

150836

**MODEL BUILDING OF LOGISTIC REGRESSION
MODELS**

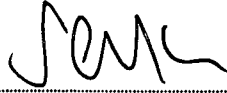
**A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of
Dokuz Eylül University
In Partial Fulfilment of the Requirements for
the Degree of Master of Science in Statistics**

**by
Özgül VUPA**

**January, 2004
İZMİR**

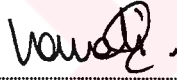
M.Sc THESIS EXAMINATION RESULT FORM

We certify that we have read this thesis, entitled “**MODEL BUILDING OF LOGISTIC REGRESSION MODELS**” completed by Özgül VUPA under supervision of Prof. Dr. Serdar KURT and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.



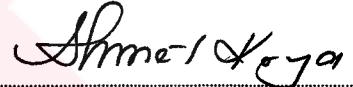
Prof. Dr Serdar KURT

Supervisor



Yrd. Doç. Dr. Hamdi EMEÇ

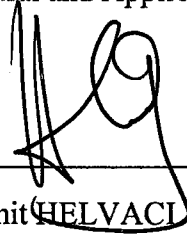
(Committee Member)



Yrd. Doç. Dr. Ahmet KAYA

(Committee Member)

Approved by the
Graduate School of Natural and Applied Sciences



Prof. Dr. Cahit HELVACI

Director

ACKNOWLEDGMENTS

I wish to express my sincere gratitude to my supervisor Prof. Dr. Serdar KURT for his guidance throughout the course of this work.

I am grateful to Research Ass. Hatice ULUER and Research Ass. Selma GÜRLER for their continual encouragement and support all throughout the work.

I also wish to express my deepest gratitude to my family for their encouragement and support during my studies.

Özgül VUPA

ABSTRACT

Logistic regression analysis is the most popular regression techniques available for modeling dichotomous dependent variables. Logistic regression is a mathematical modeling approach that is used to describe the relationship of several predictor (independent) variables X_1, X_2, \dots, X_p to a dichotomous dependent variable Y , where Y is typically coded as 0 or 1 for its two possible categories. Here, a set of independent variables may be continuous, discrete, dichotomous (binary) or a mixture of any of these.

In this study, a logistic regression model is concerned. In other words, this study investigates the simple and multiple logistic regression model forms and several of their key features, particularly how an odds ratio can be estimated from them. 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 (Forward, Backward), Odds Ratio, Likelihood Ratio Test (G).

ÖZET

Lojistik regresyon analizi ikili bağımlı değişkenleri modellemek için uygulanabilen en popüler regresyon tekniklerinden biridir. Lojistik regresyon X_1, X_2, \dots, X_p gibi bağımsız değişkenleri ile iki olası kategori için 0 veya 1 gibi kodlanmış Y ikili bağımlı değişkeni arasındaki ilişkiyi tanımlamakta kullanılan matematiksel modelleme yaklaşımıdır. Burada bağımsız değişkenler seti sürekli, kesikli, ikili veya bunların karışımı olabilir.

Bu çalışmada, lojistik regresyon ile ilgilenilmektedir. Başka bir deyişle, bu çalışmada basit ve çoklu lojistik regresyon model yapıları ve onların bazı anahtar özellikleri özellikle de odds oranının bu modellerden nasıl hesaplanabildiği araştırılmaktadır. 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 (İleriye Doğru, Geriye Doğru), Odds Oranı, Olabilirlik Oran Testi (G).

CONTENTS

	Page
Contents.....	vii
List of Tables.....	x
List of Figures.....	xii

Chapter One INTRODUCTION

1.1 Introduction	1
------------------------	---

Chapter Two

GENERAL INFORMATION ON LOGISTIC REGRESSION

2.1 Regression Models with Binary Response Variable	3
2.2 Meaning of Response Function When Response Variable is Binary.....	3
2.3 Special Problems When Response Variable is Binary.....	5
2.4 Introduction to Logistic Response Function.....	7
2.5 Fitting of Simple Logistic Regression Model.....	8
2.5.1 Likelihood Function.....	9
2.5.2 Maximum Likelihood Estimation Method.....	11
2.5.3 Testing for the Significance of the Coefficient.....	11

2.5.3.1	Likelihood Ratio Test.....	12
2.5.3.2	Wald Test.....	15
2.5.3.3	Score Test.....	17
2.6	Fitting of Multiple Logistic Regression Model.....	17
2.6.1	The Design Variable.....	20
2.6.2	Testing for the Significance of the Model.....	21
2.6.2.1	Likelihood Ratio Test.....	21
2.6.2.2	Wald Test.....	23
2.6.2.3	Score Test.....	25
2.7	Interpretation of the Coefficients	25
2.7.1	Dichotomous Independent Variable.....	26
2.7.2	Polytomous Independent Variable.....	29
2.7.3	Continuous Independent Variable.....	31
2.7.4	Multivariate Case.....	32

Chapter Three

MODEL BUILDING STRATEGIES

3.1	Model Building Procedures	35
3.1.1	The Univariate Analysis.....	36
3.1.2	The Stepwise Logistic Regression Method.....	40
3.1.3	The Best Subsets Logistic Regression Method.....	46
3.2	Goodness of Fit Tests	46
3.2.1	Pearson Chi-Square Test and Deviance Test.....	47
3.2.2	The Hosmer-Lemeshow Test	49

Chapter Four

APPLICATION

4.1	General Information on the Data	51
4.2	The Univariate Analysis.....	54

4.3	The Stepwise Analysis.....	62
4.3.1	The Forward Selection.....	62
4.3.2	The Backward Elimination.....	75
4.4	Goodness of Fit Test.....	77

Chapter Five

Conclusion

5.1	Conclusion.....	80
5.2	Further Research.....	82
	References.....	83



LIST OF TABLES

		Page
Table 2.1	The Probability Distribution.....	4
Table 2.2	The Coding of the Design Variables for Race.....	21
Table 2.3	Values of the Logistic Regression Model When the Independent Variable is Dichotomous.....	26
Table 2.4	The Coding of the Design Variables for Independent Variables Coded Four Levels.....	29
Table 2.5	Cross-Classification of Independent Variables and Status for 95 Subjects.....	30
Table 2.6	Descriptive Statistics for Two Groups of 70.....	33
Table 2.7	Results of Fitting the Logistic Regression Model.....	33
Table 3.1	Observed and Estimated Expected Frequencies.....	50
Table 4.1	Categorical Variables Coding.....	52
Table 4.2	Y*SEX Cross Tabulation Count.....	53
Table 4.3	Y*EDU Cross Tabulation Count.....	53
Table 4.4	Y*YOS Cross Tabulation Count.....	53
Table 4.5	Y*AOFS Cross Tabulation Count.....	53
Table 4.6	Y*NOPPY Cross Tabulation Count.....	53
Table 4.7	Y*DOGUS Cross Tabulation Count.....	53
Table 4.8	AGE Situation.....	53
Table 4.9	Univariate Logistic Regression Models for Case to Have or Don't Have Cancer.....	55
Table 4.10	Multivariate Model Containing Variables Identified in the Univariate Analysis.....	56

Table 4.11	Multivariate Model without AOIS.....	57
Table 4.12	Multivariate Model without DOGUS.....	58
Table 4.13	Results of the Quartile Analysis of AGE.....	59
Table 4.14	Multivariate Model of Linearity for AGE.....	59
Table 4.15.a	Variables in the Model (Only Constant).....	63
Table 4.15.b	Variables not in the Model (SEX, EDU, AGE, YOS, AOIS, NOPPY, DOGUS).....	63
Table 4.16.a	Variables in the Model (Constant, NOPPY).....	64
Table 4.16.b	Variables not in the Model (SEX, EDU, AGE, YOS, AOIS, DOGUS).....	65
Table 4.17.a	Variables in the Model (Constant, NOPPY, SEX).....	66
Table 4.17.b	Variables not in the Model (EDU, AGE, YOS, AOIS, DOGUS).....	66
Table 4.18.a	Variables in the Model (Constant, NOPPY, SEX, EDU).....	67
Table 4.18.b	Variables not in the Model (AGE, YOS, AOIS, DOGUS).....	68
Table 4.19.a	Variables in the Model (Constant, NOPPY, SEX, EDU,AGE)	69
Table 4.19.b	Variables not in The Model (YOS, AOIS, DOGUS).....	69
Table 4.20.a	Variables in the Model (Constant, NOPPY, SEX, EDU, AGE, DOGUS).....	70
Table 4.20.b	Variables not in the Model (YOS, AOIS).....	71
Table 4.21.a	Variables in the Model (Without DOGUS).....	72
Table 4.21.b	Variables not in the Model (Without DOGUS).....	72
Table 4.22.a	Variables in the Multivariate Model of Linearity for AGE.....	73
Table 4.22.b	Variables not in the Multivariate Model of Linearity for AGE	73
Table 4.23.a	Variables in the Model (Only Constant).....	76
Table 4.23.b	Variables not in the Model (SEX, EDU, AGE, YOS, AOIS, NOPPY, DOGUS).....	76
Table 4.24	Observed and Estimated Expected Frequencies (Forward Selection).....	78
Table 4.25	Observed and Estimated Expected (Backward Elimination).....	78

LIST OF FIGURES

	Page
Figure 2.1 Logistic Regression Function.....	5



CHAPTER ONE

INTRODUCTION

There are many statistical approaches in predictive probability modeling. Five popular techniques of predictive modeling are examined in literature. These are the density transfer, the density regression, the significance regression, the discriminant function analysis and the logistic regression analysis, respectively. The most popular of these is known as the logistic regression analysis. It can be said that the use of logistic regression is easier than the others. Because, its assumptions are less constrained. Researchers are often interested in performing regression when the response variable (outcome, dependent) is categorical. The logistic regression is used when the response variable has only two categories. When the response variable has more than two categories, methods that extend the technique of logistic regression are available. The choice of method depends on whether the response variable is measured on an ordinal or nominal scale.

A nominal scale has categories that are not ordered. For example, an education researcher may be interested in a nominal response variable such as what each student decides to do after high school: attend collage, find a job, join the military or something else. Logistic regression for a nominal response variable is called nominal, multinomial, or polytomous logistic regression.

An ordinal scale has categories that are ordered. Letter grades for a particular class (A, B, C, or D) and academic tracks (remedical, regular, or advanced placement) are examples of ordinal variables.

In logistic regression, the odds ratio values of variables are examined to interpret of the coefficients. This procedure is so easy in logistic model.

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 stepwise logistic regression, selection or elimination of variables from the model is based on a statistical algorithm that checks for “importance” of variables, and either includes or excludes them on the basis of a fixed decision rule. The likelihood ratio (chi-square) test, G , is used to assess significance in logistic regression since the errors are assumed to follow a binomial distribution. This test assigns a p-value to each variable to assess significance. Therefore, the most important variable is the one with the smallest p-value. So, the final model with appropriate variables is stated.

After fitting the model, three tests are used to determine the fit of the model. These are the Pearson chi-square test, Deviance test and Hosmer-Lemeshow test.

This study contains five chapters. In chapter one, whole study is summarized shortly. In chapter two, basic features of a logistic regression model is examined. The general informations and the interpretation of coefficients of simple and multiple logistic regression models are given. In chapter three, model building strategies and methods for logistic regression model are given. In addition, some diagnostic measures are used to look at the fit of the model. In chapter four, the univariate analysis and two commonly used variable selection methods are examined and the differences between them and their procedure steps are presented in the tables. In chapter five, the conclusion of this study is given.

CHAPTER TWO

GENERAL INFORMATION ON LOGISTIC REGRESSION

2.1 Regression Models with Binary Response Variable

In a variety of regression applications, the response variable has only two possible outcomes, and it can be represented by a binary indicator variable taking on values 0 and 1. Binary response variables are known as binary responses or dichotomous responses. These response variables are measured on a binary scale. For example, the responses may be alive or dead, or present or absent. These outcomes can be coded 0 and 1, respectively. (Freund and Wilson, 1996)

2.2 Meaning of Response Function When Response Variable is Binary

The simple linear regression model is written as:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, 2, \dots, n \quad (2.1)$$

where the outcome Y_i is binary taking on the value of either 0 or 1. The expected response $E\{Y_i\}$ has a special meaning in this case. Since $E\{\varepsilon_i\} = 0$, it can be written as:

$$E\{Y_i\} = \beta_0 + \beta_1 X_i \quad (2.2)$$

When Y_i is a Bernoulli random variable, then its the probability distribution can be written as follows:

