DOKUZ EYLÜL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

# THE PROBLEM OF MISSING DATA IN REGRESSION ANALYSIS

by

Neslihan DEMİREL

February, 2007

İZMİR

# THE PROBLEM OF MISSING DATA IN REGRESSION ANALYSIS

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

**by**
**Neslihan DEMİREL**

**February, 2007**
**İZMİR**

## Ph.D. THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"THE PROBLEM OF MISSING DATA IN REGRESSION ANALYSIS"** completed by **NESLİHAN DEMİREL** under supervision of **PROF. DR. SERDAR KURT** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy.

Prof. Dr. Serdar KURT

Supervisor

Prof. Dr. İsmihan BAYRAMOĞLU

Thesis Committee Member

Assoc. Prof. Dr. Halil ORUÇ

Thesis Committee Member

Prof. Dr. Gülay KIROĞLU

Examining Committee Member

Assoc. Prof. Dr. C. Cengiz ÇELİKOĞLU

Examining Committee Member

Prof. Dr. Cahit HELVACI
Director
Graduate School of Natural and Applied Sciences

# ACKNOWLEDGEMENTS

Above all, I would like to thank to my dissertation chair Prof. Dr. Serdar Kurt, who has been supporting my scientific career as my supervisor since 2000, in which I started my master's thesis with him. Not only has he been invaluable for the development of both my master's and my PhD thesis, but it has always been a great pleasure to work with him.  If it hadn't been for his true mentorship and academic guidance this dissertation would not have been written.

I am very thankful to members of my committee who generously contributed me; namely to Prof. Dr. İsmihan Bayramoğlu for the contributions and perspectives and to Assoc. Prof. Dr. Halil Oruç for his suggesting me many helpful revisions. Their effect certainly improved my perspective, and I hope that I have carried out their very helpful suggestions in this dissertation.

Special thanks to all my friends, especially my roommate Selma Gürler who intimately and promptly shared her experiences, Alper Vahaplar for his infinite patience and help throughout the work of my dissertation and Uğraş Erdoğan who helped me for preparing the C# code. I am very thankful to Ayşe Övgü Kınay for her encouragement and support. Finally I am deeply appreciative of the contributions to Şeyda Eraslan and Pelin Şulha.

I wish to utter my special appreciation to my parents, Zuhal and Nihat Ortabaş, who have unfailingly supported me through all my life and for taking care of my education. My sister, Nihan Özesen has provided constant encouragement and positive attitudes, which I will never forget for good. Lastly, I owe a debt a gratitude to my husband, Hakan Demirel who *lived up to his part of the bargain* to do whatever he could and more to help me throughout my dissertation.

Neslihan DEMİREL

# THE PROBLEM OF MISSING DATA IN REGRESSION ANALYSIS

## ABSTRACT

The subject of missing data analysis consists of a data matrix in which some of the values in the matrix are not observed. Missing data analysis is one of the most important topics in applied statistics. It destroys the randomness of the sample and causes serious bias in the parameter estimates.

The regression analysis is one of the most important procedures used for estimation in multivariate statistical analysis. For this reason, in this study, missing data mechanism designed by missing at random (MAR) for independent variable in regression analysis simulation study is performed for the data set. When missing data can be ignored, model based methods such as EM algorithm, multiple imputation method and protective estimator are compared. In this thesis, C# code is improved to calculate of the protective regression coefficients, standard error of regression coefficients and mean square error.

**Keywords** : Missing data, Regression analysis, EM algorithm, Multiple imputation, Protective estimator.

# REGRESYON ÇÖZÜMLEMESİNDE KAYIP VERİ SORUNU

## ÖZ

Kayıp veri çözümlemesinin konusu veri matrisindeki bazı değerlerin gözlenmemiş olmasıdır. Kayıp veri çözümlemesi özellikle uygulamalı istatistiğin çok önemli konularından birini oluşturmaktadır. Kayıp veriyi yok saymak, örneklemin rasgeleliğini bozarak yanlı parametre tahminleri elde edilmesine neden olabilmektedir.

Regresyon çözümlemesi, tahmin amaçlı kullanılan önemli çok değişkenli istatistiksel çözümlemelerin başında gelmektedir. Bu nedenle bu çalışmada, regresyon çözümlemesinde, bağımsız değişkende kayıp veri mekanizması rassal kayıp (MAR) olacak şekilde, veri seti üzerinde benzetim çalışması yapılmıştır. Kayıp veri göz ardı edilebilir olduğunda model esaslı yöntemler arasında yer alan, EM algoritması, çoklu atıf ve geliştirilen koruyucu kestirim yöntemleri karşılaştırmalı olarak incelenmiştir. Bu çalışmada, koruyucu kestirim katsayıları, regresyon katsayıların standart hataları ve hata kareler ortalamasını hesaplamak üzere C# kodu geliştirilmiştir.

**Anahtar sözcükler** : Kayıp veri, Regresyon çözümlemesi, EM algoritması, Çoklu atıf, Koruyucu kestirim

## CONTENTS

**Page**

# CHAPTER ONE
# INTRODUCTION AND LITERATURE REVIEW

Twenty four years ago Greenlees et al. (1982) wrote that “there is a large literature on the problem of parameter estimation, but with few exceptions this literature treats the case in which the missing values are missing at random”. Although substantial advances have been made, this statement continues to be valid. (Pastor, 2003)

In the last twenty years, many researchers have assessed the requirements of different methods for the analysis of incomplete data, showing that single imputation (unconditional or conditional mean, stochastic regression, hot deck, artificial neural networks, etc.), complete-case or listwise analysis, available-case or pairwise analysis, maximum likelihood (Expectation Maximisation (EM) algorithm, Structural Equation Modelling (SEM), Raw Maximum Likelihood (RML)) and multiple imputation (MI) methods require, for generalizable results, that the missing values be missing completely at random or at least missing at random (Little and Rubin, 1987; Little and Rubin, 1989; Navarro and Losilla, 2000; Rubin, 1987; Schafer, 1997; Simonoff, 1988). In the estimation of an explanatory linear regression model, many studies have shown that the best procedures (less biased and more efficient) for the treatment of incomplete data with missing values completely at random or missing at random are maximum likelihood estimation and multiple imputation (Gold and Bentler, 2000; Graham et al., 1994; Graham et al., 1996; Othuon, 1999). Graham et al. (1996) showed the superiority of Maximum Likelihood and Multiple Imputation in the analysis of incomplete data with nonrandom missing values obtained with planned missing value patterns. Graham et al. (1997) and Wothke (1998) also suggest the use of these techniques even when the missing values are not at random, since they produce less biased results than other traditional approaches. Kromrey and Hines (1994) investigated the effects of nonrandom missing data in one of the two variables acting as predictors in a linear regression model. Hippel (2004) investigated biases in SPSS 12.0 missing value analysis when normally distributed values are missing at random.

The study of missing data is one of the most important topics in applied statistics, especially in survey problem and medical and biological data. Standard statistical methods are designed for rectangular data sets.

$$\begin{array}{ll} & \text{Variables} \\ \text{Cases} \quad Y_{ij} = & \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1k} \\ y_{21} & y_{22} & \cdots & y_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ y_{i1} & y_{i2} & \cdots & y_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nk} \end{bmatrix} \end{array}$$

k variables measured for each of n units (cases, observations, subjects).
$i = 1,2,\ldots,n \quad j = 1,2,\ldots,k$

The subject of this analysis is such a data matrix when some of the values in the matrix are not observed. Standard methods are not directly applicable if there is missing data (nonresponse), i.e. some $y_{ij}$ values in the matrix are not observed. Complete case analysis treat missingness by omitting cases with any variables missing. This is occasionally appropriate, but more often leads to inefficiency and biased estimation. The aim is clarify the limitations of complete case analysis and to suggest improved methods of analysis which take missingness into account.

## 1.1 Sources of Missing Data

Two main sources of missing data can be distinguished.

*Item nonresponse* (some but not all variables missing for a case): refusal, don't know, interviewer error, equipment failure, response out of range and edited out.

*Unit nonresponse* (all variables missing): refusal, not at home, not contacted.

Sometimes missingness may be deliberate, i.e. under the control of the researcher. Example is double sampling where some variables are measured for all cases in the sample but some only for a smaller subsample. Typically this is done to reduce costs.

Sampling itself leads to missingness in the sense that variables are not recorded for units not sampled. However, this is under the control of the sampler and is not normally thought of as missingness.

Assume that missingness hides a well-defined, meaningful true value e.g., `Don't know' to a question about income is missingness. `Don't know' to a question on political views may be missingness (refusal) but may also indicate lack of opinion; unclear whether to treat as missing value. Some case-variable combinations are never observed because they are not applicable; e.g., prostate cancer incidence for women, length of current employment for unemployed.

## 1.2 Missing Data Pattern

### *1.2.1 Univariate Missingness*

The missingness is confined to a single variable. (?: missing)

$$\begin{bmatrix} y_{11} & y_{12} & \cdots & ? \\ y_{21} & y_{22} & \cdots & ? \\ \vdots & \vdots & \ddots & \vdots \\ y_{i1} & y_{i2} & \cdots & ? \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & ? \end{bmatrix}$$

### *1.2.2 Unit Nonresponse*

All variables missing for some cases but we may have background variables.

$$\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1k} \\ y_{21} & y_{22} & \cdots & y_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ y_{i1} & y_{i2} & ??? & ? \\ \vdots & \vdots & ??? & ? \\ y_{n1} & y_{n2} & ??? & ? \end{bmatrix}$$

### *1.2.3 Monotone Missing Data*

Longitudinal studies collect information on a set of cases repeatedly over time. The subject are drop out prior to the end of the study and do not return.

$$\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1k} \\ y_{21} & y_{22} & \cdots & y_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ y_{i1} & y_{i2} & \cdots & ? \\ \vdots & \vdots & ??? & ? \\ y_{n1} & y_{n2} & ??? & ? \end{bmatrix}$$

## 1.3 Missing Data Mechanisms

A different issue concerns the mechanisms that lead to missing data is related to the underlying values of the variables in the data set. Missing-data mechanisms are crucial since the properties of missing data methods depend very strongly on the nature of the dependencies in these mechanisms. The crucial role of the mechanism in the analysis of data with missing values was largely ignored until the concept was formalized in the theory of Rubin (1976), through the simple device of treating the missing data indicators as random variables and assigning them a distribution.

Let $Y = (y_{ij})$ denote an (*n x k*) rectangular data set without missing values, with *i*th row $y_i = (y_{i1}, \ldots, y_{ik})$ where $(y_{ij})$ is the value of variable $Y_j$ for subject *i*. With missing data, define the missing data indicator matrix $R = (r_{ij})$, such that $r_{ij} = 1$ if $y_{ij}$ present and $r_{ij} = 0$ if $y_{ij}$ is missing. The matrix *R* then defines the pattern of missing data. The missing data mechanism is characterized by the conditional distribution of *R* given *Y*, say $f(R \mid Y, \phi)$ where $\phi$ denotes unknown parameters. If missingness does not depend on the values of the data $Y$, missing or observed, that is, if

$$f(R \mid Y, \phi) = f(R \mid \phi) \text{ for all } Y \text{ and } \phi,$$

then the data are called missing completely at random (MCAR). This assumption does not mean that the pattern itself is random, but rather that missingness does not depend on the data values.

Let $Y_{obs}$ denote the observed components or entries of $Y$ and $Y_{mis}$ the missing components. An assumption less restrictive than MCAR is that missingness depends only on the components $Y_{obs}$ of $Y$ that are observed, and not on the components that are missing. That is,

$$f(R \mid Y, \phi) = f(R \mid Y_{obs}, \phi) \text{ for all } Y_{mis} \text{ and } \phi\textbf{.}$$

In this case, the missing data mechanism is than called missing at random (MAR).

The mechanism is called not missing at random (NMAR) if the distribution of $R$ depends on the missing values in the data matrix $Y$.

Perhaps the simplest data structure is a univariate random sample for which some units are missing. Let $Y = (y_1, \ldots, y_n)^T$ where $y_i$ denotes the value of a random variable for unit $i$, and let $R = (R_1, \ldots, R_n)^T$ where $R_i = 1$ for units that are observed and $R_i = 0$ for units that are missing. Suppose the joint distribution of $(y_i, R_i)$ is independent across units, so in particular the probability that a unit is observed does not depend on the values of $Y$ or $R$ for other units. Then,

$$f(Y, R \mid \theta, \phi) = f(Y \mid \theta) f(R \mid Y, \phi) = \prod_{i=1}^{n} f(y_i \mid \theta) \prod_{i=1}^{n} f(R_i \mid y_i, \phi)$$

where $f(y_i \mid \theta)$ denotes the density of $y_i$ indexed by unknown parameters $\theta$, and $f(R_i \mid y_i, \phi)$ is the density of a Bernoulli distribution for the binary indicator $R_i$ with the probability $\Pr(R_i = 0 \mid y_i, \phi)$ that $y_i$ is missing. If missingness is independent of $Y$, that is if $f(R_i = 0 \mid y_i, \phi) = \phi$, a constant that does not depend on $y_i$, then the missing data mechanism is MCAR (or in this case equivalently MAR). If the mechanism depends on $y_i$, then the mechanism is NMAR since it depends on $y_i$ that are missing.

### *1.3.1 Missing Completely at Rrandom*

Data elements are missing for reasons that are unrelated to any chracateristics or responses for the subject, including the value of the missing data, where it to be known. Examples, include missing laboratory measurements because of a dropped test tube (if it was not dropped because of knowledge of any measurements) and a survey in which a subject omitted her response to a question for reasons unrelated to the response she would have made or to any other of her chracteristics.

### *1.3.2 Missing at Random*

Data elements are not missing at random, but the probability that a value is missing depends on values of variables that were actually measured. As an example, consider a survey in which females are less likely to provide their personal income in general (but the likelihood of responding is independent of her actual income). If we know the sex of every subject and have income levels for some of the females, unbiased sex-specific income estimates can be made. That is because the incomes we do have for some of the females are a random sample of all females incomes.

### *1.3.3 Not Missing at Random*

Elements are more likely to be missing if their true values of the variable in question are systematically higher or lower. In an interview this situation can be given as an example of not missing at random mechanism when subjects with lower income levels or very high incomes are less likely to provide their personal income.

These distinctions of mechanisms are important, because when missing data mechanism is MCAR unbiased estimates will be produced even with rather primitive analysis methods. When missing data mechanism is MAR, unbiased estimates will be produced if a model and estimation technique is used that renders the missingness mechanism ignorable. When missing data mechanism is NMAR, an analysis method must be used that includes both a model for the observed data, and a model for the missingness mechanism. For missing data that are MCAR or MAR, general modeling software is available, that produces unbiased using all the available information. For missing data that are NMAR, there are no easy solutions.

Often, however, it is impossible to eliminate completely missing data. Then we need to use missing data estimation methods which base estimation on the observed (non-rectangular) data only. Rests of the chapters are about such methods.

## 1.4 Thesis Outline

This thesis consists of six chapters that investigate missing data estimation methods. We first present some important subjects of missing data such as introduction to missing data, sources of missing data, missing data pattern and missing data mechanisms. Chapter 2 presents missing data methods. Because of the missing data mechanism is assumed as MAR, Expectation Maximization (EM) algorithm and Multiple Imputation (MI) methods which are the model-based methods will be examined. In Chapter 3, we present a simulation study to compare MI and EM algorithm. In Chapter 4, the Protective Estimator (PE) is proposed when the missing data mechanism is MAR for the linear regression parameters. Chapter 5 presents simulation study to compare the estimates obtained using the complete cases (CC), EM algorithm (EM), proposed Protective Estimator (PE) and Multiple Imputation (MI) for various imputations. Finally in Chapter 6 conclusions of this thesis will be given.

# CHAPTER TWO
# MISSING DATA METHODS

The literature on the analysis of partially missing data is comparatively recent. Review papers include Afifi and Elashoff (1966), Hartley and Hocking (1971), Orchard and Woodbury (1972), Dempster, Laird and Rubin (1977), Little and Rubin (1983), Little and Schenker (1994), and Little (1997). Methods proposed in this literature can be usefully grouped into the following categories:

## 2.1 Methods Based on Completely Recorded Units

When some variables are not recorded for some of the units, it can be done to discard incompletely recorded units and to analyze only with complete data. This is generally easy to carry out and may be satisfactory with small amounts of missing data. But it can lead to serious biases, however, and it is not usually very efficient.

## 2.2 Weighting Methods

To give weights to observed cases, so that they represent not only themselves but also `similar' missing cases. Randomization inferences from sample survey data without nonresponse commonly weight sampled units by their design weights, which are inversely proportional to their probabilities of selection. For example, let $y_i$ be the value of a variable $Y$ for unit *i* on the population. Then the population mean is often estimated by Horvitz-Thompson estimator: $\left(\sum_{i=1}^{n}\pi_i^{-1}y_i\right)\left(\sum_{i=1}^{n}\pi_i^{-1}\right)^{-1}$ , where the sums over sampled units, and $\pi_i$ is known probability of inclusion in the sample for unit *i*. Weighting procedures for nonresponse modify the weights in an attempt to adjust for nonresponse as if it were part of the sample design. The resultant estimator is replaced by $\sum_{i=1}^{n}(\pi_i\hat{p}_i)^{-1}y_i \Big/ \sum_{i=1}^{n}(\pi_i\hat{p}_i)^{-1}$ where the sums are now over sampled units that respond, and $\hat{p}_i$ is an estimate of the probability of response for unit *i*, usually the proportion of responding units in a subclass of the sample. (Little, R. J. A., & Rubin, D. B. (2002)).

## 2.3 Imputation-Based Methods

`Impute' (fill in) values for the missing cases to create a rectangular data set and use it for analysis. Need to be careful with the choice of the imputation model. Commonly used procedures for imputation include; *Hot deck imputation*, which involves substituting individual values drawn from “similar” responding units. Hot deck imputation is common in survey practice and can involve very elaborate schemes for selecting units that are similar for imputation. *Mean imputation*, where means from the responding units in the sample are substituted. The means may be formed within cells or classes analogous to the weighting classes. Mean imputation then leads to estimates similar to those found by weighting, provided the sampling weights are constant within weighting classes. *Regression imputation* replaces missing 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. Mean imputation can be regarded as a special case of regression imputation where the predictor variables are dummy indicator variables for the cell within which the means are imputed. Multiple Imputation is a subject of model-based methods.

## 2.4 Model-Based Methods

In Maximum Likelihood estimation of the observed data (nonrectangular), likelihood is based on statistical models for the complete data and the nonresponse. Theoretically it is the most satisfying approach, because it is based on, and can rely on, general likelihood-theory methods and results. Disadvantage is computational complexity in some cases. Dependence on model assumptions may also be regarded as a disadvantage; however, note that other missing data methods also make assumptions, even though they may be implicit rather then explicit as in model-based methods.

In this thesis, missing data mechanism will be assumed as MAR. That is the missing data mechanism does not depend on the set of missing values though it may possibly depend on the set of observed values. Then the missing data mechanism is said to be ignorable (Little and Rubin, 1987). Little (1992) suggests that model-based methods, such as Maximum Likelihood (ML), Bayesian methods and Multiple Imputation are best among the current methods for dealing with missing values. For that reason MI and EM algorithm will be examined in this study.

### *2.4.1 EM Algorithm*

General missing data patterns can be handled by a method called the EM algorithm. (Dempster, Laird, & Rubin, 1977). EM algorithm is a very general iterative algorithm. It is called EM because each iteration of the EM algorithm consists two steps: an expectation (E) and a maximization (M) step. These two steps are repeated as:

1. Replace missing values by estimated values.
2. Estimate parameters.
3. Re-estimate the missing values assuming the new parameter estimates are correct.
4. Re-estimate parameters.

and so forth, iterating until convergence.

Many multivariate statistical analysis, including multiple linear regression are based on the initial summary of the data matrix into the sample mean and covariance matrix of the variables. Thus the efficient estimation of these quantities for an arbitrary pattern of missing values is a particularly important problem. ML estimation of the mean and covariance matrix from an incomplete multivariate normal sample is discussed, assuming the missing data mechanism is ignorable. Although the assumption of multivariate normality may appear restrictive, the methods can provide consistent estimates under weaker assumptions about the underlying distribution. Furthermore, the normality will be relaxed in linear regression.

Suppose that ($Y_1$,$Y_2$,…,$Y_k$) have a $k$-variate normal distribution with mean $\mu = (\mu_1, \mu_2, ..., \mu_k)$ and covariance matrix $\sum = (\sigma_{jl})$. $Y$=($Y_{obs}$, $Y_{mis}$), where $Y$ represents a random sample of size $n$ on ($Y_1$,$Y_2$,…,$Y_k$), $Y_{obs}$ the set of observed values, and $Y_{mis}$ the missing data. It follows that,

$$Y_{obs} = (y_{obs,1}, y_{obs,2}, ..., y_{obs,n})$$

where $y_{obs,i}$ represents the set of variables observed for observation $i$, $i$=1,2,…,$n$. The loglikelihood based on observed data is then:

$$L(\mu, \Sigma \mid Y_{obs}) = const - \frac{1}{2}\sum_{i=1}^{n} \ln\left|\Sigma_{obs,i}\right| - \frac{1}{2}\sum_{i=1}^{n}(y_{obs,i} - \mu_{obs,i})^T \Sigma_{obs,i}^{-1}(y_{obs,i} - \mu_{obs,i}) \quad (2.1)$$

where $\mu_{obs,i}$ and $\Sigma_{obs,i}$ are the mean and covariance matrix of the observed components of $Y$ for observation $i$.

To derive the EM algorithm for maximizing Equation (2.1), note that the hypothetical complete data $Y$ belong to the regular exponential family with sufficient statistics,

$$S = \left( \sum_{i=1}^{n} y_{ij} \quad j = 1,2,...,k; \quad \sum_{i=1}^{n} y_{ij} y_{il} \quad j,l = 1,2,...,k \right).$$

At the $t^{th}$ iteration of EM, let $\theta^{(t)} = (\mu^{(t)}, \Sigma^{(t)})$ denote the current estimates of the parameters. The E step of the algorithm consists in calculating,

$$E(\sum_{i=1}^{n} y_{ij} \mid Y_{obs}, \theta^{(t)}) = \sum_{i=1}^{n} y_{ij}^{(t)} \quad j = 1,...,k$$

and

$$E(\sum_{i=1}^{n} y_{ij} y_{il} \mid Y_{obs}, \theta^{(t)}) = \sum_{i=1}^{n} (y_{ij}^{(t)} y_{il}^{(t)} + c_{jli}^{(t)}) \quad j,l = 1,...,k \tag{2.2}$$

where

$$y_{ij}^{(t)} = \begin{cases} y_{ij} & ,if\ y_{ij} \text{ is observed;} \\ E(y_{ij} \mid y_{obs,i}, \theta^{(t)}) & ,if\ y_{ij} \text{ is missing;} \end{cases} \tag{2.3}$$

and

$$c_{jli}^{(t)} = \begin{cases} 0 & ,if\ y_{ij} \text{ or } y_{il} \text{ are observed;} \\ \text{cov}(y_{ij}, y_{il} \mid y_{obs,i}, \theta^{(t)}) & ,if\, y_{ij} \text{ and } y_{il} \text{ are missing;} \end{cases} \tag{2.4}$$

Missing values $y_{ij}$ are thus replaced by the conditional mean of $y_{ij}$ given the set of values, $y_{obs,i}$ observed for that observation. These conditional means and the nonzero conditional covariances are easily found from the current parameter estimates by sweeping the augmented covariance matrix so that the variables $y_{obs,i}$ are predictors in the regression equation and the remaining variables are outcome variables.

The M step of the EM algorithm is straightforward. The new estimates $\theta^{(t+1)}$ of the parameters will be estimated. (Little, R. J. A., & Rubin, D. B. (2002)). That is,

$$\mu_j^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} y_{ij}^{(t)} \quad j = 1,...,k; \tag{2.5}$$

$$\begin{aligned} \sigma_{jl}^{(t+1)} &= \frac{1}{n} E(\sum_{i=1}^{n} y_{ij} y_{il} \mid Y_{obs}) - \mu_j^{(t+1)} \mu_l^{(t+1)} \\ &= \frac{1}{n} \sum_{i=1}^{n} [(y_{ij}^{(t)} - \mu_j^{(t+1)})(y_{il}^{(t)} - \mu_l^{(t+1)}) + c_{jli}^{(t)}] \quad j,l = 1,...,k \end{aligned}$$

### *2.4.2 Multiple Imputation*

In the EM algorithm the missing values are “imputed” in the E-step and complete data methods are applied on the M-step. Thus the EM algorithm besides providing MLEs of parameters also provides estimates for the missing values. Although ML represents a major advance over conventional approaches to missing data, it has its limitations. ML theory and software are readily available for linear models and log-linear models, but beyond that, either theory or software is generally lacking. Although these imputed values may be good for the limited purpose of point estimation, using them for other purposes like testing hypothesis may not be suitable. The method of Multiple Imputation (MI) is a solution to this problem. (McLachlan, G.J., Krishnan, T. (1997)). It has the same optimal properties as ML, but removes some of these limitations. More specifically, MI, when used correctly, produces estimates that are consistent, asymptotically efficient, and asymptotically normal when data are MAR. Unlike ML, MI can be used with virtually for any kind of data and any kind of model, and the analysis can be done with modified conventional software. Of course MI has its own drawbacks. It can be cumbersome to implement and it is easy to do it the wrong way. Both of these problems can be substantially alleviated by using good software to do the imputations. A more fundamental drawback is that MI produces different estimates (hopefully, only slightly different) every time you use it. That can lead to awkward situations in which different researchers get different numbers from the same data using the same methods. (Allison, 2002)

Instead of imputing a single value for each missing value, MI is a technique designed to handle missing data, which fills in the missing values several times, and then creating several completed data sets for analysis. Each data set is analyzed separately using techniques designed for complete data, and the results are then combined in such a way that the variability due to imputation may be incorporated. In the notation of Rubin, let $Y_{obs}$ be the set of observed values and $Y_{mis}$ be the set of missing values. Then the posterior density of a population quantity $Q$ can be written as

$$h(Q \mid Y_{obs}) = \int g(Q \mid Y_{obs}, Y_{mis}) f(Y_{mis} \mid Y_{obs}) dY_{mis} \tag{2.6}$$

where $f(.)$ is the posterior density of the missing values and $g(.)$ is the complete data posterior density of $\theta$. Therefore, multiple imputations are simulated draws from the posterior distribution of the missing data.

The values of complete data statistics $\hat{Q}$ and $U$ calculated on the $s$ completed data sets are $\hat{Q}_1,...,\hat{Q}_s$ and $U_1,...,U_s$. The repeated-imputation estimate is

$$\overline{Q}_s = \frac{1}{s}\sum_{l=1}^{s}\hat{Q}_l \tag{2.7}$$

and the associated variance-covariance of $\overline{Q}_s$ is

$$T_s = \overline{U}_s + \frac{s+1}{s}B_s \tag{2.8}$$

where

$$\overline{U}_s = \frac{1}{s}\sum_{l=1}^{s}U_l \text{ within-imputation variability} \tag{2.9}$$

and

$$B_s = \frac{1}{s-1}\left(\sum_{l=1}^{s}\hat{Q}_l - \overline{Q}_s)(\hat{Q}_l - \overline{Q}_s)^T\right)\text{between-imputation variability.} \tag{2.10}$$

The large $s$ repeated-imputation inference treats $(Q-\overline{Q}_s)$ as a normal distribution with variance-covariance matrix $T_s$. Letting $s=\infty$, we have

$$(Q-\overline{Q}_\infty) \sim N(0,T_\infty) \tag{2.11}$$

where $T_\infty = \overline{U}_\infty + B_\infty$. (Atkinson & Cheng, (2000)). (2.12)

# CHAPTER 3
# A SIMULATION STUDY COMPARING EM and MI

Atkinson and Cheng (2000) have a simulation study to compare EM algorithm and Multiple Imputation. In their study, $X$ matrix generated from the multivariate normal distribution with dimension p=4 and 10%, 20%, 30% and 40% of the element of $X$ matrix be randomly missing with sample sizes n=100 and n=200. Additional to Atkinson and Cheng (2000), Demirel and Kurt (2005) carried out to verify the characteristics of the EM algorithm and MI when the assumption is not valid. In this study, $X$ matrix generated from the multivariate normal distribution $MN(O, I_p)$. All parameters of regression coefficients are assigned to 1, and $\varepsilon_i \sim N(0,1)$. Once the data are generated, let 12%, 24% and 36% of the elements of the $X$ matrix be randomly missing. Two kinds of data are generated: symmetric and skewed with sample size n=100 and dimension p=4. The statistical criteria to compare the methods are the regression coefficients and Mean Square Error (MSE) of regression model. For these purposes the following steps were followed:

1. Symmetric population is generated.
2. A sample of size n=100 is selected from the population.
3. 12% of $X$ matrix be randomly missing.
4. Apply 2, 5 and 10 repeat imputations and EM algorithm to sampled data.
5. The regression coefficients and MSE are computed.
6. Step 2,3,4 and 5 are repeated for n=300.
7. Step 2,3,4,5 and 6 are repeated for missing proportions 24% and 36%.
8. Step 2,3,4,5,6 and 7 are repeated for skewed population.

