

DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED
SCIENCES

EVALUATION OF OBESITY RISK FACTORS
USING LOGISTIC REGRESSION AND
ARTIFICIAL NEURAL NETWORKS

by
Ayça EFE

September, 2012
İZMİR

EVALUATION OF OBESITY RISK FACTORS USING LOGISTIC REGRESSION AND ARTIFICIAL NEURAL NETWORKS

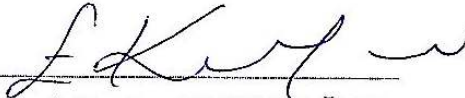
**A Thesis Submitted to the Graduate School of Natural and Applied Sciences
of Dokuz Eylül University In Partial Fulfillment of the Requirements for the
Degree of Master of Science in Statistics**

**by
Ayça EFE**

**September, 2012
İZMİR**

M. Sc THESIS EXAMINATION RESULT FORM

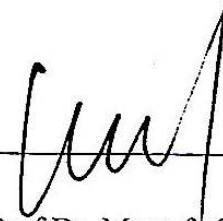
We have read the thesis entitled “EVALUATION OF OBESITY RISK FACTORS USING LOGISTIC REGRESSION AND ARTIFICIAL NEURAL NETWORKS” completed by AYÇA EFE under supervision of ASST. PROF. DR. EMEL KURUOĞLU and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.


Asst. Prof. Dr. Emel KURUOĞLU

Supervisor


Doc. Dr. Aylin Kantarci
(Jury Member)


Doc. Dr. Aylin ALIN
(Jury Member)


Prof. Dr. Mustafa SABUNCU
Director
Graduate School of Natural and Applied Sciences

ACKNOWLEDGMENTS

I would like to express my full gratitude to my supervisor Asst. Prof. Dr. Emel KURUOĞLU for guiding me throughout my studies.

Also I wish to thank my dearest husband Özgür EFE, my mother İffet ŞAHİN, my little boy Yiğit EFE and my genuine friends Derya ÖZKAN and FatmaŞAHİNER.

Ayça EFE

EVALUATION OF OBESITY RISK FACTORS USING LOGISTIC REGRESSION AND ARTIFICIAL NEURAL NETWORKS

ABSTRACT

In this study, two widely used techniques in a situation where outcome variable is dichotomous, while classifying observations, logistic regression and artificial neural network are examined. The data from obesity survey which is answered by 12th graders of the Anatolian and State high schools in the province of Gazıemir, İzmir is analyzed by using MATLAB, and of the considered methods the predictive abilities are evaluated. The logistic regression coefficients have been determined by using maximum likelihood method. According to the data from obesity survey, whether each relation between obesity risk factor and the outcome variable is significant or not has been determined by using univariate analysis. In the feed forward neural network, for adjusting connection weights, a backpropagation learning rule has been used.

Keywords: Logistic regression, artificial neural network, obesity.

OBEZİTE RİSK FAKTÖRLERİNİN LOJİSTİK REGRESYON ve YAPAY SİNİR AĞLARI KULLANILARAK DEĞERLENDİRİLMESİ

ÖZ

Bu çalışmada yanıt değişkeninin iki kategorili olduğu durumda, gözlemlerin sınıflandırılmasında yaygın olarak kullanılan iki temel teknik olan logistic regression ve yapay sinir ağları incelenmiştir. İzmir ili Gaziemir ilçesinde bulunan Anadolu lisesi ve düz lise statüsündeki 3 lisenin 12 nci sınıf öğrencilerinin yanıtladığı obezite anket formu verileri, MATLAB programı kullanılarak analiz edilmiş ve her iki tekniğin sonuç çıktısını tahminleme yeterlilikleri değerlendirilmiştir. Logistic regresyon modeli katsayı değerleri en çok olabilirlik yöntemi kullanılarak belirlenmiştir. Obezite anket formu verilerine göre her bir obezite risk faktörünün yanıt değişkeni ile ilişkisinin istatistiksel olarak anlamlı olup olmadığı tek değişkenli analiz tekniği ile belirlenmiştir. Çok katmanlı ileri sürümlü yapay sinir ağında, bağlantı ağırlıklarının sonuç çıktısına göre ayarlanmasında öğrenme kuralı olarak geriye yayılım öğrenme algoritması kullanılmıştır.

Anahtar Sözcükler: Lojistik regresyon, yapay sinir ağları, obezite.

CONTENTS

	Page
THESIS EXAMINATION RESULT FORM.....	ii
ACKNOWLEDGEMENTS.....	iii
ABSTRACT.....	iv
ÖZ.....	v
 CHAPTER ONE-INTRODUCTION.....	 1
 CHAPTER TWO- MAIN FEATURES OF LOGISTIC REGRESSION	 4
 2.1 Meaning of Response Function when Outcome Variable is Dichotomous	 4
2.2 Special Problems when Outcome Variable is Dichotomous	5
2.3 Simple Logistic Regression Model.....	6
2.3.1 Fitting the Simple Logistic Regression Model.....	8
2.3.1.1 Likelihood Function.....	9
2.3.1.2 Fitted Simple Logistic Regression Model.....	10
2.3.1.3 Testing for the Significance of the Coefficients.....	11
2.3.1.3.1 Likelihood Ratio Test.....	11
2.3.1.3.2 Wald Test.....	12
2.3.1.3.3 Score Test.....	13
2.4 Multiple Logistic Regression Model.....	14
2.4.1 Dummy Variable.....	14
2.4.2 Fitting the Multiple Logistic Regression Model.....	15
2.4.2.1 Likelihood Function.....	15
2.4.2.2 Fitted Multiple Logistic Regression Model	16
2.4.2.3 Testing for the Significance of the Coefficients	16
2.4.2.3.1 Likelihood Ratio Test.....	17
2.4.2.3.2 Wald Test.....	17
2.4.2.3.3 Score Test.....	18
2.4.3 Confidence Interval of the Coefficients.....	18
2.5 Interpretation of the Coefficients	19

2.5.1 Dichotomous Independent Variable.....	20
2.5.2 Polytomous Independent Variable.....	22
2.5.3 Continuous Independent Variable.....	24
2.5.4 Multivariate Case.....	25
2.6 Model Building Strategies and Methods	27
2.6.1 Univariate Analysis.....	28
2.6.2 Stepwise logistic regression	29
2.6.3 Best Subsets Selection Method	31
2.7 Assessing the Fit of the Model.....	31
2.7.1 Pearson Chi-Square and Deviance.....	32
2.7.2 The Hosmer-Lemeshow Tests.....	33
CHAPTER THREE- ARTIFICIAL NEURAL NETWORK.....	35
3.1 History of Neural Networks	36
3.2 Biological Neural Networks.....	37
3.3 Artificial Neuron Models	37
3.4 Single Layer Feedforward Networks.....	39
3.5 Multi-Layer Feedforward Networks.....	40
CHAPTER FOUR- APPLICATION.....	42
4.1 Univariate Analysis	50
4.2 Artificial Neural Network.....	57
CHAPTER FIVE- CONCLUSION.....	58
REFERENCES.....	59
APPENDIX A.....	61
APPENDIX B.....	64
APPENDIX C.....	68

CHAPTER ONE

INTRODUCTION

Logistic regression and artificial neural networks (ANNs) are used increasingly in many applications. Logistic regression and ANNs allow you to develop predictive models for categorical outcomes with two or more categories. In logistic regression, predictor variables can be either categorical or continuous, or a combination of these in the one model. The strength of a modeling technique lies in its ability to model many variables but our primary goal is to obtain the best fitting model while minimizing the number of parameters.

A categorical variable has two primary types of scales. Nominal scale is the one which is used to group the characteristic to be examined according to its presence or absence in a case. For nominal variables, the order of listing the categories is irrelevant. The statistical analysis does not depend on that ordering (Agresti, 2007). Examples are gender (male, female), smoking status (smoker, nonsmoker), etc. The other categorical type of scale is called ordinal. Ordinal scales list the examined characteristic qualitatively. Therefore distances between categories are unknown. Examples are socio economic status (high, medium, low), education level (primary, secondary, high, university), etc.

In logistic regression, the outcome or response variable, Y , can take two possible values denoted by 0 and 1 where 1 represents the occurrence of the event and 0 represents the absence of the event.

Since the results of the method of least square which is used for the coefficient estimation of linear regression is meaningless when it is used for a dichotomous variable, the method of maximum likelihood has been used. The maximum likelihood estimate of a parameter is the parameter value for which the probability of the observed data takes its greatest value. It is the parameter value at which the likelihood function takes its maximum. (Agresti, 2007)

There are three different procedure to be applied to determine significant variables which should be included into the logistic regression model: the univariate analysis, stepwise logistic regression method and best subsets logistic regression method. The selection process becomes more challenging as the number of explanatory variables increases, because of the rapid increase in possible effects and interactions. There are two competing goals: The model should be complex enough to fit the data well, but simpler models are easier to interpret (Agresti, 2007). After the model building is completed, how well suited the fitted logistic regression model to the observed data should be determined through one of Pearson chi square test, Deviance test and Hosmer-Lemeshow test.

Like logistic regression, artificial neural network is one of the non-linear multivariate predictive methods. The nodes are the basic units of the artificial neuron model. A general artificial neuron model has an input layer, one or more hidden layers, and the output layer. The input layer has only the role of distributing the inputs to the hidden layer. Each of these nodes in the hidden layer computes a weighted sum of the inputs, adds a constant and runs an activation function. Several iterative algorithms can be used but the most widely used is the back-propagation method. Backpropagation uses supervised learning in which the network is trained using data for which inputs as well as desired outputs are known. Once trained, the network weights are frozen and can be used to compute output values for new input samples. (Mehrotra, Mohan, Ranka, 2000)

In this study as a result of the obesity survey conducted on 12th grader high school students most of whom are 18 years old it is estimated whether students may become obese or not by the two techniques mentioned about using independent variables which have an effect to bring about obesity.

Obesity has been one of the most influential health problem across the world recently. Obesity may result in diabetes, hypertension, some forms of cancer and cardiovascular diseases. Furthermore, rapid changes in diets and lifestyles that have

occurred with industrialization, urbanization, economic development and market globalization, have accelerated over the past decade. This is having a significant impact on the health and nutritional status of populations, particularly in developing countries and in countries in transition. While standards of living have improved, food availability has expanded and become more diversified, and access to services has increased, there have also been significant negative consequences in terms of inappropriate dietary patterns, decreased physical activities and increased tobacco use, and a corresponding increase in diet-related chronic diseases, especially among poor people. (World Health Organization [WHO], 2003)

Many authorities agree that genetic predisposition, physical inactivity, and poor dietary choices are primary contributors to the problem of overweight children. The problem of obesity is multifactorial and thought to be a convergence of factors favoring an imbalance between energy consumed and expended. Patterns of physical activity, as well as a sedentary lifestyle, appear to play important roles in long-term weight regulation. (Mota, Ribeiro, Santos & Helena Gomes, 2006)

This study has five main chapters. In the first, the whole study is introduced. In the second chapter the main features of logistic regression is explained. In the third the artificial neural network is presented. In the fourth, the application based on the data from obesity survey is carried out. The MATLAB was used for data analysis for logistic regression and artificial neural network techniques. In logistic regression, data were examined through the univariate analysis procedure to determine the candidate variables. In multi layer feed forward neural network for training data we use the back-propagation learning rule. The final chapter, the implications of findings are discussed.

CHAPTER TWO

MAIN FEATURES OF LOGISTIC REGRESSION

Logistic regression is a mathematical modeling approach that can be used to group the observations and describe the relationship of several independent variables to a categorical dependent variable. Logistic regression method provides an easy interpretation for the users and mathematical flexibility which draws interest of the researchers. Early uses were in biomedical studies but the past 20 years have also seen much use in biostatistics, social science and marketing researches.

There are many research situations, however, when the outcome variable of interest is categorical (e.g. win/lose; fail/pass; diseased/not diseased ;dead/alive). These outcomes may be coded 1 and 0 respectively. Because of the outcome variable in logistic regression is dichotomous the choice of parametric model and the assumptions are different from linear regression.

2.1 Meaning of Response Function when Response Variable is Dichotomous

The simple linear regression model is:

$$Y_i = B_0 + B_1 X_i + \epsilon_i \quad i=1,2,\dots,n \quad 2.1$$

where the response variable Y_i is binary with possible values of 0 or 1. Since the expected value of the error is zero which is $E\{\epsilon_i\} = 0$, then we obtain the equation

2.2.

$$E\{Y_i\} = B_0 + B_1 x_i \quad 2.2$$

Because Y_i is a bernoulli random variable, the probability distribution is written as:

Table 2.1 The probability distribution of binary Y_i

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

By the definition of the expected value of a random variable we obtain the equation 2.3.

$$E\{Y_i\} = 1 \cdot \pi(x_i) + 0 \cdot (1 - \pi(x_i)) \quad 2.3$$

It is seen that the expected value of Y_i always represents the probability that $Y=1$. Equating 2.2 and 2.3 we reached the equation 2.4.

$$E\{Y_i\} = B_0 + B_1x_i = \pi(x_i) \quad 2.4$$

The conditional mean is the mean value of the response variable, given the value of the independent variable. It can be expressed as $E(Y|x)$ where Y denotes the outcome variable and x denotes a value of the independent variable. Thus we reached the equation 2.5.

$$E\{Y|x\} = B_0 + B_1x_i = \pi(x_i) \quad 2.5$$

2.2 Special Problems when Response Variable is Dichotomous

First problem is the assumption is that the error terms are normally distributed for linear regression is not valid for the dichotomous outcome, each error terms can take on only two values. If $Y=1$ then $\varepsilon_i = 1 - (B_0 + B_1X_i)$ with probability $\pi(x_i)$ and if $Y=0$ then $\varepsilon_i = -(B_0 + B_1X_i)$ with probability $1 - \pi(x_i)$. Since ε_i can take on only two values, the distribution of the error terms is binomial instead of normal distribution.

The second problem is that the error terms do not have equal variances when the response variable is dichotomous. To see this we shall obtain $\sigma^2\{\varepsilon_i\}$ for the simple linear regression model is as follows:

$$\begin{aligned} \sigma^2\{\varepsilon_i\} &= E\{(\varepsilon_i^2)\} = P(Y=0)(-\pi(x_i))^2 + P(Y=1)(1-\pi(x_i))^2 \\ \sigma^2\{\varepsilon_i\} &= (1-\pi(x_i))(-\pi(x_i))^2 + \pi(x_i)(1-\pi(x_i))^2 = \pi(x_i)(1-\pi(x_i)) \\ \sigma^2\{\varepsilon_i\} &= \pi(x_i)(1-\pi(x_i)) \end{aligned} \quad 2.6$$

Thus the variance of ε_i varies as a function of level of x .