**Table 2.1 : The Probability
Distribution of Y_i**

Y_i	Probability
1	$P(Y_i = 1) = \pi_i$
0	$P(Y_i = 0) = 1 - \pi_i$

Thus, π_i is the probability that $Y_i=1$ and $(1 - \pi_i)$ is the probability that $Y_i=0$.
Expected value of a Bernoulli random variable is

$$E\{Y_i\} = 1(\pi_i) + 0(1 - \pi_i) = \pi_i \quad (2.3)$$

So, $E\{Y_i\}$ is found as follows:

$$E\{Y_i\} = \beta_0 + \beta_1 X_i = \pi_i \quad (2.4)$$

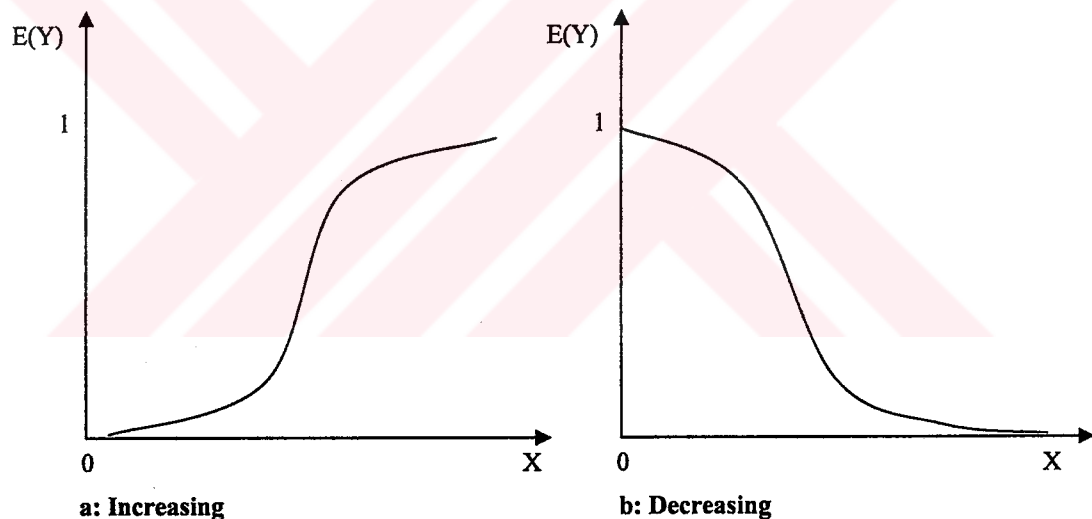
As seen , when the response variable is 0 or 1 indicator variable, the mean response always represents the probability that $Y_i=1$ for the given level of the independent (predictor) variable (X_i). (Neter and Kutner, 1996)

The mean value of the response variable is called the conditional mean and it can be expressed as “ $E(Y|x)$ ”. Here, Y denotes the response variable and x denotes a value of the independent (predictor) variable. $E(Y|x)$ is called the expected value of Y for given a value of $X = x$. In linear regression this mean may be expressed as an equation linear in x , such as

$$E(Y|x) = \beta_0 + \beta_1 X \quad (2.5)$$

This expression can take any value between $-\infty$ and ∞ . Here X , β_0 , β_1 and Y are called independent variable, intercept, regression coefficient and response (dependent, outcome) variable respectively. On the other hand, since the response variable is binary (dichotomous), the mean value of the response variable is greater than or equal to 0 and less than or equal to 1 in the logistic regression. This can be expressed as $0 \leq E(Y|x) \leq 1$. The change in the $E(Y|x)$ per unit change in x becomes smaller as the conditional mean gets closer to 0 or 1. In the logistic regression, the relationship between the independent (predictor) and response variable is not a linear function. This can be seen in Figure 2.1 and this shape of the curve is said to be S shaped and resembles the plot of a cumulative distribution function of a random variable.

Figure 2.1 : Logistic Regression Function (a: increasing, b: decreasing)



2.3 Special Problems When Response Variable is Binary

There is no doubt, there are some special problems when the response variable is binary (dichotomous). 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 $\epsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$ 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 $\pi(x_i)$. On the other hand, 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.

Another problem with the error terms (ε_i) is that they do not have equal variances when the response variable is 0 or 1 indicator variable. The variance of Y_i for the simple linear regression model can be determined as follows: (Neter and Kutner, 1996)

$$V(Y_i) = E(Y_i - E(Y_i))^2 = (1 - \pi_i)^2 \pi_i + (0 - \pi_i)^2 (1 - \pi_i) = \pi_i (1 - \pi_i) \quad (2.6)$$

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. So, it can be determined in a similar manner as follows:

$$V(\varepsilon_i) = E(Y_i)(1 - E(Y_i)) = \pi_i (1 - \pi_i) \quad (2.7)$$

As seen from the equations (2.4) and (2.7), $V(\varepsilon_i)$ depends on x_i . The error variances will differ at different levels of X .

The last problem is related with constraints on response function. Since the response function represents probabilities, the mean responses should be constrained as follows:

$$0 \leq E(Y_i) = \pi_i \leq 1 \quad (2.8)$$

2.4 Introduction to Logistic Response Function

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

$$\pi(x_i) = E(Y|x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \quad (2.9)$$

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

$$g(x_i) = \ln \left[\frac{\pi(x_i)}{1 - \pi(x_i)} \right] = \ln \left[\frac{E(Y=1|x_i)}{E(Y=0|x_i)} \right] \quad (2.10)$$

$$1 - \pi(x_i) = 1 - \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} = \frac{1}{1 + \exp(\beta_0 + \beta_1 x_i)} \quad (2.11)$$

Hence

$$g(x_i) = \ln \left\{ \frac{\exp(\beta_0 + \beta_1 x_i) / (1 + \exp(\beta_0 + \beta_1 x_i))}{1 / (1 + \exp(\beta_0 + \beta_1 x_i))} \right\} = \ln(e^{\beta_0 + \beta_1 x_i}) = \beta_0 + \beta_1 x_i \quad (2.12)$$

Here, the ratio $\frac{\pi(x_i)}{1 - \pi(x_i)}$ in the logit transformation is called odds.

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 .

2.5 Fitting of Simple Logistic Regression Model

There is a sample of n independent observations and it is expressed as (x_i, y_i) . Here y_i denotes the value of a binary response variable and x_i is the value of the independent variable for the i th subject. In simple logistic regression model, there is only one independent variable. The unknown parameters which are β_0 and β_1 for fitting the logistic regression model in the equation (2.9) are estimated.

In linear regression model, the least squares method is used to estimate the unknown parameters (β_0 and β_1). This method is based on minimizing the sum of squared deviations of the observed values of the response variable from the predicted values based upon the model.

In logistic regression model, when the method of the least squares is applied to the model with a dichotomous outcome, the estimators do not have the same properties in linear regression model. (Hosmer and Lemeshow, 1989) The general methods of estimation in logistic regression are investigated in three main concepts. These are given below.

1. The Maximum Likelihood Method.
2. Iteratively Reweighted Least Squares Method.
3. The Minimum Logit Chi-Square Method.

In this study, the maximum likelihood estimation method will be used. Firstly, the likelihood function is constructed in order to apply this method. This likelihood function express the probability of the observed data as a function of the unknown parameters. The maximum likelihood estimates of these parameters are chosen to be those values which maximize likelihood function. As a result of this, the resulting estimators are those which agree most closely with the observed data.

2.5.1 Likelihood Function

The parameters β_0 and β_1 are estimated through the method of maximum likelihood. 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 $1 - \pi(x_i)$. The pairs (x_i, y_i) the contribution to the likelihood function is shown in equation (2.13). In addition, this method isolates the values of parameters that maximize the likelihood function, where the likelihood function $L(\beta_0, \beta_1)$ is defined as the joint probability distribution for all of the data points. Since Y_i 's have a Bernoulli distribution, the probability density function can be defined as:

$$P(Y = y_i) = f_i(y_i) = \pi(x_i)^{y_i} (1 - \pi(x_i))^{1-y_i} \quad (2.13)$$

where $y_i = 0$ or $y_i = 1$ for $i = 1, 2, \dots, n$

Since the observations Y_i are assumed to be independent, the likelihood function can be defined as follows:

$$\begin{aligned} L(\beta_0, \beta_1) &= g(y_1, y_2, \dots, y_n) \\ &= \prod_{i=1}^n f_i(y_i) \\ &= \prod_{i=1}^n \pi(x_i)^{y_i} (1 - \pi(x_i))^{1-y_i} \end{aligned} \quad (2.14)$$

In order to maximize this function, the derivative must be taken with respect to each of the parameters. Then, the resulting 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, being that maximizing the natural log of the function would result in the same value as maximizing the likelihood function itself. Obtaining the log-likelihood function is expressed as:

$$\begin{aligned}
\ln L(\beta_0, \beta_1) &= \ln \prod_{i=1}^n \pi(x_i)^{y_i} (1 - \pi(x_i))^{1-y_i} \\
&= \sum_{i=1}^n \left\{ \ln \left[\pi(x_i)^{y_i} (1 - \pi(x_i))^{1-y_i} \right] \right\} \\
&= \sum_{i=1}^n \left\{ \ln \left[\pi(x_i)^{y_i} \right] + \ln \left[(1 - \pi(x_i))^{1-y_i} \right] \right\} \\
&= \sum_{i=1}^n \left\{ y_i \ln \left[\pi(x_i) \right] + (1 - y_i) \ln \left[1 - \pi(x_i) \right] \right\} \\
&= \sum_{i=1}^n \left\{ y_i \ln \left(\frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right) + (1 - y_i) \ln \left(1 - \left(\frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right) \right) \right\} \\
&= \sum_{i=1}^n \left\{ y_i \left[\beta_0 + \beta_1 x_i - \ln(1 + e^{\beta_0 + \beta_1 x_i}) \right] + (1 - y_i) \left[-\ln(1 + e^{\beta_0 + \beta_1 x_i}) \right] \right\} \\
&= \sum_{i=1}^n \left\{ y_i (\beta_0 + \beta_1 x_i) - \ln(1 + e^{\beta_0 + \beta_1 x_i}) \right\} \tag{2.15}
\end{aligned}$$

Now taking the derivative, first with respect to β_0 and then with respect to β_1 and setting each equal to zero, the following likelihood equations are formed as follows:

$$\begin{aligned}
\frac{\partial \ln L(\beta_0, \beta_1)}{\partial \beta_0} &= \sum_{i=1}^n \left\{ y_i - \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right\} = 0 \\
&= \sum_{i=1}^n [y_i - \pi(x_i)] = \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)] = 0 \tag{2.16}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \ln L(\beta_0, \beta_1)}{\partial \beta_1} &= \sum_{i=1}^n \left\{ y_i x_i - \frac{x_i e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right\} = 0 \\
&= \sum_{i=1}^n x_i [y_i - \pi(x_i)] = \sum_{i=1}^n x_i [y_i - (\beta_0 + \beta_1 x_i)] = 0 \tag{2.17}
\end{aligned}$$

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

2.5.2 Maximum Likelihood Estimation Method

Maximum likelihood estimation method is used to calculate the logit coefficients. The value of $\beta = (\beta_0, \beta_1)$ given by the solution to likelihood equations (2.16) and (2.17) is called the maximum likelihood estimate and is denoted as $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$. $\hat{\pi}(x_i)$ is the maximum likelihood estimate of $\pi(x_i)$ and it estimates the conditional probability that Y_i is equal to 1, given $X = x_i$. In addition, the sum of the observed values of y_i is equal to the sum of the expected values and it is expressed as follows: (Hosmer and Lemeshow, 1989)

$$\sum_{i=1}^n y_i = \sum_{i=1}^n \hat{\pi}(x_i) \quad (2.18)$$

If these estimated values are substituted into the response function (in equation 2.9), the fitted response function is obtained. The fitted value for the i th case is expressed as $\hat{\pi}(x_i)$.

Also, fitted simple logistic response function for the i th case is follows:

$$\hat{\pi}(x_i) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)} \quad (2.19)$$

2.5.3 Testing for the Significance of the Coefficients

After estimating the coefficients, an assessment of significance of the variable in the fitted model is concerned. This involves formulation and testing of statistical hypothesis to determine whether the independent variable in the model is significantly related to the response variable. (Hosmer and Lemeshow, 1989)

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. If the predicted values with the variable in the model are better or clearer, than when the variable is not in the model, then the variable in question is said to be significant. The comparison is based on the log-likelihood. In addition, it is not important question of whether the predicted values that are obtained from saturated model have accurate relation or representation of the observed values of response variable in an absolute sense or not. This is concerned in goodness of fit.

In logistic regression model, there are three commonly used tests for hypothesis testing.

1. Likelihood Ratio Test (G Statistic).
2. Wald Test.
3. Score Test.

2.5.3.1 Likelihood Ratio Test

Comparison of observed to predicted values is based on the log-likelihood function in logistic regression. The model which includes all possible terms (including interactions) is called as saturated model. In addition, a saturated model is one that contains as many parameters as there are data points. The current model is the subset of the saturated model. The current model does not include the variable investigated by the researcher. The likelihood ratio test statistic is -2 times of the difference between the log-likelihoods of saturated and current model. The distribution of the likelihood ratio test statistic is closely approximated by the chi-square distribution for large sample sizes. The degree of freedom (df) of the approximating chi-square distribution is equal to the difference in the number of regression coefficients in the two models. (NCSS)

The comparison of observed to predicted values using the likelihood function is based on the following expression:

$$D = -2 \ln \left[\frac{\text{likelihood of the current model}}{\text{likelihood of the saturated model}} \right] \quad (2.20)$$

$$D = -2 [\ln(\text{likelihood of the current model}) - \ln(\text{likelihood of the saturated model})]$$

This expression is called likelihood ratio. 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. Recalling the log-likelihood function is necessary to obtain deviance statistic. This equation is written as follows:

$$\begin{aligned} \ln L(\beta_0, \beta_1) &= \ln \prod_{i=1}^n \pi(x_i)^{y_i} (1 - \pi(x_i))^{1-y_i} \\ &= \sum_{i=1}^n \{ y_i \ln[\pi(x_i)] + (1 - y_i) \ln[1 - \pi(x_i)] \} \end{aligned} \quad (2.21)$$

This equation can be substituted into the formula for the deviance and then manipulated in order to get the following equations:

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

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

$$D = -2 \sum_{i=1}^n \left[y_i \ln \left(\frac{\hat{\pi}(x_i)}{y_i} \right) + (1 - y_i) \ln \left(\frac{1 - \hat{\pi}(x_i)}{1 - y_i} \right) \right] \quad (2.22)$$

These equations are called the deviance. The last equation is usually used more than the others. The deviance for logistic regression model plays the same role as SSE in linear regression.