The data are generated and the elements of the $X$ matrix be randomly missing with Minitab package program. Multiple imputations are applied by using SOLAS and EM algorithms are applied by using SPSS. After these methods the missing values are estimated so the full data are analyzed, the regression coefficients and MSE are recorded. The results are summarized in Table 3.1.

Table 3.1 The mean of regression coefficients, standard errors of regression coefficients and MSE of the model for symmetric data with n=300 repeats.

| Prop. of Missing Values % | Methods | $E(\hat{\beta}_0)$<br>( $S_{\hat{\beta}_0}$ ) | $E(\hat{\beta}_1)$<br>( $S_{\hat{\beta}_1}$ ) | $E(\hat{\beta}_2)$<br>( $S_{\hat{\beta}_2}$ ) | $E(\hat{\beta}_3)$<br>( $S_{\hat{\beta}_3}$ ) | Standard Error of MSE |
|---|---|---|---|---|---|---|
| 12% | MI(2) | 1.00234<br>(0.11360) | 0.89951<br>(0.13070) | 0.89484<br>(0.13485) | 0.89360<br>(0.11900) | 0.2679 |
| | MI(5) | 1.00614<br>(0.12378) | 0.90321<br>(0.14728) | 0.88632<br>(0.14077) | 0.88617<br>(0.14478) | 0.2636 |
| | MI(10) | 1.00861<br>(0.11324) | 0.88136<br>(0.14065) | 0.88074<br>(0.14200) | 0.89486<br>(0.13295) | 0.3066 |
| | EM | 0.99596<br>(0.09739) | 0.97504<br>(0.11040) | 0.97468<br>(0.10519) | 0.97649<br>(0.10353) | 0.2036 |
| | | | | | | |
| 24% | MI(2) | 1.01883<br>(0.13665) | 0.77319<br>(0.16179) | 0.75485<br>(0.15405) | 0.75536<br>(0.17101) | 0.3489 |
| | MI(5) | 1.01902<br>(0.13837) | 0.77400<br>(0.16151) | 0.74761<br>(0.16345) | 0.78096<br>(0.15766) | 0.3433 |
| | MI(10) | 1.01577<br>(0.14104) | 0.75798<br>(0.15866) | 0.74450<br>(0.17125) | 0.76661<br>(0.16511) | 0.3591 |
| | EM | 1.01670<br>(0.12940) | 0.76839<br>(0.25075) | 0.75050<br>(0.26198) | 0.76876<br>(0.25945) | 0.6803 |
| | | | | | | |
| 36% | MI(2) | 1.02200<br>(0.14122) | 0.65712<br>(0.17048) | 0.64930<br>(0.16580) | 0.66161<br>(0.17325) | 0.3673 |
| | MI(5) | 1.02815<br>(0.14659) | 0.65937<br>(0.16465) | 0.64781<br>(0.16662) | 0.65456<br>(0.18045) | 0.4222 |
| | MI(10) | 1.03076<br>(0.14862) | 0.65609<br>(0.17962) | 0.64791<br>(0.18280) | 0.64215<br>(0.17473) | 0.4154 |
| | EM | 1.01221<br>(0.13198) | 0.76841<br>(0.14844) | 0.76253<br>(0.14292) | 0.77451<br>(0.15264) | 0.3391 |

The population regression coefficients are 1 so when the Table 3.1 is examined, EM algorithm is given the minimum MSE and mean of the $\hat{\beta}_j$ are close to 1 when the missing proportion 12% and 36%. MI(5) is given the minimum MSE when missing proportion 24%. Atkinson and Cheng (2000) found that MI values are closer to 1 than EM algorithm. In their studies, the imputations repeated 5 and 10 times in MI have better results than do only two imputations. In our study it is obtained that mean of $\hat{\beta}_0$ values are bigger than 1, the mean of $\hat{\beta}_1, \hat{\beta}_2$ and $\hat{\beta}_3$ values are smaller than 1. The results of multiple imputations methods are similar but the 5 repeat multiple imputations is the best in this study.

Table 3.2 The mean of regression coefficients, standard errors of regression coefficients and MSE of the mode for skewed data with n=300 repeats.

| Prop. of Missing Values % | Methods | $E(\hat{\beta}_0)$<br>$(S_{\hat{\beta}_0})$ | $E(\hat{\beta}_1)$<br>$(S_{\hat{\beta}_1})$ | $E(\hat{\beta}_2)$<br>$(S_{\hat{\beta}_2})$ | $E(\hat{\beta}_3)$<br>$(S_{\hat{\beta}_3})$ | Standard Error of MSE |
|---|---|---|---|---|---|---|
| 12% | MI(2) | 0.98865<br>(0.11860) | 0.92644<br>(0.14967) | 0.88079<br>(0.14408) | 0.85505<br>(0.14429) | 0.3788 |
| | MI(5) | 0.98816<br>(0.11946) | 0.92963<br>(0.15634) | 0.87729<br>(0.14150) | 0.84995<br>(0.14253) | 0.3843 |
| | MI(10) | 0.98227<br>(0.11797) | 0.92023<br>(0.15610) | 0.88627<br>(0.13554) | 0.85671<br>(0.14329) | 0.3734 |
| | EM | 0.99248<br>(0.11086) | 0.96604<br>(0.13602) | 0.90449<br>(0.13278) | 0.88561<br>(0.12704) | 0.3499 |
| | | | | | | |
| 24% | MI(2) | 0.99684<br>(0.14136) | 0.79486<br>(0.16362) | 0.76363<br>(0.15342) | 0.75197<br>(0.16239) | 0.4417 |
| | MI(5) | 0.99279<br>(0.13227) | 0.80806<br>(0.17032) | 0.75760<br>(0.16906) | 0.75450<br>(0.18137) | 0.4479 |
| | MI(10) | 1.00350<br>(0.12876) | 0.81109<br>(0.18312) | 0.76535<br>(0.16704) | 0.74960<br>(0.16559) | 0.4391 |
| | EM | 0.995387<br>(0.14343) | 0.77138<br>(0.34559) | 0.73023<br>(0.32510) | 0.729011<br>(0.30076) | 0.8214 |
| | | | | | | |
| 36% | MI(2) | 0.98886<br>(0.14097) | 0.67761<br>(0.19223) | 0.66218<br>(0.19182) | 0.64490<br>(0.18511) | 0.5248 |
| | MI(5) | 0.98766<br>(0.15672) | 0.68323<br>(0.20389) | 0.66941<br>(0.18420) | 0.64564<br>(0.19014) | 0.4848 |
| | MI(10) | 0.99866<br>(0.14883) | 0.66894<br>(0.20154) | 0.65555<br>(0.20469) | 0.63507<br>(0.19116) | 0.5589 |
| | EM | 0.98744<br>(0.14389) | 0.80658<br>(0.15471) | 0.78319<br>(0.14598) | 0.76185<br>(0.15699) | 0.4572 |

The population regression coefficients are 1 so when the Table 3.2 is examined, EM algorithm is given the minimum MSE and mean of the $\hat{\beta}_j$ are close to 1 when the missing proportion 12% and 36%. MI(10) is given the minimum MSE when missing proportion 24%. It is obtained that mean of $\hat{\beta}_j$ values are smaller than 1 for skewed data. The results of multiple imputations methods are similar but the 10 repeat multiple imputations is the best.

As a result, the statistical criteria to compare the methods are the expected values of regression coefficients values close to 1, standard error of regression coefficients and MSE are given the minimum. EM algorithm is given the minimum mean square error and mean of the $\hat{\beta}_i$ are close to 1 when the missing proportion 12% and 36% for symmetric and skewed data. MI(5) is given the minimum MSE when missing proportion 24% for symmetric data and for skewed data MI(10) is given. Consequently, when the assumption is not valid, EM algorithm is not affected, but imputations should be increased for Multiple Imputation.

# CHAPTER FOUR
# PROTECTIVE ESTIMATOR

## 4.1 Introduction

Lipsitz, S.R., Molenberghs, G., Fitzmaurice, G.M. and Ibrahim, J.G. (2004) propose a method for estimating the regression parameters in a linear regression model for Gaussian data when the outcome variable is missing for some subjects and missingness is thought to be nonignorable. That missingness is restricted to the outcome variable and that the independent variables are fully observed. Although maximum likelihood estimation of the regression parameters is possible once joint models for outcome variable and the nonignorable missing data mechanism have been specified, these models are fundamentally nonidentifiable unless unverifiable modeling assumptions are imposed. In their study rather than explicitly modeling the nonignorable missingness mechanism, they consider the use of a “protective” estimator of the regression parameters. To implement the proposed method, it is necessary to assume that the outcome variable and one of the independent variables have an approximate bivariate normal distribution, conditional on the remaining independent variables. In addition, it is assumed that the missing data mechanism is conditionally independent of this independent variable, given the outcome variable and the remaining independent variables; the latter is referred to as the “protective” assumption. A method of moments approach is used to obtain the protective estimator of the regression parameters; the jackknife method is used to estimate the variance.

In this study, the protective estimator is proposed when the missing data mechanism is MAR. To implement the proposed method, it is necessary to assume that the outcome variable and one of the independent variable have approximate bivariate normal distributions, conditional on the remaining independent variable. That missing data is restricted to the independent variable and that the outcome variable and the remaining independent variable are fully observed. A method of moments approach is used to obtain the protective estimator of the regression parameters and the variance.

### 4.2 Notation and Maximum Likelihood

Consider a linear regression model with $n$ independent subjects, $i = 1,2,...,n$. Let $Y_i$ denote the outcome variable for the $i$th subject and let $X_{ij}$ $i = 1,2,...,n$, $j = 1,2,..., p$ denote a $nxp$ matrix of independent variables.

$$X_{ij} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \vdots & \vdots & & \ddots & & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & \ddots & & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix}$$

The primary interest is the estimation of the vector of regression coefficients $\beta$ for the linear regression model

$$\mu = E[Y] = X\beta \tag{4.1}$$

Note that maximum likelihood estimation of $\beta$ (and $\sigma^2$) requires specification of the conditional distribution of $y_i$ given $x_i$. It is assumed that $y_i$ given $x_i$ is normal

$$f(y_i \mid x_i, \beta, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_i - \mu_i}{\sigma}\right)^2} \tag{4.2}$$

where $\sigma^2 = Var[Y_i \mid x_i]$ and $\mu_i = \mu_i(\beta)$ is given by Equation 4.1. However, since $X_i$ can be missing, also define the indicator random variable $R_i$, which equals 1 if $X_i$ is observed and 0 if $X_i$ is missing. With missing data mechanism is MAR, propose using the joint distribution $(y_i, r_i \mid x_i)$ to estimate $\beta$, that is,

$$f(r_i, y_i \mid x_i, \alpha, \beta, \sigma^2) = f(y_i \mid x_i, \beta, \sigma^2) f(r_i \mid x_i, y_i, \alpha)$$

where $\alpha$ is the parameter vector of the 'missing data mechanism' $f(r_i \mid x_i, y_i, \alpha)$.

### 4.3 Protective Estimator

To develop the protective estimator we must assume that one of the independent variables, say $x_{i1}$, has a normal distribution. In particular, we partition $x_i$ into $x_i' = [x_{i1}, x_{i2}]$, and assume that $f(\ y_i, x_{i1} \mid x_{i2}\ )$ has a bivariate normal distribution. Next, consider the distribution of $(y_i, x_{i1})$ given $x_{i2}$ when no data are missing. The density $f(\ y_i, x_{i1} \mid x_{i2}\ )$ is given by

$$\left( \begin{matrix} Y_i \\ X_{i1} \end{matrix} \middle| x_{i2} \right) \sim N\left[ \begin{pmatrix} \theta_0 + \theta_1 x_{i2} \\ \gamma_0 + \gamma_1 x_{i2} \end{pmatrix}, \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{12} & \sigma_{22}^2 \end{pmatrix} \right] \tag{4.3}$$

Then, in terms of the parameters in Equation (4.3), the regression model $\mu_i = E[Y_i | x_i, \beta] = \beta_0 + \beta x_i'$ is given by,

$$E(Y_i | x_i) = (\theta_0 + \theta_1 x_{i2}) + \frac{\sigma_{12}}{\sigma_{11}\sigma_{22}} \frac{\sigma_{11}}{\sigma_{22}} (x_{i1} - \gamma_0 - \gamma_1 x_{i2})$$

$$= \left( \theta_0 - \frac{\sigma_{12}}{\sigma_{22}^2} \gamma_0 \right) + \frac{\sigma_{12}}{\sigma_{22}^2} x_{i1} + \left( \theta_1 - \frac{\sigma_{12}}{\sigma_{22}^2} \gamma_1 \right) x_{i2} \tag{4.4}$$

$$= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

where

$$\beta_0 = \theta_0 - \frac{\sigma_{12}}{\sigma_{22}^2} \gamma_0$$

$$\beta_1 = \frac{\sigma_{12}}{\sigma_{22}^2}$$

$$\beta_2 = \theta_1 - \frac{\sigma_{12}}{\sigma_{22}^2}\gamma_1$$

Further, the conditional variance is

$$Var(Y_i|x_{i1}, x_{i2}) = \sigma_{11}^2(1-\rho^2) = \sigma_{11}^2\left(1 - \frac{\sigma_{12}^2}{\sigma_{11}^2\sigma_{22}^2}\right) = \sigma_{11}^2 - \frac{\sigma_{12}^2}{\sigma_{22}^2} \tag{4.5}$$

In the presence of missing at random of $x_{i1}$, if the parameters $(\theta_0, \theta_1, \gamma_0, \gamma_1, \sigma_{11}^2, \sigma_{12}, \sigma_{22}^2)$ in Equation (4.3) can be consistently estimated, they can be substituted in Equation (4.4) to consistently estimate the regression parameters of interest. The protective estimator of $\beta$ uses the conditional distributions of $f(y_i|x_{i2})$ and $f(x_{i1}|y_i, x_{i2})$ to estimate these parameters. Since $x_{i2}$ and $y_i$ are both fully observed, it is straightforward to estimate $f(y_i|x_{i2})$ using all observations.

From an examination of Equation (4.3), note that the conditional mean of $Y_i$ given $x_{i2}$ is

$$E[Y_i|x_{i2}, \theta] = \theta_0 + \theta_1 x_{i2} \tag{4.6}$$

with conditional variance

$$V[Y_i|x_{i2}] = \sigma_{11}^2 \tag{4.7}$$

Since there are no missing data on $Y_i$ or $x_{i2}$, $(\theta_0, \theta_1, \sigma_{11}^2)$ can be consistently estimated using ordinary least squares, where the outcome variable is $Y_i$ and the regression model is given by Equation (4.6). Suppose we denote the ordinary least squares estimate of these parameters by $\left(\hat{\theta}_0,\hat{\theta}_1,\hat{\sigma}_{11}^2\right)$. Estimation of the remaining parameters, $\left(\gamma_0,\gamma_1,\sigma_{12},\sigma_{22}^2\right)$ can be based on the conditional distribution $f\left(x_{i1}|y_i,x_{i2}\right)$.

However, since $x_{i1}$ is observed when $r_i=1$, it is not straightforward to estimate $f\left(x_{i1}|y_i,x_{i2}\right)$ unless missing at random mechanism of $x_{i1}$.

However, the missing data mechanism is called missing at random (MAR) (Rubin, 1976) that the missing data mechanism does not depend on the set of missing values though it may possibly depend on the set of observed values. If the missingness mechanism does not depend on the parameters of the model, this assumption is called distinct. Moreover, if both MAR and distinctness hold, then the missing data mechanism is said to be ignorable (Little and Rubin, 1987). So, it is possible to estimate the relationships between $X_{i1}$ and other variables only when $X_{i1}$ is observed $\left(R_i=1\right)$. Consider the density

$$f\left(x_{i1}|y_i,x_{i2},R_i=1\right)=f\left(x_{i1}|y_i,x_{i2}\right) \tag{4.8}$$

The result in Equation (4.8) implies that the complete cases ($R_i=1$) can be used to consistently estimate the parameters of the conditional distribution of $X_{i1}$ given ($y_i,x_{i2}$).

In particular

$$\begin{aligned}
E\left[X_{i1}|y_i,x_{i2}\right] &= \left(\gamma_0+\gamma_1 x_{i2}\right)+\left(\frac{\sigma_{12}}{\sigma_{11}\sigma_{22}}\right)\left(\frac{\sigma_{22}}{\sigma_{11}}\right)\left(y_i-\theta_0-\theta_1 x_{i2}\right)\\
&=\left(\gamma_0-\frac{\sigma_{12}}{\sigma_{11}^2}\theta_0\right)+\frac{\sigma_{12}}{\sigma_{11}^2}y_i+\left(\gamma_1-\frac{\sigma_{12}}{\sigma_{11}^2}\theta_1\right)x_{i2}\\
&=\phi_0+\phi_1 y_i+\phi_2 x_{i2}
\end{aligned} \tag{4.9}$$

where

$$\phi_1 = \frac{\sigma_{12}}{\sigma_{11}^2} \qquad \phi_0 = \gamma_0 - \frac{\sigma_{12}}{\sigma_{11}^2}\theta_0 = \gamma_0 - \phi_1\theta_0$$

$$\phi_2 = \gamma_1 - \frac{\sigma_{12}}{\sigma_{11}^2}\theta_1 = \gamma_1 - \phi_1\theta_1$$

Also, the conditional variance is given by

$$Var\left(X_{i1}\middle|y_i,x_{i2}\right) = \sigma_{22}^2\left(1-\rho^2\right) = \sigma_{22}^2\left(1 - \frac{\sigma_{12}^2}{\sigma_{11}^2\sigma_{22}^2}\right) = \sigma_{22}^2 - \frac{\sigma_{12}^2}{\sigma_{11}^2} \tag{4.10}$$

Then the parameters $\left[\phi_0,\phi_1,\phi_2,Var\left(X_{i1}\middle|y_i,x_{i2}\right)\right]$ can be estimated, based on the complete cases, via ordinary least squares regression with outcome variable $X_{i1}$ and independent variables $\left[y_i,x_{i2}\right]$. Given the ordinary least squares estimates $\left[\hat{\phi}_0,\hat{\phi}_1,\hat{\phi}_2,\widehat{Var}\left(X_{i1}\middle|y_i,x_{i2}\right)\right]$ from the latter regression model, and the estimate $\left(\hat{\theta}_0,\hat{\theta}_1,\hat{\sigma}_{11}^2\right)$ from the regression model in Equation (4.6), $\left(\gamma_0,\gamma_1,\sigma_{22}^2,\sigma_{12}\right)$ can be estimated as follows,

$$\hat{\gamma}_0 = \hat{\phi}_0 + \hat{\phi}_1\theta_0 \qquad \hat{\gamma}_1 = \hat{\phi}_2 + \hat{\phi}_1\hat{\theta}_1$$

From an examination of the residual variance in Equation (4.10), note that

$$\frac{\sigma_{12}^2}{\sigma_{11}^2} = \sigma_{22}^2 - Var\left[X_{i1}\middle|y_i,x_{i2}\right]$$

so that

$$\sigma_{12} = \frac{\sigma_{12}^2/\sigma_{11}^2}{\sigma_{12}/\sigma_{11}^2} = \frac{\sigma_{12}^2/\sigma_{11}^2}{\phi_1} = \frac{\sigma_{22}^2 - Var\left[X_{i1}\middle|y_i,x_{i2}\right]}{\phi_1}$$

then $\sigma_{12}$ can be estimated using

$$\hat{\sigma}_{12} = \frac{\hat{\sigma}_{22}^{2} - \hat{Var}\left[X_{i1} \middle| y_i, x_{i2}\right]}{\hat{\phi}_1}$$

and

$$\hat{\phi}_1 = \frac{\hat{\sigma}_{12}}{\hat{\sigma}_{11}^{2}},$$

$\sigma_{22}^{2}$ can be estimated using

$$\sigma_{22}^{2} = \hat{Var}\left[X_{i1} \middle| y_i, x_{i2}\right] + \hat{\sigma}_{12}\hat{\phi}_1 = \hat{Var}\left[X_{i1} \middle| y_i, x_{i2}\right] + \frac{\hat{\sigma}_{12}^{2}}{\hat{\sigma}_{11}^{2}}$$

Then, the protective estimator of $\beta' = [\beta_0, \beta_1, \beta_2']$ in Equation (4.4) is given by

$$\hat{\beta}_0 = \hat{\theta}_0 - \frac{\hat{\sigma}_{12}}{\hat{\sigma}_{22}^{2}}\hat{\gamma}_0$$

where $\hat{\gamma}_0 = \hat{\phi}_0 + \hat{\phi}_1\theta_0$

then $\hat{\beta}_0 = \hat{\theta}_0 - \hat{\beta}_1(\hat{\phi}_0 + \hat{\phi}_1\hat{\theta}_0)$

$$\hat{\beta}_1 = \frac{\hat{\sigma}_{12}}{\hat{\sigma}_{22}^{2}}$$

$$\hat{\beta}_2 = \hat{\theta}_1 - \frac{\hat{\sigma}_{12}}{\hat{\sigma}_{22}^{2}}\hat{\gamma}_1$$

where $\hat{\gamma}_1 = \hat{\phi}_2 + \hat{\phi}_1\hat{\theta}_1$

then $\hat{\beta}_2 = \hat{\theta}_1 - \hat{\beta}_1(\hat{\phi}_2 + \hat{\phi}_1\hat{\theta}_1)$

When the assumptions about the missing data mechanism and the specification of $f(y_i, x_{i1} | x_{i2})$ are correct, results from method of moments can be used to show that $\hat{\beta}$ is consistent and has an asymptotic multivariate normal distribution, with mean vector $\beta$ and a covariance matrix that can be consistently estimated using the Equation (4.9). Because $X$ matrix is a complicated function of $[\phi_0, \phi_1, \phi_2, Var(X_{i1} | y_i, x_{i2})]$ and ordinary least squares regression is computationally demanding.

# CHAPTER FIVE
# APPLICATION

## 5.1 Introduction

Computer simulation studies are considered for investigating attrition bias as unavailable data in field studies is known by the investigator. This information allows the computation of the correct parameter estimates and a direct comparison of the true and observed estimates. In other words, the correct distribution of the data and the attrition mechanisms are comprehended because they were formed by investigator. In this respect, simulation studies allow us to understand of impact that methods used to account for attrition on our real-world results.

In this chapter, simulation study will be given a modest to compare the estimates obtained using the complete cases (CC), EM algorithm (EM), proposed protective estimate (PE) and Multiple Imputation (MI) from 1 to 10 repeat imputations.

## 5.2 Simulation Study

In the simulation study, there are two covariates ( $X_{i1}, X_{i2}$ ). The distribution of $Y_i$ given ( $x_{i1}, x_{i2}$ ) is assumed to be normal with mean $\mu_i = E(Y_i|x_{i1}, x_{i2}) = 1 + x_{i1} + x_{i2}$ and variance 2, so that $\beta' = (\beta_0, \beta_1, \beta_2) = (1,1,1)$. The variance-covariance matrix and the correlation matrix of the data are given in Table 5.1.

Table 5.1 Covariance and Correlation Coefficients Between Variables

| Covariance Matrix | | | Correlation Matrix | | |
|---|---|---|---|---|---|
| Y | X1 | X2 | Y | X1 | X2 |
| 2,11963 | | | 1,00000 | | |
| 1,06161 | 1,00696 | | 0,72666 | 1,00000 | |
| 1,05101 | 0,05333 | 0,99681 | 0,72306 | 0,05323 | 1,00000 |

In the simulation study, once the data are generated. Because of the missing data mechanism MAR, the missingness of $X_1$ depends on $Y$. Two different types of missing data are formed. In type 1, the absolute value of $Y$ is taken first, and then $X_1$ corresponding the minimum values of $Y$ $(\min(|Y|))$ are missing. In type 2, the absolute value of $Y$ is taken first, and then $X_1$ corresponding the maximum values of $Y$ $(\max(|Y|))$ are missing. Note that in both cases, the proportion of missing values are let 6%, 9%, 12% and 15% respectively. For each of n=50, n=75 and n=100, this process is performed 500 times. The plan of simulation study is given Table 5.2.

Table 5.2 The plan of simulation study

| Sample size | Missing at random MAR<br>$X_1$ is missing, where … | Proportion of missing values (%) | Methods |
|---|---|---|---|
| n=50, 75 ,100 | $\min(\lvert Y\rvert)$, $\max(\lvert Y\rvert)$ | 6% | CC, EM, PE,<br>MI(1)…MI(10) |
| | | 9% | CC, EM, PE,<br>MI(1)…MI(10) |
| | | 12% | CC, EM, PE,<br>MI(1)…MI(10) |
| | | 15% | CC, EM, PE,<br>MI(1)…MI(10) |

In this simulation study, Multiple Imputations are applied by using SOLAS and EM algorithms are applied by using SPSS. Complete Case analysis are applied by using Minitab. Macro program is written for Protective Estimator by using Minitab commands. The macro program is given in Appendix A.

The simulation study results are given in the following tables.

