ANADOLU ÜNİVERSİTESİ BİLİM VE TEKNOLOJİ DERGİSİ
ANADOLU UNIVERSITY JOURNAL OF SCIENCE AND TECHNOLOGY
Cilt/Vol.:7 Sayı/No: 1 : 127-141 (2006)

ARAŞTIRMA MAKALESİ/**RESEARCH ARTICLE**

MODEL BUILDING IN LOGISTIC REGRESSION MODELS ABOUT LUNG CANCER DATA

Özgül VUPA[1], C. Cengiz ÇELİKOĞLU[2]

ABSTRACT

In this study, a simple and multiple logistic regression model forms, several of their key features and model building procedures are concerned. Maximum likelihood procedures are used to estimate the model parameters of a logistic model. Interpretation of the coefficients is explained by using odds ratio values.

When the model includes more variables than needed, the greater estimated standard errors become. For this reason, there are some methods to find the best fitting through variables for the model. The final model equations of these methods can be different from each others. Here, the aim is to determine the "best" model.

A logistic regression model is developed by using a database of 1200 patients with lung cancer in İzmir. In order to obtain a solution, univariate analysis, forward selection and backward elimination methods are applied to cancer data. The SPSS software package is used and results are evaluated.

Keywords : Binary variable, Stepwise logistic regression, Odds ratio, Likelihood ratio Test (G), Lung cancer.

AKCİĞER KANSERİ VERİLERİ İLE İLGİLİ OLARAK LOJİSTİK REGRESYON MODELLERİNDE MODEL KURMA

ÖZ

Bu çalışmada, basit ve çoklu lojistik regresyon model yapıları, onların bazı anahtar özellikleri ve model kurma yöntemleri ile ilgilenilmektedir. En çok olabilirlik yöntemleri lojistik modelin parametrelerini tahmin etmek için kullanılır. Katsayıların yorumu odds oran değerleri kullanılarak yapılmaktadır.

Model gereğinden fazla değişken içerdiği zaman, daha büyük standart hatalar elde edilmektedir. Bu nedenle, değişkenler arasındaki en iyi modeli bulmak için bazı yöntemler kullanılmaktadır. Bu yöntemlerin son model denklemleri birbirinden farklılık gösterebilmektedir. Burada amaç "en iyi" modeli bulabilmektir.

Çalışmada lojistik regresyon modeli, İzmir ilindeki akciğer kanserli 1200 hastaya ilişkin veriler kullanarak geliştirilmiştir. Çalışmada tek değişkenli lojistik regresyon çözümlemesi, ileriye doğru seçim ve geriye doğru eleme yöntemleri uygulanmıştır. Çözümlemeler SPSS paket programı kullanılarak yapılmış ve elde edilen sonuçlar tartışılmıştır.

Anahtar Kelimeler : İkili değişken, Adımsal lojistik regresyon, Odds oranı, Olabilirlik oran testi (G), Akciğer kanseri.

[1] Ress. Ass. Dokuz Eylul University, Faculty of Arts and Sciences, Statistics Department, Buca/İZMİR
E-mail: ozgul.vupa@deu.edu.tr
[2] Ass. Prof. Dokuz Eylul University, Faculty of Arts and Sciences, Statistics Department, Buca/İZMİR

Received: 14 October 2004 **Reviced**: 23 February 2005 **Accepted**: 27 September 2005

1. INTRODUCTION

The logistic regression is used when the response variable is measured on a nominal scale. In other words, logistic regression analysis is used to study the association between a qualitative variable Y and a quantitative or qualitative variables X (Rao, 1984). In logistic regression, the odds ratio values of variables are examined to interpret of the coefficients. In addition, there are three statistical methods that are often employed in determining which variables to include in a model: the univariate method, the stepwise logistic regression method and the best subsets logistic regression method. The stepwise logistic regression method contains two analyses: the forward selection and the backward elimination. In this study, the minimum logit chi-square method (the likelihood ratio chi-square test) is used. So, the final model with appropriate variables is stated for lung cancer patients. After fitting the model, Hosmer-Lemeshow test is used to determine the fit of the model for lung cancer patients.

2. REGRESSION MODELS WITH BINARY RESPONSE VARIABLE

The response variable has only two possible outcomes, and it can be represented by a binary indicator variable taking on values 0 and 1. These response variables are measured on a binary scale. For example, the responses may be alive or dead, or present or absent (Freund and Wilson, 1996).

The simple linear regression model is written as: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, $i = 1, 2, \ldots, n$. The expected response, $E\{Y_i\}$, has a special meaning in this case. Since $E\{\varepsilon_i\} = 0$, it is written as: $E\{Y_i\} = \beta_0 + \beta_1 X_i$. When Y_i is a Bernoulli random variable, there are two probabilities. π_i is the probability that Y_i=1 and $(1-\pi_i)$ is the probability that Y_i=0. The expected value of a Bernoulli random variable is $E\{Y_i\} = 1(\pi_i) + 0(1-\pi_i) = \pi_i$. So, $E\{Y_i\}$ is written as $E\{Y_i\} = \beta_0 + \beta_1 X_i = \pi_i$.

2.1 Meaning of Response Function When Response Variable is Binary

The error terms in linear regression model are assumed to have a normal distribution with a constant variance for all levels of X. However, when the response variable is 0 or 1 indicator variable, error terms are not only distributed normal but also they don't have constant variance. The error term $\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$ can take on only two values. If Y_i=1, then the error term takes the value as $\varepsilon_i = 1 - \pi(x_i) = 1 - \beta_0 - \beta_1 X_i$ with the probability $\pi(x_i)$. If Y_i=0, then the error term takes the value as $\varepsilon_i = -\pi(x_i) = -\beta_0 - \beta_1 X_i$ with probability $1-\pi(x_i)$. Thus, the assumption of normality does not hold for this model. It is not appropriate (Neter and Kutner, 1996). Another problem with the error terms (ε_i) is that they do not have equal variances. The variance of Y_i for the simple linear regression model can be determined as follows: (Neter and Kutner, 1996)

$$V(Y_i) = E(Y_i - E(Y_i))^2 = (1-\pi_i)^2 \pi_i + (0-\pi_i)^2 (1-\pi_i) = \pi_i (1-\pi_i) \qquad \textbf{(2.1)}$$

Also, the variance of the error terms (ε_i) is the same as that of Y_i, because ε_i is equal to $(Y_i - \pi_i)$ and π_i is a constant. The last problem is related with constraints on response function. Since the response function represents probabilities, the mean responses should be constrained as follows: $0 \le E(Y_i) = \pi_i \le 1$

3. THE MODEL

The conditional mean, $\pi(x_i)$, is expressed as follows:

$$\pi(x_i) = E(Y|x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \qquad \textbf{(3.1)}$$

This specific form is called logistic response function. A transformation of $\pi(x_i)$ is the logit transformation. This transformation is expressed as follows:

$$g(x_i) = \ln\left[\frac{\pi(x_i)}{1-\pi(x_i)}\right] = \ln(e^{\beta_0 + \beta_1 x_i}) = \beta_0 + \beta_1 x_i \qquad \textbf{(3.2)}$$

The importance of this transformation is that $g(x_i)$ has many of the desirable properties of a linear regression model. The logit transformation is linear in its parameters and it may be continuous. In addition, the logit may have range from $-\infty$ to ∞, depending on the range of x_i (Hosmer and Lemeshow, 1989).

4. FITTING OF SIMPLE AND MULTIPLE LOGISTIC REGRESSIOIN MODELS

4.1 Fitting Of Simple Logistic Regression Model

The general methods of estimation in logistic regression model are investigated in three main concepts. These are the Maximum Likelihood Method, Iteratively Reweighted Least Squares Method and the Minimum Logit Chi-Square Method.