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. Therefore, the test statistic, G, is expressed as follows:

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

$$G = -2 \ln \left(\frac{\text{likelihood of current model without } \beta_1}{\text{likelihood of saturated model}} \right)$$

$$+ 2 \ln \left(\frac{\text{likelihood of current model with } \beta_1}{\text{likelihood of saturated model } \beta_1} \right)$$

$$G = -2 \ln \left(\frac{\text{likelihood of current model without } \beta_1}{\text{likelihood of current model with } \beta_1} \right)$$

$$G = -2 \ln(\text{likelihood of current model without } \beta_1) + 2 \ln(\text{likelihood of current model with } \beta_1) \quad (2.23)$$

The likelihood of the saturated model is common to both values of the D. It is different computation of G.

$$G = -2 \ln \left[\frac{\text{likelihood without the variable}}{\text{likelihood with the variable}} \right] \quad (2.24)$$

It is easy to determine the maximum likelihood estimate of β_0 when the single independent variable is not in the model. The maximum likelihood estimate of β_0 is $\ln(n_1/n_0)$ where $n_1 = \sum_{i=1}^n y_i$ and $n_0 = \sum_{i=1}^n (1 - y_i)$. The predicted value is constant and it is expressed as n_1/n . G is written as follows:

$$G = -2 \ln \left[\frac{\left(\frac{n_1}{n}\right)^{n_1} \left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^n \hat{\pi}_i^{y_i} (1 - \hat{\pi}_i)^{(1-y_i)}} \right] \quad (2.25)$$

$$G = 2 \left\{ \sum_{i=1}^n [y_i \ln(\hat{\pi}_i) + (1 - y_i) \ln(1 - \hat{\pi}_i)] - [n_1 \ln(n_1) + n_0 \ln(n_0) - n \ln(n)] \right\} \quad (2.26)$$

In checking the significance of the coefficient, the following null and alternative hypotheses are written as follows:

$$H_0 : \beta_1 = 0 \quad H_1 : \beta_1 \neq 0 \quad (2.27)$$

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. But this independent variable must be found important by the researcher too.

2.5.3.2 Wald Test

The other test for significance of a variable is Wald test. Wald test is based on the comparison 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 diagonal element of the covariance matrix $V(\hat{\beta})$. This test for the logistic regression model is as follows:

$$W = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \quad (2.28)$$

Under the hypothesis that $\beta_1 = 0$, W is a standard normal distribution. Two tailed p-value is evaluated by $P(|Z| > W)$. Here, Z denotes a random variable following the standard normal distribution. If this p-value is less than given α -value, then the null hypothesis is rejected. Generally, this α -value is taken on 0.05. In addition, p-value can be defined as follows:

$$\begin{aligned} \text{p-value} &= P(|Z| > \text{the observed test statistic}) \text{ or} \\ \text{p-value} &= 2P(Z > \text{the observed test statistic}) \end{aligned} \quad (2.29)$$

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 can be written as follows:

$$W^2 = \left(\frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \right)^2 \quad (2.30)$$

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.

For a single variable model, using the Wald test is so easy. But the iterative computation needed to obtain the maximum likelihood estimates can be considerable for large data sets with many variables. (Hosmer and Lemeshow, 1989)

2.5.3.3 Score Test

The score test is the another test for the significance of a coefficient. The most important advantage of this test is to reduce computational effort to the other tests. 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:

$$ST = \frac{\sum_{i=1}^n x_i (y_i - \bar{y})}{\sqrt{\bar{y}(1 - \bar{y}) \sum_{i=1}^n (x_i - \bar{x})^2}} \quad (2.31)$$

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.

2.6 Fitting of Multiple Logistic Regression Model

Here, multiple logistic regression model for the case of more than one independent variable is fitted. Also, this model is called “the multivariate case”.

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

$$\pi(\tilde{x}) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)} = \frac{\exp(g(\tilde{x}))}{1 + \exp(g(\tilde{x}))} \quad (2.32)$$

$$g(\tilde{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p \quad (2.33)$$

There is a sample of n independent observations and it is expressed as (\tilde{x}_i, y_i) . Where y_i denotes the value of a dichotomous response variable and \tilde{x}_i is the value

of the independent variables for the i th subject. As in the univariate case, the maximum likelihood estimates of the parameters are used and it is shown as: $\tilde{\beta} = (\beta_0, \beta_1, \dots, \beta_p)$.

The likelihood function for the multiple logistic regression model is expressed as:

$$L(\tilde{\beta}) = \prod_{i=1}^n \pi(\tilde{x}_i)^{y_i} (1 - \pi(\tilde{x}_i))^{(1-y_i)} \quad (2.34)$$

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. The likelihood equations which result are expressed as follows:

$$\sum_{i=1}^n [y_i - \pi(\tilde{x}_i)] = 0 \quad (2.35)$$

$$\sum_{i=1}^n x_{ij} [y_i - \pi(\tilde{x}_i)] = 0 \quad j=1, 2, \dots, p \quad (2.36)$$

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).

Let $\hat{\beta}$ denote the solution to these likelihood equations. Here, the fitted values for the multiple logistic regression model are the $\hat{\pi}(\tilde{x}_i)$, the value of

$$\hat{\pi}(\tilde{x}_i) = \frac{\exp(\hat{g}(\tilde{x}_i))}{1 + \exp(\hat{g}(\tilde{x}_i))} \quad (2.37)$$

is computed using $\hat{\beta}$ and \tilde{x}_i .

Standard errors for the coefficients, $S\hat{E}(\hat{\beta}_j)$, are given along with the $\hat{\beta}_j$ ($j=1, 2, \dots, p$).

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.

If we let

$$X = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix}_{n \times (p+1)} \quad \text{and} \quad (2.38)$$

$$V = \begin{bmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & 0 & \dots & 0 \\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix}_{n \times n} = \text{diag}[\hat{\pi}_i(1-\hat{\pi}_i)] \quad (2.39)$$

then $\hat{I}(\hat{\beta}) = [X'VX]_{(p+1) \times (p+1)}$ is called information matrix. The variances and covariances of the estimated coefficients are obtained from the inverse of this matrix.

The estimated variance and the confidence interval of the estimated coefficients are denoted as follows:

$$\hat{\text{Var}}(\hat{\beta}) = [X'VX]^{-1} \quad (2.40)$$

$$\hat{\beta}_j \pm Z_{1-\alpha/2} S\hat{E}(\hat{\beta}_j) \quad S\hat{E} = \sqrt{\hat{\text{Var}}} \quad (2.41)$$

The estimated logit is $\hat{g}(\tilde{x}) = \sum_{j=0}^p \hat{\beta}_j x_j$ and the estimate of its variance is $\text{Vâr}[\hat{g}(\tilde{x})] = x' \left\{ \text{Vâr}(\hat{\beta}) \right\} x$. Hence the confidence interval for the logit is evaluated as follows:

$$\hat{g}(\tilde{x}) \pm Z_{1-\alpha/2} \text{SÊ}(\hat{g}(\tilde{x})) \quad (2.42)$$

This is used to obtain a confidence interval for the fitted value or estimated logistic probability as follows:

$$\frac{\exp\left\{\hat{g}(\tilde{x}) \pm Z_{1-\alpha/2} \text{SÊ}(\hat{g}(\tilde{x}))\right\}}{1 + \exp\left\{\hat{g}(\tilde{x}) \pm Z_{1-\alpha/2} \text{SÊ}(\hat{g}(\tilde{x}))\right\}} \quad (2.43)$$

2.6.1 Design Variable

If some of the independent variables are discrete, ordinal or nominal scaled variable (categorical variable) with more than two levels, then the model differs from general formula in equation (2.33). 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. For example, one of the independent variables is race and that is coded as "white", "black" or "other". Here, two design variables are necessary. When the respondent is "white", the two design variables, D_1 and D_2 , would both be set equal to zero; when the respondent is "black", D_1 would be set equal to 1 while D_2 would still equal 0; when the respondent is "other", D_2 would be set equal to 1 while D_1 would still equal 0. (Hosmer and Lemeshow, 1989). It is shown in Table 2.2.

Table 2.2 : The Coding of the Design Variables for Race

Design Variable		
Race	D_1	D_2
White	0	0
Black	1	0
Other	0	1

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} , $m = 1, 2, \dots, k_j - 1$. The logit for a model with p independent variables and the j th independent variable being discrete is expressed as:

$$g(\tilde{x}) = \beta_0 + \beta_1 x_1 + \dots + \sum_{m=1}^{k_j-1} \beta_{jm} D_{jm} + \beta_p x_p \quad (2.44)$$

2.6.2 Testing for the Significance of the Model

The analysis of the test of significance in multiple logistic regression model (multivariate case) is similar to simple logistic regression model (univariate case). Three tests are used for the hypothesis testing. These tests are the same as univariate case.

2.6.2.1 Likelihood Ratio Test

The parameters in the multiple setting are once again determined through maximum likelihood estimation method, because Y is still a Bernoulli random variable with the same probability distribution. In addition, the derivation of the maximum likelihood estimators remains the same, with the exception of the

inclusion of more parameters. Thus, the log-likelihood equation takes the form as follows:

$$\begin{aligned} \ln L(\beta_0, \beta_1, \dots, \beta_p) &= \ln L(\tilde{\beta}) \\ &= \sum_{i=1}^n \{y_i (\beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi}) - \ln(1 + \exp(\beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi}))\} \end{aligned} \quad (2.45)$$

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. *“The only difference is that the fitted values, $\hat{\pi}$, under the model are based on the vector containing p+1 parameters, $\tilde{\beta}$. Under the null hypothesis that p “slope” coefficients for the covariates in the model are equal to zero, the distribution of G will be chi-square with p degrees of freedom.”* (Hosmer, D. W. & Lemeshow S., 1989, *Applied Logistic Regression*, John Wiley & Sons p:31)

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). The test statistic, G, is calculated as follows:

$$G = -2 \ln \left[\frac{\text{likelihood without the variable (s)}}{\text{likelihood with the variable (s)}} \right] \quad (2.46)$$

In addition, G may be calculated in other ways as follows:

$$G = 2 \left\{ \sum_{i=1}^n [y_i \ln(\hat{\pi}_i) + (1 - y_i) \ln(1 - \hat{\pi}_i)] - [n_1 \ln(n_1) + n_0 \ln(n_0) - n \ln(n)] \right\}$$

$$G = 2\{\log - \text{likelihood with the variable(s)} - [n_1 \ln(n_1) + n_0 \ln(n_0) - n \ln(n)]\}$$

$$G = -2\{(\log - \text{likelihood for reduced model}) - (\log - \text{likelihood for full model})\} \quad (2.47)$$

In checking the significance of the model, the following null and alternative hypotheses are written as follows:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$$

$$H_1 : \text{At least one of the } \beta_p \neq 0 \quad (2.48)$$

The statistic G has a chi-square distribution with $(v_2 - v_1)$ degrees of freedom. Here, v_2 equals to the number of variables in the full model +1 and v_1 equals to the number of variables in the reduced model +1. For this test, the decision rule requires that p-value is $P\{\chi^2_{(1-\alpha, df=(v_2-v_1))} > G\}$. If this p-value is less than α -value, H_0 is rejected. This means that the model would be deemed significant. Here, any or all of the coefficients are nonzero. α -value is usually accepted as 0.05. For this reason, p-value is compared with $\alpha = 0.05$ level. On the other hand, 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. In addition, if the statistic G is greater than $\chi^2_{(1-\alpha, df=(v_2-v_1))}$, then H_0 is rejected. The model is accepted as significant.

2.6.2.2 Wald Test

After testing the significance of the model, at least one or perhaps all p coefficients can be different from zero. The Wald test statistics are used to see which variables are significant. These statistics have the standard normal distribution and they are evaluated as follows:

$$W_j = \frac{\hat{\beta}_j}{\widehat{SE}(\hat{\beta}_j)} \sim Z(\alpha/2) \quad (2.49)$$

Under the hypothesis that $\beta_j = 0$, two tailed p – value is evaluated by $P(|Z| > W)$. Here, Z denotes a random variable following the standard normal distribution. If this p-value is less than given α -value, then the null hypothesis is rejected. Generally this α -value is taken as 0.05. For this test, p-value can be defined by equation (2.29).

For multivariate case, Wald test is used in statistical package programs. This W value is then squared, yielding a Wald statistic with a chi-square distribution. However, several authors have identified problems with use of the Wald statistics. Menard warns that for large coefficients, standard error is inflated, lowering the Wald statistic (chi-square) value (Menard, Scott and Stanley Lemeshow, 1996, Applied Logistic Regression Analysis, Sage Publications Series: Quantitative Applications in the Social Sciences, Mo:106). Agresti states that the likelihood ratio test is more reliable for large sample sizes than the Wald test. (Agresti, Alan, 1996, An Introduction to Categorical Data Analysis, Jhon Wiley and Sons, Inc) The Wald test is obtained from the following vector-matrix calculation.

$$W = \hat{\beta}' \left[\sum \hat{\beta} \right]^{-1} \hat{\beta} \quad (2.50)$$

W has a chi-square distribution with p+1 degrees of freedom under the hypothesis that each of the p+1 coefficients are equal to zero. A similar situation can be done with excluding $\hat{\beta}_0$ from the analysis, then W will be distributed as chi-square with p degrees of freedom.