Table 5.3 Summary of results when missing proportion 6% and n=50

| | $\min(\mid Y \mid)$ | | | | $\max(\mid Y \mid)$ | | | |
|---|---|---|---|---|---|---|---|---|
| Methods | $E(\hat{\beta}_0)$<br>$E(S_{\hat{\beta}_0})$ | $E(\hat{\beta}_1)$<br>$E(S_{\hat{\beta}_1})$ | $E(\hat{\beta}_2)$<br>$E(S_{\hat{\beta}_2})$ | $E(MSE)$ | $E(\hat{\beta}_0)$<br>$E(S_{\hat{\beta}_0})$ | $E(\hat{\beta}_1)$<br>$E(S_{\hat{\beta}_1})$ | $E(\hat{\beta}_2)$<br>$E(S_{\hat{\beta}_2})$ | $E(MSE)$ |
| CC | 1,0022<br>0,0103 | 1,0019<br>0,0103 | 1,0008<br>0,0104 | 0,00483 | 0,9993<br>0,0105 | 1,0001<br>0,0112 | 0,9999<br>0,0112 | 0,00486 |
| PE | 1,0020<br>0,0097 | 1,0017<br>0,0098 | 1,0008<br>0,0098 | 0,00452 | 0,9998<br>0,0097 | 1,0007<br>0,0099 | 1,0006<br>0,0098 | 0,00455 |
| EM | 1,0020<br>0,0097 | 1,0021<br>0,0098 | 1,0008<br>0,0098 | 0,00452 | 0,9998<br>0,0097 | 1,0012<br>0,0099 | 1,0006<br>0,0098 | 0,00456 |
| MI(1) | 1,0017<br>0,0102 | 1,0012<br>0,0103 | 1,0016<br>0,0103 | 0,00502 | 1,0019<br>0,0103 | 1,0019<br>0,0104 | 1,0037<br>0,0104 | 0,00509 |
| MI(2) | 1,0038<br>0,0098 | 1,0022<br>0,0099 | 0,9988<br>0,0099 | 0,00464 | 1,0035<br>0,0100 | 1,0075<br>0,0102 | 1,0047<br>0,0102 | 0,00486 |
| MI(3) | 1,0016<br>0,0102 | 1,0022<br>0,0103 | 1,0005<br>0,0104 | 0,00507 | 1,0009<br>0,0103 | 1,0040<br>0,0105 | 1,0027<br>0,0105 | 0,00515 |
| MI(4) | 1,0088<br>0,0104 | 0,9976<br>0,0105 | 0,9979<br>0,0106 | 0,00527 | 1,0005<br>0,01023 | 1,0052<br>0,0104 | 1,0068<br>0,0104 | 0,00507 |
| MI(5) | 1,0060<br>0,0099 | 1,0002<br>0,0100 | 0,9985<br>0,0101 | 0,00477 | 1,0011<br>0,0098 | 1,0020<br>0,0100 | 1,0010<br>0,00998 | 0,00464 |
| MI(6) | 1,0016<br>0,0098 | 1,0026<br>0,0099 | 1,0006<br>0,0100 | 0,00466 | 0,9987<br>0,0098 | 1,0001<br>0,0100 | 0,9988<br>0,0100 | 0,00469 |
| MI(7) | 1,0053<br>0,0099 | 1,0003<br>0,0100 | 0,9992<br>0,0100 | 0,00473 | 1,0030<br>0,0099 | 1,00500<br>0,01040 | 1,0047<br>0,0101 | 0,00478 |
| MI(8) | 1,0034<br>0,0098 | 1,0021<br>0,0099 | 0,9992<br>0,0100 | 0,00469 | 1,0037<br>0,0101 | 1,0075<br>0,01034 | 1,0054<br>0,0102 | 0,00495 |
| MI(9) | 0,9989<br>0,0101 | 1,0031<br>0,0102 | 1,0025<br>0,0102 | 0,00489 | 0,9981<br>0,0099 | 0,9986<br>0,0101 | 0,9989<br>0,0101 | 0,00477 |
| MI(10) | 1,0002<br>0,0098 | 1,0035<br>0,0099 | 1,0010<br>0,0099 | 0,00464 | 0,9988<br>0,0098 | 1,0012<br>0,0099 | 0,9989<br>0,0099 | 0,00464 |

As for Table 5.3, PE and EM give the lowest MSE as for the missing data corresponding to minimum $Y$ values. PE gives the lowest MSE as for the missing data corresponding to maximum $Y$ values. EM is following this method. In relation to this, the smallest values in standard errors of coefficients ($S_{\hat{\beta}_j}$) are obtained from these two methods. At the same time, despite the fact that $\hat{\beta}_0$ coefficient at MI(10), $\hat{\beta}_1$ coefficient at MI(5) and $\hat{\beta}_2$ coefficient at MI(3) are found approximately 1.When the nearness of $\hat{\beta}_j$ coefficients to 1 is considered, it is observed that PE gives the better results. MI(4) maximizes the MSE and $S_{\hat{\beta}_j}$ as for the missing data corresponding to minimum $Y$ values. As for the missing data corresponding to maximum $Y$ values, MI(3) maximizes the MSE, but CC maximizes $S_{\hat{\beta}_j}$ .

Table 5.4 Summary of results when missing proportion 9% and n=50

| | min(\| $Y$ \|) | | | | max(\| $Y$ \|) | | | |
|---|---|---|---|---|---|---|---|---|
| Methods | $E(\hat{\beta}_0)$<br>$E(S_{\hat{\beta}_0})$ | $E(\hat{\beta}_1)$<br>$E(S_{\hat{\beta}_1})$ | $E(\hat{\beta}_2)$<br>$E(S_{\hat{\beta}_2})$ | $E(MSE)$ | $E(\hat{\beta}_0)$<br>$E(S_{\hat{\beta}_0})$ | $E(\hat{\beta}_1)$<br>$E(S_{\hat{\beta}_1})$ | $E(\hat{\beta}_2)$<br>$E(S_{\hat{\beta}_2})$ | $E(MSE)$ |
| CC | 1,0003<br>0,0106 | 1,0018<br>0,0106 | 1,0012<br>0,0108 | 0,00486 | 0,9988<br>0,0108 | 1,0000<br>0,0117 | 0,9990<br>0,0117 | 0,00475 |
| PE | 1,0001<br>0,0095 | 1,0016<br>0,0096 | 1,0011<br>0,0097 | 0,00435 | 0,9997<br>0,0094 | 1,0010<br>0,0094 | 1,0001<br>0,0095 | 0,00425 |
| EM | 1,0001<br>0,0095 | 1,0022<br>0,0096 | 1,0010<br>0,0097 | 0,00435 | 0,9997<br>0,0094 | 1,0017<br>0,0094 | 1,0001<br>0,0095 | 0,00425 |
| MI(1) | 1,0012<br>0,0101 | 1,0001<br>0,0102 | 1,0016<br>0,0103 | 0,00491 | 1,0048<br>0,0102 | 1,0069<br>0,0103 | 1,0077<br>0,0103 | 0,00502 |
| MI(2) | 1,0061<br>0,0106 | 0,9982<br>0,0107 | 0,9979<br>0,0108 | 0,00541 | 0,9995<br>0,0103 | 0,9964<br>0,0103 | 0,9966<br>0,0104 | 0,00513 |
| MI(3) | 1,0035<br>0,0103 | 1,0000<br>0,0104 | 0,9992<br>0,0106 | 0,00516 | 1,0039<br>0,0103 | 1,0082<br>0,0104 | 1,0061<br>0,0104 | 0,00512 |
| MI(4) | 0,9969<br>0,0096 | 1,0036<br>0,0098 | 1,0026<br>0,0099 | 0,00450 | 0,9947<br>0,0095 | 0,9958<br>0,0096 | 0,9943<br>0,0097 | 0,00441 |
| MI(5) | 0,9962<br>0,0098 | 1,0025<br>0,0099 | 1,0043<br>0,0101 | 0,00468 | 0,9937<br>0,0097 | 0,9922<br>0,0097 | 0,9938<br>0,0099 | 0,00459 |
| MI(6) | 0,9921<br>0,0113 | 1,0056<br>0,0115 | 1,0035<br>0,0116 | 0,00623 | 0,9942<br>0,0108 | 0,9962<br>0,0108 | 0,9937<br>0,0109 | 0,00566 |
| MI(7) | 0,9951<br>0,0104 | 1,0045<br>0,0105 | 1,0026<br>0,0106 | 0,00522 | 0,9934<br>0,0101 | 0,9929<br>0,0101 | 0,9913<br>0,0103 | 0,00500 |
| MI(8) | 0,9986<br>0,0109 | 1,0030<br>0,0111 | 1,0003<br>0,0112 | 0,00577 | 0,9976<br>0,0108 | 1,0018<br>0,0109 | 0,9977<br>0,0109 | 0,00565 |
| MI(9) | 1,0037<br>0,0108 | 1,0012<br>0,0110 | 0,9971<br>0,0111 | 0,00568 | 1,0016<br>0,0106 | 1,0049<br>0,0108 | 1,0005<br>0,0108 | 0,00549 |
| MI(10) | 1,0023<br>0,0098 | 1,0012<br>0,0100 | 0,9995<br>0,0101 | 0,00470 | 1,0008<br>0,0097 | 1,0031<br>0,0098 | 1,0009<br>0,0099 | 0,00459 |

As for Table 5.4, PE and EM give the lowest average of MSE. In relation to this, the smallest values of $S_{\hat{\beta}_j}$ are obtained from these two methods. For both cases, MI(4) is the closest result to these two methods. When the nearness of $\hat{\beta}_j$ coefficients to 1 for the missing data corresponding to maximum $Y$ is considered, it is observed that PE gives the better results. For both cases, MI(6) maximizes the MSE and the standard errors of regression coefficients.

Table 5.5 Summary of results when missing proportion 12% and n=50

| | $\min(\lvert Y \rvert)$ | | | | $\max(\lvert Y \rvert)$ | | | |
|---|---|---|---|---|---|---|---|---|
| Methods | $E(\hat{\beta}_0)$<br>$E(S_{\hat{\beta}_0})$ | $E(\hat{\beta}_1)$<br>$E(S_{\hat{\beta}_1})$ | $E(\hat{\beta}_2)$<br>$E(S_{\hat{\beta}_2})$ | $E(MSE)$ | $E(\hat{\beta}_0)$<br>$E(S_{\hat{\beta}_0})$ | $E(\hat{\beta}_1)$<br>$E(S_{\hat{\beta}_1})$ | $E(\hat{\beta}_2)$<br>$E(S_{\hat{\beta}_2})$ | $E(MSE)$ |
| CC | 1,0006<br>0,0107 | 1,0023<br>0,0106 | 1,0004<br>0,0107 | 0,00478 | 0,9991<br>0,0110 | 0,9985<br>0,0122 | 0,9990<br>0,0122 | 0,00478 |
| PE | 1,0003<br>0,0093 | 1,0021<br>0,0094 | 1,0003<br>0,0094 | 0,00417 | 1,0001<br>0,0093 | 0,9998<br>0,0094 | 1,0003<br>0,0095 | 0,00418 |
| EM | 1,0003<br>0,0093 | 1,0028<br>0,0094 | 1,0002<br>0,0094 | 0,00418 | 1,0002<br>0,0093 | 1,0006<br>0,0094 | 1,0003<br>0,0095 | 0,00419 |
| MI(1) | 1,0023<br>0,0099 | 1,0003<br>0,0100 | 1,0005<br>0,0101 | 0,00477 | 1,0068<br>0,0102 | 1,0076<br>0,0104 | 1,0102<br>0,0104 | 0,00505 |
| MI(2) | 1,0056<br>0,0105 | 0,9987<br>0,0105 | 0,9980<br>0,0106 | 0,00529 | 0,9979<br>0,0104 | 0,9922<br>0,0104 | 0,9934<br>0,0106 | 0,00524 |
| MI(3) | 1,0050<br>0,0102 | 1,0005<br>0,0103 | 0,9973<br>0,0104 | 0,00506 | 1,0057<br>0,0103 | 1,0077<br>0,0105 | 1,0077<br>0,0105 | 0,00512 |
| MI(4) | 0,9960<br>0,0095 | 1,0047<br>0,0096 | 1,0023<br>0,0096 | 0,00438 | 0,9938<br>0,0095 | 0,9931<br>0,0095 | 0,9929<br>0,0097 | 0,00438 |
| MI(5) | 0,9955<br>0,0097 | 1,0031<br>0,0098 | 1,0042<br>0,0098 | 0,00454 | 0,9928<br>0,0097 | 0,9896<br>0,0097 | 0,9929<br>0,0099 | 0,00454 |
| MI(6) | 0,9932<br>0,0112 | 1,0057<br>0,0114 | 1,0024<br>0,0114 | 0,00609 | 0,9961<br>0,0109 | 0,9982<br>0,0111 | 0,9965<br>0,0112 | 0,00583 |
| MI(7) | 0,9941<br>0,0103 | 1,0054<br>0,0104 | 1,0026<br>0,0104 | 0,00510 | 0,9925<br>0,0101 | 0,9911<br>0,0102 | 0,9900<br>0,0104 | 0,00498 |
| MI(8) | 1,0004<br>0,0109 | 1,0036<br>0,0111 | 0,9979<br>0,0111 | 0,00579 | 0,9996<br>0,0109 | 1,0025<br>0,0110 | 1,0002<br>0,0111 | 0,00574 |
| MI(9) | 1,0063<br>0,0109 | 1,0009<br>0,0111 | 0,9947<br>0,0111 | 0,00580 | 1,0041<br>0,0108 | 1,0060<br>0,0110 | 1,0029<br>0,0110 | 0,00563 |
| MI(10) | 1,0029<br>0,0097 | 1,0019<br>0,0098 | 0,9982<br>0,0098 | 0,00454 | 1,0014<br>0,0096 | 1,0020<br>0,0098 | 1,0008<br>0,0099 | 0,00452 |

As for Table 5.5, PE gives the lowest average of MSE. The closest result to this method is obtained from EM. In relation to this, the smallest values of $S_{\hat{\beta}_j}$ are obtained from these two methods. The closest result to these two methods for both cases is obtained from MI(4). When the nearness of $\hat{\beta}_j$ coefficients to 1 is considered for the missing data corresponding to maximum $Y$ values, it is observed that PE gives the better results. For both cases, MI(6) maximizes the MSE. MI(6) maximizes standard errors of regression coefficients as for the missing data corresponding to minimum $Y$ values. CC maximizes standard errors of regression coefficients as for the missing data corresponding to maximum $Y$ values.

Table 5.6 Summary of results when missing proportion 15% and n=50

| | $\min(\lvert Y \rvert)$ | | | | $\max(\lvert Y \rvert)$ | | | |
|---|---|---|---|---|---|---|---|---|
| Methods | $E(\hat{\beta}_0)$<br>$E(S_{\hat{\beta}_0})$ | $E(\hat{\beta}_1)$<br>$E(S_{\hat{\beta}_1})$ | $E(\hat{\beta}_2)$<br>$E(S_{\hat{\beta}_2})$ | $E(MSE)$ | $E(\hat{\beta}_0)$<br>$E(S_{\hat{\beta}_0})$ | $E(\hat{\beta}_1)$<br>$E(S_{\hat{\beta}_1})$ | $E(\hat{\beta}_2)$<br>$E(S_{\hat{\beta}_2})$ | $E(MSE)$ |
| CC | 1,0014<br>0,0110 | 1,0018<br>0,0109 | 1,0013<br>0,0109 | 0,00481 | 0,9982<br>0,0115 | 0,9984<br>0,0130 | 0,9982<br>0,0129 | 0,00480 |
| PE | 1,0010<br>0,0091 | 1,0015<br>0,0092 | 1,0011<br>0,0092 | 0,00400 | 0,9997<br>0,0091 | 1,0002<br>0,0092 | 0,9999<br>0,0092 | 0,00400 |
| EM | 1,0012<br>0,0093 | 1,0023<br>0,0094 | 1,0007<br>0,0093 | 0,00406 | 0,9997<br>0,0091 | 1,0014<br>0,0092 | 1,0002<br>0,0093 | 0,00405 |
| MI(1) | 1,0033<br>0,0099 | 0,9993<br>0,0100 | 1,0012<br>0,0100 | 0,00467 | 1,0066<br>0,0100 | 1,0065<br>0,0102 | 1,0098<br>0,0102 | 0,00487 |
| MI(2) | 0,9990<br>0,0100 | 1,0035<br>0,0101 | 1,0012<br>0,0101 | 0,00474 | 0,9994<br>0,0098 | 1,0031<br>0,0100 | 1,0002<br>0,0100 | 0,00466 |
| MI(3) | 1,0049<br>0,0095 | 0,9990<br>0,0096 | 1,0003<br>0,0096 | 0,00430 | 1,0026<br>0,0095 | 1,0009<br>0,0096 | 1,0036<br>0,0097 | 0,00439 |
| MI(4) | 0,9972<br>0,0099 | 1,0043<br>0,0101 | 1,0020<br>0,0100 | 0,00468 | 0,9956<br>0,0097 | 0,9980<br>0,00980 | 0,9953<br>0,0099 | 0,00455 |
| MI(5) | 1,0009<br>0,0118 | 0,9988<br>0,0119 | 1,0024<br>0,0120 | 0,00663 | 0,9896<br>0,0118 | 0,9874<br>0,0118 | 0,9903<br>0,0120 | 0,00670 |
| MI(6) | 1,0083<br>0,0098 | 0,9975<br>0,0099 | 0,9983<br>0,0099 | 0,00458 | 1,0005<br>0,0098 | 0,9965<br>0,0099 | 0,9985<br>0,0100 | 0,00466 |
| MI(7) | 1,0045<br>0,0106 | 1,0010<br>0,0107 | 0,9980<br>0,0107 | 0,00533 | 1,0033<br>0,0107 | 1,0054<br>0,0109 | 1,0025<br>0,0108 | 0,00552 |
| MI(8) | 1,0047<br>0,0099 | 0,9996<br>0,0100 | 0,9994<br>0,0100 | 0,00464 | 0,9985<br>0,0099 | 0,9967<br>0,0100 | 0,9975<br>0,0100 | 0,00471 |
| MI(9) | 0,9956<br>0,0099 | 1,0040<br>0,0100 | 1,0039<br>0,0100 | 0,00462 | 0,9986<br>0,0099 | 1,0021<br>0,0100 | 1,0021<br>0,0100 | 0,00471 |
| MI(10) | 1,0045<br>0,0099 | 0,9998<br>0,0100 | 0,9994<br>0,0100 | 0,00469 | 1,0019<br>0,0099 | 1,0000<br>0,0100 | 1,0009<br>0,0100 | 0,00470 |

As for Table 5.6, PE gives the lowest average of MSE. The closest result to this method is obtained from EM. In relation to this, the smallest values of $S_{\hat{\beta}_j}$ are obtained from PE. As for the missing data corresponding to maximum $Y$ values, $\hat{\beta}_j$ coefficients are found approximately 1 at the method of PE. For both cases, MI(5) maximizes the MSE. MI(5) maximizes standard errors of regression coefficients as for the missing data corresponding to minimum $Y$ values. CC maximize standard errors of regression coefficients as for the missing data corresponding to maximum $Y$ values.

Table 5.7 Summary of results when missing proportion 6% and n=75

| | min(\| $Y$ \|) | | | | max(\| $Y$ \|) | | | |
|---|---|---|---|---|---|---|---|---|
| Methods | $E(\hat{\beta}_0)$<br>$E(S_{\hat{\beta}_0})$ | $E(\hat{\beta}_1)$<br>$E(S_{\hat{\beta}_1})$ | $E(\hat{\beta}_2)$<br>$E(S_{\hat{\beta}_2})$ | $E(MSE)$ | $E(\hat{\beta}_0)$<br>$E(S_{\hat{\beta}_0})$ | $E(\hat{\beta}_1)$<br>$E(S_{\hat{\beta}_1})$ | $E(\hat{\beta}_2)$<br>$E(S_{\hat{\beta}_2})$ | $E(MSE)$ |
| CC | 1,0014<br>0,0084 | 1,0012<br>0,0084 | 1,0004<br>0,0084 | 0,00486 | 1,0005<br>0,0086 | 0,9999<br>0,0092 | 0,9999<br>0,0093 | 0,00489 |
| PE | 1,0013<br>0,0078 | 1,0010<br>0,0079 | 1,0004<br>0,0079 | 0,00452 | 1,0010<br>0,0079 | 1,0007<br>0,0080 | 1,0007<br>0,0080 | 0,00455 |
| EM | 1,0013<br>0,0078 | 1,0014<br>0,0079 | 1,0003<br>0,0079 | 0,00452 | 1,0011<br>0,0079 | 1,0011<br>0,0080 | 1,0007<br>0,0080 | 0,00455 |
| MI(1) | 1,0020<br>0,0082 | 1,0002<br>0,0082 | 1,0005<br>0,0082 | 0,00489 | 1,0040<br>0,0083 | 1,0044<br>0,0084 | 1,0055<br>0,0085 | 0,00505 |
| MI(2) | 1,0048<br>0,0084 | 0,9991<br>0,0084 | 0,9986<br>0,0084 | 0,00515 | 1,0014<br>0,0083 | 0,9987<br>0,0084 | 0,9994<br>0,0085 | 0,00508 |
| MI(3) | 1,0034<br>0,0083 | 1,0000<br>0,0083 | 0,9992<br>0,0083 | 0,00504 | 1,0036<br>0,0084 | 1,0036<br>0,0085 | 1,0047<br>0,0085 | 0,00513 |
| MI(4) | 0,9992<br>0,0079 | 1,0023<br>0,0080 | 1,0014<br>0,0080 | 0,00462 | 0,9980<br>0,0080 | 0,9971<br>0,0080 | 0,9966<br>0,0081 | 0,00465 |
| MI(5) | 0,9987<br>0,0080 | 1,0017<br>0,0081 | 1,0023<br>0,0081 | 0,00473 | 0,9973<br>0,0081 | 0,9950<br>0,0081 | 0,9961<br>0,0082 | 0,00476 |
| MI(6) | 0,9967<br>0,0087 | 1,0031<br>0,0088 | 1,0021<br>0,0088 | 0,00559 | 0,9975<br>0,0086 | 0,9970<br>0,0086 | 0,9960<br>0,0087 | 0,00541 |
| MI(7) | 0,9984<br>0,0083 | 1,0026<br>0,0083 | 1,0014<br>0,0083 | 0,00502 | 0,9973<br>0,0083 | 0,9953<br>0,0083 | 0,9948<br>0,0084 | 0,00501 |
| MI(8) | 1,0003<br>0,0085 | 1,0018<br>0,0086 | 1,0001<br>0,0086 | 0,00537 | 0,9997<br>0,0086 | 1,0005<br>0,0087 | 0,9986<br>0,0087 | 0,00540 |
| MI(9) | 1,0035<br>0,0085 | 1,0006<br>0,0086 | 0,9982<br>0,0086 | 0,00530 | 1,0023<br>0,0085 | 1,0030<br>0,0086 | 1,0011<br>0,0087 | 0,00531 |
| MI(10) | 1,0024<br>0,0080 | 1,0010<br>0,0081 | 0,9995<br>0,0081 | 0,00474 | 1,0017<br>0,0081 | 1,0020<br>0,0082 | 1,0012<br>0,0082 | 0,00477 |

As for Table 5.7, PE and EM give the lowest average of MSE. In relation to this, the smallest values of $S_{\hat{\beta}_j}$ are obtained from these two methods. The closest result to these two methods for both cases is obtained from MI(4). When the nearness of $\hat{\beta}_j$ coefficients to 1 is considered, as for the missing data corresponding to minimum $Y$ values, it is observed that MI(8) and as for the missing data corresponding to maximum $Y$ values, it is observed that CC gives the better results. For both cases, MI(6) maximizes the MSE. MI(6) maximizes standard errors of regression coefficients as for the missing data corresponding to minimum $Y$ values. CC maximizes standard errors of regression coefficients as for the missing data corresponding to maximum $Y$ values.

Table 5.8 Summary of results when missing proportion 9% and n=75

| | min($\|Y\|$) | | | | max($\|Y\|$) | | | |
|---|---|---|---|---|---|---|---|---|
| Methods | $E(\hat{\beta}_0)$<br>$E(S_{\hat{\beta}_0})$ | $E(\hat{\beta}_1)$<br>$E(S_{\hat{\beta}_1})$ | $E(\hat{\beta}_2)$<br>$E(S_{\hat{\beta}_2})$ | $E(MSE)$ | $E(\hat{\beta}_0)$<br>$E(S_{\hat{\beta}_0})$ | $E(\hat{\beta}_1)$<br>$E(S_{\hat{\beta}_1})$ | $E(\hat{\beta}_2)$<br>$E(S_{\hat{\beta}_2})$ | $E(MSE)$ |
| CC | 1,0007<br>0,0085 | 1,0010<br>0,0085 | 1,0008<br>0,0086 | 0,00480 | 0,9994<br>0,0087 | 0,9996<br>0,0095 | 0,9985<br>0,0094 | 0,00481 |
| PE | 1,0005<br>0,0077 | 1,0008<br>0,0078 | 1,0007<br>0,0078 | 0,00433 | 1,0002<br>0,0077 | 1,0006<br>0,0078 | 0,9996<br>0,0078 | 0,00435 |
| EM | 1,0005<br>0,0077 | 1,0014<br>0,0078 | 1,0007<br>0,0078 | 0,00433 | 1,0002<br>0,0077 | 1,0012<br>0,0078 | 0,9995<br>0,0078 | 0,00435 |
| MI(1) | 1,0018<br>0,0080 | 0,9996<br>0,0081 | 1,0008<br>0,0082 | 0,00472 | 1,0046<br>0,0082 | 1,0059<br>0,0083 | 1,0062<br>0,0082 | 0,00491 |
| MI(2) | 0,9994<br>0,0081 | 1,0018<br>0,0082 | 1,0009<br>0,0082 | 0,00478 | 0,9998<br>0,0081 | 1,0015<br>0,0082 | 0,9991<br>0,0081 | 0,00480 |
| MI(3) | 1,0030<br>0,0078 | 0,9994<br>0,0079 | 1,0001<br>0,0080 | 0,00449 | 1,0030<br>0,0079 | 1,0037<br>0,0080 | 1,0037<br>0,0079 | 0,00455 |
| MI(4) | 0,9979<br>0,0081 | 1,0026<br>0,0082 | 1,0016<br>0,0082 | 0,00477 | 0,9981<br>0,0080 | 0,9987<br>0,0081 | 0,9970<br>0,0081 | 0,00471 |
| MI(5) | 1,0000<br>0,0089 | 0,9997<br>0,0090 | 1,0015<br>0,0091 | 0,00581 | 0,9956<br>0,0089 | 0,9958<br>0,0090 | 0,9952<br>0,0090 | 0,00584 |
| MI(6) | 1,0051<br>0,0080 | 0,9985<br>0,0081 | 0,9988<br>0,0082 | 0,00469 | 1,0026<br>0,0079 | 1,0027<br>0,0080 | 1,0022<br>0,0080 | 0,00460 |
| MI(7) | 1,0026<br>0,0084 | 1,0002<br>0,0085 | 0,9992<br>0,0085 | 0,00511 | 1,0035<br>0,0084 | 1,0060<br>0,0086 | 1,0039<br>0,0085 | 0,00521 |
| MI(8) | 1,0023<br>0,0080 | 0,9999<br>0,0081 | 1,0001<br>0,0081 | 0,00466 | 0,9987<br>0,0080 | 0,9979<br>0,0080 | 0,9968<br>0,0080 | 0,00466 |
| MI(9) | 0,9972<br>0,0079 | 1,0023<br>0,0081 | 1,0026<br>0,0082 | 0,00466 | 0,9979<br>0,0079 | 0,9987<br>0,0080 | 0,9975<br>0,0080 | 0,00461 |
| MI(10) | 1,0030<br>0,0080 | 0,9997<br>0,0081 | 0,9996<br>0,0082 | 0,00468 | 1,0024<br>0,0080 | 1,0023<br>0,0081 | 1,0015<br>0,0081 | 0,00475 |

As for Table 5.8, PE and EM give the lowest MSE. In relation to this, the smallest values of $S_{\hat{\beta}_j}$ are obtained from these two methods. The closest result to these two methods for both cases is obtained from MI(3). When the nearness of $\hat{\beta}_j$ coefficients to 1 is considered, as for the missing data corresponding to minimum $Y$ values, it is observed that MI(8) and as for the missing data corresponding to maximum $Y$ values, it is observed that PE gives the better results. For both cases, MI(5) maximizes the MSE. MI(5) maximizes standard errors of regression coefficients as for the missing data corresponding to minimum $Y$ values. CC maximizes standard errors of regression coefficients as for the missing data corresponding to maximum $Y$ values.