Finally, since the response variable represents probabilities when the response variable is 0 or 1, the conditional mean should be constrained as follows:

$$0 \leq E\{Y | x\} \leq 1 \quad 2.7$$

2.3 Simple Logistic Regression Model

The conditional mean can be denoted as $\pi(x)$ instead of $E\{Y | x\}$. The specific form of the logistic regression model is given equation 2.8 :

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \quad 2.8$$

This equation can also be written as:

$$\pi(x) = \left[1 + e^{-(\beta_0 + \beta_1 x)} \right]^{-1} \quad 2.9$$

When the response variable is binary, the shape of the response function will often be curvilinear. The curve is said to be S shaped and approximately linear except the ends. When sign of β_1 is positive the function is monotone increasing else the function is monotone decreasing.

The change in the $\pi(x)$ per-unit change in x becomes progressively smaller as the conditional mean gets closer to 0 or 1. (Hosmer and Lemeshow, 1989)

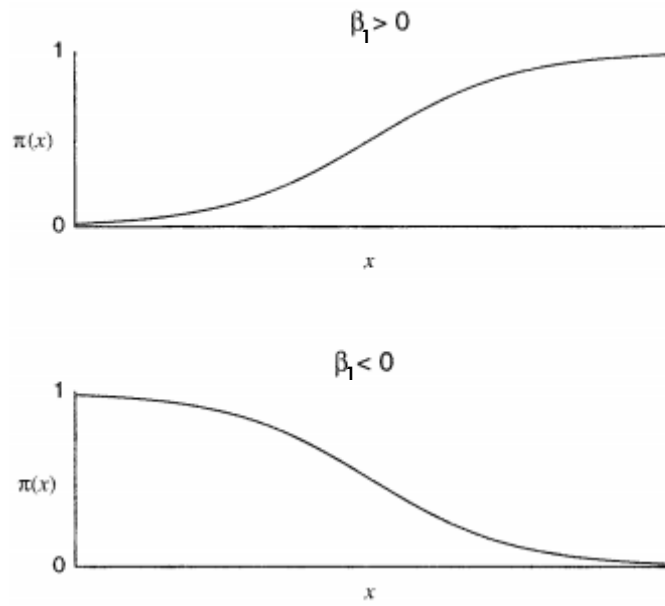


Figure 2.1 The curve of the response function is monoton increasing or decreasing depending on the sign of β_1

The alternative form of the logistic model we make a transformation as follows:

$$g(x) = \ln \left[\frac{\pi(x)}{1 - \pi(x)} \right] \quad \text{where} \quad \pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \quad 2.10$$

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

$$g(x) = \beta_0 + \beta_1 x \quad 2.11$$

The logit response function is a linear function of the independent variables and the coefficients in the logistic model are interpreted just the same as linear regression coefficients are interpreted.

The importance of this transformation is that $g(x)$ has many of the desirable properties of a linear regression model. The logit $g(x)$ is linear in its parameters, may be continuous and may range from $-\infty$ to $+\infty$, depending on the range of x (Hosmer and Lemeshow, 1989). The ratio $\pi(x)/(1-\pi(x))$ in the logit transformation is called the odds. Basically an odds is the ratio of the probability that some event will occur over the probability that the same event will not occur. The fact that the odds are greater than one indicates that the event has a probability of occurring greater than one-half. Conversely, if the odds are less than one, the event has probability of occurring less than one-half. (Christensen, 1997)

2.3.1 Fitting the Simple Logistic Regression Model

To fit the simple logistic regression model we first estimate the unknown parameters (β_0 and β_1). In linear regression the method that is used for estimating parameters is least square. In this method, unknown parameters are chosen in a way the sum of squared differences is minimum between predicted values obtained from the model and observed values. Under the usual assumptions for linear regression the method of least squares yields estimators with a number of desirable statistical properties. When the method of least square is applied to a model with a dichotomous outcome the estimators no longer have these properties. (Hosmer and Lemeshow, 1989)

There are a few methods when outcome variable is dichotomous in determining the parameter values. These methods are: Maximum Likelihood, Reweighted Iterative Least Square and Minimum Logit Chi-square. The maximum likelihood estimation method is used for this study.

Maximum likelihood method choose values for the unknown parameters which maximize the probability of observed data. To accomplish this we must first construct the likelihood function.

2.3.1.1 Likelihood Function

Since each Y_i observation is an ordinary Bernoulli variable, where: $P(Y_i = 1) = \pi_i$ and $P(Y_i = 0) = 1 - \pi_i$, we can represent its probability distribution as follows:

$$f_i(Y_i) = \pi_i^{Y_i} [1 - \pi_i]^{1-Y_i} \quad 2.12$$

Since the observations are assumed to be independent, the likelihood function is calculated as the product of the terms given in equation 2.12 as follows.

$$L(\pi_0, \pi_1) = \prod_{i=1}^n \pi_i^{Y_i} [1 - \pi_i]^{1-Y_i} \quad 2.13$$

It is easier to find the maximum likelihood estimates by working with the logarithm of the joint probability function:

$$\begin{aligned} \ln L(\pi_0, \pi_1) &= \sum_{i=1}^n \left[Y_i \ln \pi_i + (1 - Y_i) \ln (1 - \pi_i) \right] \\ \ln L(\pi_0, \pi_1) &= \sum_{i=1}^n Y_i \ln \pi_i + \sum_{i=1}^n (1 - Y_i) \ln (1 - \pi_i) \end{aligned} \quad 2.14$$

$$\begin{aligned} &= \sum_{i=1}^n Y_i \ln \left(\frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right) + \sum_{i=1}^n (1 - Y_i) \ln \left(1 - \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right) \\ &= \sum_{i=1}^n Y_i \left[\ln e^{\beta_0 + \beta_1 x_i} - \ln (1 + e^{\beta_0 + \beta_1 x_i}) \right] + \sum_{i=1}^n (1 - Y_i) \left[-\ln (1 + e^{\beta_0 + \beta_1 x_i}) \right] \\ \ln L(\pi_0, \pi_1) &= \sum_{i=1}^n Y_i (\beta_0 + \beta_1 x_i) - \sum_{i=1}^n (1 + Y_i) (\beta_0 + \beta_1 x_i) \end{aligned} \quad 2.15$$

Now to maximize the likelihood function, we take the derivative first with respect to β_0 and equal to zero. The equation is as follows:

$$\begin{aligned} \frac{\partial \ln L(\pi_0, \pi_1)}{\partial \beta_0} &= \sum_{i=1}^n \left\{ Y_i - \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right\} = 0 \\ \sum_{i=1}^n [Y_i - \pi_i] &= 0 \end{aligned} \quad 2.16$$

The maximum likelihood estimation of $\pi(x_i)$ is denoted by $\hat{\pi}(x_i)$. If we put this quantity to the right hand side of the equation we reached the equation 2.17 that the sum of the observed values of Y_i is equal to the sum of the predicted values $\hat{\pi}(x_i)$.

$$\sum_{i=1}^n Y_i = \sum_{i=1}^n \hat{\pi}(x_i) \quad 2.17$$

Now by taking the derivative with respect to β_1 and setting equal to zero, we obtain the equation 2.18.

$$\frac{\partial \ln L(\beta_0, \beta_1)}{\partial \beta_1} = \sum_{i=1}^n \left[Y_i x_i - \frac{x_i e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right] = 0$$

$$\sum_{i=1}^n [x_i (Y_i - \hat{\pi}(x_i))] = 0 \quad 2.18$$

For logistic regression the expressions 2.16 and 2.18 are nonlinear in β_0 and β_1 thus solving these equations simultaneously requires an iterative numerical method.

2.3.1.2 Fitted Simple Logistic Regression Model

Once the maximum likelihood estimates of the unknown parameters $\hat{\beta}_0$ and $\hat{\beta}_1$ are found, we substitute these values into the response function in 2.8 to obtain the fitted response function.

$$\hat{\pi}(x) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x}} \quad 2.19$$

The fitted value for the i^{th} case:

$$\hat{\pi}(x_i) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_i}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_i}} \quad 2.20$$

and the fitted logit response function is:

$$\hat{g}(x) = \ln \left[\frac{\hat{\pi}(x)}{1 - \hat{\pi}(x)} \right] \quad \text{where} \quad \hat{g}(x) = \ln \left[\frac{\hat{\pi}(x)}{1 - \hat{\pi}(x)} \right] \quad 2.21$$

2.3.1.3 Testing for the Significance of the Coefficients

After the coefficient estimates of the variables in the model are conducted, whether independent variables have a significant relation with outcome variable is determined. One approach to testing for the significance of the coefficient of a variable in any model relates to the following question: Does the model that includes the variable in question tell us more about the outcome variable than does a model that does not include that variable? This question is answered comparing the observed values of the response variable to those predicted by each of two models; the first with and the second without the variable in question. If the predicted values with the variable the model are better or more accurate in some sense, than when the variable not in the model, then we feel that the variable in question is “significant” (Hosmer and Lemeshow, 1989)

There are 3 basic tests to determine the significance of the variables in the logistic model. These tests are Likelihood ratio test, Wald test and Score test respectively.

2.3.1.3.1 Likelihood Ratio Test. Likelihood ratio is a significance test based on the likelihood function defined in equation 2.14. It tests whether a current model which is the model without the variable in question as good as the saturated model that is the model including all the variables. The likelihood ratio test is calculated as twice the difference between the saturated model and the current model. The likelihood ratio has approximately chi-square distribution with degrees of freedom which is equal to difference in the number of parameters in the two models. The comparison observed to predicted values using the likelihood function is based on the following expression:

$$D = -2\ln \left[\frac{\text{likelihood of the current model}}{\text{likelihood of the saturated model}} \right] \quad 2.22$$

The statistic D is called the deviance and plays the same role as the residual sum of squares in linear regression. Using equation 2.14 and 2.22 it becomes:

$$D = -2 \left[\sum_{i=1}^n Y_i \ln \left(\frac{\hat{\pi}(x_i)}{Y_i} \right) + (1 - Y_i) \ln \left(\frac{1 - \hat{\pi}(x_i)}{1 - Y_i} \right) \right] \quad 2.23$$

The assessment of significance of a variable in question we compare the value of D with and without the variable in the equation. This is obtained as follows:

$$G^2 = D(\text{for the model without the variable}) - D(\text{for the model with the variable})$$

$$\begin{aligned} G^2 &= -2 \ln \left[\frac{\text{likelihood of the current model without the variable}}{\text{likelihood of the saturated model}} \right] \\ &+ 2 \ln \left[\frac{\text{likelihood of the current model with the variable}}{\text{likelihood of the saturated model}} \right] \\ G^2 &= -2 \ln \left[\frac{\text{likelihood without the variable}}{\text{likelihood with the variable}} \right] \end{aligned} \quad 2.24$$

The statistic G^2 plays the same role in logistic regression with the partial F test in linear regression. If the p value is associated with this test is less than the alpha level then the null hypothesis is rejected that is the variable has a significant relationship with the response variable.

In a case where the single independent variable is not in the model, the maximum likelihood estimate of β_0 is $\ln(n_1 / n_0)$ where $n_1 = \sum y_i$ and $n_0 = \sum (1 - y_i)$ and the predicted value is constant n_1 / n . In this case the value of G^2 is as follows:

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

$$G^2 = 2 \left\{ \sum_{i=1}^n Y_i \ln \left(\frac{\hat{\pi}(x_i)}{Y_i} \right) + (1 - Y_i) \ln \left(\frac{1 - \hat{\pi}(x_i)}{1 - Y_i} \right) \right\} \quad 2.26$$

2.3.1.3.2 Wald Test. The other test for significance for variable in question is the Wald test. It tests whether a independent variable has a significant relationship with the dependent variable. To do so the Wald test statistic is obtained using maximum

likelihood estimate of the slope parameter $\hat{\beta}_1$ divided by its standard error. The ratio, under the hypothesis that β_1 is equal to zero ($H_0: \beta_1 = 0$), will follow a standard normal distribution. Standard error of $\hat{\beta}_1$ is obtained from the square root of corresponding diagonal element of the covariance matrix, $V(\hat{\beta})$. The test statistic is:

$$W = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \quad \text{where} \quad SE(\hat{\beta}_1) = \sqrt{V(\hat{\beta}_1)} \quad 2.27$$

For this test, two tailed p value is evaluated by $P(|Z| > W)$. If the p value is less than the alpha level then the null hypothesis is rejected.

An alternative form of the Wald statistic is a square wald statistic has a chi-square distribution with one degrees of freedom.

$$(W)^2 = \left(\frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \right)^2 \quad 2.28$$

Both the likelihood ratio test G^2 , and the Wald test, W , require the computation of the maximum likelihood estimate for β_1 . A test for the significance of a variable which does not require these computations is Score test.

2.3.1.3.3 Score Test. The other test for the significance of the coefficient is Score test. The test statistic for the Score test is:

$$ST = \frac{\sum_{i=1}^n x_i (y_i - \bar{y})}{\sqrt{\bar{y}(1 - \bar{y}) \sum_{i=1}^n (x_i - \bar{x})^2}} \quad 2.29$$

Under the hypothesis that β_1 is equal to zero, the test statistic has standard normal distribution. It can be used z to the standard normal table to obtain two-tailed p value. If the p value is less than the alpha level then the null hypothesis is rejected.

2.4 Multiple Logistic Regression Model

Let a collection of p independent variables denoted by the vector $x' = (x_1, x_2, \dots, x_p)$. If we assume that all the independent variables at least interval scaled a model for single independent variable in equation 2.8 can be extended for multiple logistic regression model as follows:

$$\pi(x) = \frac{e^{\beta'x}}{1 + e^{\beta'x}} \quad \text{where} \quad \beta'x = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p \quad 2.30$$

and the logit response function is:

$$g(x) = \frac{e^{\beta'x}}{1 + e^{\beta'x}} \quad \text{where} \quad \beta'x = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p \quad 2.31$$

Like the simple logistic response function in equation 2.8, multiple logistic response function in equation 2.30 is monotonic and sigmoidal in shape. Also the predictor variables may be quantitative or qualitative which is represented by indicator variables. This flexibility makes the multiple logistic regression model very attractive.

2.4.1 Dummy Variable

It is not appropriate to include nominal and ordinal scaled variables as if they were interval scaled variables, because the code values are not meaningful numerically. In this situation the method of choice is to use a collection of dummy variables. In general, if a nominal scaled variable has k possible values, then $k-1$ dummy variables will be created.

Suppose let one of the independent variables is “marital status” which has been categorized as single, married and the other. In this situation two dummy variables are generated. One possible coding strategy is that when the respondent is “married” two dummy variables, D_1 and D_2 , would both be set equal to zero; when the respondent is “single” D_1 would be set equal to 1 while D_2 would still equal 0; when

the respondent is “other” we would use $D_1 = 0$ and $D_2 = 1$ would still equal 0. Here, the reference group is the group whose both dummy variables are 0.