As said, the model containing all the variables is called the full model and the model containing some variables thought to be significant is called the reduced model. The next step is to compare reduced and full model. The comparison of two models is evaluated by G statistic. After the analysis, if it is found that the reduced

model is not equivalent to the full model (if we reject H_0), then the variables excluded in the reduced model are considered necessary. In this case the full model is used.

2.6.2.3 Score Test

Score test is based on the conditional distribution of the p derivatives of $L(\tilde{\beta})$ with respect to $\tilde{\beta}$. The computation of the score test is as complicated as the Wald test.

2.7 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. Interpretation requires two issues. First issue is to determine the functional relationship between dependent variable and the independent variable(s). The second issue is to define of the unit of change for 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. y is linear in the parameters and it is shown as $y = \beta_0 + \beta_1 x$. In logistic regression model, the link function is the logit transformation and its expression is shown in equations (2.10 and 2.12).

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 , for any value of x . It is expressed as follows:

$$\beta_1 = |y(x = x + 1) - y(x = x)| \quad (2.51)$$

Let $y(x) = \beta_0 + \beta_1 x$ then $y(x+1) = \beta_0 + \beta_1(x+1) = \beta_0 + \beta_1 x + \beta_1$ so $y(x+1) - y(x) = \beta_0 + \beta_1(x) + \beta_1 - (\beta_0 + \beta_1 x) = \beta_1$. Hence, β_1 is equal to change in y for a unit change in x .

In logistic regression model, $g(x+1)$ is expressed as follows:

$$g(x+1) = \ln\left\{\frac{\pi(x+1)}{1-\pi(x+1)}\right\} = \beta_0 + \beta_1(x+1) = \beta_0 + \beta_1 x + \beta_1 \quad (2.52)$$

so $g(x+1) - g(x) = g(x+1) - (\beta_0 + \beta_1 x) = \beta_1$. The logit difference is equal to β_1 .

To interpret the coefficient β_1 , the meaning on the difference between logits in logistic regression model must be investigated. This depends on the nature of the independent variable. Independent variable is investigated in three issues.(binary=dichotomous, polytomous, continuous)

2.7.1 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 $1 - \pi(x)$. These are shown in Table 2.3.

Table 2.3 : Values of the Logistic Regression Model When the Independent Variable is Dichotomous

		Independent Variable X	
		x=1	x=0
Outcome Variable Y	y=1	$a = \pi(1) = \frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)}$	$b = \pi(0) = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}$
	y=0	$c = 1 - \pi(1) = \frac{1}{1 + \exp(\beta_0 + \beta_1)}$	$d = 1 - \pi(0) = \frac{1}{1 + \exp(\beta_0)}$
Total		1	0

The odds of the outcome being present among individuals with $x = 1$ is expressed as:

$$\frac{P(y=1|x=1)}{P(y=0|x=1)} = \frac{\pi(1)}{1-\pi(1)} \quad (2.53)$$

The odds of the outcome being present among individuals with $x = 0$ is expressed as:

$$\frac{P(y=1|x=0)}{P(y=0|x=0)} = \frac{\pi(0)}{1-\pi(0)} \quad (2.54)$$

The logit is defined to be the logarithm (natural exponential) of the odds. In other words, “log of the odds” or “log odds” is called logit. The logit is expressed as $\ln(\pi(x)/(1-\pi(x)))$. They are defined by $g(1)$ and $g(0)$ for dichotomous independent variable and shown as follows:

$$g(1) = \ln\left(\frac{\pi(1)}{1-\pi(1)}\right) \quad g(0) = \ln\left(\frac{\pi(0)}{1-\pi(0)}\right) \quad (2.55)$$

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:

$$OR = \frac{\pi(1)/(1-\pi(1))}{\pi(0)/(1-\pi(0))} \quad (2.56)$$

The log of the odds ratio is called logit difference (log odds ratio) and it is expressed as:

$$\ln(OR) = \ln[\pi(1)/(1-\pi(1))] - \ln[\pi(0)/(1-\pi(0))] = g(1) - g(0) \quad (2.57)$$

OR and log odds ratio (logit difference) are expressed for a dichotomous independent variable in below respectively,

$$\begin{aligned} \text{OR} &= \frac{[\exp(\beta_0 + \beta_1)/(1 + \exp(\beta_0 + \beta_1))] [1/(1 + \exp(\beta_0))(1 + \exp(\beta_0))]}{[\exp(\beta_0)/(1 + \exp(\beta_0))] [1/(1 + \exp(\beta_0 + \beta_1))]} \\ &= \frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)} = \exp(\beta_1) = e^{\beta_1} \end{aligned} \quad (2.58)$$

$$\ln(\text{OR}) = \beta_1 \quad (2.59)$$

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 $\hat{\text{OR}}$ 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 $\hat{\text{OR}}$ 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 in the case when X is dichotomous as follows:

$$\text{Vâr}(\hat{\beta}_1) = \left[\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} \right] \quad (2.60)$$

a,b,c,d are cell frequencies in the 2×2 table of Y \times X .

In addition, $\hat{\beta}_1$ and $\hat{\text{OR}}$ are evaluated without using MLE in logistic regression model as follows:

$$\hat{\text{OR}} = \frac{a/c}{b/d} \quad \hat{\beta}_1 = \ln(\hat{\text{OR}}) \quad (2.61)$$

The distribution of the estimate of OR tends to be skewed to the right; it is clearly not normally distributed. \hat{OR} is taken to be normal for large sample sizes. Thus, confidence interval is usually based on $\hat{\beta}_1$ which is closer to being normally distributed. $\hat{\beta}_1 \sim N(\beta_1, \text{Var}(\hat{\beta}_1))$

The confidence interval for the odds ratio is given by

$$\exp\{\hat{\beta}_1 \pm Z_{1-\alpha/2} \text{SE}(\hat{\beta}_1)\} \quad (2.62)$$

2.7.2 Polytomous Independent Variable

In this case, if the independent variable takes three or more levels, then, it is called polytomous independent variable. The design variables to represent the categories of the polytomous independent variable are created. For example, nominal scale variable X is coded at 4 levels. Thus, $(4-1)=3$ design variables are created and they are shown in Table 2.4.

Table 2.4 : The Coding of the Design Variables for Independent Variables

Coded Four Levels

X	Design Variables		
	D ₁	D ₂	D ₃
A (1)	0	0	0
B (2)	1	0	0
C (3)	0	1	0
D (4)	0	0	1

Here, A is called the reference group. The design variable D₁ gives the comparison of group B to the reference group A and D₂ gives the comparison of group C to the reference group A and also D₃ gives the comparison of group D to the reference group A. This procedure continues to k groups for any variable and k-1 design variables are created.

The unknown parameters and the odds ratio values are determined through maximum likelihood estimation in logistic regression. But the odds ratio values can be evaluated without using logistic regression. They are shown in Table 2.5.

Table 2.5 : Cross-Classification of Independent Variables and Status for 95 Subjects

	A	B	C	D	Total
Present	5	15	5	5	45
Absent	10	15	20	15	50
Total	30	30	15	20	95
Odds Ratio (OR)	1.0	2.0	0.5	0.67	
ln(OR)	0.0	0.69	-0.69	-0.40	

For B the estimated odds ratio is $(15 \times 10) / (15 \times 5) = 2$ with using A as the reference group. The log of the odds ratios are given in the last row of this table.

When the logistic regression model to the data using design variables is obtained, the same solution for the coefficients is found. This does not happen by chance. The calculation of the logit difference is shown simply. For the comparison of C to A this is as follows:

$$\begin{aligned}
 \ln\{\hat{OR}(C, A)\} &= \hat{g}(C) - \hat{g}(A) \\
 &= \left\{ \hat{\beta}_0 + \hat{\beta}_{11} \times (D_1 = 0) + \hat{\beta}_{12} \times (D_2 = 1) + \hat{\beta}_{13} \times (D_3 = 0) \right\} \\
 &\quad - \left\{ \hat{\beta}_0 + \hat{\beta}_{11} \times (D_1 = 0) + \hat{\beta}_{12} \times (D_2 = 0) + \hat{\beta}_{13} \times (D_3 = 0) \right\} \\
 &= \hat{\beta}_{12}
 \end{aligned} \tag{2.63}$$

The estimated standard error of the estimated coefficient for design variables is found by using the cell frequencies from the contingency table. This is expressed as follows:

$$\hat{SE}(\hat{\beta}_{12}) = \left[\frac{1}{5} + \frac{1}{20} + \frac{1}{5} + \frac{1}{10} \right]^{1/2} \tag{2.64}$$

After these computations, $100(1 - \alpha)\%$ confidence interval for the coefficient can be obtained using the following formula.

$$\hat{\beta}_{ij} \pm Z_{1-\alpha/2} \{SE(\hat{\beta}_{ij})\} \quad (2.65)$$

Here, i refers to the reference group subscript, and j is the subscript of the group which is compared to the referent group ($j=1, 2, \dots, k-1$). $100(1 - \alpha)\%$ confidence interval for the odds ratio is obtained as follows:

$$\exp\{\hat{\beta}_{ij} \pm Z_{1-\alpha/2} (SE(\hat{\beta}_{ij}))\} \quad (2.66)$$

2.7.3 Continuous Independent Variable

In this case, when there is an independent continuous variable in the model, the unit of this variable should be defined. Under the assumption that the logit is linear in the continuous variable, X , then it is expressed as: $g(x) = \beta_0 + \beta_1 x$. Here, β_1 represents the change in log odds ratio for an increase of 1 unit in X and it is shown as follows:

$$g(x + 1) = \beta_0 + \beta_1(x + 1) = \beta_0 + \beta_1 x + \beta_1 \quad (2.67)$$

$$g(x + 1) - g(x) = \beta_1 \quad \text{for any value of } X.$$

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. For this reason, the unit of independent variable is very important.

The log odds ratio for a change of c units in X , odds ratio and variance of the variable are expressed respectively as follows:

$$x = g(x + c) - g(x) = c\beta_1 \quad (2.68)$$

$$OR(x + c, x) = e^{c\beta_1} \quad (2.69)$$

$$V\hat{a}r\{\ln(OR\hat{R}(x + c, x))\} = c^2 V\hat{a}r\hat{\beta}_1 \quad (2.70)$$

100% confidence interval is evaluated as:

$$\exp(c\hat{\beta}_1 - Z_{1-\alpha/2} c S\hat{E}(\hat{\beta}_1)) \leq OR \leq \exp(c\hat{\beta}_1 + Z_{1-\alpha/2} c S\hat{E}(\hat{\beta}_1)) \quad (2.71)$$

2.7.4 Multivariate Case

In general, there are more than one independent variable in logistic regression models. To statistically adjust the estimated effects of each variable in the model for differences in the distributions and associations among the other independent variables is the goal of multivariate case. Applying this situation to a multivariate logistic regression model, each estimated coefficient provides an estimate of the log odds adjusting for all other variables included in the model.

The model which includes two independent variables is assumed. One of them is continuous and the other is dichotomous variable. Primary interest is focused on the effect of the dichotomous variable. The logistic regression model with two variables is expressed as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \quad (2.72)$$

X_1 , X_2 are continuous and dichotomous variables, respectively. Here, β_1 explains the difference of Y between two different groups ($X_1 = 0$, $X_1 = 1$) and β_2 explains the rate of change in Y per 1 unit of change in X_2 . The mean of continuous variable (X_2) for each of two groups are written as a_1 , a_2 , respectively. y_i is the

mean value of Y for group i (i=1, 2). Comparison of the mean value of Y for two groups is expressed by the following difference:

$$y_2 - y_1 = \beta_1 + \beta_2(a_2 - a_1) \quad (2.73)$$

Here, the value of Y for two groups of X_1 with the adjustment for X_2 are compared. Then, the log odds ratio obtained that is expressed as $\exp(\beta_1)$ is called the ' X_2 ' adjusted odds ratio.

For example, the data summarized below show the basis for an example of evaluating the estimated logistic regression coefficient for a dichotomous variable when the coefficient is adjusted for a continuous variable. In addition, results of logistic regression model are shown in Table 2.7.

Table 2.6 : Descriptive Statistics for Two Groups of 70

Variable	Group 1		Group 2	
	Mean	Sd.	Mean	Sd.
Cancer	0.30	0.46	0.80	0.40
Age	40.18	5.34	48.45	5.02

Table 2.7 : Results of Fitting the Logistic Regression Model

Variable	Estimated Coefficient	Standard Error
Group	1.559	0.557
Age	0.096	0.048
Constant	-4.739	1.998

According to Table 2.6, the univariate log odds ratio for group 2 versus group 1 is $\ln(\hat{OR}) = \ln(0.80/0.20) - \ln(0.30/0.70) = 2.234$, and the unadjusted odds ratio is $\hat{OR} = 9.34$.

There is a considerable difference age distribution of two groups, with women in group 2 being on average nearly 8 years older than group 1. It is also easy to

determine in logistic regression with logit difference. An approximation to the unadjusted odds ratio is obtained by exponentiating the difference $y_2 - y_1$.

$$\begin{aligned}y_2 - y_1 &= [-4.739 + 1.559 + 0.096(48.45)] - [-4.739 + 0.096(40.18)] \\ &= 1.559 + 0.096(48.45 - 40.18) = 2.35292\end{aligned}$$

$$\text{OR}^\hat{=} = 2.35292 = 10.52$$

The discrepancy between 10.52 and the actual unadjusted odds ratio, 9.34, is due to the fact that the above comparison is based on the difference in the average logit.



CHAPTER THREE

MODEL BUILDING STRATEGIES

3.1 Model Building Procedures

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 developing include more complex situations. Determining a strategy and a method are very necessary for handling these 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, such as:

1. The Univariate Analysis.
2. The Multivariate Analysis.
 - a) Stepwise Logistic Regression Methods.
 - i) Forward Selection.
 - ii) Backward Elimination.
 - b) Best Subset Logistic Regression Method.