Table 5.9 Summary of results when missing proportion 12% and n=75

| | $\min(\lvert Y \rvert)$ | | | | $\max(\lvert Y \rvert)$ | | | |
|---|---|---|---|---|---|---|---|---|
| Methods | $E(\hat{\beta}_0)$<br>$E(S_{\hat{\beta}_0})$ | $E(\hat{\beta}_1)$<br>$E(S_{\hat{\beta}_1})$ | $E(\hat{\beta}_2)$<br>$E(S_{\hat{\beta}_2})$ | $E(MSE)$ | $E(\hat{\beta}_0)$<br>$E(S_{\hat{\beta}_0})$ | $E(\hat{\beta}_1)$<br>$E(S_{\hat{\beta}_1})$ | $E(\hat{\beta}_2)$<br>$E(S_{\hat{\beta}_2})$ | $E(MSE)$ |
| CC | 1,0014<br>0,0087 | 1,0023<br>0,0086 | 1,0009<br>0,0086 | 0,00483 | 0,9999<br>0,0089 | 0,9989<br>0,0098 | 0,9988<br>0,0098 | 0,00482 |
| PE | 1,0011<br>0,0076 | 1,0021<br>0,0076 | 1,0007<br>0,0077 | 0,00422 | 1,0009<br>0,0076 | 1,0002<br>0,0076 | 1,0001<br>0,0076 | 0,00423 |
| EM | 1,0011<br>0,0076 | 1,0028<br>0,0076 | 1,0007<br>0,0077 | 0,00423 | 1,0010<br>0,0076 | 1,0009<br>0,0076 | 1,0002<br>0,0076 | 0,00423 |
| MI(1) | 1,0018<br>0,0079 | 1,0009<br>0,0080 | 1,0015<br>0,0080 | 0,00465 | 1,0055<br>0,0088 | 1,0049<br>0,0089 | 1,0068<br>0,0089 | 0,00479 |
| MI(2) | 1,0031<br>0,0081 | 1,0011<br>0,0081 | 0,9999<br>0,0081 | 0,00478 | 1,0053<br>0,0082 | 1,0049<br>0,0082 | 1,0051<br>0,0082 | 0,00489 |
| MI(3) | 1,0020<br>0,0080 | 1,0008<br>0,0080 | 1,0012<br>0,0081 | 0,00470 | 0,9984<br>0,0080 | 0,9952<br>0,0080 | 0,9966<br>0,0080 | 0,00470 |
| MI(4) | 0,9995<br>0,0079 | 1,0023<br>0,0080 | 1,0022<br>0,0080 | 0,00464 | 0,9989<br>0,0079 | 0,9976<br>0,0079 | 0,9986<br>0,0080 | 0,00460 |
| MI(5) | 1,0039<br>0,0080 | 0,9998<br>0,0080 | 1,0005<br>0,0081 | 0,00469 | 1,0000<br>0,0080 | 0,9956<br>0,0080 | 0,9975<br>0,0080 | 0,00467 |
| MI(6) | 0,9996<br>0,0079 | 1,0033<br>0,0079 | 1,0012<br>0,0079 | 0,00455 | 0,9989<br>0,0079 | 0,9996<br>0,0079 | 0,9987<br>0,0079 | 0,00454 |
| MI(7) | 1,0046<br>0,0080 | 1,0029<br>0,0080 | 0,9977<br>0,0081 | 0,00469 | 1,0025<br>0,0080 | 1,0029<br>0,0080 | 0,9997<br>0,0080 | 0,00465 |
| MI(8) | 0,9976<br>0,0086 | 1,0019<br>0,0087 | 1,0036<br>0,0087 | 0,00545 | 0,9986<br>0,0085 | 0,9948<br>0,0085 | 0,9973<br>0,0085 | 0,00527 |
| MI(9) | 0,9949<br>0,0083 | 1,0053<br>0,0083 | 1,0032<br>0,0083 | 0,00500 | 0,9931<br>0,0082 | 0,9915<br>0,0081 | 0,9905<br>0,0082 | 0,00490 |
| MI(10) | 1,0025<br>0,0088 | 1,0023<br>0,0089 | 0,9984<br>0,0089 | 0,00576 | 1,0012<br>0,0088 | 1,0013<br>0,0089 | 0,9982<br>0,0089 | 0,00575 |

As for Table 5.9, PE gives the lowest MSE as for the missing data corresponding to minimum $Y$ values. EM is following this method. In relation to this, the smallest values of $S_{\hat{\beta}_j}$ are obtained from these two methods. PE and EM give the lowest MSE as for the missing data corresponding to maximum $Y$ values. In relation to this, the smallest values of $S_{\hat{\beta}_j}$ are obtained from these two methods. When the nearness of $\hat{\beta}_j$ coefficients to 1 is considered as for the missing data corresponding to maximum $Y$ values, it is observed that PE gives the better results. MI(10) maximizes the MSE and $S_{\hat{\beta}_j}$ as for the missing data corresponding to minimum $Y$ values. As for the missing data corresponding to maximum $Y$ values, MI(10) maximizes the MSE, but CC maximizes $S_{\hat{\beta}_j}$ .

Table 5.10 Summary of results when missing proportion 15% and n=75

| | min(\| $Y$ \|) | | | | max(\| $Y$ \|) | | | |
|---|---|---|---|---|---|---|---|---|
| Methods | $E(\hat{\beta}_0)$<br>$E(S_{\hat{\beta}_0})$ | $E(\hat{\beta}_1)$<br>$E(S_{\hat{\beta}_1})$ | $E(\hat{\beta}_2)$<br>$E(S_{\hat{\beta}_2})$ | $E(MSE)$ | $E(\hat{\beta}_0)$<br>$E(S_{\hat{\beta}_0})$ | $E(\hat{\beta}_1)$<br>$E(S_{\hat{\beta}_1})$ | $E(\hat{\beta}_2)$<br>$E(S_{\hat{\beta}_2})$ | $E(MSE)$ |
| CC | 1,0005<br>0,0089 | 1,0018<br>0,0088 | 1,0010<br>0,0088 | 0,00485 | 0,9987<br>0,0092 | 0,9985<br>0,0102 | 0,9988<br>0,0102 | 0,00480 |
| PE | 1,0001<br>0,0075 | 1,0016<br>0,0076 | 1,0008<br>0,0076 | 0,00412 | 1,0001<br>0,0074 | 1,0000<br>0,0075 | 1,0004<br>0,0076 | 0,00408 |
| EM | 1,0001<br>0,0076 | 1,0024<br>0,0077 | 1,0008<br>0,0077 | 0,00417 | 1,0004<br>0,0075 | 1,0011<br>0,0075 | 1,0004<br>0,0076 | 0,00408 |
| MI(1) | 0,9995<br>0,0081 | 1,0008<br>0,0082 | 1,0023<br>0,0082 | 0,00475 | 1,0037<br>0,0080 | 1,0027<br>0,0081 | 1,0051<br>0,0081 | 0,00472 |
| MI(2) | 1,0030<br>0,0081 | 1,0003<br>0,0082 | 0,9996<br>0,0082 | 0,00477 | 1,0085<br>0,0084 | 1,0117<br>0,0085 | 1,0105<br>0,0085 | 0,00517 |
| MI(3) | 0,9970<br>0,0080 | 1,0023<br>0,0081 | 1,0031<br>0,0081 | 0,00466 | 0,9948<br>0,0085 | 0,9940<br>0,0084 | 0,9932<br>0,0086 | 0,00527 |
| MI(4) | 1,0033<br>0,0081 | 0,9992<br>0,0082 | 1,0005<br>0,0082 | 0,00476 | 1,0075<br>0,0088 | 1,0081<br>0,0089 | 1,0063<br>0,0089 | 0,00563 |
| MI(5) | 0,9974<br>0,0080 | 1,0033<br>0,0080 | 1,0020<br>0,0081 | 0,00464 | 0,9925<br>0,0086 | 0,9881<br>0,0085 | 0,9896<br>0,0087 | 0,00540 |
| MI(6) | 0,9974<br>0,0080 | 1,0033<br>0,0081 | 1,0020<br>0,0081 | 0,00461 | 0,9986<br>0,0079 | 0,9965<br>0,0079 | 0,9979<br>0,0080 | 0,00457 |
| MI(7) | 1,0040<br>0,0080 | 1,0016<br>0,0081 | 0,9974<br>0,0081 | 0,00466 | 1,0057<br>0,0080 | 1,0069<br>0,0081 | 1,0079<br>0,0081 | 0,00467 |
| MI(8) | 0,9946<br>0,0089 | 1,0021<br>0,0089 | 1,0046<br>0,0090 | 0,00570 | 1,0021<br>0,0080 | 1,0043<br>0,0081 | 1,0031<br>0,0081 | 0,00468 |
| MI(9) | 0,9931<br>0,0083 | 1,0055<br>0,0084 | 1,0035<br>0,0084 | 0,00499 | 0,9982<br>0,0081 | 0,9988<br>0,0081 | 0,9955<br>0,0082 | 0,00480 |
| MI(10) | 1,0036<br>0,0092 | 1,0010<br>0,0092 | 0,9973<br>0,0092 | 0,00607 | 0,9979<br>0,0077 | 0,9968<br>0,0078 | 0,9969<br>0,0079 | 0,00440 |

As for Table 5.10, PE gives the lowest MSE as for the missing data corresponding to minimum $Y$ values. EM is following this method. In relation to this, the smallest values of $S_{\hat{\beta}_j}$ are obtained from PE. PE and EM give the lowest MSE as for the missing data corresponding to maximum $Y$ values. In relation to this, the smallest values of $S_{\hat{\beta}_j}$ are obtained from these two methods. When the nearness of $\hat{\beta}_j$ coefficients to 1 is considered, as for the missing data corresponding to minimum $Y$ values, it is observed that MI(2) and as for the missing data corresponding to maximum $Y$ values, it is observed that PE gives the better results. MI(10) maximizes the MSE and $S_{\hat{\beta}_j}$ as for the missing data corresponding to minimum $Y$ values. As for the missing data corresponding to maximum $Y$ values, MI(4) maximizes the MSE, but CC maximizes $S_{\hat{\beta}_j}$ .

Table 5.11 Summary of results when missing proportion 6% and n=100

| | min($\lvert Y \rvert$) | | | | max($\lvert Y \rvert$) | | | |
|---|---|---|---|---|---|---|---|---|
| Methods | $E(\hat{\beta}_0)$<br>$E(S_{\hat{\beta}_0})$ | $E(\hat{\beta}_1)$<br>$E(S_{\hat{\beta}_1})$ | $E(\hat{\beta}_2)$<br>$E(S_{\hat{\beta}_2})$ | $E(MSE)$ | $E(\hat{\beta}_0)$<br>$E(S_{\hat{\beta}_0})$ | $E(\hat{\beta}_1)$<br>$E(S_{\hat{\beta}_1})$ | $E(\hat{\beta}_2)$<br>$E(S_{\hat{\beta}_2})$ | $E(MSE)$ |
| CC | 1,0015<br>0,0072 | 1,0017<br>0,0072 | 1,0006<br>0,0008 | 0,00477 | 1,0003<br>0,0074 | 1,0002<br>0,0078 | 0,9994<br>0,0078 | 0,00487 |
| PE | 1,0014<br>0,0067 | 1,0015<br>0,0068 | 1,0005<br>0,0007 | 0,00448 | 1,0008<br>0,0068 | 1,0009<br>0,0069 | 1,0001<br>0,0068 | 0,00458 |
| EM | 1,0014<br>0,0067 | 1,0019<br>0,0068 | 1,0005<br>0,0007 | 0,00448 | 1,0008<br>0,0068 | 1,0012<br>0,0069 | 1,0000<br>0,0068 | 0,00458 |
| MI(1) | 1,0024<br>0,0070 | 1,0007<br>0,0070 | 1,0004<br>0,0007 | 0,00476 | 1,0036<br>0,0071 | 1,0046<br>0,0072 | 1,0045<br>0,0071 | 0,00499 |
| MI(2) | 1,0034<br>0,0071 | 1,0005<br>0,0071 | 0,9995<br>0,0007 | 0,00494 | 1,0004<br>0,0071 | 0,9983<br>0,0072 | 0,9982<br>0,0072 | 0,00502 |
| MI(3) | 1,0035<br>0,0070 | 1,0007<br>0,0071 | 0,9992<br>0,0007 | 0,00488 | 1,0034<br>0,0072 | 1,0050<br>0,0072 | 1,0037<br>0,0072 | 0,00504 |
| MI(4) | 1,0003<br>0,0069 | 1,0015<br>0,0070 | 1,0017<br>0,0007 | 0,00471 | 0,9980<br>0,0070 | 0,9965<br>0,0070 | 0,9968<br>0,0070 | 0,00484 |
| MI(5) | 1,0035<br>0,0070 | 1,0007<br>0,0070 | 0,9993<br>0,0007 | 0,00479 | 1,0022<br>0,0070 | 1,0033<br>0,0071 | 1,0018<br>0,0070 | 0,00486 |
| MI(6) | 1,0020<br>0,0068 | 1,0016<br>0,0069 | 1,0001<br>0,0007 | 0,00457 | 1,0012<br>0,0069 | 1,0019<br>0,0069 | 1,0004<br>0,0069 | 0,00468 |
| MI(7) | 1,0015<br>0,0068 | 1,0018<br>0,0068 | 1,0004<br>0,0007 | 0,00455 | 1,0007<br>0,0069 | 1,0011<br>0,0069 | 0,9999<br>0,0069 | 0,00465 |
| MI(8) | 1,0026<br>0,0070 | 1,0012<br>0,0070 | 0,9996<br>0,0007 | 0,00478 | 1,0011<br>0,0070 | 1,0010<br>0,0071 | 0,9997<br>0,0071 | 0,00487 |
| MI(9) | 1,0023<br>0,0070 | 1,0015<br>0,0070 | 0,9997<br>0,0007 | 0,00481 | 1,0023<br>0,0071 | 1,0033<br>0,0071 | 1,0020<br>0,0071 | 0,00495 |
| MI(10) | 1,0020<br>0,0072 | 1,0007<br>0,0072 | 1,0005<br>0,0008 | 0,00510 | 0,9979<br>0,0073 | 0,9945<br>0,0073 | 0,9953<br>0,0074 | 0,00530 |

As for Table 5.11, PE and EM give the lowest MSE . In relation to this, the smallest values of $S_{\hat{\beta}_j}$ are obtained from these two methods. The closest result to these two methods for both cases is obtained from MI(7). When the nearness of $\hat{\beta}_j$ coefficients to 1 is considered, as for the missing data corresponding to minimum $Y$ values, it isn't observed any method, but for the missing data corresponding to maximum $Y$ values, it is observed that CC gives the better results. For both cases, MI(10) maximizes the mean square error and CC maximizes the standard errors of regression coefficients.

Table 5.12 Summary of results when missing proportion 9% and n=100

| | min($\lvert Y \rvert$) | | | | max($\lvert Y \rvert$) | | | |
|---|---|---|---|---|---|---|---|---|
| Methods | $E(\hat{\beta}_0)$<br>$E(S_{\hat{\beta}_0})$ | $E(\hat{\beta}_1)$<br>$E(S_{\hat{\beta}_1})$ | $E(\hat{\beta}_2)$<br>$E(S_{\hat{\beta}_2})$ | $E(MSE)$ | $E(\hat{\beta}_0)$<br>$E(S_{\hat{\beta}_0})$ | $E(\hat{\beta}_1)$<br>$E(S_{\hat{\beta}_1})$ | $E(\hat{\beta}_2)$<br>$E(S_{\hat{\beta}_2})$ | $E(MSE)$ |
| CC | 1,0010<br>0,0074 | 1,0020<br>0,0073 | 1,0011<br>0,0073 | 0,00481 | 0,9999<br>0,0075 | 0,9996<br>0,0081 | 0,9991<br>0,0082 | 0,00483 |
| PE | 1,0008<br>0,0067 | 1,0018<br>0,0067 | 1,0010<br>0,0067 | 0,00437 | 1,0007<br>0,0067 | 1,0006<br>0,0067 | 1,0001<br>0,0068 | 0,00439 |
| EM | 1,0008<br>0,0067 | 1,0023<br>0,0068 | 1,0010<br>0,0068 | 0,00441 | 1,0007<br>0,0067 | 1,0011<br>0,0067 | 1,0001<br>0,0068 | 0,00439 |
| MI(1) | 1,0015<br>0,0070 | 1,0012<br>0,0071 | 1,0011<br>0,0070 | 0,00472 | 1,0039<br>0,0070 | 1,0042<br>0,0071 | 1,0048<br>0,0071 | 0,00480 |
| MI(2) | 1,0002<br>0,0070 | 1,0025<br>0,0071 | 1,0011<br>0,0070 | 0,00476 | 1,0038<br>0,0071 | 1,0040<br>0,0071 | 1,0040<br>0,0072 | 0,00490 |
| MI(3) | 1,0021<br>0,0069 | 1,0010<br>0,0069 | 1,0008<br>0,0069 | 0,00459 | 0,9990<br>0,0069 | 0,9973<br>0,0070 | 0,9980<br>0,0070 | 0,00473 |
| MI(4) | 0,9981<br>0,0070 | 1,0034<br>0,0071 | 1,0021<br>0,0071 | 0,00480 | 0,9991<br>0,0069 | 0,9984<br>0,0069 | 0,9988<br>0,0070 | 0,00467 |
| MI(5) | 0,9998<br>0,0075 | 1,0012<br>0,0076 | 1,0020<br>0,0076 | 0,00552 | 1,0002<br>0,0069 | 0,9976<br>0,0069 | 0,9986<br>0,0070 | 0,00471 |
| MI(6) | 1,0008<br>0,0071 | 1,0021<br>0,0072 | 1,0008<br>0,0071 | 0,00490 | 0,9992<br>0,0069 | 0,9999<br>0,0069 | 0,9989<br>0,0070 | 0,00463 |
| MI(7) | 1,0038<br>0,0071 | 1,0007<br>0,0071 | 0,9993<br>0,0071 | 0,00485 | 1,0020<br>0,0069 | 1,0030<br>0,0070 | 1,0004<br>0,0070 | 0,00471 |
| MI(8) | 0,9998<br>0,0071 | 1,0016<br>0,0072 | 1,0021<br>0,0071 | 0,00491 | 0,9990<br>0,0072 | 0,9961<br>0,0072 | 0,9977<br>0,0073 | 0,00515 |
| MI(9) | 0,9996<br>0,0069 | 1,0032<br>0,0070 | 1,0010<br>0,0070 | 0,00470 | 0,9948<br>0,0071 | 0,9937<br>0,0071 | 0,9926<br>0,0072 | 0,00491 |
| MI(10) | 0,9965<br>0,0071 | 1,0031<br>0,0072 | 1,0038<br>0,0072 | 0,00494 | 1,0010<br>0,0075 | 1,0017<br>0,0075 | 0,9993<br>0,0076 | 0,00548 |

As for Table 5.12, PE gives the lowest MSE as for the missing data corresponding to minimum $Y$ values. EM is following this method. In relation to this, the smallest values of $S_{\hat{\beta}_j}$ are obtained from PE. PE and EM give the lowest MSE as for the missing data corresponding to maximum $Y$ values. In relation to this, the smallest values of $S_{\hat{\beta}_j}$ are obtained from these two methods. When the nearness of $\hat{\beta}_j$ coefficients to 1 is considered, it isn't observed any method for both cases. MI(5) maximizes the MSE and $S_{\hat{\beta}_j}$ as for the missing data corresponding to minimum $Y$ values. MI(10) maximizes the MSE and CC maximizes $S_{\hat{\beta}_j}$ as for the missing data corresponding to maximum $Y$ values.

Table 5.13 Summary of results when missing proportion 12% and n=100

| | min($\|Y\|$) | | | | max($\|Y\|$) | | | |
|---|---|---|---|---|---|---|---|---|
| Methods | $E(\hat{\beta}_0)$<br>$E(S_{\hat{\beta}_0})$ | $E(\hat{\beta}_1)$<br>$E(S_{\hat{\beta}_1})$ | $E(\hat{\beta}_2)$<br>$E(S_{\hat{\beta}_2})$ | $E(MSE)$ | $E(\hat{\beta}_0)$<br>$E(S_{\hat{\beta}_0})$ | $E(\hat{\beta}_1)$<br>$E(S_{\hat{\beta}_1})$ | $E(\hat{\beta}_2)$<br>$E(S_{\hat{\beta}_2})$ | $E(MSE)$ |
| CC | 1,0010<br>0,0075 | 1,0021<br>0,0074 | 1,0006<br>0,0075 | 0,00488 | 0,9994<br>0,0078 | 0,9992<br>0,0085 | 0,9983<br>0,0084 | 0,00484 |
| PE | 1,0007<br>0,0066 | 1,0019<br>0,0066 | 1,0005<br>0,0067 | 0,00427 | 1,0005<br>0,0066 | 1,0005<br>0,0066 | 0,9996<br>0,0066 | 0,00425 |
| EM | 1,0007<br>0,0067 | 1,0025<br>0,0067 | 1,0005<br>0,0067 | 0,00432 | 1,0005<br>0,0066 | 1,0011<br>0,0066 | 0,9996<br>0,0066 | 0,00425 |
| MI(1) | 0,9997<br>0,0070 | 1,0017<br>0,0070 | 1,0018<br>0,0071 | 0,00477 | 1,0020<br>0,0069 | 1,0010<br>0,0070 | 1,0016<br>0,0070 | 0,00474 |
| MI(2) | 1,0024<br>0,0070 | 1,0017<br>0,0071 | 0,9993<br>0,0071 | 0,00484 | 1,0067<br>0,0071 | 1,0089<br>0,0072 | 1,0072<br>0,0072 | 0,00501 |
| MI(3) | 0,9975<br>0,0073 | 1,0036<br>0,0074 | 1,0015<br>0,0074 | 0,00526 | 0,9956<br>0,0072 | 0,9944<br>0,0072 | 0,9930<br>0,0073 | 0,00514 |
| MI(4) | 1,0032<br>0,0073 | 1,0014<br>0,0074 | 0,9984<br>0,0074 | 0,00527 | 1,0051<br>0,0074 | 1,0053<br>0,0074 | 1,0032<br>0,0074 | 0,00535 |
| MI(5) | 1,0023<br>0,0073 | 1,0007<br>0,0073 | 0,9999<br>0,0074 | 0,00518 | 0,9955<br>0,0073 | 0,9932<br>0,0072 | 0,9929<br>0,0073 | 0,00519 |
| MI(6) | 1,0017<br>0,0069 | 1,0012<br>0,0070 | 1,0005<br>0,0070 | 0,00470 | 1,0003<br>0,0069 | 0,9999<br>0,0069 | 0,9994<br>0,0069 | 0,00470 |
| MI(7) | 1,0025<br>0,0069 | 1,0011<br>0,0069 | 1,0000<br>0,0070 | 0,00465 | 1,0049<br>0,0069 | 1,0059<br>0,0070 | 1,0058<br>0,0069 | 0,00470 |
| MI(8) | 0,9997<br>0,0070 | 1,0029<br>0,0070 | 1,0007<br>0,0070 | 0,00475 | 1,0026<br>0,0070 | 1,0047<br>0,0070 | 1,0027<br>0,0070 | 0,00478 |
| MI(9) | 1,0008<br>0,0071 | 1,0031<br>0,0071 | 0,9993<br>0,0071 | 0,00486 | 0,9999<br>0,0070 | 1,0009<br>0,0071 | 0,9974<br>0,0071 | 0,00485 |
| MI(10) | 1,0016<br>0,0069 | 1,0017<br>0,0069 | 1,0001<br>0,0069 | 0,00457 | 0,9991<br>0,0068 | 0,9986<br>0,0068 | 0,9977<br>0,0068 | 0,00450 |

As for Table 5.13, PE give the lowest MSE as for the missing data corresponding to minimum $Y$ values. EM is following this method. In relation to this, the smallest values of $S_{\hat{\beta}_j}$ are obtained from PE. PE and EM give the lowest MSE as for the missing data corresponding to maximum $Y$ values. In relation to this, the smallest values of $S_{\hat{\beta}_j}$ are obtained from these two methods. When the nearness of $\hat{\beta}_j$ coefficients to 1 is considered, it isn’t observed any method for both cases. MI(3) and MI(4) maximize the MSE as for the missing data corresponding to minimum $Y$ values and MI(4) maximizes MSE as for the missing data corresponding to maximum $Y$ values. For both cases, CC maximizes $S_{\hat{\beta}_j}$.

Table 5.14 Summary of results when missing proportion 15% and n=100

| | min(\| $Y$ \|) | | | | max(\| $Y$ \|) | | | |
|---|---|---|---|---|---|---|---|---|
| Methods | $E(\hat{\beta}_0)$<br>$E(S_{\hat{\beta}_0})$ | $E(\hat{\beta}_1)$<br>$E(S_{\hat{\beta}_1})$ | $E(\hat{\beta}_2)$<br>$E(S_{\hat{\beta}_2})$ | $E(MSE)$ | $E(\hat{\beta}_0)$<br>$E(S_{\hat{\beta}_0})$ | $E(\hat{\beta}_1)$<br>$E(S_{\hat{\beta}_1})$ | $E(\hat{\beta}_2)$<br>$E(S_{\hat{\beta}_2})$ | $E(MSE)$ |
| CC | 1,0010<br>0,0077 | 1,0021<br>0,0075 | 1,0006<br>0,0076 | 0,00483 | 0,9994<br>0,0080 | 0,9984<br>0,0090 | 0,9989<br>0,0089 | 0,00486 |
| PE | 1,0007<br>0,0064 | 1,0019<br>0,0065 | 1,0005<br>0,0065 | 0,00408 | 1,0009<br>0,0065 | 1,0002<br>0,0065 | 1,0006<br>0,0065 | 0,00412 |
| EM | 1,0007<br>0,0065 | 1,0025<br>0,0065 | 1,0005<br>0,0066 | 0,00413 | 1,0011<br>0,0065 | 1,0012<br>0,0065 | 1,0008<br>0,0066 | 0,00415 |
| MI(1) | 0,9997<br>0,0070 | 1,0017<br>0,0070 | 1,0018<br>0,0070 | 0,00476 | 1,0052<br>0,0071 | 1,0054<br>0,0071 | 1,0076<br>0,0071 | 0,00491 |
| MI(2) | 1,0024<br>0,0071 | 1,0017<br>0,0071 | 0,9993<br>0,0071 | 0,00486 | 1,0037<br>0,0073 | 1,0035<br>0,0073 | 1,0045<br>0,0074 | 0,00522 |
| MI(3) | 0,9975<br>0,0071 | 1,0036<br>0,0071 | 1,0015<br>0,0071 | 0,00485 | 1,0037<br>0,0072 | 1,0035<br>0,0072 | 1,0045<br>0,0072 | 0,00505 |
| MI(4) | 1,0032<br>0,0072 | 1,0014<br>0,0072 | 0,9984<br>0,0072 | 0,00499 | 0,9994<br>0,0070 | 0,9952<br>0,0070 | 0,9974<br>0,0070 | 0,00479 |
| MI(5) | 1,0023<br>0,0076 | 1,0007<br>0,0076 | 0,9999<br>0,0076 | 0,00559 | 1,0003<br>0,0069 | 0,9988<br>0,0069 | 0,9996<br>0,0070 | 0,00471 |
| MI(6) | 1,0017<br>0,0073 | 1,0012<br>0,0073 | 1,0005<br>0,0074 | 0,00521 | 0,9989<br>0,0074 | 1,0008<br>0,0074 | 0,9977<br>0,0075 | 0,00535 |
| MI(7) | 1,0025<br>0,0072 | 1,0011<br>0,0072 | 1,0000<br>0,0072 | 0,00504 | 1,0048<br>0,0070 | 1,0061<br>0,0071 | 1,0075<br>0,0071 | 0,00485 |
| MI(8) | 0,9997<br>0,0069 | 1,0029<br>0,0069 | 1,0007<br>0,0069 | 0,00462 | 1,0009<br>0,0071 | 1,0002<br>0,0071 | 0,9996<br>0,0072 | 0,00495 |
| MI(9) | 1,0008<br>0,0069 | 1,0031<br>0,0069 | 0,9993<br>0,0069 | 0,00458 | 0,9988<br>0,0070 | 0,9988<br>0,0070 | 0,9997<br>0,0070 | 0,00475 |
| MI(10) | 1,0016<br>0,0068 | 1,0017<br>0,0068 | 1,0001<br>0,0069 | 0,00456 | 0,9978<br>0,0068 | 0,9971<br>0,0068 | 0,9952<br>0,0068 | 0,00450 |

As for Table 5.14, PE gives the lowest MSE. In relation to this, the smallest values of $S_{\hat{\beta}_j}$ are obtained from this method. The closest result to this method is obtained from EM. When the nearness of $\hat{\beta}_j$ coefficients to 1 is considered, as for the missing data corresponding to minimum $Y$ values, it isn't observed any method and as for the missing data corresponding to maximum $Y$ values, it is observed that MI(5) and MI(8) give the better results. MI(5) maximizes the MSE as for the missing data corresponding to minimum $Y$ values and MI(6) maximizes the MSE as for the missing data corresponding to maximum $Y$ values. For both cases, CC maximizes the $S_{\hat{\beta}_j}$.

The matrix plots of mean square error are given from Figure 5.1 to 5.6.

