 1997, Comparing Stepwise Variable Selection and Stepwise Discriminant Analysis as Variable Selection Method, PhD thesis, Ankara

Şahin M., 1999, Logistic Regression and Its Use in Biological Fields, Kahramanmaraş,

Özdamar K., 2002, Statistical Data Analysis Using Package Programs–I, 4. Edition, Kaan Bookstore, Eskişehir

ÜNAL, 1996

Wang, P., Putterman, M. L., 1998, Mixed logistic regression models. Journal of Agriculture, Biological and Environmental Statistics, 3(2): 175-200.

http://epidemiyoloji.org/moodle/mod/wiki/view.php?id=741&page=Lojistik+regresyon&MoodleSession=16b88071cfe0a1c9581788013d2eb068

http://www.deu.edu.tr/userweb/k.yaralioglu/dosyalar/ver_mad.doc

http://www.sayisalyontemler.com/?q=content/cok-kategorili-lojistik-regresyon-analizi

http://fikretgultekin.com/yukseklisans/Regresyon%20Analizi.pdf

http://www.yildiz.edu.tr/~tastan/teaching/tahminyont_slides.pdf

http://oak.cats.ohiou.edu/~milesd/logistic.ppt#283,1,Alternative Methods of Regression

http://en.wikipedia.org/wiki/Logistic_regression