Table 2.2 The coding of dummy variables of “marital status”

Dummy Variable		
Marital Status	D_1	D_2
Married	0	0
Single	1	0
Other	0	1

If that the j^{th} independent variable, X_j has k_j levels. The $k_j - 1$ dummy variables will be denoted as D_{ju} and the coefficients for these dummy variables will be denoted as B_{ju} $u = 1, 2, \dots, k_j - 1$. The formulation of the logit for a model with p variables and the j^{th} variable being discrete would be:

$$g(x) = \ln \left(\frac{P_1}{P_0} \right) = \beta_0 + \sum_{j=1}^p \beta_j x_j + \sum_{j=1}^p \sum_{u=1}^{k_j-1} \beta_{ju} D_{ju} \quad 2.32$$

2.4.2 Fitting the Multiple Logistic Regression Model

Assume that we have a sample of n independent observations of the pair (x_i, y_i) , $i = 1, 2, \dots, n$. As in the univariate case, fitting the model requires that we obtain estimates of the vector $\beta' = (\beta_0, \beta_1, \dots, \beta_p)$. We utilize the method of maximum likelihood to estimate the unknown parameters.

2.4.2.1 Likelihood Function

The log-likelihood function for simple logistic regression in 2.15 extends directly for multiple logistic regression:

$$\ln L(\beta) = \sum_{i=1}^n y_i \ln(\pi_i) + \sum_{i=1}^n (1 - y_i) \ln(1 - \pi_i) \quad 2.33$$

There will be $p+1$ equations which are obtained by differentiating the log likelihood function with respect to the $p+1$ coefficients. The likelihood equations that result may be expressed as follows:

$$\sum_{i=1}^n [Y_i - \pi(x_i)] \quad 2.34$$

$$\sum_{i=1}^n x_{ij} [y_i - \pi(x_i)] \quad j=1,2,\dots,p \quad 2.35$$

2.4.2.2 Fitted Multiple Logistic Regression Model

Numerical search procedures are used to find values of $\beta' = (\beta_0, \beta_1, \dots, \beta_p)$ that maximize the likelihood function. The fitted values of the multiple logistic regression model is denoted by $\hat{\beta}' = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)$.

The fitted multiple logistic response function as follows :

$$\hat{\pi}(x) = \frac{e^{(\hat{\beta}'x)}}{1 + e^{(\hat{\beta}'x)}} = \left[1 + e^{-(\hat{\beta}'x)} \right]^{-1} \quad 2.36$$

The fitted value for the i th case as follows :

$$\hat{\pi}(x_i) = \frac{e^{(\hat{\beta}'x_i)}}{1 + e^{(\hat{\beta}'x_i)}} \quad 2.37$$

The fitted multiple logit response function is as follows :

$$\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_p x_{p-1} \quad 2.38$$

2.4.2.3 Testing for the Significance of the Coefficients

After fitting the multiple logistic regression model, the first step is to determine the significance of the variables in the model. As in the univariate case, there are three basic tests to determine the significance of the variable in question.

2.4.2.3.1 Likelihood Ratio Test. The same procedure is performed for the multivariate case as in the univariate case. The only difference there is $p+1$ parameters to be estimated. The likelihood ratio G^2 statistic is used for comparing models. G^2 has a chi-square distribution with $v_2 - v_1$ degrees of freedom which v_2 is the number of variables of saturated model plus one and v_1 is the number of variables of reduced model plus one. To assess the significance of the model the null and the alternative hypothesis are stated as follows:

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

$$H_1: \text{At least one of the } \beta_p \neq 0$$

The test statistic G^2 is calculated as follows:

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

Alternatively the following equation can be used for computing the G statistic

$$G^2 = 2 \left\{ \sum_{i=1}^n Y_i \ln \frac{\hat{\pi}_i}{\pi_i} \right\} = 2 \left\{ \sum_{i=1}^n Y_i \ln \frac{\hat{\pi}_i}{\pi_i} \right\} = 2 \left\{ \sum_{i=1}^n Y_i \ln \frac{\hat{\pi}_i}{\pi_i} \right\}$$

For G^2 statistic, the decision rule is that p value is $P \left\{ \chi^2_{(1-(1-\alpha)(v_2-v_1))} \geq G^2 \right\}$. If the p value is less than the alpha level then H_0 is rejected and it is concluded that at least one and perhaps all coefficients are different from zero.

2.4.2.3.2 Wald Test. Under the hypothesis that β_j is equal to zero ($H_0: \beta_j = 0$), these statistics will follow the standart normal distribution. P value can be defined by $P(|Z| > W)$. If the p value is less than the alpha level, H_0 is rejected.

The Wald statistic is as follows:

$$W_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \quad 2.39$$

The multivariate form of the Wald test is obtained from the following vector-matrix calculation.

$$W = \hat{\beta}' \left[\begin{array}{c} \text{---} \\ \text{---} \\ \text{---} \end{array} \right] \quad 2.40$$

which is distributed as chi-square with $p+1$ degrees of freedom. Tests for just the p slope coefficients are obtained by eliminating $\hat{\beta}_0$ from $\hat{\beta}$ and the relevant row (first) and column (first) from $X'VX$.

The next step is to determine whether the reduced model is as good as the full model (model contains all the variables). For this comparison, the G^2 statistic with $v_2 - v_1$ degrees of freedom is used. If the p value for the G^2 statistic exceeds 0.05, we conclude that the reduced model is as good as the full model.

2.4.2.3.3 Score Test. The multivariate form of the Score test is based on the conditional distribution of the p derivatives of $L(\cdot)$ with respect to $\hat{\beta}$. The computation of this test is of the same order of complication as the Wald test.

2.4.3 Confidence Interval of the Coefficients

The method of estimating the variances and covariances of the estimated coefficients follows from well developed theory of maximum likelihood estimation. This theory states that the estimators are obtained from the matrix of second partial derivatives of the log-likelihood function (Hosmer and Lemeshow, 1989).

If we let

$$X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & & & \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix}_{n \times (p+1)} \quad \text{and}$$

$$V = \begin{pmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & 0 & \dots & 0 \\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & \dots & 0 \\ \vdots & & & \\ 0 & \dots & \dots & \hat{\pi}_n(1-\hat{\pi}_n) \end{pmatrix}_{n \times n}$$

where X is $n \times (p+1)$ matrix containing the data for each subject and V is an $n \times n$ diagonal matrix with general element $\hat{\pi}_i (1 - \hat{\pi}_i)$

$$\hat{I}(\hat{\beta}) = X' V X \quad 2.41$$

$\hat{I}(\hat{\beta})$ is a size of $(p+1)$ by $(p+1)$ matrix called information matrix. The estimated variance covariance matrix is the inverse of the information matrix. The estimated variance is denoted as follows:

$$\text{Var}(\hat{\beta}) = [\hat{I}(\hat{\beta})]^{-1} \quad 2.42$$

Confidence interval of the estimated coefficients are denoted as follows:

$$\hat{\beta}_j \pm Z_{1-\alpha/2} \text{SE}(\hat{\beta}_j) \quad \text{SE}(\hat{\beta}_j) = \sqrt{\text{Var}(\hat{\beta}_j)} \quad 2.43$$

2.5 Interpretation of the Coefficients

The estimated coefficients for the independent variables represent the slope or rate of change of a function of the dependent variable per unit of change in the independent variable.

Interpretation of the coefficients involves the following two steps: Firstly, by determining the linear functional relationship between the dependent variable and the independent variable. This is called the link function. In the logistic regression model the link function is the logit transformation $g(x) = \ln\left(\frac{\pi}{1-\pi}\right)$. Then, defining the unit of change for the independent variable. In linear regression the slope coefficient β_1 is the value that the difference between the value of the dependent variable that is taken at $x+1$ and x for any value of x .

In the logistic regression model $\beta_1 = g(x+1) - g(x)$, that is the slope coefficient represents the change in the logit for a change of one unit in the independent variable of x .

2.5.1 Dichotomous Independent Variable

We assume that x is coded as 0 or 1. In this situation there are two values of $\pi(x)$ and equivalently two values of $1 - \pi(x)$.

Table 2.3 Values of the logistic model when the independent variable is dichotomous

	X=1	X=0
Y=1	$\pi(1) = \frac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}}$	$\pi(0) = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$
Y=0	$1 - \frac{1}{1 + e^{\beta_0 + \beta_1}}$	$1 - \frac{1}{1 + e^{\beta_0}}$

The odds of the outcome being present among individuals with $x=1$ is denoted as $\pi(1)/1 - \pi(1)$ and the odds of the outcome being present among individuals with $x=0$ is denoted as $\pi(0)/1 - \pi(0)$.

The odds ratio is the ratio of the odds for $x=1$ to the odds for $x=0$, denoted by θ and given by the following equation:

$$\theta = \frac{\pi(1)/1 - \pi(1)}{\pi(0)/1 - \pi(0)} \quad 2.44$$

The odds ratio can equal any nonnegative number. When X and Y are independent $\pi(1) = \pi(0)$ and $\theta = 1$. When $\theta > 1$, the odds of success are higher when $x=1$ than $x=0$. For instance, when $\theta = 3$, the odds of success for $x=1$ are three times the odds of success for $x=0$. Thus, subjects for $x=1$ are more likely to have successes than are subjects for $x=0$; that is, $\pi(1) > \pi(0)$. When $\theta < 1$, a success is less likely for $X=1$ than for $X=0$ that is, $\pi(1) < \pi(0)$.

The log of the odds are as follows:

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

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

The log of the odds ratio termed log-odds ratio or log-odds, is

$$\ln\left(\frac{\pi(1)/1-\pi(1)}{\pi(0)/1-\pi(0)}\right)$$

$$\ln\left(\frac{e^{\beta_0+\beta_1}}{e^{\beta_0}}\right)$$

which is the logit difference.

Using the expressions for the logistic regression model shown in Table 2.3 the odds ratio is:

$$\theta = \left[\frac{\pi(1)/1-\pi(1)}{\pi(0)/1-\pi(0)} \right] = \left[\frac{\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}} \frac{1}{1+e^{\beta_0}}}{\frac{1}{1+e^{\beta_0+\beta_1}} \frac{e^{\beta_0}}{1+e^{\beta_0}}} \right] = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

$$\theta = e^{\beta_1} \quad 2.45$$

and the logit difference or log odds is, $\ln \theta = \beta_1$.

In 2x2 table the sample odds ratio also equals the ratio of the sample odds in the two rows, which is:

$$\hat{\theta} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}} \quad 2.46$$

When the sample size is not large enough the sampling distribution of the odds ratio is skewed. Because of this skewness, statistical inference for the odds ratio uses an alternative measure its natural logarithm, $\ln \hat{\theta}$. The sample log odds ratio has normal distribution with mean of $\ln \hat{\theta}$ and a standard error of $\ln \hat{\theta}$ is:

$$\hat{SE} = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} \quad 2.47$$

Because the sampling distribution of $\ln \hat{\theta}$ is closer to normal distribution than the sampling distribution of $\hat{\theta}$ it is better to construct confidence intervals for $\ln \hat{\theta}$ and exponentiating endpoints of this confidence interval to obtain limit of the $\hat{\theta}$. The confidence interval is:

$$\left[\hat{\beta}_{ij} \pm z_{1-\alpha/2} \hat{SE}(\hat{\beta}_{ij}) \right] \quad 2.48$$

$$\exp \left[\hat{\beta}_{ij} \pm z_{1-\alpha/2} \hat{SE}(\hat{\beta}_{ij}) \right] \quad 2.49$$

where i is the reference group subscript and j is the subscript of the group. If the confidence interval for $\hat{\theta}$ does not contain 1, it is the odds of outcome being different for each group.

2.5.2 Polytomous Independent Variable

In the event that nominal scaled independent variable consist of more than 2 level ($k > 2$), independent variable is called polytomous. Since it is inappropriate to model a nominal scaled variable as if it were interval scaled, $k-1$ dummy variables are created. The dummy variables created for a polytomous independent variable with a four-level “marital status” and the group of “married” being chosen as the reference is shown on Table 2-4.

Table 2.4 The coding of dummy variables when the independent variable is polytomous

Dummy Variable			
Marital Status	D_1	D_2	D_3
Married	0	0	0
Single	1	0	0
Divorced	0	1	0
Other	0	0	1

The hypothetical summarized data, which is about the study where the relationship between having a heart attack and marital status is examined, is shown in Table 2-4.

Table 2-5:Hypothetical data on marital status and having a heart attack for 100 subjects

	Married	Single	Divorced	Other	Total
Present	4	5	10	12	31
Absent	20	18	16	15	69
Total	24	23	26	27	100
$\hat{\theta}$	1,0	1,39	3,13	4,0	
$\ln(\hat{\theta})$	0,0	0,33	1,14	1,39	

Odds ratio values of the levels of the independent variable can be found by choosing a reference group with the help of the frequencies in the cells without using the likelihood function. For example the estimated odds ratio for the “Single” group is : $(5*20)/(18*4)=1,39$

When the likelihood function is used estimated coefficients are equal to the values that are obtained from cross-classification table. Comparing the married and single groups when the design variables on Table 2.4 are used, the equation can be written as follows:

$$\begin{aligned}
 \ln \hat{\theta} &= \ln \left(\frac{f_{11} f_{02}}{f_{12} f_{01}} \right) \\
 &= \left[\begin{matrix} \ln f_{11} & \ln f_{02} & \ln f_{12} & \ln f_{01} \end{matrix} \right] - \left[\begin{matrix} \ln f_{01} & \ln f_{11} & \ln f_{02} & \ln f_{12} \end{matrix} \right] \\
 &= \begin{matrix} \ln \\ 11 \end{matrix}
 \end{aligned}$$

Table 2.6 Results of fitting the logistic regression model to the hypothetical data in table 2.5 using the design variables in table 2-4

Variable	B	S.E	Wald	Exp(B)
Single	0,329	0,745	0,194	1,389
Divorced	1,139	0,680	2,807	3,125
Other	1,386	0,671	4,271	4,00
Constant	-1,609	0,548	8,634	0,20

The process of the computation of the standard error by using a cross-classification table is the same as the univariate case. For example the standard error for the Single group is : $= (1/4 + 1/5 + 1/18 + 1/20)^{1/2} = 0,75$

Again first the confidence intervals for $\ln \hat{\theta}$ is found and exponentiating endpoints of this confidence interval to obtain limit of the $\hat{\theta}$. The confidence interval is:

$$\left[\hat{\beta}_{ij} \pm z_{1-\alpha/2} \cdot \hat{SE}(\hat{\beta}) \right] \quad 2.50$$

$$\exp \left[\hat{\beta}_{ij} \pm z_{1-\alpha/2} \cdot \hat{SE}(\hat{\beta}) \right] \quad 2.51$$

where i is the reference group subscript and j is the subscript of the group.