Stepwise logistic regression method is most often used in situations where the “important” independent variables are not known and the associations with the outcome is not understood well (AIDS, Cancer etc...). For this reason, it is not known whether variable is significant or not in the model. Stepwise logistic regression method offers a fast and effective means of screening a large number of variables and simultaneously fit a number of logistic regression equations. The selection

process becomes harder as the number of variables increases, because of the rapid increase in possible associations and interactions (Agresti, 1990). For this reason, the selection of any variable is very important. In this study, the univariate analysis and the stepwise logistic regression method from multivariate analysis will be used for variable selection.

3.1.1 The Univariate Analysis

The variable selection process begins with univariate analysis of each variable. For categorical (nominal or ordinal) and continuous variables with few integer values, the univariate analysis is done with a contingency table of outcome ($y = 0, 1$) versus the k levels of the independent variable. The value of the likelihood ratio test for the significance of the coefficients for the $k-1$ design variables in a univariate logistic regression model that contains single independent variable is exactly equal to the likelihood ratio chi-square test with $k-1$ degrees of freedom. In addition, it is a good method to estimate the individual odds ratios and their confidence limits using one of the levels as a reference group for the variables exhibiting at least a moderate level of association.

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. Some methods are used. For this situation, one of them is to eliminate the category completely, another one is to collapse the categories of the independent variable in some sensible fashion to eliminate the zero cell and the last is to model as if it were continuous in the case that the variable is ordinal.

The univariate analysis involves estimation of slope coefficient for the univariate logistic regression model containing only one variable, estimation of standard error of the estimated slope coefficient, the Wald test, the likelihood ratio test (G) for the

significance of the coefficient, the individual odds ratio, the p-value of the coefficient, the 95% CI for the odds ratio and the p-value of the likelihood ratio test.

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. Here, the confidence interval estimate of odds ratios of selected variables should contain 1 value (Ryan and Thomas, 1997). Why is the p-value less than 0.25? The empirical evidences are represented by Bendel and Afifi (1977) for linear regression and by Mickey and Greenland (1989) for logistic regression. 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. For this reason, it is important to determine variables added to model before a decision making to the final model.

The importance of each variable included in the multivariate logistic regression model should be verified. This includes an examination of the Wald statistic for each variable and a comparison that should be made between each estimated coefficient with the coefficient from the univariate model based on that variable only. How will it be investigated for this situation? Variables that do not contribute to the model are eliminated from new model. The new model are compared to the old model through the likelihood ratio test. Also, the estimated coefficients for the remaining variables are compared to those from the full model. 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. According to Wald statistic, it is taken that the reference value is equal to 2. If the independent variable is coded as 3 levels using the design variables, then there would be 2 design variables. If the Wald statistics values for both coefficients exceed 2, then it is concluded that the design variables are significant. But, if one of the coefficients is equal to 4.0 (it is greater than 2) and the other is equal to 0.4 (it is less than 2), then we can not be sure about the contribution of the variable to the

model. In this case, the likelihood ratio test (G) is used. *“The likelihood ratio test statistic for a particular regressor is the difference between two deviance statistics: the deviance without the regressor in the model minus the deviance with the regressor in the model”* (Ryan and Thomas, 1997). It is expressed as follows:

$$G = D(\text{for the model without the regressor}) - D(\text{for the model with the regressor})$$

Here, the statistic G has a chi-square distribution with ν degrees of freedom and the calculation of ν is shown as follows:

$$\nu = \text{the number of } \hat{\beta} \text{ in the full model} - \text{the number of } \hat{\beta} \text{ in the reduced model}$$

Using this notation, the p-value associated with this test is $P(\chi_{\nu}^2 > G) < 0.05$, 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. In other words, any variable with corresponding p-value > 0.05 in the multivariate logistic regression model should be considered for removing from the model. The variables that do not contribute to the model should be eliminated. Here, likelihood ratio chi-square test is used. The aim for this test is to determine the difference between two deviance statistics.

The new model (after removing the variables with large p-value) should be compared to the old model. The regression coefficients are checked in the new model. If some of them are remarkably changed in magnitude, it implies that the excluded variables may be important as confounding variables. Finally, any variable deselected for the multivariate logistic regression model should be added back into the model to identify potential confounding variables. For example, AGE is not significant variable; but this variable is required or found necessary by researcher.

“If the univariate analysis yields an extremely large number of possible variables then it may be employed a stepwise or best subsets method.” (Hosmer & Lemeshow, 1989, p:87) The stepwise logistic regression will be investigated in the forward topics.

The question of the appropriate categories for discrete variables should have been addressed at the univariate stage. The linearity in the logit for continuous scaled variables should has been checked. How will we do this check procedure? Three of them are commonly employed. These are lowess (locally weighted squares regression), dummy (design) variable method and fractional polynomials. Here, dummy variable method will be used. 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.

Because, it is necessary to transform them to logits in logistic regression, a coefficient must be equal to zero (0) for the first group. In addition, the odds ratio is equal to 1 for the first group. The four plotted points are connected in order to inspect the most logical functional form for the scale of selected variable. It may be linear, quadratic, binary or other nonlinear function. At the end of these steps, the model using the possible functional form of the variable suggested by the graph is refitted. The odds ratio values are plotted according to the groups. If there is no linear relationship that can be increasing or decreasing between them, then dummy variable method is used.

An alternative procedure for scale identification in logistic regression is the Box-Tidwell transformation (Box and Tidwell, 1962). Box and Tidwell approach adds a term of the form $x \ln(x)$ to the model. If the coefficient for this variable is significant, there is an evidence for non-linearity in the logit. But, this procedure has low power in detecting small departures from linearity.

After determining that each of the continuous variables is in the correct scale, interactions are need to be checked in the model. An interaction between two variables in any model implies that the effect of one of the variables is not constant over levels of the others. For example, an interaction between age and sex would imply that the slope coefficient for age is different for males and females. After including interaction terms in a model, their significance is then decided using a likelihood ratio chi-square test. *“By significance we mean interactions must contribute to the model. For example, inclusion of an interaction term in the model whose sole effect is to increase the estimated standard errors without changing the point estimate would not be helpful. In general, for an interaction term to alter both point and interval estimates, the estimated coefficient for the interaction term must attain at least a moderate level of statistical significance. The final decision as to whether an interaction term should be based on statistical as well as practical considerations. That is, the interaction term should also make sense form a biologic perspective.”* (Hosmer and Lemeshow, 1989, p:91)

3.1.2 The Stepwise Logistic Regression Method

Stepwise logistic regression is an extremely popular method for model building. Why do we use this method? Because, many possible covariates are collected and employing a stepwise selection procedure provides a fast and effective means to eliminate a large number of variables. Stepwise procedure for selection or deletion of variables from a model is based on a statistical algorithm. This algorithm checks for the importance of variables.

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 stepwise linear regression an F-test is used since the errors are assumed to be normally distributed. 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. Possible differences in degrees of freedom between variables are accounted at any procedure based on the likelihood ratio test statistic. Assessing significance for G is done by p-value. Here, description and illustration algorithm for forward selection and backward elimination procedures will be investigated.

The stepwise logistic regression begins with a base model containing only the intercept parameter. It then adds variables significant to the model until there are no remaining significant variables left to be added. The nice feature of this procedure is that at each step after a variable has been added to the model, all of the variables included in previous steps are retested in order to see whether they are still significant or not. The inclusion and extraction of variables from the model in the stepwise logistic regression method is based upon the likelihood ratio (chi-square) test. Normally, 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 a round 0.20. This will help in avoiding possibly significant variables from being overlooked or removed unnecessarily from the model.

This method will be described by considering the statistical computations that the computer must perform at each step of the procedure. 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. $j = 1, 2, \dots, p$

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 $j=1, 2, \dots, p$ 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 $p_j^{(0)} = \Pr(\chi_v^2 > G_j^{(0)})$.

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) $v=1$ if x_j is continuous.
- b) $v = k - 1$ 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 $p_{e_1}^{(0)}$ with a pre-specified significance level p_E .

- a) If $p_{e_1}^{(0)} < p_E$, move on the next step.
- b) If $p_{e_1}^{(0)} \geq p_E$, 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 0.25.

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 , $j=1, 2, \dots, p$ 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 $p_j^{(1)} = \Pr(\chi_v^2 > G_j^{(1)})$.

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 $p_{-e_j}^{(2)} = \Pr(\chi_v^2 > G_{-e_j}^{(2)})$ 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 $p_{e_2} > p_R$, the variable x_{e_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 $j = 1, 2, \dots, p, j \neq e_1, e_2$.

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 $p_j^{(2)} = \Pr(\chi_v^2 > G_j^{(2)})$ for $j = 1, 2, \dots, p, j \neq e_1, e_2$.

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 be the best model. It depends on the researcher and the status of data.

“Disadvantage 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 files with large numbers of variables this can be quite costly both in terms of time and money.” (Hosmer and Lemeshow, 1989, p:111)

3.1.3 The Best Subsets Logistic Regression Method

The other selection method of variables for a model is the best subsets selection. A parallel theory has been worked out for nonnormal errors models for this method. A number of models containing one, two, three and so on... variables which are considered the “best” with respect to some predetermined criterias are examined. Here, likelihood ratio test is used to select the variables. Also, this method is used in linear regression models. But, its useage is more difficult because the logistic regression includes more iteration procedures.

Stepwise and best subsets logistic regression methods are criticized because they can yield a biologically implausible model and they can select irrelevant variables. The problem is not the fact that the computer can select such model. The main problem is that the analyst fails to carefully scrutinize the resulting model and reports that the final model is as the best model.

3.2 Goodness of Fit Tests

After fitting the logistic regression model, it is useful to test its effectiveness by using goodness of fit tests. In addition, it is decided whether the fit of the model is adequate by using goodness of fit tests or not. One of them is deviance test and the other is Hosmer-Lemeshow test. 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 $y^* = (y_1, y_2, \dots, y_n)$ and the fitted values of the outcome variable in vector form as \hat{y} where $\hat{y}^* = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n)$. If summary measures of the distance between y and \hat{y} are small or each pair (y_i, \hat{y}_i) to these summary measures is not systematic and it is small relative to the error structure of the model, the fitted model is accepted well. $(y_i - \hat{y}_i)$ is defined to be residual and its value must be small ($i=1, 2, \dots, n$). The fitted model contains p independent variables ($x^* = (x_1, x_2, \dots, x_p)$) and J denotes the number of distinct observed values of x . If some subjects have the same value of x then $J < n$. Here, the number of subjects with $x = x_j$ is denoted by m_j and it is accepted as $\sum m_j = n$ ($j=1, 2, \dots, J$). y_j is denoted the number of positive responses, $y=1$, among the m_j subjects with $x = x_j$. The total number of subjects with $y=1$ is denoted by $\sum_j y_j = n_1$. The distribution of the goodness of fit statistics is obtained by assumption that n becomes large.

3.2.1 Pearson Chi-Square Test and Deviance Test

$(y_i - \hat{y}_i)$ is called residual. The fitted values are calculated for each covariate pattern in logistic regression and depend on the estimated probability for that covariate pattern. The fitted value is denoted by \hat{y}_j .

$$m_j \hat{\pi}_j = m_j \left[\frac{\exp(\hat{g}(x_j))}{1 + \exp(\hat{g}(x_j))} \right] \quad (3.1)$$

where $\hat{g}(x_j)$ is the estimated logit function.

Two measures of the difference between the observed and the fitted values are investigated. These are the Pearson residual and the deviance residual. The Pearson residual and the Pearson chi-square statistic are defined as follows:

$$r(y_j, \hat{\pi}_j) = \frac{(y_j - m_j \hat{\pi}_j)}{\sqrt{m_j \hat{\pi}_j (1 - \hat{\pi}_j)}} \quad j=1, 2, \dots, J \quad (3.2)$$

$$\chi^2 = \sum_{j=1}^J r(y_j, \hat{\pi}_j)^2 \quad (3.3)$$

Otherwise, the deviance residual is defined as follows:

$$d(y_j, \hat{\pi}_j) = \pm \left\{ 2 \left[y_j \ln \left(\frac{y_j}{m_j \hat{\pi}_j} \right) + (m_j - y_j) \ln \left(\frac{(m_j - y_j)}{m_j (1 - \hat{\pi}_j)} \right) \right] \right\}^{1/2} \quad (3.4)$$

where the sign is the same as the sign of $(y_j - m_j \hat{\pi}_j)$. For covariate patterns with $y_j = 0$, the deviance residual is defined as follows:

$$d(y_j, \hat{\pi}_j) = -\sqrt{2m_j |\ln(m_j (1 - \hat{\pi}_j))|} \quad (3.5)$$

In addition, for covariate patterns with $y_j = m_j$, the deviance residual is defined as follows:

$$d(y_j, \hat{\pi}_j) = \sqrt{2m_j |\ln(m_j \hat{\pi}_j)|} \quad (3.6)$$

The summary statistic based on the deviance residuals is the deviance and it is shown as follows:

$$D = \sum_{j=1}^J d(y_j, \hat{\pi}_j)^2 \quad (3.7)$$

The distribution of the statistics χ^2 and D is chi-square with $J - (p + 1)$ degrees of freedom.

3.2.2 The Hosmer-Lemeshow Test

The aim of the Hosmer-Lemeshow test is to make a group of the values of the estimated probabilities. Here, J is equal to n . Two grouping strategies are proposed as follows:

- a) Grouping based on the percentiles of the estimated probabilities.
- b) Grouping based on the fixed values of the estimated probabilities.

