Dealing with Missing Data in Cross Sectional Data on Transport

Susan Fred Rumisha

promotor :
dr. Niel HENS, dr. Elke MOONS, dr. Filip
VAN DEN BOSSCHE

Eindverhandeling voorgedragen tot het bekomen van de graad
Master of Science in Biostatistics

universiteit hasselt

University Hasselt
Center for Statistics

Dealing with Missing Data in Cross Sectional Data on Transport

Student: Susan F. Rumisha

Internal Supervisor: dr. Niel Hens
External supervisors: dr. Elke Moons
dr. Filip Van Den Bossche

Thesis submitted in partial fulfillment of the requirements for degree of Master of Science in Biostatistics
September 2007

ACKNOWLEDGEMENTS

It is still miraculous how this thesis was accomplished. It was a piece of work, done with a lot of aspiration and willingness to learn whatever new thing was coming up. Those who facilitated are highly appreciated and a lot of thanks is given to them.

I would like to thank my fellow students for all their enthusiasm and encouragement while working on my project. Their moral support was such a useful tool to bring back the energy whenever I was feeling down and think that, I can't do it.

It will not be fair if I will not mention special thanks to my family, especially my beloved parents, Fredrick and Beatrice, who, though couldn't put a physical hand to help, were pulling the spiritual rope and give a moral support making sure that all are going well from the other hand. With that, I didn't lose hope. Thank you.

I cannot list them all, but I appreciate the financial support from Vlaamse Interuniversitaire Raad (VLIR) for my stay in Belgium, the patience from my colleagues at the National Institute for Medical Research (NIMR), Tanzania, my lovely friends, and finally the whole team of Center for Statistics at the University Hasselt.

It was said that, learning is a progressive discovery of our ignorance. No one can prove that until it reaches a point of no-more-thinking, but, a mind once stretched by a new idea never regains its original dimensions.

Bless be the name of the Lord
Susan F. Rumisha
August 2007

CONTENTS

List of Tables

List of Figures

LIST OF ABBREVIATIONS

ASE	Averaged Squared Error
CC	Complete Cases
CI	Confidence Interval
CMI	Conditional Mean Imputation
GAM	Generalized additive Model
LCI	Length of the Confidence Interval
LL	Lower Limit of the Confidence Interval
LM	Linear Model
MAR	Missing at Random
MASE	Mean Averaged Squared Errors
MCAR	Missing Completely at Random
MI	Parametric Multiple Imputation
MNAR	Missing Not at Random
OD	Original Data
PMM	Predictive Mean Matching
SD	Standard Deviation
SE	Standard Error
SMI	Single Mean Imputation
UL	Upper Limit of the Confidence Interval
WHO	World Health Organization

ABSTRACT

In sample surveys and most research work non-response is often a major problem, this means, sometimes the required data are not obtained for all elements that are selected for observation, and this leads to missing data. Missingness can occur in cross-sectional, longitudinal or multivariate studies. Different imputation methods are available and have been used to fill-in the missing data (either response or covariates) and the produced data is expected, under certain conditions, to lead to valid inference. This study explores efficiency of several imputation methods in cross-sectional data, including parametric and nonparametric, in estimating the effect of covariates in linear models. Simple and advanced imputation methods, such as multiple imputations were considered. Since our data was from a cross-sectional study, univariate patterns and behaviors of missingness were used. Two main scenarios were considered, including a case where the missingness is in the response variable and when the missingness occurs in the covariate. An approach followed was that, a new data was generated, missingness was invocated using different types of missingness models depending on the assumed mechanism, and then imputation was employed to the missing values. Assessment of the accuracy was done by comparing results with the true estimates, which were obtained from original generated data. The focus was in the regression model parameters estimates (with their SE) and the variability introduced in the response values. To evaluate the efficiency of methods and variability of parameters of interest, simulation studies were done. With the runs obtains, MASE values were calculated for each method and compared. Parametric methods for imputation were found to be not adequate, especially when the missing proportion in the response is high. Results from nonparametric methods were good despite slight over or underestimation of the variability in the data. For the case of missingness in the covariate, unbiased results were obtained under MCAR and MAR and biased results under MNAR. However, in this case, single parametric methods seem to perform better than multiple imputation methods or nonparametric ones. It was observed that missingness mechanism could be influenced by the magnitude of the effect of covariate in the fitted model or in the missingness model involved. In other words, one can say that, the strength of the relationship between covariates and the response variable plays a role in manipulating the missingness mechanism. These results were observed using simple exploration hence more research is needed to provide more support.

Keywords: transport, traffic, missingness, imputation, parametric, nonparametric, simulation study.

1. INTRODUCTION AND BACKGROUND

Transport or transportation is the movement of people and goods from one place to another. The term is derived from the Latin *trans* ("across") and *portare* ("to carry"). The field of transport has several aspects: these include infrastructure, vehicles, and operations. Infrastructure includes the transport networks (roads, railways, airways, waterways, canals, pipelines, etc.) that are used, as well as the nodes or terminals (such as airports, railway stations, bus stations and seaports). The vehicles generally ride on the networks, such as automobiles, bicycles, buses, trains and aircrafts. The operations deal with the way the vehicles are operated on the network and the procedures set for this purpose including the legal environment (Laws, Codes, Regulations, etc.) Policies, such as how to finance the system (e.g., use of tolls or gasoline taxes) may be considered part of the operations.

Road safety continues to be one of the nation's most serious public health issues—it affects everyone, whether you drive, walk or cycle. Road traffic accidents kill or injure thousands of people every day. Most of developed and developing countries do not have national road safety programmes. Lack of these programmes results to less efficient follow up of what is happening in the traffic and transport field, which leads to less road safety for the population involved. Thousand of pages have been written on the problem of road safety, and it has been identified as a worldwide problem. It causes a lot of consequences in public health, social life and economic prosperity of the country. The number of people killed in road traffic crashes each year is estimated to be around 1.2 million and with increased efforts, this number is expected to rise by 65% between 2000 and 2020 (WHO report, April 2007)

Most countries experienced enough of these tragedies, hence, to reduce the statistics, a range of laws, regulations, penalties and initiatives on the road users are placed. These include things like speed cameras, road-side drug testing, audible line markings on roads, double demerit points for repeat speed offenders, vehicle impoundment and alcohol ignition interlocks for repeat drink drivers. In addition, in other countries reduction of the road toll was targeted through new licensing rules, regulations and better education for young drivers. For the European Union, transport is one of the community's earliest common policies and has focused on removing obstacles at the borders between Member States so as to facilitate the free movement of persons and goods. The last White Paper on transport policy constitutes a genuine action plan aimed at improving the quality and efficiency of European transport. The ultimate objective is to shift the balance between the various modes of transport by 2010 through an active policy to revitalize the railways, promote transport by sea and inland waterway and develop intermodality (Activities of the European Union, 2005).

It is very clear that, despite the efforts done, most of the users are still not following the rules and hence contributing to the high statistics of crashes and accidents reported. Therefore, provision and increasing of knowledge and skill on safety is vital and this can bring a need to establish an on-going monitor of public perception and attitudes towards road safety issues. Regular surveys on transport field might help to evaluate the effectiveness of public education campaigns, as well as identify areas requiring further attention.

Cross sectional studies are commonly used in traffic/transport surveys. The setup is good for descriptive studied and when one wants to estimate the burden at a specific place and at one specified moment in time. The results of these kind of studies may lead to hypothesis generation, which could be tested by, e.g., intervention studies, or more formally by random control study. Like other studies traffic/transport surveys face same problems like high cost, low response rate, unrealistic responses and most of time missing data. Therefore, in most cases, modeling traffic/transport data involves modeling incomplete data (Jesson, 2001).

Missing data may occur for several reasons, for instance errors in the data, inadequate data collection process, refusal from participants in providing data for reasons such as fatigue or the sensitive nature of the information or insufficient sampling. However, sometimes issues of underreporting due to settlements between drivers without any registration of accident occurrence or insurance-related non-reporting are likely to occur. Ignoring these and work on what was brought in the desk of a statistician can results in missing the targeted goals (Hawthorne and Elliott, 2005).

Missing or incomplete data is a common and an important problem in many fields of research, and there are various ways to deal with it. Incomplete datasets may lead to results that are different from those that would have been obtained from a complete dataset, hence is important to handle it careful. Different reasons for missing data give rise to different types of missingness. Generally, there are two important types described by Little and Rubin (1987) and Schafer (1997) as ignorable and non-ignorable. Non-ignorable is where the probability of a missing datum is dependent upon its value (i.e. cannot be reliably predicted from other dataset variables) and ignorable missing data is where the probability of a missing datum is not dependent upon its value and inference about the measurement mechanism can be made without addressing missingness. Even if the 'true' ignorability status of intermittent missing data is unknown, most missing data can be recoverable through several methods like imputation. Imputation is a method to fill in missing data with plausible values to produce a complete data set.

Key concepts of missingness differ according to how the missingness occurs. Few missingness mechanisms include Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR). In details, MCAR occurs when the probability of an observation being missing is independent of both unobserved and observed data. This assumes that the probability of response for a variable of interest, say y, is the same for all units in the population. MAR is the most general condition under which a valid analysis can be done by using only the observed data. It occurs if, conditional on the observed data, the mechanism for missingness does not depend on the unobserved. In this case the probability of response to a variable of interest is related to covariates only. Lastly, MNAR occurs where neither MCAR nor MAR hold. This means even after accounting for all the data in hand, the reason for the observation being missing still depends on what was not observed. MCAR and MAR are ignorable while MNAR is non-ignorable (Molenberghs and Verbeke, 2005).

Nevertheless, in most practices missingness issue is ignored and most researchers concentrate only on complete case analysis, by applying a method referred to as "list-wise deletion". List-wise deletion is the simplest procedure and is the default in many statistical packages and this removes a case from an analysis if a datum is missing for case i on any variable that is included in the analysis. The shortcomings of this strategy have been well documented. It ignores possible systematic differences between complete cases and in-complete cases, standard errors will generally be large in the reduced sample because less information is utilized and biased estimates will be obtained if the reduced sample is not a random sub-sample of the original sample (Little and Rubin, 1987). If the discarded cases form a representative and relatively small portion of the entire dataset, then case deletion approach may be reasonable. However, it leads to valid inferences in general only when missing data are MCAR.

Other approaches to deal with missing data are those of single-imputation, where missing values are filled in by a plausible estimate such as the mean or median for that variable on other participants, or stratify and sort by a key covariates then replace missing data from another record in the same strata. However, these methods cannot provide valid standard errors and confidence intervals, since ignores the uncertainty implicit in the fact that the imputed values are not the actual values (Little and Rubin, 1987; Molenberghs and Verbeke, 2005).

To improve the single imputation mentioned above, a conditional mean imputation (CMI) can be done. This can be done by replacing the missing values with predicted values from a fitted model (say regression model). The method might be very efficient for point estimation; however, the inference can be seriously distorted (Rubin 1987).

On recent, much research on missing data analysis has focused on multi-imputation techniques for addressing the issues arise in single, conditional and single random imputation procedures, (Little *et al*. 1987; Zhang 2003). Little *et al*. (1987) proposed a multiple imputation procedure to replace each missing value with a set of plausible values that represent the uncertainty about the right value to impute. Actually the procedure uses Monte Carlo simulation to produce a number (say 10) of complete datasets derived from the initial dataset with missing values. The multiple-imputed-data sets are then analyzed using a standard procedure for complete data and combining the results from these analyses to produce means and confidence intervals which reflect the uncertainty from the missing data in the original dataset. The method requires MAR or MCAR assumption.

However, it should be noted that, a naive inappropriate imputation method might creates more problems than those it can solve (Little and Rubin, 1987). If not well implemented, even the multiple imputation method can be a vague procedure despite all the positive stories about it. Most applied multiple imputation techniques are parametric hence implemented by stating several strong assumptions about both the distribution of the data and about underlying regression relationships. But, if such parametric assumptions do not hold, the multiply imputed data are not appropriate and might produce inconsistent estimators and thus misleading results.

Due to uncertainty that might occur when applying different methods of imputation, parametrically, a simple nonparametric method was applied. Generalized additive model (GAM) with integrated smoothness estimation was used and the missing values were replaced by the predicted values from this model. GAMs represent a method of fitting a smooth relationship between two or more variables through a scatterplot of data points. These models are useful when the relationship between the variables is expected to be of a complex form, not easily fitted by standard models or one wants the data to suggest the appropriate functional form. One of the main reasons for using GAMs is that they do not involve strong assumptions about the relationship that is implicit in standard parametric models like regression.

In this report, results of model parameter estimates (or/and other parameters of interest) based on the data filled-in using parametric (single and/or multiple) imputation methods and nonparametric methods are compared and discussed.

2. OBJECTIVE

To explore the effect of missingness in estimation of regression relationship between variables in a cross sectional study.

2.1. Specific objectives

- Explore missingness in the data on transport under well specified mechanism.
- Apply methods to correct for missingness focusing on the effect of using parametric imputation over nonparametric methods.
- Use simulation studies to evaluate stability of parameters estimated and assess accuracy of different imputation methods used.
- Explore the effect of the magnitude of parameter estimate in the missingness mechanism.

3. MATERIALS AND METHODS

3.1. The dataset

The data were collected in Flanders from January 2000-January 2001 using individual questionnaire and an activity-diary. People were asked to write down for two consecutive days, activities they conducted, where, when, with whom, time spent and type of transport mode used to arrive at the location of the activity. The survey based on a random sample of 2823 households, including 7638 people who were more than 6 years old. Most of the interviewees were students and workers. It contains about 40 covariates and 2 main response variables that have information on the transport and traveling behaviors of the population in Flanders. Some of the variables in the dataset explores type of mode/equipment used for traveling/moving from one place to another, distance covered say from place of residence to where a mean of transport can be obtained, or from the stop (bus, train, …) to the designated destination (considering school and/or place of work), etc. Other information collected were on driving license, specifically on its availability and time from when it was obtained. Demographical and other personal information like age, sex, occupation, level of income, level of education, profession of the individual, marital status, number of people in the household, name and type of municipality of residence, availability of transport modes, were also collected (Moons and Wets, 2007). List of all variables with their descriptions can be seen in Table A, Appendix.

3.2. Exploration

To be able to select variables to be used for this study, the amount of missingness in each of the variable was observed, the correlation between variables was checked and some summary statistics were done. Figures and tables presenting some patterns, trends and important features from the data were provided. For few variables, cross tabulations were done to see frequency distribution and obtained results were summarized.

The univariate regression models were done between each covariate and responses and to quantify the relationship, a value of Coefficient of Determination was observed. Later, a multiple regression model was fitted with few selected variables (see next section). Since our main objective is to study missingness, the response variable with the highest proportion of missingness was used.

3.3. Multiple Regression Analysis

A regression model was fitted and estimates were obtained. Due to the large number of covariates, backward and stepwise automatic model selection procedures were applied. However, available information from the literatures on the factors that can influence travel distances were considered. The covariates selected include age, sex, level of education, use of bicycle (as a mode of transport), number of members in the household younger than 6 years and average number of trips made. The response used was the Total Distance travelled by a specific individual. The regression model fitted with *p-1* covariates has the form:

$$E[Y \mid X] = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_{p-1} X_{p-1}$$

where β_i's are the parameter estimates and X_{i1}'s are the covariates.

3.4. Data generation and analysis of original data

Using the obtained conditional mean $E[Y]$ (from the multiple regression model) and the variance of the response variable from the available data, σ^2, new response values (say Y^*) were randomly generated from a normal distribution, i.e., $Y^* \sim N(E[Y], \sigma^2)$. This new data will be referred to this report as original data and will be used to attain study objectives. No generation of covariates was done rather the original data from the survey was used.

From the original data, a multiple regression model was fitted and parameter estimates with their standard errors and 95% Confidence Intervals (CI) were obtained. This model is referred to as model from Original Data. To simplify the exercise, model fitted here used fewer variables than the previous model. Variables used here were Age, Sex and Average number of trips made by an individual (abbrv. AVERP).

3.5. Invoking missingness

For the given data, $Y_1, Y_2, Y_3, ..., Y_n$ of size n, assume that the indicator for missingness is defined as follows:

$$R_i = \begin{cases} 1 & \text{if } Y_i \text{ is observed} \\ 0 & \text{otherwise} \end{cases}$$

Then one can assume $R \sim B(1, \pi)$ where π is a missingness probability and can be defined as a function of covariates only, covariates and/or the response variable, or none of them depending on the assumed mechanism of missingness.

Specifically for each mechanism of missingness and with l number of covariates, π was defined in a 'missingness model' as,

$$\pi(\mathrm{x}) = expit(\varphi_0) \qquad \text{for MCAR,}$$

$$\pi(\mathrm{x}) = expit(\varphi_0 + \varphi_1 X_1 + \varphi_2 X_2 + ... + \varphi_l X_l) \qquad \text{for MAR,}$$

and

$$\pi(\mathrm{x}) = expit(\varphi_0 + \varphi_1 X_1 + \varphi_2 X_2 + ... + \varphi_l X_l + \varphi_j Y) \qquad \text{for MNAR}$$

where

$$expit(x) = \frac{exp(x)}{1 + exp(x)}$$

To produce a specific missingness level in the original data (like 30%, 50%, …or an approximate), values of $\varphi_i's$ were randomly selected, then substituted in the missingness model (choice differs for each mechanism).

3.5.1. Missingness models

In missingness generation, two scenarios that differ by type of missingness models used were considered. In the first scenario, two missingness models were combined. The process went like this; probability for missingness for each individual was generated from two different models.

$$P_1 = expit(\varphi_{01} + \varphi_{11} X_1 + \varphi_{21} X_2 + \varphi_{31} X_3 + \varphi_{41} y)$$

and

$$P_2 = expit(\varphi_{02} + \varphi_{21} X_1 + \varphi_{22} X_2 + \varphi_{32} X_3 + \varphi_{42} y)$$

N.B: Components in the model change according to the corresponding missingness mechanism

Then, two sets of missingness indicators, R_1 and R_2 were generated respectively from each function (model). Now each observation in the dataset has two indicators for missingness, i.e. one from each set. The combination of the results was done in such away that, an observation with missing indicator value $R_i = 1$ in both R_1 and R_2 was taken as missing. In the second scenario only one function (model) was considered.

$$P_1 = expit(\varphi_0 + \varphi_1 X_1 + \varphi_2 X_2 + \varphi_3 X_3 + \varphi_4 y)$$

The two scenarios are expected to generate different missingness patterns hence allow to study the effect of missingness model and missingness pattern in the estimation of model parameters.

3.6. Analyses methods

After invoking missingness in the data, the following methods/analyses were done,

- Complete Cases (CC)
- Single Mean Imputation (SMI)
- Condition Mean Imputation (CMI)
- Multiple Imputation (MI)
- Single Imputation using Generalized Additive Model (GAM)
- Multiple Imputation using GAM

The parameter estimates for each covariate, standard errors and their 95% CI were calculated for each component and compared with the ones obtained from original data.

3.7. Description of imputation methods

As the main objective of the study mentioned, the statistical part of this report focused on different ways of dealing with missing data and doing imputation for the missing values. Short descriptions of different algorithms to conduct parametric and nonparametric imputations are presented here:

3.7.1. *Mean Imputation*

This is a single imputation method and was done by replacing missing values with the arithmetic (unconditional) mean of the observed data.

3.7.2. *Conditional Mean Imputation, using regression model*

Let $\mu(\theta)$ be the vector with elements $\mu_i(\theta),\ i \in mis$; that is

$$\mu(\theta) = E(Y_{mis} \mid X, Y_{obs}, \theta)$$

The approach seeks to fill in the missing data with one set of "best" values might choose $\mu(\hat{\theta})$ has been referred to as conditional mean imputation.

Models used for the conditional mean imputation were as follows:

$E[Y] = \beta_0 + \beta_1 Sex + \beta_2 Age + \beta_3 AVERP + \beta_4 Sex * AVERP$ ……………for the first scenario

$E[Y] = \beta_0 + \beta_1 Sex + \beta_2 Age + \beta_3 AVERP + \beta_4 Age^2 + \beta_5 Age^3$ …………for the second scenario

After fitting the specified regression models, the missing values were then replaced by the predicted values estimated from the model. Different models were defined for each case to explore the effect of the imputation model used.

3.7.3. *Single Imputation using PMM*

The imputation method used here is based on the normal-theory linear regression which assumes existence of a linear relationship between covariates and the response. During the process, the linear regression model is fitted to the complete cases and parameter estimates β_i's and variability in the data σ are obtained by drawn from their posterior distribution, then given the drawn values, a set of imputes for missing values were drawn using Predictive Mean-Matching (PMM) method (Lazzeroni, L.C. *et al)*. PMM refers to, for each incomplete case, a random case is drawn from a set of complete cases having conditional predictive means close to that of the incomplete case and imputed to the missing value. In this report, this method will be referred to as PMM-I

3.7.4. Multiple Imputation using PMM

The procedure is similar to what was explained in the *Single Imputation using PMM*, rather here the process is done in a multiple way. The number of imputations, *m* considered were 5. Results of the models fitted from the 5 sets of imputed data were averaged taking into account between and within variability of the estimates and SEs. The method will be referred to as PMM-II.

Let Q be the estimate of the parameter for a given covariate, $\hat{Q}_i$ and $\hat{U}_i$ be the point and variance estimates from the i^{th} imputed dataset. Then the point estimates for Q from multiple imputations is the average of the m complete-data estimates:

$$\overline{Q} = \frac{1}{m}\sum_{i=1}^{m}\hat{Q}_i$$

Let $\overline{U}$ be the within-imputation variance, which is the average of the m complete-data estimates

$$\overline{U} = \frac{1}{m}\sum_{i=1}^{m}\hat{U}_i$$

and the between-imputation variance B is calculated as

$$B = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{Q}_i - \overline{Q})^2$$

Then the variance estimate associated with $\overline{Q}$ is the total variance

$$T = \overline{U} + (1+\frac{1}{m})B$$

The test statistics calculated to check significance of the estimates is approximately to follow a t-distribution with modified degrees of freedom. To check the efficiency of the imputation, the fraction of missing information about Q was assessed. For 5 imputations, a fraction of missingness of up to 50% reported to produce estimates with efficiency of above 90% (Rubin, 1987).

3.7.5. Single Imputation with Generalized Additive Model

Generalized Additive Model works by replacing the coefficients found in parametric models by a smoother. A smoother(s) is a tool for summarizing the trend of a response variable (Y) as a function of one or more predictors $(X_1,...,X_p)$. The model fitted has a general form,

$$g(E[Y \mid X]) = s_0 + s_1 X_{i1} + s_2 X_{i2} + ... + s_{p-1} X_{p-1}$$

By applying a smoother, s_i, the model produces an estimate of the trend that is less variable, i.e. smoother than original Y. Smoothing takes place by local averaging, that is averaging the Y-values of observations having predictor values close to a target value. Prediction was done based on the obtained model and the predicted values were used to fill-in the missing ones. The method is referred to as GAM-I in this report.

3.7.6. Multiple Imputation with Generalized Additive Model

In this case, GAM was fitted and the predicted value for each observation was obtained. Using the conditional mean for each observation and the variance of the data based on the complete cases, for each observation, random values were generated from a normal distribution and the missing values were replaced 5 times using generated values. Following the procedure of multiple imputations (section 3.7.4), 5 models were fitted and the results were averaged to obtain final model. The method is referred to as GAM-II in this report.

3.8. Missingness data pattern

For all the mentioned methods, this study consider univariate missing data pattern where at first, is a situation where some of the variables (say all covariates) are fully observed and some involve missing measurements (say only the response variable), and second, a situation where there is missingness in a covariate (Figure 1).

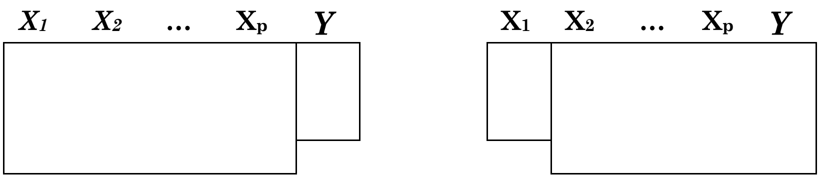


Figure 1: Missing data patterns considered in sample

The parameter of interest is the regression relationship between a partially observed response variable and fully observed covariates and the variability within the response variable or relationship between fully observed response with partially observed covariate where other covariates are full observed.

3.9. Simulation study

To illustrate and compare the performance of imputation methods used simulation studies were carried out. This allows evaluation of the variability of the results obtained from the methods explained above using single sequence data. For the first scenario, a total of 1000 runs were done while for the second scenario 200 runs were used. Each run is expected to produce a slight different pattern of missingness of same proportion hence allow to study the variation and stability of the estimates. For each simulation, means of parameter estimates $\hat{\mu}$, for each variable were computed, estimated standard error $S(\hat{\mu})$ and a 95% confidence interval, i.e. $\hat{\mu} \pm 1.96S(\hat{\mu})$ were calculated. Average length of the CI was also calculated. Boxplots for the estimates and SEs were plotted to assess distribution.