2.5.3 Continuous Independent Variable

Under the assumption that the logit is linear in the continuous covariate, X, then it is expressed as $g(x) = \beta_0 + \beta_1 x$ and β_1 represents the change in log odds ratio for an increase of 1 unit in x. It is shown as follows:

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

$$g(x) = \beta_0 + \beta_1 x$$

$$g(x+1) - g(x) = \beta_1$$

It is important that the interpretation of the coefficient of the continuous independent variable depends on the unit. For example an increase of 1 year age or 1 mm-Hg in systolic blood pressure may not be a meaningful. But a change of 5 years or 10 mm/Hg may be more meaningful. The log odds for a change of c units in x is obtained from the logit difference $g(x+c) - g(x)$ and the associated odds ratio is obtained by exponentiating this logit difference,

$$g(x+c) = \beta_0 + \beta_1(x+c)$$

$$g(x) = \beta_0 + \beta_1 x$$

$$g(x+c) - g(x) = c \beta_1 \quad 2.52$$

$$\theta(c) = \theta(x + c, x) = \exp(c\beta_1) \quad 2.53$$

An estimate may be obtained by replacing β_1 with its maximum likelihood estimate $\hat{\beta}_1$. Standard error estimation is obtained by multiplying the estimated standard error $\hat{SE}(\hat{\beta}_1)$ by c .

The confidence intervals for $\theta(c)$ are:

$$\exp\left[\hat{c} \pm \hat{SE}(\hat{\beta}_1) \cdot c\right] \quad 2.54$$

2.5.4 Multivariate Case

There is a multivariate case in models where there are more than one type of scaled variable. One goal of such an analysis is to statistically adjust the estimated effects of each variable in the model for differences in the distributions of and associations among the other independent variables. (Hosmer and Lemeshow, 1989)

Let say we have one dichotomous (X_1) which is coded 0 and 1 and one continuous (X_2), two variable multivariate model. It can be written as follows:

$Y = B_0 + B_1X_1 + B_2X_2$. Our primary interest is focused on the effect of the dichotomous variable. It would not be possible to determine the effect of group without first eliminating the discrepancy in continuous independent variable between groups.

Suppose the mean value for the continuous independent variable for group one and two are respectively $\bar{a}_1$ and $\bar{a}_2$. The statistical model where $x=0$ for group one is $y_1 = B_0 + B_1(x=0) + B_2\bar{a}_1$, and the statistical model where $x=1$ for group two is $y_2 = B_0 + B_1(x=1) + B_2\bar{a}_2$. The difference between the groups is as follows:

$$y_2 - y_1 = B_0 + B_1(x = 1) + B_2\bar{a}_2 - (B_0 + B_1(x = 0) + B_2\bar{a}_1)$$

$$y_2 - y_1 = B_1 + B_2(\bar{a}_2 - \bar{a}_1) \quad 2.55$$

As we can see comparison involves not only the true difference between two groups β_1 , but a component $\beta_2(\bar{a}_2 - \bar{a}_1)$. The process of statistically adjusting for continuous variable involves comparing the two groups at some common value of that variable. The value usually used is the mean of the two groups which for example is denoted by $\bar{a}$. In terms of the model this yields a comparison of y_4 to y_3 .

$$y_4 - y_3 = \beta_1 \quad 2.56$$

Here the β_1 is the true difference of the two groups.

Consider the same situation when the outcome variable being is dichotomous. That is under the model the logit, the logit difference of the groups is given by the equation as follows:

$$g(x = 1, \bar{a}) - g(x = 0, \bar{a}) = \beta_1 \quad 2.57$$

It is shown with an example how effect of the continuous variable is adjusted in Table 2.7 and Table 2.8.

Table 2.7 Descriptive statistics for the two groups on AGE and dieting (1=yes, 0=no)

Variable	Group 1		Group 2	
	Mean	SD	Mean	SD
Diet	0,30	0,46	0,80	0,40
Age	40,18	5,34	48,45	5,02

The univariate log odds ratio for group 2 versus group 1 is:

$$\ln(\hat{\theta}) = \frac{y_2 - y_1}{x_2 - x_1} = \frac{0,80 - 0,30}{48,45 - 40,18} = 0,0934$$

The unadjusted estimated odds ratio is: $\hat{\theta} = 9,34$

We can also see that there is a considerable difference in age distribution of two groups. Does much of the difference between the two groups due to age?

Analyzing the data with a bivariate model using a coding of 0 for group 1 and 1 for group 2, we obtain the regression coefficients shown in Table 2.8.

Table 2.8 Results of fitting the logistic regression model to the data summarized in Table 2-7

Variable	B	SE	Wald
Group	1,559	0,557	2,80
AGE	0,096	0,048	2
Constant	-4,379	1,998	-2,37

Here the age adjusted odds ratio is $\hat{\theta} = e^{1,559} = 4,75$. It is seen that much of the apparent difference between the two groups is due to differences in age.

The unadjusted odds ratio is obtained by exponentiating the difference $y_2 - y_1$. In terms of the fitted logistic regression model shown in Table 2.8 this difference is $y_2 - y_1 = B_1 + B_2(\bar{a}_2 - \bar{a}_1) = 1,559 + 0,096(48,45 - 40,18)$ and the value of the odds ratio is $e^{[1,559 + 0,096(48,45 - 40,18)]} = 10,48$. The age adjusted odds ratio is obtained by exponentiating the difference $y_4 - y_3$, which is equal to the estimated coefficient for group. In the example this difference is: $y_4 - y_3 = 1,559 + 0,096(44,32 - 44,32) = 1,559$

The method of adjustment when the variables are all dichotomous, polytomous, continuous, or a mixture of these is identical to that just explained for the dichotomous-continuous variable case.

2.6 Model Building Strategies and Methods

The number of variables thought to be significant within the scientific concept of the problem may be too large. But as the variables included in a model increase, so the estimated standard errors become larger and dependent the model becomes more on the observed data.

The goal of the model building is to seek the most parsimonious model that still explains the data. The variable selection methods for multiple logistic regression model are Univariate Analysis and Multivariate Analysis. Two different techniques are used for multivariate analysis: stepwise logistic regression and best subset logistic regression method. Stepwise logistic regression is conducted with in two different ways, namely forward selection and backward elimination.

2.6.1 Univariate Analysis

The selection process begins with the univariate analysis of each variable. For categorical (ordinal and nominal) and continuous variables with few integer values, the univariate analysis is done with contingency table of outcome ($y=0,1$) versus the k levels of the independent variable. The likelihood ratio chi-square test with $k-1$ degrees of freedom is exactly equal to 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 that single independent variable. (Hosmer and Lemeshow, 1989).

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.

Variable selection process with univariate analysis starts with testing the meaningfulness of each variables. As a result of univariate analysis while the variables which has the p value is smaller than 0,25 are chosen as a candidate variable for multivariate model, the variables greater than 0,25 are excluded from the multivariate model.

Special attention should be placed to any contingency table with a zero cell. This will produce a univariate point estimate for one of the odds ratios of either zero or infinity. Strategies for dealing with the zero cell include: collapsing the categories of the independent variable in some sensible fashion to eliminate the zero cell;

eliminating the category completely; or if the variable is ordinal scaled, modeling the variable as if it were continuous.

As a result of univariate analysis the following parameters are found: (1) estimated slope coefficient(s) for the univariate logistic model containing only that variable, (2) standard error estimation of the slope coefficient, (3) the likelihood ratio test statistic (G), (4) p value of the likelihood ratio test statistic, (5) the estimated odds ratio, (6) the 95% confidence limit for the odds ratio are obtained.

Following the fit of the multivariate model, the importance of each variable included in the model should be verified. This should include (a) an examination of the Wald statistic for each variable and (b) a comparison of each estimated coefficient with the coefficient from the univariate model containing only that variable. Variables that do not contribute to the model based on these criteria should be eliminated and a new model fit. The new model should be compared to the old model through the likelihood ratio test. Also the estimated coefficients for the remaining variables should be compared to those from the full model. In particular we should be concerned about variables whose coefficients have changed markedly in magnitude. This would indicate that one or more of the excluded variables was important in the sense of providing a needed adjustment of the effect of the variable that remained in the model. This process of deleting, refitting and verifying continues until it appears that all of the important variables are included in the model and those excluded are either biologically or statistically unimportant. (Hosmer and Lemeshow, 1989).

2.6.2 Stepwise logistic regression

Stepwise logistic regression is widely used procedure for model building in cases where there is a large number of potential independent variables. There are two main versions of the stepwise procedure. Forward selection and backward elimination.

This method is same as that is used in linear regression. Although forward selection and backward elimination procedures have different criteria for deciding which variables are selected. This tests are based on likelihood ratio G^2 test statistic. At each stage, the variable giving the greatest improvement in the fit is selected.

Since the magnitude of G^2 depends on its degrees of freedom, any procedure based on the likelihood ratio test statistic, G^2 , must account for possible differences in degrees of freedom between variables. This is done by assessing significance through the p value for G^2 .

The forward selection procedure begins with no variable in the model. At each stage the most significant variable is considered for added to the model. In the large likelihood ratio means small p value indicates that the variable should be included. The backward elimination procedure begins with all the variable in the model. At each stage the least significant variable is considered for elimination.

Since it is possible that once a variable has been added to the model, other variable(s) that previously added might not be important anymore . Thus, forward selection includes a check for backward elimination. In general this is accomplished by fitting models that delete one of the variables added in the previous steps and assessing the continued importance of the variable removed. These two processes continue until further additions or eliminations do not improve the fit.

The most important disadvantage of stepwise selection is the necessity of calculating maximum likelihood estimates all of the variables at every stage for the coefficients which will not be present in the final model. For large data files with large numbers of variables this can be quite expensive both as far as time and money are concerned.

2.6.3 Best Subsets Selection Method

One of the problems of the univariate approach is while the relationship between outcome variable and predictors is not significant when they are univariate, it may be significant predictor when taken together. The best subset selection technique is an effective model building strategy for identification of collection of variables having this type of association with the outcome variable.

2.7 Assessing the Fit of the Model

After the model building stage is completed we would like to know how effective the model we have is in describing the outcome variable. It will be determined with goodness of fit tests. These tests are deviance test and Hosmer-Lemeshow test.

The observed sample values of the outcome variable in vector form as y where $y' = (y_1, y_2, \dots, y_n)$. We denote the values predicted by the model or fitted values, 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 unsystematic and it is small relative to the error structure of the model, the fitted model is accepted well. The fitted model contains p independent variables $(x^* = (x_1, x_2, \dots, x_p))$ and j denote the number of distinct values of x observed. 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 $\sum m_j = n$ ($j=1, 2, \dots, j$). Let y_j denote 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 $\sum_j y_j = n_1$.

2.7.1 Pearson Chi-Square and Deviance

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

$$\hat{m}_j = \frac{\exp(\hat{g}(x_j))}{1 + \exp(\hat{g}(x_j))} \quad 2.58$$

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

There are two measures of the difference between the observed and fitted values: the Pearson residual and deviance residual. For a particular covariate pattern the Pearson residual is defined as follows:

$$r(y_j, \hat{m}_j) = \frac{(y_j - \hat{m}_j)}{\sqrt{\hat{m}_j(1 - \hat{m}_j)}} \quad 2.59$$

The summary statistic based on these residuals is the Pearson chi-square statistic which is as follows:

$$\chi^2 = \sum_{j=1}^J r(y_j, \hat{m}_j)^2 \quad 2.60$$

Also the deviance residual is defined as follows:

$$d(y_j, \hat{m}_j) = \left\{ \left[\frac{y_j}{\hat{m}_j} - \left(\frac{y_j}{\hat{m}_j} - \frac{(m_j - y_j)}{\hat{m}_j(1 - \hat{m}_j)} \right) \right] \right\}^{1/2} \quad 2.61$$

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

$$d(y_j, \hat{m}_j) = \sqrt{\left| \frac{y_j}{\hat{m}_j} - \left(\frac{y_j}{\hat{m}_j} - \frac{(m_j - y_j)}{\hat{m}_j(1 - \hat{m}_j)} \right) \right|} \quad 2.62$$

And the deviance residual when $y_j = m_j$, is

$$d(y_j, \hat{m}_j) = \sqrt{\left| \frac{y_j}{\hat{m}_j} - \left(\frac{y_j}{\hat{m}_j} - \frac{(m_j - y_j)}{\hat{m}_j(1 - \hat{m}_j)} \right) \right|} \quad 2.63$$

The summary statistic based on the deviance residuals is the deviance

$$D = \sum_{j=1}^J d(y_j, \hat{y}_j)^2 \quad 2.64$$

Under the assumption that the fitted model is correct, the distribution of the statistics χ^2 and D will follow chi-square with degrees of freedom equal to $J - (p + 1)$

2.7.2 The Hosmer-Lemeshow Tests

The Hosmer-Lemeshow tests are proposed grouping based on the values of the estimated probabilities. In this case J is equal to n and there are n columns as corresponding to the n values of the estimated probabilities, with the first column corresponding to the smallest value and the n^{th} column to the largest value. The grouping strategies were proposed: as follows: (a) Collapse the table based on percentiles of the estimated probabilities (b) collapse the table based on fixed values of the estimated probability.

For the first method, use of $g=10$ results in the number of first group is $n_1^* = n/10$ and the number of last group is $n_{10}^* = n/10$. The subjects in the first group have the smallest estimated probabilities and the subjects in the last group have the largest estimated probabilities. For the second method use of $g=10$ groups results in cutpoints defined at the values $k/10$, $k=1,2,\dots,9$, and the groups contain all subjects with estimated probabilities between adjacent points. For the $y=1$ row, estimates of the expected values are obtained by summing the estimated probabilities over all subjects in a group. For the $y=0$ row, the estimated expected value is obtained by summing, over all subjects in the group, one minus the estimated probability.

The Hosmer-Lemeshow goodness of fit statistic, $\hat{C}$, is obtained by calculating the Pearson chi-square statistic from $2 \times g$ table of observed and estimated expected frequencies. The statistic of $\hat{C}$ is:

$$\hat{C} = \sum_{k=1}^g \frac{(o_k - n_k \cdot \bar{p}_k)^2}{n_k \cdot \bar{p}_k (1 - \bar{p}_k)} \quad 2.65$$

where n'_k is the number of covariate patterns in the k^{th} group.

$$o_k = \sum_{j=1}^{n'_k} y_j \quad 2.66$$

where o_k is the number of responses among the n'_k covariate patterns. Also $\bar{\pi}_k$ is the average estimated probability and it is calculated as:

$$\bar{\pi}_k = \sum_{j=1}^{n'_k} m_j \hat{\pi}_j / n'_k \quad 2.67$$

The distribution of the statistic of $\hat{C}$ is well approximated by the chi-square distribution with $g-2$ degrees of freedom when j is equal to n . If the value of the $\hat{C}$ 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.

