DOKUZ EYLUL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

APPLICATIONS OF LOGISTIC REGRESSION WITH MISSING DATA

by
Nıver SİLAHLI

January, 2013
İZMİR

APPLICATIONS OF LOGISTIC REGRESSION WITH MISSING DATA

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

by
Nıver SİLAHLI

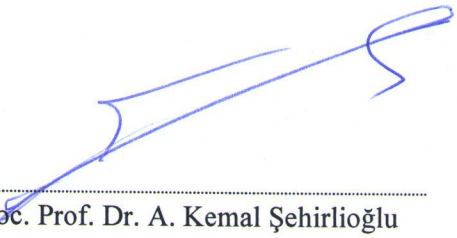
January, 2013
İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"APPLICATIONS OF LOGISTIC REGRESSION WITH MISSING DATA"** completed by **NIVER SİLAHLI** under supervision of **PROF. DR. C. CENGİZ ÇELİKOĞLU** 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.

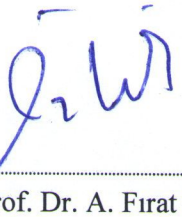
Prof. Dr. C. Cengiz ÇELİKOĞLU

(Supervisor)



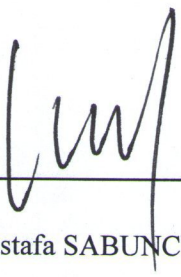
Assoc. Prof. Dr. A. Kemal Şehirlioğlu

(Jury Member)



Assist. Prof. Dr. A. Fırat ÖZDEMİR

(Jury Member)



Prof.Dr. Mustafa SABUNCU

Director

Graduate School of Natural and Applied Sciences

ACKNOWLEDMENTS

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

I am grateful to Asst. Prof. Dr. Neslihan DEMİREL and Research Ass. Özgül VUPA ÇİLENGİROĞLU for their continual encouragement and support all throughout the work.

I am also grateful to Dr. Nilgün ÖZÇAKAR and Dr. Gizem LİMNİLİ for their continual encouragement and support all throughout the work.

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

Nıver SİLAHLI

APPLICATIONS OF LOGISTIC REGRESSION WITH MISSING DATA

ABSTRACT

Missing data is a common problem in statistical studies. While ignoring missing data is an option, it is possible to contribute to study by analyzing them with various statistical methods. Missing data analysis includes methods aiming at missing data problem solving. These methods are classified as deletion (Listwise and Pairwise) and imputation (Regression imputation, Expectation Maximization and Multiple Imputation).

Logistic regression analysis method, one of the most popular methods applied for modeling two dependent variables, has two possible categories of dependent variable 0 and 1. Logistic Regression Analysis can be expanded according to the dependent variable as nominal and ordinal. There is no limitation for independent variables.

The aim of this study is to examine the methods of missing value analysis and logistic regression and to evaluate the performance of different missing value analysis methods on logistic regression.

Keywords: Missing data analysis, regression imputation, expectation maximization, multiple imputation, logistic regression analysis

KAYIP VERİ OLMASI DURUMUNDA LOJİSTİK REGRESYON UYGULAMALARI

ÖZ

Kayıp veri, istatistiksel çalışmalarda sıkça karşılaşılan problemlerden biridir. Kayıp verileri göz ardı etmek bir seçenek iken, bunları çeşitli istatistiksel yöntemlerle çözümleyerek çalışmaya katmakta mümkündür. Kayıp veri analizi araştırmacıların çok sık karşılaştıkları kayıp veri sorununa çözüm getirmeyi amaçlayan yöntemler içerir. Bu yöntemler, silme (Liste/Durum Düzeyli ve Çiftler Düzeyinde) ve atama (Regresyon Ataması, Hot Deck Ataması, Beklenti Maksimizasyonu ve Çoklu Atama) olarak sınıflandırılmıştır.

İkili bağımlı değişkeni modellemek için uygulanabilen en popüler regresyon yöntemlerinden biri olan Lojistik Regresyon Analizi'nde, bağımlı değişken 0 ve 1 gibi iki olası kategoriye sahiptir. Lojistik Regresyon Analizi, bağımlı değişkenin sınıflayıcı ve sıralı olmasına göre genişletilebilir. Burada bağımsız değişkenler için kısıtlama getirilmemiştir.

Bu çalışmanın amacı, kayıp veri analizi ve lojistik regresyon yöntemlerini inceleyerek, farklı kayıp veri analizi yöntemlerinin lojistik regresyondaki performanslarının değerlendirilmesidir.

Anahtar sözcükler: Kayıp veri analizi, regresyon ataması, beklenti maksimizasyonu, çoklu atama, lojistik regresyon analizi

CONTENTS

Page

CHAPTER ONE
INTRODUCTION

In many statistical researches the data may have missing. There are several reasons for data is missing. When applying a survey, people may not answer some questions as household income. They can refuse from telling their weight or age. In longitudinal researches, each participant may drop out the study or lose his life for some reason.

Missing values problems are a widespread problem in many domains of research, especially medical researchs, for data analysis. These missing values create a problem for analysts using some statistical approaches for data analyses. Because some statistical approaches and multivariate methods depend upon full data, a few methods have been affirmed for treating the issue of these missing values. Missing data reduce the representativeness of the sample and can therefore distort inferences about the population. For this reason, the missing data problem must be solved.

To handling for missing data problem, there are several techniques. Such as, listwise/pairwise deletion, the hot deck imputation, the regression imputation, the expectation maximization (EM) and multiple imputation (MI).

To achieve unbiased estimators from the data, the missing data situation must be solved. First, the missing data pattern must be built; second, the missingness mechanism must be figured; and third, the most suitable imputation method for imputing the missing values must be implemented.

The most popular method of the modeling the binary dependent variable is logistic regression. Because of its assumptions are not strict, use of logistic regression is easy. When the dependent variable has more than two categories, logistic regression methods can be extending. When the dependent variable is ordinal, then we can use ordinal logistic regression.

The study is a cross sectional study. In this study, we tried to imputate the missing data with the most commonly used and is considered to be the most effective methods as complete case, regression imputation, EM algorithm and multiple imputation. The data sets were modeling with the logistic regression analysis. This model was compared.

This study includes five chapters. First chapter summarizes the whole study. In chapter two, the missing value analysis and its methods are studied. In chapter three, logistic regression, ordinal regression and their characteristics are examined. In chapter four, chapter two and three are supported with the application. The missing data pattern was built, the missingness mechanism was figured. The most appropriate imputation methods are implemented. The data sets were modeling with the logistic regression analysis. This model was compared to find the best imputation methods.

CHAPTER TWO
MISSING VALUE ANALYSIS

Many times in statistical research it is hardly possible to aggregate data that is complete. Missing data can be present for a many reasons. When applying a questionnaire, for instance, people may not answer some questions. For instance, participants may not find certain questions applicable. In other situations, the questions may be acceptable but the given responses are not. Additionally, some participants may simply can reject to answer certain types of questions, and this is to say nothing of the rigorous technical dangers associated with impair databases, integrating data from different sources, and faulty input.

Missing values problems are a widespread problem in many domains of research, especially medical researches, for data analysis. These missing values create a problem for analysts using some statistical approaches for data analyses. Because some statistical approaches and multivariate methods depend upon full data, a few methods have been affirmed for treating the issue of these missing values.

To achieve unbiased estimators from the data, the missing data situation must be solved. First, the missing data pattern must be built; second, the missingness mechanism must be figured; and third, the most suitable imputation method for imputing the missing values must be implemented (Patzer, 2009).

2.1 Missing Data

In statistics, missing data consist when no value is stored for the variable in the existent observation. Data are missing for many reasons. These are ordered as:

Participants in longitudinal studies often leave before the study is finished because they have moved away of the area, pass away, no longer see personal benefit to participating, or do not like the effects of the treatment.

Questionnaires suffer missing data when participants reject, or do not know the answer to or by mistake skip an item. Some questionnaire analysts even design the study so that some questions are asked of only a subset of participants.

Empirical studies have missing values when an analyst is commonly unable to collect an observation. Bad weather conditions may render off observation impossible in area experiments. An analyst becomes sick or equipment fails. Data may be missing in any type of study due to by mistake or data faulty input. An analyst drops a tray of test tubes. A data file becomes spoilt. Most analysts can come across with one (or more) of these situations.

2.2 Missing Data Patterns

Rubin (2002) and his colleagues find it appropriate to determine the missing data pattern, that defines which values are observed in the data matrix and which values are missing, and the missing data mechanisms, that interests the relationship between missingness and the values of variables in the data matrix.

Let $X = (x_{ij})$ express a $(n \times K)$ rectangular data set without missing values, with i th row $x_i = (x_{i1}, ..., x_{iK})$ where x_{ij} is the value of variable X_j for subjects i. With missing data, assign the missing data indicator matrix $M = (m_{ij})$, such as $m_{ij} = 1$ if x_{ij} is missing and $m_{ij} = 0$ if x_{ij} is present. The matrix M then assigns the pattern of missing data.

In general, missing data pattern can be univariate**,** which means that missing data values only consist of a single dependent variable, or multivariate in the sense that missing values consist of more than one variable. Univariate and multivariate data patterns are displayed in Figure 2.1(a) and Figure 2.1(b), respectively.

X_1	X_2	X_3	X_4	X_5
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	1
0	0	0	0	1
0	0	0	0	1
0	0	0	0	1
0	0	0	0	1

Figure 2.1 (a) Univariate pattern

For univariate non response, which is diplayed in Figure 2.1 (a), the missing values occur on an item X_5 but the other items X_1, X_2, X_3 and X_4 are completely observed.

X_1	X_2	X_3	X_4	X_5
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	1	1	1
0	0	1	1	1
0	0	1	1	1
0	0	1	1	1
0	0	1	1	1

Figure 2.1 (b) Multivariate pattern

For multivariate patterns, which is displayed in Figure 2.1 (b), the missing values occur on an items X_3, X_4 and X_5 but the other items X_1 and X_2 are completely observed.

X_1	X_2	X_3	X_4	X_5
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	1
0	0	0	0	1
0	0	0	1	1
0	0	0	1	1
0	0	1	1	1
0	0	1	1	1
0	1	1	1	1
0	1	1	1	1

Figure 2.1 (c) Monotone pattern

A particular missing data pattern is a monotone pattern, which may occur by dropouts in longitudinal studies. In Figure 2.1 (c), items or item groups $X_1,...,X_p$ may be ordered in such a way that if X_j is missing for a unit, then $X_{j+1},...,X_p$ are missing as well; which shows a monotone pattern.

X_1	X_2	X_3	X_4	X_5
0	0	0	1	0
0	0	0	1	0
0	0	1	0	0
0	0	1	0	1
0	0	0	0	0
0	1	0	0	0
0	1	0	1	0
0	1	0	0	0
0	0	0	0	1
0	0	0	0	0
0	0	0	0	0

Figure 2.1 (d) General multivariate pattern

The general multivariate pattern is illustrated in Figure 2.1 (d) above. It can be seen that any set of variables may be missing in general multivariate pattern.

The selection of imputation method may be based on the underlying missing data pattern such that the exploration of the missing pattern is important and helpful (Durrant, 2005).

2.3 Missing Data Mechanisms

The missing data mechanisms are critical because of the features of missing data methods based very strongly on the nature of the dependencies in these mechanisms (Rubin, 2002). The types of missing data vary on three categories, which are based on the relationship between the missing data mechanism and the missing and the observed values. These categories are important to know because the problems arisen by missing data and the solutions to these problems are different for the three categories. These mechanisms are the Missing Completely at Random (MCAR), the Missing at Random (MAR) and the Missing not at Random (MNAR), respectively.

2.3.1 Missing Completely at Random (MCAR)

We denote the complete data as $X = (x_{ij})$ and the missing data indicator matrix as $M = (m_{ij})$. The missing data mechanism is qualified by the conditional distribution of M given X, $f(M / X, \phi)$, where ϕ denotes unknown parameters. If missingness does not relate on the values of the data X, missing or observed, that is, if

$$f(M / X, \phi) = f(M / \phi) \text{ for all } X, \phi \qquad (2.1)$$

the data are entitled missing completely at random (MCAR) (Rubin, 2002). Response variables are said to be MCAR when the distribution of missingness does not relate on the complete data $X = (x_{ij})$. Stated in other words the missing value has no dependence on any other variable.

2.3.2 Missing at Random (MAR)

If the missingness based only on the components X_{obs} of X that are observed, and not on the components that are missing. That is,

$$f(M / X, \phi) = f(M / X_{obs}, \phi) \text{ for all } X_{mis}, \phi \quad (2.2)$$

The missing data mechanism is then named the missing at random (MAR) (Rubin, 2002). Put differently, a participant's missingness may relate on his or her own values for the observed variables, but not the missing variables.

2.3.3 Missing Not at Random (MNAR)

The mechanism is named the missing not at random (MNAR) if the distribution of M depends on the missing values in the data matrix X .

$$f(M / X, \phi) = f(M / X_{mis}, \phi) \text{ for all } X_{obs}, \phi \quad (2.3)$$

The missing value depends on other missing values and thus missing data imputation cannot be performed from the existing data.

2.4 Interpretation of the Randomness in the Missing Data

It is important to define that the missing data belongs to that mechanism. Missing data mechanisms are used to determine the method to be used for missing data analysis. Independences of missing data are tested to identify which missing data fixes to which mechanism. There are various statistical methods for this.