4.1.1 Likelihood Function

Likelihood function express the probability of the observed data as a function of the unknown parameters. For pairs (x_i, y_i), since $y_i = 1$, the contribution to the likelihood function is $\pi(x_i)$. Since $y_i = 0$, the contribution to the likelihood function is

$1-\pi(x_i)$. Since Y_i's have a Bernoulli distribution, the probability density function can be defined as follows:

$$P(Y=y_i)=f_i(y_i)=\pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i} \quad \textbf{(4.1)}$$

Where $y_i=0$ or $y_i=1$ for $i=1, 2, \ldots, n$. Since the observations Y_i are assumed to be independent, the likelihood function can be defined as follows:

$$L(\beta_0,\beta_1)=\prod_{i=1}^{n}\pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i} \quad \textbf{(4.2)}$$

In order to maximize this function, the derivative must be taken with respect to each of the parameters. Then, the resulting equations which are called likelihood equations would be set equal to zero and solved simultaneously. This process can be simplified by performing the same analysis on the natural log of the likelihood function (Kleinbaum, 1994). Obtaining the log-likelihood function is expressed as:

$$\ln L(\beta_0,\beta_1)=\ln\prod_{i=1}^{n}\pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i}$$

$$=\sum_{i=1}^{n}\left\{y_i(\beta_0+\beta_1x_i)-\ln(1+e^{\beta_0+\beta_1x_i})\right\} \quad \textbf{(4.3)}$$

Likelihood equations are not linear, solving these equations simultaneously requires an iterative procedure that is normally left to a software package.

4.1.2 Maximum Likelihood Estimation Method

The maximum likelihood estimation method (MLE) is used to calculate the logit coefficients. This method yields values for the unknown parameters which maximize the probability of obtaining the observed set of data. In order to apply this method, the likelihood function is constructed. This method uses the logistic function and an assumed distribution of y to obtain estimates for the coefficients that are most consistent with the sample data.

The sum of the observed values of y_i is equal to the sum of the expected values. Fitted simple logistic response function for the i[th] case is follows:

$$\hat{\pi}(x_i)=\frac{\exp(\hat{\beta}_0+\hat{\beta}_1x_i)}{1+\exp(\hat{\beta}_0+\hat{\beta}_1x_i)} \quad \textbf{(4.4)}$$

MLE is a iterative algorithm and this procedure is complex and usually requires numerical search methods. Hence MLE of the logistic regression is done on a computer.

4.1.3 Testing For The Significance of the Coefficients

After estimating the coefficients, an assessment of significance of the variable in the fitted model is concerned. The approach in testing for the significance of the coefficient of a variable in the model is related with the following question. Does the model which includes the variable in question tell us more informations about the response variable than does a model which does not include that variable? This question is answered by comparing the observed values of the response variable to those predicted by each of two models. The comparison is based on the log-likelihood. In logistic regression model, there are three commonly used tests for hypothesis testing. These are likelihood ratio test, wald test and score test.

i) Likelihood Ratio Test

The comparison of observed to predicted values is based on the log likelihood function defined in equation (4.3). To better understand this comparison, it is helpful conceptually to think that an observed value of the response variable as also being a predicted value resulting from a saturaed model. A saturated model is one that contains as many parameters as there are data points. This comparison is obtained as follows:

$$D=-2\ln\left[\frac{\text{likelihood of the current model}}{\text{likelihood of the saturated model}}\right] \quad \textbf{(4.5)}$$

This expression is called the deviance (D). The deviance for logistic regression model plays the same role as SSE in linear regression. Using minus twice its log is necessary to obtain a quantity whose distribution is known. Also, this procedure can be used for hypothesis testing purposes. This test is called likelihood ratio test. In order to determine whether the parameter is significant to the model or not, the deviance of the model containing the independent variable is compared with the deviance of the model without the independent variable. This change in D is called G statistic. This statistic in logistic regression plays the same role as the numerator of the partial F test in linear regression. The test statistic is expressed as follows:

$$G=D(\text{for the model without the variable})-D(\text{for the model with the variable}) \quad \textbf{(4.6)}$$

$$G=-2\ln\left[\frac{\text{likelihood without the variable}}{\text{likelihood with the variable}}\right] \quad \textbf{(4.7)}$$

In checking the significance of the coefficient, the following null and alternative hypotheses are written as $H_0:\beta_1=0 \quad H_1:\beta_1\neq 0$. The statistic G has a chi-square distribution with 1 degrees of freedom under $H_0:\beta_1=0$. The p-value associated with this test is $P(\chi^2_{(df=1)}>G)$. If this p-value is less than given α-level, then the null hypothesis is rejected. This is a

statement of the statistical evidence for the independent variable.

ii) Wald Test

Wald test is based on the comparision between maximum likelihood estimate of the slope parameter $\hat{\beta}_1$ and an estimate of its standard error. Standard error of $\hat{\beta}_1$ is provided by the square root of the corresponding daigonal element of the covariance matrix $V(\hat{\beta})$. This test for the logistic regression model is as follows:

$$W = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \qquad \textbf{(4.8)}$$

Under the hypothesis that $\beta_1 = 0$, W is a standard normal distribution. This test also can be written in an alternative manner. Because the squaring a normal random variable will result in a chi-square random variable with 1 degrees of freedom. So, the Wald test statistic is written as follows:

$$W^2 = \left(\frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \right)^2 \qquad \textbf{(4.9)}$$

Where $W^2 \sim \chi^2_{(1-\alpha,1)}$. In accordance with this equation, the decision rule must be adjusted such that the null hypothesis is rejected when p-value that is evaluated by $P(|\chi^2| > W^2)$ is less than given α-value.

iii) Score Test

The score test is based on the conditional distribution theory of the derivatives of the likelihood equations. The test statistic for the score test (ST) is calculated as follows:

$$ST = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sqrt{\bar{y}(1-\bar{y}) \sum_{i=1}^{n} (x_i - \bar{x})^2}} \qquad \textbf{(4.10)}$$

Under the hypothesis that β_1 is equal to zero, the two tailed p-value is evaluated by $P(|Z| > ST) < \alpha$-level and this test statistic has a standard normal distribution.

4.2 Fitting Of Multiple Logistic Regression Model

In this setting, the vector $\tilde{x} = (x_1, x_2, \ldots, x_p)$ represents the collection of p independent variables for this model. The equations for the probability and the logit transformation can be expressed as follows:

$$\pi(\tilde{x}) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)} = \frac{\exp(g(\tilde{x}))}{1 + \exp(g(\tilde{x}))} \qquad \textbf{(4.11)}$$

$$g(\tilde{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \qquad \textbf{(4.12)}$$

There is a sample of n independent observations and it is expressed as $(\tilde{x}_i, y_i)$. The maximum likelihood estimates of the parameters are used and it is shown as: $\tilde{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$. The likelihood function for the multiple logistic regression model is expressed as:

$$L(\tilde{\beta}) = \prod_{i=1}^{n} \pi(\tilde{x}_i)^{y_i} (1 - \pi(\tilde{x}_i))^{(1-y_i)} \qquad \textbf{(4.13)}$$

In this case, there are p+1 likelihood equations which are obtained by differentiating the log-likelihood function with respect to the p+1 coefficients. As in the univariate model, the solution of the likelihood equations requires special package programs. Maximum likelihood estimates of the parameters can be found in many packages (SPSS, Minitab, SAS).