Here the first strategy will be used. In this grouping method, 10 groups are created ($g=10$). The first group contains $n_1^* = n/10$ subjects having the smallest estimated probabilities. Also, the last group contains $n_{10}^* = n/10$ subjects having the largest estimated probabilities. The each group's n_k^* equals to $n/10$ ($k=1, 2, \dots, 10$). For the $y=1$ row, the estimates of the expected values are found by summing the estimated probabilities over all subjects in a group. In addition, 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 obtained by calculating the Pearson chi-square statistic from the $2 \times g$ table of observed and estimated expected frequencies. The calculation of this test is given as follows:

$$\hat{C} = \sum_{k=1}^g \frac{(o_k - n_k^* \bar{\pi}_k)^2}{n_k^* \bar{\pi}_k (1 - \bar{\pi}_k)} \quad (3.8)$$

where n_k^* is the number of covariate patterns in the k th group.

$$o_k = \sum_{j=1}^{n_k^*} y_j \quad (3.9)$$

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

$$\bar{\pi}_k = \sum_{j=1}^{n_k^*} \frac{m_j \hat{\pi}_j}{n_k^*} \quad (3.10)$$

The distribution of the statistic \hat{C} is well approximated by the chi-square distribution with $g-2$ degrees of freedom, when J is equal to n and the fitted logistic regression model is the correct model. If the value of the Hosmer-Lemeshow goodness of fit statistic computed from “deciles of risk” table is less than the corresponding p-value computed from the chi-square distribution with 8 degrees of freedom, then the model is accepted to fit quite well.

The Hosmer-Lemeshow goodness of fit statistic is easily interpretable and it can be easily applied to data. It is illustrated as follows:

Table 3.1: Observed and Estimated Expected Frequencies

Y		Decile of Risk					Total
		1	2	...	10		
Y=1	Obs	o_{11}	o_{12}	...	o_{110}	n_1	
	Exp	$\bar{\pi}_{11}$	$\bar{\pi}_{12}$...	$\bar{\pi}_{110}$		
Y=0	Obs	o_{01}	o_{02}	...	o_{010}	n_0	
	Exp	$\bar{\pi}_{01}$	$\bar{\pi}_{02}$...	$\bar{\pi}_{010}$		
Total		$n/10$	$n/10$...	$n/10$	n	

CHAPTER FOUR

APPLICATION

4.1 General Information on the Data

This study contains 1200 patients and these data include the statement of the absence or presence of lung cancer. For this reason, 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. There are many factors for patients with lung cancer to have this disease. The factors that are obtained from clinical trials or observations increase the number of patients who have lung cancer. These factors should be controlled for this reason. Nowadays, some of them can be controlled. But majority of them can not be controlled. Because, the details of disease are not known or predicted. The risk factors affected to lung cancer should be known to decrease or to stop effects of disease. Also, in which level that is affected should be obtained. The ratio of cancerous patients can be reduced to lower levels. How can we reduce this? Here, fitting the “best” model is important. The variable selection is very important to find the “best” model. In addition, it is an important criteria to have the number of less variable. In this chapter, the logistic regression method was used to find the ratio of cancerous patients and the “best” model. The univariate analysis and the stepwise variable selection procedure were applied to cancer data. The data set was obtained from Ege University Faculty of Medicine Department of Chest Diseases in İzmir.

Here, 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. AGE variable is continuous and the others are categoric variables. They are illustrated in Table 4.1.

Table 4.1 : 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

These independent variables according to patients who have lung cancer and patients who become control group are presented in Tables 4.2 - 4.3 - 4.4 - 4.5 - 4.6 - 4.7 and 4.8.

Table 4.2 : Y * SEX Cross Tabulation Count

		SEX		
Y	Male	Female	Total	
Co	567	33	600	
Ca	576	24	600	
Total	1143	57	1200	

Table 4.3 : Y * EDU Cross Tabulation Count

		EDU				
Y	Illiterate	Primary	Secondary	High+Unv.	Total	
Co	166	327	48	59	600	
Ca	202	361	29	8	600	
Total	368	688	77	67	1200	

Table 4.4 : Y * YOS Cross Tabulation Count

		YOS				
Y	Non-Smoker	<=20	21-30	31-40	>40	Total
Co	193	61	117	113	116	600
Ca	23	16	91	192	278	600
Total	216	77	208	305	394	1200

Table 4.5 : Y * AOIS Cross Tabulation Count

		AOIS				
Y	Non-Smoker	<=10	11-15	16-19	=>20	Total
Co	193	34	122	78	173	600
Ca	23	85	209	103	180	600
Total	216	119	331	181	353	1200

Table 4.6 : Y * NOPPY Cross Tabulation Count

		NOPPY				
Y	Non-Smoker	01-10	11-20	21-30	>30	Total
Co	193	37	62	109	199	600
Ca	23	11	18	77	471	600
Total	216	48	80	186	670	1200

Table 4.7 : Y * DOGUS Cross Tabulation Count

		DOGUS				
Y	Smoke	01-05	06-11	=>11	Non-Smoker	Total
Co	263	51	31	62	193	600
Ca	437	74	26	40	23	600
Total	700	125	57	102	216	1200

Table 4.8 : AGE Situation

	N	Minimum	Maximum	Mean	Std. Deviation
Age (year)	1200	30.00	90.00	58.9583	9.8190

4.2 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 results of fitting the univariate logistic regression models to data are given in Table 4.9. In this table, the following informations for each variable listed in the first column are presented.

- (1) the estimated slope coefficient for the univariate logistic regression model containing only this variable.
- (2) the estimated standard error of the estimated slope coefficient.
- (3) the Wald statistic.
- (4) the degrees of freedom.
- (5) the p-value of the coefficient.
- (6) the estimated odds ratio.
- (7) the 95% confidence interval (CI) for the odds ratio.
- (8) the likelihood ratio test statistic (G).
- (9) the p-value of the G statistic.

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. In addition, the Wald statistic values of these variables are so high and the confidence interval of odds ratios does not contain value 1. These are some evidences that they are significant.

Table 4.9 : Univariate Logistic Regression Models for Case to Have or Don't Have Ca

Variable	$\hat{\beta}$	SE	Wald	df	p-value	Exp($\hat{\beta}$)	CI for Exp($\hat{\beta}$)		G	P-value
							Lower	Upper		
SEX (1)	-0.334	0.275	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		

Under the null hypothesis, the slope coefficients are zero, G follows the chi-square distribution with 1 degrees of freedom for SEX and AGE variables, except for the variables YOS, AOIS, NOPPY, and DOGUS where they have 4 degrees of freedom and the variable EDU where it has 3 degrees of freedom. The p-values all of the variables are less than 0.25 value. For this reason, they are found to be significant. If we select the p-value as 0.10, then the variable SEX is excluded from the model. In addition, this situation can be seen from the confidence interval of odds ratio. The value of odds ratio includes value 1.

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 4.10.

Table 4.10 : Multivariate Model Containing Variables Identified in the Univariate Analysis

Variable	$\hat{\beta}$	SE	Wald	df	p-value	Exp ($\hat{\beta}$)	CI for Exp ($\hat{\beta}$)		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										

On the basis of the output displayed in Table 4.10, 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 new model which does not contain the variable AOIS is fitted. The results of fitting this model are given in Table 4.11.

Table 4.11 : Multivariate Model without AOIS

Variable	$\hat{\beta}$	SE	Wald	df	p-value	Exp ($\hat{\beta}$)	CI for Exp ($\hat{\beta}$)		G	P-value
							Lower	Upper		
SEX(1)	1.866	0.406	21.104	1	0.000	6.463	2.915	14.330	402.575	0.000
EDU			15.080	3	0.002					
1	1.568	0.437	12.842	1	0.000	4.796	2.035	11.306		
2	1.651	0.425	15.069	1	0.000	5.210	2.264	11.990		
3	1.559	0.500	9.718	1	0.002	4.754	1.784	12.669		
AGE	0.051	0.010	23.622	1	0.000	1.052	1.031	1.074		
YOS			53.404	4	0.000					
1	2.736	0.563	23.635	1	0.000	15.430	5.120	46.503		
2	3.261	0.462	49.835	1	0.000	26.065	10.542	64.449		
3	2.546	0.404	39.614	1	0.000	12.751	5.771	28.171		
4	2.369	0.430	30.394	1	0.000	10.691	4.605	24.821		
NOPPY			40.473	3	0.000					
1	-1.967	0.490	16.090	1	0.000	0.140	0.054	0.366		
2	-1.854	0.366	25.653	1	0.000	0.157	0.076	0.321		
3	-1.520	0.300	25.654	1	0.000	0.219	0.121	0.394		
DOGUS			26.347	3	0.000					
1	1.248	0.285	19.162	1	0.000	3.483	1.992	6.091		
2	0.925	0.325	8.117	1	0.004	2.523	1.335	4.768		
3	0.186	0.384	0.236 *	1	0.627	1.205	0.568	2.556		
Constant	-7.342	0.806	82.947	1	0.000	0.001				
-2LL=1260.978										

The likelihood ratio test statistic (G) for the hypothesis that the slope coefficient is zero is obtained as minus twice the difference between the log-likelihoods for all variables in the model and the model containing all variables except for the variable AOIS. Under the null hypothesis, G value follows the chi-square distribution with 3 degrees of freedom. This is denoted by $(v_{full} - v_{reduced}) = 19 - 16 = 3$. The likelihood ratio test for the difference between the models in Tables 4.10 and 4.11 (a test for the significance of AOIS) yields a value of $G = [1260.978 - 1252.142] = 8.836$. Comparing this value to a chi-square distribution with 3 degrees of freedom yields a value of 7.81 ($\chi_{3,0.95}^2 = 7.81$). Here, 8.836 is greater than 7.81. For this reason, the variable AOIS is significant in this model.

Now, an another new model which does not contain the variable DOGUS is fitted. The results of this model are given in Table 4.12.

Table 4.12 : Multivariate Model without DOGUS

Variable	$\hat{\beta}$	SE	Wald	df	p-value	Exp ($\hat{\beta}$)	CI for Exp ($\hat{\beta}$)		G	P-value
							Lower	Upper		
SEX(1)	1.848	0.401	21.218	1	0.000	6.348	2.891	13.937	381.068	0.000
EDU			17.120	3	0.001					
1	1.605	0.433	13.761	1	0.000	4.979	2.132	11.627		
2	1.729	0.420	16.986	1	0.000	5.635	2.476	12.822		
3	1.615	0.494	10.704	1	0.001	5.026	1.910	13.221		
AGE	0.033	0.009	12.375	1	0.000	1.034	1.015	1.053		
YOS			101.927	4	0.000					
1	3.010	0.540	31.104	1	0.000	20.294	7.046	58.456		
2	3.739	0.433	74.479	1	0.000	42.073	17.996	98.456		
3	3.273	0.341	92.320	1	0.000	26.391	13.536	51.453		
4	3.208	0.353	82.451	1	0.000	24.730	12.374	49.426		
AOIS			5.329	3	0.149					
1	0.553	0.264	4.377	1	0.036	1.738	1.036	2.916		
2	0.294	0.182	2.623	1	0.105	1.342	0.940	1.915		
3	0.285	0.211	1.824	1	0.177	1.330	0.879	2.012		
NOPY			34.079	3	0.000	0.164	0.063	0.426		
1	-1.809	0.488	13.756	1	0.000	0.187	0.093	0.377		
2	-1.676	0.358	21.983	1	0.000	0.260	0.146	0.466		
3	-1.346	0.297	20.579	1	0.000	1.034	1.015	1.053		
Constant	-6.315	0.746	71.584	1	0.000	0.002				
-2LL=1282.485										

Here, the same procedure is applied as in above. The likelihood ratio test for the difference between the models in Tables 4.10 and 4.12 (a test for the significance of DOGUS) yields a value of $G = [1282.485 - 1252.142] = 30.343$. Comparing G value to a chi-square distribution with 3 degrees of freedom yields a value of 7.81 ($\chi_{3,0.95}^2 = 7.81$). Here, 30.343 is greater than 7.81. The variable DOGUS is found to be significant for this model with respect to this result.

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. The lowest quartile or the variable that has the lowest risk as the reference group is usually

selected. So, Table 4.13 is obtained. The new model with design variables is shown in Table 4.14.