1) Observations of a variable in data set should be divided into two groups as the ones having missing data and the ones not having, and it also should be analyzed whether any meaningful difference is available between these two groups according to values of the other variables related. This research can be done by t-test that tests

the significance of difference between two group means. Significant difference shows the existence of non-random missing data process. Hypothesis is set as:

$$H_0 : \mu_1 = \mu_2 \qquad p > \alpha ,$$
$$H_1 : \mu_1 \neq \mu_2$$

that means the null hypothesis cannot reject. There is no significant difference between the group means, which shows us the existence of random missing data process.

2) The coefficient of correlation, also known as the correlation coefficient, is the strength of a relationship, measured linearly, between two variables. This measure can range from -1 to 1.

If the coefficient of correlation is equal to;

(-1) We have a perfectly negative correlation. As one asset moves in a direction, the other asset will move in a perfectly different direction.

(0) We have no correlation, positive or negative.

(1) We have a perfectly positive correlation as one asset moves in a direction, the other asset will move perfectly in the same direction.

$$\rho_{xy} = \frac{Cov(X,Y)}{\sigma_x \sigma_y} \tag{2.4}$$

Variables in data set are divided into two groups as the ones having missing data and the ones not having and full data is coded as 1; missing data is coded as 0 and correlation coefficient between these variables are calculated. The correlation coefficient is indicated strength of the relationship between missing data for each variable couple. The small correlation coefficient presents randomness.

3) Little (1988) proposed a multivariate extension of the t-test approach that simultaneously evaluates mean differences on every variable in the data set. Unlike

univariate t-tests, Little's procedure is a global test of MCAR that applies to the entire data set.

Like the t-test approach, Little's test evaluates means differences across subgroups of cases that share the same missing data pattern. The test statistic is a weighted sum of the standardized differences between the subgroup mean and the grand means, as follows:

$$\chi^2 = \sum_{j=1}^{J} n_j (\hat{\mu}_j - \hat{\mu}_j^{(ML)})^T \hat{\Sigma}_j^{-1} (\hat{\mu}_j - \hat{\mu}_j^{(ML)}) \qquad (2.5)$$

where j subscript indicates that the number of elements in the parameter matrices and n_j is the number of cases in missing data pattern j, $\hat{\mu}_j$ contains the variable means for the cases in missing data pattern j, $\hat{\mu}_j^{(ML)}$ contains maximum likelihood estimates of the grand means, and $\hat{\Sigma}_j$ is the maximum likelihood estimate of the covariance matrix. χ^2 is approximately distributed as a chi-square statistic with $\sum k_j - k$ degrees of freedom, where k_j is the number of complete variables for pattern j, and k is the total number of variables (Enders, 2010).

Roderick J. A. Little's chi-square statistic for testing whether values are the missing completely at random (MCAR), the null hypothesis is that the data are missing completely at random, and the p-value is significant at the 0.05 level. If the p-value is less than 0.05, the data are not missing completely at random.

2.5 Techniques of the Imputation Missing Data

The issue of missing data occurs many times in practice in applied research settings. Imputation is a way to deal with missing data. Imputation techniques are quite used in studies that contain missing data, but the parameter estimates can be biased and variance estimates can be inappropriate. Shall the imputation technique

use does not exactly represent the variability in the data, the resulting confidence intervals will be incorrect (Rockhill, n.d.).

Imputation of missing data on a variable fills that missing by a value that is drawn from an estimate of the distribution of this variable. In single imputation, only one estimate is used. In multiple imputation, diverse estimates are used, reflecting the confusion in the estimation of this distribution.

Missing data are challenge because most statistical methods entail a value for each variable. Missing data can be inspired by missing areas in a database or incorrectly entered information. To base on the nature of the data and amount of samples available, different imputation methods are available.

The most common decision is to use the listwise deletion (also called the complete case analysis), analyzing only the cases with complete data. Individuals with data missing on any variables are removed from the analysis. The advantages of this method are easy to use, very simple, and the default in most statistical programs. But it has limitations. It can considerably lower the sample size, resulting in a severe lack of power. This is especially true if there are many variables implied in the analysis, each with data missing for a few cases. It can also lead to biased results, depending on why the data are missing.

Disadvantages stem from the potential lack of information in discarding incomplete cases. This lack of information has two outlooks lack of sensitivity, and bias when the missing data mechanism is not the MCAR.

The pairwise deletion (also known as available-case analysis) attempts to reduce the loss of data by eliminating cases on an analysis-by-analysis basis. The entire values are used by the method; its downside is that the sample base changes from variable to variable according to the pattern of missing data (Rubin, 2002).

The listwise and the pairwise deletion are by far the most common missing data handling approaches in many fields of the statistical analysis. The primary advantage of these methods is that they are convenient to implement and are standard options in statistical software packages. However, deletion methods have serious limitations that preclude their use in most situations. Most importantly, these approaches assume MCAR data and can produce distorted parameter estimates when this assumption does not obtain (Enders, 2010).

Mean imputation takes the seemingly appealing tack of filling in the missing values with the arithmetic mean of the available cases. Like other imputation techniques, mean imputation is convenient because it generates a complete data set. However, convenience is not a compelling advantage because this approach severely distorts the resulting parameter estimates, even when the data are MCAR.

The hot deck imputation is a process in that missing items are replaced with values from respondents. A model encourage such procedures is the model in that respond probabilities are assumed equal within imputation cells. An influential version of the hot deck imputation is characterized for the cell response model and a computationally influential variance estimator is given. An approximation to the entirely influential process in that a small number of values are imputed for each missing is described (Wayne & Jae, 2005).

It has some advantages: it defends the distribution of item values, it charters the use of the same sample weight for all items and results acquired from different analyses are logical with one another (Schoier, n.d.).

2.5.1 Regression Imputation

The regression imputation imputes missing values by estimated values from a regression of the missing item on items observed for the unit, usually calculated from units with both observed and missing variables present (Rubin, 2002).

The first step of the imputation process is to estimate a set of regression equations that predict variables that have missing values from the complete variables. A complete-case analysis usually produces these estimates. The second step is to produce predicted values for the missing variables. These predicted scores fill in the missing values and produce a complete data set.

Consider univariate nonresponse, with $X_1,...,X_{k-1}$ fully observed and X_k observed for the first r observations and missing for the last $n-r$ observations. Regression imputations compute the regression of X_k on $X_1,...,X_{k-1}$ based on the r complete cases, and then fills in the missing values as predictions. The missing value is imputed using the regression equation:

$$\hat{y}_{ik} = \tilde{\beta}_{k0\cdot 12...k-1} + \sum_{j=1}^{k-1} \tilde{\beta}_{kj\cdot 12...k-1} x_{ij} \qquad (2.6)$$

where $\tilde{\beta}_{k0\cdot 12...k-1}$ is the intercept and $\tilde{\beta}_{kj\cdot 12...k-1}$ is the coefficient of X_j in the regression of X_k on $X_1,...,X_{k-1}$ based on the r complete cases.

Regression imputation is largely the same with multivariate data sets but is somewhat more complicated to implement. To illustrate, think of a assumptive data set with three variables, X_1, X_2 and X_3, all of which have missing data. Not including the complete cases, there are six possible missing data patterns. The presence of multiple missing data patterns complicates the imputation process somewhat because each missing data pattern requires a unique regression equation. To illustrate, Table 2.1 shows the regression equations. $\hat{y}_1$ is the estimation of X_1, $\hat{y}_2$ is the estimation of X_2 and $\hat{y}_3$ is the estimation of X_3.

Table 2.1 Equations used by regression imputations

Missing Variables	Regression Equations
X_1	$\hat{y}_1 = \beta_0 + \beta_1 x_2 + \beta_2 x_3$
X_2	$\hat{y}_2 = \beta_0 + \beta_1 x_1 + \beta_2 x_3$
X_3	$\hat{y}_3 = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
X_1 and X_2	$\hat{y}_1 = \beta_0 + \beta_1 x_3$ and $\hat{y}_2 = \beta_0 + \beta_1 x_3$
X_1 and X_3	$\hat{y}_1 = \beta_0 + \beta_1 x_2$ and $\hat{y}_3 = \beta_0 + \beta_1 x_2$
X_2 and X_3	$\hat{y}_2 = \beta_0 + \beta_1 x_1$ and $\hat{y}_3 = \beta_0 + \beta_1 x_1$

Substituting the observed scores into the relevant regression equations produces predicted values for the incomplete variables, and these predicted scores impute in the missing values and produce a complete data set.

Under an MCAR mechanism, we can yield consistent estimates of the covariance matrix, meaning that the estimates get closer to their true population values as the sample size increases.

2.5.2 EM Algorithm

EM algorithm is widely used in last years. Algorithm is a re-iteratively method that includes the maximum likelihood estimations to calculate the parameter predictions in in-complete data problems. EM algorithm helps to find the maximum likelihood estimation that is not impossible but seems difficult. EM algorithm uses the way of possibility prediction set according to observed and missing data. It is possible to estimate the parameters of probability distribution or the probability of observed data as a function of parameters with EM algorithm by imputing the missing data by maximizing via repetition method.

The purpose of this algorithm is to achieve the maximum likelihood estimation in cases of in-complete data problems appeared in studies. EM algorithm is applied in points shall the maximum likelihood estimation of probability distribution in given sample (in case the function is complicated) cannot be directly calculated.

2.5.2.1 Formulation of EM Algorithm

We have model for the complete data X , with associated density $f(X/\theta)$ indexed by unknown parameter θ. We write $X=(X_{obs},X_{mis})$ where X_{obs} represents the observed part of X and X_{mis} denotes the missing part. If the missing data mechanism is the MAR (Missing at Random) and the objective is to maximize the ignorable likelihood

$$L(\theta/X_{obs})=\int f(X_{obs},X_{mis}/\theta)dX_{mis} \tag{2.7}$$

with respect to θ. When the likelihood is differentiable and unimodal, when likelihood equation solve, ML estimates can be found

$$D_{\ell}(\theta/X_{obs})\equiv\frac{\partial \ln L(\theta/X_{obs})}{\partial\theta}=0 \tag{2.8}$$

An alternative computing strategy for incomplete-data problems, which does not require second derives to be calculated or approximated, is the Expectation Maximization (EM) algorithm, a method that relates ML estimation of θ from $\ell(\theta/X_{obs})$ to ML estimation based on complete-data log likelihood $\ell(\theta/X)$.

There are two disadvantages, in some cases, with large fractions of missing information, it can be very slow to convergence; and in some problems, the M step is difficult and then the theoretical simplicity of EM does not convert to practical simplicity (Rubin, 2002).

2.5.2.2 The E Step and the M Step of EM Algorithm

The M step is particularly simple to describe: perform ML estimation of θ just as if there were no missing data, which is, as if they had been imputed in. Thus the M step of EM uses the same computational method as ML estimation from $\ell(\theta/X)$.

The observed data and current estimated parameters are given by the E step finds the conditional expectation of the missing data, and then substitute these expectations for the missing data.

Accurately, let $\theta^{(t)}$ be the present estimate of the parameter θ. The E step of EM achieve the expected complete-data log-likelihood if θ were $\theta^{(t)}$:

$$Q(\theta / \theta^{(t)}) = \int \ell(\theta / x) f(X_{mis} / X_{obs}, \theta = \theta^{(t)}) dX_{mis} \tag{2.9}$$

The M step of EM describes $\theta^{(t+1)}$ by maximizing this expected complete data log-likelihood:

$$Q(\theta^{(t+1)} / \theta^{(t)}) \geq Q(\theta / \theta^{(t)}), \quad \text{for all } \theta \tag{2.10}$$

In the convergence of the algorithm, refer that $\theta^{(t+1)}$ is the estimate for θ which maximizes the difference $\Delta(\theta/\theta^{(t)})$. Starting with the present estimate for θ, that is, $\theta^{(t)}$ we then had that $\Delta(\theta^{(t)}/\theta^{(t)}) = 0$. Because $\theta^{(t+1)}$ is chosen to maximize $\Delta(\theta/\theta^{(t)})$, we then have that, $\Delta(\theta^{(t+1)}/\theta^{(t)}) \geq \Delta(\theta^{(t)}/\theta^{(t)}) = 0$, so for each repetition the likelihood $L(\theta)$ is non-decreasing.

When the algorithm achieves a stable point for some $\theta^{(t)}$ the value $\theta^{(t)}$ maximizes $l(\theta / \theta^{(t)})$. Because L and l are equal at $\theta^{(t)}$ if L and l are differentiable at $\theta^{(t)}$, then $\theta^{(t)}$ must be a fixed point of L. The fixed point need not, however, be a local maximum (Borman, 2004).