CHAPTER THREE

ARTIFICIAL NEURAL NETWORK

Parallel to the advancements in technology it has been witnessed that computers which were initially used merely to transfer electronic data and perform complex computations gained in the course of time new features. A variety of performances involving intelligence or pattern recognition are extremely difficult to make automated yet they seem to be performed very easily by animals. Natural neural networks are highly complex, nonlinear systems which allow great degrees of freedom that employ a wide array of information processing from those of computers. It seems feasible that computing systems that attempt similar tasks and also simulating these processes to the extent allowed by physical limitations. This in turn necessitates the study and simulation of neural networks.

Artificial neural networks are also titled as "neural nets," "artificial neural systems" "parallel distributed processing" and "connectionist models". A neural network represents a highly parallelized dynamic system which has a directed graph topology able to receive the output information through a reaction of its state on the input actions. Processor elements are described as nodes or neurons of the neural network. The input to a node may come from other nodes or it can also come directly from the input data.

The areas artificial neural networks are commonly employed for diagnosis, classification, prediction, control, data filtering and interpretation can be listed as industrial applications, financial applications, military and defense applications as well as health-care applications.

In this part the foundations and development of artificial neural networks, the structure and basic components of artificial neural networks, architecture of artificial neural networks and learning strategies have been explained respectively.

3.1 History of Neural Networks

The modern perspective of neural networks was initiated with the study of Warren McCullough and Walter Pitts in 1943. They made us see that networks of artificial neurons could, in theory, compute any arithmetic or logical function. They have developed the first mathematical model of a single input neuron. This model has been modified and widely applied in subsequent work. Warren McCullough and Walter Pitts are followed by Donald Hebb who proposed a mechanism for learning in biological neurons. Hebb's (1949) learning rule incrementally modifies connection weights by examining whether two connected nodes are simultaneously ON or OFF.

In 1958, Frank Rosenblatt and his colleagues invented the perceptron network and associated learning rule along with first practical application that was introduced. They built a perceptron network and demonstrated its ability to perform pattern recognition. A perceptron element consists of a single node which receives weighted inputs and thresholds the results according to a rule. The perceptron is able to classify linearly separable data but is unable to handle nonlinear data.

At about the same time, Bernard Widrow and Ted Hoff introduced a new learning algorithm and use it to train adaptive linear neural networks which were similar in structure and capability to Rosenblatt's perceptron. For two decades, development in neural networks was slow as a result of the inability to find efficient methods to solve non-linearly separable problems.

In 1980 there was a rise of interest towards neural networks parallel to the increase in computing power and the development of several new algorithms and network topologies. One of these was the backpropagation multi layer perceptron (MLP) algorithm which is described by David Rumelhart and James McClelland in 1986. The MLP algorithm is still widely used today.

3.2 Biological Neural Networks

A typical biological neuron is composed of a cell body, an axon, synapses and a number of root-like dendrites which surround the body of the neuron, as illustrated in Figure 3.1. The cell body of the neuron, which integrates the neuron's nucleus is where most of the neural computation takes place.

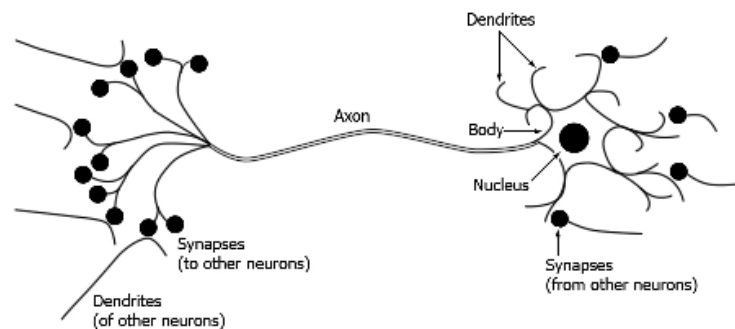


Figure 3.1 A typical biological neuron

Synapses can be identified as the connections between neural cells. These are not physical connections but rather spaces that enable transmission of electrical signals from one cell to another. These signals reach the cell where they are processed. Neural cell constructs its own electrical signal and passes it to dendrites by means of axon. Dendrites then pass these signals to synapses to enable the transmission of message to the rest of cells.

Neural activity passes from one neuron to different one in terms of electrical triggers that can travel from one cell to the another down the neuron's axon.

3.3 Artificial Neuron Models

The main operational principle of artificial neural networks puts forth that an input set is received then transformed into an output set. To accomplish that the network needs to be trained to generate the proper outputs for the presented inputs. The samples which shall be presented to the network are at first transformed into a

vector. This vector is presented to the network and network generates the required output vector for this particular vector. In every single iteration weight connections of network are regulated to generate the proper output. Figure 3.2 demonstrates a graphical illustration of artificial neuron model.

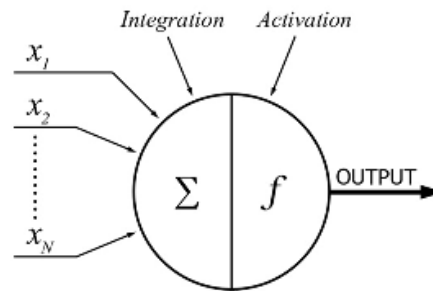


Figure 3.2 An artificial neuron

The general artificial neuron model consists of five basic elements :

Input: External information sent to an artificial neural cell from outside. Inputs are determined by the samples demanded to be learnt by network.

Connection weights: Show the strength of information received by the cell and its effect on the cell. The weights can take positive or negative values.

Net function: This function computes weighted inputs received by the cell. Each input is multiplied with its own weight . The net function is comprised of sum of weighted inputs.

$$\text{net}(w, x) = w_1x_1 + w_2x_2 + \dots + w_nx_n \quad 3.1$$

Activation function and Output: This function processes weighted inputs received by the neuron to determine the output that shall be generated by the neuron in response to this input. In the simplest case the output y is computed as:

$$y = f(\text{net}) = \begin{cases} 1 & \sum w_i x_i \geq \theta \\ 0 & \text{otherwise} \end{cases} \quad 3.2$$

where θ is a threshold level.

As in the case for summing function in the computation of output value too different activation functions can be employed. In multilayer perceptron network model which is widely used in modeling as activation function, sigmoid function is the most widely selected one. Typically the activation function is chosen by the designer and then the weight and threshold values will be adjusted some learning rule.

3.4 Single Layer Feedforward Networks

Single layer feedforward network is the simplest form of feedforward networks. The single layer feedforward network consisting of only one layer of nodes at which computation is performed, frequently with only one node in this layer. This model also called as perceptron model.

A simple perceptron is a computing unit with threshold θ which, when receiving the n neural inputs $x_1, x_2, \dots, x_n$ through edges with the associated weights $w_1, w_2, \dots, w_n$, outputs 1 if the inequality $\sum w_i x_i \geq \theta$ holds, and 0 otherwise. (Rojas, 1996)

The other form of the single layer network is Adaline. Adaline is abbreviated name of adaptive linear element. The main difference between Adaline and single

layer perceptron is seen their learning rules. Learning rule of Adaline performs classification by modifying weights in such a way as to minimize the mean square error at every iteration.

The important disadvantage of the single layer networks is they are used for only as a linear classifiers. Feedforward multilayer networks with nonlinear node functions can overcome these limitations.

3.5 Multi-Layer Feedforward Networks

Typically multilayer feed forward network consist of a set of inputs as an input layer, one or more hidden layers and one output layer. The input signals are distributed in a forward direction on a layer-by-layer. The hidden neurons enable the network to learn complex nonlinear tasks. The multilayer feed forward network model is a generalization of single layer perceptron model and it is used for many applications. A basic architecture of a multilayer networks with two hidden layers is shown in Figure 3.3.

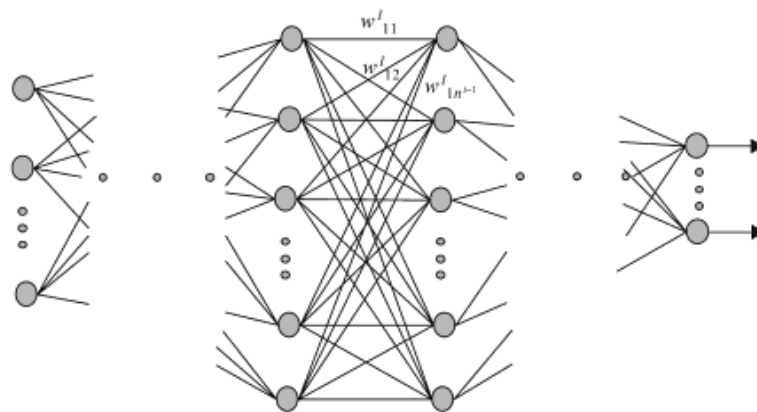


Figure 3.3 Multi layer network with two hidden layers and two outputs

The problem of learning is simply the problem of finding a set of connection strengths (weights) which allow the network to carry out desired computation. The network is provided with a set of example input/output pairs (a training set) and is to modify its connections in order to approximate the function from which the

input/output pairs have been drawn. The networks are then tested for ability to generalize (Fuller, 2000).

There are number of different learning rules. The learning rules can be categorized as supervised learning and unsupervised learning. In supervised learning or learning with a teacher, it is assumed that the learning process is supervised by a teacher who presents input-output patterns to network. The teacher also compares the attained output with desired ones. Then the network adjust its weights in such a way that the difference between the output at hand and desired outputs is reduced. In unsupervised or self-organized learning there is no teacher to supervise the learning process. Therefore, it is also called learning without a teacher. In this type of learning, no spesific examples are provided to the network. The desired output is not known. The results are unpredictable since during the learning process there is no desired output displayed.

The multi layer perceptron and many other neural networks learn by using an algorithm that is called backpropagation. The backpropogation algorithm is based on the error correction learning rule which is a generalization of least mean square algorithm that is used for single layer feedforward networks. The backpropogation algorithm consists of two phases, namely, a feed forward phase and a backpropogation phase. With backpropagation, the input data is repeatedly presented to the neural network. With each presentation the output of the neural network is compared to the desired output and an error is computed. This error is then feedback (backpropagated) to the neural network and used to adjust the weights such that the error decreases with each iteration and the neural model gets closer and closer to producing the desired output (Alexandrov, Albada, M.A.Sloot, 2006).

The backpropagation algorithm has had a major impact on the field of neural networks and has been widely applied to a large number of problems in a wide array of disciplines.

CHAPTER FOUR

APPLICATION

The purpose of this study is to examine the obesity facts among high school students using logistic regression and artificial neural networks. The MATLAB was used for both of the analysis. Data consist of all the 12th degree of high school students who study in Kipa, Nevvar Salih İsgören and Gaziemir state and anatolian high schools. This study was conducted in April during 2011-2012 academic year under the authorization of the Ministry of Education.

The total number of students was 749. The number decreases to 504 because of students not answering majority of the questions and the students who were not present at the day of the survey.

31% (157 students) of the survey data is made of Gaziemir High School, 19.4% (98 students) of the survey data is made of Nevvar Salih İsgören High School and 49.4%(249 students) of the survey data is made of Kipa Anatolian High School students. 81.5%(411 students) of the students are at the age of 18 ,16,1%(81 students) are at the age of 19, and 2,4% (12 students) are at the age of 20. The mean value of the age variable is 18.20. The survey data comprised 215 boys (%42,7) and 289 (%57,3) girls.

The presence of overweight and obesity was determined based on self reported height and weight to calculate body mass index(BMI), which is the weight in kilograms divided by the height in meters squared (kg/m^2). The international BMI cutpoints are used to define obesity and overweight. The students who have the BMI value is ($\text{BMI} > 25 \text{ kg/m}^2$) is accepted to be preobese(overweight) and the students who has the BMI value is higher than the obesity threshold level ($\text{BMI} > 30 \text{ kg/m}^2$) is accepted obese as a dependent variable.

The obesity survey consisted of 37 questions and it is made up of four main parts. Personal information(1 to 11), health information (12 to 16), dietary habits(17 to 29)

and social life (30 to 37). In the first part the questions were related to demographical and socio-economical facts, in the second part they were related to health information, in the third related to dietary habits and in the last section related to physical activity and sedentary behaviours.

The age variable is examined in two levels as 18 and more than 18. The variable about the number of people residing at home is examined in three levels. These are low for 2-3, medium for 4-5 and high for 6 and more. Since the frequency is so low for the last group, medium and high group are combined. For socio economic status variable, a score value which is calculated by using parental education, occupation and family income rates is obtained. Parental education is examined in four levels. These are primary, secondary, high school and university. Parental occupation is examined in four levels. These are professional or semi professional, skilled, semi skilled or retired, and not working. Finally family income rate is examined in five levels. 999 and less, 1000-1999, 2000-2999, 3000-3999 and 4000 and more. Accordingly the score value of socio economic status varies between 5 and 21. Thus the score value evaluated as low for 5-8, low-medium for 9-12, medium for 13-16 and high for 17-21. Since the frequency is so low for the high group, medium and high group are combined.

The question relating to whether parents are together or not is examined in two categories. The question related to last dental or physical check-up is examined in four categories. Since the frequency is so low for the second group, first two groups are combined. The question related to eating speed is examined in three levels. These are Slow, Normal and Fast. Since the frequency is so low for the slow group, slow and normal group are combined. The question related to number of major meals eaten is examined in three levels. The question related to smoking is examined in two levels. These are Never and Current /Former. The questions related to dietary habits are examined in three levels. These are High, Medium and Low. The question related to fast food consumption is examined in five levels. Since the frequency is so low for the last two group(5 or 6 - 7 and more), they are combined. Similarly the question related to breakfast consumption is examined in five levels and for the same reason

the second and third (5 or 6-3 or 4) groups are combined. The answers relating to food type and frequency of consuming these types of food is made up of seven options which range between high frequency to low frequency of consuming these food types. High group is made up of options such as more than once a day, once a day and 5-6 day a week, Medium group is made up of options such as 2-4 day a week and Low group is made up of options such as once a week, less than once a week and never. For fruit and vegetable consumption reference group is determined as high while the other food group consumption reference group is determined as low. For sedentary behaviours three habits variables are asked in three options each for weekdays and weekends. The question about physical activities has five options. These options are Low(Never and 1 or less day a week), Moderate(2-3 day a week) and High. (6-7 day a week). In the question asking the most common transportation means for travelling to and from the school the answers are active for the pedestrians and bicycle riders and passive for the bus and minibus riders. All of the independent variables that is consisted in survey is illustrated in Table 4-1.