Table 4.13 : Results of the Quartile Analysis of AGE

Quartile	1	2	3	4
Interval	=<51	[52,59]	[60,66]	>=67
Estimated Coefficient	0	1.120	1.449	1.331
Odds Ratio	1	3.063	4.259	3.786
95 % CI	-	1.986-4.724	2.576-7.041	2.171-6.603

Table 4.14 : Multivariate Model of Linearity for AGE

Variable	$\hat{\beta}$	SE	Wald	df	p-value	Exp ($\hat{\beta}$)	CI for Exp ($\hat{\beta}$)		G	P-value
							Lower	Upper		
SEX (1)	1.740	0.408	18.198	1	0.000	5.697	2.561	12.670	418.023	0.000
EDU			17.003	3	0.001					
1	1.763	0.442	15.936	1	0.000	5.830	2.453	13.854		
2	1.754	0.429	16.714	1	0.000	5.776	2.492	13.390		
3	1.620	0.505	10.307	1	0.001	5.054	1.880	13.590		
AGE			36.168	3	0.000					
1	1.120	0.221	25.651	1	0.000	3.063	1.986	4.724		
2	1.449	0.256	31.99	1	0.000	4.259	2.576	7.041		
3	1.331	0.284	21.999	1	0.000	3.786	2.171	6.603		
YOS			47.106	4	0.000					
1	2.537	0.566	20.089	1	0.000	12.640	4.168	38.327		
2	3.097	0.468	43.726	1	0.000	22.130	8.837	55.415		
3	2.224	0.420	28.009	1	0.000	9.243	4.056	21.063		
4	1.976	0.467	17.921	1	0.000	7.217	2.890	18.019		
AOIS			8.726	3	0.033					
1	0.716	0.269	7.080	1	0.008	2.047	1.208	3.470		
2	0.420	0.186	5.072	1	0.024	1.521	1.056	2.192		
3	0.303	0.217	1.948	1	0.163	1.354	0.885	2.071		
NOPPY			36.454	3	0.000					
1	-1.909	0.492	15.031	1	0.000	0.148	0.056	0.389		
2	-1.711	0.367	21.748	1	0.000	0.181	0.088	0.371		
3	-1.444	0.299	23.263	1	0.000	0.236	0.131	0.424		
DOGUS			25.453	3	0.000					
1	1.213	0.284	18.268	1	0.000	3.364	1.929	5.867		
2	0.920	0.324	8.062	1	0.005	2.509	1.330	4.735		
3	0.135	0.382	0.125	1	0.723	1.145	0.542	2.420		
Constant	-5.428	0.556	95.449	1	0.000	0.004				
-2LL=1245.530										

If the variable AGE is as linear in the logit, then it is expected to show either a linear increasing or decreasing trend in the estimated coefficient. But, the statistical evidence of linearity for variable AGE is not obtained in Tables 4.13 or 4.14. For this reason, the statement that the variable AGE is not linear in the logit is supported. This variable is used as continuous. If the estimated odds ratio values are less than 1 or near to each others, then they are combined and formed as reference group.

After these processes, the final model is accepted in Table 4.10. The logit function of this model is expressed as follows:

$$\hat{g}(x) = \beta_0 + \beta_{11}D_{11} + \beta_{21}D_{21} + \beta_{22}D_{22} + \beta_{23}D_{23} + \beta_3x_3 + \beta_{41}D_{41} + \beta_{42}D_{42} + \beta_{43}D_{43} + \beta_{44}D_{44} + \beta_{51}D_{51} + \beta_{52}D_{52} + \beta_{53}D_{53} + \beta_{61}D_{61} + \beta_{62}D_{62} + \beta_{63}D_{63} + \beta_{71}D_{71} + \beta_{72}D_{72} + \beta_{73}D_{73}$$

$$\hat{g}(x) = -7.960 + 1.892D_{11} + 1.557D_{21} + 1.660D_{22} + 1.576D_{23} + 0.060x_3 + 2.605D_{41} + 3.054D_{42} + 2.237D_{43} + 1.857D_{44} + 0.711D_{51} + 0.437D_{52} + 0.289D_{53} - 1.907D_{61} - 1.805D_{62} - 1.524D_{63} + 1.374D_{71} + 1.059D_{72} + 0.246D_{73}$$

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 cancerous.

The estimated odds ratios and the confidence intervals for the variables SEX, EDU, YOS, AOIS, NOPPY and DOGUS are given in Table 4.10. These confidence interval values show whether the variables have an important effect in the model or not. In case, NOPPY variable contains value 1. For this reason, this variable is found to be insignificant. But, this variable is significant for p-values at design variables of different levels. According to these odds ratio values, being a female has 6.631 times more risk factor than being a male. Here, being a male is the reference group. In addition, the value of the Wald test is too high. For this reason, p-value is quite small (0.000). For EDU variable, high school and university are combined and it is called the reference group. Illiterates (in category 1), people graduated from primary school (in category 2) and people graduated from secondary school (in category 3) 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 example, smokers, who are less than 21 years of smoking, in category 1 have 13.527 times more risk of having lung cancer with respect to non-smokers, smokers, who are between 21 and 30 years of smoking, in category 2 have 21.203 times more risk of having lung cancer with respect to non-smokers. But smokers, who are between 31 and 40 years of smoking,

in category 3 and smokers, who are greater than 40, in category 4 have 9.367 times and 6.402 times respectively more risk of having lung cancer with respect to non-smokers. 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.

4.3 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.

4.3.1 The Forward Selection

Forward selection procedure will be applied to the data. The results of this process are presented in Tables that will be shown below in terms of the p-values to enter and remove calculated at each step. The program is run by using $p_E = 0.15$ and $p_R = 0.20$.

Step (0):

At Step (0) the program selects as a candidate for entry at Step (1) the variable with the smallest p-value. In addition, the largest value of the score statistic should be chosen. The variable NOPPY with a p-value of 0.000 is selected and the score

statistic value is 288.009. These are shown in Tables 4.15.a and 4.15.b. Since this p-value is less than 0.15, the program proceeds to Step (1).

Table 4.15.a : Variables in the Model (Only Constant)

Variable	$\hat{\beta}$	S.E.	Wald	df	p-value	Exp ($\hat{\beta}$)
Constant	0.000	0.058	0.000	1	1.000	1.000

**Table 4.15.b : Variables not in the Model
(SEX, EDU, AGE, YOS, AOIS, NOPPY, DOGUS)**

Variables	Score	df	p-value
SEX (1)	1.492	1	0.222
EDU	48.711	3	0.000
1	5.079	1	0.024
2	3.938	1	0.047
3	5.010	1	0.025
AGE	63.718	1	0.000
YOS	250.416	4	0.000
1	28.102	1	0.000
2	3.931	1	0.047
3	27.435	1	0.000
4	99.170	1	0.000
AOIS	182.112	4	0.000
1	24.263	1	0.000
2	31.577	1	0.000
3	4.066	1	0.044
4	0.197	1	0.657
NOPPY	288.009	4	0.000 *
1	14.670	1	0.000
2	25.929	1	0.000
3	6.515	1	0.011
4	250.016	1	0.000
DOGUS	186.463	4	0.000
1	103.803	1	0.000
2	4.724	1	0.030
3	0.460	1	0.497
4	5.186	1	0.023

Step (1):

At Step (1) the program will not remove the variable just entered since $p_R > p_E$ and the p-value to remove at Step (1) is equal to the p-value to enter at Step (0). The variable with the smallest p-value to enter at Step (1) is SEX with a value of 0.000 among the variables that are not in the model. This p-value is less than 0.15. In addition, the largest value of the score statistic is for the variable SEX with a value of 25.245. So the program proceeds to Step (2). These are shown in Tables 4.16.a and 4.16.b.

**Table 4.16.a : Variables in the Model
(Constant, NOPPY)**

Variables	$\hat{\beta}$	S.E.	Wald	df	p-value	Exp ($\hat{\beta}$)	CI for Exp ($\hat{\beta}$)	
							Lower	Upper
NOPPY			231.148	4	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
Constant	-2.127	0.221	92.992	1	0.000	0.119		
-2LL= 1350.927								

**Table 4.16.b : Variables not in the Model
(SEX, EDU, AGE, YOS, AOIS, DOGUS)**

Variables	Score	df	p-value
SEX (1)	25.245	1	0.000 *
EDU	23.645	3	0.000
1	2.526	1	0.112
2	0.807	1	0.369
3	0.532	1	0.466
AGE	17.909	1	0.000
YOS	5.304	3	0.151
1	1.673	1	0.196
2	0.501	1	0.479
3	2.759	1	0.097
AOIS	3.177	3	0.365
1	2.666	1	0.102
2	0.036	1	0.849
3	0.185	1	0.668
DOGUS	11.421	3	0.010
1	7.803	1	0.005
2	0.056	1	0.812
3	4.965	1	0.026
	92.950	14	0.000

Step (2):

There are two variables in the model at this step. One of them is NOPPY and the other is SEX. At Step (2) the largest p-value to remove is 0.000 and this p-value does not exceed 0.20, thus the program moves to the variable selection phase. These are shown in Table 4.17.a. The smallest p-value to enter among the remaining variables that are not in the model is for the variable EDU and this value is 0.000. In addition, the largest value of the score statistic is for the variable EDU with a value of 22.799. These are shown in Table 4.17.b.

Table 4.17.a : Variables in the Model
(Constant, NOPPY, SEX)

Variables	$\hat{\beta}$	S.E.	Wald	df	p-value	Exp ($\hat{\beta}$)	CI for Exp ($\hat{\beta}$)	
							Lower	Upper
SEX (1)	1.840	0.389	22.377	1	0.000	6.299	2.939	13.504
NOPPY			212.976	4	0.000			
1	1.254	0.444	7.993	1	0.005	3.505	1.469	8.361
2	1.426	0.393	13.141	1	0.000	4.162	1.925	8.997
3	2.376	0.329	52.262	1	0.000	10.763	5.651	20.498
4	3.572	0.34	137.823	1	0.000	35.578	19.598	64.588
Constant	-2.724	0.293	86.389	1	0.000	0.066		
-2LL= 1327.927								

Table 4.17.b : Variables not in the Model
(EDU, AGE, YOS, AOIS, DOGUS)

Variables	Score	df	p-value
EDU	22.799	3	0.000 *
1	0.880	1	0.348
2	1.770	1	0.183
3	0.191	1	0.662
AGE	15.844	1	0.000
YOS	5.571	3	0.134
1	1.839	1	0.175
2	0.585	1	0.444
3	2.838	1	0.092
AOIS	3.243	3	0.356
1	2.659	1	0.103
2	0.033	1	0.856
3	0.078	1	0.780
DOGUS	12.146	3	0.007
1	8.229	1	0.004
2	0.082	1	0.775
3	6.018	1	0.014
	71.667	13	0.000

Step (3):

At Step (3) the largest p-value to remove is 0.000 and this p-value does not exceed 0.20, thus the program moves to the variable selection phase. This is shown in Table 4.18.a. The smallest p-value to enter at Step (4) is for the variable AGE with a value of 0.001 among the variables not in the model. The largest value of the score statistic is for the variable DOGUS with a value of 13.303. But the p-value for the variable DOGUS is larger than the variable AGE with a value of 0.004. For this reason, the variable AGE for the model is preferred. These are shown in Table 4.18.b.

Table 4.18.a : Variables in the Model
(Constant, NOPPY, SEX, EDU)

Variables	$\hat{\beta}$	S.E.	Wald	df	p-value	Exp ($\hat{\beta}$)	CI for Exp ($\hat{\beta}$)	
							Lower	Upper
SEX (1)	1.834	0.400	21.025	1	0.000	6.261	2.358	13.714
EDU			19.476	3	0.000			
1	1.815	0.423	18.404	1	0.000	6.138	2.679	14.063
2	1.792	0.414	18.694	1	0.000	6.002	2.664	13.525
3	1.592	0.488	10.626	1	0.001	4.914	1.887	12.799
NOPPY			194.601	4	0.000			
1	1.329	0.451	8.674	1	0.003	3.779	1.560	9.153
2	1.550	0.402	14.866	1	0.000	4.714	2.143	10.367
3	2.370	0.337	49.614	1	0.000	10.703	5.534	20.699
4	3.553	0.313	129.181	1	0.000	34.930	18.927	64.464
Constant	-4.439	0.508	76.229	1	0.000	0.012		
-2LL= 1303.281								

**Table 4.18.b : Variables not in the Model
(AGE, YOS, AOIS, DOGUS)**

Variables	Score	df	p-value
AGE	11.783	1	0.001 *
YOS	5.876	3	0.118
1	2.531	1	0.112
2	0.842	1	0.359
3	2.426	1	0.119
AOIS	3.445	3	0.328
1	1.945	1	0.163
2	0.226	1	0.635
3	0.012	1	0.914
DOGUS	13.303	3	0.004
1	8.827	1	0.003
2	0.049	1	0.825
3	6.024	1	0.014
	49.648	10	0.000

Step (4):

At this step, there are four variables in the model. The largest p-value to remove is 0.000, which does not exceed 0.20, thus the program moves to the variable selection phase. This is shown in Table 4.19.a. The smallest p-value among the variables not in the model is for the variable DOGUS and with a value of 0.000. In addition, the largest value of the score statistic is for the variable DOGUS with a value of 23.483. These are shown in Table 4.19.b.

Table 4.19.a : Variables in the Model
(Constant, NOPPY, SEX, EDU, AGE)

Variables	$\hat{\beta}$	S.E.	Wald	df	p-value	Exp ($\hat{\beta}$)	CI for Exp ($\hat{\beta}$)	
							Lower	Upper
SEX (1)	1.829	0.403	20.596	1	0.000	6.227	2.826	13.718
EDU			16.631	3	0.001			
1	1.590	0.431	13.637	1	0.000	4.905	2.109	11.406
2	1.703	0.418	16.551	1	0.000	5.488	2.417	12.463
3	1.617	0.494	10.723	1	0.001	5.040	1.914	13.269
AGE	0.027	0.008	11.676	1	0.001	1.028	1.012	1.044
NOPPY			171.564	4	0.000			
1	1.430	0.457	9.764	1	0.002	4.177	1.704	10.239
2	1.663	0.406	16.777	1	0.000	5.277	2.381	11.698
3	2.442	0.339	51.882	1	0.000	11.497	5.916	22.345
4	3.491	0.313	124.169	1	0.000	32.804	17.754	60.613
Constant	-5.916	0.679	75.998	1	0.000	0.003		
-2LL= 1291.443								

Tables 4.19.b : Variables not in the Model
(YOS, AOIS, DOGUS)

Variables	Score	df	p-value
YOS	3.627	3	0.305
1	1.827	1	0.177
2	2.721	1	0.099
3	0.222	1	0.637
AOIS	4.752	3	0.191
1	1.936	1	0.164
2	0.363	1	0.547
3	0.230	1	0.632
DOGUS	23.483	3	0.000 *
1	16.124	1	0.000
2	0.163	1	0.686
3	7.430	1	0.006
	39.093	9	0.000

Step (F):

At Step (F) the program finds that the maximum p-value to remove is 0.001 for the variable EDU. This value is less than 0.20. So, the variable EDU is not removed from the model. This is shown in Table 4.20.a. There is no smallest p-value to enter among the remaining variables not in the model. In addition, there is no largest score statistic value to enter among the remaining variables not in the model. These are shown in Table 4.20.b.