The method of estimating the variances and covariances of the estimated coefficients follows from theory of maximum likelihood estimation. This theory states which estimators are obtained from the matrix of second partial derivatives of the log-likelihood functions. The estimated variance and the confidence interval of the estimated coefficients are denoted as follows:

$$Var(\hat{\tilde{\beta}}) = \left[X^1 V X \right]^{-1} \qquad \hat{\tilde{\beta}}_j \pm Z_{1-\alpha/2} SE(\hat{\tilde{\beta}}_j)$$

$$SE = \sqrt{Var\left(\hat{\tilde{\beta}}\right)} \qquad \textbf{(4.14)}$$

4.2.1 Design Variable

If some of the independent variables are discreate, ordinal or nominal scaled variable (categorical variable) with more than two levels, then the model differs from general formula in lojit transformation. For example, race, sex, regions of Turkey, number of treatment groups and so on... If the number of variable categories is equal to k, then k-1 design variables must be created. The notation to indicate design variables is more different than the logistic regression model. Suppose that the j[th] independent variable x_j has k_j levels. The $k_j - 1$ design variables are needed and they are denoted as D_{jm}. In addition, the coefficients for these design variables are denoted as β_{jm}, $m = 1, 2, \ldots, k_j - 1$. The logit for a model with p independent variables and the j[th] independent variable being discrete is expressed as:

$$g(\tilde{x}) = \beta_0 + \beta_1 x_1 + \ldots + \sum_{m=1}^{k_j - 1} \beta_{jm} D_{jm} + \beta_p x_p \quad \textbf{(4.15)}$$

4.2.2 Testing For The Significance Of The Model

i) Likelihood Ratio Test

The parameters in the multiple setting are once again determined through MLE method, because Y is still a Bernoulli random variable with the same probability distribution. The derivation of the maximum likelihood estimators remains the same, with the expection of the inclusion of more parameters. The log-likelihood equation takes the form as follows:

$$\ln L(\beta_0, \beta_1, \ldots, \beta_p) = \sum_{i=1}^{n} \{ y_i (\beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}) - \ln(1 + \exp(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi})) \} \quad \textbf{(4.16)}$$

In the same manner as before, the equations resulting from taking the derivative of the log-likelihood equation with respect to each of the parameters and then setting each derivative equal to zero are solved simultaneously in order to obtain the estimates. The likelihood ratio test is used for overall significance of the p-coefficients for the independent variables in the model. The test is based on the G statistic. In order to determine whether the model is significant or not, the log-likelihood of the model without the variable(s) must be compared with the log-likelihood of the model with the variable(s) (Hosmer and Lemeshow, 1989). The test statistic, G, is calculated as follows:

$$G = -2 \ln \left[\frac{\text{likelihood without the variable(s)}}{\text{likelihood with the variable(s)}} \right] \quad \textbf{(4.17)}$$

In checking the significance of the model, the following null and alternative hypotheses are written as $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$ H_1 : at least one of the $\beta_p \neq 0$. The statistic G has a chi-square distribution with $(\nu_2 - \nu_1)$ degrees of freedom. Here, ν_2 equals to the number of variables in the full model +1 and ν_1 equals to the number of variables in the reduced model +1. For this test, the decision rule requires that p-value is $P\left\{ \chi^2_{(1-\alpha, df=(\nu_2-\nu_1))} > G \right\}$. If this p-value is less than α-value, H_0 is rejected. This means that the model would be deemed significant. if p-value is greater than α-value, then the reduced model is as good as the full model and the null hypothesis (H_0) is failed to reject.

5. INTERPRETATION OF THE COEFFICIENTS

The estimated coefficients for the independent variables give the slope or rate of change of a function of the dependent variable per unit of change in the independent variable. The function of the dependent variable yields a linear function of the independent variables. This is called a link function. In linear regression model, it is the identity function. In logistic regression model, the link function is the logit.

In linear regression model, the slope coefficient, β_1, is equal to the difference between the value of the dependent variable at $x+1$ and the dependent variable at x. It is expressed as follows:

$$\beta_1 = \left| y(x = x+1) - y(x = x) \right| \quad \textbf{(5.1)}$$

In logistic regression model it is expressed as follows:

$$g(x+1) = \ln\{\pi(x+1)/(1-\pi(x+1))\} = \beta_0 + \beta_1 (x+1) = \beta_0 + \beta_1 x + \beta_1 \quad \textbf{(5.2)}$$

Here, the logit difference is equal to β_1 and evaluated as follows:

$$g(x+1) - g(x) = g(x+1) - (\beta_0 - \beta_1 (x)) = \beta_1 \quad \textbf{(5.3)}$$

i) Dichotomous Independent Variable

In this case, independent variable (x) can take only two values and it is coded as 0, 1. In logistic regression model, there are two values of $\pi(x)$ and two values of $1-\pi(x)$. The odds of the outcome being present among individuals with $x=1$ and $x=0$ are expressed respectively

$$\frac{P(y=1|x=1)}{P(y=0|x=1)} = \frac{\pi(1)}{1-\pi(1)}, \qquad \frac{P(y=1|x=0)}{P(y=0|x=0)} = \frac{\pi(0)}{1-\pi(0)} \quad \textbf{(5.4)}$$

The logit is defined to be the logarithm (natural exponential) of the odds. They are defined by $g(1)$ and $g(0)$ for dichotomous independent variable. The "odds ratio" is defined as the ratio of the odds for $x=1$ to the odds for $x=0$ and it is expressed as follows:

$$OR = \frac{\pi(1) / (1-\pi(1))}{\pi(0) / (1-\pi(0))} \quad \textbf{(5.5)}$$

$$OR = \frac{[\exp(\beta_0 + \beta_1)/(1+\exp(\beta_0 + \beta_1))][1/(1+\exp(\beta_0))(1+\exp(\beta_0))]}{[\exp(\beta_0)/(1+\exp(\beta_0))][1/(1+\exp(\beta_0 + \beta_1))]} = \exp(\beta_1) \quad \textbf{(5.6)}$$

The log of the odds ratio is called logit difference (log odds ratio) and it is expressed as $\ln(OR) = \ln[\pi(1)/(1-\pi(1))] - \ln[\pi(0)/(1-\pi(0))] = g(1) - g(0) = \beta_1$. OR can take any value between 0 and ∞. The odds

ratio gives us the effect of a one-unit change in X on the probability that $Y=1$. If the odds ratio equals 1, the effect is estimated to equal 0. If the odds ratio is greater than 1, for example $\hat{OR}$ equals 1.3, a one-unit increase in X raises the probability of $Y=1$ by 0.3, or 30%. On the other hand, If the odds ratio is less than 1, for example $\hat{OR}$ equals 0.7, the effect of X on Y is negative: a one-unit increase in x leads to a 30% reduction in the probability of $Y=1$.

The variance is evaluated by $Var(\hat{\beta}_1)=\left[(1/a)+(1/b)+(1/c)+(1/d)\right]$. Where a,b,c,d are cell frequencies in the 2×2 table of $Y\times X$. The distribution of the estimate of OR tends to be skewed to the right. Thus, confidence interval is usually based on $\hat{\beta}_1$ which is closer to being normally distributed. $\hat{\beta}_1 \sim N(\beta_1, Var(\hat{\beta}_1))$ The confidence interval for the odds ratio is $\exp\left\{\hat{\beta}_1 \pm Z_{1-\alpha/2}\ SE(\hat{\beta}_1)\right\}$.

ii) Polytomous Independent Variable

In this case, if the independent variable takes three or more levels, then, it is called polytomous independent variable. For example, nominal scale variable X is coded at 4 levels. Thus, (4-1) =3 design variables are created.

iii) Continuous Independent Variable

In this case, when there is an independent continuous variable in the model, the unit of this variable should be defined. Most often the value of "1" is not biologically very interesting. For example, increased risk for 1 additional year of age or mmHg in systolic blood pressure or mg/100 ml of cholesterol are not very interesting. But, A change of 10 years or 5 mmHg or 25 mg/100 ml may be more meaningful. The log odds ratio for a change of c units in X, odds ratio and variance of the variable are expressed respectively as follows:

$$x = g(x+c)-g(x)=c\beta_1, \qquad OR(x+c,x)=e^{c\,\beta_1},$$
$$Var\left\{\ln(\hat{OR}(x+c,x))\right\}=c^2 Var\hat{\beta}_1) \qquad \textbf{(5.7)}$$

100% confidence interval is evaluated as:

$$\exp(c\ \hat{\beta}_1 - Z_{1-\alpha/2}\ c\ SE(\hat{\beta}_1)) \le OR \le \exp(c\ \hat{\beta}_1 + Z_{1-\alpha/2}\ c\ SE\ (\hat{\beta}_1)) \qquad \textbf{(5.8)}$$

6. MODEL BUILDING PROCEDURE

If there are more variables included in the model, then estimates of standard errors become greater. While there are many independent variables in the model, model building and devoloping include more complex situations. For this reason, to select less variables is very important. There are different ways used for variable selection in logistic regression model. These are the univariate analysis and the multivariate analysis. Multivariate analysis consists on two methods. These are stepwise logistic regression methods (Forward Selection, Backward Elimination) and best subset logistic regression method.