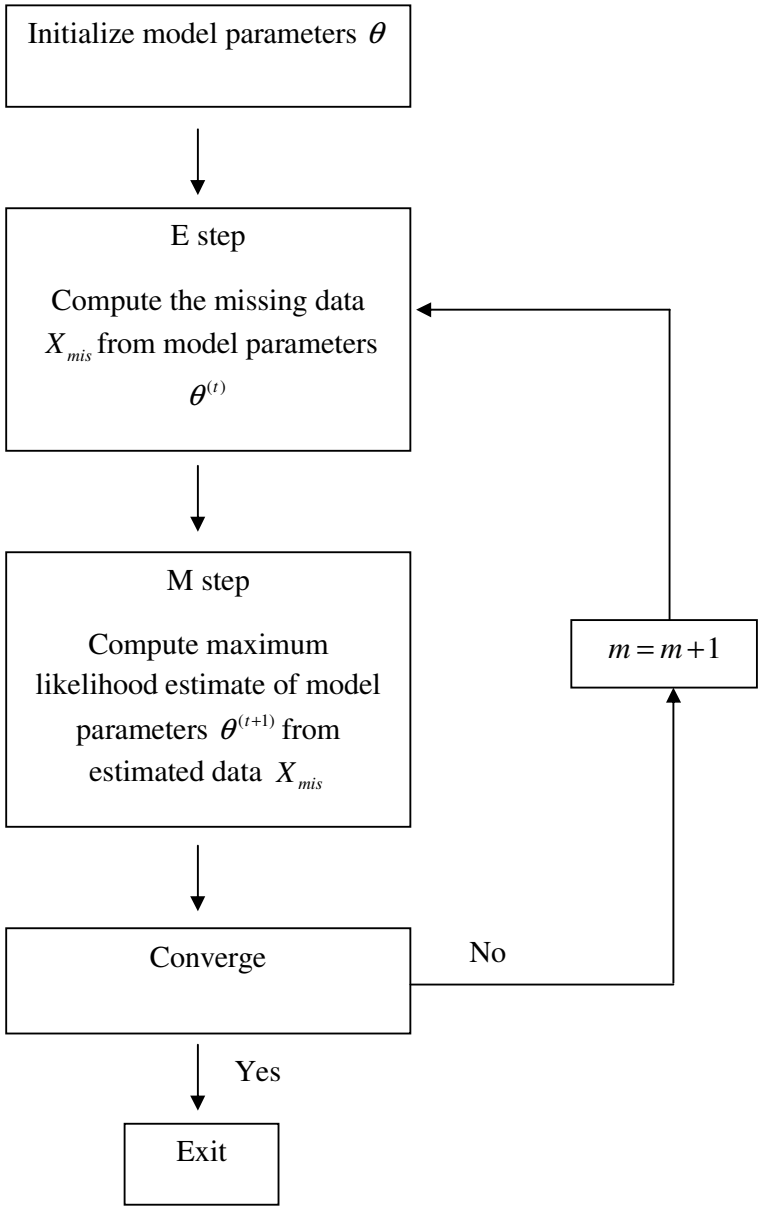


Figure 2.2 Flow chart of the EM algorithm

2.5.3 Multiple Imputation

Multiple imputation methodology was first proposed to handling with missing data by Rubin (1987). In the multiple imputation (MI) each missing value is imputed by a list of $M > 1$ values. Replacing the jth element of each list for the corresponding missing value generates M plausible alternative versions of complete data (Schafer and Graham, 2002). All of the data sets is analyzed in the same action by a complete

case method. The results are then combined using techniques offered by Rubin (1987) to give parameter estimates and standard errors that take into account the confusion due to missing values. Schafer (1999) specified that unless there are extraordinary high rates of missing data, the optimum is to use five to ten imputations. In many practical applications, the additional time and effort required to handle $M = 20$ versions than $M = 10$ has often little consequence (Schafer and Graham, 2002).

Multiple imputation for missing in public consumption files imputes each missing value by two or more valid values. The values can be selected to represent both uncertainty about which values to impute assuming the reasons for missing are familiar and ambiguity about the reasons for missing (Rubin, 1987).

Schafer and Olsen (1998) note that the multiple imputation (MI) do not have to be the MCAR mechanism but instead need only meet the less strict assumption that the missing data are missing at random (MAR).

Multiple imputation (MI) has several desirable features:

- Present appropriate random error into the imputation procedures makes it possible to get approximately unbiased estimates of all parameters. No deterministic imputation method can do this in general settings.
- Iterated imputation assigns one to get perfect estimates of the standard errors. The single imputation methods do not assign for the additional error introduced by imputation.
- Multiple imputation (MI) can be used with any kind of data and any kind of analysis without specialized software.

Of course, certain needs must be met for the multiple imputation (MI) to have these adorable features. First, the data must be missing at random (MAR), sense that the probability of missing data on a particular variable Y can depend on other observed variables, but not on Y itself (controlling for the other observed variables). Second, the model used to produce the imputed values must be “correct” in partially.

Third, the model used for the analysis must coupling, in some sense, with the model used in the imputation. All these conditions have been carefully determined by Rubin (1987, 1996).

The problem is that it's easy to disrupt these conditions in practice. There are often strong causes to distrust that the data are not MAR. Unfortunately, not much can be done about this. While it's possible to formulate and estimate models for data that are not MAR, such models are complex, untestable, and require specialized software. Hence, any general-purpose method will necessarily invoke the MAR assumption. Even when the MAR condition is satisfied, producing random imputations that yield unbiased estimates of the desired parameters is not always easy or straightforward (Allison, n.d.).

CHAPTER THREE
LOGISTIC REGRESSION

The logistic regression model is one of the most widely used models in statistics. Logistic regression is a mathematical modeling approach that can be used to model relationship between one or more independent variables (X) and a dichotomous or multicategory dependent variable (Y). To determine the relationship between the discrete dependent variable and the independent variables which can be discrete or continuous logistic regression models are used (Çolak, E. 2002). To describe logistic regression, the conditional mean of Y given X has a different interpretation, it must be reconsidered shall the independent variable is a categorical variable Researchers try to modeling probabilities in logistic regression. Logistic regression is a procedure for modeling categorical dependent variable that does not depend on the assumption that the independent variables are normally distributed (Dielman, 2001).

Logistic regression (LogR) is popular to defeat many of the limiting assumptions of ordinary least square (OLS) regression. These assumptions are ordered as follows:

- The LogR does not assume a linear relationship between the dependent and the independent variable(s).
- The dependent variable need not be normally distributed.
- The dependent variable need not be homoscedastic for each category of the independents. It means that there is no homogeneity of variance assumption.
- Normally distributed error terms are not assumed.
- The LogR does not desire that the independents be interval.

3.1 Binary Logistic Regression

In a mainly of regression applications, the dependent variable of interest has only two possible qualitative outcomes, and therefore can be corresponded by a binary indicator variable taking on values 0 and 1(Neter, 1996).

3.1.1 Fitting of Simple Logistic Regression

Let **X,** y be a data set with binary outcomes. For each experiment x_i in **X** the outcome is either $y_i = 1$ or $y_i = 0$. When the dependent variable has indicator, $y_i = 1$ that are said to belong to the positive class, or $y_i = 0$ belong to the negative class (Komarek, n.d.).

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad Y_i = 0,1 \qquad \text{i=1,...,n} \tag{3.1}$$

The expected response $E\{Y_i\}$ has a special meaning. Since $E\{\varepsilon_i\} = 0$,

$$E\{Y_i\} = b_0 + b_1 X_i \tag{3.2}$$

Y_i has Bernoulli distribution which probability distribution as follows,

$$Y_i = 1 \qquad \mathrm{P(Y_i=1)}=\pi_i \tag{3.3}$$

$$Y_i = 0 \qquad \mathrm{P(Y_i=0)}=1-\pi_i \tag{3.4}$$

Thus, π_i is the probability that $Y_i = 1$, and $1-\pi_i$ is the probability that $Y_i = 0$. From expected value of Bernoulli distribution,

$$E\{Y_i\} = \beta_0 + \beta_1 X_i = \pi_i \tag{3.5}$$

There are special problems when dependent variable is binary. One of them is ε_i can also take on only two values: $1-\beta_0-\beta_1 X_i$ if $Y_i = 1$ and $-\beta_0-\beta_1 X_i$ if $Y_i = 0$. Therefore, ε_i cannot be even approximately normally distributed (Ryan, 1997). This problem refers to nonnormal error terms by Neter, Kutner, Nachtsheim, Wasserman (1996).

Second of these problems is nonconstant error variance that error terms do not have equal variances as with linear regression.

$$\sigma^2\{Y_i\} = E\left\{\left[Y_i - E\{Y_i\}\right]^2\right\} = E\{Y_i\}\left[1 - E\{Y_i\}\right] \tag{3.6}$$

$$\varepsilon_i = Y_i - \pi_i \tag{3.7}$$

and π_i is constant. Due to this reason the variance of ε_i is the same as that of Y_i.

$$\sigma^2\{\varepsilon_i\} = E\left\{\left[Y_i - E\{Y_i\}\right]^2\right\} = E\{Y_i\}\left[1 - E\{Y_i\}\right] \tag{3.8}$$

$$\sigma^2\{\varepsilon_i\} = \left[\beta_0 + \beta_1 X_i\right]\left[1 - \beta_0 - \beta_1 X_i\right] \tag{3.9}$$

As can be seen equation (3.9) variances of ε_i depend on X_i . Thus, using ordinary least square would not be appropriate.

3.1.1.1 Logistic Response Function

Logistic response functions are used for describing the nature of the relationship between the mean response and one or more independent variables.

The response function plotted in Figure 3.1 is named logistic response function and is the form:

$$E(Y) = \frac{e^{\beta X_i}}{1 + e^{\beta X_i}} \tag{3.10}$$

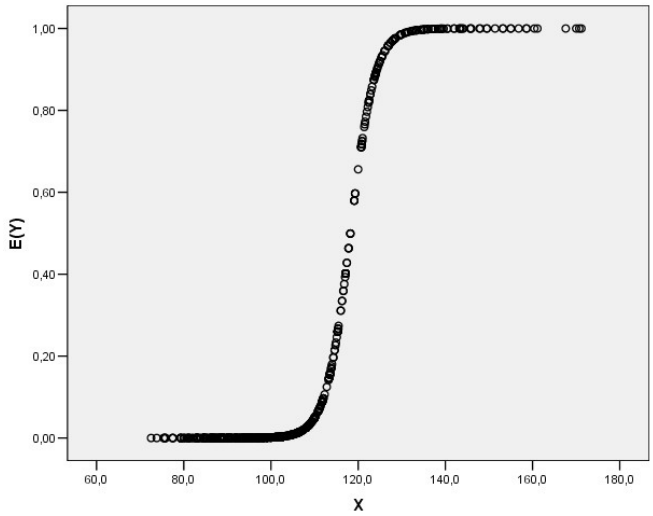


Figure 3.1 Simple logistic response function

The response function showed in Figure 3.1 is shaped either as a tilted S and, that it is approximately linear except the ends. This response function is often referred to as sigmoidal.

3.1.2 Fitting of Multiple Logistic Regression

The simple logistic regression model is easily expanded to more than one independent variable. Take into account a collection of p independent variables that will be denoted by the vector $x' = (x_1, x_2, ..., x_p)$. Multiple logistic regression model is $\beta_0 + X_1\beta_1 + X_2\beta_2 + ... + X_p\beta_p$. To simplify the formulas, we can use the matrix notation:

$$\beta' X_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + ... + X_{ip}\beta_p \tag{3.11}$$

With this notation, the simple logistic response function extends to the multiple logistic response function as follows:

$$E\{Y_i\} = \pi_i = \frac{\exp(\beta' X_i)}{1 + \exp(\beta' X_i)} \tag{3.12}$$

Like the simple logistic response function, the multiple logistic response function is monotonic and sigmoidal in shape with respect to $\beta' X$ and is almost linear when π is between 0.2 and 0.8.

3.1.2.1 Likelihood Function

The likelihood function is the joint probability (density) function of observable random variables but it is investigated as the function of the parameters given the carried out random variables.

In the binary logistic regression each Y_i observation is a Bernoulli random variable. Its probability distribution as follows:

$$f_i(Y_i) = \pi_i^{Y_i}(1-\pi_i)^{1-Y_i} \qquad Y_i = 0,1 \qquad i = 1,...,n \tag{3.13}$$

As the Y_i observations are independent, their joint probability function is:

$$g(Y_1,...,Y_n) = \prod_{i=1}^{n} f_i(Y_i) = \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i} \tag{3.14}$$

To find the maximum likelihood estimates we must working with logarithm of the joint probability function:

$$\begin{aligned} \log_e g(Y_1,...,Y_n) &= \log_e \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i} \\ &= \sum_{i=1}^{n}\left[Y_i \log_e\left(\frac{\pi_i}{1-\pi_i}\right)\right] + \sum_{i=1}^{n} \log_e(1-\pi_i) \end{aligned} \tag{3.15}$$

Since $E\{Y_i\} = \pi_i$ for a binary variable, it follows from (3.5) that:

$$1 - \pi_i = \left[1 + \exp(\beta_0 + \beta_1 X_i)\right]^{-1} \tag{3.16}$$

From (3.5) and (3.16) we obtain:

$$\log_e \left(\frac{\pi_i}{1 - \pi_i} \right) = \beta_0 + \beta_1 X_i \tag{3.17}$$

We can discovery the log-likelihood function for the simple logistic regression model as follows:

$$\log_e L(\beta_0, \beta_1) = \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 X_i) - \sum_{i=1}^{n} \log_e \left[1 + \exp(\beta_0 + \beta_1 X_i)\right] \tag{3.18}$$

The log-likelihood function for multiple logistic regression model can be also expand as follows:

$$\log_e L(\beta) = \sum_{i=1}^{n} Y_i(\beta' X_i) - \sum_{i=1}^{n} \log_e \left[1 + \exp(\beta' X_i)\right] \tag{3.19}$$

3.1.2.2 Maximum Likelihood Estimation

Method of the maximum likelihood is the most widely used method of estimating the parameters of a logistic regression model. Maximum likelihood estimators are generally obtained by maximizing the logarithm of the likelihood function. The logarithm of the likelihood function is given by (3.19) and differentiating (3.19) with respect to β_0 and then with respect to β_1 produces the two likelihood function, as follows:

$$\frac{\partial \log(L(\beta_0, \beta_1))}{\partial \beta_1 \partial \beta_0} = \sum X_i Y_i - \sum \frac{X_i \exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)} \tag{3.20}$$

The maximum likelihood estimators of β_0 and β_1 are obtained by setting the right side of equation (3.20) equal to zero, and then solving the equations simultaneously so as to produce $\hat{\beta}_0$ and $\hat{\beta}_1$. Iteration would continue until certain convergence criteria are met.