3.10. General assessment of the accuracy of imputation

Imputation methods explained might be well-known approach to treat non-response in surveys. However, they can have a number of impacts on data and other processes, but more importantly, on estimates produced from the 'filled-in' data. It is therefore important to assess the accuracy of the imputation method used. Among the best approaches that have been suggested is the calculation of the variance under imputation. These can be done either based on the model or check variability in the imputed data. For our study the accuracy of the imputations used were assessed using two main measures.

First, after generation of missingness, for all models fitted using simulated data (i.e. CC, SMI, CMI, PMM,…), the Mean Averaged Squared Error (MASE) values were calculated and compared. Depends on the imputation method used, MASE was calculated as follows:

$$MASE = \frac{1}{nn}\sum_{i=1}^{nn}(\mu_i^M - \mu_i^{OD})^2$$

where μ_i^M are fitted values obtained from models using CC or augmented data obtained from different imputation method, M, i.e. SMI, CMI, PMM,…, μ_i^{OD} are fitted values from the model fitted with original data, OD, and nn is the number of simulation runs.

Since the same regression model was used for all cases, fluctuations in the MASE values can tell the difference between the mean curve estimated from the original data and that obtained from the augmented data, hence quantify the accuracy of the imputation method used and stability of parameter estimates. The worse imputation method is expected to have high MASE value, which implies that the difference between the two curves is large.

To support information reported by MASE, the bias-variance decomposition of the ASE based on all models/analyses done was also reported. This is defined as:

$$ASE(\hat{\theta}) = Var(\hat{\theta}) + (Bias(\hat{\theta}, \theta))^2$$

where *Bias* is defined as $E(\hat{\theta} - \theta)$

For the case of this study θ were fitted values obtained from the OD and $\hat{\theta}$ were fitted values obtained from the CC or augmented data.

In addition, for all sets of data, the variance (and standard deviation) of the response values was calculated and compared to that obtained from the original data. Data with very low/high variance was taken as badly imputed data. This was done only for the case that missingness was in the response variabel since for the case where missingness was invoked in covariate, response values from OD were used.

3.11. Scheme of simulation

The flow chart in Figure 2 summarizes the scheme of simulation described.

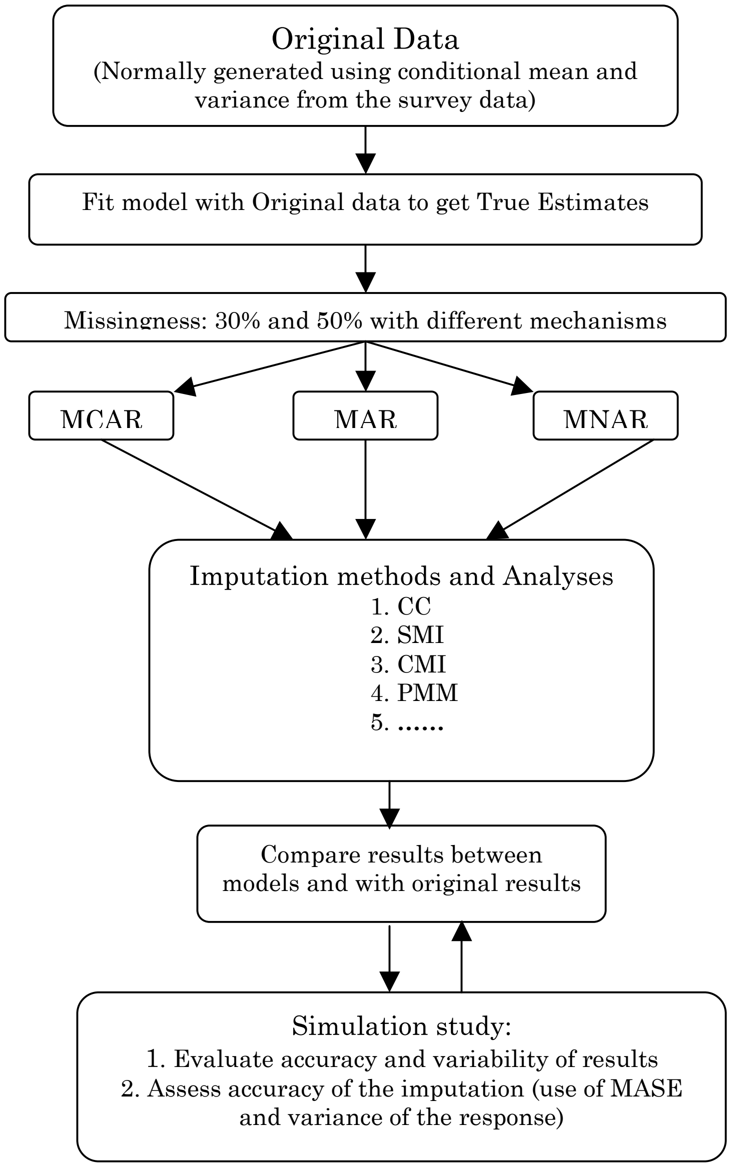


Figure 2: Scheme of Simulation used for Data on Transport

3.12. Use and dissemination of results

Whenever possible the results obtained from this study can be shared to others through local and international communication such as report, publication and presentation in meeting and conferences. No specific results will be provided to individuals involved in the study.

3.13. Tools and software

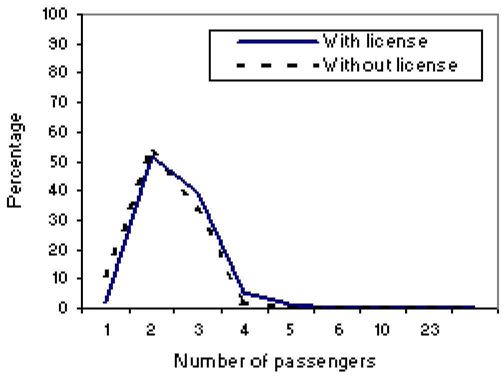
The SAS software and *R* program were used. Data manipulations were done in SAS while the actual analysis was done in *R*. Specifically in *R*, the packages *mice* which stands for 'Multivariate Imputation by Chained Equations', *mitools* which stands for 'tools for multiple imputation of missing data', *mgcv* which is a package for smoothness estimation were used. All the tests were done at 5% level of significance. Selected codes and programs used for analysis are attaches in Appendix.

4. RESULTS

4.1. Study population

A total of about 6059 individuals were involved in the study, among those 51.75% were males while females were 48.25% Mean age was almost the same for each sex and was about 39.3 years (SD=18.65). Most of the respondents were married (58.1%), followed by unmarried individuals (32.30%). Other marital status with small proportions includes living together, divorced and widow/widower. More than half of the individuals include students and employees/workers. About 51.7% of the interviewed people mentioned to have income between 501 and 1250 euros a month while 37.5% had income between 1251 and 2500 euros a month. Very low proportion had income below 500 euros (women attributes 87.3% of this) or above 2500 euros (men attributes 90.7% of this) a month. About 25% of respondents attained higher education and university level, 43% had secondary education plus other general or technical education while the remaining proportion had at most primary education. Men were observed to attain higher education more than women. For instance among those with university degree, men were 63.6% while women were only 36.6%

On traveling and driving information, it was observed that, on average most people travel for about 43.5 km. Male individuals covered higher average distance (50.7km) than females (35.7km). Among respondents used modes that requires licence, 72.5% mentioned to own a driving license while few (27.5%) had no license. The average shortest distance to the place where transport can be obtained was 500 – 999 meters. Main means of transport mentioned were cars (self drive or as a passenger), train, tram or metro, transport arranged by company or school, bus, motor, bike and on foot. The proportion of travellers with and without license on the number of passengers carried was observed to be the same (Figure 3).



It was noted that most of travellers without license are those of young age (less than 20 years) and with low education level (at most secondary education). It was also observed that the proportion of people with no license decreases as the level of income increases.

Figure 3: Proportion of travellers with or without license and number of passengers carried

4.2. Multiple regression analysis

To obtain the conditional mean for generation of the data to be used for the study (OD), a multiple regression model was fitted with selected variables. Results of the model fitted are summarized in Table 1.

Table 1: Parameter Estimates for the regression model with Total distance from field data

Variable	Parameter Estimate	Standard Error	P-value
Intercept	-13.8892	3.9005	0.0004
AVERP	4.4647	0.4830	<0.0001
Age	-0.1422	0.0551	0.0099
DIPLOMA	5.3451	0.4285	<0.0001
Sex (M=1)	15.8271	1.8726	<0.0001
Use of Bicycle	3.5771	0.7595	<0.0001
Member < 6yrs	-5.2593	1.9140	0.0060

M= males

For a quick look, it can be observed that total travel distance is highly significant associated with age of the person, gender, type of mode used, and average number of trips made. The distance decreases as age increases and males individuals have higher distances as compared to females.

The parameter estimates obtained from this model were used to define the mean of the distribution of the original data, which was used for the whole exercise. The variance of the available cases was 4595.48 that make a standard error of 67.79.

Generation of missingness was done in the original data to obtain missingness of 30% and 50% levels under different missingness mechanisms. Multiple regression analysis was then done for complete cases, single mean imputed data, conditional mean imputed data, single and multiple PMM imputed data and, single and multiple GAM imputed data.

Results from the whole exercise are presented in three parts based on the scenario of missingness models used and pattern described in the methodology. First part includes a scenario where missingness assumed to occur only in response and when a combined missingness model was used. Second part reports results when only a single missingness model was used and missingness occurred in response. The third part includes results when missingness is in covariate. Lastly, results on the effect of magnitude of coefficient of varaible (in missingness and fitted models) in missingness mechanism are reported.

4.3. Part I: Combined missingness models: Missingness in response

This part summarizes results obtained from the missingness generated using the first scenario where two missingness models were combined. Overall missingness models with their corresponding vector of missingness indicators that were considered are as follows:

$$P_1 = expit(\varphi_{01} + \varphi_{11} Sex + \varphi_{21} Age + \varphi_{31} AVERP + \varphi_{41} y) \text{ with } R_1 \sim B(1, 1 - P_1)$$

and $$P_2 = expit(\varphi_{02} + \varphi_{12} Sex + \varphi_{22} Age + \varphi_{32} AVERP + \varphi_{42} y) \text{ with } R_2 \sim B(1, 1 - P_2)$$

The two vectors of missingness indicators, R_1 and R_2 were combined (summed up) and the missing values were taken to the observation with value of $R_i = 1$ in both vectors.

Values of coefficients used in the missingness models described above for all mechanisms are reported in Table 2.

Table 2: Values of coefficients used in the missingness models-1st scenario

Mechanism	**Level**	**Parameters model 1**					**Parameters model 2**				
		φ_{01}	φ_{11}	φ_{21}	φ_{31}	φ_{41}	φ_{02}	φ_{12}	φ_{22}	φ_{32}	φ_{42}
MCAR	30%	-1.5	--	--	--	--	0.5	--	--	--	--
	50%	1.5	--	--	--	--	-0.45	--	--	--	--
MAR	30%	-90.5	5	3	0.5	--	47	37	-45.5	2	--
	50%	-110.5	-9	2.85	0.5	--	55	14	-40.5	2	--
MNAR	30%	1	21	4	1	-2	3	1	1	1	-3
	50%	0.9	1	2	1	-2	-0.2	14.45	3	2	-3

4.3.1. Analysis of the Original Data

Results of the parameter estimates from the regression model fitted using the original data are presented in Table 3.

Table 3: Parameter estimates, SE and 95% CI of the estimate for the Original Data

Parameter	**Estimate**	**SE**	**LL**	**UL**	**LCI**
Intercept	15.444	3.0059	9.5525	21.3356	11.7831
Sex (M=1)	13.775	1.8830	10.0843	17.4657	7.3814
Age	0.015	0.0515	-0.086	0.1160	0.2020
AVERP	5.9478	0.4772	5.0126	6.8831	1.8705

From Table 3 it can be seen that, sex and average trips significantly increase the total travel distance of an individual. Age was found to be not significant.

Since for this model complete data was used, these results will be referred to as true estimates. The variance of the response variable from the original data is 69.77, which is very similar to the one from the survey data.

4.3.2. *Analysis after generation of missingness and apply imputation: Parametric methods*

Results of models fitted under each missingness mechanism and for each proportion of missingness are presented in Table 4, Table 5 and Table 6 and respective plots of probability of missing with covariates involved are presented in Figure 3 and Figure 4.

i. MCAR

Table 3a summarizes results of estimates from model using CC for 30% and 50% missingness level. For all variables, results of CC for the 30% missingness are quite close to those of the original data while those under 50% are different, which might indicate the effect of level of missingness in data (Table 4a).

Table 4a: Parameter estimates, SE and CI for MCAR mechanism for CC-1st scenario

	30% missingness				*50% missingness*			
Parameter	**Estimate**	**SE**	**LL**	**UL**	**Estimate**	**SE**	**LL**	**UL**
Intercept	14.791	3.5724	7.7891	21.793	17.9541	4.3056	9.5152	26.3932
Sex (M=1)	13.79	2.2665	9.3476	18.2324	12.6192	2.6692	7.3877	17.8507
Age	0.0215	0.0622	-0.1003	0.1434	0.0046	0.0739	-0.1403	0.14941
AVERP	6.0878	0.5667	4.9771	7.1984	5.4769	0.6714	4.1610	6.7928

For both levels (i.e. 30% and 50%) the SE were overestimated hence results to wider CI. CC analysis is easy to apply but the loss of information can results into bias results.

After imputing the missing values, worse results were obtained when the missing values were replaced by the mean of the observed ones (Table 4b).

Table4b: Parameter estimates, SE and CI for MCAR mechanism for SMI and PMM-II,-1st scenario

	30% missingness				*50% missingness*			
Parameter	**Estimate**	**SE**	**LL**	**UL**	**Estimate**	**SE**	**LL**	**UL**
				SMI				
Intercept	22.8944	2.5223	17.9507	27.838	31.0857	2.1247	26.9214	35.2501
Sex (M=1)	9.9032	1.5801	6.8063	13.0001	6.0696	1.3310	3.4609	8.6783
Age	0.0215	0.0433	-0.0633	0.1063	-0.0023	0.0364	-0.0737	0.0692
AVERP	4.3873	0.4004	3.6025	5.1720	2.7286	0.3373	2.0675	3.3897
				CMI				
Intercept	16.1692	2.9396	10.4075	21.9309	15.7629	2.4698	10.9220	20.6037
Sex (M=1)	11.2868	3.2262	4.9635	17.6101	16.6922	2.7106	11.3795	22.0049
Age	0.0197	0.0430	-0.0647	0.1041	0.0070	0.0362	-0.0639	0.0779
AVERP	5.7192	0.5719	4.5983	6.8401	6.0612	0.4805	5.1194	7.0029
				PMM-II				
Intercept	12.5855	3.2066	6.2727	18.8984	19.0560	3.1644	12.8367	25.2752
Sex (M=1)	12.8570	2.0244	8.8674	16.8467	10.9665	2.6023	5.4935	16.4395
Age	0.0279	0.0563	-0.0833	0.1391	-0.0048	0.0587	-0.1216	0.1120
AVERP	6.4711	0.4807	5.5289	7.4134	5.5636	0.5191	4.5391	6.5882

The parameter estimates and SEs are very small as compared to the true ones due to that confidence intervals for the estimates are very narrow and even lie between ones obtained from analysis of original data. Results from conditional mean imputation were better as compared to the SMI but there was overestimation of SEs for some covariates. Results of the MI method under 30% level were the best for this case, since estimates and SEs were closer to the true estimates than other methods, but this was not the case for the 50% level (Table 4b). Simulation study will be done to evaluate stability of these estimates.

ii. MAR

For the case of MAR mechanism, in the CC analysis, the estimates for other covariates except Age were close to the true ones for both levels of missingness. Estimates for age were overestimated (in magnitude) and even the significance status changes. This makes the CI far different from that of the original data (Table 5a)

Table 5a: Parameter estimates, SE and CI for MAR mechanism for CC-1st scenario

	30% missingness				*50% missingness*			
Parameter	**Estimate**	**SE**	**LL**	**UL**	**Estimate**	**SE**	**LL**	**UL**
Intercept	35.8340	5.4484	25.1552	46.5127	38.6119	8.1796	22.5798	54.6440
Sex (M=1)	13.9963	2.2361	9.6136	18.3789	13.4138	2.6925	8.1364	18.6912
Age	-0.3756	0.0909	-0.5537	-0.1975	-0.4371	0.1337	-0.6993	-0.1750
AVERP	5.7447	0.5603	4.6466	6.8429	6.0169	0.6939	4.6568	7.3769

Table 5b presents results of the estimates after employing different imputation methods to the missing values for both 30% and 50% levels of missingness.

Table 5b: Parameter estimates, SE and CI for MAR mechanism for SMI, CMI and PMM-II,-1st scenario

	30% missingness				*50% missingness*			
Parameter	**Estimate**	**SE**	**LL**	**UL**	**Estimate**	**SE**	**LL**	**UL**
					SMI			
Intercept	30.7252	2.4952	25.8347	35.6158	30.8805	2.1334	26.6991	35.0620
Sex (M=1)	9.9735	1.5631	6.9099	13.0371	6.4923	1.3364	3.8728	9.1117
Age	-0.1629	0.0428	-0.2467	-0.0790	-0.0820	0.0366	-0.1537	-0.0103
AVERP	4.5670	0.3961	3.7907	5.3434	3.2555	0.3387	2.5917	3.9193
					CMI			
Intercept	34.6982	2.9053	29.0039	40.3926	36.8752	2.4703	32.0334	41.7171
Sex (M=1)	15.8711	3.1885	9.6217	22.1205	16.1774	2.7111	10.8636	21.4911
Age	-0.3733	0.0425	-0.4566	-0.2899	-0.4323	0.0362	-0.5032	-0.3614
AVERP	6.0283	0.5652	4.9205	7.1361	6.4466	0.4806	5.5047	7.3886
					PMM-II			
Intercept	32.0028	7.4121	13.5901	50.4155	25.5148	14.3799	-13.0506	64.0803
Sex (M=1)	13.9287	2.3035	9.2343	18.6230	11.8990	3.9449	2.5340	21.2640
Age	-0.3210	0.1042	-0.5684	-0.0737	-0.2925	0.2504	-0.9649	0.3799
AVERP	6.0479	0.6290	4.7367	7.3591	7.6313	0.7991	5.8531	9.4096

It can be seen that, for SMI method, the estimates and SEs were underestimated, different from what was observed in CC analysis. The underestimation was worse in the case of 50% level of missingness, which was the same case for MCAR. Since the estimates and the SEs are small, narrow CI was obtained.

To improve results from SMI a conditional mean imputation was done. As one can see from Table 5b, the results were better compared to those obtained under SMI but there is still a problem in the estimation of Age parameters. However, compared the results with those obtained from OD, the SEs were overestimated for some of the variables like Sex. Though the underestimation of variability was reduced by use of CMI method, the method is still doing single imputation hence does not acknowledge the variability between possible values of the missing values. To correct for that multiple imputation method was employed.

For the case of 30% missingness, the results for some of the estimates were close to those in the OD. Surprisingly, though best results were expected from this method, estimates for age are still different and highly overestimated. The SEs for this method were slightly higher than those under CC (Table 5b).

Conditional probabilities of missingness for 30% and 50% were plotted with the respective covariates used to generate missingness and presented in Figure 4.

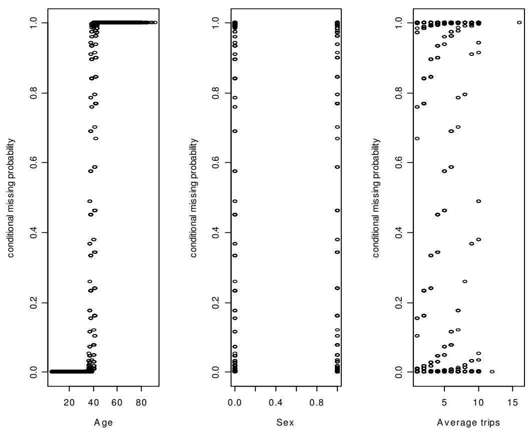


Figure 4: Probability of missingness by covariates under MAR -1st scenario

From the plots it can be observed that, for age, the probability is very low at the lower ages and increases sharply at a certain age level. Actually, the pattern generated here shows that, almost all people with low ages (say up to 40 years) have a very high chance of being missing hence missing in our dataset. Almost the same kind of pattern is seen for the average trips. The patterns were similar for both levels of missingness. These patterns could be reasons for some of bad estimates obtained.

iii. MNAR

Results obtained under MNAR were different as compared to other mechanisms. Estimates deviate from the true ones from CC analysis, which was not the case for other mechanisms.

Results of all four models fitted and for both levels of missingness are presented in Table 6.

Table 6: Parameter estimates, SE and CI for MNAR mechanism-1st scenario

	30% *missingness*				50% *missingness*			
Parameter	**Estimate**	**SE**	**LL**	**UL**	**Estimate**	**SE**	**LL**	**UL**
				CC				
Intercept	-37.9590	2.8564	-43.5574	-32.3605	-47.4157	2.7803	-52.8652	-41.9663
Sex (M=1)	10.8289	1.6925	7.5115	14.1462	6.3348	1.6770	3.0478	9.6217
Age	0.7888	0.0467	0.6972	0.8804	0.5892	0.0448	0.5014	0.6769
AVERP	3.5079	0.4557	2.6147	4.4011	2.8677	0.4703	1.9458	3.7896
				SMI				
Intercept	-19.4319	1.9218	-23.1986	-15.6652	-27.4436	1.3590	-30.1073	-24.7799
Sex (M=1)	7.7417	1.2039	5.3821	10.1013	3.4163	0.8513	1.7477	5.0850
Age	0.5424	0.0330	0.4778	0.6070	0.3079	0.0233	0.2622	0.3536
AVERP	2.1596	0.3051	1.5616	2.7575	1.1559	0.2157	0.7331	1.5788
				CMI				
Intercept	-40.1496	2.2167	-44.4943	-35.8050	-48.4190	1.5613	-51.4791	-45.3589
Sex (M=1)	14.8005	2.4327	10.0324	19.5686	8.2610	1.7134	4.9026	11.6193
Age	0.7921	0.0325	0.7285	0.8557	0.5907	0.0229	0.5459	0.6355
AVERP	4.1110	0.4312	3.2657	4.9562	3.1522	0.3037	2.5569	3.7475
				PMM-II				
Intercept	-39.1002	2.8016	-44.8443	-33.3561	-48.1016	2.3510	-52.9169	-43.2862
Sex (M=1)	10.7444	1.5836	7.5854	13.9035	5.8977	1.2496	3.4349	8.3604
Age	0.8273	0.0391	0.7506	0.9040	0.6174	0.0404	0.5346	0.7002
AVERP	3.4080	0.4445	2.4970	4.3191	2.9594	0.4505	1.9795	3.9393

As it can be seen from Table 6, in all methods except CMI, estimates and SEs were either over or underestimated, with the worse situation occurred when mean was used to fill the missing values and when the proportion of missingness is large. Estimates of CMI are larger compared to other models but still not close to the true ones. No better results were obtained even when multiple imputation method was applied.

The conditional probabilities for missingness were plotted with the covariates and the response values (Figure 5). The same pattern was observed for both 30% and 50% level.

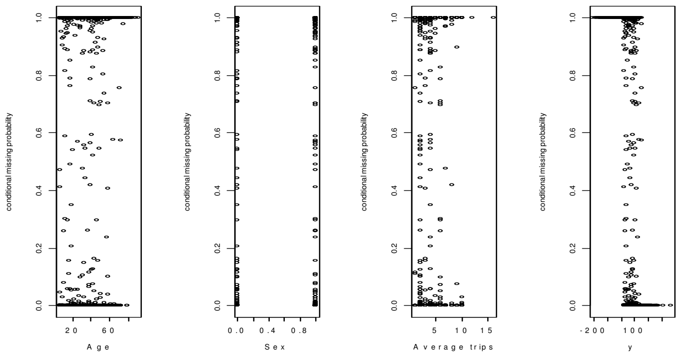


Figure 5: Probability of missingness for each covariate under MNAR -1st scenario

Almost the same pattern was observed for age and average trips, but now with some random trends for middle values. For the response, the probability is high at higher values and decreases sharply for lower ones.

4.3.3. Simulation study: Parametric Imputation

To evaluate the accuracy of the imputation procedures mentioned before, a simulation study was done with a total of 1000 runs. Summary results obtained for each of the missingness mechanism and for each level of missingness are presented in this section.

i. MCAR

Results of the estimates and SE obtained under simulation study for MCAR under 30% level of missingness are presented in Table 7a. Results are summarized for all methods.