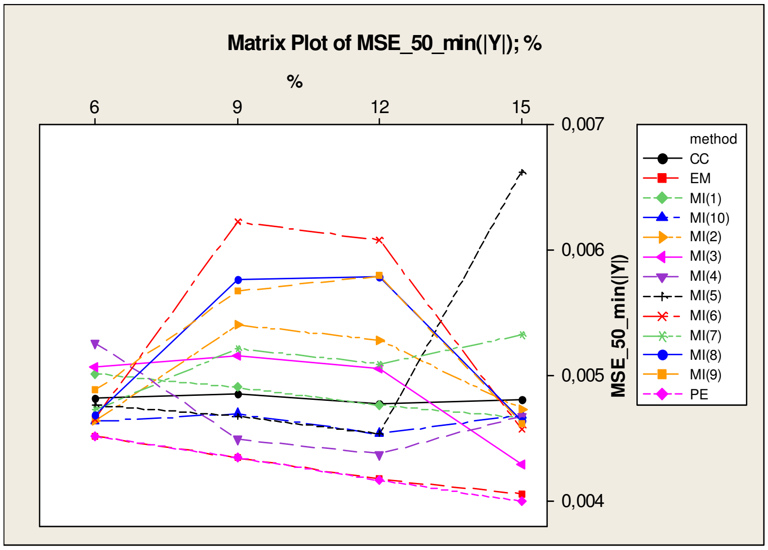


Figure 5.1 The plot of MSE values for $\min(|Y|)$ when n=50

According to the matrix plot in Figure 5.1, EM and PE give the lowest MSE. MI(2), MI(3), MI(6), MI(8) and MI(9) give the parallel results for different missing proportions. MI(5) and MI(7) give the worst results respectively when missing proportion is 15%. MI(1), MI(4) and MI(10) are similar to CC.

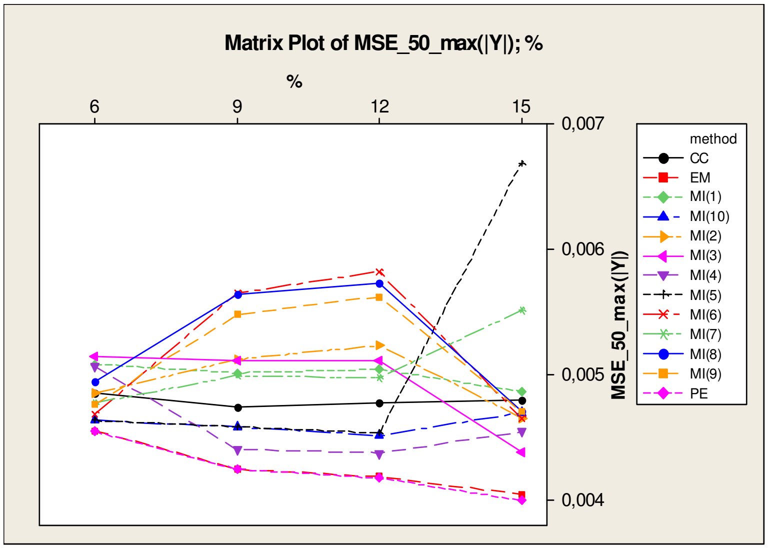


Figure 5.2 The plot of MSE values for $\max(|Y|)$ when n=50

As for Figure 5.2, EM and PE give the lowest MSE. MI(2), MI(3), MI(6), MI(8) and MI(9) give the similar results. MI(5) and MI(7) give the worst results when missing proportion is 15%. On the other hand, closest results are obtained from CC, MI(1), MI(4) and MI(10).

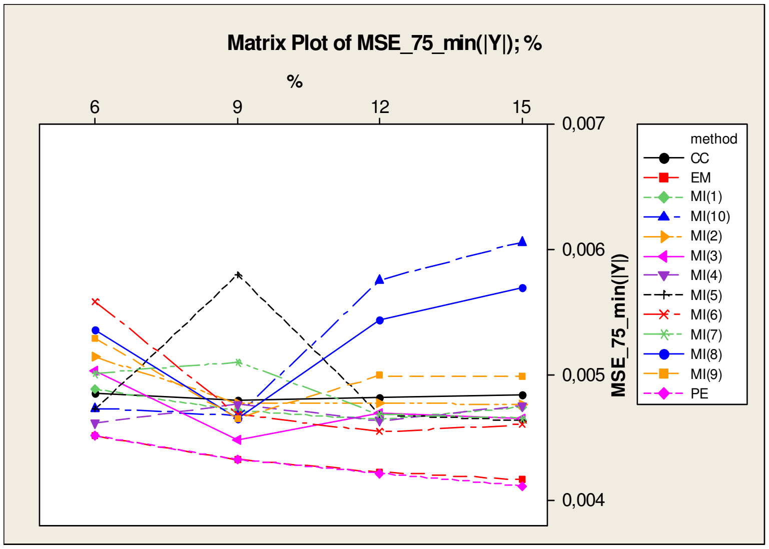


Figure 5.3 The plot of MSE values for min(| $Y$ |) when n=75

According to the matrix plot in Figure 5.3, EM and PE give the lowest MSE. MI(8) and MI(10) give the worst results at missing proportions of 12% and 15%. MI(5) and MI(7) give the worst results when missing proportion is 9%. The other methods are resemble each other.

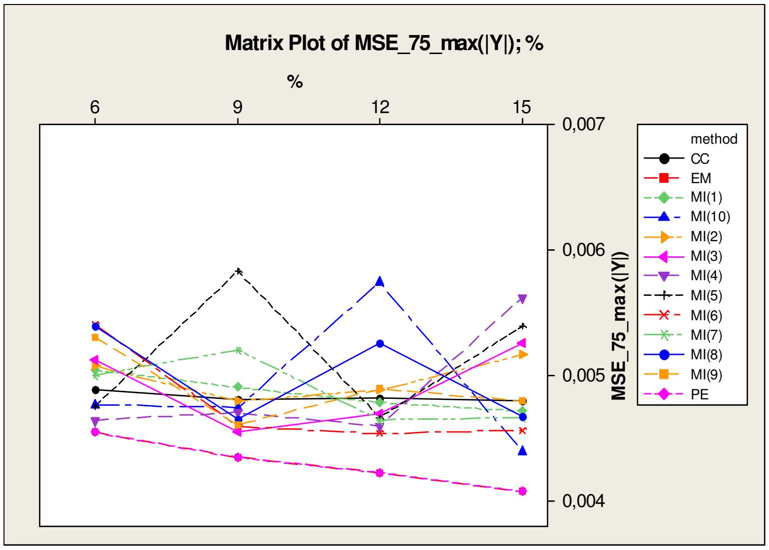


Figure 5.4 The plot of MSE values for max(| $Y$ |) when n=75

According to the matrix plot in Figure 5.4, EM and PE give the lowest MSE. MI(5) and MI(7) give the worst results when missing proportion is 9%. MI(8) and MI(10) give the worst results at missing proportion of 12%. MI(1) and MI(9) are similar to CC.

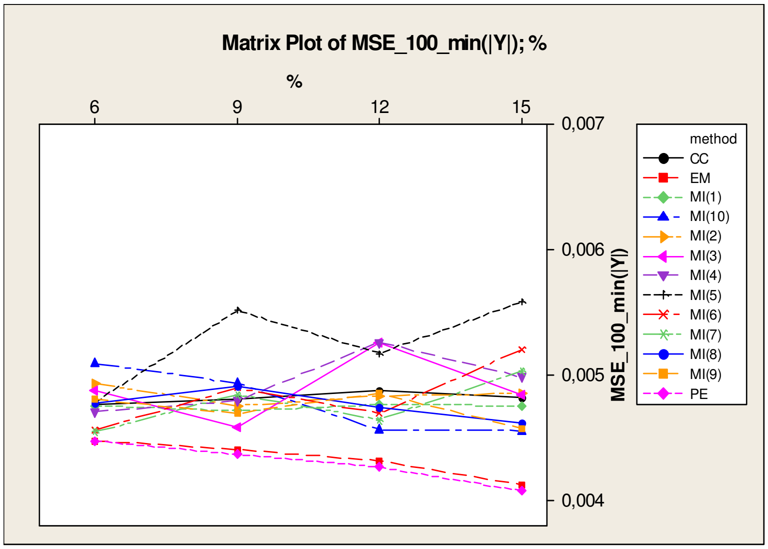


Figure 5.5 The plot of MSE values for $\min(|Y|)$ when n=100

According to the matrix plot in Figure 5.5, PE gives the lowest MSE. MI(3) and MI(4) give the worst results at missing proportion of 12%. MI(5) give the worst result when missing proportions are 9% and 15% . There is a resemblance between the remaining methods.

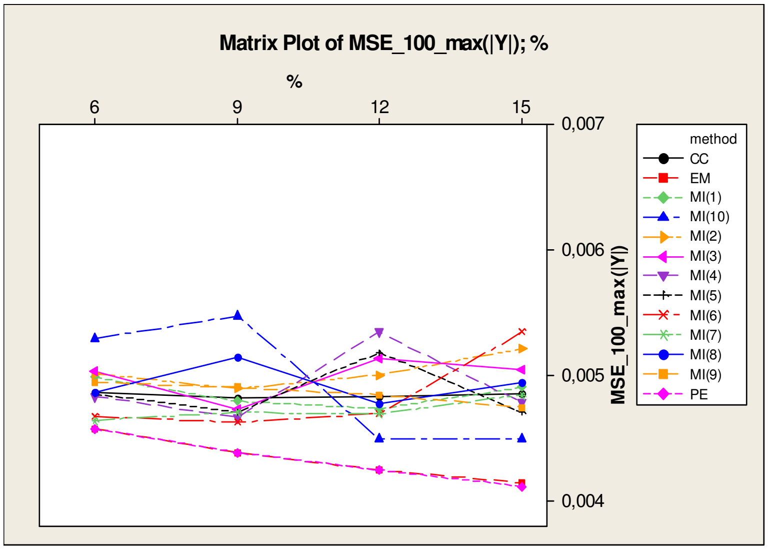


Figure 5.6 The plot of MSE values for $\max(|Y|)$ when n=100

According to the matrix plot in Figure 5.6, EM and PE give the lowest MSE. MI(10) gives the worst results when missing proportions are 6% and 9%. On the contrary MI(10) gives the closest result to PE and EM at missing proportions of 12% and 15%. Whereas CC and MI(1) give the closest results.

Matrix plots of $S_{\hat{\beta}_1}$ are given from Figure 5.7 to Figure 5.12. Matrix plots of $S_{\hat{\beta}_0}$ and $S_{\hat{\beta}_2}$ are given in Appendix C.

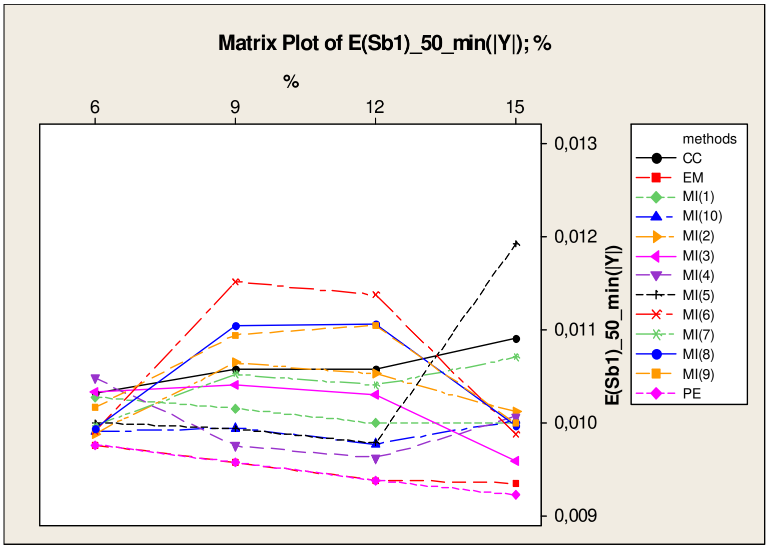


Figure 5.7 The plot of $E(S_{\hat{\beta}_1})$ values for $\min(|Y|)$ when n=50

As for Figure 5.7, the smallest values of $S_{\hat{\beta}_1}$ are obtained from PE and EM. MI(6) gives the worst results when missing proportions are 9% and 12%. MI(5) gives the worst result at missing proportion of 15%.

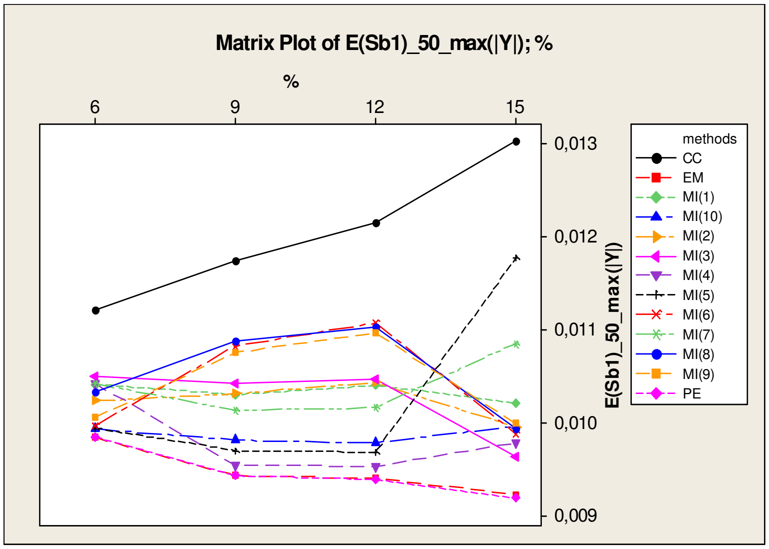


Figure 5.8 The plot of $E(S_{\hat{\beta}_1})$ values for $\max(|Y|)$ when n=50

As for Figure 5.8, the smallest values of $S_{\hat{\beta}_1}$ are obtained from PE and EM. CC gives the worst results for all missing proportions.

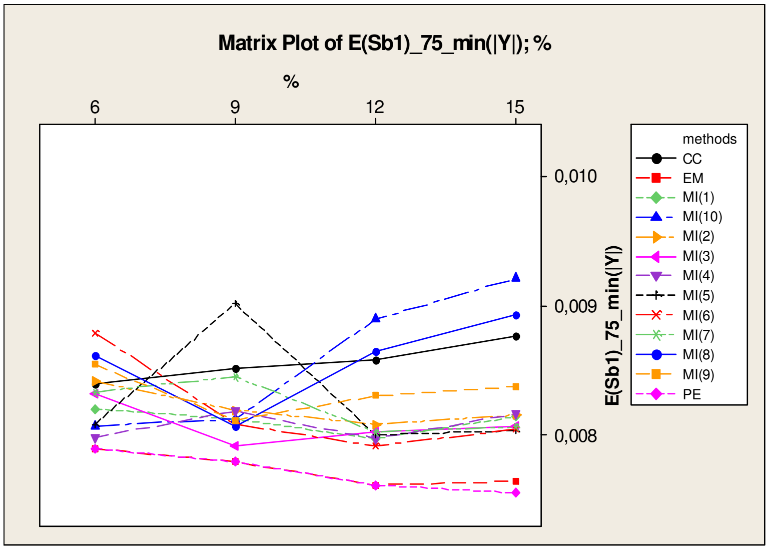


Figure 5.9 The plot of $E(S_{\hat{\beta}_1})$ values for $\min(|Y|)$ when n=75

As for Figure 5.9, the smallest values of $S_{\hat{\beta}_1}$ are obtained from PE and EM. MI(5) gives the worst result when missing proportion is 9%. MI(10) gives the worst results at missing proportions of 12% and 15%.

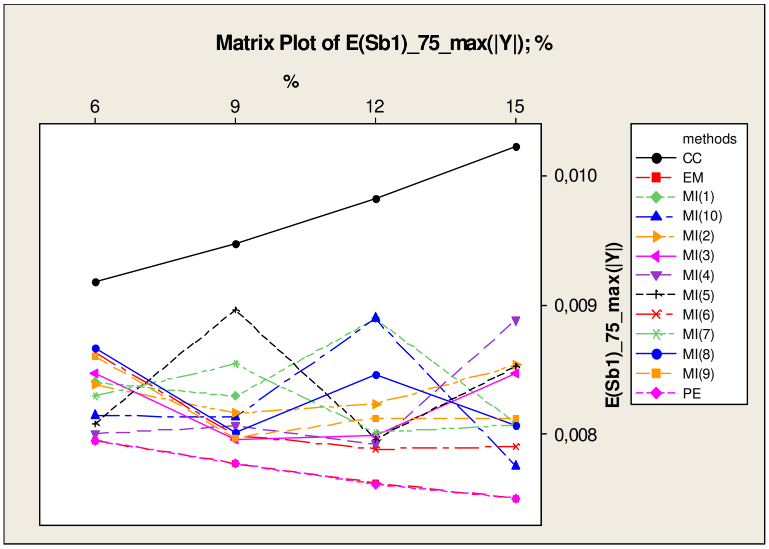


Figure 5.10 The plot of $E(S_{\hat{\beta}_1})$ values for $\max(|Y|)$ when n=75

As for Figure 5.10, the smallest values of $S_{\hat{\beta}_1}$ are obtained from PE and EM. CC gives the worst results for all missing proportions.

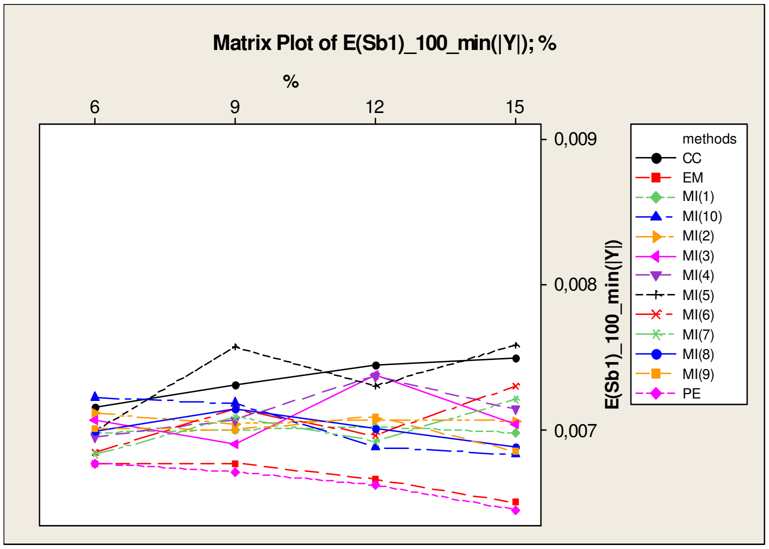


Figure 5.11 The plot of $E(S_{\hat{\beta}_1})$ values for $\min(|Y|)$ when n=100

As for Figure 5.11, the smallest values of $S_{\hat{\beta}_1}$ are obtained from PE followed by EM. MI(5) gives the worst result when missing proportions are 9% and 15%. CC gives the worst result at missing proportion of 12%.

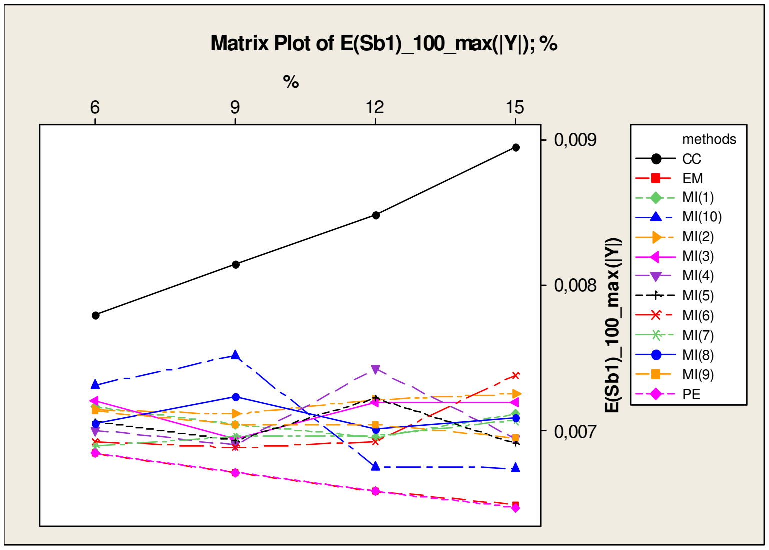


Figure 5.12 The plot of $E(S_{\hat{\beta}_1})$ values for $\max(|Y|)$ when n=100

As for Figure 5.12, the smallest values of $S_{\hat{\beta}_1}$ are obtained from PE and EM. CC gives the worst results for all missing proportions.

Matrix plots of $E(\hat{\beta}_1)$ are given from Figure 5.13 to Figure 5.18. Matrix plots of $E(\hat{\beta}_0)$ and $E(\hat{\beta}_2)$ are given in Appendix D.

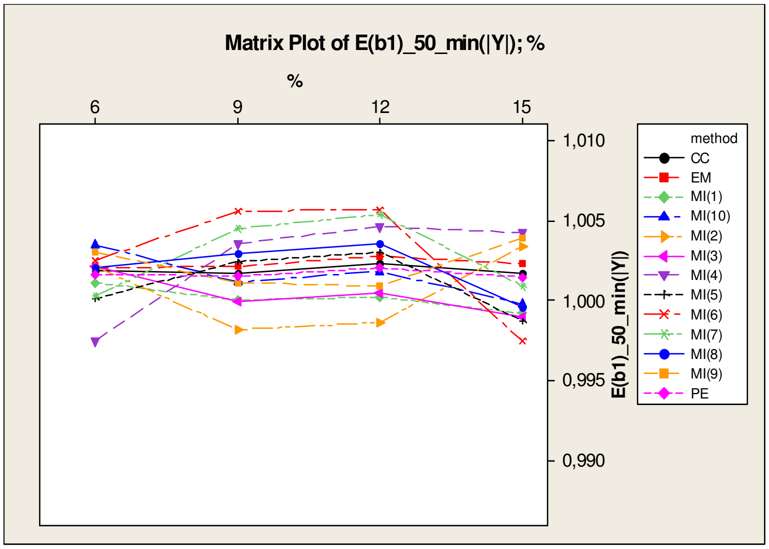


Figure 5.13 The plot of $E(\hat{\beta}_1)$ values for min(| $Y$ |) when n=50

As for Figure 5.13, $\hat{\beta}_1$ coefficient at MI(1) and MI(3) are found approximately as 1. PE, EM and CC are alike among themselves. MI(6) gives the worst results at missing proportions of 9%, 12% and 15%.

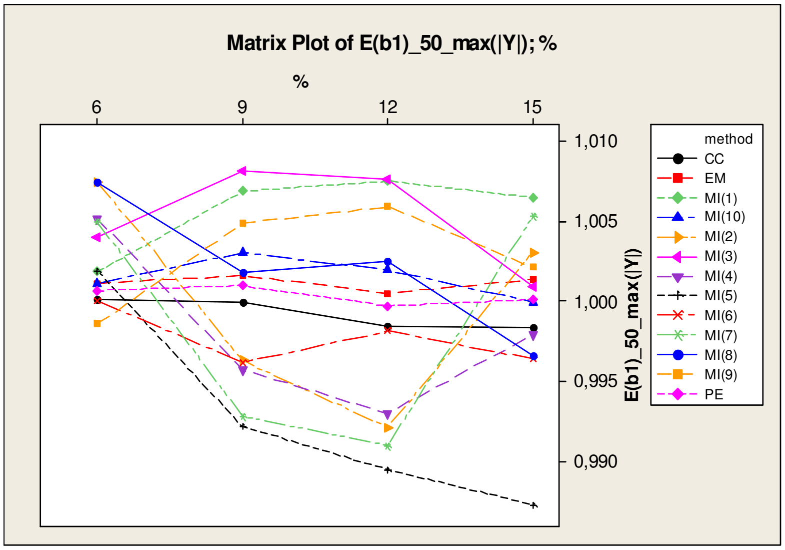


Figure 5.14 The plot of $E(\hat{\beta}_1)$ values for max(| $Y$ |) when n=50

As for Figure 5.14, $\hat{\beta}_1$ coefficient at PE is found approximately as 1. MI(5) gives the worst results at missing proportions of 9%, 12% and 15%.

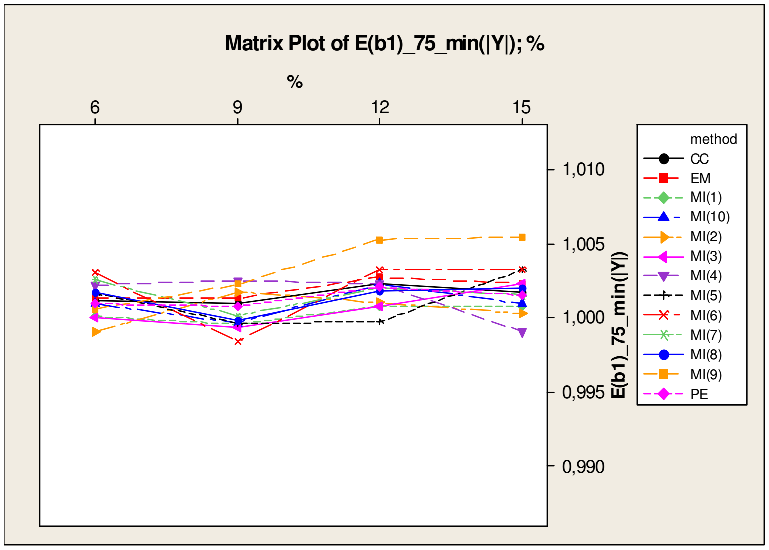


Figure 5.15 The plot of $E(\hat{\beta}_1)$ values for $\min(|Y|)$ when n=75

As for Figure 5.15, all methods except MI(9) give the parallel results for $\hat{\beta}_1$ coefficient found approximately as 1. MI(9) gives the worst results at missing proportions of 12% and 15%.

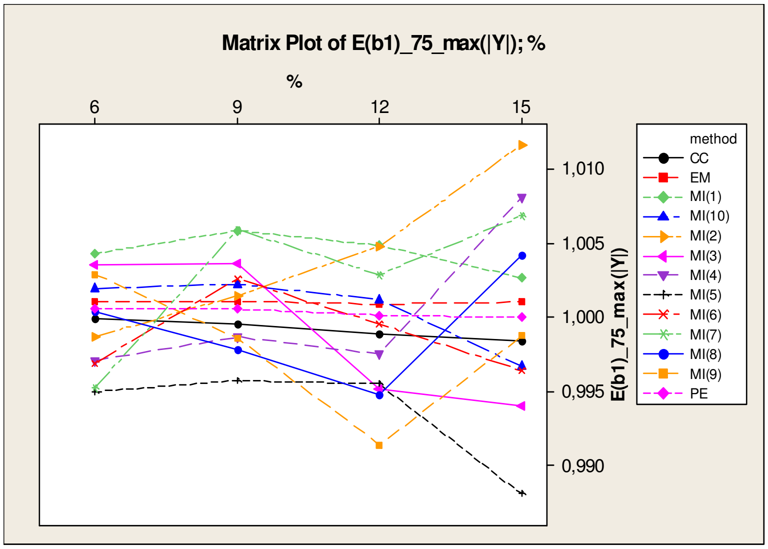


Figure 5.16 The plot of $E(\hat{\beta}_1)$ values for $\max(|Y|)$ when n=75

As for Figure 5.16, $\hat{\beta}_1$ coefficient at PE is found approximately as 1. There is an alikeness between the PE, EM and CC. MI(2) and MI(5) give the worst results at missing proportion of 15%.

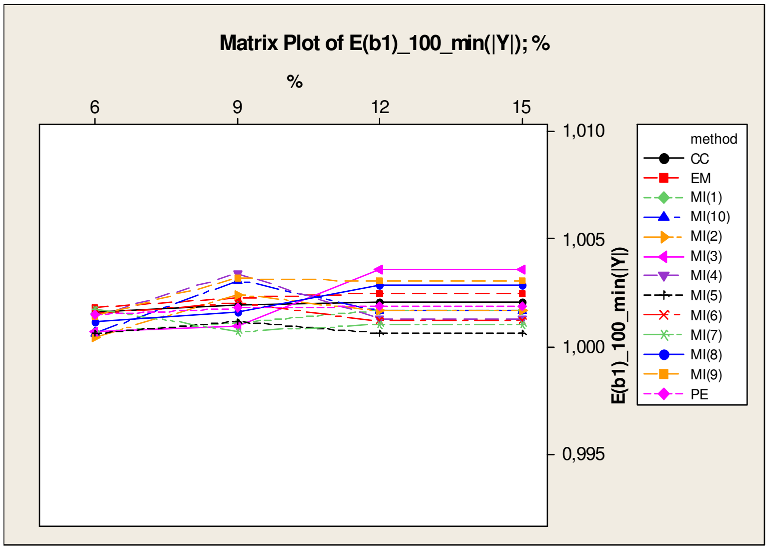


Figure 5.17 The plot of $E(\hat{\beta}_1)$ values for $\min(|Y|)$ when n=100

As for Figure 5.17, all methods except MI(3) give the similar results for $\hat{\beta}_1$ coefficient found approximately as 1. MI(3) gives the worst result at missing proportions of 12% and 15%.