3.1.3 Testing for the Significance of the Coefficients

For testing the significance of the coefficients, the likelihood ratio test and wald statistic is used.

3.1.3.1 Likelihood Ratio Test

A subset of the X variables in a multiple logistic regression model can be dropped, that is, testing whether the associated regression coefficients β_k equal zero. k is refer to number of regression coefficients. The test procedure we shall employ is a general one for use with maximum likelihood estimation, just like the general linear test procedure for linear models. The test is called the likelihood ratio test. It requires a large sample size and is based on a statistic called model deviance.

Comparing the log-likelihood of the fitted model to the log-likelihood of a model with n parameters is named the deviance of a fitted model compares that fits the n observations perfectly. Such a perfectly fitting model is called a saturated model.

We will have n parameters for the n observations and can obtain a perfect fit. It can be shown that the log-likelihood function (3.21) is maximized if $\pi_i = Y_i$. Hence, the maximum likelihood estimator of π_i for the saturated model, denoted by $\hat{\pi}_{is}$, $\hat{\pi}_{is} = Y_i$. It can be displayed that likelihood of the sample observations evaluated at $\hat{\pi}_{is} = Y_i$, denoted by $L(\hat{\pi}_{1s},...,\hat{\pi}_{ns})$ is equal to 1 so that the log-likelihood is equal to 0:

$$\log_e L(\hat{\pi}_{1s},...,\hat{\pi}_{ns}) = \sum_{i=1}^{n} \left[Y_i \log_e (Y_i) + (1-Y_i) \log_e (1-Y_i) \right] = 0 \qquad (3.21)$$

This log-likelihood value for the saturated model will now be compared with the log-likelihood value for the fitted model. The maximize log-likelihood function of the fitted model as follows:

$$\log_e L(b_0, b_1, ..., b_p) = \sum_{i=1}^{n} Y_i (b' X_i) - \sum_{i=1}^{n} \log_e \left[1 + \exp(b' X_i)\right] \tag{3.22}$$

The deviance is depending on the difference between the two log-likelihood values.

$$\begin{aligned} DEV(X_0, X_1, ..., X_p) &= 2\log_e L(\hat{\pi}_{1s}, ..., \hat{\pi}_{ns}) - 2\log_e L(b_0, b_1, ..., b_p) \\ &= -2\sum_{i=1}^{n} \left[Y_i \log_e(\pi_i) + (1 - Y_i)\log_e(1 - \pi_i)\right] \end{aligned} \tag{3.23}$$

where π_i is the i th fitted value for the logistic regression model. For logistic regression, the deviance (also known as residual deviance) is used to determine the fit of the overall model. The deviance for a logistic model can be approximated to the residual sum of squares in ordinary regression. The smaller deviance is the better to fit of the model. The deviance can be contrasted to a chi-square distribution, which approximates the distribution of the deviance.

For aims of determine the significance of an independent variable we compare the value of DEV with and without the independent variable in the equation. The variation in DEV due to including the independent variable in the model is obtained as follows:

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

This statistic plays the same role in logistic regression as does the numerator of the partial F test in linear regression. Because the likelihood of the saturated model is

common to both values of *DEV* being differenced to compute G, it can be expressed as

$$G = -2\ln\left[\frac{\text{likelihood without the variable}}{\text{likelihood with the variable}}\right] \tag{3.24}$$

The G statistic is also a likelihood ratio test. It will follow a chi-square distribution with p degrees of freedom.

3.1.3.2 Wald Statistic

In linear regression, t-statistics are used in assessing the value of individual predictor when other predictors are in the model. In logistic regression

$$W = \frac{\hat{\beta}_i}{s_{\hat{\beta}_i}} \tag{3.25}$$

is called a Wald statistic. (Hosmer, Lemeshow, 1989) There is no agreement as to the general form of what is being called a Wald statistic. Equation (3.25) is given by Hosmer and Lemeshow (1989) but $\frac{\hat{\beta}_i^2}{s_{\hat{\beta}_i}^2}$, written in a different but equivalent form, is termed a Wald statistic by Rao (1973) and also by Wald (1943).

3.1.4 Interpretation of the Coefficients of the Logistic Regression Model

After fitting model importance alterations from the computation and determine of significance of estimated coefficients to interpretation of their values. 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. Thus interpretation supposes two issues, assessing the functional relationship

between the dependent variable and the independent variable, and appropriately defining the unit of change for the independent variable.

Appropriate interpretation of the coefficient in a logistic regression model bases on being able to place meaning on the difference between two logits. In the following sections, each of the possible measurement scales of the independent variable will be consider the interpretation of the coefficients.

3.1.4.1 Dichotomous Independent Variable

We assume that x is coded as either 0 or 1. Under this model there are two values of $\pi(x)$ and equivalently two values of $1-\pi(x)$. The odds of the outcome being present among individuals with $x=1$ is assigned as $\pi(1)/[1-\pi(1)]$. In a like manner, the odds of the outcome being present among individuals with $x=0$ is defined as $\pi(0)/[1-\pi(0)]$. The log of the odds is named the logit and these are

$$g(1)=\ln\{\pi(1)/[1-\pi(1)]\}$$

and

$$g(0)=\ln\{\pi(0)/[1-\pi(0)]\}$$

The odds ratio is defined as the ratio of the odds for $x=1$ to the odds for $x=0$ and is given by the equation

$$\psi=\left[\frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]}\right] \tag{3.26}$$

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

$$\begin{aligned}\ln(\psi) &= \ln\left[\frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]}\right] \\ &= g(1)-g(0)\end{aligned} \tag{3.27}$$

which is the logit difference.

$$\psi = \frac{\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right)\left(\frac{1}{1+e^{\beta_0}}\right)}{\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right)\left(\frac{1}{1+e^{\beta_0+\beta_1}}\right)} \qquad (3.28)$$
$$= \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

For logistic regression with a dichotomous independent variable

$$\psi = e^{\beta_1} \qquad (3.29)$$

and the logit difference, or log odds, is

$$\ln(\psi) = \ln(e^{\beta_1}) = \beta_1 \qquad (3.30)$$

The odds ratio is a measure of association that has constructed wide use as it approximates how much more similarly (or unsimilarly) it is for the outcome to be present among those with $x = 1$ than among those with $x = 0$. For example, if y denotes the presence or absence of the having insurance and if x denotes whether or not the person has an accident, then $\hat{\psi} = 2$ indicates that having insurance occurs twice as often among has accident than among has not in the study.

The confidence interval of the odds ratio are

$$\exp\left[\hat{\beta}_1 \pm z_{1-\alpha/2} S\hat{E}(\hat{\beta}_1)\right] \qquad (3.31)$$

Importance of the odds ratio as a measure of association, point and interval estimates are often found in additional columns in tables presenting the results of a logistic regression. If the confidence interval contains 1 then there is no association.

3.1.4.2 Polytomous Independent Variable

Suppose that instead of two categories the independent variable has $k > 2$ distinct values. For example, we may have variables that denote the county of residence within a state, the clinic used for primary health care within a city, or race. Each of these variables has fixed number of discrete outcomes and the scale of measurement is nominal. Table 3.1 shows that frequencies of groups.

Table 3.1 Cross classification of the data on independent (X) and dependent (Y)

Y		X				Total
		1	2	3	4	
Absent	0	5	20	15	10	50
Present	1	20	10	10	10	50
Total		25	30	25	20	100
Odds Ratio		1.0	8.0	6.0	4.0	
CI			2.3-27.6	1.7-21.3	1.1-14.9	

For example, for the third group the estimated odds ratio is $(15\times 20)/(5\times 10) = 6.0$. The reference group is the first group. The odds ratio and confidence interval (CI) are shown in Table 3.1.

3.1.4.3 Continuous Independent Variable

When a logistic regression model includes a continuous independent variable, interpretation of the estimated coefficient will depend on how it is entered into the model and the accurate units of the variable.

To obtain a helpful interpretation for continuous scaled independent variables we need to develop a method for point and interval estimation for an arbitrary change of "c" units in the covariate.

The log odds for a change of c units in x is obtained from the logit difference $g(x+c)-g(x)=c\beta_1$ and the associated odds ratio is obtained by exponentiating this logit difference, $\psi(c)=\psi(x+c,x)=\exp(c\beta_1)$.

An estimate of the standard error needed for confidence interval estimation provided by multiplying the estimated standard error of $\hat{\beta}_1$ by c. The confidence interval estimate of $\psi(c)$ are

$$\exp\left[c\hat{\beta}_1 \pm z_{1-\alpha/2}cS\hat{E}(\hat{\beta}_1)\right] \tag{3.32}$$

Both the point estimate and endpoints of the confidence interval depend on the choice of c, the particular value of c should be clearly specified in all tables and calculations.

3.1.5 Logistic Regression Model Selection Methods

Method selection permits you to define how independent variables are entered into the analysis. Using different methods, you can arrange a variety of regression models from the same set of variables.

3.1.5.1 Forward Selection

The forward-selection technique starts with no variables in the model. For each of the independent variables, the method calculates F statistics that express the variable's contribution to the model if it is included. The p-values for these F statistics are compared to the α. If no F statistic has a p-value less than the α,

method stops. Otherwise, method adds the variable that has the smallest p-value less than α. F statistics are calculated again for the variables still remaining outside the model, and the evaluation process is repeated. Thus, variables are added one by one to the model until no remaining variable produces a significant F statistic.

3.1.5.2 Backward Elemination

The backward-elimination technique starts by calculating F statistics for the full model that includes all of the independent variables. Until all the variables remaining in the model produce F statistics significant at the α, the variables are deleted from the model one by one (Fomby, 2005).

3.2 Ordinal Logistic Regression

Logistic regression is most commonly used to model the relationship between a dichotomous dependent variable and independent variables. However, the dependent variable has more than two categories. Logistic regression can still be employed, by means of an ordinal logistic regression model. It is also known as polytomous logistic regression model.

In social applications of controlled learning frequently requires situations indicating an order among the different categories, e.g. a teacher always rates his/her students by giving grades on their term performance. In contrast to measurable regression problems, the grades are usually discrete. These grades are also different from the class labels in classification problems due to the existence of ranking information. For example, grade labels have the ordering F <D < C < B < A. This is a learning task of predicting variables of ordinal scale, an establishing bridging between measurable regression and classification referred to as ranking learning or ordinal regression (Chu, Ghahramani, 2005).

Ordinal logistic regression refers to the case where the dependent variable has an order; the multinomial case is contained below. The most common ordinal logistic model is the proportional odds model (Flom, n.d.).

3.2.1 Cumulative Probabilities and Their Logits

Let Y denote an ordinal dependent variable. Let $P(Y \leq j)$ describe the probability that the response falls in category j or below (i.e., in category 1, 2,..., or j). This is named a cumulative probability. With four categories, for example, the cumulative probabilities are

$$P(Y = 1),$$
$$\mathrm{P(Y} \leq 2) = P(Y = 1) + P(Y = 2),$$
$$\mathrm{P(Y} \leq 3) = P(Y = 1) + P(Y = 2) + P(Y = 3)$$

and the final cumulative probability uses the entire scale, so $\mathrm{P(Y} \leq 4) = 1$.

A c-category response has c cumulative probabilities. The order of forming the cumulative probabilities expresses the ordering of the response scale. The probabilities satisfy

$$P(Y \leq 1) \leq P(\mathrm{Y} \leq 2) \leq \cdots \leq \mathrm{P(Y} \leq c) = 1$$

The odds of response in category j or below is the ratio

$$\frac{P(Y \leq j)}{P(Y > j)} \tag{3.33}$$

For instance, when the odds equal 2.5, the probability of response in category j or below equals 2.5 times the probability of response above category j. Each cumulative probability can transform to an odds. A popular logistic model for an

ordinal response uses logits of the cumulative probabilities. With $c = 4$, for example, the logits are

$$\text{logit}\left[P(Y \le 1)\right] = \log\left[\frac{P(Y=1)}{P(Y>1)}\right] = \log\left[\frac{P(Y=1)}{P(Y=2)+P(Y=3)+P(Y=4)}\right] \tag{3.34}$$

$$\text{logit}\left[P(Y \le 2)\right] = \log\left[\frac{P(Y \le 2)}{P(Y>2)}\right] = \log\left[\frac{P(Y=1)+P(Y=2)}{P(Y=3)+P(Y=4)}\right] \tag{3.35}$$

$$\text{logit}\left[P(Y \le 3)\right] = \log\left[\frac{P(Y \le 3)}{P(Y>3)}\right] = \log\left[\frac{P(Y=1)+P(Y=2)+P(Y=3)}{P(Y=4)}\right] \tag{3.36}$$

Since the final cumulative probability accordingly equals 1, we close out it from the model. These logits of cumulative probabilities are called cumulative logits (Agresti, 2010).

3.2.2 Cumulative Logit Models for an Ordinal Response

A model can concurrently characterized the impression of an independent variable on all the cumulative probabilities for Y. For each cumulative probability, the model looks like an ordinary logistic regression, where the two outcomes are low = "category j or below" and high = "above category j." This model is

$$\text{logit}\left[P(Y \le j)\right] = \alpha_j - \beta x, \qquad j = 1, 2, ..., c-1$$

for $c = 4$, for example, this single model describes three relationships: the effect of x on the odds that $Y \le 1$ instead of $Y > 1$, the effect of x on the odds that $Y \le 2$

instead of $Y > 2$, and the impression of x on the odds that $Y \leq 3$ instead of Y>3. The model desires a different intercept parameter α_j for each cumulative probability.