Table 7a: Estimates, SE, CI and LCI obtained from the simulation study for 30% levels of missingness from CC, SMI, CMI and PMM-II analysis under MCAR-1st scenario

Parameter	Estimate	SE	LL	UL	LCI	Estimate	SE	LL	UL	LCI
			CC					***SMI***		
Intercept	15.4963	3.6164	8.4081	22.5845	14.1764	24.3605	2.5137	19.4337	29.2873	9.8536
Sex (M=1)	13.7184	2.2657	9.2776	18.1593	8.8817	9.4897	1.5747	6.4034	12.5761	6.1727
Age	0.0151	0.0620	-0.1064	0.1367	0.2431	0.0105	0.0431	-0.0740	0.0950	0.1690
AVERP	5.9364	0.5742	4.8111	7.0618	2.2507	4.1041	0.3990	3.3220	4.8863	1.5642
			CMI					***PMM-II***		
Intercept	14.9015	2.9291	9.1605	20.6425	11.4821	15.0487	3.3776	8.3183	21.7791	13.4608
Sex (M=1)	14.8093	3.2146	8.5087	21.1099	12.6012	13.7003	2.1135	9.4895	17.9110	8.4215
Age	0.0158	0.0429	-0.0682	0.0999	0.1681	0.0182	0.0578	-0.0969	0.1332	0.2301
AVERP	6.0960	0.5698	4.9791	7.2129	2.2338	6.0398	0.5250	4.9984	7.0812	2.0828

As it was observed under single analysis, results of the CC and SMI differ enormously for almost all covariates. The estimates and SE were lower for the case of SMI than in CC and the CI of SMI is almost within that of CC. Comparing the results with those of original data, estimates of CC are much closer while estimates and SEs of SMI were underestimated. Estimates from CMI and PMM-II were good but the SEs were over estimated.

In the CC scenario, the MASE value obtained was 5858.6 and the average variance of the response was 68.48, which is low compared to the one of the original data, which was 69.77. For the case of single mean imputation, the MASE value was 98,777.9, which is much higher than that of CC. The variance of the response was 57.26, which seems to be even lower than that obtained from CC. Moreover, for the cases of CMI and PMM-II the MASE values were 11,263 and 11,640 respectively. The values are very close to each other and surprising higher than that of the CC. Despite the same values of estimates for other variables observed, the high values of MASE could be influenced by the underestimation of the estimates of the intercept observed hence shifted the fitted curve. This resulted to a new fitted curve of the same shape as that of OD but in different position hence makes the difference between the fitted values.

The variance of response was 56.9 for CMI and 63.7 for PMM-II case. The implication of these results will be discussed later. For all samples, the average percentage of missingness was 30.89%.

For the case of 50% missingness, the results obtained for the CC, CMI and PMM-II are almost similar to that of 30% missingness level (Table 7b). For the case of SMI estimates and SE are lower than those in 30% and much lower as compared to the ones obtained in the original data analysis.

Table 7b: Estimates, SE, CI and LCI obtained from the simulation study for 50% levels of missingness from CC, SMI, CMI and PMM-II analysis under MCAR-1st scenario

Parameter	Estimate	SE	LL	UL	LCI	Estimate	SE	LL	UL	LCI
	CC					***SMI***				
Intercept	15.4118	4.2468	7.0880	23.7356	16.6476	29.8079	2.1465	25.6008	34.0150	8.4142
Sex (M=1)	13.8219	2.6607	8.6069	19.0369	10.4300	6.9185	1.3446	4.2830	9.5540	5.2710
Age	0.0155	0.0728	-0.1272	0.1583	0.2855	0.0079	0.0368	-0.0642	0.0801	0.1443
AVERP	5.9524	0.6745	4.6303	7.2745	2.6442	2.9783	0.3407	2.3104	3.6462	1.3357
	CMI					***PMM-II***				
Intercept	14.8420	2.4915	9.9587	19.7254	9.7667	14.8755	3.5760	7.6396	22.1114	14.4718
Sex (M=1)	14.8623	2.7343	9.5030	20.2216	10.7186	13.6332	2.2511	9.0694	18.1969	9.1275
Age	0.0162	0.0365	-0.0553	0.0877	0.1430	0.0212	0.0614	-0.1032	0.1457	0.2489
AVERP	6.1051	0.4847	5.1551	7.0551	1.9001	6.1043	0.5634	4.9663	7.2423	2.2760

MASE values for these models were 9,760; 250,505; 25,339; and 26,475 for CC, SMI, CMI, and PMM-II respectively. As it can be observed, MASE value for SMI is extremely high, showing poor performance of this method. Again for this case, the MASE values for CMI and PMM-II are very similar. The variance of the response variable for CC, SMI, CMI and PMM-II data were 68.4, 48.9, 48.4 and 60.4 respectively. One can then see that, single imputation methods seriously underestimate variability in the data. For all samples, the average percentage missingness was 49.92%.

ii. MAR

Results summarized from the simulation study under MAR mechanism are reported in Table 8a and 8b. There are clear differences between these results and those obtained under MCAR mechanism. In the case of 30% level of missingness, for some of covariates like age, the estimates are low and significant which was not the case in MCAR. For other covariates the results were almost similar as MCAR.

When results were compared to those from original data, under CC and PMM-II, the estimates for age and average trips were very close though the SEs were overestimated. As it was seen in previous analysis, SMI still underestimates SEs and produces low estimates. A lot of fluctuation is still observed for estimates for Age.

Table 8a: Estimates, SE, CI and LCI obtained from the simulation study for 30% levels of missingness from CC, SMI, CMI and PMM-II analysis under MAR-1st scenario

Parameter	Estimate	SE	LL	UL	LCI	Estimate	SE	LL	UL	LCI
	CC					***SMI***				
Intercept	36.0270	5.4493	25.3464	46.7077	21.3613	30.8416	2.4970	25.9475	35.7357	9.7881
Sex (M=1)	13.8182	2.2382	9.4314	18.2050	8.7736	9.8080	1.5642	6.7421	12.8738	6.1317
Age	-0.3761	0.0910	-0.5544	-0.1978	0.3565	-0.1619	0.0428	-0.2458	-0.0780	0.1678
AVERP	5.7262	0.5608	4.6271	6.8253	2.1982	4.5442	0.3964	3.7673	5.3211	1.5538
	CMI					***PMM-II***				
Intercept	34.8318	2.9075	29.1332	40.5305	11.3973	35.9916	6.8570	19.2656	52.7177	33.4521
Sex (M=1)	15.7931	3.1909	9.5390	22.0472	12.5082	13.3860	2.8253	7.2134	19.5586	12.3452
Age	-0.3735	0.0426	-0.4570	-0.2901	0.1669	-0.3714	0.1142	-0.6482	-0.0946	0.5536
AVERP	6.0236	0.5656	4.9149	7.1322	2.2173	5.7478	0.7009	4.2275	7.2681	3.0407

MASE values for CC, SMI, CMI and PMM-II were 88477.4, 124733.6, 378068.3 and 381980.9 respectively. Based on the variability in the response value, the lowest SD obtained was 56.5, which was under CMI.

For the case of 50%, similar trend of results was observed (Table 8b).

Table 8b: Estimates, SE, CI and LCI obtained from the simulation study for 50% levels of missingness from CC, SMI, CMI and PMM-II analysis under MAR-1st scenario

Parameter	Estimate	SE	LL	UL	LCI	Estimate	SE	LL	UL	LCI
			CC					***SMI***		
Intercept	40.6360	8.1456	24.6705	56.6014	31.9309	31.2899	2.1328	27.1097	35.4700	8.3604
Sex (M=1)	13.7254	2.6864	8.4600	18.9908	10.5308	6.5909	1.3360	3.9722	9.2095	5.2373
Age	-0.4684	0.1333	-0.7296	-0.2072	0.5224	-0.0862	0.0366	-0.1579	-0.0145	0.1434
AVERP	5.9396	0.6926	4.5821	7.2971	2.7151	3.2298	0.3386	2.5662	3.8934	1.3272
			CMI					***PMM-II***		
Intercept	38.8822	2.4690	34.0429	43.7215	9.6786	41.2894	10.5349	13.8740	68.7047	54.8307
Sex (M=1)	16.5212	2.7097	11.2102	21.8322	10.6220	14.3206	3.8010	5.3257	23.3155	17.9898
Age	-0.4635	0.0361	-0.5343	-0.3926	0.1417	-0.5186	0.1756	-0.9742	-0.0630	0.9112
AVERP	6.3728	0.4803	5.4314	7.3143	1.8829	6.5173	0.9234	4.3548	8.6798	4.3249

Despite that MI method was expected to perform better, it showed to have the highest MASE value. The lowest MASE value was obtained under CC analysis (69845.65). However, the MASE values for CMI (617759.1) and that of PMM-II (854484.7) were very close. These results bring doubts on the imputation model used under MI method and/or influence of the missingness pattern. MASE value for SMI was 244573.9.

iii. MNAR

The same analysis was done for MNAR mechanism. For the 30% level of missingness, same results were obtained as it was observed for the single analysis (ref. Table 6). There was no clear trend on the estimates or the standard errors. Compared to the true estimates, in CC, SMI and PMM-II, the estimates for Sex and Average trips were underestimated while overestimated in the case of CMI. The estimates for Age were overestimated in all methods. Results for 30% level of missingness are summarized in the Table 9a.

Table 9a: Estimates, SE, CI and LCI obtained from the simulation study for 30% levels of missingness from CC, SMI, CMI and PMM-II analysis under MNAR-1st scenario

Parameter	Estimate	SE	LL	UL	LCI	Estimate	SE	LL	UL	LCI
			CC					***SMI***		
Intercept	-37.9056	2.8571	-43.5056	-32.3056	11.1999	-19.3977	1.9243	-23.1693	-15.6260	7.5432
Sex (M=1)	10.4657	1.6934	7.1466	13.7848	6.6381	7.5414	1.2055	5.1787	9.9041	4.7254
Age	0.7915	0.0468	0.6998	0.8833	0.1835	0.5442	0.0330	0.4795	0.6089	0.1293
AVERP	3.5368	0.4557	2.6436	4.4300	1.7864	2.1826	0.3055	1.5838	2.7813	1.1975
			CMI					***PMM-II***		
Intercept	-40.0630	2.2197	-44.4136	-35.7124	8.7012	-39.6721	2.4988	-44.6602	-34.6839	9.9763
Sex (M=1)	14.3877	2.4360	9.6131	19.1624	9.5493	10.3195	1.5462	7.2415	13.3975	6.1560
Age	0.7948	0.0325	0.7311	0.8585	0.1274	0.8448	0.0419	0.7616	0.9281	0.1665
AVERP	4.1307	0.4318	3.2843	4.9770	1.6928	3.4364	0.4115	2.6069	4.2659	1.6591

Concerning MASE values, as it was expected, very high values were obtained for all analysis with the highest value obtained under PMM-II method (7296185). As it was the case for other mechanisms, the MASE value of CMI (7104445) and that of PMM-II were very close. MASE values for CC and SMI were 4124429 and 5868876 respectively. Variability in the response variable was checked using its standard deviation, this value was observed to be low for all methods. The highest was 51.53 and it was obtained under CC analysis, however, this was low compared to the original standard deviation. On average, the percentage of missingness was 29.9%.

Things were worse for the 50% level of missingness. These results might be influenced by the too much missingness in the data together with the pattern (Table 9b).

Table 9b: Estimates, SE, CI and LCI obtained from the simulation study for 50% levels of missingness from CC, SMI, CMI and PMM-II analysis under MNAR-1st scenario

Parameter	Estimate	SE	LL	UL	LCI	Estimate	SE	LL	UL	LCI
			CC					***SMI***		
Intercept	-47.8935	2.7834	-53.3491	-42.4380	10.9111	-27.6362	1.3592	-30.3002	-24.9721	5.3281
Sex (M=1)	6.4562	1.6776	3.1680	9.7443	6.5763	3.4870	0.8515	1.8181	5.1559	3.3378
Age	0.5957	0.0448	0.5080	0.6835	0.1756	0.3109	0.0233	0.2652	0.3565	0.0914
AVERP	2.8989	0.4699	1.9779	3.8200	1.8421	1.1679	0.2158	0.7450	1.5908	0.8458
			CMI					***PMM-II***		
Intercept	-48.8002	1.5608	-51.8593	-45.7411	6.1182	-47.7576	2.4396	-52.8485	-42.6666	10.1820
Sex (M=1)	8.1827	1.7129	4.8254	11.5400	6.7145	6.1945	1.4157	3.3099	9.0792	5.7693
Age	0.5971	0.0229	0.5523	0.6419	0.0896	0.6053	0.0387	0.5265	0.6840	0.1576
AVERP	3.1559	0.3036	2.5608	3.7510	1.1903	2.9756	0.4359	2.0313	3.9199	1.8886

As it can be seen from Table 9b, for some covariates, SEs were highly underestimated especially when SMI was used. It was noticed that, even results from PMM-II were very different from the true ones in terms of parameters estimates for all covariates. Overestimation of the estimate for Age was real high. On the MASE values and the variability in the data, similar pattern as for the 30% level of missingness was obtained. Values were 7587264, 16109453, 17052868 and 16684080 for CC, SMI, CMI and PMM-II respectively. The average percentage of missingness for all samples was 50.2%.

4.3.4. Analysis after generation of missingness and apply imputation: Nonparametric method

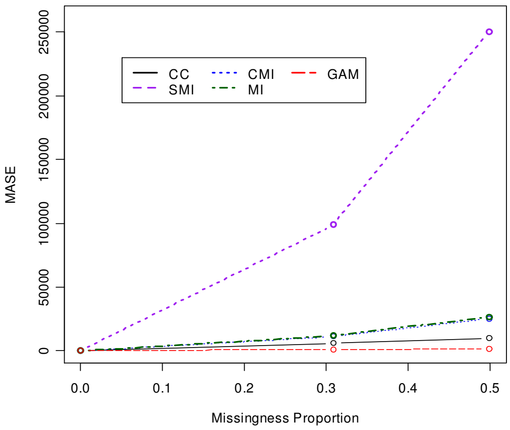
It can be observed that most of results obtained from parametric imputation methods were not very promising. There might be possibilities of misspecification of imputation models based on the data in hand or could be the effect of the observed pattern of missingness probabilities with some covariates. It was decided to apply single imputation in a nonparametric way to allow data to select the best model. Results of the GAMs, which are developed in a non-parametric way, are presented for each of the missingness mechanism (Table 10).

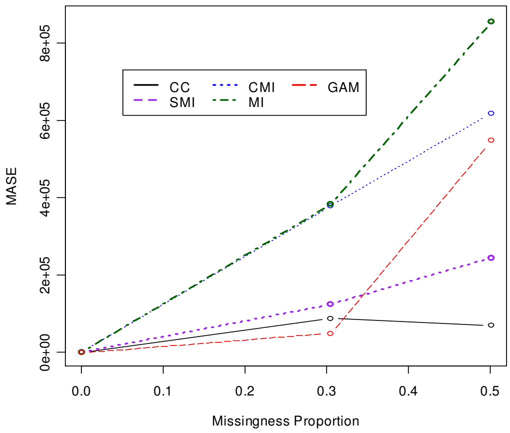
Table 10: Estimates, SE, CI and LCI obtained from GAM study for 30% and 50% levels of missingness for all mechanisms-1st scenario

	30% missingness					***50% missingness***				
Parameter	**Estimate**	**SE**	**LL**	**UL**	**LCI**	**Estimate**	**SE**	**LL**	**UL**	**LCI**
MCAR										
Intercept	14.5371	2.5152	9.6072	19.4670	9.8597	15.9613	2.1280	11.7905	20.1322	8.3417
Sex (M=1)	13.9560	1.5757	10.8678	17.0443	6.1766	13.0847	1.3331	10.4719	15.6975	5.2256
Age	0.0200	0.0431	-0.0645	0.1046	0.1691	0.0169	0.0365	-0.0546	0.0884	0.1430
AVERP	6.1481	0.3993	5.3655	6.9307	1.5652	5.8243	0.3378	5.1622	6.4864	1.3242
MAR										
Intercept	21.9941	2.4824	17.1285	26.8596	9.7311	38.2090	2.1114	34.0707	42.3473	8.2766
Sex (M=1)	13.9113	1.5551	10.8633	16.9593	6.0960	13.3165	1.3227	10.7241	15.9089	5.1848
Age	-0.1354	0.0426	-0.2189	-0.0520	0.1669	-0.4415	0.0362	-0.5124	-0.3705	0.1419
AVERP	6.0880	0.3941	5.3156	6.8604	1.5448	6.2390	0.3352	5.5821	6.8960	1.3139
MNAR										
Intercept	-40.8651	1.8975	-44.5842	-37.1459	7.4383	-48.0488	1.3410	-50.6772	-45.4205	5.2567
Sex (M=1)	10.8433	1.1887	8.5135	13.1732	4.6596	6.1419	0.8400	4.4954	7.7884	3.2930
Age	0.8505	0.0325	0.7867	0.9143	0.1275	0.6103	0.0230	0.5652	0.6554	0.0901
AVERP	3.6121	0.3012	3.0217	4.2025	1.1808	2.9281	0.2129	2.5108	3.3453	0.8345

As it can be seen from Table 10, at the MCAR and MAR cases, the estimates for Sex and Average trips were very similar to the true ones. It was noticed that results were almost the same for both levels of missingness. As it was expected, no improvement on the estimates was observed when GAM was applied under MNAR mechanism.

Figure 6 illustrates the values of MASE for each model fitted in MCAR and MAR missingness mechanism. It can be observed from the figures that, under MCAR, the CMI is doing worse as compared to other methods at both levels of missingness. The GAM has the best results followed by the CC. As it was seen before, performance of CMI and PMM-II were very similar (Figure 6).





a. **b.**

Figure 6: Plots of MASE to assess the accuracy of the imputation method used under a) MCAR and b) MAR-1st scenario

For the case of MAR, the results were different. In this case, CMI and PMM-II methods had very bad results. GAM was doing well for the 30% missingness but not good when the missingness level is around 50%. Under this scenario, the CC analysis seems to be the best when the missingness is high (Figure 6).

Plots of generated values and original data with covariates were plotted under MCAR and MAR mechanisms (Figure A, Appendix). For large values of Age, almost similar pattern was observed for both mechanisms but clearly a difference was seen for low values of Age in the imputed data in the MAR mechanism. From the plot of the fitted curve by Age, different curves were obtained for the smoothed model (GAM) and the linear model (Figure B, Appendix). From these plots it can be observed that, allowing the data to estimate the appropriate model, some of the patterns existing in the data that were not seen by parametric models were captured.

4.4. Part II: Single missingness model: missingness in response variable

For this section the same exercise was repeated using the data with missingness generated from the second scenario. In this scenario single missingness model was used and it was defined as:

$$P = \text{expit}(\varphi_0 + \varphi_1 Sex + \varphi_2 Age + \varphi_3 AVERP + \varphi_4 y)$$

N.B: Components in the model change according to the corresponding missingness mechanism

The vector of missingness indicators was then generated from $R \sim B(1, 1-P)$. Using this model different missingness patterns were obtained (as compared to ones in the first scenario) and it was our expectation to observe differences in the results obtained from different analyses performed.

Values of coefficients used in the missingness model described above for all mechanisms are reported in Table 11.

Table 11: Values of coefficients used in the missingness model-2nd scenario

Mechanism	Level	Model parameters				
		φ_0	φ_1	φ_2	φ_3	φ_4
MCAR	30%	0.89	--	--	--	--
	50%	0.005	--	--	--	--
MAR	30%	2.95	-0.005	-0.05	-0.005	--
	50%	1.97	-0.055	-0.05	0.005	--
MNAR	30%	1.5	2.05	2.15	0.02	-1
	50%	1.5	2.05	1.05	0.02	-1

Results of models fitted after employing different methods of imputation to the missing values together with the trend of Average Squared Error are reported here.

i. MCAR

Results of the complete cases analysis for the case of 30% level of missingness are presented in Table 12a.

Table 12a: Estimates, SE, CI and LCI obtained for 30% and 50% levels of missingness from CC analysis under MCAR-2nd scenario

Parameter	Estimate	SE	LL	UL	LCI	Estimate	SE	LL	UL	LCI
	30% missingness					*50% missingness*				
Intercept	14.7274	3.5770	7.7164	21.7384	14.0220	16.1553	4.2826	7.7613	24.5492	16.7880
Sex (M=1)	12.8817	2.2424	8.4866	17.2768	8.7902	15.1438	2.6934	9.8648	20.4228	10.5581
Age	0.0541	0.0613	-0.0660	0.1743	0.2403	0.0458	0.0730	-0.0973	0.1888	0.2862
AVERP	5.8226	0.5715	4.7024	6.9427	2.2404	5.6715	0.6813	4.3360	7.0069	2.6709

Results of CC analysis under 30% level were very close to the true ones as it was a case for the 1st scenario. Slight overestimation of SE for the parameter Sex was observed. The same pattern of results was obtained for the case of 50% level of missingness. Value of ASE for CC was 2862.14 and was observed to be lower compared to those obtained after imputations.

Tables 12b and 12c presents summary results of estimates obtained after imputing data using different imputation methods for 30% and 50% levels of missingness respectively.

Table 12b: Estimates, SE, CI and LCI obtained for 30% levels of missingness from SMI, CMI, PMM-I, PMM-II, GAM-I and GAM-II analysis under MCAR-2nd scenario

Parameter	Estimate	SE	LL	UL	LCI	Estimate	SE	LL	UL	LCI
			SMI					***CMI***		
Intercept	23.6538	2.5177	18.7190	28.5885	9.8695	14.9728	2.5138	10.0457	19.9000	9.8542
Sex (M=1)	9.0046	1.5772	5.9132	12.0959	6.1827	12.8938	1.5748	9.8072	15.9803	6.1731
Age	0.0392	0.0432	-0.0455	0.1238	0.1692	0.0502	0.0431	-0.0343	0.1346	0.1690
AVERP	4.0255	0.3997	3.2421	4.8089	1.5667	5.8092	0.3991	5.0270	6.5913	1.5643
			PMM-I					***GAM-I***		
Intercept	13.2903	3.0178	7.3754	19.2051	11.8296	15.0397	2.5136	10.1131	19.9663	9.8532
Sex (M=1)	12.1341	1.8905	8.4288	15.8394	7.4106	12.8943	1.5746	9.8081	15.9806	6.1725
Age	0.0687	0.0517	-0.0327	0.1702	0.2028	0.0506	0.0431	-0.0339	0.1351	0.1690
AVERP	6.2092	0.4791	5.2702	7.1481	1.8779	5.7833	0.3990	5.0012	6.5654	1.5642
			PMM-II					***GAM-II***		
Intercept	14.4757	3.5924	7.2392	21.7122	14.4729	14.4096	5.3582	3.9075	24.9117	21.0041
Sex (M=1)	12.0394	2.0337	8.0291	16.0498	8.0207	12.9766	3.4779	6.1599	19.7933	13.6334
Age	0.0576	0.0588	-0.0596	0.1748	0.2344	0.0621	0.0736	-0.0822	0.2064	0.2885
AVERP	5.9160	0.5286	4.8693	6.9628	2.0936	5.8249	0.7772	4.3016	7.3482	3.0466

Results obtained under MCAR for 30% level of missingness under this scenario do not differ much from ones obtained under the 1st scenario. Except for SMI, which still presents worse results by underestimating estimates and SE, performance of other single imputation methods (i.e. CMI, PMM-I and GAM-I) is quite similar and good (estimates close to the true ones). However, there is slightly underestimation of SE for the case of CMI and GAM-I. Results of both multiple imputation methods (PMM-I and GAM-II) were very similar with good estimates though the SEs were overestimated for the case of GAM-II. ASE values SMI, CMI, PMM-I, PMM-II, GAM-I and GAM-II were 106922.7, 3599.8, 9501.9, 7479.9, 3838.9 and 5064.9 respectively. It can be seen that, highest ASE value was obtained under SMI and the ASE values under parametric methods were lower than those under nonparametric ones.

The same pattern of results was obtained for the case of 50% level of missingness, for both analyses after filling-in the data (Table 12c). There is still underestimating of parameters and SEs when SMI was used. Other single methods were performing well though there was a slight underestimation of SEs.