Table 4.20.a : 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 4.20.b : Variables not in the Model
(YOS, AOIS)**

Variables	Score	df	p-value
YOS		3	
1	0.000	1	1.000
2	0.000	1	1.000
3	18.394	1	0.000
AOIS		3	
1	0.000	1	1.000
2	0.000	1	1.000
3	0.000	1	1.000

But the p-value of the design variable is for the variable DOGUS with a value of 0.992. This is shown in Table 4.20.a. This value is large. For this reason, this variable will be investigated. The likelihood ratio test statistic (G) is used. This is obtained as minus twice the difference between the log-likelihoods for the model that has been obtained at Step (F) and the model that does not contain for the variable DOGUS. Under the null hypothesis, G value follows the chi-square distribution with 3 degrees of freedom. The likelihood ratio test for the difference between the models in Tables 4.20.a. and 4.21.a. (a test for the significance of DOGUS) yields a value of $G = [1291.443 - 1268.458] = 22.985$. Comparing G value to a chi-square distribution with 3 degrees of freedom yields a value of 7.81 ($\chi_{3,0.95}^2 = 7.81$). Here, 22.985 is greater than 7.81. For this reason, the variable DOGUS is significant for this model.

Table 4.21.a : Variables in the Model without DOGUS

Variables	$\hat{\beta}$	S.E.	Wald	df	p-value	Exp ($\hat{\beta}$)	CI for Exp ($\hat{\beta}$)	
							Lower	Upper
SEX (1)	1.829	0.403	20.596	1	0.000	6.227	2.826	13.718
EDU			16.631	3	0.001			
1	1.590	0.431	13.637	1	0.000	4.905	2.109	11.406
2	1.703	0.418	16.551	1	0.000	5.488	2.417	12.463
3	1.617	0.494	10.723	1	0.001	5.040	1.914	13.269
AGE	0.027	0.008	11.676	1	0.001	1.028	1.012	1.044
NOPPY			17.564	4	0.000			
1	1.430	0.457	9.764	1	0.002	4.177	1.704	10.239
2	1.663	0.406	16.777	1	0.000	5.277	2.381	11.698
3	2.442	0.339	51.882	1	0.000	11.497	5.916	22.345
4	3.491	0.313	124.169	1	0.000	32.804	17.754	60.613
Constant	-5.916	0.679	75.998	1	0.000	0.003		
-2LL= 1291.443								

Table 4.21.b : Variables not in the Model without DOGUS

Variables	Score	df	p-value
YOS	3.627	3	0.305
1	1.827	1	0.177
2	2.721	1	0.099
3	0.222	1	0.637
AOIS	4.752	3	0.191
1	1.936	1	0.164
2	0.363	1	0.547
3	0.230	1	0.632

After fitting the model, the variable AGE that has been modeled as continuous to obtain the correct scale in the logit is needed to be examined. Three design variables based on the quartiles of age are formed and replaced as AGE (continuous) with these design variables in the model. It is shown in Table 4.13 as before.

The new model with design variables is shown in Tables 4.22.a and 4.22.b.

Table 4.22.a : Variables in the Multivariate Model of Linearity for AGE

Variables	$\hat{\beta}$	S.E.	Wald	df	p-value	Exp ($\hat{\beta}$)	CI for Exp ($\hat{\beta}$)	
							Lower	Upper
SEX (1)	1.716	0.395	18.850	1	0.000	5.562	2.563	12.067
AGE			32.794	3	0.000			
1	0.883	0.203	18.996	1	0.000	2.419	1.626	3.599
2	1.145	0.214	28.562	1	0.000	3.143	2.065	4.784
3	0.980	0.214	20.931	1	0.000	2.665	1.751	4.056
NOPPY			84.783	4	0.000			
1	0.867	0.482	3.235	1	0.072	2.380	0.925	6.121
2	1.009	0.449	5.051	1	0.025	2.742	1.138	6.609
3	1.858	0.391	22.569	1	0.000	6.412	2.979	13.803
4	2.704	0.384	49.607	1	0.000	14.943	7.041	31.713
DOGUS			21.437	3	0.000			
1	0.915	0.248	13.587	1	0.000	2.496	1.535	4.059
2	0.620	0.298	4.324	1	0.038	1.858	1.036	3.333
3	-0.075	0.365	0.042	1	0.837	0.928	0.453	1.898
Constant	-3.465	0.331	109.810	1	0.000	0.031		
-2LL= 1282.588								

Table 4.22.b : Variables not in the Multivariate Model of Linearity for AGE

Variables	Score	df	p-value
EDU		3	
1	0.000	1	1.000
2	0.000	1	1.000
3	0.000	1	1.000
YOS		3	
1	0.000	1	1.000
2	0.000	1	1.000
3	30.195	1	0.000
AOIS		3	
1	0.000	1	1.000
2	0.000	1	1.000
3	0.000	1	1.000

If the variable AGE is linear in logit, then it is seen to be either a linear increasing or decreasing trend in the estimated coefficient. But there is no evidence of linearity for AGE. It is shown as in Table 4.22.a. So, this variable is used as continuous. The final model is given in Table 4.20.a.

The logit function of this model is expressed as follows:

$$\hat{g}(x) = \beta_0 + \beta_{11}D_{11} + \beta_{21}D_{21} + \beta_{22}D_{22} + \beta_{23}D_{23} + \beta_3x_3 + \beta_{41}D_{41} + \beta_{42}D_{42} + \beta_{43}D_{43} + \beta_{44}D_{44} + \beta_{51}D_{51} + \beta_{52}D_{52} + \beta_{53}D_{53}$$

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

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. For EDU variable, the risk of having lung cancer varies according to education status. 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. Here, people graduated from high school or university are the 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 cigarettes consumption per year increases, the risk of having lung cancer also increases with respect to non-smokers. When the number of packages of cigarettes consumption per year is less than 11, this category contains the value of 2.598 times more risk of having lung cancer with respect to non-smokers. According to this situation, when the number of packages of cigarettes consumption per year is greater than 31, this category contains the value of 14.675 times more risk of having lung cancer 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. In addition, the confidence interval of odds ratio for every variable does not contain value 1.

4.3.2 The Backward Elimination

Backward elimination procedure was applied too. At Step (0) the program selects as a candidate for remove at Step (1) the variable that has the largest p-value. In addition, the smallest value of the score statistic should be chosen. But there is no available variable to these criterias. For this reason, all variables take place in the model. These are shown in the following tables.

Step (0):

Table 4.23.a : Variables in the Model (Only Constant)

Variable	$\hat{\beta}$	S.E.	Wald	df	p-value	Exp ($\hat{\beta}$)
Constant	0.000	0.058	0.000	1	1.000	1.000

**Table 4.23.b : Variables not in the Model
(SEX, EDU, AGE, YOS, AOIS, NOPPY, DOGUS)**

Variables	Score	df	p-value
SEX (1)	1.492	1	0.222
EDU	48.711	3	0.000
1	5.079	1	0.024
2	3.938	1	0.047
3	5.010	1	0.025
AGE	63.718	1	0.000
YOS	250.416	4	0.000
1	28.102	1	0.000
2	3.931	1	0.047
3	27.435	1	0.000
4	99.170	1	0.000
AOIS	182.112	4	0.000
1	24.263	1	0.000
2	31.577	1	0.000
3	4.066	1	0.044
4	0.197	1	0.657
NOPPY	288.009	4	0.000
1	14.670	1	0.000
2	25.929	1	0.000
3	6.515	1	0.011
4	250.016	1	0.000
DOGUS	186.463	4	0.000
1	103.803	1	0.000
2	4.724	1	0.030
3	0.460	1	0.497
4	5.186	1	0.023

Step (1):

At Step (1) the model includes all variables. That is shown in Table 4.10 as before. All values of the model are the same as the univariate analysis. Interpretation of them have been explained as before.

The logit function and logistic regression function of this model is expressed as follows:

$$\hat{g}(x) = \beta_0 + \beta_{11}D_{11} + \beta_{21}D_{21} + \beta_{22}D_{22} + \beta_{23}D_{23} + \beta_3x_3 + \beta_{41}D_{41} + \beta_{42}D_{42} + \beta_{43}D_{43} + \beta_{44}D_{44} + \beta_{51}D_{51} + \beta_{52}D_{52} + \beta_{53}D_{53} + \beta_{61}D_{61} + \beta_{62}D_{62} + \beta_{63}D_{63} + \beta_{71}D_{71} + \beta_{72}D_{72} + \beta_{73}D_{73}$$

$$\hat{g}(x) = -7.960 + 1.892D_{11} + 1.557D_{21} + 1.660D_{22} + 1.576D_{23} + 0.060x_3 + 2.605D_{41} + 3.054D_{42} + 2.237D_{43} + 1.857D_{44} + 0.711D_{51} + 0.437D_{52} + 0.289D_{53} - 1.907D_{61} - 1.805D_{62} - 1.524D_{63} + 1.374D_{71} + 1.059D_{72} + 0.246D_{73}$$

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

4.4 Goodness of Fit Test

The values of the Hosmer-Lemeshow goodness of fit test statistic computed from the frequencies in Tables 4.24 and 4.25 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. 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.

Table 4.24 : 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

$$\hat{C} = \frac{(5 - 4.582)^2}{4.582} + \dots + \frac{(81 - 91.135)^2}{91.135} + \frac{(115 - 115.418)^2}{115.418} + \dots + \frac{(30 - 19.865)^2}{19.865}$$

$$\hat{C} = 13.590$$

$\hat{C} = 13.590 < \chi_{8,0.05}^2 = 15.507$. For this reason, the final model obtained from forward selection method fits data.

The logit function is shown as follows:

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

Table 4.25 : 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

$$\hat{C} = \frac{(5-3.683)^2}{3.683} + \dots + \frac{(86-96.620)^2}{96.620} + \frac{(115-116.317)^2}{116.317} + \dots + \frac{(28-17.380)^2}{17.380}$$

$$\hat{C} = 15.769$$

$\hat{C} = 15.769 > \chi_{8,0.05}^2 = 15.507$. For this reason, the final model obtained from backward elimination method does not fit data.

The logit function is shown as follows:

$$\begin{aligned} \hat{g}(x) = & -7.960 + 1.892D_{11} + 1.557D_{21} + 1.660D_{22} + 1.576D_{23} + 0.060x_3 + 2.605D_{41} \\ & + 3.054D_{42} + 2.237D_{43} + 1.857D_{44} + 0.711D_{51} + 0.437D_{52} + 0.289D_{53} \\ & - 1.907D_{61} - 1.805D_{62} - 1.524D_{63} + 1.374D_{71} + 1.059D_{72} + 0.246D_{73} \end{aligned}$$

CHAPTER FIVE

CONCLUSION

5.1 Conclusion

There are many statistical approaches to predictive probability modeling. In this study, a logistic regression model was investigated. Because the logistic regression model is used to explain the relationship between the response variable and independent variables, when the response variable was observed into two or more categories. Here, the response variable is observed into two categories. These categories were being lung cancer (Ca) or control group (Co). 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. In this study, univariate case and multivariate case were investigated. The risks of being lung cancer were determined by using variable selection methods. These methods are forward selection and backward elimination.

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 cancer by 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.

5.2 Further Research

Assessing goodness of fit in logistic regression model can be problematic. Deviance and Pearson chi-square statistics do not have approximate chi-square distributions, under the null hypothesis of no lack of fit, when continuous covariates are modelled. In addition, Hosmer-Lemeshow test is used mostly. What is the main difference between Hosmer-Lemeshow test and Deviance or Pearson chi-square statistics? For this reason, further research can be done about differences between goodness of fit tests or goodness of fit tests with continuous covariates. Another further research can be done about outliers or zero cells. When data consist on many zero cells, which procedures can be done? These zero cells can be ignore, combine or anything else.



REFERENCES

- Agresti, A., (1990). Categorical Data Analysis. John Wiley & Sons.
- Akgül, A., (2003). Tıbbi Araştırmalarda İstatistiksel Analiz Teknikleri. Emek Ofset.
- Cristensen, R., (1997). Log-Linear Models and Logistic Regression. (2th edition) Springer-Verlag.
- Dobson, A.J, (1990). An Intoduction to Generalized Linear Models. Chapman & Hall.
- Fermanian, J., Batista, G., Meindinger, A., Payan, C., & Cremniter, D., (2001). Predictors of short-term deterioration and compliance in psychiatric emergency patients: Aprospective study of 457 patients referred to the emergency room of a general hospital. Psychiatry Research, 104,49-59.
- Freund, R. J., & Wilson, W.J., Regression Analysis (Statistical modeling of a response variable).
- Grouven, U., & Bender, R., (1998). Using binary logistic regression models for ordinary data with non-proportional odds. J. Clin. Epidemiol, 51,809-816.
- Hosmer, D., & Lemeshow, S., (1989). Applied Logistic Regression. John Wiley & Sons.