Table 4.1 Categorical variable coding

Question	Name	Categories	Dummy Variable		
			D1	D2	D3
Date of Birth	AGE	18	0		
		More than 18	1		
Sex	SEX	Male	0		
		Female	1		
Number of people residing at home	NOPRAH	2-3(Low)	0		
		4-5(Medium)+ 6 and more (High)	1		
Socio Economic Status	SES	High+Medium	0	0	
		Low-Med	1	0	
		Low	0	1	
Whether parents are together or not?	PDD	Together	0		
		Not together	1		
Any chronical disease diagnosed by a doctor?	ACDDD	No	0		
		Yes	1		
Using medication due to a disease?	UMACD	No	0		
		Yes	1		
Last dental or physical check-up	LDPC	Do not remember +Years ago	0	0	
		One year ago	1	0	
		3 months or less ago	0	1	
Smoking status	SMOKE	Never	0		
		Former and Current	1		
Obesity among parents	OAP	None	0	0	
		Father or Mother	1	0	
		Both	0	1	
Eating a lot between meals	ELBM	No	0		
		Yes	1		
Dieting	DIET	No	0		
		Yes	1		

Table 4.1 continued

Number of major meals eaten	NMME	4+	0	0	
		3	1	0	
		1 and 2	0	1	
Eating speed	ES	Slow+ Normal	0		
		Fast	1		
Fast food consumption of past week	FFC	Never	0	0	0
		1-2	1	0	0
		3-4	0	1	0
		5-6 / 7and more	0	0	1
Breakfast consumption	BC	Everyday	0	0	0
		5-6 and 3-4	1	0	0
		1-2	0	1	0
		Never	0	0	1
Fruit intake(day/week)	FIN	High	0	0	
		Medium	1	0	
		Low	0	1	
Vegetable intake	VIN	High	0	0	
		Medium	1	0	
		Low	0	1	
Cake, cram cake intake(day/week)	CIN	Low	0	0	
		Medium	1	0	
		High	0	1	
Sweets and chocolates intake (day/week)	SCIN	Low	0	0	
		Medium	1	0	
		High	0	1	
Pastry cooks intake (day/week)	PCOOK	Low	0	0	
		Medium	1	0	
		High	0	1	
Potato chips intake (day/week)	PCIN	Low	0	0	
		Medium	1	0	
		High	0	1	
Coke and sugar sweetened beverages intake (day/week)	CSSBIN	Low	0	0	
		Medium	1	0	
		High	0	1	
Physical activity	PACT	High	0	0	
		Moderate	1	0	
		Low	0	1	
Commuting to and from school	CTFS	Active	0		
		Passive	1		
Studying (hour/weekday)	SSBWD	Low	0	0	
		Medium	1	0	
		High	0	1	
Studying (hour/weekend)	SSBWND	Low	0	0	
		Medium	1	0	
		High	0	1	
Using computer (hour/weekday)	UCWD	Low	0	0	
		Medium	1	0	
		High	0	1	
Using computer (hour/weekend)	UCWND	Low	0	0	
		Medium	1	0	
		High	0	1	
Watching TV (hour/weekday)	WTWD	Low	0	0	
		Medium	1	0	
		High	0	1	
Watching TV (hour/weekend)	WTWND	Low	0	0	
		Medium	1	0	
		High	0	1	

The crosstabs, independent variables versus the situation of being preobese and obese or not are illustrated on Table 4.2 to Table 4.32.

Table 4.2 Obesity status * SEX Crosstabulation

	SEX	
	Male	Female
Nonobese	168	260
Preobese and Obese	47	29

Table 4.3 Obesity status * AGE Crosstabulation

	AGE	
	18 and less	More than 18
Nonobese	351	77
Preobese and Obese	60	16

Table 4.4 Obesity status * NOPRAH Crosstabulation

	NOPRAH	
	Low	Medium
Nonobese	96	332
Preobese and Obese	15	61

Table 4.5 Obesity status * SES Crosstabulation

	SES		
	High+ Medium	Low Medium	Low
Nonobese	129	167	132
Preobese and Obese	22	32	22

Table 4.6 Obesity status * PDD Crosstabulation

	PDD	
	Together	Not together
Nonobese	392	36
Preobese and Obese	65	11

Table 4.7 Obesity status * ACDDBD Crosstabulation

	ACDDBD	
	No	Yes
Nonobese	355	73
Preobese and Obese	60	16

Table 4.8 Obesity status * UMDACD Crosstabulation

	UMDACD	
	No	Yes
Nonobese	392	36
Preobese and Obese	65	11

Table 4.9 Obesity status * LDPC Crosstabulation

	LPDC		
	Do not remember	One year ago	3 months or less ago
Nonobese	157	65	206
Preobese and Obese	34	10	32

Table 4.10 Obesity status * SMOKE Crosstabulation

	SMOKE	
	Never	Ever
Nonobese	339	89
Preobese and Obese	55	21

Table 4.11 Obesity status * OAP Crosstabulation

	OAP		
	None	Father or Mother	Both
Nonobese	360	57	11
Preobese and Obese	51	18	7

Table 4.12 Obesity status * ELBM Crosstabulation

	ELBM	
	No	Yes
Nonobese	165	263
Preobese and Obese	31	45

Table 4.13 Obesity status * DIET Crosstabulation

	DIET	
	No	Yes
Nonobese	400	28
Preobese and Obese	59	17

Table 4.14 Obesity status * NMME Crosstabulation

	NMME		
	>4 or 4	3	1 or 2
Nonobese	100	241	87
Preobese and Obese	12	47	17

Table 4.15 Obesity status * ES Crosstabulation

	ES	
	Slow and Normal	Fast
Nonobese	290	138
Preobese and Obese	50	26

Table 4.16 Obesity status * FFC Crosstabulation

	FFC			
	Never	1 or 2	3 or 4	5 or 6 / 7 or +
Nonobese	141	198	58	31
Preobese and Obese	30	26	12	8

Table 4.17 Obesity status * BC Crosstabulation

	BC			
	Everyday	5 or 6 / 3 or 4	1 or 2	Never
Nonobese	182	77	82	87
Preobese and Obese	27	16	17	16

Table 4.18 Obesity status * FIN Crosstabulation

	FIN		
	Low	Medium	High
Nonobese	196	124	108
Preobese and Obese	28	14	34

Table 4.19 Obesity status * VIN Crosstabulation

	VIN		
	Low	Medium	High
Nonobese	156	160	112
Preobese and Obese	21	27	28

Table 4.20 Obesity status * CIN Crosstabulation

	CIN		
	Low	Medium	High
Nonobese	229	139	60
Preobese and Obese	52	17	7

Table 4.21 Obesity status * SCIN Crosstabulation

	SCIN		
	Low	Medium	High
Nonobese	121	134	173
Preobese and Obese	31	22	23

Table 4.22 Obesity status * PCOOK Crosstabulation

	PCOOK		
	Low	Medium	High
Nonobese	227	110	91
Preobese and Obese	43	19	14

Table 4.23 Obesity status * PCIN Crosstabulation

	PCIN		
	Low	Medium	High
Nonobese	317	65	46
Preobese and Obese	56	12	8

Table 4.24 Obesity status * CSSBIN Crosstabulation

	CSSBIN		
	Low	Medium	High
Nonobese	205	107	116
Preobese and Obese	37	14	25

Table 4.25 Obesity status * PACT Crosstabulation

	PACT		
	High	Moderate	Low
Nonobese	91	84	253
Preobese and Obese	13	14	49

Table 4.26 Obesity status * CTFS Crosstabulation

	CTFS	
	Active	Passive
Nonobese	167	261
Preobese and Obese	31	45

Table 4.27 Obesity status * SSBWD Crosstabulation

	SSBWD		
	Low	Medium	High
Nonobese	102	159	167
Preobese and Obese	25	23	28

Table 4.28 Obesity status * SSBWND Crosstabulation

	SSBWND		
	Low	Medium	High
Nonobese	116	160	152
Preobese and Obese	28	22	26

Table 4.29 Obesity status * UCSBWD Crosstabulation

	UCSBWD		
	Low	Medium	High
Nonobese	277	94	57
Preobese and Obese	47	18	11

Table 4.30 Obesity status * UCSBWND Crosstabulation

	UCSBWND		
	Low	Medium	High
Nonobese	232	139	57
Preobese and Obese	37	26	13

Table 4.31 Obesity status * WTSBWD Crosstabulation

	WTSBWD		
	Low	Medium	High
Nonobese	217	164	47
Preobese and Obese	38	23	15

Table 4.32 Obesity status * WTSBWND Crosstabulation

	WTSBWND		
	Low	Medium	High
Nonobese	178	188	62
Preobese and Obese	27	35	14

4.1 Univariate Analysis

The candidate variables which include multivariate model are determined by using univariate analysis. The result of univariate analysis is shown in Table 4.33 that consists of the following information: (1) The estimated slope coefficient(s) for the univariate logistic regression model containing only this variable, (2) the estimated standard error of the estimated slope coefficient(s), (3) degrees of freedom for each variable, (4) the estimated odds ratio, which is obtained by exponentiating the estimated coefficient, (5) the 95% CI for the odds ratio, (6) the loglikelihood value (7) the likelihood ratio test statistic, (8) the p value of the loglikelihood for the univariate model.

Table 4.33 Univariate Logistic Regression Models for Obesity Status

Variable	B	SE	df	OR	95%CI	Log-likelihood	G ²	p
SEX	-0.9196	0.2560	1	0.3987	0.2414-0.6585	-207.0728	13.330	0.000
AGE	0.1952	0.3082	1	1.2156	0.6644-2.2241	-213.5425	0.3910	0.5318
NOPRAH	0.162	0.3106	1	1.1759	0.640-2.162	-213.5987	0.2785	0.5977
SES			2			-213.6076	0.2608	0.8777
SES(1)	0.1165	0.3007	1	1.1236	0.6232-2.0258			
SES(2)	-0.023	0.3259	1	0.9773	0.5159-1.8512			
PDD	0.6113	0.3696	1	1.8427	0.8930-3.8027	-212.4839	2.5081	0.1133
ACDDBD	0.2599	0.3093	1	1.2968	0.7072-2.3778	-213.3968	0.6824	0.4088
UMDADC	0.6113	0.3696	1	1.8427	0.8930-3.8027	-212.4839	2.5081	0.1133
LDPC			2			-212.8623	1.7513	0.4166
LDPC(1)	-0.3419	0.3888	1	0.7104	0.3316-1.5222			
LDPC(2)	-0.3323	0.2681	1	0.7173	0.4241-1.2132			
SMOKE	0.3746	0.2828	1	1.4543	0.8355-2.5317	-212.8941	1.6877	0.1939
OAP			2			-207.4807	12.515	0.0019
OAP(1)	0.8016	0.3090	1	2.2291	1.216-4.085			
OAP(2)	1.5023	0.5061	1	4.4920	1.666-12.1130			
ELBM	-0.0935	0.2537	1	0.9107	0.5539-1.4973	-213.6703	0.1354	0.7129
DIET	1.4149	0.3376	1	4.116	2.124-7.978	-205.9070	15.662	0.000
NMME			2			-212.5936	2.2887	0.3184
NMME(1)	0.4856	0.3446	1	1.6252	0.8271-3.1933			
NMME(2)	0.4876	0.4045	1	1.6284	0.7369-3.5983			
ES	0.0887	0.2630	1	1.0928	0.6526- 1.8297	-213.6815	0.1130	0.7367
FFC			3			-211.6946	4.0868	0.2522
FFC(1)	-0.4826	0.2897	1	0.6172	0.3498- 1.0890			
FFC(2)	-0.0280	0.3755	1	0.9724	0.4658- 2.0299			
FFC(3)	0.1930	0.4446	1	1.2129	0.5074- 2.8993			
BC			3			-213.0113	1.4534	0.6931
BC(1)	0.3370	0.3435	1	1.4007	0.7144- 2.7464			
BC(2)	0.3347	0.3370	1	1.3975	0.7219- 2.7051			
BC(3)	0.2149	0.3414	1	1.2397	0.6349- 2.4204			
FIN			2			-207.8568	11.762	0.0028
FIN(1)	-0.2353	0.3469	1	0.7903	0.4005-1.5597			
FIN(2)	0.7901	0.2819	1	2.2037	1.2681-3.8295			

Table 4.33 continued

VIN			2			-211.7242	4.0275	0.1335
VIN(1)	0.2260	0.3120	1	1.2536	0.6802-2.3104			
VIN(2)	0.6190	0.3141	1	1.8571	1.0034-3.4374			
CIN			2			-210.7440	5.9879	0.0501
CIN(1)	-0.6188	0.2994	1	0.5386	0.2995-0.9685			
CIN(2)	-0.6660	0.4279	1	0.5138	0.2221-1.1886			
SCIN			2			-211.2243	5.0273	0.0810
SCIN(1)	-0.4450	0.3057	1	0.6408	0.3520-1.1667			
SCIN(2)	-0.6560	0.2996	1	0.5189	0.2884-0.9336			
PCOOK			2			-213.5281	0.4197	0.8107
PCOOK(1)	-0.0923	0.2990	1	0.9118	0.5075-1.6384			
PCOOK(2)	-0.2081	0.3318	1	0.8122	0.4239-1.5562			
PCIN			2			-213.7283	0.0194	0.9903
PCIN(1)	0.0441	0.3460	1	1.0451	0.5304-2.0591			
PCIN(2)	-0.0156	0.4096	1	0.9845	0.4411-2.1970			
CSSBIN			2			-212.7399	1.9962	0.3686
CSSBIN(1)	-0.3217	0.3357	1	0.7249	0.3755-1.3997			
CSSBIN(2)	0.1774	0.2838	1	1.1941	0.6847-2.0825			
PACT			2			-213.2777	0.9205	0.6311
PACT(1)	0.1542	0.4138	1	1.1667	0.5184-2.6254			
PACT(2)	0.3043	0.3351	1	1.3557	0.7030-2.6145			
CTFS			1	0.9288	0.5650-1.5268	-213.6957	0.0845	0.7712
SSBWD			2			-212.2779	2.9201	0.2322
SSBWD(1)	-0.5273	0.3155	1	0.5902	0.3180-1.0954			
SSBWD(2)	-0.3797	0.3025	1	0.6841	0.3781-1.2376			
SSBWND			2			-212.0505	3.3749	0.1850
SSBWND1	-0.5627	0.3099	1	0.5696	0.3103-1.0457			
SSBWND2	-0.3444	0.2990	1	0.7086	0.3944-1.2732			
UCSBWD			2			-213.6223	0.2313	0.8908
UCSBWD1	0.1209	0.3018	1	1.1286	0.6246-2.0390			
UCSBWD2	0.1287	0.3652	1	1.1374	0.5560-2.3266			
UCSBWND			2			-213.2053	1.0653	0.5871
UCSBWND1	0.1594	0.2775	1	1.1729	0.6809-2.0204			
UCSBWND2	0.3577	0.3547	1	1.4301	0.7136-2.8660			
WTSBWD			2			-211.3837	4.7086	0.0950
WTSBWD(1)	-0.2221	0.2837	1	0.8009	0.4592-1.3966			
WTSBWD(2)	0.6002	0.3448	1	1.8225	0.9272-3.5822			
WTSBWND			2			-213.0896	1.2967	0.5229
WTSBWND1	0.2049	0.2767	1	1.2273	0.7136-2.1109			
WTSBWND2	0.3979	0.3608	1	1.4886	0.7339-3.0196			

Under the null hypothesis that the slope coefficients are zero the quantity of G^2 follows the chi square distribution and corresponding degrees of freedom for each variable are shown in Table 4.33.