It is so clear that modeling is a useful process both for prediction of future observables and for describing the relationship between variables. Large models reproduce the data on which they were fitted better than smaller models. The saturated model provides a perfect fit of the data. However, smaller models have more powerful interpretations and are often better predictive tools than large models. Often, the main goal is to find the smallest model that fits the data (Cristensen, 1997).

6.1 The Univariate Analysis

The variable selection process begins with univariate analysis of each variable. If a cell contains no observation, this cell is called "the zero cell" and this situation should be paid extra attention. The zero cell yields a univariate point estimate for one of the odds ratios of either zero or infinity. The observations should be designed before making a univariate analysis.

The variables are selected for the multivariate analysis after fitting the univariate analysis. Any variable whose univariate test has a p-value ≤ 0.25 is considered as candidate for the multivariate model along with all variables of known clinical importance. Otherwise, if any variable's p-value is greater than 0.25, then this variable is excluded from the model (Ryan and Thomas, 1997). Why is the p-value less than 0.25? If we set the threshold too low, we often fail to identify variable known to be important. If we set the threshold too high, then the model consists of variables that are of questionable importance.

The importance of each variable included in the multivariate logistic regression model should be verified. Variables that do not contribute to the model are eliminated from the model and the new model is constructed. The new model are compared to the old model through the likelihood ratio test. Variables whose coefficients have changed markedly in magnitude are concerned. Thus, the value of these statistics may give us an indication of which variables in the model may or may not be significant. In this case, the likelihood ratio test (G) is used. Using this notation, the p-value associated with this test is $P(\chi^2_\nu > G) < 0.05$, thus there is a strong evidence that the investigated variable is a significant variable in predicting Y. This is the statistical evidence for this variable.

The question of the appropriate categories for discrete variables should have been addressed at the univariate stage. The linearity in the logit for continuous scalled variables should has been checked. How will we do this check procedure? It is checked

with dummy (design) variable method. The stages of the dummy variable method are as follows: 1. Obtain the quartiles of the designed variable, 2. Create a categorical variable with 4 levels using the 3 quartile values as the cutt-off points, 3. Create 3 design variables with the lowest quartile serving as the reference group, 4. Fit the multiple logistic regression using the dummy variables, 5. Plot the odds ratio values of the estimated coefficients according to groups. If there is no linear relationship that can be increasing or decreasing between them, then dummy variable method is used.

6.2 The Stepwise Logistic Regression Method

Stepwise logistic regression is an extremely popular method for model building. Stepwise logistic regression is used when the outcome being studied is relatively new (AIDS, some cancer types…) and the important covariates may not be known and associations with the outcome not well understood. In these situations, most studies collect many possible covariates and determine them for significant associations. For this reason, stepwise procedures provide a useful, fast and effective means to determine a large number of variables and fit a number of logistic regression equations (Hosmer and Lemeshow, 1989).

Stepwise procedures assume an initial model and then use rules for adding or delating terms to arrive at a final model (Cristensen, 1997). There are two procedures for model building in the stepwise logistic regression method. The forward selection process adds variables sequentially to the model until further additions do not improve the fit. At each stage, the variable giving the greatest improvement in the fit is selected. The maximum p-value for the final model is a sensible criterion. A stepwise variation of this procedure retests, at each stage, variables added at previous stages to see if they are still needed. The backward elimination process begins with a complex model and sequentially removes variables. At each stage, the variable with least damaging effect on the model is removed. The process stops when any further deletion leads to a significantly poorer-fitting model.

In logistic regression the errors are assumed to follow a Binomial distribution, and significance is assessed with respect to the likelihood ratio (chi-square) test. So, the variable that produces the greatest change in the log-likelihood at any step in the procedure will be most important variable in statistical terms. There are k-1 design variables for discrete variables with k levels. The importance of G depends on its degrees of freedom.

For likelihood ratio (chi-square) test accepted α-level such as 0.05 or 0.10 is chosen as the critical value for the entry of variables into the model. For this model building process, this cutoff value for the entry or removal of a variable can be increased to around 0.20. This will help in avoiding possibly significant variables from being overlooked or removed unnecesssarily from the model. This method will be described by considering the statistical computations that the computer must perform at each step of the procedure (Vupa, 2004). Before starting a procedure, it is necessary to give some informations and abbreviations where possible.

p_E: The probability value for enter to a model, p_R: The probability value for removal from a model, j : The number of independent variables. $j = 1, 2, \ldots, p$

Step (0):

1) Fit a model with intercept only and evaluate the value of its log-likelihood, L_0.

2) Fit each of the p possible univariate logistic regression models, denote the log-likelihood value by $L_j^{(0)}$ for $j = 1, 2, \ldots, p$ and compare their respective log-likelihoods.

L : Log-likelihood statistic, $L_j^{(0)}$: The subscript j refers to that variable which has been added to the model and the subscript 0 refers to the step.

3) Evaluate the value of likelihood ratio statistic for the model containing x_j versus the intercept only, denote the likelihood ratio statistic by $G_j^{(0)} = 2(L_j^{(0)} - L_0)$ and compute the p-value by $p_j^{(0)} = \Pr(\chi_\nu^2 > G_j^{(0)})$.

G : Likelihood ratio statistic, $G_j^{(0)}$: The subscript j refers to that variable which has been added to the model and the subscript 0 refers to the step.

a) $\nu = 1$ if x_j is continuous.

b) $\nu = k - 1$ if x_j is a categorical variable with k levels.

4) Find the variable with smallest p-value, denote this variable by x_{e_1} and find minimum p-value by $p_{e_1}^{(0)} = \min(p_j^{(0)})$.

The subscript e_1 is used to denote that the variable is a candidate for entry at Step 1. For example, if variable x_3 had the smallest p-value, then $p_3^{(0)} = \min(p_j^{(0)})$ and $e_1 = 3$.

5) Determine whether this variable will enter or not into the model, compare $p_{e_j}^{(0)}$ with a pre-specified significance level p_E.

a) If $p_{e_j}^{(0)} < p_E$, move on the next step.

b) If $p_{e_j}^{(0)} \geq p_E$, stop the procedure.

It is different from the hypothesis test where the pre-specified significance level is commonly selected as 0.05, 0.10 or 0.15.

Step (1)

Fit the logistic regression model containing the variable x_{e_1}, denote the log-likelihood of this model by $L_{e_1}^{(1)}$.

1) Determine whether any of the remaining p-1 variables are important once the variable x_{e_1} is in the model. Fit p-1 logistic regression models which contain only the x_{e_1} and one other variable x_j, $j=1,2,\ldots,p$ and $j \neq e_1$. Denote the corresponding log-likelihood value by $L_{e_1,j}^{(1)}$. Compute the likelihood ratio statistic by $G_j^{(1)} = 2(L_{e_1,j}^{(1)} - L_{e_1}^{(0)})$ and its corresponding p-value by $p_j^{(1)} = \Pr(\chi_\nu^2 > G_j^{(1)})$.