Table 12c: Estimates, SE, CI and LCI obtained for 50% levels of missingness from SMI, CMI, PMM-I, PMM-II, GAM-I and GAM-II analysis under MCAR-2nd scenario

Parameter	Estimate	SE	LL	UL	LCI	Estimate	SE	LL	UL	LCI
			SMI					***CMI***		
Intercept	30.8390	2.1666	26.5924	35.0857	8.4932	16.8266	2.1621	12.5889	21.0643	8.4754
Sex (M=1)	7.7774	1.3573	5.1171	10.4376	5.3205	15.3446	1.3544	12.6900	17.9993	5.3093
Age	0.0208	0.0372	-0.0520	0.0936	0.1456	0.0378	0.0371	-0.0348	0.1105	0.1453
AVERP	2.8805	0.3439	2.2063	3.5546	1.3483	5.6379	0.3432	4.9652	6.3106	1.3454
			PMM-I					**GAM-I**		
Intercept	16.5797	3.0158	10.6686	22.4907	11.8221	16.2055	2.1628	11.9664	20.4446	8.4783
Sex (M=1)	16.3435	1.8892	12.6406	20.0464	7.4058	15.3044	1.3549	12.6488	17.9600	5.3111
Age	-0.0002	0.0517	-0.1015	0.1012	0.2027	0.0346	0.0371	-0.0381	0.1073	0.1454
AVERP	5.8857	0.4787	4.9474	6.8241	1.8767	5.8736	0.3433	5.2006	6.5465	1.3459
			PMM-II					**GAM-II**		
Intercept	14.9151	4.0893	6.3709	23.4593	17.0885	15.5667	5.4321	4.9198	26.2136	21.2938
Sex (M=1)	14.4181	1.9922	10.5028	18.3335	7.8307	15.3878	3.5258	8.4772	22.2984	13.8211
Age	0.0778	0.0744	-0.0807	0.2363	0.3170	0.0462	0.0747	-0.1002	0.1926	0.2928
AVERP	5.9394	0.6231	4.6540	7.2249	2.5709	5.9157	0.7879	4.3714	7.4600	3.0886

ASE values have similar pattern. Values were 10240.2, 254314.6, 27545.5, 23525.8, 34100.7, 26515, and 28040.6 for CC, SMI, CMI, PMM-I, PMM-II, GAM-I and GAM-II respectively.

ii. MAR

Results obtained under MAR are much better compared to ones obtained under the 1st scenario. For this case, estimates were closer to the true ones and same significance status was obtained for Age. Results of CC analysis for both levels are presented in Table 13a.

Table 13a: Estimates, SE, CI and LCI obtained for 30% and 50% levels of missingness from CC analysis under MAR-2nd scenario

			30% missingness					*50% missingness*		
Parameter	**Estimate**	**SE**	**LL**	**UL**	**LCI**	**Estimate**	**SE**	**LL**	**UL**	**LCI**
Intercept	12.6057	3.5061	5.7337	19.4776	13.7439	12.1372	4.1333	4.0359	20.2385	16.2026
Sex (M=1)	13.8753	2.2833	9.4001	18.3505	8.9504	16.3314	2.7416	10.9578	21.7049	10.7471
Age	0.1191	0.0674	-0.0131	0.2512	0.2643	0.1636	0.0817	0.0033	0.3238	0.3205
AVERP	5.8386	0.5750	4.7117	6.9656	2.2538	5.0682	0.6868	3.7221	6.4143	2.6922

From Table 13a, it can be seen that estimates for the case of 30% were closer to the true ones than those of 50% level, which were bit larger. There was more over estimation of SEs for the case of 50% level hence wider CI. ASE value for 30% was 12225.1 and for 50% was 29814.5.

Results obtained after applying imputations to the missing values using different methods for 30% level, are summarized in Table 13b.

Table 13b: Estimates, SE, CI and LCI obtained for 30% levels of missingness from SMI, CMI, PMM-I, PMM-II, GAM-I and GAM-II analysis under MAR-2nd scenario

Parameter	Estimate	SE	LL	UL	LCI	Estimate	SE	LL	UL	LCI
			SMI					***CMI***		
Intercept	21.5922	2.5415	16.6109	26.5736	9.9627	14.8832	2.5343	9.9160	19.8503	9.9343
Sex (M=1)	9.5164	1.5921	6.3959	12.6369	6.2410	13.9765	1.5876	10.8649	17.0882	6.2233
Age	0.0883	0.0436	0.0028	0.1737	0.1708	0.0468	0.0435	-0.0384	0.1319	0.1703
AVERP	4.1307	0.4035	3.3399	4.9215	1.5815	5.7900	0.4023	5.0015	6.5785	1.5770
			PMM-I					***GAM-I***		
Intercept	10.9939	3.0259	5.0632	16.9246	11.8614	14.8939	2.5334	9.9284	19.8594	9.9310
Sex (M=1)	14.7323	1.8955	11.0171	18.4476	7.4305	14.0294	1.5870	10.9188	17.1400	6.2212
Age	0.1298	0.0519	0.0281	0.2315	0.2034	0.0348	0.0434	-0.0504	0.1199	0.1703
AVERP	5.8193	0.4803	4.8779	6.7608	1.8829	5.8987	0.4022	5.1104	6.6869	1.5765
			PMM-II					***GAM-II***		
Intercept	12.4538	3.7839	4.7565	20.1511	15.3945	14.2553	5.4303	3.6119	24.8987	21.2868
Sex (M=1)	14.0027	2.2573	9.4705	18.5348	9.0642	14.1127	3.5247	7.2043	21.0211	13.8168
Age	0.1032	0.0623	-0.0221	0.2285	0.2506	0.0464	0.0746	-0.0998	0.1926	0.2924
AVERP	5.9816	0.5136	4.9714	6.9917	2.0203	5.9408	0.7876	4.3971	7.4845	3.0874

It can be observed from Table 13b that, SMI still had worse performance with low estimates and Ses, and highest value of ASE (103213.5). Other single imputation methods performed quite well with best results obtained under GAM-I procedure. ASE values for other methods were 2773.2, 25725.1, 16472.2, 1047.3 and 2134.5 for CMI, PMM-I, PMM-II, GAM-I and GAM-II respectively. Similarly, comparing results obtained under MI methods, the best results were obtained when nonparametric method was used.

Some differences were observed for the case of 50% level of missingness. The obtained results are summarized in Table 13c.

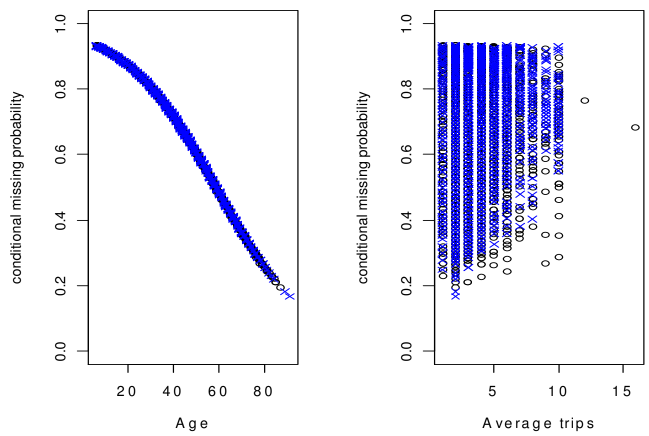
Table 13c: Estimates, SE, CI and LCI obtained for 50% levels of missingness from SMI, CMI, PMM-I, PMM-II, GAM-I and GAM-II analysis under MAR-2nd scenario

Parameter	Estimate	SE	LL	UL	LCI	Estimate	SE	LL	UL	LCI
			SMI					***CMI***		
Intercept	27.8769	2.1925	23.5795	32.1743	8.5947	18.2592	2.1914	13.9641	22.5542	8.5901
Sex (M=1)	7.8006	1.3735	5.1086	10.4927	5.3841	16.5246	1.3728	13.8340	19.2152	5.3812
Age	0.0741	0.0376	0.0004	0.1477	0.1474	-0.0338	0.0376	-0.1074	0.0399	0.1473
AVERP	2.5643	0.3481	1.8821	3.2465	1.3644	5.0681	0.3479	4.3863	5.7499	1.3636
			PMM-I					***GAM-I***		
Intercept	12.9152	3.1048	6.8299	19.0005	12.1706	17.5686	2.1913	13.2737	21.8636	8.5899
Sex (M=1)	15.3798	1.9449	11.5677	19.1919	7.6242	16.6794	1.3727	13.9888	19.3699	5.3811
Age	0.0935	0.0532	-0.0108	0.1979	0.2087	-0.0330	0.0376	-0.1067	0.0406	0.1473
AVERP	5.6076	0.4929	4.6415	6.5736	1.9320	5.2554	0.3479	4.5736	5.9372	1.3636
			PMM-II					***GAM-II***		
Intercept	13.4046	4.3496	4.1940	22.6152	18.4211	16.9211	5.5079	6.1256	27.7166	21.5910
Sex (M=1)	17.0660	2.7918	11.1070	23.0250	11.9180	16.7639	3.5750	9.7569	23.7709	14.0140
Age	0.0502	0.0856	-0.1394	0.2399	0.3793	-0.0213	0.0757	-0.1697	0.1271	0.2967
AVERP	5.6762	0.6830	4.2349	7.1175	2.8825	5.2981	0.7990	3.7321	6.8641	3.1321

As it was observed for other cases, SMI method had very low estimates and SEs as compared to the true ones. The CMI, PMM-I and GAM-I results were better than SMI though there was still underestimation of SEs in some of the methods. Performance of both multiple imputation methods was good (Table 13c).

Almost similar pattern was obtained in terms of ASE values as it was a case under 30% level. For SMI, CMI, PMM-I, PMM-II, GAM-I and GAM-II the ASE values were 290897.2, 32345.6, 18053.0, 18899.6, 26795.5 and 24785.7 respectively.

The conditional missingness probabilities were plotted with age and average trips (by gender) and the plots are presented in Figure 7.



As it can be seen from Figure 7, the probability changes gradually with Age and is not as steep as the one obtained under the 1st scenario. This pattern could influence the performance of imputation methods used as it was reflected in the results. Same pattern was obtained for both missingness levels.

Figure 7: Plots for conditional probability with covariates for MAR-2nd scenario

For the stability of the estimates, simulation study will be performed.

iii. MNAR

In this section results obtained under MNAR mechanism are presented. As it was the case in the 1st scenario, results were worse from the CC analysis for both levels. Most of the estimates and SEs were very low as compared to the true ones (Table 14a).

Table 14a: Estimates, SE, CI and LCI obtained for 30% and 50% levels of missingness from CC analysis under MNAR-2nd scenario

	30% missingness					*50% missingness*				
Parameter	**Estimate**	**SE**	**LL**	**UL**	**LCI**	**Estimate**	**SE**	**LL**	**UL**	**LCI**
Intercept	-37.1589	2.8997	-42.8424	-31.4755	11.3669	-47.1470	2.8069	-52.6486	-41.6454	11.0032
Sex (M=1)	6.2805	1.7123	2.9243	9.6367	6.7124	5.0356	1.6879	1.7274	8.3439	6.6165
Age	0.8434	0.0478	0.7498	0.9370	0.1872	0.6261	0.0451	0.5376	0.7145	0.1768
AVERP	3.3161	0.4631	2.4084	4.2239	1.8155	2.5072	0.4774	1.5716	3.4428	1.8712

Generally for all cases, very high values of ASE were obtained. The lowest values were obtained under CC analyses, which were 4131147 for 30% level and 7549357 for 50% level. This indicates poor performance of all methods under this missingness mechanism.

Tables 14b and 14c summarize results obtained after imputing the missing values for 30% and 50% level respectively. All methods produced estimates that were very different (low) when compared to the true ones and SE were underestimated.

Table 14b: Estimates, SE, CI and LCI obtained for 30% levels of missingness from SMI, CMI, PMM-I, PMM-II, GAM-I and GAM-II analysis under MNAR-2nd scenario

Parameter	Estimate	SE	LL	UL	LCI	Estimate	SE	LL	UL	LCI
			SMI					***CMI***		
Intercept	-18.1424	1.9372	-21.9393	-14.3455	7.5939	-40.9173	1.9160	-44.6726	-37.1620	7.5106
Sex (M=1)	4.9846	1.2135	2.6060	7.3631	4.7571	6.2552	1.2002	3.9027	8.6077	4.7050
Age	0.5681	0.0332	0.5030	0.6332	0.1302	0.9266	0.0329	0.8622	0.9910	0.1288
AVERP	1.9958	0.3075	1.3930	2.5985	1.2055	3.3814	0.3042	2.7853	3.9776	1.1923
			PMM-I					***GAM-I***		
Intercept	-38.4622	2.2236	-42.8205	-34.1039	8.7166	-40.9021	1.9145	-44.6544	-37.1497	7.5047
Sex (M=1)	6.6485	1.3930	3.9182	9.3787	5.4604	6.3124	1.1993	3.9618	8.6630	4.7012
Age	0.8831	0.0381	0.8084	0.9578	0.1495	0.9223	0.0328	0.8580	0.9867	0.1287
AVERP	3.2245	0.3530	2.5327	3.9164	1.3837	3.4245	0.3039	2.8288	4.0201	1.1913
			PMM-II					***GAM-II***		
Intercept	-39.5711	2.3807	-44.2621	-34.8802	9.3819	-41.3818	4.0815	-49.3815	-33.3821	15.9995
Sex (M=1)	6.0217	1.5439	2.9622	9.0811	6.1189	6.3750	2.6491	1.1828	11.5672	10.3845
Age	0.9102	0.0420	0.8272	0.9933	0.1661	0.9311	0.0561	0.8211	1.0411	0.2199
AVERP	3.2372	0.3609	2.5293	3.9451	1.4157	3.4561	0.5921	2.2956	4.6166	2.3210

Underestimation of the estimates and SEs was extremely high for the case of 50% level of missingness (Table 14c).

Table 14c: Estimates, SE, CI and LCI obtained for 50% levels of missingness from SMI, CMI, PMM-I, PMM-II, GAM-I and GAM-II analysis under MNAR-2nd scenario

Parameter	Estimate	SE	LL	UL	LCI	Estimate	SE	LL	UL	LCI
			SMI					***CMI***		
Intercept	-27.1151	1.3640	-29.7884	-24.4417	5.3467	-48.1423	1.3436	-50.7759	-45.5088	5.2671
Sex (M=1)	2.8164	0.8544	1.1417	4.4911	3.3494	4.8764	0.8417	3.2267	6.5262	3.2995
Age	0.3256	0.0234	0.2797	0.3714	0.0917	0.6537	0.0230	0.6086	0.6989	0.0903
AVERP	0.9795	0.2165	0.5552	1.4039	0.8488	2.5784	0.2133	2.1603	2.9964	0.8361
			PMM-I					***GAM-I***		
Intercept	-44.7976	1.8666	-48.4560	-41.1391	7.3170	-47.9423	1.3437	-50.5761	-45.3086	5.2675
Sex (M=1)	3.8186	1.1693	1.5268	6.1105	4.5837	4.8761	0.8418	3.2262	6.5260	3.2998
Age	0.6117	0.0320	0.5489	0.6744	0.1255	0.6507	0.0230	0.6055	0.6958	0.0903
AVERP	2.3030	0.2963	1.7223	2.8838	1.1615	2.5613	0.2133	2.1432	2.9794	0.8362
			PMM-II					***GAM-II***		
Intercept	-47.5519	2.0833	-51.6891	-43.4146	8.2744	-48.3426	-48.3426	-55.0101	-41.6751	13.3351
Sex (M=1)	5.2504	1.3209	2.6206	7.8802	5.2596	4.9284	4.9284	0.6005	9.2563	8.6558
Age	0.6387	0.0416	0.5526	0.7247	0.1721	0.6580	0.6580	0.5665	0.7495	0.1831
AVERP	2.6750	0.3807	1.8905	3.4595	1.5690	2.5877	2.5877	1.6206	3.5548	1.9341

In terms of ASE, the highest values were obtained from CMI and nonparametric methods implying poor performance. Actual ASE values for all methods under 30% level of missingness were SMI (5925395), CMI (7723129), PMM-I (7456308), PMM-II (7666275), GAM-I (7692435) and GAM-II (7713957) while those under 50% level were SMI (16135365), CMI (17208642), PMM-I (17078522), PMM-II (16839854), GAM-I (17192571) and GAM-II (17201508). For better assessment simulation study will be performed.

Plots of missingness probabilities with covariates and response, are presented in Figure 8.

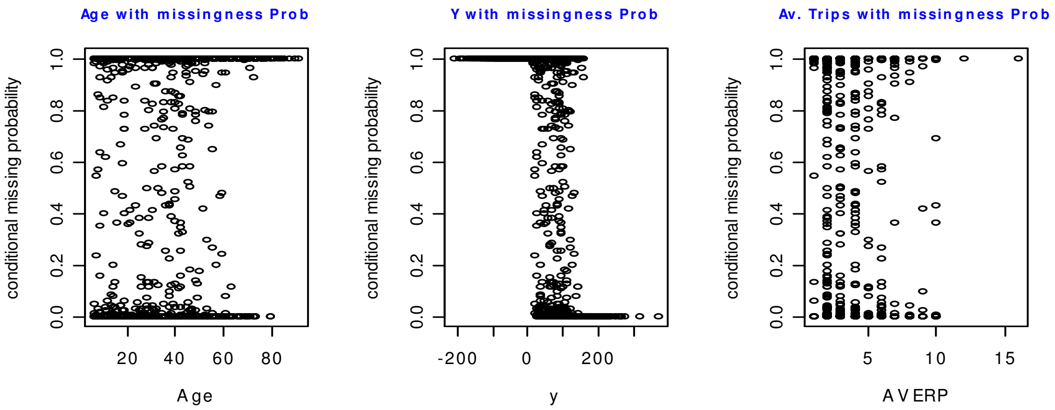


Figure 8: Plots for conditional probability with covariates for MNAR-2nd scenario

As it can be observed from Figure 8, there is a clear pattern of missingness as far as the response and age are concerned. For instance for the response, individuals with low values have higher chance to be missing as compared to those who had high values.

Table 15 presents values of variance and bias² decomposed from ASE under the 2nd scenario, for all methods, mechanisms and both levels of missingness.

Table 15: Bias-variance decomposition based on the CC and imputation methods used – 2nd scenario

Estimate	% Miss	CC	SMI	CMI	PMM-I	PMM-II	GAM-I	GAM-II
				MCAR				
Variance	**30%**	170.16	82.54	171.11	185.38	170.86	169.97	172.61
Bias²		0.0084	0.015	0.0018	0.0007	0.0975	0.0024	0.0017
Variance	**50%**	184.2	46..85	180.4	199.15	188.0	190.5	193.2
Bias²		2.72	2.74	3.96	2.72	4.98	4.23	4.26
				MAR				
Variance	**30%**	184.37	89.74	177.34	187.72	188.17	182.44	185.08
Bias²		0.167	0.141	0.052	0.0064	0.4861	0.033	0.036
Variance	**50%**	170.57	41.74	166.77	181.07	195.98	175.5	177.58
Bias²		0.196	0.11	0.62	0.028	0.0064	0.49	0.48
				MNAR				
Variance	**30%**	270.34	126.64	331.86	304.02	317.97	330.43	336.70
Bias²		854.9	929.9	1134.04	1107.24	1127.10	1132.10	1131.69
Variance	**50%**	159.78	40.42	169.62	144.93	166.01	168.01	171.75
Bias²		2675.2	2881.1	3039.7	3020.5	2981	3037.7	3037.2

It can be observed from Table 15 that, under MCAR mechanism PMM-I method had the lowest bias in both levels of missingness while the highest bias was observed under PMM-II method. The SMI method still present lowest variance estimate for both missigness levels. Meanwhile, for the case of MAR the method with highest variability was PMM-II for both missingness levels, and PMM-II had lowest bias for the case of 50% level. In MNAR mechanism, CC and SMI analyses had better results in terms of bias, though high values of bias obtained suggest poor perfromance.

4.4.1. Simulation study: second scenario

To assess the stability of the results obtained under single analysis, a simulation study was done. A total of 200 runs were obtained for each of the analysis/ imputation method used and the results were averaged. To assess the efficiency of the imputation method used, the values of MASE obtained from each analysis were plotted and compared. Results are presented for each of the missingness mechanism. These results reported the average of estimates and SE over all simulations.

i. MCAR

Table 16a summarizes results of the CC analysis obtained from simulation study for both levels of missingness.

Table 16a: Estimates, SE, CI and LCI obtained from the simulation study for 30% and 50% levels of missingness from CC analysis under MCAR-2nd scenario

	30% missingness					*50% missingness*				
Parameter	**Estimate**	**SE**	**LL**	**UL**	**LCI**	**Estimate**	**SE**	**LL**	**UL**	**LCI**
Intercept	15.6643	3.5729	8.6613	22.6672	14.0060	15.4371	4.2515	7.1041	23.7700	16.6658
Sex (M=1)	13.7338	2.2379	9.3476	18.1201	8.7725	13.7643	2.6631	8.5447	18.9839	10.4393
Age	0.0118	0.0613	-0.1083	0.1319	0.2402	0.0186	0.0729	-0.1243	0.1615	0.2858
AVERP	5.9397	0.5672	4.8281	7.0513	2.2232	5.9170	0.6746	4.5948	7.2391	2.6443

Results obtained under CC analysis were very similar for both cases and estimates were very close to the true ones. However, the SEs were over estimated for both cases.

Results obtained after imputing the missing values using different methods are presented in Table 16b and 16c for the case of 30% level and 50% level respectively.

Table 16b: Estimates, SE, CI and LCI obtained from the simulation study for 30% levels of missingness from SMI, CMI, PMM-I, PMM-II, GAM-I and GAM-II analysis under MCAR-2nd scenario

Parameter	Estimate	SE	LL	UL	LCI	Estimate	SE	LL	UL	LCI
			SMI					***CMI***		
Intercept	23.9970	2.5434	19.0120	28.9821	9.9702	15.6537	2.5368	10.6816	20.6259	9.9443
Sex (M=1)	9.7396	1.5933	6.6167	12.8625	6.2457	13.7259	1.5892	10.6112	16.8407	6.2295
Age	0.0082	0.0436	-0.0773	0.0937	0.1710	0.0122	0.0435	-0.0730	0.0975	0.1705
AVERP	4.2073	0.4038	3.4160	4.9987	1.5827	5.9384	0.4027	5.1490	6.7277	1.5786
			PMM-I					***GAM-I***		
Intercept	15.2805	3.0033	9.3940	21.1669	11.7729	15.5245	2.5370	10.5520	20.4970	9.9450
Sex (M=1)	13.7149	1.8814	10.0274	17.4024	7.3750	13.7482	1.5893	10.6332	16.8632	6.2300
Age	0.0161	0.0515	-0.0849	0.1170	0.2019	0.0124	0.0435	-0.0729	0.0976	0.1705
AVERP	6.0476	0.4768	5.1131	6.9820	1.8689	5.9728	0.4027	5.1834	6.7621	1.5787
			PMM-II					***GAM-II***		
Intercept	15.1416	3.3545	8.5668	21.7164	13.1496	14.8914	5.3822	4.3423	25.4405	21.0982
Sex (M=1)	13.7893	2.1083	9.6571	17.9215	8.2644	13.8308	3.4935	6.9836	20.6780	13.6944
Age	0.0166	0.0572	-0.0955	0.1288	0.2242	0.0239	0.0740	-0.1211	0.1688	0.2899
AVERP	6.0469	0.5284	5.0113	7.0825	2.0712	6.0145	0.7806	4.4845	7.5446	3.0601

It can be seen from Table 16b that, SMI underestimate the estimates and the SEs for the parameters. Other single imputation methods perform well though there was a slight underestimation of SE when CMI and GAM-I were used. For MI methods, the best results were obtained under PMM-II. Estimates under GAM-II were very close to the true ones but the SEs were overestimated. Moreover, SMI, CMI and GAM methods underestimate the variability in the response. Similar pattern of results was obtained for the case of 50% level of missingness (Table 16c).