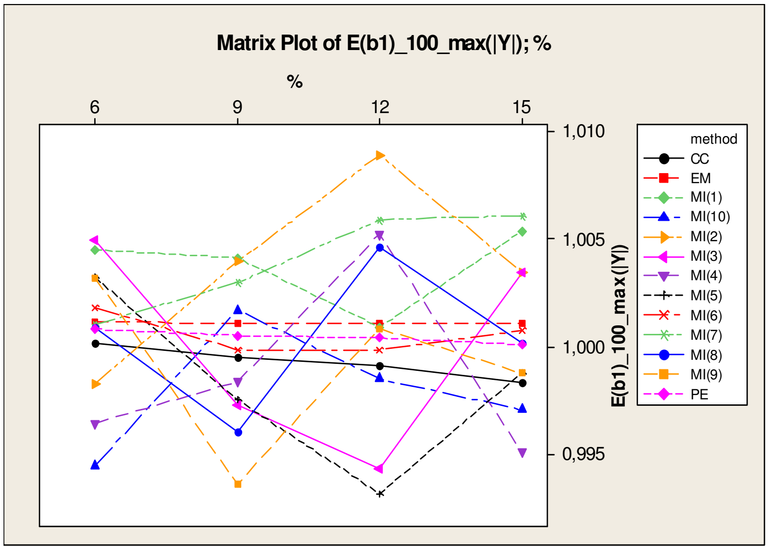


Figure 5.18 The plot of $E(\hat{\beta}_1)$ values for $\max(|Y|)$ when n=100

As for Figure 5.18, $\hat{\beta}_1$ coefficient at PE is found approximately as 1. MI(9) gives the worst result at missing proportion of 9% and MI(2) gives the worst result at missing proportion of 12%.

# CHAPTER SIX
# CONCLUSIONS

Missing data represent a general problem in many scientific fields. Records coming from surveys, physical experiments, and a secondary sources often show some missing data. The impact of the missing data on the results of statistical analysis depends on the mechanism that made the data to be missing and the way in which the data analyst deals with them. In scientific literature this problem has been investigated only recently and almost always with reference to population surveys, market research and census problem.

In this thesis, missing data mechanism are assumed as MAR. That is the missing data mechanism does not depend on the set of missing values though it may possibly depend on the set of observed values. Then the missing data mechanism is said to be ignorable (Little and Rubin, 1987). The current methods for dealing with missing values such as MI and EM algorithm are chosen from the model-based methods. In this thesis, two different simulation study are applied. The first is conducted to compare with the methods MI and EM algorithm. Before the second simulation study, we consider the protective estimator which is used method of moments approach to obtain the regression parameters and the variance. In addition, a simulation study is carried out to verify the characteristics of the MI, EM algorithm and PE.

From the first simulation study, when $X$ matrix generated from the multivariate normal distribution $MN(O, I_p)$ and all parameters of regression coefficients are assigned to 1 with $\varepsilon_i \sim N(0,1)$, we have seen that EM algorithm is given the minimum mean square error and mean of the $\hat{\beta}_j$ are close to 1 when the missing proportion 12% and 36% for symmetric and skewed data. MI(5) is given the minimum MSE when missing proportion 24% for symmetric data and for skewed data MI(10) is given. Consequently, when the assumption is not valid, EM algorithm is not affected, but imputations should be increased for Multiple Imputation.

In the second simulation study, it is necessary to assume that the outcome variable and one of the independent variable have approximate bivariate normal distributions, conditional on the remaining independent variable. That missing data is restricted to the independent variable and that the outcome variable and the remaining independent variable are fully observed.

In this study, C# code is named as PEA that is improved to calculate of the protective regression coefficients, standard error of regression coefficients and mean square error. The link of this programme is http://kisi.deu.edu.tr/neslihan.ortabas/.

General results for the methods which give the best consequences are summarized from Table 6.1 to 6.4.

Table 6.1 Summary of results for which methods give the lowest MSE for $\min(|Y|)$

| | Missing proportion | | | |
|---|---|---|---|---|
| Sample size | 6% | 9% | 12% | 15% |
| n=50 | EM, PE | EM, PE | PE | PE |
| n=75 | EM, PE | EM, PE | PE | PE |
| n=100 | EM, PE | PE | PE | PE |

According to Table 6.1, PE and EM give the lowest average of MSE as for the missing data corresponding to minimum $Y$ values. When the missing proportion and sample size increase, PE gives the lower average of MSE than EM. As a result it can be said that PE is the best for all cases.

Table 6.2 Summary of results for which methods give the lowest MSE for $\max(|Y|)$

| | Missing proportion | | | |
|---|---|---|---|---|
| Sample size | 6% | 9% | 12% | 15% |
| n=50 | PE | EM, PE | PE | PE |
| n=75 | EM, PE | EM, PE | EM, PE | EM, PE |
| n=100 | EM, PE | EM, PE | EM, PE | PE |

According to Table 6.2, PE and EM give the lowest average of MSE as for the missing data corresponding to maximum $Y$ values. Again for this criteria PE is the best for all cases.

Table 6.3 Summary of results for which methods give $\hat{\beta}_1$ coefficients nearly 1 for $\max(|Y|)$

| | Missing proportion | | | |
|---|---|---|---|---|
| Sample size | 6% | 9% | 12% | 15% |
| n=50 | MI(5), MI(7) | MI(1), MI(3) | MI(1), MI(3) | MI(8),MI(10) |
| n=75 | MI(1), MI(3) | MI(7), MI(8) | MI(5) | MI(2) |
| n=100 | MI(2) | MI(7) | MI(5) | MI(5) |

As regards the results in Table 6.3, $\hat{\beta}_1$ coefficients are found approximately 1 at the method of MI as for the missing data corresponding to minimum $Y$ values.

Table 6.4 Summary of results for which methods give $\hat{\beta}_1$ coefficients nearly 1 for $\min(|Y|)$

| | Missing proportion | | | |
|---|---|---|---|---|
| Sample size | 6% | 9% | 12% | 15% |
| n=50 | CC, MI(6) | CC, PE | PE | PE, MI(10) |
| n=75 | CC | CC, PE | PE, MI(6) | PE |
| n=100 | CC, PE | CC, MI(6) | PE, MI(6) | PE, MI(8) |

As regards the results in Table 6.4, $\hat{\beta}_1$ coefficients are found approximately 1 at the methods of CC and PE as for the missing data corresponding to maximum $Y$ values. In general it can be said that PE is the best approximately all cases.

In the light of the results obtained from the given tables and figures, the following generalizations may be made: The MSE and standard errors of coefficients decreases as the size of the sample increases. As for the missing data corresponding to minimum $Y$ values when the nearness of $\hat{\beta}_j$ coefficients to 1 are taken into account it is close at MI, far at CC. As for the missing data corresponding to maximum $Y$ values as regards the nearness of $\hat{\beta}_j$ coefficients to 1 it is close at PE, far at MI. In multiple imputation, a very small value of imputations 2 and 3 are given the best results for standard errors of coefficients. Other imputations such as 5,6,7 are given inconsistent results. In multiple imputation, a very small value of imputations are usually suffice. When the nearness of $\hat{\beta}_1$ coefficients to 1 is considered, PE method gives the better results than EM method. Most of the time PE and EM methods are given parallel results, but when the missing proportion increases PE gives the better results than EM.

As a result, PE method is the only method among the methods that have been explored. It gives the lowest average of MSE as for the missing data corresponding to minimum $Y$ values when the missing proportion and sample size increase. $\hat{\beta}_1$ coefficients are found approximately 1 at the method of PE as for the missing data corresponding to maximum $Y$ values. As for the missing data corresponding to minimum $Y$ values when the nearness of $\hat{\beta}_j$ coefficients to 1 are taken into account PE gives the consistent results.

## REFERENCES

Affifi, A.A., & Elashoff, R.M. (1996). Missing observations in multivariate statistics: Review of the literature, *J. Am. Statist. Assoc*. 61,595-604

Allison, P. D. (2002). *Missing data*, Sage Publications, USA.

Atkinson, A.C., & Cheng, T-C. (2000). On robust linear regression with incomplete data. *Computational Statistics & Data Analysis*, 33, 361-380.

Demirel, N., & Kurt, S. (2005). RegresyonÇözümlemesinde Kayıp Veri Sorunu. İstatistik Araştırma Dergisi (4), In press

Dempster, A.P., Laird, N.M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society,* Series B (Methodological) , Vol.39, No.1, 1-38.

Gold, M. S., & Bentler, P. M. (2000). Treatment of missing data: A Monte Carlo comparison of RBHDI, iterative stochastic regression imputation, and expectation-maximization. *Structural Equation Modelling* 7, 319–355.

Graham, J. W., Hofer, S. M., & Piccinin, A. M. (1994). Analysis with missing data in drug prevention research. In L. M. Collins & L. A. Seitz (eds). Advances in Data Analysis for Prevention Intervention Research. *NIDA Research Monograph. Series* (#142), Washington, DC: National Institute on Drug Abuse.

Graham, J. W., Hofer, S. M., & Mackinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: an application of maximum likelihood procedures. *Multivariate Behavioral Research* 31: 197–218.

Graham, J. W., Hofer, S. M., Donaldson, S. I., MacKinnon, D. P., & Schafer, J. L. (1997), Analysis with missing data in prevention research. In K. Bryant, M. Windle & S. West (eds). *The Science of Prevention: Methodological Advances from Alcohol and Substance Abuse Research.* Washington, DC: American Psychological Association.

Greenlees, W.S., Reece, J.S., & Zieschang, K.D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *J. Am Statist. Assoc.* 77,251-261

Hartley, H. O., & Hocking, R.R. (1971) The analysis of incomplete data. *Biometrics* 27, 783-808.

Hippel, P. T. V. (2004). *Biases in SPSS 12.0 Missing Value Analysis*, The American Statisticion, May 2004, Vol.58, No.2.

Kromrey, J. D., & Hines, C. V. (1994). Nonrandomly missing data in multiple regression: An empirical comparison of common missing-data treatments. *Educational and Psychological Measurement* 54: 573–593.

Lipsitz, S.R., Molenberghs, G., Fitzmaurice, G.M., & Ibrahim, J.G. (2004). Protective estimator for linear regression with nonignorably missing Gaussian outcomes. *Statistical Modeling*,4,3-17.

Little, R.J.A., & Rubin, D.B. (1983). Incomplete data. *Encyclopedia of the Statistical Sciences* 4, 46-53.

Little, R.J.A., & Rubin, D.B. (1987). *Statistical analysis with missing data*. John Wiley & Sons, Inc., USA.

Little, Roderick J. A., & Donald B. Rubin (1989). The analysis of social science data with missing values. *Sociological Methods and Research* 18: 292-326.

Little, R.J.A., (1992). Regression with missing X's : A review. *Journal of American Statistical Association*, Vol.87, No.420, 1227-1237

Little, R.J.A., & Schenker, N. (1994). Missing data. *In Handbook for Statistical Modeling in the Social and Behavioral Sciences* (G. Arminger, C.C. Clogg and M.E. Sobel, eds.) pp.39-75. New York:

Little, R.J.A (1997). Biostatistical analysis with missing data. *Encyclopedia of Biostatistics* (P.Armitage and T. Calton, eds.) London: Wiley.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data.* (2nd ed.). A John Wiley & Sons, Inc. USA.

McLachlan, G.J., & Krishnan, T. (1997). *The EM algorithm and extensions*. John Wiley & Sons, Inc., USA.

Navarro, J. B., & Losilla, J. M. (2000). Analysis of incomplete data with artificial neural networks: A simulation study. *Psicothema* 12: 503–510.

Orchard, T., & Woodbury, M.A. (1972). A missing information principle: theory and applications. *Proc. 6th Berkeley Symposium on Math. Statist. and Prob.* 1,697-715.

Othuon, L. O. (1999). The accuracy of parameter estimates and coverage probability of population values in regression models upon different treatments of systematically missing data. *Dissertation Abstracts International Section* A 59: 4359.

Pastor, J.B.N. (2003). Methods for the analysis of explanatory linear regression models with missing data not at random. *Quality&Quantity*, 37, 363-376.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, Vol.63, 581-592.

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

Schafer, J. L. (1997). *Analysis of incomplete multivariate data.* Chapman & Hall, USA

Simonoff, J. S. (1988). Regression diagnostics to detect non-random missingness in linear regression. *Technometrics* 30: 205–214.

Wothke, W. (1998). Longitudinal and multi-group modelling with missing data. In T. D. Little, K. U. Schnabel & J. Baumert (eds). Modeling Longitudinal and Multiple Group Data: Practical Issues, Applied Approaches and Specific Examples. Mahwah, NJ: Lawrence Erlbaum Associates.

# APPENDICES

## APPENDIX A – Minitab Macro Program for Protective Estimator

We can extend Minitab's capabilities by writing macros that do sets of Minitab commands for us, or that create new Minitab commands. Minitab macro capabilities fall into two categories. Execs were created for earlier releases of Minitab, but Minitab now has a more robust programming language that allows us to create %macros, which are more powerful and flexible than Execs. There are two kinds of %macros: global macros, which are simple and local macros, a more sophisticated form of macro that allows much more flexibility. Global and local macros share many qualities; they are both invoked by typing %, they end in the file extension of MAC, and they can use many of the same macro statements.

Protective.mac is created for calculated protective regression coefficients, standard error of regression coefficients and mean square error. This program is executed 500 times by itself.

PROTECTIVE.MAC

```
gmacro

protective

retrieve  'C:\Documents and Settings\Desktop\statistics.mtw'  # This is an empty
worksheet for saving coeffcients after the program is finished.

retrieve  'C:\Documents and Settings\Desktop\data.mtw' # This is the data worksheet.
it has 500 different data sets.

let k1=1
let k2=3
let k3=2

while k1<1500
while k2<1500
while k3<1500

regress ck1 1 ck2;         # Regress Y on X2 for the first data set.
coefficients c1510;        # Stores the coefficients  in c1510. (coefficients are θ0 and
                             θ1)
```

```
mse k100.               # Stores the mean square error in k100. (k100=$\sigma_{11}^2$)

regress ck3 2 ck1 ck2;  # Regress $X_1$ on Y and $X_2$  for the first data set.
coefficients c1511;     # Stores the coefficients  in c1511. (coefficients are $\phi_0$,
                          $\phi_1$ and $\phi_2$)
mse k101.               # Stores the mean square error($\sigma_{X_1.YX_2}^2$) as k101.

let k102=k100*c1511(2)              # Stores $\sigma_{12} = \sigma_{11}^2\phi_1$ as k102

let k103=k101+((k102**2)/k100)      # Stores $\sigma_{22}^2 = (\sigma_{X_1.YX_2}^2 + \sigma_{12}^2/\sigma_{11}^2)$ as k103

let k104=c1511(1)+c1511(2)*c1510(1) # Stores $\hat{\gamma}_0 = \hat{\phi}_0 + \hat{\phi}_1\hat{\theta}_0$ as k104
let k105=c1511(3)+c1511(2)*c1510(2) # Stores $\hat{\gamma}_1 = \hat{\phi}_2 + \hat{\phi}_{11}\hat{\theta}_1$ as k105

let k106=k102/k103                  # Stores $\hat{\beta}_1 = \hat{\sigma}_{12}/\hat{\sigma}_{22}^2$ as k106

let k107=c1510(1)-k106*k104         # Stores $\hat{\beta}_0 = \hat{\theta}_0 - \hat{\beta}_1\hat{\gamma}_0$ as k107

let k108=c1510(2)-k106*k105         # Stores $\hat{\beta}_2 = \hat{\theta}_1 - \hat{\beta}_1\hat{\gamma}_1$ as k108

copy k107 k106 k108 c1512           # Stores the coefficients  in c1512.
                                      (coefficients are $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$)

do k4=92:100                        # $X_1$ has a nine missing observation
let ck3(k4)=c1511(1)+c1511(2)*ck1(k4)+c1511(3)*ck2(k4)  # $X_1$ is imputed
enddo

set c1509
100(1)
end

copy c1509 ck3 ck2 m1               # Stores X matrix as m1
transpose m1 m2                     # Stores $X^T$ matrix as m2
multiply m2 m1 m3                   # Stores $X^TX$ matrix as m3
inverse m3 m4                       # Stores $(X^TX)^{-1}$ matrix as m4

let c1513=c1512(1)+c1512(2)*ck3+c1512(3)*ck2   # Stores $\hat{Y}_i$ as c1513

let c1514=ck1-c1513                            # Stores $e_i = Y_i - \hat{Y}_i$ as c1514
let c1515=c1514**2                             # Stores $e_i^2$ as c1515

let k109=(sum(c1515)/(count(c1515)-3))        # Stores MSE as k109
```

```
multiply k109 m4 m5                 # Stores MSE*$(X^TX)^{-1}$  as m5

copy m5 c1520-c1522


let k110=sqrt(c1520(1))             # Stores $S_{\hat{\beta}_0}$ as k110
let k111=sqrt(c1521(2))             # Stores $S_{\hat{\beta}_1}$ as k111
let k112=sqrt(c1522(3))             # Stores $S_{\hat{\beta}_2}$ as k112

copy k107 k106 k108 c1525           # Stores the coefficients  in
                                    c1525. (coefficients are $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$)

copy k109 c1527                     # Stores MSE in c1527
let k115=sqrt(k109)                 # Stores S as k115
copy k115 c1528                     # Stores S in c1528

copy k110-k112 c1526                # Stores $S_{\hat{\beta}_i}$  in c1526


name c1525 'coef'
name c1527 'mse'
name c1528 'S'
name c1526 's_coef'

copy c1525 c1526 c1527 c1528;
after "statistics.mtw";
varnames.

worksheet "data.mtw"

let k1=k1+3  }                      # The program is run for the 2. data
let k2=k2+3  }                      set and then 3. it goes to 500.
let k3=k3+3  }

print k1-k3

endwhile
endwhile
endwhile

endmacro
```

## APPENDIX B – C# Code for Protective Estimator

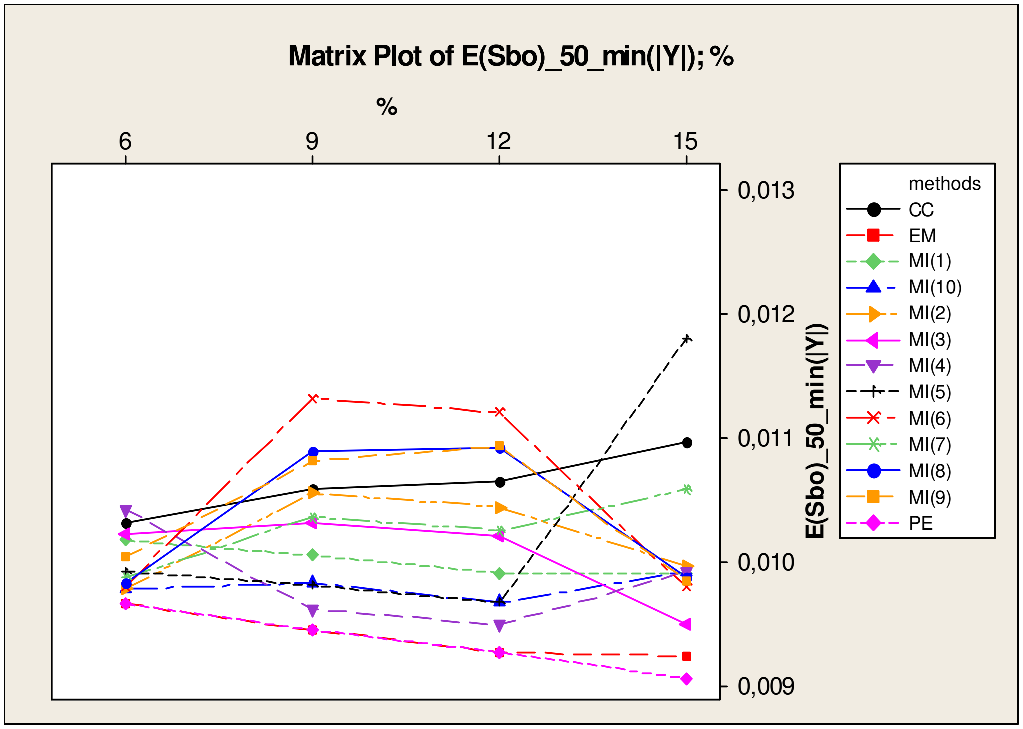
C# (pronounced C sharp) is a programming language designed for building a wide range of enterprise applications that run on the .NET Framework. An evolution of Microsoft C and Microsoft C++, C# is simple, modern, type safe, and object oriented. C# code is compiled as managed code, which means it benefits from the services of the common language runtime. These services include language interoperability, garbage collection, enhanced security, and improved versioning support.

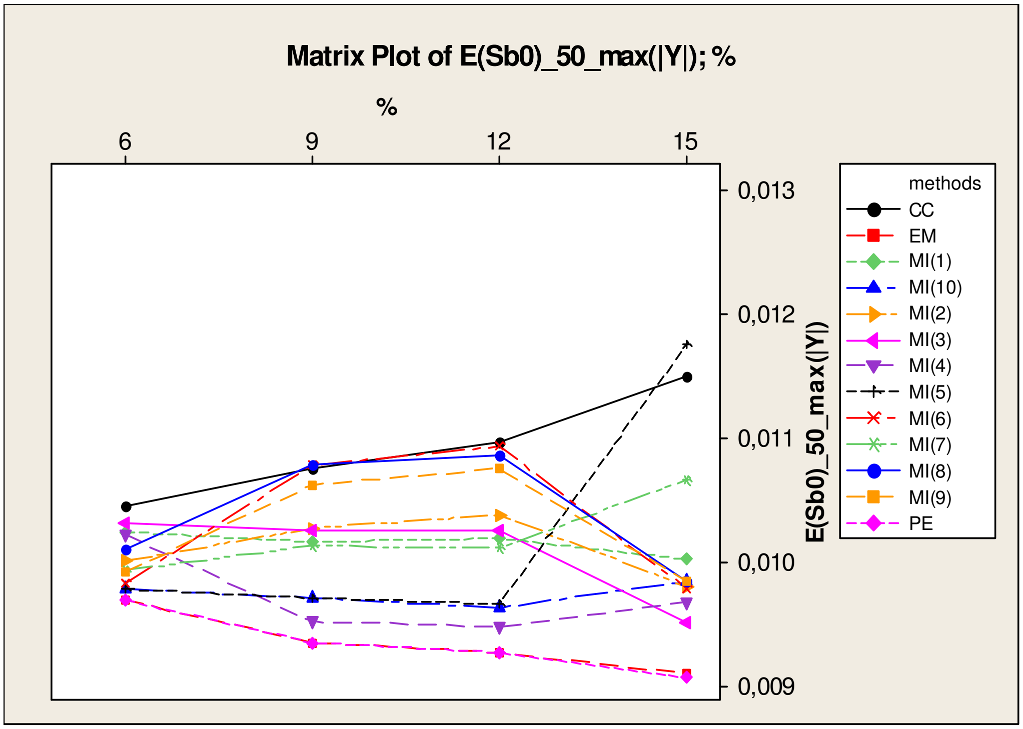
ASP.NET is a set of web application development technologies marketed by Microsoft. Programmers can use it to build dynamic web sites, web applications and XML web services.

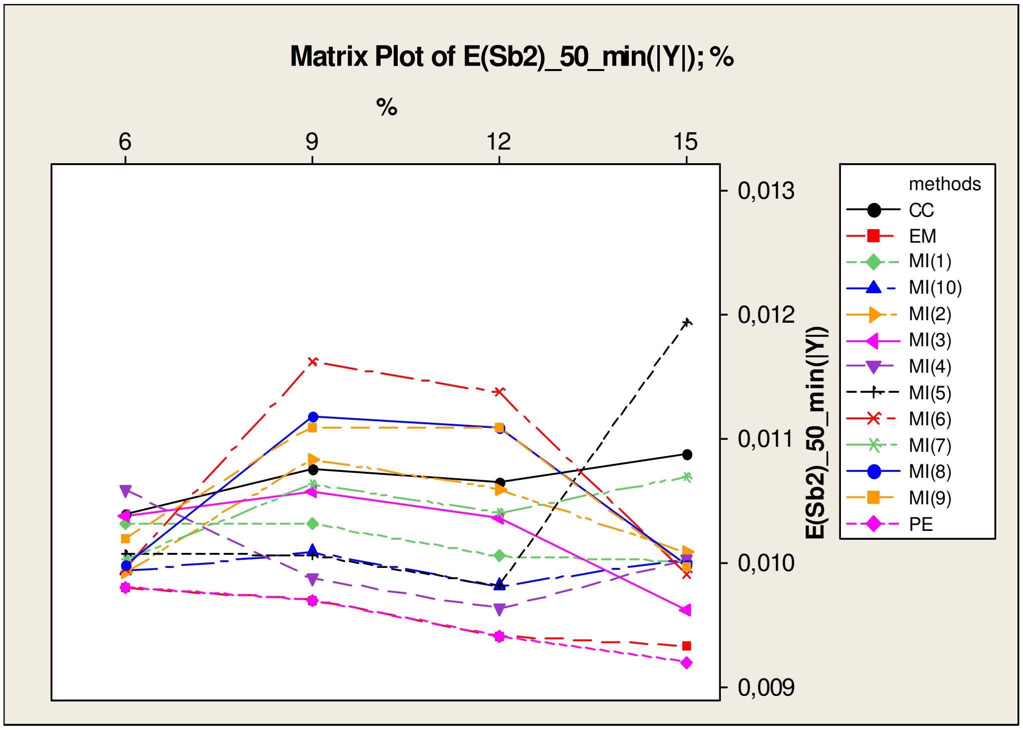
Microsoft .NET is an umbrella term that applies to a wide collection of products and technologies from Microsoft. Most have in common a dependence on the Microsoft .NET Framework, a component of the Windows operating system.

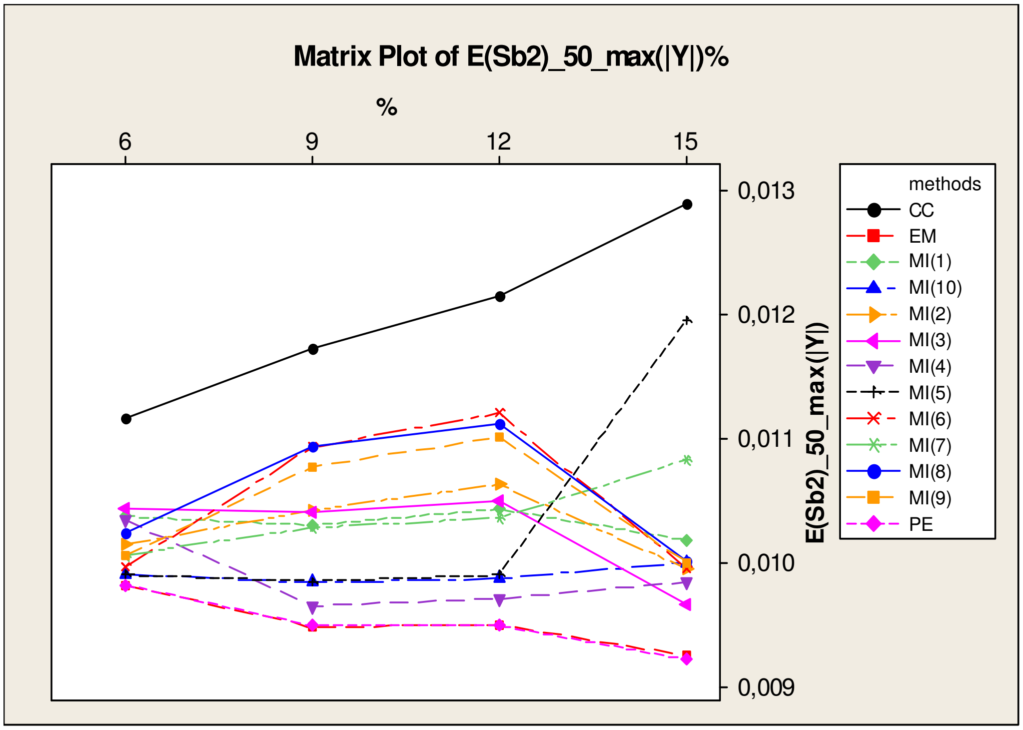
In this thesis, C# code is named as PEA that is improved to calculate of the protective regression coefficients, standard error of regression coefficients and mean square error. The link of this programme is http://kisi.deu.edu.tr/neslihan.ortabas/.