2) Let x_{e_2} corresponds $p_{e_2}^{(1)} = \min(p_j^{(1)})$.

a)If $p_{e_2}^{(1)} < p_E$, grow the model by including x_{e_2} and move on the next step.

b) Otherwise, stop the procedure.

Step (2)

Backward elimination and forward variable selection.

1) Fit a model containing both x_{e_1} and x_{e_2}.

2) Remove variable x_{e_j} from the model just established in Step 2, $j=1,2$ and denote the log-likelihood value for the reduced model by $L_{-e_j}^{(2)}$ and evaluate the corresponding log-likelihood ratio statistic by $G_{-e_j}^{(2)} = 2(L_{e_1,e_2}^{(1)} - L_{-e_j}^{(2)})$.

3) Calculate p-value by $p_{-e_j}^{(2)} = \Pr(\chi_\nu^2 > G_{-e_j}^{(2)})$ and select the variable x_{r_2} with $p_{r_2}^{(2)} = \max(p_{-e_1}^{(2)}, p_{-e_2}^{(2)})$.

The subscript r_2 is used to denote that the variable is a candidate for removal at Step 2. For example, if variable x_3 had the largest p-value, then $p_3^{(2)} = \max(p_{-e_1}^{(2)}, p_{-e_2}^{(2)})$ and $r_2 = 3$.

4) For a pre-specified significance level p_R, if $p_{r_2} > p_R$, the variable x_{r_2} should be removed from the model against the situation that the variable just being added is possibly eliminated, $p_R > p_E$ should be selected. If excluding any variables once they have entered is not required, then $p_R = 0.90$ is chosen.

5) Fit $p-2$ logistic regression models containing x_{e_1}, x_{e_2} and x_j for $j=1,2,\ldots,p$, $j \neq e_1, e_2$.

6) Evaluate the likelihood ratio statistic and its corresponding p-value by $G_j^{(2)} = 2(L_{e_1,e_2,j}^{(2)} - L_{e_1,e_2}^{(1)})$, and $p_j^{(2)} = \Pr(\chi_\nu^2 > G_j^{(2)})$ for $j=1,2,\ldots,p$, $j \neq e_1, e_2$.

7) Denote $p_{e_3}^{(2)} = \min(p_j^{(2)})$.

a)If $p_{e_3}^{(2)} < p_E$, enter variable x_{e_3} into the model.

b) Otherwise, stop the procedure.

Step (3)

Continue the cycle backward elimination followed by forward selection identical to the procedure in Step 2 until the last step.

Step (F)

There are possibly a few scenarios.

a)All variables have entered the model.

b) All variables in the model have p-values that are less than p_R to remove, and the variables not included in the model have p-values that are larger than p_E to enter.

The variables at the Step F are only important relative to criterias of p_E and p_R. The final model may or may not the best model. It depends on the researcher and the status of data.

Disadvantege of this procedure is that the maximum likelihood estimates for the coefficients of all variables not in the model must be calculated at each step. For large data sets, this is quite costly both in terms of time and money (Hosmer and Lemeshow, 1989).

6.3 Goodness of Fit Test

After fitting the logistic regression model, it is useful to test its effectiveness by using goodness of fit tests. Here, the null hypothesis is that the model of interest fits well. The observed values of the outcome variable in vector form is denoted as y where $y^* = (y_1, y_2, \ldots, y_n)$ and the fitted values of the outcome variable in vector form as $\hat{y}$ where $\hat{y}^* = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)$. $(y_i - \hat{y})$ is defined to be residual and its value must be small $(i=1,2,\ldots,n)$.

6.3.1 The Hosmer-Lemeshow Test

The aim of the Hosmer-Lemeshow test is to make a group of the values of the estimated probabilities. 10 groups are created $(g=10)$. The first group contains $n_1^* = n/10$ subjects having the smallest estimated probabilities. The last group contains $n_{10}^* = n/10$ subjects having the largest estimated probabilities. The each group's n_k^* equals to $n/10$ $(k=1,2,\ldots,10)$. For the $y=1$ row, the estimates of the expected values are found by summing the estimated probabilities over all subjects in a group. For $y=0$ row, the estimates of the expected values are found by subtracting from 1 (1-the estimated probabilities over all subjects in a

group). The Hosmer-Lemeshow goodness of fit statistic is denoted by $\hat{C}$ and it is evaluated as follows:

$\hat{C} = \sum_{k=1}^{g} \frac{(o_k - n_k^* \bar{\pi}_k)^2}{n_k^* \bar{\pi}_k (1-\bar{\pi}_k)}$ where n_k^* is the number of covariate patterns in the k[th] group. $o_k = \sum_{j=1}^{n_k^*} y_j$ where o_k is the number of responses among n_k^* covariate patterns. In addition, $\bar{\pi}_k$ is the average estimated probability and it is calculated as $\bar{\pi}_k = \sum_{j=1}^{n_k^*} \frac{m_j \hat{\pi}_j}{n_k^*}$. The distribution of the statistic $\hat{C}$ is well approximated by the chi-square distribution with $g-2$ degrees of freedom,

7. APPLICATION

This study contains 1200 patients and these data include the statement of the absence or presence of lung cancer. Response variable is observed into two categories. The number of patients who have lung cancer (Ca) is 600. The reference group is the control group (Co) that patients in this group do not have lung cancer. The data set was obtained from Ege University Faculty of Medicine Department of Chest Diseases in İzmir. There are seven independent variables. These are sex (SEX), education (EDU), age (AGE), years of smoking (YOS), age of initial smoking (AOIS), number of packages per year (NOPPY) and duration of giving up smoking (DOGUS), respectively. They are illustrated in Table 1 in Appendix.

7.1 The Univariate Analysis

The application of the logistic regression model is started with a univariate analysis of each variable by using SPSS. This analysis will be used for setting multivariate models after finding candidates with univariate analysis. The candidate variables with using these informations are decided easily. If the p-value of the variable is less than 0.25, then this variable is found to be significant. Otherwise, this variable is not significant and it is excluded from the model. This situation is not seen in these data. For this reason, all of the variables are found to be significant. This is shown in Table 2 in Appendix.

Under the null hypothesis, the slope coefficients are zero. If we select the p-value as 0.10, then the variable SEX is excluded from the model. The multivariate logistic regression analysis will be done by using the variables found to be significant in the univariate case. The results of fitting this model are given in Table 3 in Appendix.

On the basis of the output displayed in Table 3 in Appendix, it appears that all of the variables except for AOIS and DOGUS demonstrate considerable importance in the multivariate model. Here, p-values of both of them are greater than 0.05. These p-values are denoted by 0.177 and 0.524. For this reason, these variables should be investigated. If the Wald statistic values are greater than 2, then the variable is significant. Here, the Wald statistic values of both of them are less than 2. They are denoted by 1.822 and 0.406. For this reason, they are not found significant. First of all, a first model which does not contain the variable AOIS is fitted and the secod model which does not contain the variable DOGUS is fitted. We can see that they are significant according to G statistic. The likelihood ratio test for the variable AOIS yields a value of $G = [1260.978 - 1252.142] = 8.836$ and the likelihood ratio test for the variable DOGUS yields a value of $G = [1282.485 - 1252.142] = 30.343$. Comparing these values to a chi-square distribution with 3 degrees of freedom $((\nu_{full} - \nu_{reduced}) = 19 - 16 = 3)$ yields a value of 7.81 $(\chi^2_{3,0.95} = 7.81)$. Here, these values are greater than 7.81. For this reason they are found significant.