If $\beta > 0$, when x higher cumulative probabilities are lower. But cumulative probabilities being lower means it is less likely to observe relatively low values and thus more likely to observe higher values of Y. So, this parameterization accords with the usual formulation of a positive association, in manner of speaking that a positive β corresponds to a positive association (higher x tending to occur with higher Y).

The parameter of main interest, β defines the impressions of x on Y. When $\beta = 0$, each cumulative probability does not change as x changes, and the variables are independent. If $|\beta|$ increases, then the impression of x increases. In this model, β does not have a j subscript. It has the same value for each cumulative logit. To put it in a different way, the model assumes that the impression of x is the same for each cumulative probability. This cumulative logit model with this common impression is often named the proportional odds model.

3.2.3 Odds Ratio

After the proportional odds model is fit and the parameters estimated, the process for computing the odds ratio is the same as in standard logistic regression. We will first evaluate the special case where sustain is the only independent variable and is coded 1 and 0. Recall that the odds comparing $Y \leq j$ vs. $Y > j$ is e to the α_j minus β_1 times X_1. To determine the impression of the exposure on the outcome, we formulate the ratio of the odds of $Y \leq j$ for comparing $X_1 = 1$ and $X_1 = 0$.

$$\begin{aligned} odds(Y \leq j) &= \frac{P(Y \leq j / X_1)}{P(Y > j / X_1)} \\ &= \exp(\alpha_j - \beta_1 X_1) \end{aligned} \tag{3.37}$$

The odds ratio is

$$\begin{aligned}\psi &= \frac{P(Y \le j / X_1 = 1)/P(Y > j / X_1 = 1)}{P(Y \le j / X_1 = 0)/P(Y > j / X_1 = 0)} \\ &= \frac{\exp\left[\alpha_j - \beta_1(1)\right]}{\exp\left[\alpha_j - \beta_1(0)\right]} = \frac{\exp(\alpha_j - \beta_1)}{\exp(\alpha_j)} \\ &= e^{\beta_1}\end{aligned} \tag{3.38}$$

Confidence interval estimation is the similar to standard logistic regression.

$$\exp\left[\hat{\beta}_1 \pm z_{1-\alpha/2} S\hat{E}(\hat{\beta}_1)\right] \tag{3.39}$$

If the 95% confidence interval contain the value 1, the association is not statistically significant at α.

3.2.4 Likelihood Function

In the proportional odds model, we model the probability of $Y \le j$. To access an expression for the probability of $Y = j$, can be use the relationship that the probability ($Y = j$) is equal to the probability of $Y \le j$ minus the probability of $Y < j$. For instance, the probability that $Y = 2$ is equal to the probability that $Y \le 2$ minus the probability that $Y < 2$. In this way we can use the model to access an expression for the probability that an individual is in a specific outcome category for a given pattern of covariates (X).

$$\prod_{i=1}^{n}\prod_{j=0}^{J-1} P(Y = j / X)^{g_{ij}} \qquad g_{ij} = \begin{cases} 1 \text{ if the } j\text{th subject has Y=j} \\ 0 \text{ if otherwise} \end{cases} \tag{3.40}$$

3.2.5 Testing of Parallel Lines

The test of parallel lines is designed to make a test concerning the adequacy of the model. The null hypothesis establishes that the related regression coefficient is equal across all categories of the dependent variable. The alternative hypothesis establishes that the related regression coefficients are different across all categories of dependent variables. According to the test of parallel lines results, we make interpretation whether there are significant or are not significant difference for the corresponding regression coefficients across the dependent variable categories.

3.2.6 Pseudo-R^2

As a starting point, recall that a non-pseudo R-squared is a statistic expanded in ordinary least squares (OLS) regression that is often used as a goodness-of-fit measure. In OLS,

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y}_i)^2} \tag{3.41}$$

where n is the number of observations in the model, y is the dependent variable, $\overline{y}$ is the mean of the y values, and $\hat{y}$ is the value predicted by the model. The numerator of the ratio is the sum of the squared differences between the actual y values and the predicted y values. The denominator of the ratio is the sum of squared differences between the actual y values and their mean.

When analyzing data with a logistic regression, an equivalent statistic to R-squared does not exist. The model estimates from a logistic regression are maximum likelihood estimates arrived at through an iterative process. They are not calculated to minimize variance, so the OLS approach to goodness-of-fit does not apply. However, to evaluate the goodness-of-fit of logistic models, several pseudo R-squared have been developed (Long, Freese, 2006)

3.2.6.1 Mc Fadden's

The log likelihood of the intercept model ($\log \hat{L}(M_{Intercept})$) is treated as a total sum of squares, and the log likelihood of the full model ($\log \hat{L}(M_{Full})$) is treated as the sum of squared errors. The ratio of the likelihoods suggests the level of improvement over the intercept model offered by the full model.

$$R^2 = 1 - \frac{\log \hat{L}(M_{Full})}{\log \hat{L}(M_{Intercept})} \tag{3.42}$$

A likelihood falls between 0 and 1, so the log of a likelihood is less than or equal to zero. If a model has a very low likelihood, then the log of the likelihood will have a larger magnitude than the log of a more likely model. Thus, a small ratio of log likelihoods indicates that the full model is a far better fit than the intercept model.

If comparing two models on the same data, McFadden's would be higher for the model with the greater likelihood.

3.2.6.2 Cox and Snell

The ratio of the likelihoods reflects the improvement of the full model over the intercept model (the smaller the ratio, the greater the improvement).

$$R^2 = 1 - \left[\frac{L(M_{Intercept})}{L(M_{Full})} \right]^{2/N} \tag{3.43}$$

Consider the definition of *L(M).* *L(M)* is the conditional probability of the dependent variable given the independent variables. If there are N observations in the dataset, then *L(M)* is the product of N such probabilities. Thus, taking the n^{th} root of the product *L(M)* provides an estimate of the likelihood of each Y value. Cox &

Snell's presents the R-squared as a transformation of the *-2ln[L($M_{Intercept}$)/L(M_{Full})]* statistic that is used to determine the convergence of a logistic regression.

Note that Cox & Snell's pseudo R-squared has a maximum value that is not 1: if the full model predicts the outcome perfectly and has a likelihood of 1, Cox & Snell's is then 1-$L(M_{Intercept})^{2/N}$, which is less than one.

3.2.6.3 Nagelkerke

Nagelkerke adjusts Cox & Snell's so that the range of possible values extends to 1. To achieve this, the Cox & Snell R-squared is divided by its maximum possible value, 1-$L(M_{Intercept})^{2/N}$. Then, if the full model perfectly predicts the outcome and has a likelihood of 1, Nagelkerke R-squared = 1. When $L(M_{Full}) = 1$, then $R^2 = 1$, when $L(M_{Full}) = L(M_{Intercept})$, then $R^2 = 0$.

$$R^2 = \frac{1 - \left[\frac{L(M_{Intercept})}{L(M_{Full})}\right]^{2/N}}{1 - L(M_{Intercept})^{2/N}} \tag{3.44}$$

CHAPTER FOUR
APPLICATION

4.1 Introduction

Data of Family Practice Thesis of Gizem LİMNİLİ, doctor of Dokuz Eylül University Faculty of Family Medicine Department, have been used in application. Data has obtained from four high schools in İzmir, Balçova. Doctor Gizem LİMNİLİ has used this data in her thesis of "*the prevalence of obesity in high school students aged 15-17 in Balçova regions and the relationship between obesity and health promoting behaviors*". This data is used only with the purpose of statistical analysis.

The purpose of this application is to fill the missing in data with the methods of regression imputation, expectation maximization and multiple imputation, and to set mathematical models using logistic regression analysis for dependent variable.

4.2 Description of Data Set

The study is cross sectional and sample width is 1089. Dependent variable is Body Mass Index group of teen in high school (BMIgrp). Body mass index has divided into 3 groups based on World Health Organization limits. The groups are defined according to body mass indexes as; the ones smaller than 18.5 are the 1^{st} group(low-weight), the ones between 18.5 and 23.9 are the 2^{nd} group (normal weight) and the ones bigger than 23.9 are the 3^{rd} group (obese). There are 15 independent variables in data set. The continuous ones of independent variables are lined up with units as age(years), skin folds(centimeter), waist largeness(centimeter), haunch largeness(centimeter) and relative weight ((weight/ideal weight)*100).

Obesity is a medical condition in that overload body fat has cumulative to expand that it may have a reverse effect on health, leading to reduce life expectancy or increased health problems. Body Mass Index (BMI) is a value calculated from an individual's weight and height. BMI obtains a confidential indicator of body fatness

for most individuals and is used to screen weight categories and health problems associated with weight.

For adults, the ones with body mass index (BMI) bigger than 25 are described as over-weight and the ones bigger than 30 are obese. Children with percentile >85 are classified as over-weight and children with percentile >90 are as obese by using BMI percentile curves according to age and gender. Other diagnostic procedures used are body weight according to age, weight according to height and measurement of skin fold thickness. Table 4.1 shows the discrete independent variables in data set.

Table 4.1 Discrete independent variables and their codes and frequency distributions

Name	Codes	Frequencies
Gender	1= Girl	656
	2= Boy	433
N		1089
Missing		0
Class	1 = High School 1	546
	2 = High School 2	385
	3 = High School 3	158
N		1089
Missing		0
Mother Obese	1= Yes	26
	0= No	1063
N		1089
Missing		0
Father Obese	1= Yes	21
	0= No	1068
N		1089
Missing		0
Mother's Education Level	1= High School	943
	2= University	145
N		1088
Missing		1
Father's Education Level	1= High School	816
	2= University	272
N		1088
Missing		1

Table 4.1 is continued

Name	Codes	Frequencies
Health Perceptiveness	1= Top of Average	265
	2= Average and sub of Average	823
N		1088
Missing		1
Carbonated Drink	1= Yes	488
	0= No	600
N		1088
Missing		1
Fastfood	1= Yes	738
	0= No	350
N		1088
Missing		1
Percentile	1= <15p	158
	2= 15p-50p	319
	3=50p-85p	193
	4= >85p	78
N		848
Missing		241

Table 4.2 – 4.11 crosstabs show frequencies of independent variables according to dependent variables and chi-square values belong to them.

Table 4.2 BMIgrp versus class tabulation count

		BMIgrp			Total
		<18.5	18.5< <23.9	> 23.9	
Class	High School 1	65	276	76	417
	High School 2	50	202	60	312
	High School 3	12	80	27	119
Total		127	558	163	848
Chi-square = 3.294			df = 4		p-value = 0.510

Table 4.3 BMIgrp versus gender tabulation count

		BMIgrp			Total
		<18.5	18.5< <23.9	> 23.9	
Gender	Girl	68	341	85	494
	Boy	59	217	78	354
Total		127	558	163	848
Chi-square = 5.532			df = 2		p-value = 0.632

Table 4.4 BMIgrp versus mother obese tabulation count

		BMIgrp			Total
		<18.5	18.5< <23.9	> 23.9	
Mother Obese	No	124	547	156	827
	Yes	3	11	7	21
Total		127	558	163	848
Chi-square = 2.827			df = 2		p-value = 0.243

Table 4.5 BMIgrp versus father obese tabulation count

		BMIgrp			Total
		<18.5	18.5< <23.9	> 23.9	
Father Obese	No	126	547	160	833
	Yes	1	11	3	15
Total		127	558	163	848
Chi-square = 0.841			df = 2		p-value = 0.657

Table 4.6 BMIgrp versus mother education tabulation count

		BMIgrp			Total
		<18.5	18.5< <23.9	> 23.9	
Mother Education	High School	112	498	125	735
	University	15	60	38	113
Total		127	558	163	848
Chi-square = 17.527			df = 2		p-value = 0.000

Table 4.7 BMIgrp versus father education tabulation count

		BMIgrp			Total
		<18.5	18.5< <23.9	> 23.9	
Father Education	High School	101	428	108	637
	University	26	130	55	211
Total		127	558	163	848
Chi-square = 8.917			df = 2		p-value = 0.012

Table 4.8 BMIgrp versus fast food tabulation count

		BMIgrp			Total
		<18.5	18.5< <23.9	> 23.9	
Fast food	No	40	203	46	289
	Yes	87	355	117	559
Total		127	558	163	848
Chi-square = 4.182		df = 2		p-value = 0.124	

Table 4.9 BMIgrp versus carbonated drink tabulation count

		BMIgrp			Total
		<18.5	18.5< <23.9	> 23.9	
Carbonated Drink	No	67	324	91	482
	Yes	60	234	72	366
Total		127	558	163	848
Chi-square = 1.273		df = 2		p-value = 0.529	

Table 4.10 BMIgrp versus health tabulation count

		BMIgrp			Total
		<18.5	18.5< <23.9	> 23.9	
Health	Average and sub of average	30	145	43	218
	Top of Average	97	413	120	630
Total		127	558	163	848
Chi-square = 0.350		df = 2		p-value = 0.839	

Table 4.11 BMIgrp versus percentile tabulation count

		BMIgrp			Total
		<18.5	18.5< <23.9	> 23.9	
Percentile	< 15p	54	98	6	158
	15p-50p	65	239	24	319
	50p-85p	16	206	71	293
	> 85p	1	15	62	78
Total		127	558	163	848
Chi-square = 292.254		df = 6		p-value = 0.000	

Pearson chi-square test analyzes independence of variables. Null hypothesis is set as “variables are independent” and alternative hypothesis is as “variables are not independent”. Degrees of freedom is (r-1)*(c-1) and r defines row number, c defines column number. $p < \alpha$ shows the hypothesis is rejected and therefore variables are not independent.