Table 16c: Estimates, SE, CI and LCI obtained from the simulation study for 50% levels of missingness from SMI, CMI, PMM-I, PMM-II, GAM-I and GAM-II analysis under MCAR-2nd scenario

Parameter	Estimate	SE	LL	UL	LCI	Estimate	SE	LL	UL	LCI
			SMI					***CMI***		
Intercept	29.8184	2.1479	25.6086	34.0283	8.4197	15.4501	2.1433	11.2493	19.6509	8.4017
Sex (M=1)	6.8789	1.3455	4.2417	9.5161	5.2745	13.7808	1.3426	11.1492	16.4123	5.2631
Age	0.0095	0.0368	-0.0627	0.0817	0.1444	0.0187	0.0368	-0.0534	0.0907	0.1441
AVERP	2.9648	0.3410	2.2965	3.6331	1.3366	5.9115	0.3402	5.2446	6.5784	1.3337
			PMM-I					***GAM-I***		
Intercept	15.0643	3.0016	9.1811	20.9475	11.7664	15.1955	2.1433	10.9946	19.3964	8.4018
Sex (M=1)	13.4647	1.8803	9.7792	17.1501	7.3709	13.8165	1.3427	11.1849	16.4481	5.2632
Age	0.0196	0.0515	0.0812	0.1205	0.2018	0.0190	0.0368	-0.0530	0.0910	0.1441
AVERP	6.0863	0.4765	5.1524	7.0203	1.8679	5.9797	0.3402	5.3128	6.6466	1.3337
			PMM-II					***GAM-II***		
Intercept	14.9821	3.6042	7.9179	22.0463	14.1284	14.5623	5.3830	4.0115	25.1130	21.1015
Sex (M=1)	13.5791	2.2485	9.1721	17.9861	8.8140	13.8991	3.4940	7.0508	20.7474	13.6966
Age	0.0218	0.0616	-0.0989	0.1425	0.2413	0.0305	0.0740	-0.1145	0.1755	0.2899
AVERP	6.0716	0.5590	4.9760	7.1672	2.1912	6.0214	0.7808	4.4912	7.5517	3.0606

One shouldn't rely on the averaged values but rather study distribution of estimates and standard errors obtained under the simulation for better comparison of the imputation methods. For this case boxplots were plotted for each parameter (Figure 9). True estimates (SE) for sex, age and AVERP were 13.78 (1.88), 0.015 (0.05) and 5.95 (0.48) respectively.

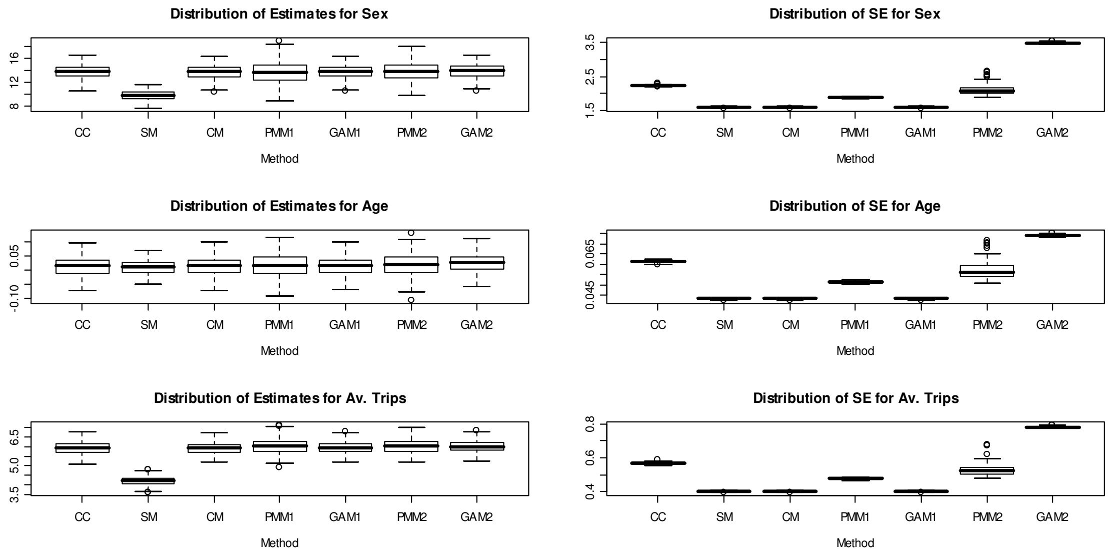


Figure 9: Boxplots of simulated estimates and SE for each parameter under MCAR- 2nd scenario

It can be seen that, SMI performs worse in terms of estimates and GAM-II is doing bad in terms of SE of the estimates. The performance in the SEs might be influenced by the fluctuation (increase) of the variability in the multiple imputed response values.

For assessment of the accuracy of the imputation methods, plot of MASE values obtained for each analysis were plotted (Figure 10).

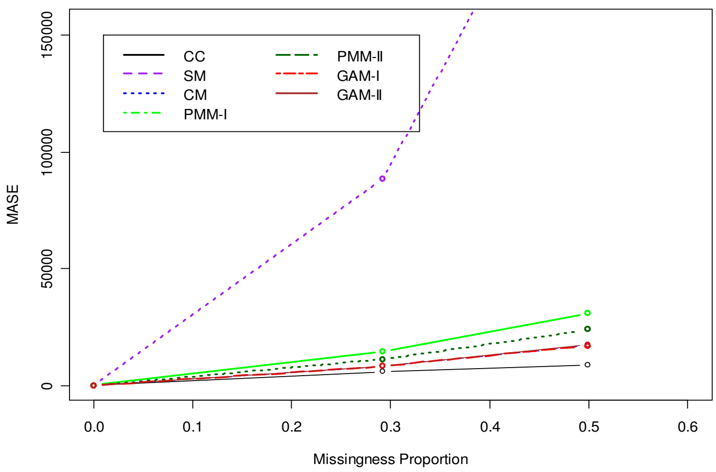


Figure 10: MASE values for different analysis under MCAR- 2nd scenario

It can be seen from Figure 10 that SMI performs poorly compared to other methods. Despite bad performance of GAM-II in terms of SEs for the estimates, its performance in terms of MASE is quite well. This tells us that the difference between the filled-in values (under GAM-II) and their corresponding mean is quite similar to that of OD (recall calculations of coefficients in regression model) thus makes the estimates similar hence closer fitted curve. Boxplots of simulated MASE-values for the different methods can be seen in Figure C, Appendix.

ii. MAR

Results obtained under MAR for different analysis done are presented in Tables 17a, 17b and 17c.

Table 17a: Estimates, SE, CI and LCI obtained from the simulation study for 30% and 50% levels of missingness from CC analysis under MAR-2nd scenario

	30% missingness					50% missingness				
Parameter	**Estimate**	**SE**	**LL**	**UL**	**LCI**	**Estimate**	**SE**	**LL**	**UL**	**LCI**
Intercept	12.7951	3.4681	5.9977	19.5926	13.5949	10.3488	3.9704	2.5668	18.1308	15.5640
Sex (M=1)	13.8973	2.2630	9.4619	18.3328	8.8709	14.0288	2.6690	8.7976	19.2601	10.4626
Age	0.1176	0.0658	-0.0114	0.2466	0.2580	0.2008	0.0804	0.0432	0.3585	0.3152
AVERP	5.7761	0.5693	4.6603	6.8918	2.2315	5.6999	0.6691	4.3885	7.0114	2.6229

Again for CC results were quite similar for both levels of missingness. Except for Age, estimates for other covariates were close to the true ones though SEs were overestimated (Table 17a).

There was a lot of improvement on the estimation of parameters for Age after imputing the missing values compared to CC analysis. Results for both levels are summarized in Tables 17b and 17c.

Table 17b: Estimates, SE, CI and LCI obtained from the simulation study for 30% levels of missingness from SMI, CMI, PMM-I, PMM-II, GAM-I and GAM-II analysis under MAR-2nd scenario

Parameter	**Estimate**	**SE**	**LL**	**UL**	**LCI**	**Estimate**	**SE**	**LL**	**UL**	**LCI**
	SMI					***CMI***				
Intercept	21.7989	2.5250	16.8500	26.7478	9.8978	15.9520	2.5176	11.0175	20.8866	9.8692
Sex (M=1)	9.4248	1.5817	6.3246	12.5250	6.2004	13.9044	1.5772	10.8132	16.9956	6.1824
Age	0.0898	0.0433	0.0049	0.1746	0.1697	0.0148	0.0432	-0.0698	0.0994	0.1692
AVERP	4.0785	0.4008	3.2929	4.8641	1.5712	5.7980	0.3997	5.0146	6.5813	1.5667
	PMM-I					***GAM-I***				
Intercept	14.7496	3.0044	8.8610	20.6383	11.7773	15.9389	2.5177	11.0042	20.8736	9.8694
Sex (M=1)	13.6963	1.8821	10.0074	17.3852	7.3778	13.9327	1.5772	10.8414	17.0240	6.1826
Age	0.0374	0.0515	-0.0636	0.1384	0.2019	0.0101	0.0432	-0.0745	0.0948	0.1692
AVERP	5.9804	0.4769	5.0456	6.9152	1.8696	5.8470	0.3997	5.0636	6.6303	1.5667
	PMM-II					***GAM-II***				
Intercept	14.7228	3.3317	8.1928	21.2528	13.0601	15.3052	5.3888	4.7431	25.8673	21.1242
Sex (M=1)	13.7012	2.1121	9.5616	17.8409	8.2793	14.0154	3.4977	7.1598	20.8710	13.7111
Age	0.0406	0.0618	-0.0805	0.1618	0.2423	0.0217	0.0741	-0.1235	0.1668	0.2903
AVERP	5.9666	0.5304	4.9270	7.0062	2.0792	5.8888	0.7816	4.3568	7.4208	3.0640

As it can be seen from Table 17b, SMI still presents worse results with very low estimates and SEs as compared to the true ones and even changes the significance status for Age. Other single imputation methods were performing well with very similar results obtained between CMI and GAM-I. Multiple imputation methods perform best in terms of both, estimates and SEs.

Few differences were observed under 50% level. Underestimation of SEs under single imputation methods was a bit high as compared to 30% level.

Table 17c: Estimates, SE, CI and LCI obtained from the simulation study for 50% levels of missingness from SMI, CMI, PMM-I, PMM-II, GAM-I and GAM-II analysis under MAR-2nd scenario

Parameter	Estimate	SE	LL	UL	LCI	Estimate	SE	LL	UL	LCI
			SMI					***CMI***		
Intercept	26.3429	2.1549	22.1193	30.5665	8.4472	15.8234	2.1496	11.6103	20.0366	8.4263
Sex (M=1)	6.7440	1.3499	4.0981	9.3898	5.2917	14.0661	1.3466	11.4268	16.7054	5.2786
Age	0.1018	0.0370	0.0294	0.1743	0.1448	0.0201	0.0369	-0.0522	0.0923	0.1445
AVERP	2.9582	0.3421	2.2877	3.6287	1.3410	5.7442	0.3412	5.0754	6.4130	1.3376
			PMM-I					***GAM-I***		
Intercept	13.3912	3.0064	7.4987	19.2837	11.7850	15.8196	2.1500	11.6057	20.0336	8.4279
Sex (M=1)	13.4574	1.8833	9.7661	17.1487	7.3826	14.0995	1.3468	11.4597	16.7393	5.2796
Age	0.0870	0.0516	-0.0140	0.1881	0.2021	0.0119	0.0369	-0.0604	0.0841	0.1445
AVERP	5.9657	0.4772	5.0303	6.9011	1.8708	5.8263	0.3413	5.1573	6.4952	1.3379
			PMM-II					***GAM-II***		
Intercept	13.2758	3.6219	6.1768	20.3747	14.1979	15.1853	5.3937	4.6136	25.7570	21.1434
Sex (M=1)	13.4803	2.3409	8.8920	18.0686	9.1765	14.1823	3.5009	7.3206	21.0441	13.7235
Age	0.0910	0.0712	-0.0485	0.2306	0.2790	0.0234	0.0741	-0.1219	0.1687	0.2905
AVERP	5.9651	0.5839	4.8208	7.1095	2.2888	5.8681	0.7823	4.3347	7.4015	3.0668

Similarly, MI methods had better results with best estimates obtained under nonparametric method (i.e. GAM-II). For clear evaluation of the performance of methods used, Figure 11 presents the distribution of parameter estimates and SEs obtained for each analysis in the simulation runs.

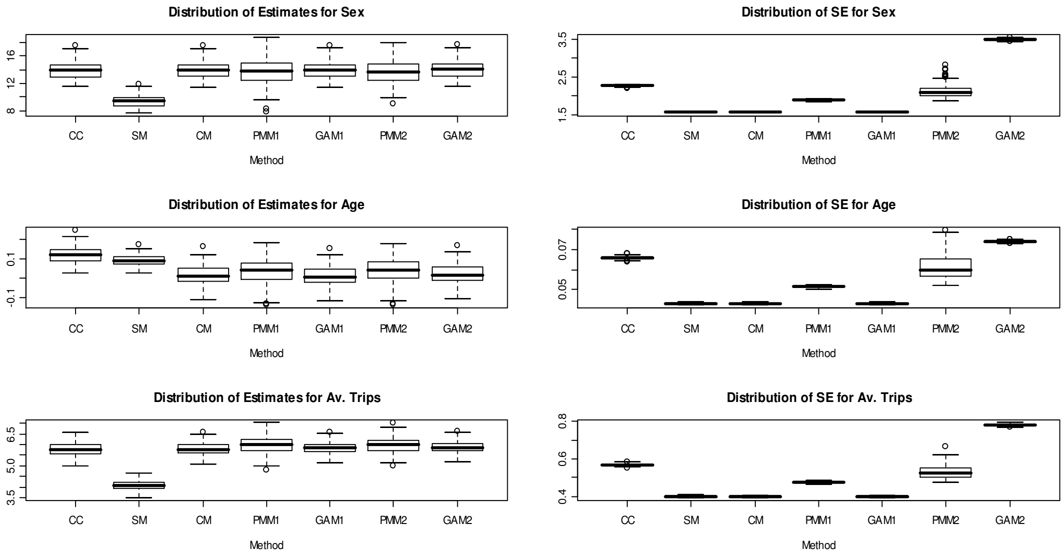


Figure 11: Boxplots of simulated estimates and SE for each parameter under MAR- 2nd scenario

Estimates from SMI were very different from other methods and the SEs from nonparametric multiple imputation method were higher than those from other methods. Figure 12 shows the MASE values for the different methods used.

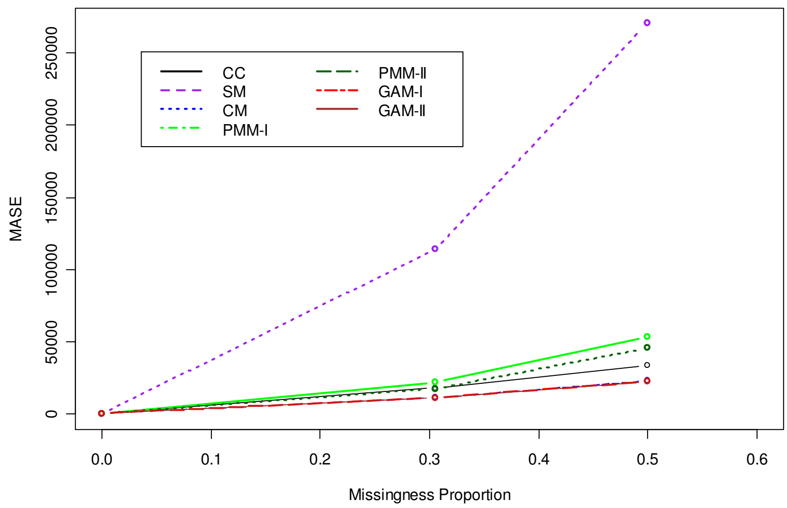


Figure 12: MASE values for different analysis under MAR- 2nd scenario

Performance of all methods except SMI was quite well under 30% level of missingness but differs when the level of missingness increases. Boxplots of MASE-values can be seen in Figure D, Appendix.

iii. MNAR

As it was observed in the single analysis, results obtained under MNAR case were not very promising. The estimates and SEs were biased from the CC analysis. Summarize results are presented in Table 18a, 18b and 18c.

Table 18a: Estimates, SE, CI and LCI obtained from the simulation study for 30% and 50% levels of missingness from CC analysis under MNAR-2nd scenario

	30% missingness					*50% missingness*				
Parameter	**Estimate**	**SE**	**LL**	**UL**	**LCI**	**Estimate**	**SE**	**LL**	**UL**	**LCI**
Intercept	-37.2121	2.8955	-42.8873	-31.5370	11.3503	-47.4749	2.8058	-52.9743	-41.9755	10.9989
Sex (M=1)	6.1717	1.7113	2.8174	9.5259	6.7085	4.9220	1.6875	1.6145	8.2295	6.6150
Age	0.8412	0.0477	0.7477	0.9347	0.1870	0.6272	0.0451	0.5389	0.7155	0.1767
AVERP	3.3705	0.4624	2.4641	4.2768	1.8127	2.5696	0.4759	1.6368	3.5025	1.8656

In both missingness levels, there was a highly overestimation of Age parameters and Sex, while parameters for AVERP were underestimated.

Even after imputing the missing values, no improvement was seen rather the results were still poor. The estimates and SEs were still very low for all methods for both levels of missingness (Table 18b and 18c)

Table 18b: Estimates, SE, CI and LCI obtained from the simulation study for 30% levels of missingness from SMI, CMI, PMM-I, PMM-II, GAM-I and GAM-II analysis under MNAR-2nd scenario

Parameter	**Estimate**	**SE**	**LL**	**UL**	**LCI**	**Estimate**	**SE**	**LL**	**UL**	**LCI**
			SMI					***CMI***		
Intercept	-18.2426	1.9360	-22.0371	-14.4481	7.5890	-40.8957	1.9146	-44.6483	-37.1430	7.5054
Sex (M=1)	4.9477	1.2128	2.5706	7.3247	4.7540	6.1379	1.1994	3.7871	8.4887	4.7017
Age	0.5672	0.0332	0.5022	0.6323	0.1301	0.9233	0.0328	0.8590	0.9877	0.1287
AVERP	2.0303	0.3073	1.4280	2.6327	1.2047	3.4311	0.3039	2.8354	4.0268	1.1914
			PMM-I					***GAM-I***		
Intercept	-39.3324	2.2138	-43.6714	-34.9933	8.6781	-40.8859	1.9131	-44.6356	-37.1361	7.4995
Sex (M=1)	5.8875	1.3868	3.1693	8.6057	5.4363	6.1925	1.1985	3.8435	8.5415	4.6980
Age	0.9020	0.0380	0.8276	0.9764	0.1488	0.9189	0.0328	0.8546	0.9832	0.1286
AVERP	3.3097	0.3514	2.6209	3.9985	1.3776	3.4774	0.3037	2.8821	4.0726	1.1905
			PMM-II					***GAM-II***		
Intercept	-39.3365	2.5074	-44.2511	-34.4219	9.8292	-41.3653	4.0791	-49.3604	-33.3703	15.9901
Sex (M=1)	5.9068	1.5308	2.9065	8.9071	6.0007	6.2551	2.6475	1.0659	11.4442	10.3783
Age	0.9032	0.0422	0.8204	0.9859	0.1655	0.9276	0.0561	0.8177	1.0375	0.2198
AVERP	3.2858	0.4107	2.4809	4.0907	1.6098	3.5090	0.5917	2.3492	4.6688	2.3195

All methods underestimate the variability of the response values to almost 50% less. Plot of all standard deviations obtained for all analysis from MCAR to MNAR case can be seen in Figure E, Appendix.

Table 18c: Estimates, SE, CI and LCI obtained from the simulation study for 50% levels of missingness from SMI, CMI, PMM-I, PMM-II, GAM-I and GAM-II analysis under MNAR-2nd scenario

Parameter	Estimate	SE	LL	UL	LCI	Estimate	SE	LL	UL	LCI
			SMI					***CMI***		
Intercept	-27.2845	1.3614	-29.9528	-24.6161	5.3366	-48.3713	1.3406	-50.9990	-45.7436	5.2553
Sex (M=1)	2.7358	0.8528	1.0642	4.4073	3.3431	4.7543	0.8398	3.1082	6.4003	3.2922
Age	0.3256	0.0233	0.2798	0.3713	0.0915	0.6530	0.0230	0.6079	0.6980	0.0901
AVERP	1.0059	0.2161	0.5823	1.4295	0.8472	2.6368	0.2128	2.2197	3.0539	0.8343
			PMM-I					***GAM-I***		
Intercept	-47.2626	1.8521	-50.8927	-43.6326	7.2601	-48.2737	1.3402	-50.9005	-45.6469	5.2536
Sex (M=1)	4.5979	1.1602	2.3239	6.8719	4.5480	4.7581	0.8396	3.1126	6.4037	3.2911
Age	0.6384	0.0318	0.5762	0.7007	0.1245	0.6512	0.0230	0.6062	0.6963	0.0901
AVERP	2.5989	0.2940	2.0227	3.1752	1.1525	2.6271	0.2128	2.2101	3.0441	0.8340
			PMM-II					***GAM-II***		
Intercept	-47.2583	2.3566	-51.8773	-42.6394	9.2379	5.2536	3.3982	-55.3340	-42.0132	13.3208
Sex (M=1)	4.5480	1.4295	1.7462	7.3498	5.6036	3.2911	2.2058	0.4870	9.1336	8.6466
Age	0.6384	0.0387	0.5624	0.7143	0.1519	0.0901	0.0467	0.5670	0.7500	0.1830
AVERP	2.6051	0.4193	1.7833	3.4269	1.6436	0.8340	0.4928	1.6876	3.6194	1.9319

Boxplots of the estimates and standard errors obtained under each analysis and plot of MASE values can be seen in Figure 13 and Figure 14 respectively.

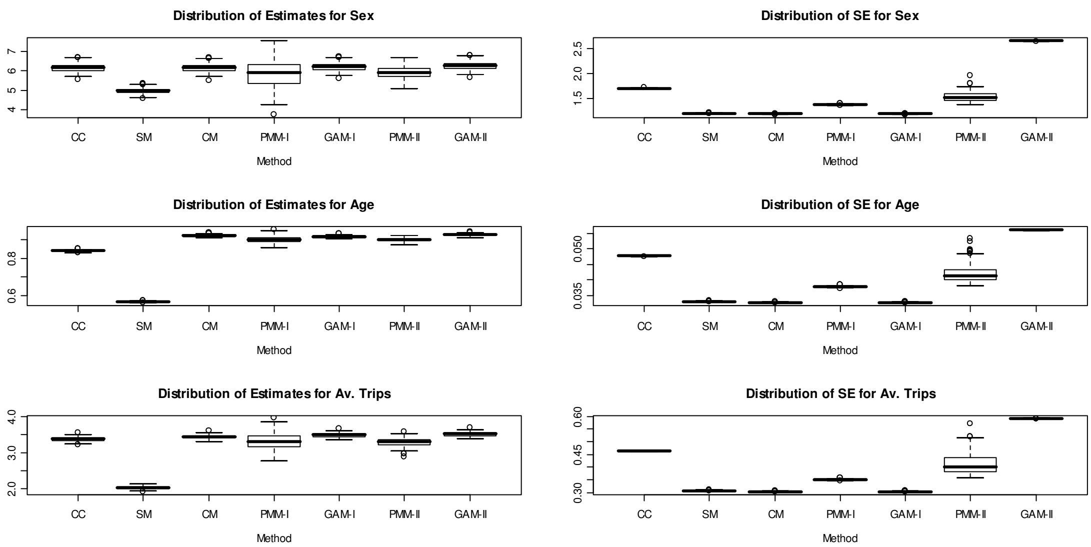


Figure 13: Boxplots of simulated estimates and SE for each parameter under MNAR- 2nd scenario

For the distribution, similar pattern was observed for both levels of missingness. It can be seen, the variability of the estimates and SEs within runs is not very high in all covariates but the estimates under SMI were very different compared to other methods

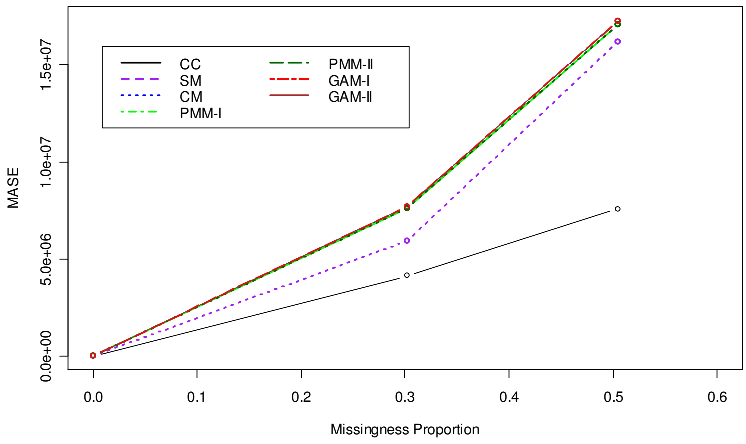


Figure 14: MASE values for different analysis under MNAR- 2nd scenario

It can be seen from Figure 14 that performance of CC is more reliable than any of the imputation method. Boxplots for the MASE values are present in Figure F, Appendix

4.5. Part III: Missingness in covariates

The same exercise was repeated for the case where the missingness is in a covariate (Age). As it has been said before, missingness model of a 2nd scenario (single function) was used to generate missingness probabilities then the missingness indicators for each observation (similar to what is explained in part II). Results presented in here reports average of the estimates and SEs for all models fitted from the simulation runs for each of the missingness mechanism.

i. MCAR

Results of the CC analysis obtained from simulation study for both levels of missingness are summarized in Table 19a. Generally, values of MASE were very small as compared to the situation when missingness was in response, which implies better performance of the methods used.