After the model is complicated, the examination of the variable AGE that has been modeled as continuous to obtain the correct scale in the logit will be needed. To examine this situation, three design variables based on the quartiles of AGE are formed and they are replaced as variable AGE (continuous) in the model. If the variable AGE is as linear in the logit, then it is expected to show either a linear incresing or decreasing trend in the estimated coefficient. But, the statistical evidence of linearity for variable AGE is not obtained. For this reason, the statement that the variable AGE is not linear in the logit is supported. This variable is used as continuous.

After these processes, the final model is accepted in Table 3 in Appendix. The logit function of this model is expressed as follows:

$$\hat{g}(x) = \beta_0 + \beta_{11}D_{11} + \beta_{21}D_{21} + \beta_{22}D_{22} + \beta_{23}D_{23} + \beta_3 x_3 + \beta_{41}D_{41} + \beta_{42}D_{42} + \beta_{43}D_{43} + \beta_{44}D_{44} + \beta_{51}D_{51} + \beta_{52}D_{52} + \beta_{53}D_{53} + \beta_{61}D_{61} + \beta_{62}D_{62} + \beta_{63}D_{63} + \beta_{71}D_{71} + \beta_{72}D_{72} + \beta_{73}D_{73}$$

$$\hat{g}(x) = -7.960 + 1.892D_{11} + 1.557D_{21} + 1.660D_{22} + 1.576D_{23} + 0.060x_3 + 2.605D_{41} + 3.054D_{42} + 2.237D_{43} + 1.857D_{44} + 0.711D_{51} + 0.437D_{52} + 0.289D_{53} - 1.907D_{61} - 1.805D_{62} - 1{,}524D_{63} + 1.374D_{71} + 1.059D_{72} + 0.246D_{73}$$

For example, we can calculate the probability of being Ca of any person with respect to his characteristic features. Some special features are shown as follows: SEX: woman, EDU: primary, AGE: 50 years old, YOS: 25 years, AOIS: 25 years old, NOPPY: 35 packages, DOGUS: smoker. According to these features, logit function and logistic regression function are evaluated as follows:

$$\hat{g}(x) = -7.960 + 1.892 + 1.660*1 + 0.060*50 + 3.054*1 + 0.289*1 - 1.805*1 + 1.374*1 = 1.504$$

$$\hat{\pi}(x) = \frac{\exp(\hat{g}(x))}{1+\exp(\hat{g}(x))} = 0.82$$

If logistic regression function value is greater than 0.50, then we conclude that patient is being lung cancer.

According to odds ratio values, being a female has 6.631 times more risk factor than being a male. Illiterates, people graduated from primary school and people graduated from secondary school have respectively 4.747 times, 5.258 times, 4.834 times more risk of being lung cancer with respect to reference group. For AGE variable, a one unit increase in age raises the probability of having lung cancer by 0.06 or 6%. For YOS variable, one unit increase in year of smoking rises risk of having lung cancer with respect to non-smokers. But this rise is more until 30 years of smoking (in categories 1 and 2) and less after 30 years of smoking (in categories 3 and 4). For AOIS variable, an increase in age of initial smoking decreases the risk of having lung cancer. In other words, smokers, who are less than 11 age of initial smoking, in category 1 have more risk of having lung cancer with respect to smokers in category 2 and 3. This situation can be seen from decrease of odds ratio from 2.037 to 1.335. For NOPPY variable, it can not be determined that the increase in number of packages per year rises risk of having lung cancer with respect to non-smokers. This can be shown in odds ratio values of being very similar related to each other. For DOGUS variable, smokers have more risk of having lung cancer with respect to non-smokers.

7.2 The Stepwise Analysis

Most of the statistical software packages contain of the stepwise analysis method. In this study, SPSS statistical software will be used to build a model. Here, two sub-methods will be used. One of them is the forward selection and the other is the backward elimination. Finally, these two methods will be compared. p_E: The probability value for enter to a model, p_R: The probability value for removal from a model, j: The number of independent variables. $j = 1, 2, \ldots, p$.

7.2.1 The Forward Selection

Forward selection procedure is applied to the data. The results of this process are explained in section 3.2. The program is run by using $p_E = 0.15$ and $p_R = 0.20$. The final model is represented in Table 4 in Appendix. The logit function of this model is expressed as follows:

$$\hat{g}(x) = \beta_0 + \beta_{11}D_{11} + \beta_{21}D_{21} + \beta_{22}D_{22} + \beta_{23}D_{23} + \beta_3 x_3 + \beta_{41}D_{41} + \beta_{42}D_{42} + \beta_{43}D_{43} + \beta_{44}D_{44} + \beta_{51}D_{51} + \beta_{52}D_{52} + \beta_{53}D_{53}$$

$$\hat{g}(x) = -6.659 + 1.852D_{11} + 1.597D_{21} + 1.663D_{22} + 1.595D_{23} + 0.039x_3 + 0.955D_{41} + 1.028D_{42} + 1.738D_{43} + 2.686D_{44} + 0.998D_{51} + 0.710D_{52} + 0.004D_{53}$$

For example, we can calculate the probability of being Ca of any person with respect to his characteristic features. Some special features are shown as follows: SEX: woman, EDU: primary, AGE: 50 years old, NOPPY: 35 packages, DOGUS: smoker. According to these features, logit function and logistic regression function are evaluated as follows:

$$\hat{g}(x) = -6.659 + 1.852 * 1 + 1.663 * 1 + 0.039 * 50 + 2.686 * 1 + 0.998 * 1 = 2.49$$

$$\hat{\pi}(x) = \frac{\exp(\hat{g}(x))}{1+\exp(\hat{g}(x))} = 0.92$$

The probability of having lung cancer is so high according to these features. Because 0.92 value is greater than 0.50 value.

SEX is the important risk factor for patients with lung Ca. For SEX variable, being a female has 6.372 times more risk factor than being a male. Illiterates, people graduated from primary school and people graduated from secondary school have respectively 4.940 times, 5.383 times and 4.928 times more risk of having cancer with respect to reference group. For AGE variable, a one unit increase in age raises the probability of having lung cancer by 0.04 or 4%. For NOPPY variable, when the number of packages of cigarattes consumption per year increases, the risk of having lung cancer also increases with respect to non-smokers. For DOGUS variable, smokers have more risk of having lung cancer. Smokers have 2.713 times more risk of having lung cancer with respect to non-smokers.

7.2.2 The Backward Elimination

The result of the backward elimination method is the same as the univariate analysis.

7.3 Goodness of Fit Test

The values of the Hosmer-Lemeshow goodness of fit test statistic computed from the frequencies in Tables 5 and 6 in Appendix are 13.590 and 15.769 and the corresponding p-values computed from the chi-square distribution with 8 degrees of freedom are 0.093 and 0.046, respectively. These calculation are evaluated in below respectively. This indicates that the model obtained from forward selection method seems better than the model obtained from backward elimination method. Here, any computation is made to form risk group that contains 10 subjects. This computation is expressed as $1200/10 = 120$. But, the values in Tables are different from the value of 120. Because, predicted probability values for each subject is listed to ascending from descending order.

For forward selection method;

$$\hat{C} = \frac{(5-4.582)^2}{4.582} + \cdots + \frac{(30-19.865)^2}{19.865} = 13.590,$$
$$\hat{C} = 13.590 < \chi^2_{8,0.05} = 15.507.$$

The final model obtained from forward selection method fits data.

For backward elimination method;

$$\hat{C} = \frac{(5-3.683)^2}{3.683} + \cdots + \frac{(28-17.380)^2}{17.380} = 15.769,$$
$$\hat{C} = 15.769 > \chi^2_{8,0.05} = 15.507$$

For this reason, the final model obtained from backward elimination method does not fit data.

8. CONCLUSION

There are many statistical approaches to predictive probability modeling. In this study, a logistic regression model was investigated. To find "best" model is very important. At the same time, this "best" model should explain the relationship between response and independent variables. This "best" model is found by using variable selection methods.