Table 4.12 shows descriptive statistics and Table 4.13 shows correlation coefficients of continuous independent variables.

Table 4.12 Descriptive statistics of continuous independent variables

Variables	N	Missing	Mean	Std. Dev.
Age	1089	0	17.84	0.77
Skin Folds	848	241	13.51	5.27
Waist	848	241	72.44	8.19
Haunch	848	241	96.15	7.67
Relative Weight	848	241	105.16	16.26

Table 4.13 Correlation table of continuous independent variables

	Age	Skin Folds	Waist	Haunch	Relative Weight
Age	1				
Skin Folds	-0.050	1			
Waist	0.051	0.350	1		
Haunch	0.081	0.509	0.768	1	
Relative Weight	0.032	0.601	0.721	0.789	1

In Table 4.12 there is no missing observation in “age” variable however there are 241 units of loss in other continuous independent variables. In Table 4.1 there are 241 units of loss in “percentile” variable in discrete independent variables. Besides, there is no loss in variables in discrete independent variables.

Histograms in Figure 4.1 show values of continuous independent variables.

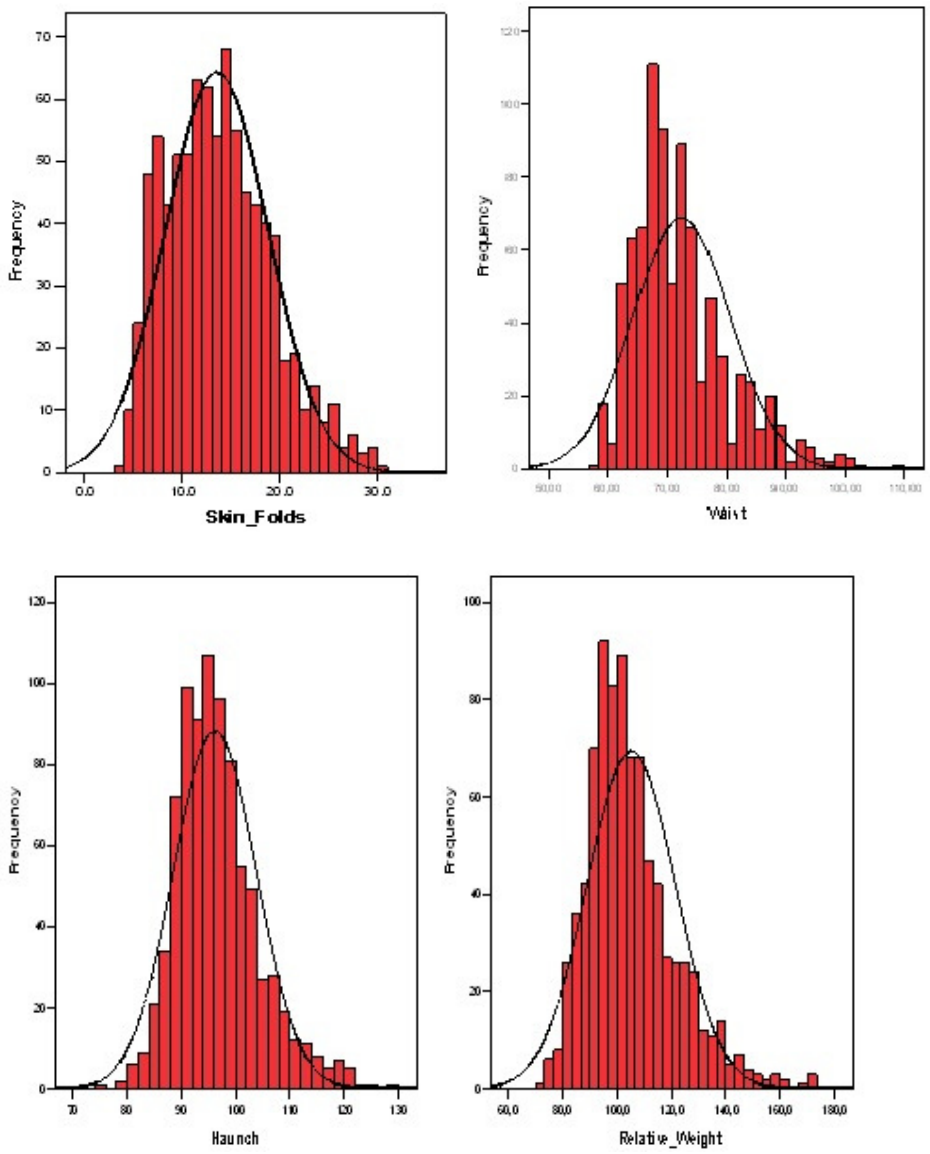


Figure 4.1 Histograms of continuous independent variables

Table 4.14 Results of Kolmogorov Smirnov test

Variables	Statistic	df	p-value
Skin Folds	0.052	848	0.000
Waist	0.136	848	0.000
Haunch	0.095	848	0.000
Relative Weight	0.090	848	0.000

Figure 4.1 shows us the independent variables are not normally distributed. It can be supported by Kolmogorov Smirnov test which is the normality test. The null hypothesis is the sample data are not significantly different than a normal population. In Table 4.14 shows us the independent variables are significantly different than a normal population.

4.3 Missing Data Analysis

Determining missing data mechanism play role effectually in choosing the correct analysis for the solution of missing data problem and choosing the right analysis is important for reaching the right result for data evaluation. Table 4.15 shows which missing data is on which variable and unit.

Table 4.15 Missingness data patterns

Missing Patterns							
Number of Cases	Waist	Haunch	Relative Weight	Skin Folds	Percentile	BMIgrp	Complete if
848							848
241	X	X	X	X	X	X	1089

Table 4.15 shows the dependent variable “BMIgrp” and there are 241 missing observations in independent variables and 848 observations have seen in all variables. Missing values in variables are in same units. This missing data pattern fits to multivariate pattern.

There can be several reasons for these missing values. For instance, person has not responded the question for any reason or known the reply. In medical studies, there may be cases like discontinuity, refusing treatment, death, or etc.

4.3.1 Questioning Missing Data Process

There are three different methods for questioning missing data process. The first method used for determining the missing data mechanism is to examine whether

there is any significant difference between or not the values of observed and missing data for specific variables. This research can be done with “t test” that analyzes the significant of difference between two means. Significant difference shows the presence of missing not at random data process. Table 4.16 shows “t-test” and “p-values”.

Table 4.16 Separate variance t-tests

		Age	Skin Folds	Waist	Haunch	Relative Weight
Skin Folds	t	0.958	.	.	.	.
	df	377.387	.	.	.	.
	# Present	848	848	848	848	848
	# Missing	241	0	0	0	0
	Mean(Present)	17.851	13.513	72.435	96.152	105.162
	Mean(Missing)	17.797	.	.	.	.
	p-value	0.330				
Waist	t	0.958	.	.	.	.
	df	377.387	.	.	.	.
	# Present	848	848	848	848	848
	# Missing	241	0	0	0	0
	Mean(Present)	17.851	13.513	72.435	96.152	105.162
	Mean(Missing)	17.797	.	.	.	.
	p-value	0.330				
Haunch	t	0.958	.	.	.	.
	df	377.387	.	.	.	.
	# Present	848	848	848	848	848
	# Missing	241	0	0	0	0
	Mean(Present)	17.851	13.513	72.435	96.152	105.162
	Mean(Missing)	17.797	.	.	.	.
	p-value	0.330				
Relative Weight	t	0.958	.	.	.	.
	df	377.387	.	.	.	.
	# Present	848	848	848	848	848
	# Missing	241	0	0	0	0
	Mean(Present)	17.851	13.513	72.435	96.152	105.162
	Mean(Missing)	17.797	.	.	.	.
	p-value	0.330				

$$H_0 : \mu_{\text{skinfolds}_{\text{obs}}} = \mu_{\text{skinfolds}_{\text{mis}}}$$
$$H_1 : \mu_{\text{skinfolds}_{\text{obs}}} \neq \mu_{\text{skinfolds}_{\text{mis}}}$$

When independent variable "age" values in places where "skinfolds" independent variable is observed and not observed are compared, $p > \alpha$, namely 0.330 >0.05 means that there is no significant difference between them. This result means that the reason for the data of missing is not depended on the variable observed or not observed. In the light of this, missing data mechanism fits MAR (missing at random) mechanism. As the missingness are in same unit in all independent variables, average values of groups are same for the other independent variables.

Second method for questioning randomness is to analyze the correlation of variables. Variables in data set are categorized as the ones having missing value and the ones not having missing value. Present values are coded as 1 and missing values as 0 and correlation coefficients between these variables are calculated. Obtained correlation coefficients define the degree of relationship between missing values for each variable pair. Small correlation coefficient means randomness. Table 4.17 shows us the correlation coefficients of variables. (See the appendix for Table 4.17.).

Chi-Square statistics has been developed to test whether Roderick J. A. Little's missing data mechanism present or not.

Like the t-test approach, Little's test evaluates means differences across subgroups of cases that share the same missing data pattern. The test statistic is a weighted sum of the standardized differences between the subgroup mean and the grand means. Table 4.18 has SPSS output of Little's MCAR tests.

Table 4.18 Continuous independent variables' covariance and Little's MCAR test

	Skin_Folds	Waist	Haunch	Relative_Weight
Skin_Folds	27.728			
Waist	15.100	67.004		
Haunch	20.574	48.199	58.839	
Relative_Weight	51.414	95.963	98.468	264.381

Little's MCAR test: chi-square = 0.952 *d*f=1 p-value=0.329

H_0:The data are missing completely at random
H_1:The data are not missing completely at random

According to the Little's MCAR test, $p > \alpha$ namely 0.329>0.05 means missing data mechanism fits to MCAR mechanism. Missingness in data set can be filled as well as be ignored.

4.3.2 Regression Imputation

Regression imputation imputes missing values by estimated values from a regression of the missing item on items observed for the unit, usually calculated from units with both observed and missing variables present (Rubin, 2002). Table 4.19 shows correlation table of continuous independent variables on result of regression imputation.

Table 4.19 Correlation of regression imputation

	Skin Folds	Waist	Haunch	Relative Weight
Skin Folds	1			
Waist	0.285	1		
Haunch	0.430	0.649	1	
Relative Weight	0.481	0.547	0.601	1

Analyzing the correlation among continuous independent variables in regression imputation (Table 4.13) and correlation among continuous independent variables in complete case, relationship ratios among variables applied regression imputation is decreased.

When the assumption of the MCAR mechanism is satisfied and imputations depend on the present values of other independent variables then the coefficients of ordinary least square is consistent. It means the estimations are almost unbiased.

4.3.3 EM Algorithm

EM algorithm estimates parameter for missing values in estimated distribution of E step and continues in M step till the estimated parameter's likelihood function become non decreasing.

Table 4.20 Correlation of EM algorithm

	Skin Folds	Waist	Haunch	Relative Weight
Skin Folds	1			
Waist	0.350	1		
Haunch	0.509	0.768	1	
Relative Weight	0.600	0.721	0.789	1

Correlation among the continuous independent variables in complete case and correlation among the continuous independent variables in EM algorithm (Table 4.13) are seen not to be different.

4.3.4 Multiple Imputation

Every missing data is valued as M>1 in multiple imputation. In our application, M=5 is used and multiple imputation method SOLAS package software has been used. Table 4.21 and 4.22 show standard deviation and estimated averages of independent variables filled by different methods.

Table 4.21 Summary of estimated means

	Skin Folds	Waist	Haunch	Relative Weight
Complete Case	13.513	72.435	96.150	105.162
EM	13.516	72.430	96.140	105.156
Regression	13.510	72.429	96.050	105.278
M=1	13.495	72.183	96.000	104.850
M=2	13.507	72.323	96.117	105.010
M=3	13.482	72.519	96.146	105.120
M=4	13.540	72.476	96.220	105.330
M=5	13.498	72.500	96.116	105.140

Table 4.22 Summary of estimated standard deviations

	Skin Folds	Waist	Haunch	Relative Weight
Complete Case	5.266	8.185	7.670	16.260
EM	5.266	8.186	7.671	16.260
Regression	5.294	8.126	7.692	16.451
M=1	5.327	8.462	7.712	16.306
M=2	5.269	8.355	7.716	16.525
M=3	5.306	8.184	7.853	16.280
M=4	5.241	8.128	7.585	16.291
M=5	5.206	8.098	7.559	16.373

Missing 241 units could not be predicted by missing regression analysis in dependent variable due to their presence in same units for all variables. Continuous form of dependent variable has been filled with EM and regression imputation method, and then they have been converted into ordinal data. In multiple imputation method, dependent variables are filled ordinally.

Table 4.23 shows ANOVA values of continuous independent variables filled by different methods.

$$H_0 : \mu_{CC} = \mu_{EM} = \mu_{REG} = \mu_{M=1} = ... = \mu_{M=5}$$
$$H_1 : \text{at least one variable is different}$$

Table 4.23 One way ANOVA

		Sum of Squares	df	Mean Square	F	p-value
Skin Folds	Between Groups	2.251	7	0.322	0.012	1.000
	Within Groups	228426.536	8458	27.007		
	Total	228428.786	8465			
Waist	Between Groups	92.044	7	13.149	0.200	0.985
	Within Groups	555013.373	8458	65.620		
	Total	555105.417	8465			
Haunch	Between Groups	32.348	7	4.621	0.081	0.999
	Within Groups	480655.691	8458	56.829		
	Total	480688.039	8465			
Relative Weight	Between Groups	169.637	7	24.234	0.093	0.999
	Within Groups	2196197.789	8458	259.659		
	Total	2196367.426	8465			

When group averages of variables filled with different methods are compared, it can be said that there is no difference between averages for all variables as $p > \alpha$.