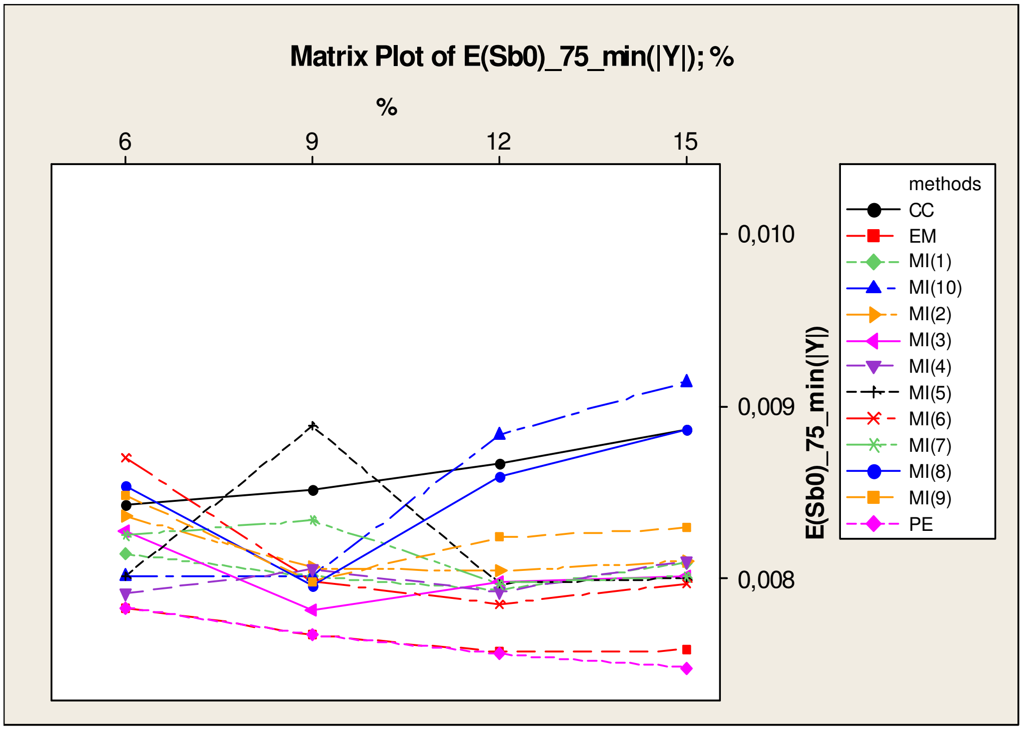
## APPENDIX C – Rest of the Matrix Plots for $E(S_{\hat{\beta}_i})$ ($i = 0,2$)

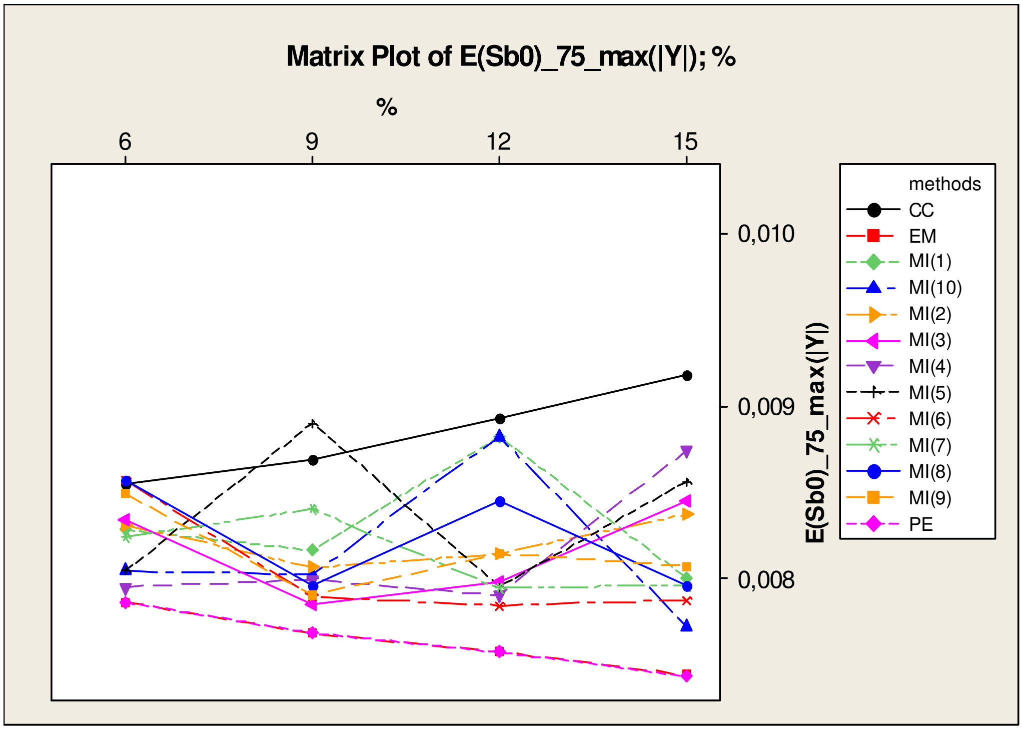


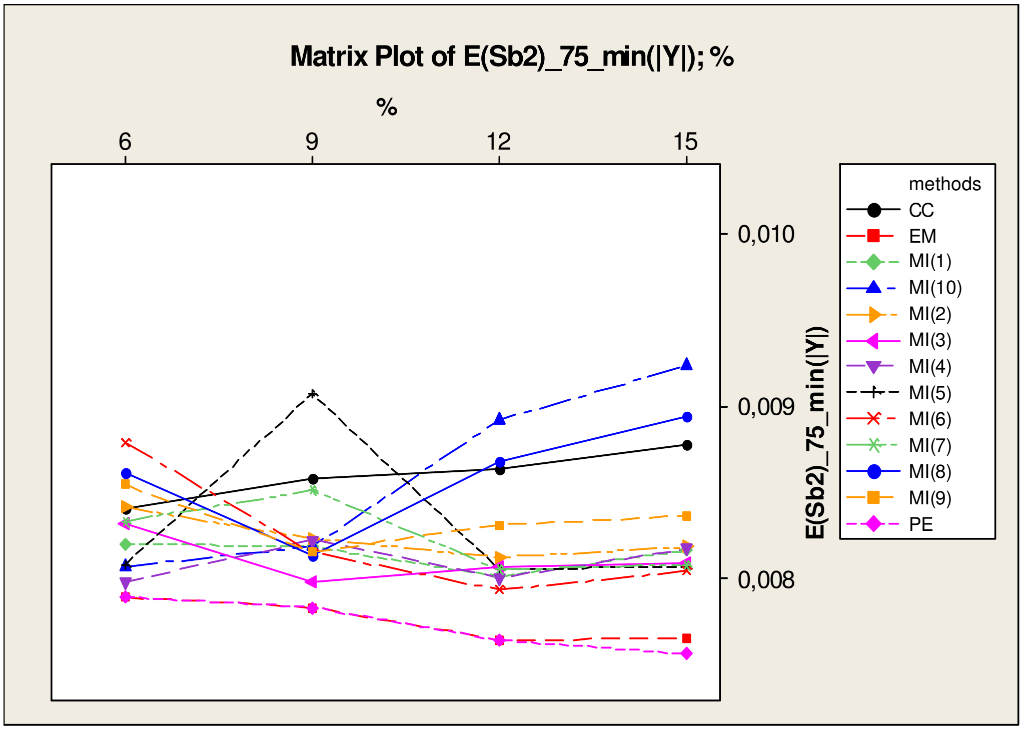


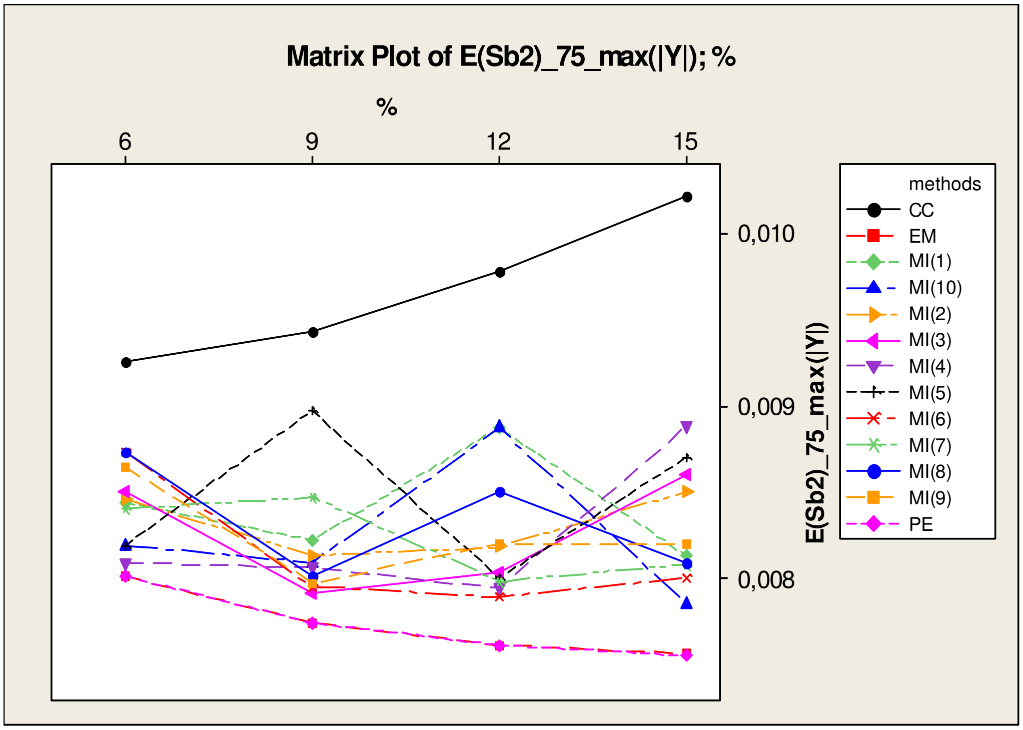

Matrix Plot of E(Sb2)_50_min(|Y|); %
%
6
9
12
15
0,013
0,012
0,011
0,010
0,009
E(Sb2)_50_min(|Y|)
methods
CC
EM
MI(1)
MI(10)
MI(2)
MI(3)
MI(4)
MI(5)
MI(6)
MI(7)
MI(8)
MI(9)
PE

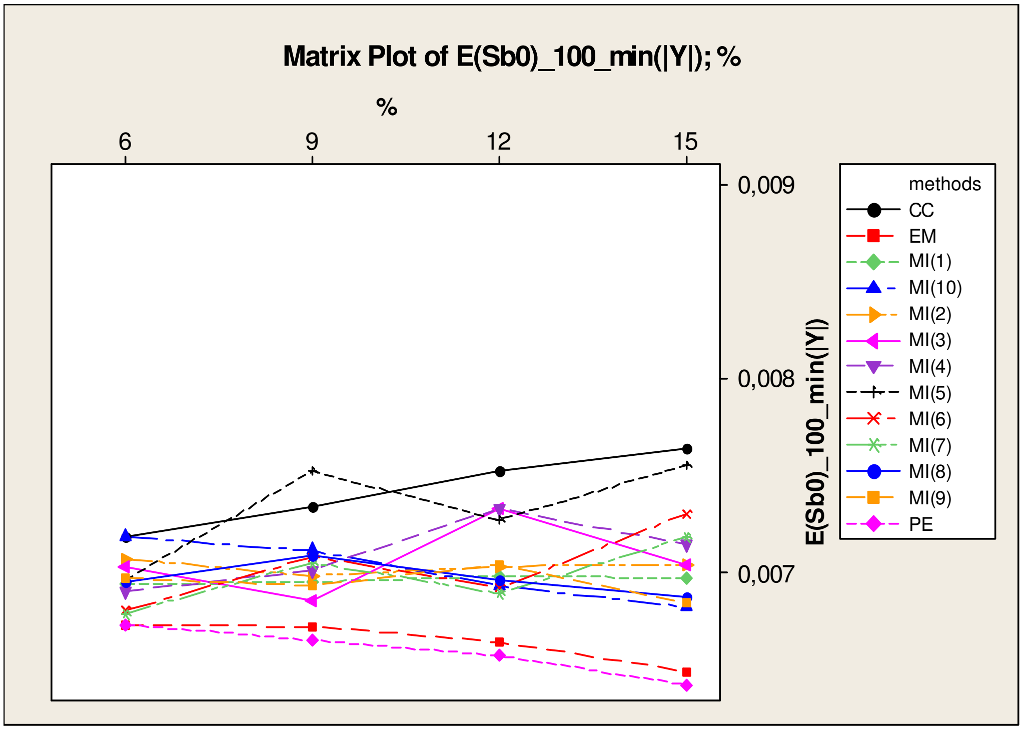

Matrix Plot of E(Sb2)_50_max(|Y|)%
%
6
9
12
15
0,013
0,012
0,011
0,010
0,009
E(Sb2)_50_max(|Y|)
methods
CC
EM
MI(1)
MI(10)
MI(2)
MI(3)
MI(4)
MI(5)
MI(6)
MI(7)
MI(8)
MI(9)
PE

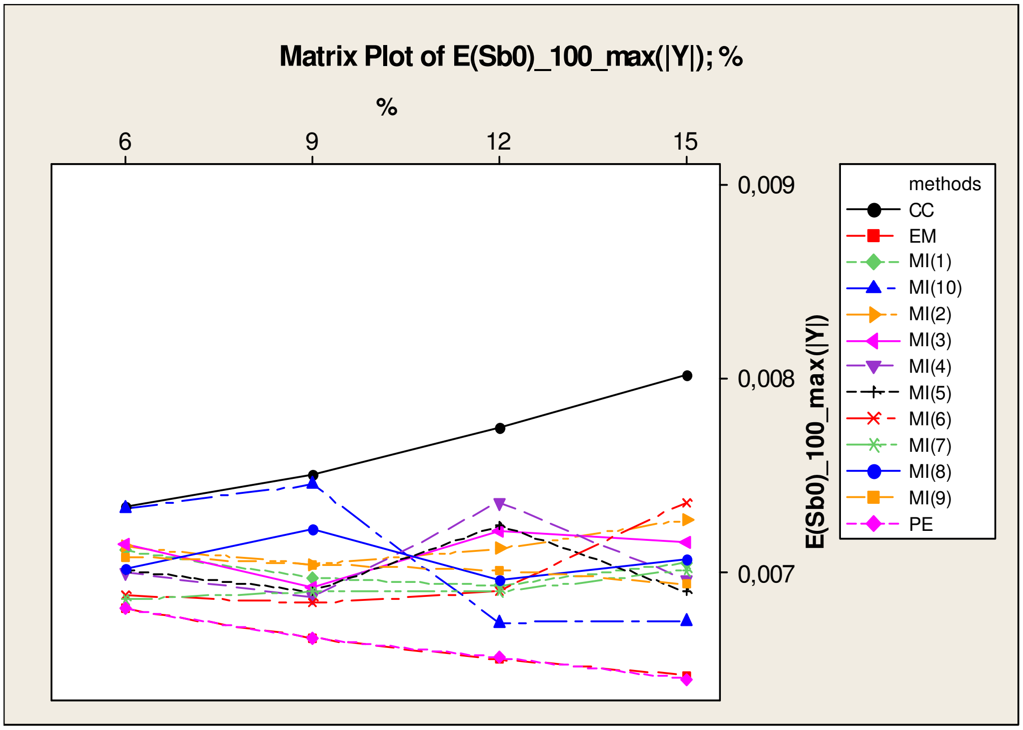

Matrix Plot of E(Sb0)_75_min(|Y|); %
%
6
9
12
15
0,010
0,009
0,008
E(Sb0)_75_min(|Y|)
methods
CC
EM
MI(1)
MI(10)
MI(2)
MI(3)
MI(4)
MI(5)
MI(6)
MI(7)
MI(8)
MI(9)
PE

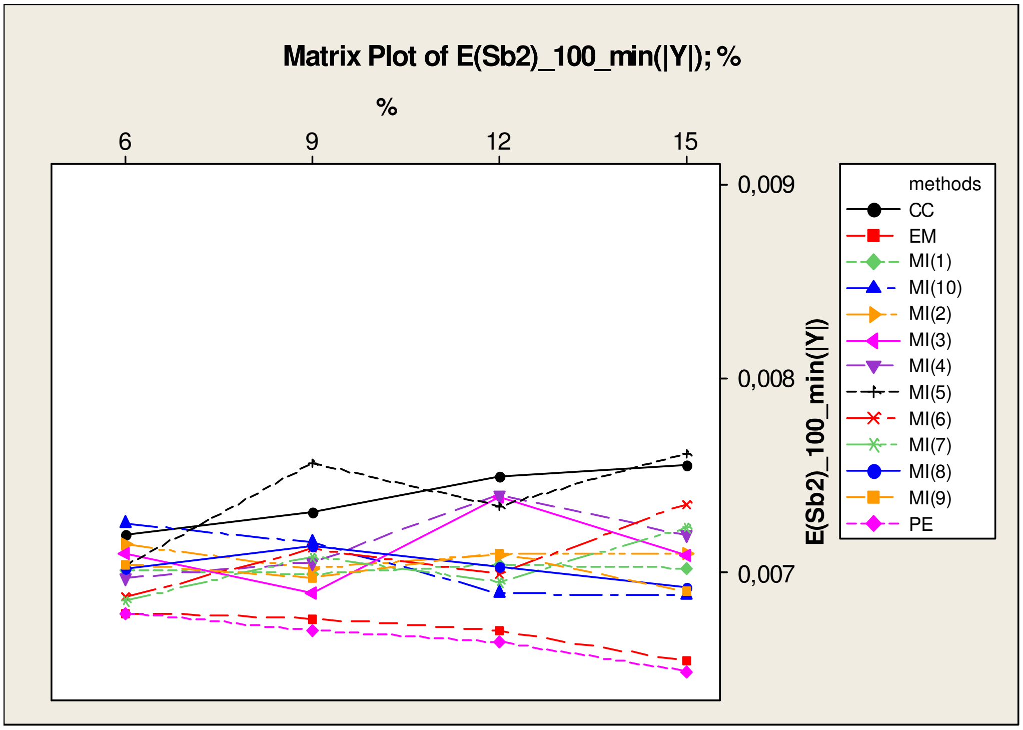

Matrix Plot of E(Sb0)_75_max(|Y|); %
%
6
9
12
15
0,010
0,009
0,008
E(Sb0)_75_max(|Y|)
methods
CC
EM
MI(1)
MI(10)
MI(2)
MI(3)
MI(4)
MI(5)
MI(6)
MI(7)
MI(8)
MI(9)
PE

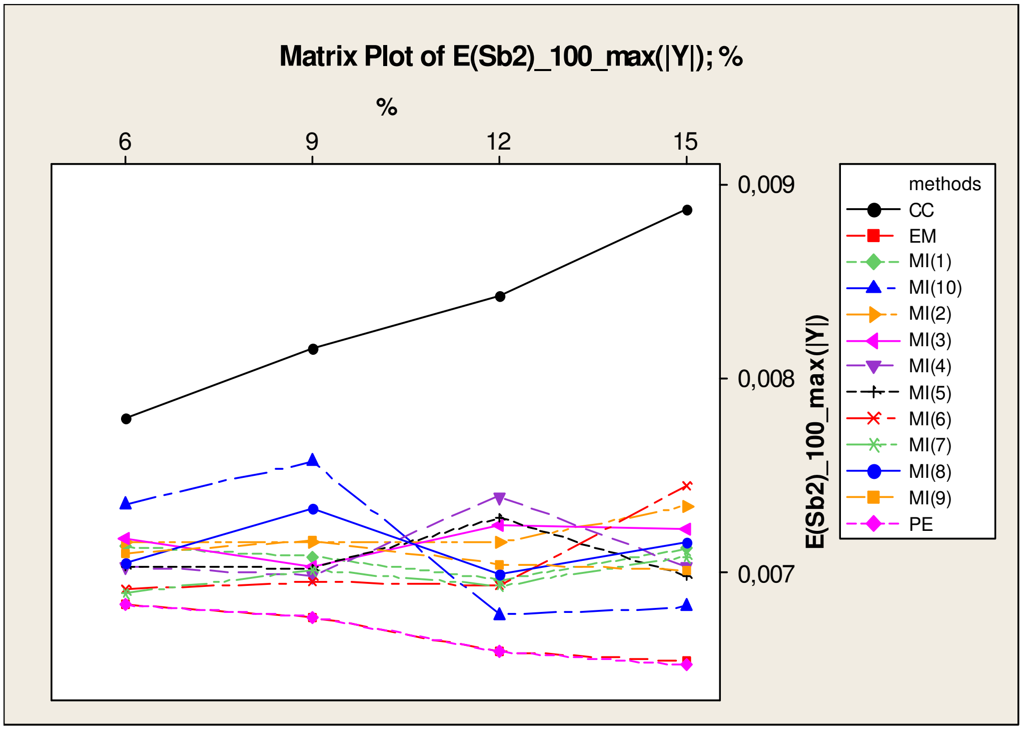

Matrix Plot of E(Sb2)_75_min(|Y|); %
%
6
9
12
15
0,010
0,009
0,008
E(Sb2)_75_min(|Y|)
methods
CC
EM
MI(1)
MI(10)
MI(2)
MI(3)
MI(4)
MI(5)
MI(6)
MI(7)
MI(8)
MI(9)
PE


Matrix Plot of E(Sb2)_75_max(|Y|); %
%
6
9
12
15
0,010
0,009
0,008
E(Sb2)_75_max(|Y|)
methods
CC
EM
MI(1)
MI(10)
MI(2)
MI(3)
MI(4)
MI(5)
MI(6)
MI(7)
MI(8)
MI(9)
PE

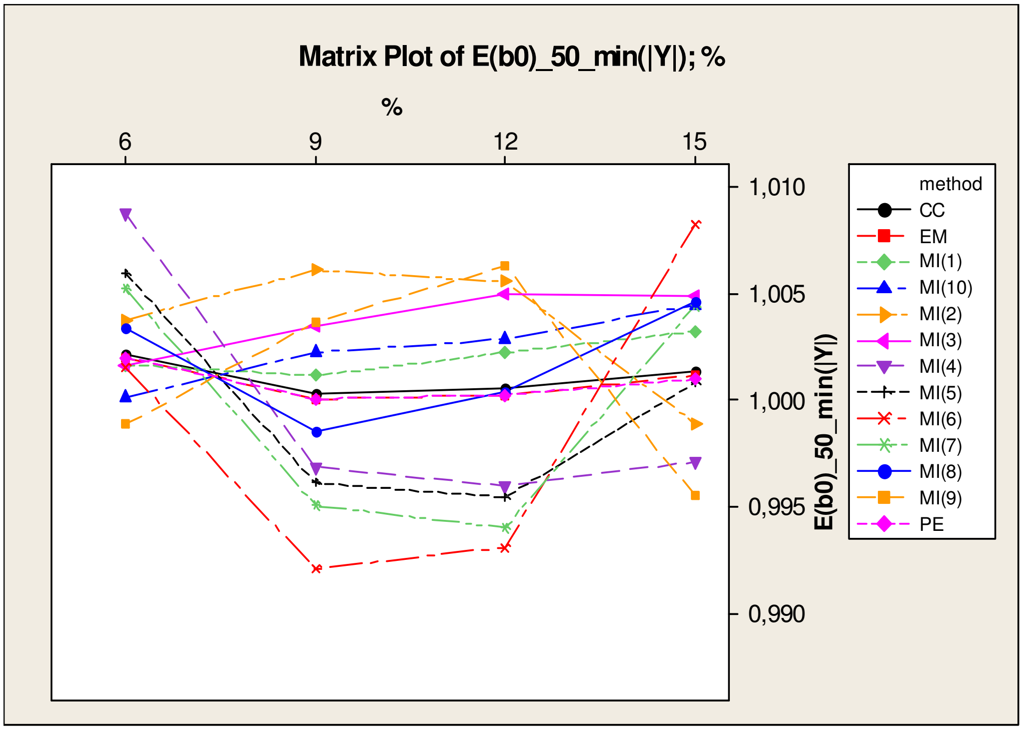

Matrix Plot of E(Sb0)_100_min(|Y|); %
%
6
9
12
15
0,009
0,008
0,007
E(Sb0)_100_min(|Y|)
methods
CC
EM
MI(1)
MI(10)
MI(2)
MI(3)
MI(4)
MI(5)
MI(6)
MI(7)
MI(8)
MI(9)
PE

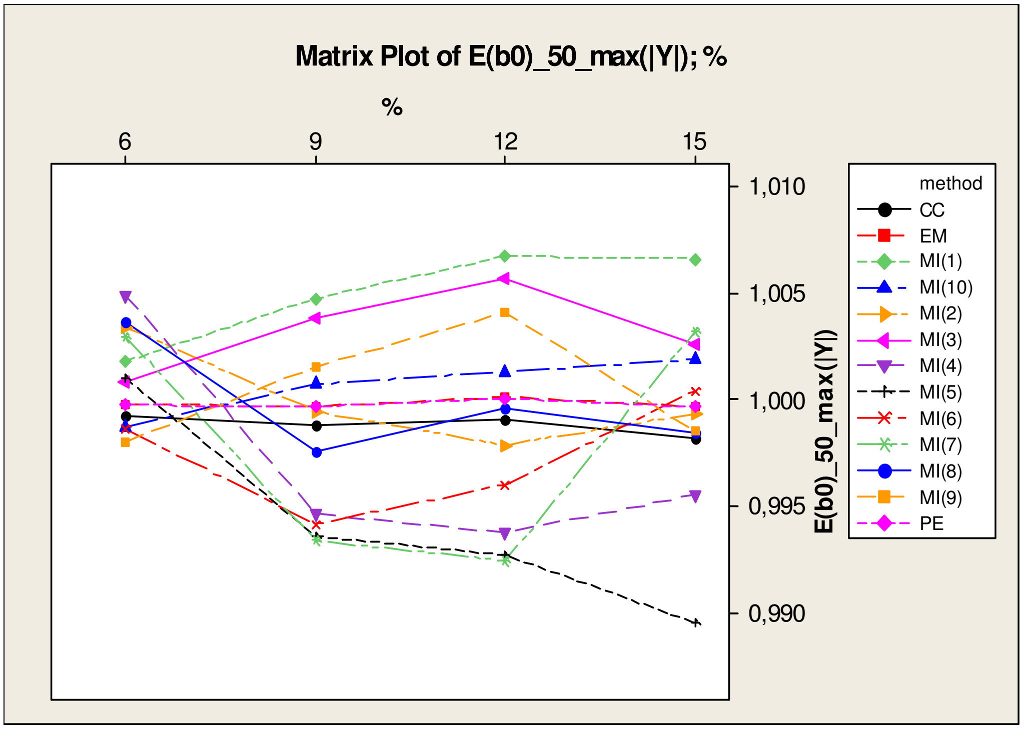

Matrix Plot of E(Sb0)_100_max(|Y|); %
%
6
9
12
15
0,009
0,008
0,007
E(Sb0)_100_max(|Y|)
methods
CC
EM
MI(1)
MI(10)
MI(2)
MI(3)
MI(4)
MI(5)
MI(6)
MI(7)
MI(8)
MI(9)
PE