To describe the application of logistic regression method, it was studied on clinical data to determine important risk factors of being lung cancer. Stepwise logistic regression method was applied to these data with this aim. Some results between forward selection method and backward elimination method varied. For example, being a female has more risk factor than being a male for every two methods. Their risk are almost the same. The value of 6.631 obtained from backward elimination method is greater than the value of 6.372 obtained from forward selection method. The risk of being lung cancer according to education status obtained from forward selection method is almost the same risk of being lung cancer according to education status obtained from backward elimination method. For AGE variable, a one unit increase in age raises the probability of being lung cancerby 0.06 or 6% in backward elimination method. This risk decreases to 4% from 6% in forward selection method. In this phase, forward selection method can be better than backward elimination method. For NOPPY variable, when the number of packets of cigarettes consumption per year increases, there is no evidence the risk about being lung cancer in backward elimination method. But the risk of being lung cancer increases with respect to non-smokers in forward selection method. The values of odds ratio for backward elimination are denoted by 0.149, 0.164, 0.218. The values of odds ratio for forward selection are denoted by 2.598, 2.796, 5.687 and 14.675. Here, forward selection method can be better than backward elimination method. In addition, the number of variables in backward elimination method are more than forward selection method. The logistic regression model obtained from forward selection method does not include YOS and AOIS variables. Duration of giving up smoking for patients is important. Giving up smoking early is more advantageous with respect to smokers. This situation is valid for every two methods. Forward selection method is better than backward elimination method with respect to goodness of fit tests.

Finally, the final model of forward selection method is biologically acceptable, this model can be used for determining risk factors. For this reason, the model obtained from forward selection method is called "best" model. Nowadays, the differences between final model and best model are accepted by researchers. Model fitting is based on science, experimentations and statistical methods. They can not be separated from each other.

REFERENCES

Agresti, A. (1996). *An Introduction to Categorical Data Analysis.* John Wiley & Sons, New York, USA.

Akgül, A. (2003). *Tıbbi Araştırmalarda İstatistiksel Analiz Teknikleri*. Emek Ofset, Ankara.

Cristensen, R. (1997). *Log-Linear Models and Logistic Regression.* 2nd edition, Springer-Verlag, New York.

Dobson, A.J. (1990) *An Intoduction to Generalized Linear Models*, Chapman & Hall, London.

Freund, R. J. and Wilson, W.J (1998). *Regression Analysis (Statistical modeling of a response variable)*, New York, Academic Press.

Grouven, U.. and Bender, R. (1998). Using binary logistic regression models for ordinary data with non-proportional odds, *J. Clin. Epidemiol*, 51, 809-816.

Hosmer, D. and Lemeshow, S. (1989). *Applied Logistic Regression*. John Wiley & Sons, Canada.

Kleinbaum, D. (1994). *Logistic Regression : A Self Learning Text*, 2nd edition, Springer-Verlag, New York.

Menard S. (2001). *Applied Logistic Regression Analysis*, 2nd edition, Sage Publications.

Mendenhall, W. and Sincich, T. (1996). *A Second Course in Statistics*, 5th edition, Prentice Hall, New Jersey.

Neter, J., Kutner, M.H., Nachsheim, C.J. and Wasserman, W. (1996). *Applied Linear Regression Methods*; 4th edition, The McGraw-Hill Irwin, The United State of America.

O'byrne, K.K. and Haddock, C.K. and (2002). Parenting style and adolescent smoking. *Journal of Adolescent Smoking* 30, 418-425.

Rao, J.N.K and Scott, A.J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *Annals of Statistics*. 12, 46-60.

Ryan, P.T. (1997). *Modern Regression Methods*, John Wiley & Sons, Canada.

Steyerberg, E.W., Eijkemans, M. J. and Habbema, J.F. (1999). Stepwise selection in small data sets: A simulation study of bias in logistic regression analysis. *J. Clin. Epidemiol*. 52, 935-942.

Tatlıdil, H. (1996). *Uygulamalı Çok Değişkenli İstatistiksel Analiz*, Cem Ofset, Ankara.

Vupa, Ö. (2004). *Model Building of Logistic Regression Models*, Msc. Thesis, Dokuz Eylul University, Faculty of Arts and Sciences, Statistics Department, İzmir.

WEB_1. (2003) www.maxwell.syr.edu. How do I interpret odds ratios in logistic regression?

WEB_2. (2003) www.Sfsu.edu. Logistic regression

WEB_3. (2003) www.spss.com. Logistic regression.

WEB_4. (2003) www.stata.com. Logistic regression, part III: Hypothesis testing comparisons to OLS.

WEB_5. (2003) www.maxwell.syr.edu. Logistic Regression and Odds Ratio.

Özgül VUPA, was born in 1978 in İzmir. She completed her high school education in İzmir. She took her Bachelor degree from Department of Statistics in Faculty of Arts and Sciences of Dokuz Eylül University in 2001. She took her Master of Science degree from Graduate School of Natural and Applied Sciences of Dokuz Eylül University in 2004. She has been working in Department of Statistics in Faculty of Arts and Sciences of Dokuz Eylül University since 2002. She has started her Postgraduate education in the same year. She has papers and studies concerning Regression Models, Applies Statistics and Biostatistics areas.

Cengiz ÇELİKOĞLU, took his Bachelor degree from Ege University Faculty of Sciences, Department of Mathematics in 1980, his Master of Science degree from Ege University Institute of Applied Analysis Science in Applied Statistics area in 1983, and his Postgraduate degree from Graduate School of Natural and Applied Sciences of Anadolu University, Industrial Engineering area in 1989. He worked in Industrial Engineering department in Engineering Faculty of Anadolu University, between the years 1981-1989. After that, he worked in department of Statistics in Faculty of Sciences of Ege University between the years 1990-1997. He has been working in Department of Statistics in Faculty of Arts and Sciences of Dokuz Eylül University since 1998. Cengiz ÇELİKOĞLU who took the title of Associate Professor in 2005 has papers and studies concerning Applies Statistics, Industrial Engineering and Operations Research areas.

Appendix

Table 1. Categorical Variable Coding

		1	2	3	4
SEX	Male (0)	0.000			
	Female (1)	1.000			
EDU	Illiterate(1)	1.000	0.000	0.000	
	Primary (2)	0.000	1.000	0.000	
	Secondary (3)	0.000	0.000	1.000	
	High+Unv. (0)	0.000	0.000	0.000	
YOS	Non-Smoker (0)	0.000	0.000	0.000	0.000
	<=20 (1)	1.000	0.000	0.000	0.000
	21-30 (2)	0.000	1.000	0.000	0.000
	31-40 (3)	0.000	0.000	1.000	0.000
	>40 (4)	0.000	0.000	0.000	1.000
AOIS	Non-Smoker (0)	0.000	0.000	0.000	0.000
	<=10 (1)	1.000	0.000	0.000	0.000
	11-15 (2)	0.000	1.000	0.000	0.000
	16-19 (3)	0.000	0.000	1.000	0.000
	=>20 (4)	0.000	0.000	0.000	1.000
NOPPY	Non-Smoker (0)	0.000	0.000	0.000	0.000
	01-10 (1)	1.000	0.000	0.000	0.000
	11-20 (2)	0.000	1.000	0.000	0.000
	21-30 (3)	0.000	0.000	1.000	0.000
	>30 (4)	0.000	0.000	0.000	1.000
DOGUS	Smoker (1)	1.000	0.000	0.000	0.000
	01-05 (2)	0.000	1.000	0.000	0.000
	06-11 (3)	0.000	0.000	1.000	0.000
	=>11 (4)	0.000	0.000	0.000	1.000
	Non-Smoker (0)	0.000	0.000	0.000	0.000