4.4 Logistic Regression Analysis

After obtaining descriptive statistics of data set, it is determined that missings are in which variable and where. Logistic Regression analysis will be applied to four different types of data. The first of these is the data set obtained by ignoring the losses rather than any missing data filling process. Second is data set whose losses are filled by Expectation Maximization, third is by Regression Imputation and fourth is by Multiple Imputation.

Shall Wald statistics of dependent variables' in model are bigger than 2, this means that variables are significant (Neter, 1996). P-value of Wald statistics also supports this significant. Confidence interval of odds ratio not containing 1 shows that there is a meaningful relationship between independent variable and dependent variable.

It can be deduced from p-value of G statistics whether model is appropriate or not. Test of parallel lines is a test analyzing whether the calculated independent variables correlations are same or not with all levels of dependent variable. Null hypothesis is set as regression coefficients are equal for all levels of dependent variable and alternative hypothesis is set regression coefficients are not equal for all levels of dependent variable.

4.4.1 Logistic Regression Model of Complete Case Analysis

Before applying the logistic regression to data, we select the independent variables using methods of forward selection. Table 4.24 has been obtained in conclusion of logistic regression analysis done with these variables.

Table 4.24 Logistic regression model of complete case

		Coef (SE coef)	Z (p-value)	Wald	Odds ratio	OR CI Lower	OR CI Upper
Threshold	BMIgrp=1	42.398 (3.254)	13.03 (0.000)	169.715			
	BMIgrp=2	54.642 (4.039)	13.53 (0.000)	183.062			
Location	Waist	0.138 (0.025)	5.52 (0.002)	30.439	1.148	1.093	1.206
	Relative Weight	0.360 (0.028)	13.06 (0.000)	170.574	1.433	1.358	1.514
	Skin folds	0.095 (0.031)	3.07 (0.000)	9.935	1.099	1.035	1.169

Considering Table 4.24, odds ratio of "Skin folds" variable is 1.099. Unit of continuous independent variables is very important for odds ratio interpretation. 10 cm increase in skin folds results in $e^{c\beta_{\text{skin folds}}}$ namely $e^{10*0.095} = 2.59$ unit change in

dependent variable. Confidence interval of odds ratio not containing 1 shows that there is a meaningful relationship between it and dependent variable.

Table 4.25 has information on model fitting for complete case Table 4.26 gives values of test whether regression coefficients are parallel or not.

Table 4.25 Model fitting information of complete case

Model	-2 Log Likelihood	Chi-Square	df	P-value
Intercept Only	1485.569			
Final	424.969	1060.600	3	0.000

$$G = 1458.596 - 424.969$$
$$= 1060.600$$

Regarding p-value of G statistics distributed chi-square with 3 degree of freedom, it is said that the model is appropriate.

Table 4.26 Test of parallel lines of complete case

Model	-2 Log Likelihood	Chi-Square	df	P-value
Null Hypothesis	424.969			
General	423.139	1.829	3	0.609

Considering Table 4.26, null hypothesis cannot be rejected as $p > \alpha$. Regression coefficients are same for all levels of dependent variable.

4.4.2 Logistic Regression Model of EM Algorithm

Missing values of original data are filled using EM algorithm and Table 4.26 has been achieved.

Table 4.27 Logistic regression model of imputed by EM algorithm

		Coef (SE coef)	Z (p-value)	Wald	Odds ratio	OR CI Lower	OR CI Upper
Threshold	BMIgrp=1	43.721 (3.162)	13.83 (0.000)	191.162			
	BMIgrp=2	56.353 (3.908)	14.42 (0.000)	207.979			
Location	Waist	0.142 (0.025)	5.67 (0.000)	32.136	1.15	1.097	1.212
	Relative Weight	0.371 (0.027)	13.83 (0.000)	191.152	1.45	1.376	1.606
	Skin folds	0.099 (0.031)	3.15 (0.002)	9.942	1.11	1.038	1.175

In Table 4.27, the odds ratio of "Relative Weight" variable is 1.45. Unit of continuous independent variables is very important for odds ratio commentary. %25 increase in "Relative Weight" results in $e^{0.25*0.371} = 1.097$ unit change in dependent variable. . Confidence interval of odds ratio not containing 1 shows that there is a relationship between it and dependent variable.

Table 4.28 gives information on model fitting for EM algorithm. Table 4.29 gives results whether regression coefficients are ame or not of the model obtained by data set filled with EM algorithm.

Table 4.28 Model fitting information of EM algorithm

Model	-2 Log Likelihood	Chi-Square	df	P-value
Intercept Only	1658.404			
Final	427.344	1231.060	3	0.000

$$G = 1658.404 - 427.344$$
$$= 1231.060$$

Regarding p-value of G statistics distributed chi-square with 3 degree of freedom, it is said that the model is appropriate.

Table 4.29 Test of parallel lines of EM algorithm

Model	-2 Log Likelihood	Chi-Square	df	P-value
Null Hypothesis	427.344			
General	426.088	1.256	3	0.740

Null hypothesis cannot be rejected as the α is bigger than p-value. This means, regression coefficients are same for all levels of dependent variable.

4.4.3 Logistic Regression Model of Regression Imputation

Missing values of original data are filled using regression imputation and Table 4.30 has been achieved.

Table 4.30 Logistic regression model of imputed by regression imputation

		Coef (SE coef)	Z (p-value)	Wald	Odds ratio	OR CI Lower	OR CI Upper
Threshold	BMIgrp=1	39.093 (3.548)	11.01 (0.000)	121.400			
	BMIgrp=2	50.977 (4.172)	12.22 (0.000)	149.267			
Location	Haunch	0.101 (0.030)	3.37 (0.001)	11.345	1.11	1.042	1.174
	Relative Weight	0.358 (0.028)	12.79 (0.000)	164.177	1.43	1.354	1.509
	Percentile=1	2.250 (0.699)	3.22 (0.001)	10.349	9.49	2.408	37.338
	Percentile=2	1.806 (0.664)	2.72 (0.006)	7.409	6.09	1.659	22.354
	Percentile=3	1.519 (0.624)	1.86 (0.015)	5.926	4.57	1.344	15.503
	Percentile=4	0	.	.	.	.	.

Missing data of data set has been filled with regression imputation method and logistic regression analysis applied to obtained data is shown in Table 4.30. Percentile independent variable is a categorical variable. If the first level of this variable increases 9.49 units then dependent variable will be 1 increase unit.

Table 4.31 gives information on model fitting for regression imputation. Table 4.32 gives results whether regression coefficients are same or not of the model obtained by data set filled with regression imputation.

Table 4.31 Model fitting information of regression imputation

Model	-2 Log Likelihood	Chi-Square	df	P-value
Intercept Only	1459.804			
Final	404.831	1054.973	5	0.000

$$G = 1459.804 - 404.831$$
$$= 1054.973$$

Regarding p-value of G statistics distributed chi-square with 5 degree of freedom, it is said that the model is appropriate.

Table 4.32 Test of parallel lines of regression imputation

Model	-2 Log Likelihood	Chi-Square	df	P-value
Null Hypothesis	404.831			
General	397.038	7.793	5	0.168

Null hypothesis cannot be rejected as the p-value is bigger than α. This means, regression coefficients are same for all levels of dependent variable.

4.4.4 Logistic Regression Model of Multiple Imputation

5 data set (M=1, M=2 … M=5) has been obtained from variables filled with multiple imputation method. Each data set is applied logistic regression. Table 4.33 gives results of M=1 data set obtained by multiple imputation.

Table 4.33 Multiple imputation M=1

		Coef (SE coef)	Z (p-value)	Wald	Odds ratio	OR CI Lower	OR CI Upper
Threshold	BMIgrp=1	31.820 (2.086)	15.25 (0.000)	232.654			
	BMIgrp=2	41.088 (2.496)	16.46 (0.000)	271.034			
Location	Relative Weight	0.283 (0.018)	15.72 (0.000)	247.404	1.33	1.281	1.376
	Skin Folds	0.056 (0.024)	2.33 (0.018)	5.573	1.06	1.010	1.108
	Waist	0.065 (0.018)	3.61 (0.000)	12.559	1.07	1.029	1.106
	Father Obese=0	1.585 (0.662)	2.39 (0.017)	5.736	4.88	1.334	17.832
	Father Obese=1	0	.	.	.	.	.
	Helath=1	0.476 (0.231)	2.06 (0.039)	4.264	1.61	1.024	2.529
	Health=2	0	.	.	.	.	.

Considering odds ratio of Father obese variable in Table 4.33, we can deduce that 1 unit change in this variable result in 4.88 unit change in dependent variable. Children of high school whose fathers are obese are more prone to be obese.

Table 4.34 gives information on model fitting for M=1. Table 4.35 gives results whether regression coefficients are same or not of the model obtained by the first data set filled with multiple imputation.

Table 4.34 Model fitting information of M=1

Model	-2 Log Likelihood	Chi-Square	df	P-value
Intercept Only	1979.061			
Final	709.201	1269.859	5	0.000

$$G = 1979.061 - 709.201$$
$$= 1269.859$$

From Table 4.34, regarding p-value of G statistics distributed chi-square with 5 degree of freedom, it is said that the model is appropriate.

Table 4.35 Test of parallel lines of M=1

Model	-2 Log Likelihood	Chi-Square	df	P-value
Null Hypothesis	709.201			
General	708.367	0.834	5	0.975

Null hypothesis cannot be rejected as the p-value is bigger than α. This means, regression coefficients are same for all levels of dependent variable.

Results of M=2 data set obtained with multiple imputation is given on Table 4.36.

Table 4.36 Multiple imputation M=2

		Coef (SE coef)	Z (p-value)	Wald	Odds ratio	OR CI Lower	OR CI Upper
Threshold	BMIgrp=1	28.518 (2.102)	13.57 (0.000)	184.144			
	BMIgrp=2	38.455 (2.528)	15.22 (0.000)	231.339			
Location	Waist	0.048 (0.020)	2.40 (0.015)	5.931	1.05	1.009	1.090
	Relative Weight	0.302 (0.019)	15.89 (0.000)	258.103	1.35	1.303	1.404
	Percentile=1	1.666 (0.515)	3.24 (0.001)	10.458	5.29	1.927	14.527
	Percentile=2	1.381 (0.481)	2.87 (0.004)	8.264	3.98	1.553	10.206
	Percentile=3	0.983 (0.449)	2.19 (0.029)	4.792	2.67	1.108	6.443
	Percentile=4	0	.	.	.	.	.

Percentile independent variable is a categorical variable. If the first level of this variable increases 5.29 units then the dependent variable will be increase 1 unit.

Table 4.37 gives information on model fitting for M=2. Regarding p-value of G statistics distributed chi-square with 5 degree of freedom, it is said that the model is appropriate.

Table 4.37 Model fitting information of M=2

Model	-2 Log Likelihood	Chi-Square	df	P-value
Intercept Only	1927.515			
Final	642.030	1285.485	5	0.000

$$G = 1927.515 - 642.030$$
$$= 1285.485$$

Table 4.38 Test of parallel lines of M=2

Model	-2 Log Likelihood	Chi-Square	df	P-value
Null Hypothesis	642.030			
General	637.521	4.508	5	0.479

Table 4.38 gives results whether regression coefficients are same or not of the model obtained by data set filled with multiple imputation. Null hypothesis cannot be rejected as the p-value is bigger than α. This means, regression coefficients are same for all levels of dependent variable.

Table 4.39 shows results of M=3 data set obtained with multiple imputation. 10 cm increase in Haunch independent variable results in $e^{10*0.084} = 2.32$ unit increase in dependent variable.

Table 4.39 Multiple imputation M=3

		Coef (SE coef)	Z (p-value)	Wald	Odds ratio	OR CI Lower	OR CI Upper
Threshold	BMIgrp=1	35.069 (2.435)	14.40 (0.000)	207.461			
	BMIgrp=2	44.918 (2.875)	15.62 (0.000)	244.030			
Location	Haunch	0.084 (0.022)	3.82 (0.000)	14.111	1.088	1.041	1.136
	Relative Weight	0.309 (0.020)	1.55 (0.000)	247.797	1.362	1.310	1.414

Table 4.40 gives information on model fitting for M=2. Regarding p-value of G statistics distributed chi-square with 2 degree of freedom, it is said that the model is appropriate.

Table 4.40 Model fitting information of M=3

Model	-2 Log Likelihood	Chi-Square	df	P-value
Intercept Only	1901.694			
Final	607.925	1293.769	2	0.000

$$G = 1901.694 - 607.925$$
$$= 1293.769$$

Table 4.41 Test of parallel lines of M=3

Model	-2 Log Likelihood	Chi-Square	df	P-value
Null Hypothesis	607.925			
General	606.017	1.908	2	0.385

Table 4.41 gives results whether regression coefficients are same or not of the model obtained by data set filled with multiple imputation. Null hypothesis cannot be

rejected as the p-value is bigger than α. This means, regression coefficients are same for all levels of dependent variable.

Table 4.42 shows results of M=4 data set obtained with multiple imputation.