Table 19a: Estimates, SE, CI and LCI obtained from the simulation study for 30% and 50% levels of missingness from CC analysis under MCAR-missing in covariate

	30% missingness					*50% missingness*				
Parameter	**Estimate**	**SE**	**LL**	**UL**	**LCI**	**Estimate**	**SE**	**LL**	**UL**	**LCI**
Intercept	15.6643	3.5729	8.6613	22.6672	14.0060	15.6062	4.2501	7.2759	23.9365	16.6605
Sex (M=1)	13.7338	2.2379	9.3476	18.1201	8.7725	13.5376	2.6635	8.3172	18.7580	10.4408
Age	0.0118	0.0613	-0.1083	0.1319	0.2402	0.0151	0.0729	-0.1278	0.1579	0.2858
AVERP	5.9397	0.5672	4.8281	7.0513	2.2232	5.9627	0.6748	4.6401	7.2853	2.6452

As it can be observed from Table 19a, results from CC analysis were very close to the true ones for both levels of missingness. There was just a little overestimation of SEs for the 50% level. For the case of MASE values, CC reported the highest value.

Results obtained after imputing the missing values using different methods are presented in Table 16b and 16c for the case of 30% level and 50% level respectively.

Table 19b: Estimates, SE, CI and LCI obtained from the simulation study for 30% levels of missingness from SMI, CMI, PMM-I, PMM-II, GAM-I and GAM-II analysis under MCAR-missing in covariate

Parameter	**Estimate**	**SE**	**LL**	**UL**	**LCI**	**Estimate**	**SE**	**LL**	**UL**	**LCI**
	SMI					***CMI***				
Intercept	15.5797	3.2731	9.1644	21.9950	12.8305	15.4555	3.0074	9.5611	21.3500	11.7889
Sex (M=1)	13.7855	1.8826	10.0956	17.4753	7.3797	13.7754	1.8830	10.0847	17.4662	7.3815
Age	0.0117	0.0612	-0.1082	0.1317	0.2400	0.0147	0.0516	-0.0864	0.1158	0.2023
AVERP	5.9441	0.4769	5.0093	6.8789	1.8696	5.9477	0.4772	5.0124	6.8829	1.8705
	PMM-I					***GAM-I***				
Intercept	15.7702	3.0126	9.8655	21.6750	11.8095	15.3662	3.2882	8.9214	21.8109	12.8896
Sex (M=1)	13.7896	1.8827	10.0996	17.4796	7.3800	13.7739	1.8836	10.0821	17.4656	7.3835
Age	0.0069	0.0516	-0.0942	0.1080	0.2022	0.0169	0.0612	-0.1030	0.1368	0.2398
AVERP	5.9434	0.4772	5.0081	6.8786	1.8705	5.9489	0.4775	5.0129	6.8848	1.8719
	PMM-II					***GAM-II***				
Intercept	15.7608	3.2881	9.3161	22.2055	12.8894	14.0125	3.1856	7.7688	20.2562	12.4875
Sex (M=1)	13.7889	1.8833	10.0975	17.4802	7.3827	13.7316	1.8838	10.0394	17.4238	7.3844
Age	0.0071	0.0610	-0.1123	0.1266	0.2389	0.0503	0.0575	-0.0625	0.1630	0.2255
AVERP	5.9436	0.4775	5.0077	6.8795	1.8718	5.9672	0.4775	5.0314	6.9031	1.8716

Different from what has been observed in previous parts of this report, single imputation methods performs very well this time. Both estimates and covariates were very similar to the true ones. Results from the multiple imputation methods were also very good.

Table 19c: Estimates, SE, CI and LCI obtained from the simulation study for 50% levels of missingness from SMI, CMI, PMM-I, PMM-II, GAM-I and GAM-II analysis under MCAR-missing in covariate

Parameter	Estimate	SE	LL	UL	LCI	Estimate	SE	LL	UL	LCI
			SMI					***CMI***		
Intercept	15.4647	3.6165	8.3764	22.5531	14.1768	14.8372	3.6515	7.6802	21.9942	14.3140
Sex (M=1)	13.7862	1.8823	10.0969	17.4755	7.3786	13.7532	1.8844	10.0597	17.4467	7.3870
Age	0.0147	0.0728	-0.1279	0.1574	0.2853	0.0301	0.0729	-0.1127	0.1729	0.2856
AVERP	5.9435	0.4767	5.0091	6.8779	1.8688	5.9555	0.4779	5.0188	6.8922	1.8734
			PMM-I					***GAM-I***		
Intercept	15.6611	3.0113	9.7590	21.5632	11.8042	14.9291	3.6437	7.7874	22.0708	14.2834
Sex (M=1)	13.7916	1.8825	10.1020	17.4813	7.3793	13.7571	1.8844	10.0636	17.4506	7.3870
Age	0.0095	0.0516	-0.0916	0.1106	0.2021	0.0278	0.0726	-0.1145	0.1701	0.2846
AVERP	5.9454	0.4771	5.0103	6.8806	1.8703	5.9545	0.4779	5.0178	6.8911	1.8734
			PMM-II					***GAM-II***		
Intercept	15.5825	3.6535	8.4217	22.7434	14.3218	13.9355	3.1840	7.6949	20.1761	12.4812
Sex (M=1)	13.7865	1.8839	10.0940	17.4789	7.3849	13.7266	1.8841	10.0338	17.4194	7.3856
Age	0.0115	0.0723	-0.1302	0.1531	0.2833	0.0523	0.0575	-0.0605	0.1650	0.2255
AVERP	5.9462	0.4780	5.0093	6.8830	1.8738	5.9676	0.4775	5.0318	6.9034	1.8716

There was a slight overestimation of the SEs for the case of 50% level both parametric and nonparametric methods, nevertheless the results were more less similar to the true ones (Table 19c).

To assess the accuracy of the imputations done, plot of comparison of MASE values obtained from different methods is presented in Figure 15.

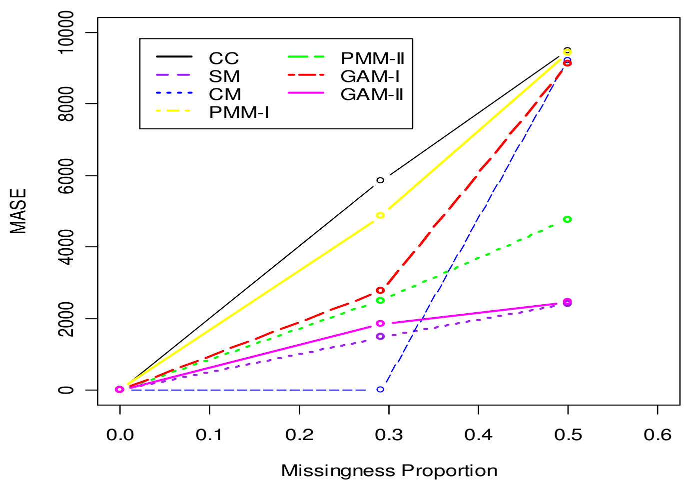


Figure 15: MASE values for different analysis under MCAR- missing in covariate

It can be seen that, CC analysis in doing worse as compared to other methods. Actually, looking at the MASE values, single methods for imputation are performing better than the multiple imputation methods (Figure 15). Boxplots of the MASE values for all methods under MCAR can be seen in Figure G, Appendix. Also, those of estimates and SE obtained under each analysis for 30 and 50% levels of missingness are presented in Figure H, Appendix.

ii. MAR

This part reports results obtained under MAR mechanism. Table 20a showed results from the CC analysis for both 30% and 50% levels of missingness. It was observed that, results of other covariates except Age were close to the true ones. Estimates for Age were over estimated in almost all methods and the significance status was distorted.

Table 20a: Estimates, SE, CI and LCI obtained from the simulation study for 30% and 50% levels of missingness from CC analysis under MAR-missing in covariate

	30% missingness					*50% missingness*				
Parameter	**Estimate**	**SE**	**LL**	**UL**	**LCI**	**Estimate**	**SE**	**LL**	**UL**	**LCI**
Intercept	12.7951	3.4681	5.9977	19.5926	13.5949	10.5477	3.9658	2.7747	18.3207	15.5459
Sex (M=1)	13.8973	2.2630	9.4619	18.3328	8.8709	14.0264	2.6663	8.8004	19.2524	10.4519
Age	0.1176	0.0658	-0.0114	0.2466	0.2580	0.1997	0.0803	0.0423	0.3570	0.3147
AVERP	5.7761	0.5693	4.6603	6.8918	2.2315	5.6681	0.6687	4.3574	6.9788	2.6214

After imputation, similar results were obtained for almost all methods. The estimates for Age were overestimated for about ten (10) times more as compared to the true estimates, but the SE were well estimate (Table 20b).

Table 20b: Estimates, SE, CI and LCI obtained from the simulation study for 30% levels of missingness from SMI, CMI, PMM-I, PMM-II, GAM-I and GAM-II analysis under MAR-missing in covariate

Parameter	**Estimate**	**SE**	**LL**	**UL**	**LCI**	**Estimate**	**SE**	**LL**	**UL**	**LCI**
	SMI					***CMI***				
Intercept	11.9782	3.1395	5.8247	18.1316	12.3069	10.2473	3.1358	4.1011	16.3936	12.2925
Sex (M=1)	13.7249	1.8814	10.0373	17.4125	7.3752	13.6503	1.8813	9.9630	17.3376	7.3746
Age	0.1179	0.0656	-0.0108	0.2465	0.2573	0.1683	0.0656	0.0397	0.2969	0.2572
AVERP	5.9468	0.4762	5.0134	6.8802	1.8668	5.9534	0.4761	5.0203	6.8866	1.8663
	PMM-I					***GAM-I***				
Intercept	12.4038	2.8772	6.7644	18.0431	11.2788	10.2944	3.1361	4.1478	16.4411	12.2933
Sex (M=1)	13.7099	1.8814	10.0223	17.3974	7.3751	13.6532	1.8813	9.9659	17.3406	7.3746
Age	0.1056	0.0546	-0.0015	0.2127	0.2142	0.1667	0.0655	0.0383	0.2952	0.2569
AVERP	5.9478	0.4762	5.0145	6.8812	1.8667	5.9543	0.4761	5.0211	6.8875	1.8663
	PMM-II					***GAM-II***				
Intercept	12.1905	3.1195	6.0763	18.3047	12.2283	10.5965	3.0591	4.6007	16.5924	11.9918
Sex (M=1)	13.7047	1.8820	10.0161	17.3934	7.3773	13.6645	1.8842	9.9714	17.3575	7.3860
Age	0.1115	0.0646	-0.0151	0.2381	0.2532	0.1578	0.0613	0.0378	0.2779	0.2401
AVERP	5.9506	0.4763	5.0171	6.8842	1.8672	5.9528	0.4767	5.0185	6.8872	1.8687

Similar pattern of the results was observed for the case of 50% level of missingness. Nevertheless, the overestimation of the estimate for Age is more. In addition, for this case, performance of single nonparametric methods had the worse results (Table 20c).

Table 20c: Estimates, SE, CI and LCI obtained from the simulation study for 50% levels of missingness from SMI, CMI, PMM-I, PMM-II, GAM-I and GAM-II analysis under MAR-missing in covariate

Parameter	Estimate	SE	LL	UL	LCI	Estimate	SE	LL	UL	LCI
			SMI					***CMI***		
Intercept	9.8014	3.3208	3.2927	16.3101	13.0174	3.8112	3.2868	-2.6309	10.2533	10.2533
Sex (M=1)	13.7540	1.8805	10.0682	17.4399	7.3717	13.6240	1.8775	9.9441	17.3039	17.3039
Age	0.1991	0.0800	0.0422	0.3559	0.3137	0.3955	0.0798	0.2391	0.5519	0.5519
AVERP	5.9221	0.4761	4.9889	6.8553	1.8664	5.8694	0.4754	4.9375	6.8012	6.8012
			PMM-I					***GAM-I***		
Intercept	10.6003	2.7976	5.1170	16.0837	10.9667	3.9764	3.2851	-2.4624	10.4151	12.8775
Sex (M=1)	13.7310	1.8799	10.0464	17.4156	7.3693	13.6321	1.8776	9.9521	17.3122	7.3601
Age	0.1748	0.0565	0.0640	0.2856	0.2216	0.3896	0.0796	0.2336	0.5457	0.3121
AVERP	5.9153	0.4760	4.9824	6.8483	1.8659	5.8717	0.4754	4.9398	6.8036	1.8638
			PMM-II					***GAM-II***		
Intercept	10.4941	3.2080	4.2064	16.7819	12.5755	8.6040	2.9743	2.7743	14.4337	11.6594
Sex (M=1)	13.7290	1.8819	10.0405	17.4175	7.3770	13.6978	1.8865	10.0002	17.3954	7.3953
Age	0.1787	0.0756	0.0305	0.3269	0.2964	0.2402	0.0635	0.1158	0.3646	0.2489
AVERP	5.9120	0.4765	4.9780	6.8459	1.8679	5.8965	0.4771	4.9615	6.8315	1.8701

Looking at the distribution of the estimates and SEs of covariates, using boxplots, it was seen that only estimates for Age had some variability, but estimates for other covariates were very similar between simulation runs.

For graphical assessment of the accuracy of the imputation method used, the plot of MASE values is given in Figure 16 and their distributions can be viewed using boxplots in Figure I, Appendix.

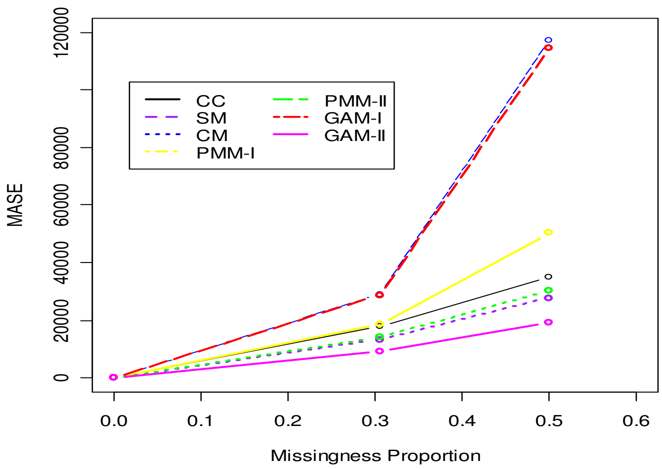


Figure 16: MASE values for different analysis under MAR- missing in covariate

iii. MNAR

In general, very poor estimates were obtained under MNAR and this was for all methods and for both levels of missingness. Summarized results are presented in Table 21a, 21b and 21c for CC analysis, imputation at 30% level and imputation at 50% level, respectively.

Table 21a: Estimates, SE, CI and LCI obtained from the simulation study for 30% and 50% levels of missingness from CC analysis under MNAR-missing in covariate

	30% missingness					*50% missingness*				
Parameter	**Estimate**	**SE**	**LL**	**UL**	**LCI**	**Estimate**	**SE**	**LL**	**UL**	**LCI**
Intercept	-37.2121	2.8955	-42.8873	-31.5370	11.3503	-47.4602	2.8065	-52.9609	-41.9596	11.0013
Sex (M=1)	6.1717	1.7113	2.8174	9.5259	6.7085	4.9316	1.6875	1.6241	8.2392	6.6151
Age	0.8412	0.0477	0.7477	0.9347	0.1870	0.6270	0.0451	0.5386	0.7153	0.1767
AVERP	3.3705	0.4624	2.4641	4.2768	1.8127	2.5642	0.4760	1.6312	3.4972	1.8660

Table 21b: Estimates, SE, CI and LCI obtained from the simulation study for 30% levels of missingness from SMI, CMI, PMM-I, PMM-II, GAM-I and GAM-II analysis under MNAR-missing in covariate

Parameter	**Estimate**	**SE**	**LL**	**UL**	**LCI**	**Estimate**	**SE**	**LL**	**UL**	**LCI**
	SMI					***CMI***				
Intercept	-21.2438	3.4487	-28.0032	-14.4844	13.5188	-64.3729	3.2062	-70.6571	-58.0887	12.5683
Sex (M=1)	12.2825	1.8521	8.6525	15.9126	7.2600	8.4548	1.7307	5.0626	11.8470	6.7844
Age	0.8496	0.0617	0.7287	0.9705	0.2419	1.7557	0.0549	1.6482	1.8633	0.2152
AVERP	6.2832	0.4687	5.3644	7.2019	1.8374	6.5874	0.4366	5.7316	7.4431	1.7114
	PMM-I					***GAM-I***				
Intercept	-68.4682	2.8749	-74.1031	-62.8334	11.2697	-84.1028	2.6262	-89.2502	-78.9553	10.2948
Sex (M=1)	5.9526	1.6670	2.6853	9.2199	6.5346	5.3259	1.5431	2.3014	8.3503	6.0489
Age	1.8108	0.0460	1.7206	1.9010	0.1803	2.1578	0.0418	2.0759	2.2396	0.1637
AVERP	6.6078	0.4194	5.7857	7.4299	1.6442	5.6739	0.3885	4.9125	6.4354	1.5229
	PMM-II					***GAM-II***				
Intercept	-69.2348	3.8966	-76.8721	-61.5976	15.2745	-65.3797	3.9402	-73.1026	-57.6569	15.4457
Sex (M=1)	6.4325	1.7677	2.9678	9.8972	6.9294	6.8748	1.9977	2.9593	10.7902	7.8308
Age	1.8115	0.0703	1.6738	1.9493	0.2755	1.7556	0.0530	1.6518	1.8595	0.2077
AVERP	6.7544	0.4751	5.8231	7.6856	1.8625	5.7087	0.4859	4.7564	6.6611	1.9047

Estimates for Age were very biased (extremely large compared to the true ones) for both, single and multiple imputation methods. The boxplots for the estimates and SEs obtained from the simulation runs are given in Figure J, Appendix. Similar pattern was observed for both levels of missingness.

Table 21c: Estimates, SE, CI and LCI obtained from the simulation study for 50% levels of missingness from SMI, CMI, PMM-I, PMM-II, GAM-I and GAM-II analysis under MNAR-missing in covariate

Parameter	Estimate	SE	LL	UL	LCI	Estimate	SE	LL	UL	LCI
			SMI					***CMI***		
Intercept	-11.9184	3.7718	-19.3112	-4.5256	14.7856	-98.7407	3.2590	-105.1284	-92.3530	12.7754
Sex (M=1)	13.0432	1.8691	9.3798	16.7067	7.3269	6.2426	1.6294	3.0490	9.4362	6.3873
Age	0.6395	0.0708	0.5008	0.7782	0.2774	2.3625	0.0549	2.2548	2.4701	0.2153
AVERP	6.1480	0.4733	5.2204	7.0757	1.8553	6.8688	0.4107	6.0639	7.6738	1.6099
			PMM-I					***GAM-I***		
Intercept	-100.2339	2.5670	-105.2652	-95.2027	10.0625	-89.5526	1.5193	-92.5304	-86.5748	5.9556
Sex (M=1)	2.0748	1.4646	-0.7957	4.9454	5.7411	0.7979	1.0352	-1.2310	2.8269	4.0579
Age	2.1618	0.0362	2.0909	2.2328	0.1419	1.8979	0.0170	1.8646	1.9313	0.0667
AVERP	7.3323	0.3683	6.6106	8.0541	1.4436	3.4220	0.2614	2.9097	3.9343	1.0246
			PMM-II					***GAM-II***		
Intercept	-102.0603	3.7430	3.7430	-94.7241	14.6725	-74.9029	2.8442	-80.4774	-80.4774	11.1492
Sex (M=1)	2.5578	2.3228	2.3228	7.1106	9.1056	2.5576	1.6631	-0.7022	-0.7022	6.5194
Age	2.1862	0.0606	0.0606	2.3051	0.2377	1.6358	0.0251	1.5866	1.5866	0.0984
AVERP	7.1852	0.9086	0.9086	8.9660	3.5616	3.7553	0.3923	2.9864	2.9864	1.5379

The values of ASE for the models fitted from simulated data were very large, with largest values obtained under single nonparametric methods followed by the PMM method. The SMI had the lowest values of ASE. For pictorial presentation of the imputation accuracy, the plot of MASE values by missingness level is given in Figure 17.

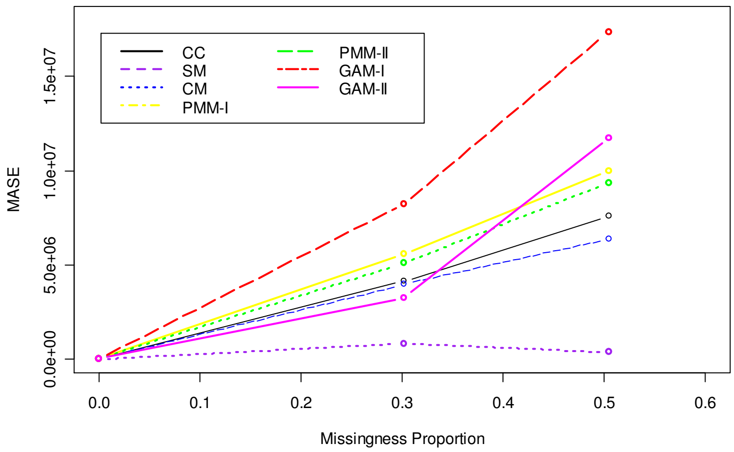


Figure 17: MASE values for different analysis under MAR- missing in covariate

4.6. *Effect of coefficient of missingness model and fitted model on the MAR mechanism*

The idea is to explore if the magnitude of the coefficient (effect) for the covariate in the fitted model and/or in the missingness model can influence the probability of missingness hence influence the missingness mechanism. Age is used for this exercise.

From the missingness model under MAR which is defined as:

$$P = expit(\varphi_0 + \varphi_1 Sex + \varphi_2 Age + \varphi_3 AVERP) \text{..} (i)$$

and the fitted model defined as

$$\hat{Y} = b_0 + b_1 Sex + b_2 Age + b_3 AVERP \text{..} (ii)$$

where the $b_i's$ are unbiased estimators. Hence:

$$E\{b_0\} = \beta_0,\ E\{b_1\} = \beta_1,\ E\{b_2\} = \beta_2 \text{ and } E\{b_3\} = \beta_3$$

From model (ii), one can equate Age as

$$Age = \frac{1}{b_2}(\hat{Y} - b_0 - b_1 Sex - b_3 AVERP)$$

Substituting this in equation (i), we have

$$P^* = expit(\varphi_0 + \varphi_1 Sex + \left(\frac{\varphi_2}{b_2}\right)(\hat{Y} - b_0 - b_1 Sex - b_3 AVERP)) + \varphi_3 AVERP) \text{.................} (iii)$$

Now, lets assume the probabilities for missingness, $P^* = P(R=1)$, are generated from model (iii), where R is the missingness indicator. We would like to explore the change in the missingness pattern as a function of the ratio φ_2 / b_2. After generating individual missingness probabilities, the average was taken over the whole sample and plotted against the ratio φ_2 / b_2. Two scenarios were considered: In the first scenario, the value of φ_2 was kept constant and values for b_2 were changed while in the second scenario the vise versa was done. The original data was used and for the value of b_2 the parameter estimate for Age obtained from the regression model fitted using OD (ref. Table 2) was taken while φ_2 was the coefficient for Age used in the missingness model for generating 30% level under MAR.

For the first case, when φ_2 was fixed, the value of b_2 was changes by multiplying with the sequence of numbers from 0.4 to 1.3 with the interval of 0.1, which gives a total of 10 points.

For the 2nd case, b_2 was kept constant and φ_2 was changed by multiplying with a sequence of numbers from 0.8 to 1.7 with the same interval of 0.1. It should be noted that, no any criteria was used for the selection of the interval, rather was just a random selection.

Figure 18 showed plots of P^* with the fraction φ_2 / b_2 for both cases.