The variables which have p values are smaller than 0,25 are found to be significant. These are: SEX, PDD, UMDACD, SMOKE, OAP, DIET, FIN, VIN, CIN, SCIN, SSBWD, SSBWND, WTSBWD. The other evidence of the significance that the confidence intervals of the odds ratio values for most variables either do not contain 1 or just barely do.

The multivariate logistic regression analysis will be done by using the variables found to be significant in the univariate case. The result of the fitted model is given in Table 4.34.

Table 4.34 Multivariate logistic regression model containing significant variables determined by the univariate analysis

Variable	$\hat{\beta}$	SE	Wald	OR	95%CI		G^2	p value
					Lower	Upper		
Constant	-1.659	0.491	3.379	0.190			68.259	0.000
SEX	-1.162	0.320	-3.630	0.313	0.167	0.586		
PDD	0.732	0.428	1.709	2.080	0.898	4.816		
UMDACD	0.748	0.451	1.659	2.113	0.873	5.114		
SMOKE	-0.139	0.340	-0.409	0.870	0.447	1.694		
OAP(1)	0.817	0.355	2.303	2.265	1.129	4.541		
OAP(2)	1.436	0.579	2.481	4.202	1.352	13.063		
DIET	1.376	0.392	3.508	3.959	1.835	8.542		
FIN(1)	-0.592	0.398	-1.487	0.553	0.254	1.207		
FIN(2)	0.536	0.338	1.583	1.709	0.880	3.316		
VIN(1)	0.084	0.362	0.231	1.087	0.535	2.210		
VIN(2)	0.307	0.394	0.779	1.359	0.628	2.939		
CIN(1)	-0.474	0.348	-1.365	0.622	0.315	1.230		
CIN(2)	-0.526	0.511	-1.029	0.591	0.217	1.610		
SCIN(1)	-0.149	0.347	-0.430	0.861	0.436	1.700		
SCIN(2)	-0.307	0.373	-0.823	0.736	0.354	1.529		
SSBWD(1)	0.238	0.432	0.551	1.269	0.544	2.960		
SSBWD(2)	0.269	0.518	0.520	1.309	0.475	3.612		
SSBWND(1)	-0.274	0.419	-0.655	0.760	0.335	1.727		
SSBWND(2)	-0.108	0.489	-0.220	0.898	0.345	2.341		
WTSBWD(1)	-0.103	0.318	-0.325	0.902	0.483	1.683		
WTSBWD(2)	0.804	0.405	1.982	2.234	1.009	4.944		
-2LL= 359.216781								

The multivariate model results containing variables identified in the univariate analysis are illustrated in Table 4.34. Under the null hypothesis that the p slope coefficients for the independent variables are equal to zero, the G^2 follows chi-square distribution with 21 degrees of freedom. The p value for the test is significant at the 0,05 level. Thus we may conclude that at least one, and perhaps all p coefficients are different from zero.

The Wald statistic is used to determine which of the variables are significant in the multivariate model. The critical value for wald statistic is 2 which would lead to an approximate level of significance of 0,05.

The variables which have the Wald statistics are more than 2 are found to be significant. These variables are SEX, OAP and DIET. Also there are some possibly significant variables. Because their wald statistic values are nearly 2 and their confidence interval does not contain 1 or just barely do. These variables are PDD, UMDACD, FIN and WTSBWD. Thus a new model which does not exist SMOKE, VIN, CIN, SCIN, SSBWD, SSBWND are fitted. The results are given in Table 4.35.

Table 4.35 Multivariate logistic regression model containing the variables SEX, OAP, DIET, PDD, UMDACD, FIN and WTSBWD

Variable	$\hat{\beta}$	SE	Wald	OR	95%CI		G^2	p value
					Lower	Upper		
Constant	-1.903	0.305	-6.232	0.149			61.572	0.000
SEX	-1.139	0.288	-3.955	0.320	0.182	0.563		
PDD	0.661	0.415	1.591	1.937	0.858	4.372		
UMDACD	0.729	0.429	1.701	2.074	0.895	4.805		
OAP(1)	0.803	0.347	2.316	2.231	1.131	4.401		
OAP(2)	1.396	0.572	2.442	4.038	1.317	12.379		
DIET	1.424	0.386	3.690	4.154	1.950	8.851		
FIN(1)	-0.510	0.377	-1.352	0.601	0.287	1.258		
FIN(2)	0.670	0.304	2.206	1.955	1.078	3.546		
WTSBWD(1)	-0.109	0.305	-0.355	0.897	0.493	1.633		
WTSBWD(2)	0.698	0.382	1.828	2.010	0.951	4.248		
-2LL= 365.903784								

The likelihood ratio test statistic for the difference between the full and reduced model is $G^2 = (365,903784 - 359,216781) = 6.687$. Comparing this test statistic to a chi square distribution with 11 degrees of freedom which is $(v_{full} - v_{reduced} = 21 - 10 = 11)$, yields a p value of 0,8238. since the p value is exceeding the 0,05 it is concluded that the reduced model is as good as the full model. Also this is supported by the fact that the values of the estimated coefficients for the other variables in Table 4.34 are nearly same.

Since the both coefficients for wald statistics of the WTSBWD and FIN variables are not significant we can not be sure about the contributions of these variables to the model. For this reason we first fit a new model which does not contain WTSBWD. The results are given in Table 4.36.

Table 4.36 Multivariate logistic regression model containing the variables SEX, OAP, DIET, PDD, UMDACD and FIN

Variable	$\hat{\beta}$	SE	Wald	OR	95%CI		G^2	p value
					Lower	Upper		
Constant	-1.816	.269	-6.763				57.453	0.000
SEX	-1.153	.286	-4.034	.316	.180	.553		
PDD	.643	.417	1.542	1.902	.840	4.306		
UMDACD	.649	.418	1.554	1.915	.844	4.344		
OAP(1)	.795	.343	2.321	2.215	1.132	4.336		
OAP(2)	1.381	.566	2.437	3.977	1.310	12.072		
DIET	1.453	.381	3.811	4.275	2.025	9.025		
FIN(1)	-.506	.375	-1.350	.603	.289	1.257		
FIN(2)	.657	.302	2.176	1.929	1.067	3.485		
-2LL= 370.023200								

The likelihood ratio test statistic for the difference between two model is $G^2 = (370.023200 - 365.903784) = 4.1194$. Comparing this test statistic to a chi square distribution with 2 degrees of freedom which is $(v_{full} - v_{reduced} = 10 - 8 = 2)$, yields a p value of 0,1275. since the p value is exceeding the 0,05 it is concluded that the reduced model is as good as the full model. Thus there is no advantage to including WTSBWD in the model.

Now another model which does not contain FIN variable is fitted. The results are given in Table 4.37.

Table 4.37 Multivariate logistic regression model containing the variables SEX, OAP, DIET, PDD, and UMDACD

Variable	$\hat{\beta}$	SE	Wald	OR	95%CI		G^2	p value
					Lower	Upper		
Constant	-1.697	.197	-8.617	0.183			46.130	0.000
SEX	-1.138	0.278	-4.085	0.321	0.186	0.553		
PDD	0.584	0.415	1.407	1.793	0.795	4.047		
UMDACD	0.652	0.408	1.600	1.920	0.863	4.272		
OAP(1)	0.760	0.338	2.249	2.138	1.103	4.144		
OAP(2)	1.311	0.544	2.413	3.712	1.279	10.771		
DIET	1.454	0.376	3.864	4.281	2.047	8.950		
-2LL= 381.345855								

The likelihood ratio test statistic for the difference between two model is $G^2 = (381.345855 - 370.023200) = 11.3227$. Comparing this test statistic to a chi square distribution with 2 degrees of freedom which is $(v_{full} - v_{reduced} = 8 - 6 = 2)$,

yields a p value of 0,0035. since the p value is not exceeding the 0,05 it is concluded that the FIN variable is significant for the model. Now we need to check the contribution of the other possibly significant variables namely PDD and UMDACD. First the new model is fitted without UMDACD. The results are given in Table 4.38.

Table 4.38 Multivariate logistic regression model containing the variables SEX, OAP, DIET, FIN and PDD

Variable	$\hat{\beta}$	SE	Wald	OR	95%CI		G^2	P value
					Lower	Upper		
Constant	-0.343	0.433	-0.793	0.709			55.202	0.000
SEX	-1.090	0.281	-3.885	0.336	0.194	0.583		
PDD	0.679	0.418	1.624	1.971	0.869	4.471		
OAP(1)	0.795	0.339	2.342	2.214	1.138	4.304		
OAP(2)	1.445	0.564	2.561	4.241	1.403	12.817		
DIET	-1.454	0.378	-3.841	0.234	0.111	0.491		
FIN(1)	-0.477	0.373	-1.277	0.621	0.299	1.290		
FIN(2)	0.679	0.301	2.256	1.972	1.093	3.558		
-2LL= 372.273706								

The likelihood ratio test statistic for the difference between two model is $G^2 = (372.273706 - 370.023200) = 2.25051$. Comparing this test statistic to a chi square distribution with 1 degrees of freedom which is $(v_{full} - v_{reduced} = 8 - 7 = 1)$ yields a p value of 0.1336. Since the p value is exceeding the 0.05 it is concluded that the UMDACD variable is not significant for the model. Finally we fit a new model which does not exist PDD. The results are given in Table 4.39.

Table 4.39 Multivariate logistic regression model containing the variables SEX, OAP, DIET and FIN

Variable	$\hat{\beta}$	SE	Wald	OR	95%CI		G^2	P value
					Lower	Upper		
Constant	-1.747	0.264	-6.604	0.174			52.748	0.000
SEX	-1.074	0.280	-3.840	0.342	0.197	0.591		
OAP(1)	0.745	0.336	2.218	2.106	1.091	4.068		
OAP(2)	1.410	0.561	2.512	4.095	1.363	12.298		
DIET	1.507	0.374	4.031	4.514	2.169	9.394		
FIN(1)	-0.429	0.369	-1.164	0.651	0.316	1.341		
FIN(2)	0.698	0.300	2.324	2.009	1.115	3.618		
-2LL= 374.727271								

The likelihood ratio test statistic for the difference between two model is $G^2 = (374.727271 - 372.273706) = 2.4536$. Comparing this test statistic to a chi square distribution with 1 degrees of freedom which is $(v_{full} - v_{reduced} = 7 - 6 = 1)$ yields a p

value of 0.1173. Since the p value is exceeding the 0.05 it is concluded that the PDD variable is not significant for the model.

The final model is accepted in Table 4.39. The estimated logit function is given by the following equation:

$$\hat{g}(X) = \begin{matrix} \beta_0 & \beta_{11} & \beta_{12} & \beta_{21} & \beta_{22} & \beta_{31} & \beta_{32} & \beta_{41} & \beta_{42} \\ 0 & 11 & 11 & 21 & 21 & 22 & 22 & 31 & 31 & 41 & 41 & 42 & 42 \end{matrix}$$

$$\hat{g}(X) = -1.747 - 1.074D_{11} + 0.745D_{21} + 1.410D_{22} + 1.507D_{31} - 0.429D_{41} + 0.698D_{42}$$

95% confidence interval for the odds ratio of SEX variable does not include an odds ratio of 1 and the wald statistic of this variable is greater than value of 2. Thus we would conclude that the SEX variable is significantly related with the obesity risk. Based on the coefficients in Table 4.39 the odds of being obese for females are 34.2% (odds ratio=0.342) of the odds for males for being obese or another explanation the odds of being obese are 65% (0.342-1=0.658) lower for females than for males.

95% confidence interval for the odds ratio of OAP variable does not include 1. Also the wald statistics of OAP variable is greater than value of 2. For this reason we conclude that the OAP variable is significant. For OAP variable the odds ratio is 2.106 for the first group which is the group one of the parents of whom is obese. Thus we would say that having an obese parent was 2.106 times more risk factor than neither of the parents are obese. The odds ratio for the second group which is the group both parents of whom are obese is 4.095. Thus we would say that having both parents obese was 4.095 times more risk factor than having neither parents obese.

The odds ratio is 4.514 for the DIET variable. Since the 95% confidence interval does not include 1 and also the wald statistic of this variable greater than 2, the relationship is found to be significant. Thus we would say that the ones going on a diet run the 4.514 times more risk of becoming obese than the ones not going on a diet.

Although the confidence interval of one of the groups for FIN variable contains value of 1 and also the wald statistic value is smaller than 2, the likelihood ratio where this variable is important is determined via testing. The odds ratio for the first group is .651. Accordingly, the odds of being obese are 35% ($0.651-1=0.349$) lower for the medium level of fruit consuming group than the frequent fruit consumers. And the odds ratio for the second group is 2.009 which means that low level of fruit consuming group 2.009 more risk factor than frequent fruit consumers.

The Hosmer Lemeshow test statistic is found by using SPSS (Version 20) program. For $\alpha = 0,05$ and $g-2=7$, the Hosmer-Lemeshow goodness-of-fit test statistic is found 4.834 and the corresponding p value is 0.680. Since 0.680 is larger than 0.05, we conclude that the final model fits the data well.

4.2 Artificial Neural Network

The neural network that is used in this study is a feed-forward back-propagation neural network with three layers: an input layer, three hidden layers and an output layer. The input layer consisted of 52 input neurons, the hidden layer consisted of fifty two, fifteen and one hidden neurons respectively. Sigmoidal function is used as the activation function.

In predicting outcome variable in logistic regression the whole data set has been used. For neural networks 2/3 and 1/3 data set are divided as training and testing respectively. The learning rate and momentum constant for network training were chosen respectively to 0.25 and 0.95.

The true classification rate for logistic regression and artificial neural network have been found 86.1% and 86.3% respectively.

CHAPTER FIVE

CONCLUSION

In this study the performances of logistic regression and artificial neural network techniques are evaluated according to predictive ability of independent variables of dependent variable which is being obese or not obese.

Recently especially in studies conducted on health researches while predicting response variable both techniques are used. Both these methods have advantages and disadvantages. One such advantage is that artificial neural network model find patterns despite missing data. It does not mean the performance will drop when artificial neural networks are operating with missing data. The importance of missing data is responsible for the drop in the performance.