Table 4.42 Multiple imputation M=4

		Coef (SE coef)	Z (p-value)	Wald	Odds ratio	OR CI Lower	OR CI Upper
Threshold	BMIgrp=1	26.329 (1.548)	17.00 (0.000)	289.430			
	BMIgrp=2	35.656 (2.002)	17.81 (0.000)	317.321			
Location	Skin Folds	0.074 (0.026)	2.85 (0.004)	8.182	1.077	1.023	1.133
	Relative Weight	0.293 (0.017)	17.24 (0.000)	291.985	1.341	1.296	1.385
	Gender=1	0.645 (0.217)	2.97 (0.003)	8.835	1.906	1.246	2.915
	Gender=2	0	.	.	.	.	.

If the gender independent variable increases 1 unit then the dependent variable increases 1 unit. Boys in high school are more prone to be obese than girls in high school.

Table 4.43 gives information on model fitting for M=4. Regarding p-value of G statistics distributed chi-square with 3 degree of freedom, it is said that the model is appropriate.

Table 4.43 Model fitting information of M=4

Model	-2 Log Likelihood	Chi-Square	df	P-value
Intercept Only	1929.957			
Final	710.623	1219.334	3	0.000

$$G = 1929.957 - 710.623$$
$$= 1219.334$$

Table 4.44 Test of parallel lines of M=4

Model	-2 Log Likelihood	Chi-Square	df	P-value
Null Hypothesis	710.623			
General	708.942	1.680	3	0.641

Table 4.44 gives results whether regression coefficients are same or not of the model obtained by data set filled with multiple imputation. Null hypothesis cannot be rejected as the p-value is bigger than α. This means, regression coefficients are same for all levels of dependent variable.

Table 4.45 shows results of M=5 data set obtained with multiple imputation.

Table 4.45 Multiple imputation M=5

		Coef (SE coef)	Z (p-value)	Wald	Odds ratio	OR CI Lower	OR CI Upper
Threshold	BMIgrp=1	26.340 (1.735)	15.18 (0.000)	230.372			
	BMIgrp=2	36.113 (2.193)	16.47 (0.000)	271.267			
Location	Relative Weight	0.313 (0.019)	16.47 (0.000)	283.370	1.37	1.319	1.419
	Percentile=1	1.834 (0.482)	3.80 (0.000)	14.501	6.26	2.435	16.087
	Percentile=2	1.338 (0.446)	3.00 (0.003)	9.022	3.81	1.592	9.134
	Percentile=3	1.028 (0.429)	2.39	5.735	2.79	1.206	6.482
	Percentile=4	0	.	.	.	.	.

In Table 4.45, the odds ratio of "Relative Weight" variable is 1.37. Unit of continuous independent variables is very important for odds ratio commentary. %25 increase in "Relative Weight" results in $e^{0.25*0.313} = 1.081$ unit change in dependent variable. . Confidence interval of odds ratio not containing 1 shows that there is a relationship between it and dependent variable.

Table 4.46 gives information on model fitting for M=5. Regarding p-value of G statistics distributed chi-square with 4 degree of freedom, it is said that the model is appropriate.

Table 4.46 Model fitting information of M=5

Model	-2 Log Likelihood	Chi-Square	df	P-value
Intercept Only	1814.411			
Final	555.544	1258.867	4	0.000

$$G = 1814.411 - 555.544$$
$$= 1258.867$$

Table 4.47 Test of parallel lines of M=5

Model	-2 Log Likelihood	Chi-Square	df	P-value
Null Hypothesis	555.544			
General	552.909	2.635	4	0.621

Table 4.47 gives results whether regression coefficients are same or not of the model obtained by data set filled with multiple imputation. Null hypothesis cannot be rejected as the p-value is bigger than α. This means, regression coefficients are same for all levels of dependent variable.

4.4.5 Comparison of methods

Pearson chi-square and deviance statistics are shown Table 4.48. These statistics are intended to test whether the model is appropriate.

The model M=2 and the model M=5 is not appropriate. Those statistics are both sensitive to empty cells. When estimating models with continuous independent variables, there are often many empty cells, as in this application. Therefore, you

shouldn't rely on either of these test statistics with such models. Because of the empty cells, you can't be sure that these statistics will really follow the chi-square distribution, and the significance values won't be accurate.

Table 4.48 Compare of methods

	Goodness of fits		R_square		
	Pearson	Deviance	Cox and Snell	Nagelkerke	McFadden
Complete case	1097.608 (1.000)	423.582 (1.000)	0.714	0.863	0.713
EM	1212.079 (1.000)	425.958 (1.000)	0.677	0.866	0.742
REG	1255.615 (1.000)	378.256 (1.000)	0.712	0.861	0.709
M=1	2220.893 (0.165)	707.815 (1.000)	0.689	0.822	0.641
M=2	2240.071 (0.000)	628.846 (1.000)	0.693	0.833	0.662
M=3	1567.781 (1.000)	561.962 (1.000)	0.696	0.833	0.661
M=4	1252.754 (1.000)	702.069 (1.000)	0.674	0.810	0.629
M=5	1877.310 (0.000)	486.792 (1.000)	0.686	0.824	0.649

As can be seen also from Table 4.47, Cox and Snell R^2 determination coefficient is found to be 71.4% in complete-case situation, 67.7% in EM analysis and 71.2% in regression imputation analysis in method comparisons. When situations compared, the lowest determination coefficient is seen in EM algorithm. For all methods, the highest determination coefficient is the complete case analysis. The highest determination coefficient of Cox and Snell R^2 is 71.4% in complete case. Nagelkerke and Mc Fadden have the highest determination coefficient in EM.

CHAPTER FIVE
CONCLUSION

Missing data is a common problem in statistical studies. While ignoring missing data is an option, it is possible to contribute to study by analyzing them with various statistical methods. Before the applying the methods, we have to determine the missing data mechanism. Data set contains 15 independent variables. The continuous independent variables; Skin folds, Waist, Haunch and Relative weight have missing values. Discrete independent variable Percentile has missing value and dependent variable also has missing values. There are 241 missing observations in independent variables and 848 observations have seen in all variables. Missing values in variables are in same units. This missing data pattern fits to multivariate pattern. Missing data mechanism fits to MCAR mechanism.

The methods used here are complete case analysis, EM algorithm, regression imputation and multiple imputation. The complete case analysis also called listwise deletion, analyzed only the cases with complete data. Individuals with data missing on any variables were dropped from the analysis. EM Algorithm is a re-iteratively method that includes the maximum likelihood estimations to calculate the parameter predictions in in-complete data problems. The normal distribution was used as a likelihood function. In the regression imputation the first step of the imputation process is to estimate a set of regression equations that predict the in complete variables from the complete variables. While the regression imputation was applied, the linear regression model was used. A complete-case analysis usually generates these estimates. The second step is to generate predicted values for the incomplete variables. These predicted scores fill in the missing values and produce a complete data set. In the multiple imputation each missing value is replaced by M >1 values.

After that imputation operation logistic regression was applied. Dependent variable is BMIgrp which denotes the body mass index group. The groups are defined according to body mass indexes as; the ones smaller than 18.5 are the 1st group (low-weight), the ones between 18.5 and 23.9 are the 2nd group (normal

weight) and the ones bigger than 23.9 are the 3rd group (obese). The regression equations were found for the 8 data set. The regression models were compared. Pearson chi-square and deviance statistics are intended to test whether the model is appropriate. The model M=2 and the model M=5 is not appropriate although the regression coefficients are significant.

Cox and Snell R^2 determination coefficient is found to be %71.4 in complete-case situation, %67.7 in EM analysis and %71.2 in regression imputation analysis in method compares. When situations compared, the lowest determination coefficient are seen in EM algorithm. For all methods, the highest determination coefficient is complete case analysis. The highest determination coefficient of Cox and Snell R^2 is 71.4% in complete case. Nagelkerke and Mc Fadden have the highest determination coefficient in EM.

In this study, there is no difference between the methods of “missing value analysis”. However it does not mean that any difference will not appear. It is concluded that the obtained results can differ according to the characteristics of the study and the data, number of missing data/observation when the different methods are applied. Therefore it is understood that strict interpretations while judging results should be avoided especially for cases close to meaningfulness in researches that “imputation” technique is used and bear such situations.

REFERENCES

Agresti, A. (2010). *Analysis of ordinal categorical data* (2nd ed.). NewYork: Wiley.

Allison, P. (n.d), *Multiple imputation for missing data: A cautionary tale.* Retrieved March 15, 2012, from http://www.ssc.upenn.edu/~allison/MultInt99.pdf.

Chu, W., & Ghahramani, Z. (2005). Gaussian processes for ordinal regression. *Journal of Machine Learning Research*, 6, 1-48.

Çolak, E. (2002). *Koşullu ve sınırlandırılmış lojistik regresyon yöntemlerinin karşılaştırılması ve bir uygulama*, Osmangazi Üniversitesi, Sağlık Bilimleri Enstitüsü, Biyoistatistik Anabilim Dalı, Yüksek Lisans Tezi.

Dielman, T. (2001). *Applied regression analysis for business and economics* (3rd ed.). USA: Duxbury.

Enders, C. K. (2010). *Applied missing data analysis*. NewYork: Guilford Press.

Flom, P. (n.d.). *Multinomial and ordinal logistic regression using PROC LOGISTIC.* Retrieved July 04, 2011, from http://www.nesug.org/proceedings/nesug05/an/an2.pdf .

Fomby, T. (2005). Best subset methods in regression modeling. *SAS/STAT User's Guide, 6* (2)*,* 1397-1400.

John, N., & Kutner, M. H., & Nachtsheim, C. J., &Wasserman, W. (1996). *Applied linear statistical models* (4th ed.). NewYork: McGraw- Hill.

Komarek, P. (n.d). *Logistic regression for data mining and high-dimensional classification,* Department of Math Sciences Carnegie Mellon University.

Long, J. S., & J. Freese. (2006). *Regression models for categorical dependent variables using stata*. (2nd. ed.) College Station, Tex: Stata Corp LP.

Patzer, S. (2009). *Missing data analysis: A case study of a randomized controlled trail.* Retrieved (n.d.), from http://drum.lib.umd.edu/bitstream/1903/9366/1/Patzer_umd_0117N_10421.pdf.

Rockhill, A. P. (n.d.) *Imputation of missing observations in forest inventories.* Retrieved (n.d.) from http://repository.lib.ncsu.edu/ir/bitstream/1840.16/2887/1/etd.pdf.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys.* New York: John Wiley & Sons.

Ryan, T. P. (1997). *Modern regression methods*. Canada: John Wiley & Sons.

Schafer, J.L. and Graham, J.W. (2002). Missing data: Our view of the state of art. *Psychological Methods*, *7*(2), 147-177.

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing data problems: A data analyst's perspective. *Multivariate Behavioral Research*, *33*(4), 545-571.

Schoier, G. (n.d.). *On partial nonresponse situations: the hot deck imputation method.* (n.d.), http://www.stat.fi/isi99/proceedings/arkisto/varasto/scho0502.pdf.

The likelihood function, maximum likelihood estimator (MLE), logit & probit, count data regression models, (n.d.). Retrieved October 10, 2012, from http://econweb.rutgers.edu/tsurumi/liklihood.pdf.

Wayne, F. & Jae, K. (2005). Hot deck imputation for the response model. *Survey Methodology, 2* (31), 139-149.

APPENDIXES

		mother_edu valid/missing	father_edu valid/missing	fastfood valid/missing	carbonated valid/missing	health valid/missing	skinfolds valid/missing	percentile valid/missing	waist valid/missing	haunch valid/missing	RA valid/missing
mother_edu valid/missing	Corr. Coeff.	1.000									
	p-value	0.000									
	N	1089									
father_edu valid/missing	Corr. Coeff.	1.000	1.000								
	p-value	0.000									
	N	1089	1089								
fastfood valid/missing	Corr. Coeff.	1.000	1.000	1.000							
	p-value	0.000	0.000								
	N	1089	1089	1089							
carbonated valid/missing	Corr. Coeff.	1.000	1.000	1.000	1.000						
	p-value	0.000	0.000	0.000							
	N	1089	1089	1089	1089						
health valid/missing	Corr. Coeff.	1.000	1.000	1.000	1.000	1.000					
	p-value	0.000	0.000	0.000	0.000						
	N	1089	1089	1089	1089	1089					
skinfolds valid/missing	Corr. Coeff.	0.057	0.057	0.057	0.057	0.057	1.000				
	p-value	0.061	0.061	0.061	0.061	0.061					
	N	1089	1089	1089	1089	1089	1089				
percentile valid/missing	Corr. Coeff.	0.057	0.057	0.057	0.057	0.057	1.000	1.000			
	p-value	0.061	0.061	0.061	0.061	0.061	0.000				
	N	1089	1089	1089	1089	1089	1089	1089			
waist valid/missing	Corr. Coeff.	0.057	0.057	0.057	0.057	0.057	1.000	1.000	1.000		
	p-value	0.061	0.061	0.061	0.061	0.061	0.000	0.000			
	N	1089	1089	1089	1089	1089	1089	1089	1089		
haunch valid/missing	Corr. Coeff.	0.057	0.057	0.057	0.057	0.057	1.000	1.000	1.000	1.000	
	p-value	0.061	0.061	0.061	0.061	0.061	0.000	0.000	0.000		
	N	1089	1089	1089	1089	1089	1089	1089	1089	1089	
RA valid/missing	Corr. Coeff.	0.057	0.057	0.057	0.057	0.057	1.000	1.000	1.000	1.000	1.000
	p-value	0.061	0.061	0.061	0.061	0.061	0.000	0.000	0.000	0.000	
	N	1089	1089	1089	1089	1089	1089	1089	1089	1089	1089

LIST OF TABLES

Page

LIST OF FIGURES

Page