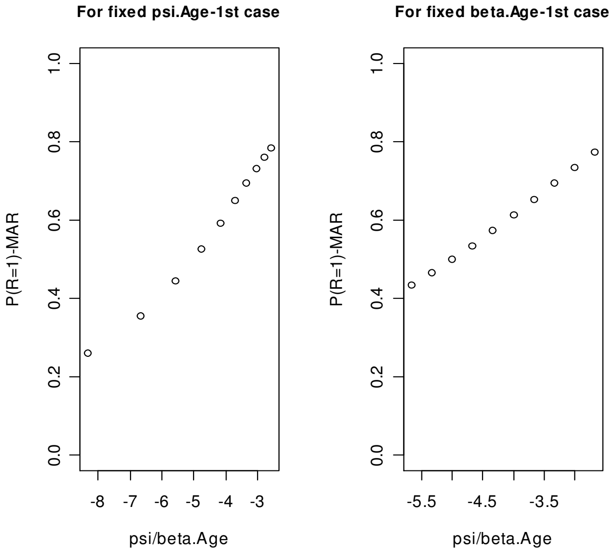
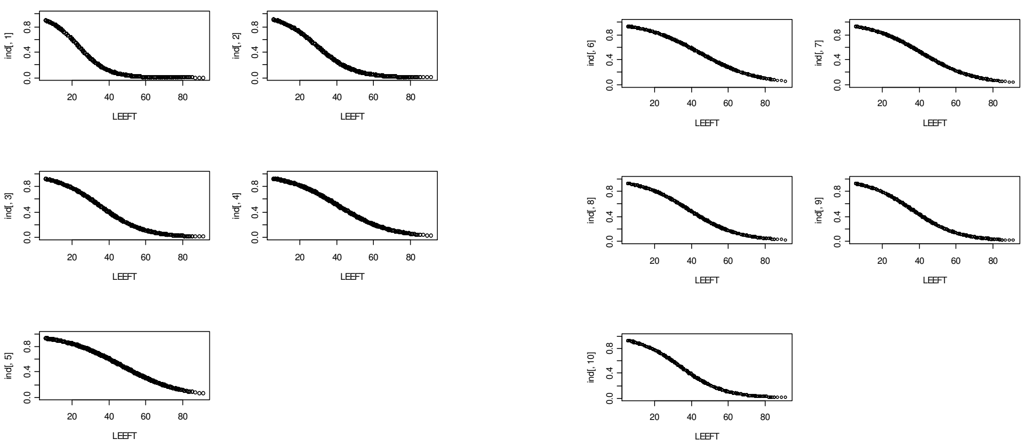


Figure 18: Fraction of the coefficient with the probability for missingness

It can be seen from the plots that, for both cases, the probability for missingness increases as the magnitude of the ratio increases. Plots of missingness probabilities obtained were plotted with Age and the following patterns were observed.



a. **b.**

Figure 19: Pattern of missingness probabilities with Age with a) fixed φ_2 and b) fixed b_2

Plots were done following the order of the sequence (increasing from left to right). Looking careful in the patterns, one can see that, within a specified case, as the value of either φ_2 or b_2 changes, there is a tendency of change in mechanism either from MAR to MNAR or the vise versa. These patterns suggest that, additional effect which is brought by the covariate (age) from the fitted values (or from the response) to the missingness model influences the original mechanism of missingness. Also, suggest that, magnitude of the effect of the covariate into the response (one can relate to collinearity between a covariate and a response) can possibly modify the missingness mechanism.

5. DISCUSSION AND CONCLUSION

Researchers are often faced with non-response problems and most are not familiar with statistical analysis methods that address the missing data problem adequately. However, a key focus of the research is not the non-response itself, but rather proper estimation of model parameters, that's why sometimes is ignored. A major issue with missing data relates to its status. If variables of interest are related to the non-response rate, then dealing with the missing values might be difficult and is important to apply adequate methods to obtain valid results. This study explore different methods of handling missing data in a cross sectional data with main focus in the effect on parameters estimated from models fitted using augumented data obtained from different imputation method. One major problem with missing data is that it is usually unknown how non-response for each variable is generated, i.e. the mechanism. In our study missingness was generated using a pre-specified model, hence assumed that the mechanism is known. This simplifies evaluation of the results.

Despite the simplicity in fitting models using complete cases, results from this type of analysis should be handled careful due to the ignorance on the possible systematic difference between the complete cases and incomplete cases due to the information lost. Lack of this knowledge might results in inference that may not be applicable to the population of all cases, especially when only a small number of complete cases were used. If one is lucky, for complete cases analysis with MCAR data, the group means and variances are likely to stay the same since it is assumed the missing values to be just random values and not depend on anything unobserved or observed. But if this is not the case, then one will be in trouble. It was observed from this study that, the mechanism and the level of missingness available in the data determine the accuracy of complete case analysis (Little and Rubin, 2002).

Use of simple single imputations like filling missing values with mean, median, or conditional mean can sometimes be more dangerous than the complete cases analysis. For instance, mean substitution is conservative because the sample mean does not change and the variance is underestimated. The approach treats missing values as if they were known in the complete-data analyses. Single imputation does not reflect the uncertainty about the predictions of the unknown missing values, and the resulting estimated variances of the parameter estimates will be biased toward zero. However, for complete cases analysis with MCAR data the group means and variances are likely to stay the same hence it not surprising that the method perform better. Results of mean imputation applied to this dataset, showed poor performance for all scenarios and all missingness mechanisms, hence not recommended.

Conditional mean imputation is known to work best when one has missingness in covariates, since the idea behind it is to develop regression model to predict missing covariates from the observed one. However, the method does not functioning well when level of missingness is high. Under CMI, regression imputation that uses the relationship between two or more variables was used. In that way a missing value of response is estimated from the overall relationship between response and other covariates present in the dataset that reduces residual variance. Potential disadvantage of regression imputation is that the method may be sensitive to model misspecification of the regression model (Schenker and Taylor, 1996, Little and Rubin, 2002). Nonparametric method can be used to address some of these issues. In this study, performance conditional mean imputation was questionable especially in the case of MAR and MNAR when the missingness was in the response. However, as it has been mentioned most references, the performance was better in the case of missing in covariate and when the level of missingness is small. When the imputation model was improved better estimates were obtained. This shows that for better and accurate estimate of parameters of interest, choice of imputation model is crucial. Nevertheless, use of nonparametric method like generalized additive model, shows to be worthwhile since in most of the analysis done when missingness was in response, these methods resulted in best estimates. This was not the case, when missingness was in the covariate.

One of the known best imputation methods is multiple imputation. Actually what happened in this method is not estimating each missing value through simulated values but rather to represent a random sample of the missing values. In most cases, this procedure results in valid statistical inferences that properly reflect the uncertainty due to missing values; for example, valid confidence intervals for parameters. However, for this study several concerns raised on its application. Different patterns of the missingness probability with covariates was observed to influence accuracy of this method. For instance if all individuals of low ages (as a covariate) were missing from the study, then there will be no data to be used to make up the imputation, for this case the method performs poorly. Also if the imputation model used is not rich enough to capture the relationship existing between covariates and response the poor imputation will occur. The effect is serious when the fraction of missing information is high and the sample size is large.

Predictive-Mean-Matching method was used to perform multiple imputations in our study. This method imputes the missing value using conditional predictive means close to that of incomplete case. Is the best method among those defined in most of the statistical software that can perform MI but some studies reported increase of bias for this method when applied to big datasets (Lazzeroni, L.C. *et al*). PMM can be seen as a semi-parametric method since it combines elements of regression, nearest-neighbour and hot deck imputation and is also assumed to be less sensitive to misspecifications of the underlying model than for example regression imputation (Schenker and Taylor, 1996). Results of PMM method obtained under MCAR and MAR were good when the missingness was in response and when a moderate probability pattern was used (from the 2nd scenario) and especially when the level of missingness is low. The method showed to perform well even when a single imputation is used. Moreover, under MNAR, PMM showed poor results, in both levels of missingness and all scenarios.

It is quite clear that existence of software that facilitates its use requires the analyst to be careful about the verification of missingness assumptions, the robustness of imputation models, and the appropriateness of inferences to be able to obtain accurate results (Nicholas and Stuart, 2001). So, for anyone who would like to perform imputation of either type should make a note that if the imputation model is seriously flawed in terms of capturing the missing data mechanism, then so will be any analysis based on such imputations. This problem can be avoided by carefully investigating each specific application, by making the best use of knowledge and data about the missing-data mechanism, and by performing various model checking procedures (Barnard and Meng, 1999).

It is not enough to rely on information on parameters computed from a single data sequence since barely reproduce true values of parameters of interest. It is therefore necessary to study any variation available in the estimates, due to this simulation studies have become of great use. For all cases and scenarios, results obtained from simulation studies matches well with what was obtained under single data sequence. However, age had a very low estimate which was also not significant. It was noted that, there was higher variation of age estimates and SEs within simulation runs compared to other significant estimates (sometimes even turn to be significant). It might be possible that, if the number of simulation runs used is not large enough, the estimates can be easly distorted during averaging. This could be the influence of the filled-in data that are assumed to vary between each run.

This study explores the relationship between the strength of the covariate effect on the response and the missingness mechanism. It was surprising to see there is a possibility of change in mechanism from MAR to MNAR as the strength of the effect increases or decreases, i.e. indirect MNAR. These results might explain why for some datasets even the best imputation methods can produce unexpected results. It might be therefore necessary to explore the relationship between variables in the dataset prior to imputing missing ones to avoid vague assumptions which might lead to bias results.

In conclusion, it was observed that, parametric methods for imputing missing values do not always perform well as most of researchers assume. It is then a high time to explore the use nonparametric methods especially when the missingness is available in the response variable. When the missingness is in covariates only and the variability in the data is not high, single imputation methods showed good performance, which helps avoiding use of complicated algorithms to do imputation. It was also observed that missingness mechanism could be influenced by the magnitude of the effect of the covariate in the fitted model or the missingness model involved. However, results from this study should not be generalized in data with other settings than cross sectional to avoid invalid conclusion.

6. RECOMMENDATIONS

- In every research work, effort should be made to collect full and complete datasets to avoid complications of dealing with missing data.
- Assessment of the imputation model used for imputation should be done before applying imputations. Whenever necessary, this model can be constructed manually instead of depending on the built-in models from software/programs.
- Application of other nonparametric methods for doing single or multiple imputations should be emphasized, especially when some of the assumptions used by parametric methods are not fulfilled.
- More research is needed on the observed indirect influence of models parameters and the relationship between covariates and response on the missingness mechanism of the data.

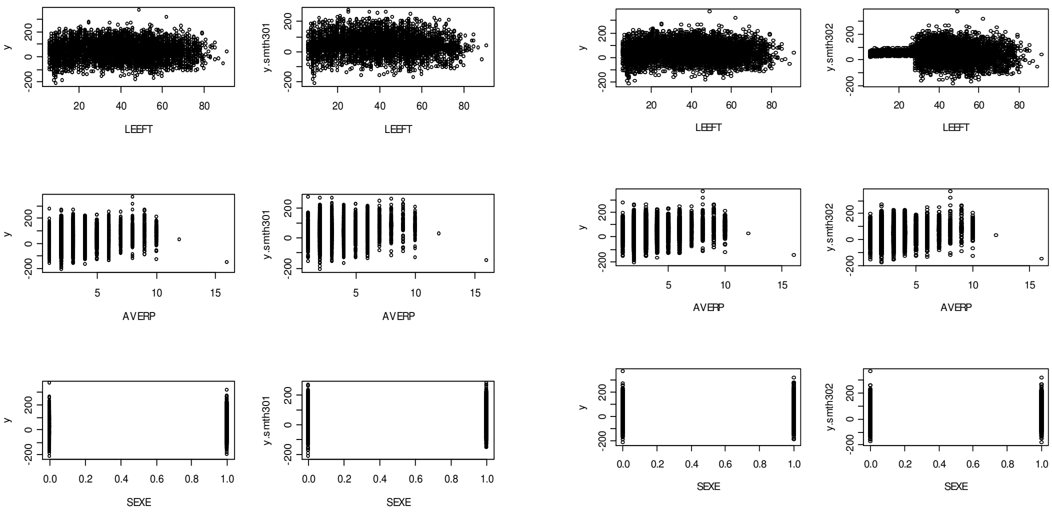
7. REFERENCES

- Activities of the European Union, 2005. http://europa.eu/scadplus/leg/en/lvb/l24040.htm; Accessed July, 2007.
- Barnard, J. and Meng, X.L. (1999). Applications of Multiple Imputation in Medical Studies: From AIDS to NHANES. *Statistical Methods in Medical Research*, **8**, 17–36.
- Hawthorne, G and Elliott, P. (2005). Imputing cross-sectional missing data: comparison of common techniques. *Australian and New Zealand Journal of Psychiatry*. **39**, 583–590.
- Jesson, J. (2001). Cross-sectional studies in prescribing research. *Journal of Clinical Pharmacy and Therapeutics* **26**, 397±403.
- Lazzeroni, L.C., Schenker, N. and Taylor, J.M.G. *Robustness of Multiple-Imputation Techniques to Model Misspecification.* UCLA Dept. of Biostatistics, Los Angeles, CA 90024-1772.
- Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*, New York: John Wiley & Sons, Inc.
- Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, New York.
- Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. New York: Springer-Verlag.
- Moons, E. and Wets, G. (2007). Examining the Impact of Household Interactions when Modelling Travel Duration. *Transportation Research Institute, Hasselt University* (unpublished).
- Nicholas J.H. and Stuart R.L. (2001). Statistical Computing Software Reviews; Multiple Imputation in Practice: Comparison of Software Packages for Regression Models With Missing Variables. *The American Statistician*, **55** (3).
- Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley & Sons, Inc.
- Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*, Kluwer Academic Publishers Group, London.
- Schenker, N. and Taylor, J.M.G. (1996). Partially Parametric Techniques for Multiple Imputation, *Computational Statistics and Data Analysis*, **22**, 425-446.
- World Health Organization (WHO): Road traffic injuries report, April 2007. http://www.who.int/violence_injury_prevention/road_traffic/en/
- Zhang, P. (2003). Multiple Imputation: Theory and Method, *International Statistical Review*, **71**(3), 581-592 (with discussions).

8. APPENDIX

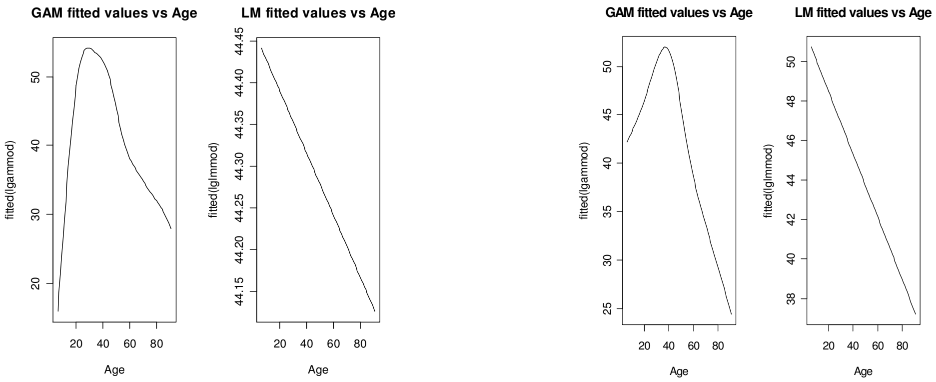
Table A: Summary statistics and percentage of missingness for all variables

Variable	Label	N	Mean	Std Dev	% missing
ABTMH	**Kortste afstand thuis - bus,tram,metro**	5691	2.756	1.728	6.466
ALIJN	**Afstand halte lijnbus tot werk of school**	3653	3.083	2.138	65.864
ALIJNH	**Afstand bushalte**	5691	2.797	1.764	6.466
AMETRO	**Afstand halte metro tot werk of school**	3653	6.588	1.269	65.864
ATRAM	**Afstand halte tram tot werk of school**	3653	6.237	1.712	65.864
ATRAMH	**Afstand tramhalte**	5691	6.419	1.513	6.466
ATREIN	**Afstand halte trein tot werk/school**	3653	5.165	1.635	65.864
ATREINH	**Afstand station**	5691	5.137	1.453	6.466
AVERP	**Average number of trips**	6059	3.471	1.953	0.000
BESTELA	**Aantal bestelwagens**	5691	0.071	0.309	6.466
BROMA	**Aantal bromfietsen**	5691	0.080	0.310	6.466
DIPLOMA	**Hoogst behaald diploma**	5803	5.513	2.340	4.412
FIETSA	**Aantal fietsen**	5385	3.273	1.984	12.516
GACAR	**Gebruik van de autocar**	5486	2.577	1.917	10.445
GAUTO	**Gebruik van de auto**	5793	2.353	0.702	4.592
GBRSNOR	**Gebruik van brom- en snorfiets**	6036	1.141	0.632	0.381
GFIETS	**Gebruik van de fiets**	5847	3.062	1.261	3.626
GLIJN	**Gebruik van de lijnbus**	5672	2.531	1.795	6.823
GMOTOR	**Gebruik van de motor**	6048	1.130	0.638	0.182
GTAXI	**Gebruik van de taxi**	5796	1.436	1.241	4.538
GTRAM	**Gebruik van de tram**	5463	2.305	1.790	10.910
GTREIN	**Gebruik van de trein**	5593	3.316	1.876	8.332
GVLIEG	**Gebruik van het vliegtuig**	5467	2.651	1.960	10.829
HVMWERK	**Hoofdvervoermiddel naar werk/school**	3519	3.816	3.466	72.180
INKCAT	**Gemiddeld maandelijks netto inkomen**	3994	2.379	0.673	51.703
LEDENA	**Aantal leden in huishouden**	5680	3.239	1.308	6.673
LEDENA6	**Aantal leden jonger dan 6 jaar**	5691	0.184	0.487	6.466
LEEFT	**AGE**	6039	39.283	18.653	0.331
LIGGING	**Ligging van de woonplaats**	5680	1.606	0.610	6.673
MOTORA	**Aantal motoren**	5691	0.074	0.279	6.466
PERSWAGA	**Aantal personenwagens**	5691	1.459	0.818	6.466
RGZ	**Relatie gezinshoofd**	6017	1.885	0.844	0.698
RYJAREN	**Aantal jaren bezit rijbewijs**	4256	21.791	11.297	42.364
SNORA	**Aantal snorfietsen**	5691	0.059	0.254	6.466
STAT12	**Individual Profession**	5961	4.835	2.796	1.644
TOTDIST	**Total distance covered**	5484	43.513	65.982	10.485
TOTINK	**Categorie van totale huishoudeninkomen**	5202	2.751	0.815	16.474
TOTTIME	**Total travel time**	5692	75.250	79.155	6.448
VASTKM	**Afstand vast werk/school tot woonplaats**	3512	152.960	211.706	72.523
BS	**Burgerlijke staat**	6034	----	----	0.414
HHNR	**Nummer van het huishouden**	6059	----	----	0.000
HUISPOST	**Postnummer van woonplaats**	6059	----	----	0.000
PERSID	**Persoonsnummer**	6059	----	----	0.000
RYBEWYS	**Bezit rijbewijs om auto te besturen**	6014	----	----	0.748
SEXE	**geslacht**	6031	----	----	0.464
WEEKDAG	**dag van de week (1=maandag)**	6059	----	----	0.000
HUISGEM	**Gemeente van woonplaats**	6059	----	----	0.000
HUISGMTP	**Type gemeente huishouden**	6056	----	----	0.050



a. **b.**

Figure A: Original and nonparametric imputed data with covariates for a) MCAR and b) MAR, first scenario



MCAR MAR

Figure B: Plots of fitted curves for GAM and LM with age by missingness mechanism

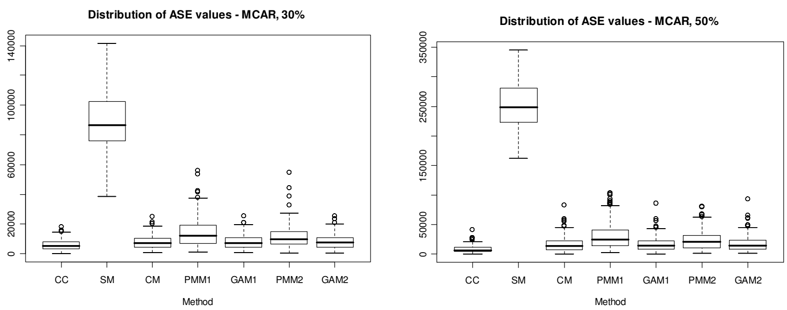


Figure C: Boxplots of simulated MASE-values under MCAR- 2nd scenario

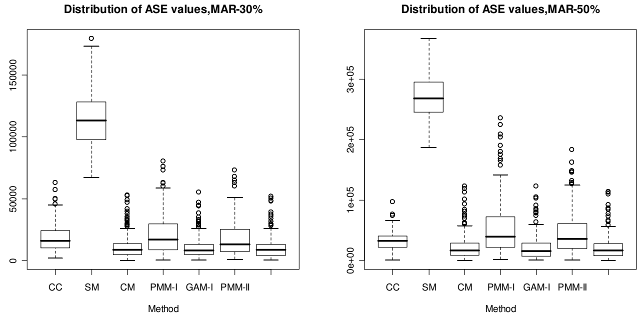


Figure D: Boxplots of simulated MASE-values under MAR- 2nd scenario

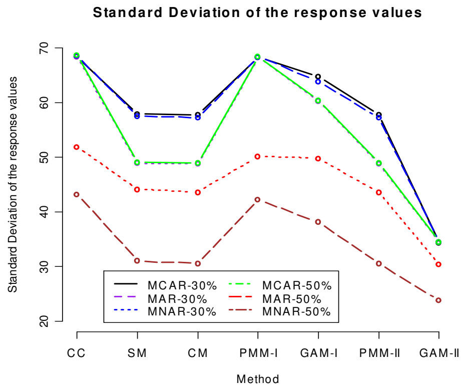


Figure E: Standard deviation of the response values from all methods, 2nd scenario

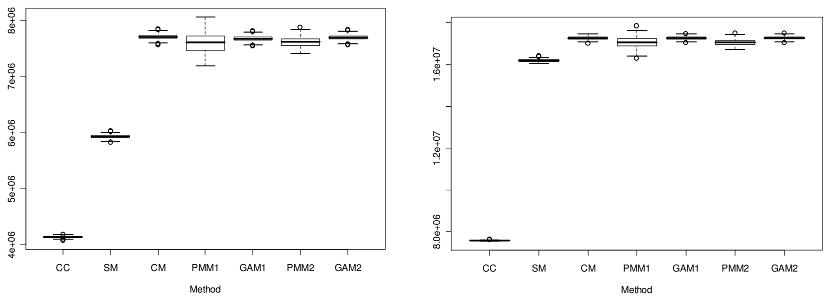


Figure F: Boxplots of simulated MASE-values under MNAR- 2nd scenario

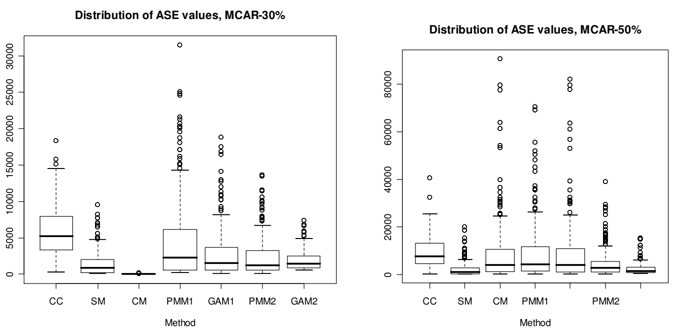


Figure G: Boxplots of simulated MASE-values under MCAR- missing in covariate

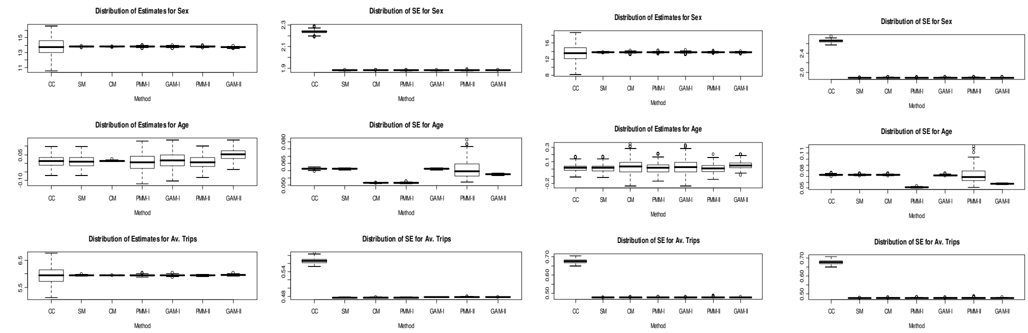


Figure H: Boxplots of estimates and SE, all methods for a) 30% and b) 50% level under MCAR when missingness is in covariate