One of the most important advantages of ANN is that it can work with large data sets with innumerable variables. Another advantage of the neural network is that there is not normality assumption that need to be verified before the models can be constructed.

On the other hand the architecture which will enable us to reach optimal results are found via trial and error method as there is no accepted theory showing the number of hidden layers and neurons in the hidden layers. The most common type of artificial neural networks is the feed-forward back propagation multiperceptron model which is also used in this study.

Another limitation of neural network models is that coefficients and odds ratios corresponding to each variable cannot be easily calculated and presented as they are in logistic regression.

REFERENCES

- Agresti, A. (2007). *An introduction to categorical data analysis* (Second edition). USA: John Wiley&Sons.
- Al-Isa, A. N. (1999). Obesity among Kuwait university students: An explorative study. *The journal of the royal society for the promotion of health*. 119 (4). 223-227.
- Alexandrov, V.N., Ablada, G.D., & Sloot, P.M. (2006). *Computational Science*. Germany: Springer.
- Christensen, R. (1997). *Log-Linear models and logistic regression* (Second edition). NY: Springer.
- Fuller, R. (2000). *Introduction to neuro-fuzzy systems* . Germany: Springer.
- Graupe, D. (2007). *Principles of artificial neural networks* (Second edition). USA: World Scientific.
- Hosmer, D.,& Lemeshow, S. (1989). *Applied logistic regression*. NY: John Wiley&Sons
- Jaimes, F., Farbiarz J., Alvarez, D., Martinez, C. (2005). Comparison between logistic regression and neural networks to predict death in patients with suspected sepsis in the emergency room. *Critical Care*. (9). 150-156.
- Janssen, I., Katzmarzyk, P. T., Boyce, W. F., Vereecken, C., Mulvihill, C., Roberts, C. et al. (2005). Comparison of overweight and obesity prevalence in school aged youth from 34 countries and their relationships with physical activity and dietary patterns. *Obesity Reviews* (6). 123-132.

- Kalaycıoğlu, S., Çelik, K., Çelen, Ü., Türkyılmaz, S. (2010). Temsili Bir Örneklemde Sosyo-Ekonomik Statü (SES) Ölçüm Aracı Geliştirilmesi: Ankara Kent Merkezi Örneği. *Sosyoloji Araştırmaları Dergisi*, 13 (1), 183-220.
- Mehrotra, K., Mohan, C., & Ranka, S. (2000). *Elements of artificial neural networks* (Second edition). USA: A Bradford Book.
- Menard, S. (2010). *Logistic Regression: From introductory to advanced concepts and applications*. USA: Sage.
- Mota, J., Ribeiro, R., Santos, M. P., & Gomez, H. (2006). Obesity, physical activity, computer use and TV viewing in portuguese adolescents. *Pediatric Exercise Science* (17), 113-121.
- Neter, J., Kutner, M., Nachsteim, C.J., and Wasserman, W. (1996). *Applied linear statistical models* (Fourth edition). USA: Irwin.
- Neumark-Sztainer, D., Paxton, S.J., Hannan, P.J., Stat, M., Haines, J., Story, M., et al. (2006). Does body satisfaction matter? Five-year longitudinal associations between body satisfaction and health behaviors in adolescent females and males. *Journal of adolescent health* (39), 244-251.
- Öztemel, E. (2006). Yapay sinir ağları (2nci basım). İstanbul: Papatya Yayıncılık.
- Rojas, R. (1996). *Neural Networks: A systematic introduction*. Germany: Springer.
- Tatlıdil, H. (1996). *Uygulamalı çok değişkenli istatistiksel analiz*. Ankara: Akademi.
- WHO Technical Report Series (2003). Diet, nutrition and the prevention of chronic diseases.

APPENDIX A

Obezite Anket Formu

Bu soru formu obezite oluşumuna neden olan etkenleri belirlemek amacıyla düzenlenmiştir ve elde edilen veriler sadece bu maksatla kullanılacaktır. İlgi ve katkılarınız için teşekkür ederiz.

Dokuz Eylül Üniversitesi FBE İstatistik Bölümü Öğrencisi

Kişisel Bilgiler

Doğum Yılıınız		Cinsiyetiniz	☐ Erkek ☐ Kadın
Boyunuz (cm)		Kilonuz (kg)	
Babanızın Mesleği		Annenizin Mesleği	
Evinizde Yaşayan Kişi Sayısı	 (siz dahil)		
Babanızın Eğitim Durumu	☐ İlkokul ☐ Ortaokul ☐ Lise ☐ Üniversite	Annenizin Eğitim Durumu	☐ İlkokul ☐ Ortaokul ☐ Lise ☐ Üniversite
Anne ve Babanızla İlgili Olarak	☐ İkisi birlikte yaşıyor ☐ İkisi ayrı yaşıyor ☐ Annem hayatta değil ☐ Babam hayatta değil	Ailenizin Toplam Geliri	☐ 999 TL ve altı ☐ 1000 TL - 1999 TL ☐ 2000 TL - 2999 TL ☐ 3000 TL - 3999 TL ☐ 4000 TL ve üstü

Sağlık Bilgileri

Doktor tarafından tanısı konulmuş bir rahatsızlığınız var mı?	☐ Var ☐ Yok	En son ne zaman bir sağlık kontrolünden (diş yada genel) geçtiniz	☐ Hatırlamıyorum ☐ Yıllar önce ☐ Bir yıl önce ☐ Üç ay ve daha kısa süre önce
Kronik bir rahatsızlığa bağlı olarak düzenli kullandığınız ilaç var mı?	☐ Var ☐ Yok		
Sigara kullanma alışkanlığınızla ilişkili olarak en uygun olanı işaretleyiniz.	☐ Evet, kullanıyorum ☐ Kullanıyordum, bıraktım ☐ Hayır, hiç kullanmadım	Ebeveynleriniz arasında fazla kilolu yada obez olan var mı?	☐ İkiside değil ☐ Babam ☐ Annem ☐ Her ikisi de

Sosyal Yaşam

Haftada kaç gün bir saat ya da daha fazla süren bir fiziksel aktivitede bulunuyorsunuz?

- ☐ Hiç
☐ 1 ve daha az
☐ 2-3 gün
☐ 4-5 gün
☐ 6-7 gün

Ev ve okul arasında en sık kullandığınız ulaşım şekli hangisidir?

- ☐ Yaya olarak
☐ Bisikletle
☐ Servis ya da başka bir taşıtla

Hafta içi günlerde aşağıdaki aktivitelere ne kadar zaman ayırırsınız?

1 saat veya daha az 2-3 Saat 4 saat veya daha fazla

Ders çalışmak	☐	☐	☐
Bilgisayar kullanmak	☐	☐	☐
Televizyon izlemek	☐	☐	☐

Hafta sonu günlerde aşağıdaki aktivitelere ne kadar zaman ayırırsınız?

1 saat veya daha az 2-3 Saat 4 saat veya daha fazla

Ders çalışmak	☐	☐	☐
Bilgisayar kullanmak	☐	☐	☐
Televizyon izlemek	☐	☐	☐

İlginize çok teşekkür ederiz...

APPENDIX B

Unianalysis.m

```

clear all,clc
load 'C:\MatLab\y.txt';
n=1;
[r,c] = size(y);

% The loglikelihood for intercept only model is calculated,

ones_=sum(y);
zero_=r-ones_;
x=zeros(size(y,1),0);
b=glmfit(x,[y ones(size(y))],'binomial','logit');
L0=(ones_*log(ones_)+zero_*log(zero_)-(r)*log(r));

% %%%
% Univariate analysis
% %%%

files=dir('C:\MatLab\independent\*.txt');
for i=1:numel(files);
    func1=get(i);
    [b,S,loglikelihood,W,OR,U,G,p,L,v]=coef(func1,L0);
    Uni(i,1)=p; %Univariate
    Uni(i,2)=loglikelihood; %%%results
end

indis=1;
for s=1:i
    if Uni(s,1)<=0.25 %univariate results
        Uni_result(indis,1)=s; %except insignificant variables
        Uni_result(indis,2)=Uni(s,1);
        Uni_result(indis,3)=Uni(s,2);
        indis=indis+1;
    end
end

% The data set for multivariate model are obtained,

indis=1;
for j = 1:size(Uni_result,1),
    i=Uni_result(j,1);
    [r,v]=size(get(i));
    size_v(j,1)=i;
    size_v(j,2)=v;
end
DATA=zeros(size(y,1),0);
DATA=data(Uni_result,y);
index=1;
[b,S,loglikelihood,W,OR,U,G,p,L,v]=coef(DATA,L0);
result(index,1)=loglikelihood;
result(index,2)=v;
result(index,3)=G;
result(index,4)=p;
indis=1;
for j = 1:size(size_v,1)
    for i=indis:indis+size_v(j,2)-1
        Multi(indis,1)=size_v(j,1)
        Multi(indis,2)=W(indis+1,1)
        indis=indis+1
    end
end
end
% →

```

```

% ←
% Variables with Wald statistic values larger than 1.5 will be selected

Multi=abs(Multi);
indis=1;
j=1;
while j<=size(Multi,1)
    Multi(j,2)>1.5
    Multinew(indis,1)=Multi(j,1)
    Multinew(indis,2)=Multi(j,2)
    if j<size(Multi,1)
        while Multi(j,1)==Multi(j+1,1)
            indis=indis+1;
            j=j+1;
            Multinew(indis,1)=Multi(j,1)
            Multinew(indis,2)=Multi(j,2)
        end
    end
    store=j;
    if j>1
        while Multi(j,1)==Multi(j-1,1)& Multi(j-1,2)<=1.5
            indis=indis+1;
            j=j-1;
            Multinew(indis,1)=Multi(j,1)
            Multinew(indis,2)=Multi(j,2)
        end
    end
    j=store
    indis=indis+1;
    end
    j=j+1;
end

% Likelihood ratio procedure for selected variables.

DATA=zeros(size(y,1),0);
DATA=data(Multinew,y);
[b,S,loglikelihood,W,OR,U,G,p,L,v]=coef(DATA,L0);
index=index+1;
result(index,1)=loglikelihood;
result(index,2)=v;
result(index,3)=G;
result(index,4)=p;

[p_val,G]=pvalue(result(index,1),result(index-1,1),result(index,2),result(index-1,2));
result(index,5)=p_val;
indis=1;
eliminate(indis,1)=0;
Multinew=sortrows(abs(Multinew),2)
for j = 1:size(Multinew,1)
    if Multinew(j,2)<1,96 & Multinew(j,1)~=eliminate(:,1)
        eliminate(indis,1)=Multinew(j,1);
        indis=indis+1;
    end
end
j=1;
key=0;
while j<=size(eliminate,1)
    DATA=zeros(size(y,1),0);
    pointer=j;
    for i=1:pointer
        eliminated(i,1)=eliminate(i,1)
    end
    DATA=data1(Multinew,eliminated,y);
    [b,S,loglikelihood,W,OR,U,G,p,L,v]=coef(DATA,L0);
    index=index+1;
    result(index,1)=loglikelihood;
    result(index,2)=v;
    result(index,3)=G;
    result(index,4)=p;
end
% →

```

```
% ←
    if key==0
        result(index,6)=result(index-1,2)-v;
        [p_val,G]=pvalue(result(index,1),result(index-1,1),
            result(index,2),result(index-1,2));
        result(index,5)=p_val;
    else
        result(index,6)=result(index-2,2)-v;
        [p_val,G]=pvalue(result(index,1),
            result(index-2,1),result(index,2),result(index-2,2));
        result(index,5)=p_val;
    end

    if result(index,5)<0.05
        eliminate(j.:[]);
        key=1;
    else
        j=j+1;
    end
    key=0;
end
end
```

get.m

```
%it is used for combining path names and locations

function variables=get(i);
yol='C:\MatLab\independent\independent';
dosyadi=strcat(yol,int2str(i));
dosyadi=strcat(dosyadi,'.txt');
disp(dosyadi);
a=load(dosyadi);
variables=a;
end
```

coef.m

```
%computes the coefficients ,fitted values, likelihood,G and p values

function [b,S,loglikelihood,W,OR,U,G,p,L,v]=coef(a,L0);
n=1;
[t,v]=size(a);
load 'C:\MatLab\y,t.txt';
b=glmfit(a,[y ones(size(y))],'binomial','logit');
fitted=glmval(b,a,'logit');
loglikelihood=sum(log(binopdf(y,n,fitted)));
G=2*(loglikelihood-L0);
p=1-chi2cdf(G,v);
x=[ones(size(y)) a];
V=eye(t);
for I = 1:t
    V(I,I)=fitted(I)*(1-fitted(I));
end
I=transpose(x)*V*x;
V=sqrt(inv(I));
S=diag(V);
W=b./S;
OR=exp(b);
Up=b+S,*1.96;
Low=b-S,*1.96;
U=exp(Up);
L=exp(Low);
end
```

data.m

```
%determine the variables which has more than one column

function DATA=data(Uni,y);
DATA=zeros(size(y,1),0);
a=get(Uni(1,1))
DATA=[DATA a];
for j = 2:size(Uni,1),
    if Uni(j,1)~=Uni(j-1,1)
        a=get(Uni(j,1))
        DATA=[DATA a];
    end
end
end
```

data1.m

```
%determine the variables whether they are in eliminated matrix or not.

function DATA=data1(Uni,eliminate,y);
DATA=zeros(size(y,1),0);
if Uni(1,1)~=eliminate(:,1)
    a=get(Uni(1,1));
    DATA=[DATA a];
end
for j = 2:size(Uni,1)
    if Uni(j,1)~=Uni(j-1,1)&Uni(j,1)~=eliminate(:,1)
        a=get(Uni(j,1));
        DATA=[DATA a];
    end
end
end
```

pvalue.m

```
%determine the significant variables

function [p_val,G]=pvalue(loglikelihood1,loglikelihood2,v1,v2);
G=2*(loglikelihood2-loglikelihood1);
v=v2-v1;
p_val=1-chi2cdf(G,v);
end
```

APPENDIX C

artneuralnet.m

```
%artificial neural net with backpropagation algorithm
clc;
load independents.txt
load output.txt

p=independents;
p=p';
T=output;
T=T';

net=newff(minmax(p),[52 15 1],{'logsig' 'logsig' 'purelin'},'trainscg');
net.initFcn='initlay';
net.performFcn='sse';           % SSE performance function
net.trainParam.goal=0.01;       % SSE goal
net.trainParam.show=20;         % frequency of progress displays (in epochs)
net.trainParam.epochs=1000;     % max number epochs to train
net.trainParam.mc=0.95;         % 0.95 is a momentum constant
net.trainParam.lr=0.25;

net = init(net);
[net,tr] = train(net,p,T);

load testFile.txt;
y=testFile;
y=y';
x=sim(net,y);
round(x')
```