Table 2. Univariate Logistic Regression Models for Case to Have or Don't Have Ca

Variable	$\hat{\beta}$		**Wald**	**df**	**p-value**	**Exp** $(\hat{\beta})$	**CI for Exp** $(\hat{\beta})$		**G**	**p-value**
							Lower	**Upper**		
SEX (1)	-0.334	75	**1.480**	1	**0.224** *	0.716	**0.418**	**1.227**	1.498	**0.221**
EDU			**37.420**	3	**0.000** *	8.958			53.819	**0.000**
1	2.193	0.391	**31.477**	1	0.000	8.127	**4.165**	**19.270**		
2	2.095	0.384	**29.746**	1	0.001	4.448	**3.828**	**17.256**		
3	1.492	0.444	**11.302**	1	0.000	1.050	**1.863**	**10.617**		
AGE	0.049	0.006	**59.434**	1	**0.000 ***	1.050	**1.037**	**1.063**	65.135	**0.000**
YOS			**196.073**	4	**0.000 ***				273.580	**0.000**
1	0.789	0.357	**4.879**	1	0.027	2.201	**1.093**	**4.432**		
2	1.876	0.261	**51.600**	1	0.000	6.527	**3.912**	**10.888**		
3	2.657	0.250	**112.588**	1	0.000	14.258	**8.727**	**23.293**		
4	3.001	0.247	**147.959**	1	0.000	20.110	**12.399**	**32.616**		
AOIS			**134.226**	4	**0.000** *				202.272	**0.000**
1	3.043	0.300	**103.108**	1	0.000	20.978	**11.658**	**37.748**		
2	2.666	0.248	**115.263**	1	0.000	14.375	**8.831**	**23.385**		
3	2.405	0.267	**81.264**	1	0.000	11.081	**6.568**	**18.693**		
4	2.167	0.245	**78.261**	1	0.000	8.731	**5.402**	**14.111**		
NOPPY			**231.148**	4	**0.000** *				312.626	**0.000**
1	0.914	0.408	**5.016**	1	0.025	2.495	**1.121**	**5.552**		
2	0.890	0.347	**6.588**	1	0.010	2.436	**1.234**	**4.808**		
3	1.780	0.266	**44.721**	1	0.000	5.928	**3.519**	**9.987**		
4	2.989	0.236	**160.060**	1	0.000	19.861	**12.500**	**31.556**		
DOGUS			**138.993**	4	**0.000** *				206.137	**0.000**
1	2.635	0.234	**126.814**	1	0.000	13.943	**8.814**	**22.056**		
2	2.499	0.286	**76.389**	1	0.000	12.176	**6.951**	**21.326**		
3	1.951	0.346	**31.895**	1	0.000	7.038	**3.576**	**13.853**		
4	1.689	0.300	**92.992**	1	0.000	5.414	**3.009**	**9.740**		

Table 3. Multivariate Model Containing Variables Identified in the Univariate Analysis

Variable	$\hat{\beta}$	**SE**	**Wald**	**df**	**p-value**	**Exp** $(\hat{\beta})$	**CI forExp** $(\hat{\beta})$		**G**	**p-value**
							Lower	**Upper**		
SEX(1)	1.892	0.407	**21.568**	1	**0.000**	6.631	2.984	14.735	411.411	**0.000**
EDU			**15.255**	3	**0.002**					
1	1.557	0.439	12.566	1	0.000	4.747	2.006	11.230		
2	1.660	0.426	15.194	1	0.000	5.258	2.282	12.112		
3	1.576	0.500	9.926	1	0.000	4.834	1.814	12.882		
AGE	0.060	0.11	**29.223**	1	0.000	1.062	1.039	1.086		
YOS				4	**0.000**					
1	2.605	0.567	21.071	1	0.000	13.527	4.448	41.132		
2	3.054	0.471	41.979	1	0.000	21.203	8.417	53.413		
3	2.237	0.421	28.232	1	0.000	9.367	4.104	21.378		
4	1.857	0.465	15.918	1	0.000	6.402	2.572	15.939		
AOIS			**8.755**	3	**0.033**					
1	0.711	0.272	6.864	1	0.009	2.037	1.196	3.469		
2	0.437	0.187	5.464	1	0.019	1.549	1.073	2.235		
3	0.289	0.214	**1.822 ***	1	**0.177 ***	1.335	**0.877**	**2.032**		
NOPPY			**39.350**	3	**0.000**					
1	-1.907	0.492	15.028	1	0.000	0.149	0.057	0.390		
2	-1.805	0.366	24.382	1	0.000	0.164	0.080	0.337		
3	-1.524	0.301	25.609	1	0.000	0.218	0.121	0.393		
DOGUS			**29.468**	3	**0.000**					
1	1.374	0.293	21.918	1	0.000	3.949	2.222	7.019		
2	1.059	0.331	10.273	1	0.000	2.884	1.509	5.513		
3	0.246	0.386	**0.406 ***	1	**0.524 ***	1.279	**0.600**	**2.723**		
Constant	-7.960	0.853	**87.070**	1	**0.000**	0.000				
-2LL=1252.142										

Table 4. Variables in the Model (Constant, NOPPY, SEX, EDU, AGE, DOGUS)

Variables	$\hat{\beta}$	**S.E.**	**Wald**	**df**	**p-value**	**Exp** $(\hat{\beta})$	**CI for Exp** $(\hat{\beta})$	
							Lower	**Upper**
SEX (1)	1.852	0.405	**20.903**	1	**0.000**	**6.372**	2.881	14.095
EDU			**15.814**	3	**0.001**			
1	1.597	0.436	13.429	1	0.000	**4.940**	2.102	11.607
2	1.663	0.423	15.801	1	0.000	**5.383**	2.347	12.344
3	1.595	0.499	10.228	1	0.001	**4.928**	1.854	13.098
AGE	0.039	0.039	**21.177**	1	**0.000**	1.040	1.023	1.058
NOPPY			**83.099**	4	**0.000**			
1	0.955	0.489	3.808	1	0.051	**2.598**	0.996	6.776
2	1.028	0.459	5.009	1	0.025	**2.796**	1.136	6.878
3	1.738	0.397	19.127	1	0.000	**5.687**	2.610	12.392
4	2.686	0.392	47.056	1	0.000	**14.675**	6.812	31.614
DOGUS			**22.771**	3	**0.000**			
1	0.998	0.255	15.271	1	0.000	**2.713**	1.644	4.474
2	0.710	0.305	5.438	1	0.020	**2.034**	1.120	3.695
3	0.004	0.372	0.000	1	0.992	**1.004**	0.484	2.081
Constant	-6.659	0.715	86.624	1	0.000	0.001		
-2LL= 1268.458								

Table 5.Observed and Estimated Expected Frequencies (Forward Selection)

Y		1	2	3	4	5	6	7	8	9	10	Total
Y=1	Obs	5	11	32	46	69	76	83	101	96	81	600
	Exp	4.582	13.405	32.445	47.027	63.935	77.606	86.562	90.118	93.195	91.135	
Y=0	Obs	115	109	88	75	51	44	41	22	25	30	600
	Exp	115.418	106.597	87.555	73.973	56.065	42.394	37.438	32.882	27.805	19.865	
Total		120	120	120	121	120	120	124	123	121	111	1200

Table 6. Observed and Estimated Expected Frequencies (Backward Elimination)

Y		1	2	3	4	5	6	7	8	9	10	Total
Y=1	Obs	5	9	33	47	62	75	87	96	100	86	600
	Exp	3.683	13.282	31.712	47.370	64.467	78.636	82.614	88.191	93.437	96.620	
Y=0	Obs	115	111	87	75	58	47	33	25	21	28	600
	Exp	116.317	106.718	88.288	74.630	55.533	43.364	37.386	32.809	27.563	17.380	
Total		120	120	120	122	120	122	120	121	121	114	1200