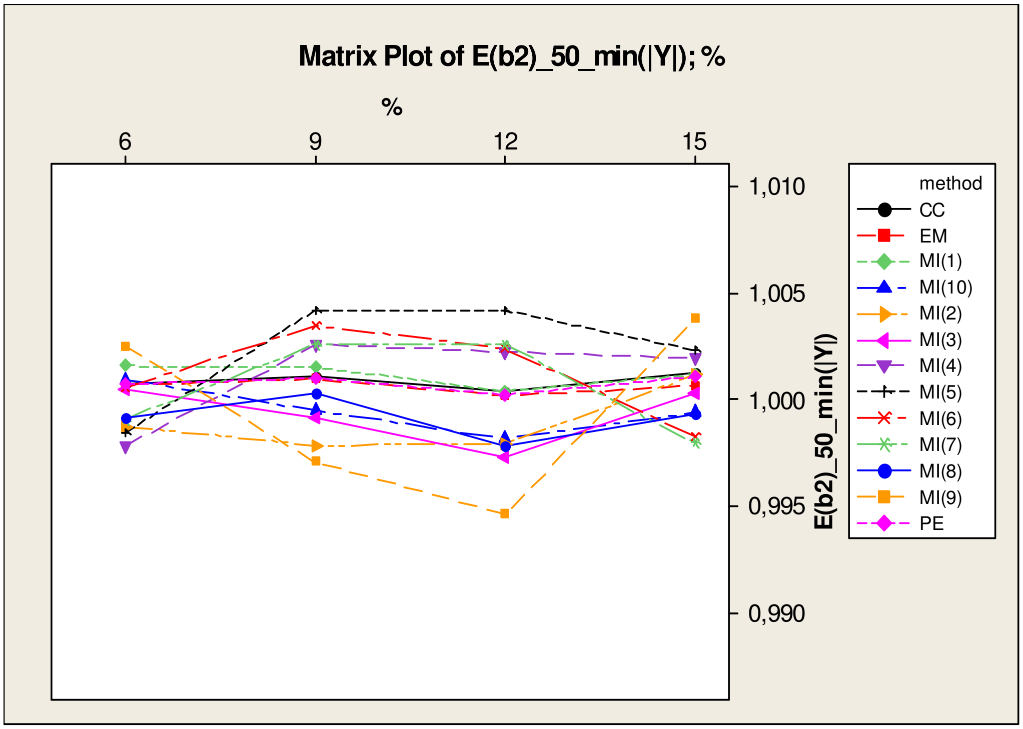

Matrix Plot of E(Sb2)_100_min(|Y|); %
%
6
9
12
15
0,009
0,008
0,007
E(Sb2)_100_min(|Y|)
methods
CC
EM
MI(1)
MI(10)
MI(2)
MI(3)
MI(4)
MI(5)
MI(6)
MI(7)
MI(8)
MI(9)
PE

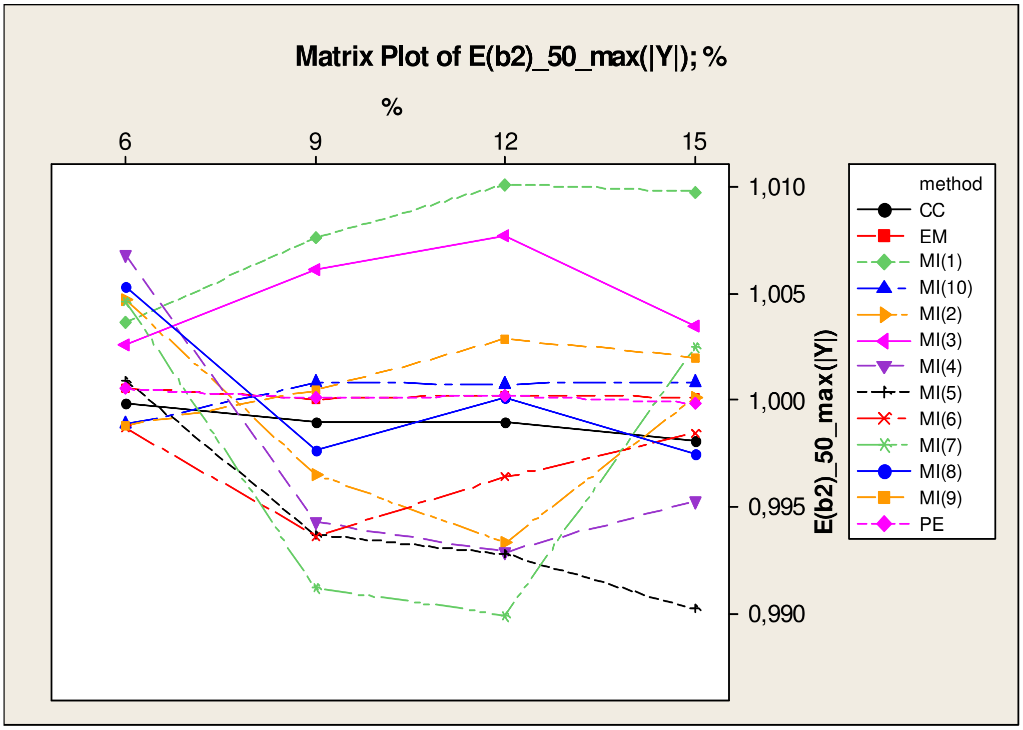

Matrix Plot of E(Sb2)_100_max(|Y|); %
%
6
9
12
15
0,009
0,008
0,007
E(Sb2)_100_max(|Y|)
methods
CC
EM
MI(1)
MI(10)
MI(2)
MI(3)
MI(4)
MI(5)
MI(6)
MI(7)
MI(8)
MI(9)
PE

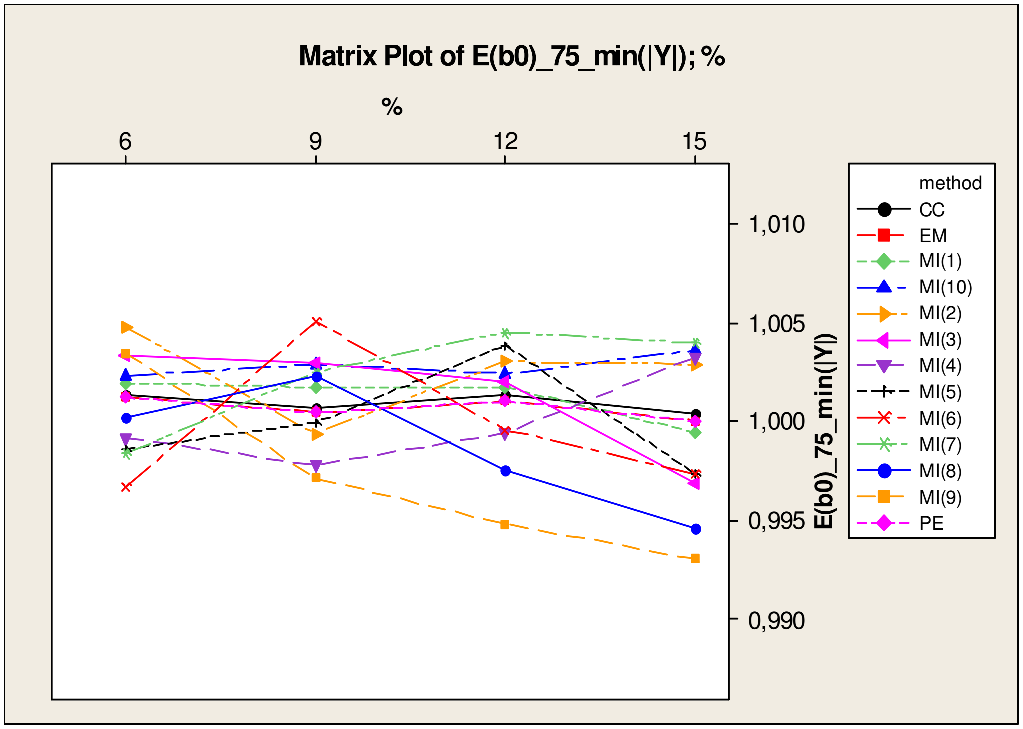
## APPENDIX D – Rest of the Matrix Plots for $E(\hat{\beta}_i)$ ($i = 0,2$)

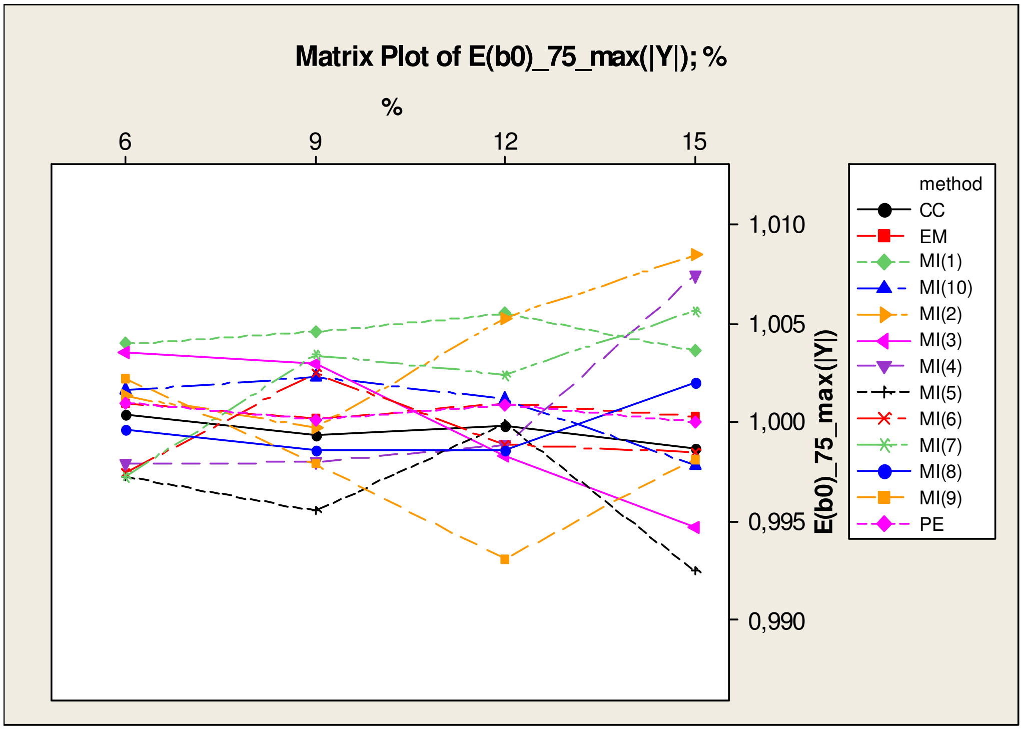


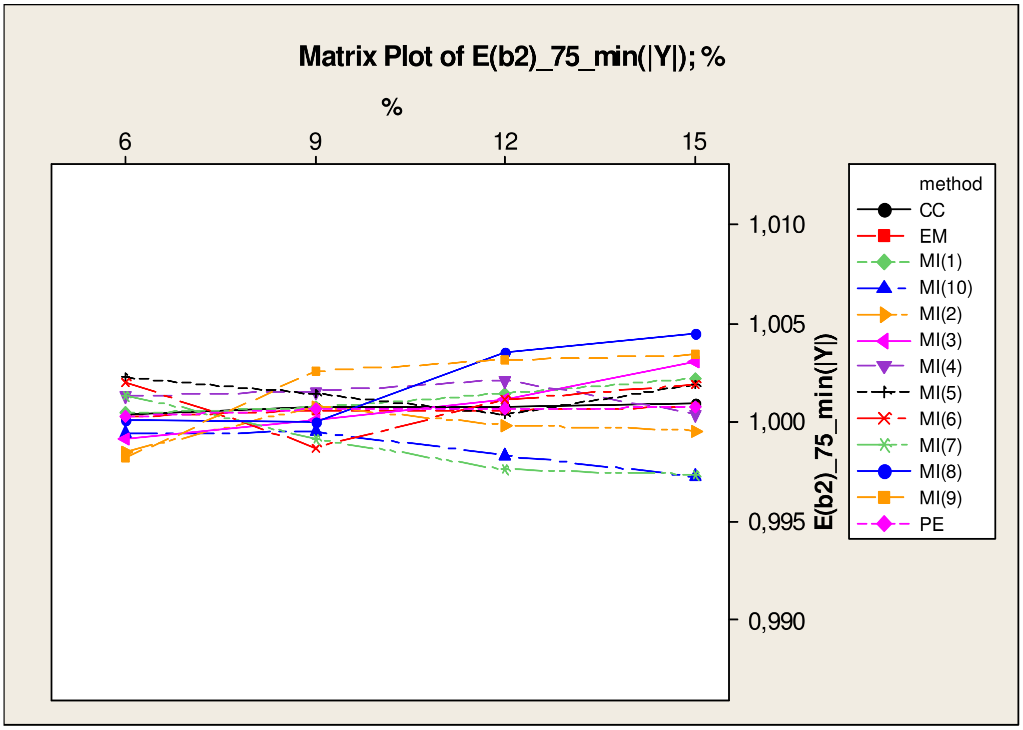


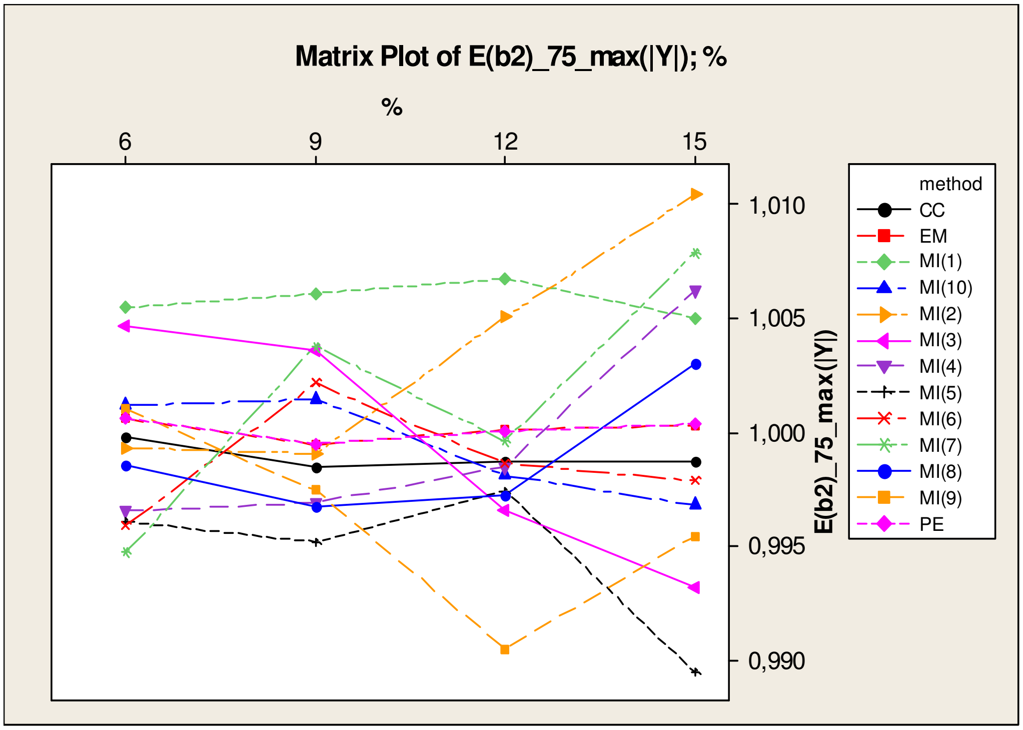

Matrix Plot of E(b2)_50_min(|Y|); %
%
6
9
12
15
1,010
1,005
1,000
0,995
0,990
E(b2)_50_min(|Y|)
method
CC
EM
MI(1)
MI(10)
MI(2)
MI(3)
MI(4)
MI(5)
MI(6)
MI(7)
MI(8)
MI(9)
PE

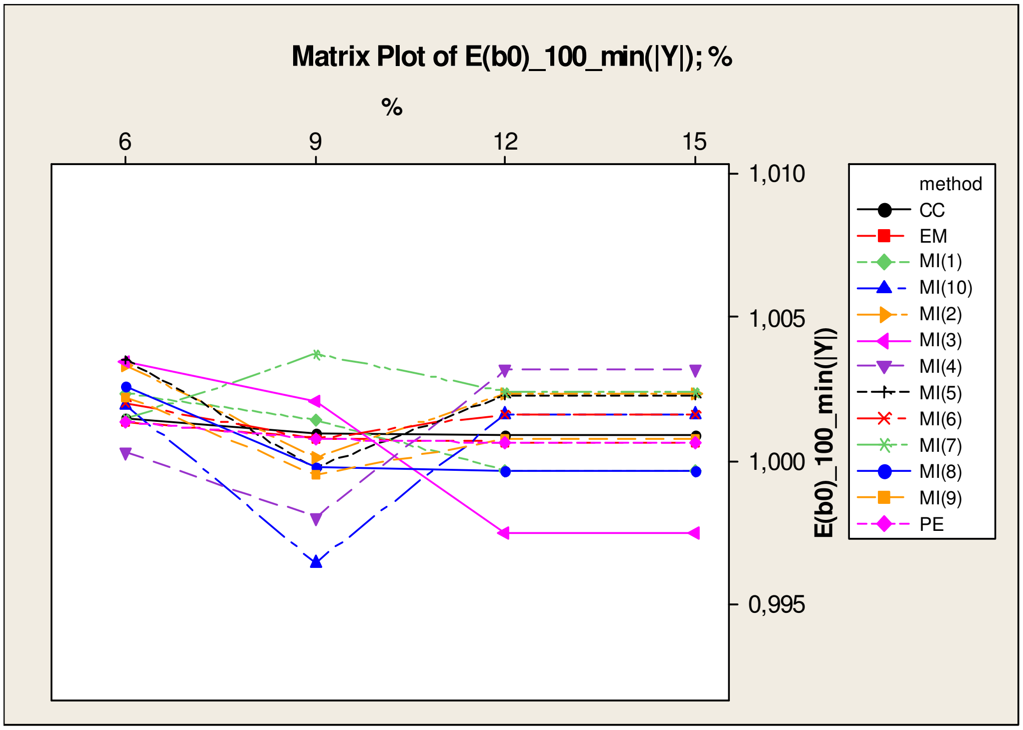

Matrix Plot of E(b2)_50_max(|Y|); %
%
6
9
12
15
1,010
1,005
1,000
0,995
0,990
E(b2)_50_max(|Y|)
method
CC
EM
MI(1)
MI(10)
MI(2)
MI(3)
MI(4)
MI(5)
MI(6)
MI(7)
MI(8)
MI(9)
PE

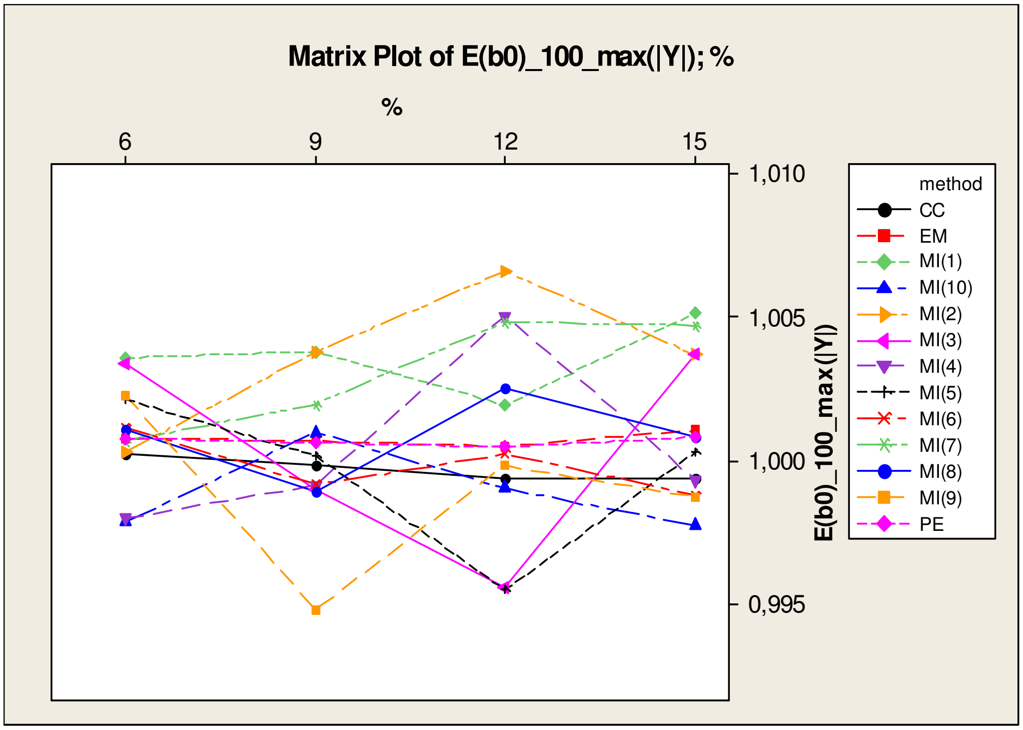

Matrix Plot of E(b0)_75_min(|Y|); %
%
6
9
12
15
1,010
1,005
1,000
0,995
0,990
E(b0)_75_min(|Y|)
method
CC
EM
MI(1)
MI(10)
MI(2)
MI(3)
MI(4)
MI(5)
MI(6)
MI(7)
MI(8)
MI(9)
PE

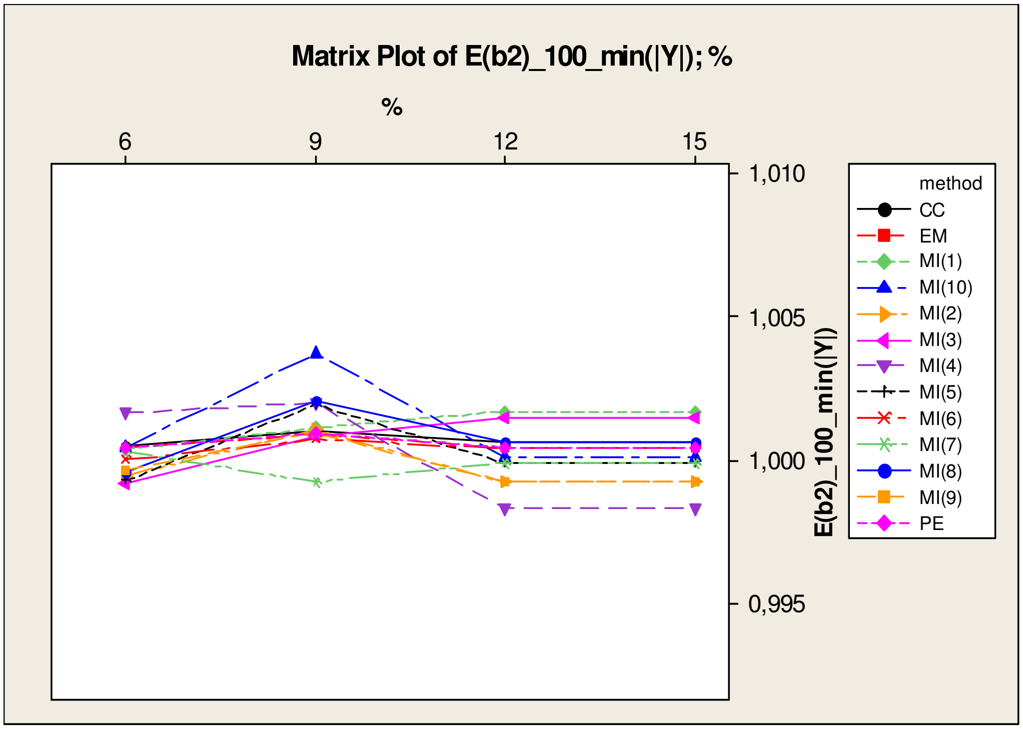

Matrix Plot of E(b0)_75_max(|Y|); %
%
6
9
12
15
1,010
1,005
1,000
0,995
0,990
E(b0)_75_max(|Y|)
method
CC
EM
MI(1)
MI(10)
MI(2)
MI(3)
MI(4)
MI(5)
MI(6)
MI(7)
MI(8)
MI(9)
PE

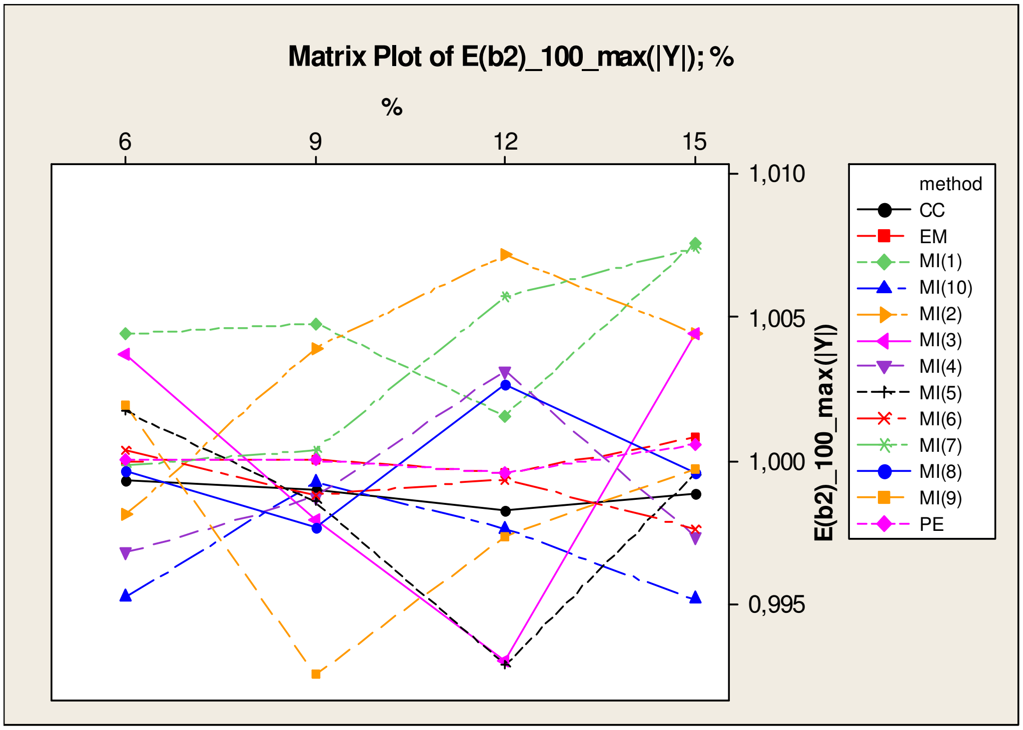

Matrix Plot of E(b2)_75_min(|Y|); %
%
6
9
12
15
1,010
1,005
1,000
0,995
0,990
E(b2)_75_min(|Y|)
method
CC
EM
MI(1)
MI(10)
MI(2)
MI(3)
MI(4)
MI(5)
MI(6)
MI(7)
MI(8)
MI(9)
PE


Matrix Plot of E(b2)_75_max(|Y|); %
%
6
9
12
15
1,010
1,005
1,000
0,995
0,990
E(b2)_75_max(|Y|)
method
CC
EM
MI(1)
MI(10)
MI(2)
MI(3)
MI(4)
MI(5)
MI(6)
MI(7)
MI(8)
MI(9)
PE


Matrix Plot of E(b0)_100_min(|Y|); %
%
6
9
12
15
1,010
1,005
1,000
0,995
E(b0)_100_min(|Y|)
method
CC
EM
MI(1)
MI(10)
MI(2)
MI(3)
MI(4)
MI(5)
MI(6)
MI(7)
MI(8)
MI(9)
PE


Matrix Plot of E(b0)_100_max(|Y|); %
%
6
9
12
15
1,010
1,005
1,000
0,995
E(b0)_100_max(|Y|)
method
CC
EM
MI(1)
MI(10)
MI(2)
MI(3)
MI(4)
MI(5)
MI(6)
MI(7)
MI(8)
MI(9)
PE


Matrix Plot of E(b2)_100_min(|Y|); %
%
6
9
12
15
1,010
1,005
1,000
0,995
E(b2)_100_min(|Y|)
method
CC
EM
MI(1)
MI(10)
MI(2)
MI(3)
MI(4)
MI(5)
MI(6)
MI(7)
MI(8)
MI(9)
PE


Matrix Plot of E(b2)_100_max(|Y|); %
%
6
9
12
15
1,010
1,005
1,000
0,995
E(b2)_100_max(|Y|)
method
CC
EM
MI(1)
MI(10)
MI(2)
MI(3)
MI(4)
MI(5)
MI(6)
MI(7)
MI(8)
MI(9)
PE