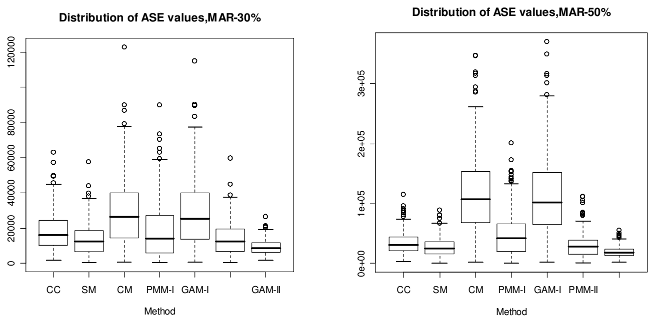


Figure I: Boxplots of simulated MASE-values under MAR- missing in covariate

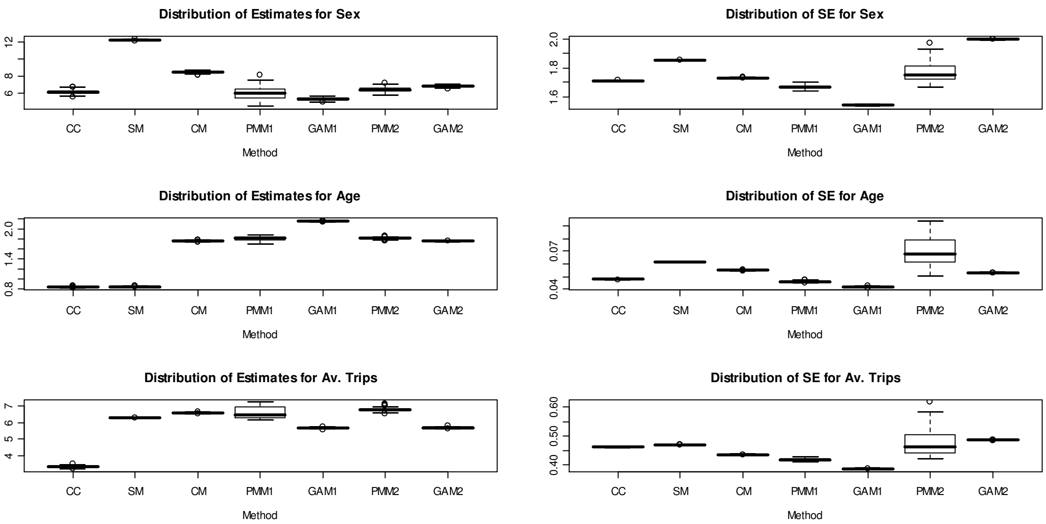


Figure J: Boxplots of estimates and SE, all methods under MNAR when missingness is in covariate

R codes used for Analysis

```
### calling libraries to be used
library(MASS); library(nnet); library(mice)
library(mitools); library(stats); library(mgcv)

#### calling data to fit lm model to obtaine conditional mean for data
generation #####
traffcc<-read.table("D:\\School life in
Belgium\\Biostatistic\\Project\\ANALY2\\traffic.txt", head=T)
SEXE<-traffcc[,1]; DIPLOMA<-traffcc[,2]; GFIETS<-traffcc[,3]
LEEFT<-traffcc[,4]; LEDENA6<-traffcc[,5]; AVERP<-traffcc[,6]
TOTDIST <-traffcc[,7]

## creating new variables- higher orders
Age2<-LEEFT *LEEFT ; Age3<-Age2*LEEFT
trafcc.lm <-lm(TOTDIST ~ LEEFT + SEXE+ DIPLOMA+GFIETS
+LEDENA6+AVERP, data= traffcc)

##### data generation ####
#Coefficients:
beta0<--13.88919 # for Intercept; beta1<--0.14218 # for LEEFT
beta2<-15.82709 # for SEXE ; beta3<-5.34507  # for DIPLOMA
beta4<-3.57714  # for GFIETS   ; beta5<--5.25932 # for LEDENA6
beta6<-4.46466  # for AVERP
#sigmacc<-sd(TOTDIST,na.rm=TRUE)
mucc<-
beta0+beta1*LEEFT+beta2*SEXE+beta3*DIPLOMA+beta4*GFIETS+beta5*L
EDENA6+beta6*AVERP
sigmacc<-67.79

# Generating data to use: Original Data
set.seed(3344)
n= nrow(traffcc) ##5304
y<-rnorm(n,mucc,sigmacc) # we call this original y
y.star<-matrix(y,1,n)

# function to generate the probability for missingness
expit<-function(x){return(exp(x)/(1+exp(x)))}
```

Generating probability for missingness
Part 1: Single analysis: combined missingness models

```
##  MCAR: 30% MISSINGNESS ###
set.seed(235)
psi0LEEFT<-(-1.5); psi0SEXE<-0.5
pLEEFT<-expit(psi0LEEFT); pSEXE<-expit(psi0SEXE)
rLEEFT<-rbinom(n,1,1-pLEEFT); rSEXE<-rbinom(n,1,1-pSEXE)
sum(rLEEFT==1&rSEXE==1)/n

##  MCAR: 50% MISSINGNESS ###
set.seed(236)
psi0LEEFT<--1.5; psi0SEXE<--0.45
pLEEFT<-expit(psi0LEEFT); pSEXE<-expit(psi0SEXE)
rLEEFT<-rbinom(n,1,1-pLEEFT); rSEXE<-rbinom(n,1,1-pSEXE)
sum(rLEEFT==1&rSEXE==1)/n ##### [1] 0.504902 == missingness
##************
##  MAR: 30% MISSINGNESS ###
set.seed(3557)
psi0LEEFT<-(-90.5); psi1LEEFT<-3; psi2LEEFT<-5;
psi3LEEFT<-0.5; psi0SEXE<-47; psi1SEXE<-(-45.5)
psi2SEXE<-37; psi3SEXE<-2
pLEEFT<-expit(psi0LEEFT+ psi1LEEFT*LEEFT+ psi2LEEFT*SEXE+
psi3LEEFT*AVERP)
pSEXE<-expit(psi0SEXE+ psi1SEXE*LEEFT+ psi2SEXE*SEXE+
psi3SEXE*AVERP)
rLEEFT<-rbinom(n,1,1-pLEEFT); rSEXE<-rbinom(n,1,1-pSEXE)
sum(rLEEFT==1&rSEXE==1)/n

##  MAR: 50% MISSINGNESS ###
set.seed(3664)
psi0LEEFT<--110.5; psi1LEEFT<-2.85; psi2LEEFT<--9
psi3LEEFT<-0.5; psi0SEXE<-55; psi1SEXE<--40.5
psi2SEXE<-14; psi3SEXE<-2
pLEEFT<-expit(psi0LEEFT+ psi1LEEFT*LEEFT+ psi2LEEFT*SEXE+
psi3LEEFT*AVERP)
pSEXE<-expit(psi0SEXE+ + psi1SEXE*LEEFT+ psi2SEXE*SEXE+
psi3SEXE*AVERP)
rLEEFT<-rbinom(n,1,1-pLEEFT); rSEXE<-rbinom(n,1,1-pSEXE)
sum(rLEEFT==1&rSEXE==1)/n

##  MNAR: 30% MISSINGNESS ###
set.seed(7642)
psi0LEEFT<-1; psi1LEEFT<-4; psi2LEEFT<-21
psi3LEEFT<-1; psi4LEEFT<-(-2); psi0SEXE<-3
psi1SEXE<-1; psi2SEXE<-1; psi3SEXE<-1; psi4SEXE<-(-3)
pLEEFT<-expit(psi0LEEFT+ psi1LEEFT*LEEFT+ psi2LEEFT*SEXE+
psi3LEEFT*AVERP+psi4LEEFT*y)
pSEXE<-expit(psi0SEXE+ + psi1SEXE*LEEFT+ psi2SEXE*SEXE+
psi3SEXE*AVERP+ psi4SEXE*y)
rLEEFT<-rbinom(n,1,1-pLEEFT); rSEXE<-rbinom(n,1,1-pSEXE)
sum(rLEEFT==1&rSEXE==1)/n

##  MNAR: 50% MISSINGNESS ###
set.seed(6742)
psi0LEEFT<-0.9; psi1LEEFT<-2; psi2LEEFT<-1
psi3LEEFT<-1; psi4LEEFT<-(-2); psi0SEXE<-(-0.2)
psi1SEXE<-3; psi2SEXE<-14.45; psi3SEXE<-2; psi4SEXE<-(-3)
pLEEFT<-expit(psi0LEEFT+ psi1LEEFT*LEEFT+ psi2LEEFT*SEXE+
psi3LEEFT*AVERP+psi4LEEFT*y)
pSEXE<-expit(psi0SEXE+ + psi1SEXE*LEEFT+ psi2SEXE*SEXE+
psi3SEXE*AVERP+ psi4SEXE*y)
rLEEFT<-rbinom(n,1,1-pLEEFT); rSEXE<-rbinom(n,1,1-pSEXE)
sum(rLEEFT==1&rSEXE==1)/n
```

complete code with summary results: presented only for 30% MCAR, the rest are similar

```
# creating the matrix with missingness probabilities
t<-rLEEFT+rSEXE
a<-matrix(t,1,n)
r<-matrix(NA,1,n) # matrix ya kuweka missingness indicator
for  (i in 1:n) {
if (a[,i]==2) r[,i]<-0 else r[,i]<-1 }

# creating dataset and generate missingness based on "r"
y.miss30a<-matrix(NA,1,n)
for  (i in 1:n) {
if (r[,i]==1) y.miss30a[,i]<-y.star[,i] else y.miss30a[,i]<-NA }
y.miss30a[,1:15]
y.miss301<-y.miss30a[1,] # change to a vector

# fit0: Original Data
fit.od<-lm(y~SEXE+LEEFT+AVERP)

# fit1: cc for y.miss30a
# Make a dataset to fit a model with cc of "y.miss" values
dataCC301<-matrix(NA,n,4)
# add the variable names on top
dimnames(dataCC301)<-list(1:n,c("LEEFT", "SEXE", "AVERP","y.miss301"))
dataCC301[,1]<-traffcc[,4]; dataCC301[,2]<-traffcc[,1]
dataCC301[,3]<-traffcc[,6]; dataCC301[,4]<-y.miss301
dataCC301[1:4,1:4]
dataCC301<-data.frame(dataCC301)# make it a data frame
#y.miss301<-y.miss30[1,] # change to a vector
fit.cc301<-lm(y.miss301~SEXE+LEEFT+AVERP, data=dataCC301)
#summary(fit.cc301)

# Imputation of y.miss (use a vector)
##### SINGLE IMPUTATION
# define index of TRUE and FALSE
ry<-matrix(T,1,n)
for  (i in 1:n) {
if (r[,i]==0) ry[,i]<-F else ry[,i]<-ry[,i] }

 ## replace NA with mean of the available ones ##
y.miss3011<-y.miss30a[1,]
rep.na<-function(y.miss3011, my.mean=TRUE)
{  if (my.mean) {value<-mean(y.miss3011[!is.na(y.miss3011)])}
for (i in (1:length(y.miss3011))){if (is.na(y.miss3011[i])==TRUE)
{y.miss3011[i]<-value}}
y.miss3011<<-y.miss3011 }
(y.miss3011)
y.miss301.imp<-(rep.na(y.miss3011))

# Make a dataset to fit a model with single imputed "y" values
dataS301<-matrix(NA,n,4)
# add the variable names on top
dimnames(dataS301)<-list(1:n,c("LEEFT", "SEXE",
"AVERP","y.miss301.imp"))
dataS301[,1]<-traffcc[,4]; dataS301[,2]<-traffcc[,1]
dataS301[,3]<-traffcc[,6]; dataS301[,4]<-y.miss301.imp
```

```
dataS301<-data.frame(dataS301)# make it a data frame
# fit2: single imputed y.miss30a (=y.miss301.imp)
#y.miss301<-y.miss30a[1,] # change to a vector
fit.impS301<-lm(y.miss301.imp~SEXE+LEEFT+AVERP,data=dataS301)
#summary(fit.impS301)

##### CONDITIONAL MEANS IMPUTATION
y.miss301111<-y.miss30a[1,]
fit.impCM301o<-lm(y.miss301111~SEXE+LEEFT+AVERP+SEXE*AVERP)
beta.CM301<-summary(fit.impCM301o)$coefficients
#y.miss301111.imp<-y.miss301111

### replacing using fitted values
DD<-
beta.CM301[1,1]+(beta.CM301[2,1]*SEXE)+(beta.CM301[3,1]*LEEFT)+(beta.
CM301[4,1]*AVERP)+(beta.CM301[5,1]*SEXE*AVERP)
y.miss301111.imp<- ifelse((is.na(y.miss301111)),DD,y.miss301111)

### fitting model with condition imputed values
fit.impCM301<-
lm(y.miss301111.imp~SEXE+LEEFT+AVERP+SEXE*AVERP)
#summary(fit.impCM301)

##### fit3: Statistics for Conditional mean imputation for y.miss30a
fit.impCM301<-
lm(y.miss301111.imp~SEXE+LEEFT+AVERP+SEXE*AVERP)
summ.impCM301<-summary(fit.impCM301)
#fitd.impCM301<-cbind(fit.impCM301$fitted.values)

##### MULTIPLE IMPUTATION
dataCC301[1:4,1:4]
imp.CC301<- mice(dataCC301,m=5,maxit=10, seed = 333)
#imp<-mice(dataCC301,predictorMatrix =(1 - diag(1, ncol(trafms))), seed =
3333)
#complete(imp.CC301)[1:10,2] # show some of completed data
#complete(imp.CC301) # show the first completed data matrix

fits<-lm.mids(y.miss301 ~ SEXE+LEEFT+AVERP, imp.CC301)
summary(pool(fits))
#?pool: Pools the results of m repeated complete data analysis
fit.impM301<-summary(MIcombine(fits$analyses))
#fits$analyses[1]
#complete(imp.CC301)[1:10,] # show some of completed data

### To get fitted values for Multiple imputation from 5 models
#fittd1<-cbind(fits$analyses[[1]]$fitted.values)
#fittd2<-cbind(fits$analyses[[2]]$fitted.values)
#fittd3<-cbind(fits$analyses[[3]]$fitted.values)
#fittd4<-cbind(fits$analyses[[4]]$fitted.values)
#fittd5<-cbind(fits$analyses[[5]]$fitted.values)
#fittd.impM301<-cbind(rowSums(cbind(fittd1)+cbind(fittd2)+cbind(fittd3)+
cbind(fittd4)+cbind(fittd5))/5)

# calculating confidence interval of estimates and its length
#for fit.od
CI.od1<-c(fit.od$coefficients[1]-1.96*summary(fit.od)$coef[, "Std. Error"]
[1],fit.od$coefficients[1]+1.96*summary(fit.od)$coef[, "Std. Error"] [1])
CI.od2<-c(fit.od$coefficients[2]-1.96*summary(fit.od)$coef[, "Std. Error"]
[2],fit.od$coefficients[2]+1.96*summary(fit.od)$coef[, "Std. Error"] [2])
CI.od3<-c(fit.od$coefficients[3]-1.96*summary(fit.od)$coef[, "Std. Error"]
[3],fit.od$coefficients[3]+1.96*summary(fit.od)$coef[, "Std. Error"] [3])
CI.od4<-c(fit.od$coefficients[4]-1.96*summary(fit.od)$coef[, "Std. Error"]
[4],fit.od$coefficients[4]+1.96*summary(fit.od)$coef[, "Std. Error"] [4])
avCI.od1<-sum(CI.od1)/2; avCI.od2<-sum(CI.od2)/2
avCI.od3<-sum(CI.od3)/2; avCI.od4<-sum(CI.od4)/2
avCI.od<-c(avCI.od1,avCI.od2,avCI.od3,avCI.od4)

#for fit.cc1
CI.cc3011<-c(fit.cc301$coefficients[1]-1.96*summary(fit.cc301) $coef[, "Std.
Error"] [1],fit.cc301$coefficients[1]+ 1.96*summary(fit.cc301)$coef[, "Std.
Error"] [1])
CI.cc3012<-c(fit.cc301$coefficients[2]-1.96*summary(fit.cc301) $coef[, "Std.
Error"] [2],fit.cc301$coefficients[2]+1.96*summary (fit.cc301)$coef[, "Std.
Error"] [2])
CI.cc3013<-c(fit.cc301$coefficients[3]-1.96*summary(fit.cc301) $coef[, "Std.
Error"] [3],fit.cc301$coefficients[3]+1.96*summary (fit.cc301)$coef[, "Std.
Error"] [3])
CI.cc3014<-c(fit.cc301$coefficients[4]-1.96*summary(fit.cc301) $coef[, "Std.
Error"] [4],fit.cc301$coefficients[4]+1.96*summary (fit.cc301)$coef[, "Std.
Error"] [4])
```

```
avCI.cc3011<-sum(CI.cc3011)/2; avCI.cc3012<-sum(CI.cc3012)/2
avCI.cc3013<-sum(CI.cc3013)/2;avCI.cc3014<-sum(CI.cc3014)/2
avCI.cc301<-c(avCI.cc3011,avCI.cc3012,avCI.cc3013,avCI.cc3014)

#for fit.impS301
CI.impS3011<-c(fit.impS301$coefficients[1]-1.96*summary
(fit.impS301)$coef[, "Std. Error"] [1],fit.impS301$coefficients[1]+
1.96*summary(fit.impS301)$coef[, "Std. Error"] [1])
CI.impS3012<-c(fit.impS301$coefficients[2]-1.96*
summary(fit.impS301)$coef[, "Std. Error"] [2],fit.impS301
$coefficients[2]+1.96*summary(fit.impS301)$coef[, "Std. Error"] [2])
CI.impS3013<-c(fit.impS301$coefficients[3]-
1.96*summary(fit.impS301)$coef[, "Std. Error"] [3],fit.impS301
$coefficients[3]+1.96*summary(fit.impS301)$coef[, "Std. Error"] [3])
CI.impS3014<-c(fit.impS301$coefficients[4]-
1.96*summary(fit.impS301)$coef[, "Std. Error"] [4],fit.impS301
$coefficients[4]+1.96*summary(fit.impS301)$coef[, "Std. Error"] [4])
avCI.impS3011<-sum(CI.impS3011)/2; avCI.impS3012<-sum(CI.impS3012)/2
avCI.impS3013<-sum(CI.impS3013)/2; avCI.impS3014<-sum(CI.impS3014)/2
avCI.impS301<-
c(avCI.impS3011,avCI.impS3012,avCI.impS3013,avCI.impS3014)

#for fit.impCM301
CI.impCM3011<-c(fit.impCM301$coefficients[1]-1.96*summary
(fit.impCM301)$coef[, "Std. Error"] [1],fit.impCM301$coefficients[1]
+1.96*summary(fit.impCM301)$coef[, "Std. Error"] [1])
CI.impCM3012<-c(fit.impCM301$coefficients[2]-1.96*summary
(fit.impCM301)$coef[, "Std. Error"] [2],fit.impCM301$coefficients[2]+
1.96*summary(fit.impCM301)$coef[, "Std. Error"] [2])
CI.impCM3013<-c(fit.impCM301$coefficients[3]-1.96*summary
fit.impCM301)$coef[, "Std. Error"] [3],fit.impCM301$coefficients[3]+
1.96*summary(fit.impCM301)$coef[, "Std. Error"] [3])
CI.impCM3014<-c(fit.impCM301$coefficients[4]-1.96*summary
(fit.impCM301)$coef[, "Std. Error"] [4],fit.impCM301$coefficients[4]+
1.96*summary(fit.impCM301)$coef[, "Std. Error"] [4])
avCI.impCM3011<-sum(CI.impCM3011)/2; avCI.impCM3012<-
sum(CI.impCM3012)/2
avCI.impCM3013<-sum(CI.impCM3013)/2; avCI.impCM3014<-
sum(CI.impCM3014)/2
avCI.impCM301<-c(avCI.impCM3011,avCI.impCM3012,
avCI.impCM3013,avCI.impCM3014)

# calculating length confidence interval of estimates for fit.impM301
#for fit.impM301
CI.impM3011<-c(fit.impM301[, "(lower"][1],fit.impM301[, "upper)"][1])
CI.impM3012<-c(fit.impM301[, "(lower"][2],fit.impM301[, "upper)"][2])
CI.impM3013<-c(fit.impM301[, "(lower"][3],fit.impM301[, "upper)"][3])
CI.impM3014<-c(fit.impM301[, "(lower"][4],fit.impM301[, "upper)"][4])
avCI.impM3011<-sum(CI.impM3011)/2; avCI.impM3012<-
sum(CI.impM3012)/2
avCI.impM3013<-sum(CI.impM3013)/2; avCI.impM3014<-
sum(CI.impM3014)/2
avCI.impM301<-
c(avCI.impM3011,avCI.impM3012,avCI.impM3013,avCI.impM3014)

### MAJIBU
jibu.od<-matrix(0,4,5)
#dimnames(jibu)<-list(1:4,c("Estimate", "std error",
"Llimit","Ulimit","LengthCI"))
col<-c("Estimate", "std error", "Llimit","Ulimit","LengthCI")
rows<-c("Intercept", "SEXE", "LEEFT","AVERP")
dimnames(jibu.od)<-list(rows,col)
a1<-c(fit.od$coefficients[1],summary(fit.od)$coef[,"Std.
Error"][1],CI.od1,avCI.od[1]);  jibu.od[1,]<-a1
a2<-c(fit.od$coefficients[2],summary(fit.od)$coef[,"Std.
Error"][2],CI.od2,avCI.od[2]);  jibu.od[2,]<-a2
a3<-c(fit.od$coefficients[3],summary(fit.od)$coef[,"Std.
Error"][3],CI.od3,avCI.od[3]);  jibu.od[3,]<-a3
a4<-c(fit.od$coefficients[4],summary(fit.od)$coef[,"Std.
Error"][4],CI.od4,avCI.od[4]);  jibu.od[4,]<-a4


jibu.cc301<-matrix(0,4,5)
#dimnames(jibu)<-list(1:4,c("Estimate", "std error",
"Llimit","Ulimit","LengthCI"))
col<-c("Estimate", "std error", "Llimit","Ulimit","LengthCI")
rows<-c("Intercept", "SEXE", "LEEFT","AVERP")
dimnames(jibu.cc301)<-list(rows,col)
b1<-c(fit.cc301$coefficients[1],summary(fit.cc301)$coef[,"Std.
Error"][1],CI.cc3011,avCI.cc301[1]);  jibu.cc301[1,]<-b1
```

```
b2<-c(fit.cc301$coefficients[2],summary(fit.cc301)$coef[,"Std.
Error"][2],CI.cc3012,avCI.cc301[2]);  jibu.cc301[2,]<-b2
b3<-c(fit.cc301$coefficients[3],summary(fit.cc301)$coef[,"Std.
Error"][3],CI.cc3013,avCI.cc301[3]); jibu.cc301[3,]<-b3
b4<-c(fit.cc301$coefficients[4],summary(fit.cc301)$coef[,"Std.
Error"][4],CI.cc3014,avCI.cc301[4]);  jibu.cc301[4,]<-b4

jibu.impS301<-matrix(0,4,5)
#dimnames(jibu)<-list(1:4,c("Estimate", "std error",
"Llimit","Ulimit","LengthCI"))
col<-c("Estimate", "std error", "Llimit","Ulimit","LengthCI")
rows<-c("Intercept", "SEXE", "LEEFT","AVERP")
dimnames(jibu.impS301)<-list(rows,col)
s1<-c(fit.impS301$coefficients[1],summary(fit.impS301)$coef[,"Std.
Error"][1],CI.impS3011,avCI.impS301[1]);  jibu.impS301[1,]<-s1
s2<-c(fit.impS301$coefficients[2],summary(fit.impS301)$coef[,"Std.
Error"][2],CI.impS3012,avCI.impS301[2]);  jibu.impS301[2,]<-s2
s3<-c(fit.impS301$coefficients[3],summary(fit.impS301)$coef[,"Std.
Error"][3],CI.impS3013,avCI.impS301[3]);  jibu.impS301[3,]<-s3
s4<-c(fit.impS301$coefficients[4],summary(fit.impS301)$coef[,"Std.
Error"][4],CI.impS3014,avCI.impS301[4]);  jibu.impS301[4,]<-s4

jibu.impCM301<-matrix(0,4,5)
#dimnames(jibu)<-list(1:4,c("Estimate", "std error",
"Llimit","Ulimit","LengthCI"))
col<-c("Estimate", "std error", "Llimit","Ulimit","LengthCI")
rows<-c("Intercept", "SEXE", "LEEFT","AVERP")
dimnames(jibu.impCM301)<-list(rows,col)
s1<-c(fit.impCM301$coefficients[1],summary(fit.impCM301) $coef[,"Std.
Error"][1],CI.impCM3011,avCI.impCM301[1]);  jibu.impCM301[1,]<-s1
s2<-c(fit.impCM301$coefficients[2],summary(fit.impCM301) $coef[,"Std.
Error"][2],CI.impCM3012,avCI.impCM301[2]);  jibu.impCM301[2,]<-s2
s3<-c(fit.impCM301$coefficients[3],summary(fit.impCM301) $coef[,"Std.
Error"][3],CI.impCM3013,avCI.impCM301[3]);  jibu.impCM301[3,]<-s3
s4<-c(fit.impCM301$coefficients[4],summary(fit.impCM301) $coef[,"Std.
Error"][4],CI.impCM3014,avCI.impCM301[4]);  jibu.impCM301[4,]<-s4

jibu.impM301<-matrix(0,4,5)
#dimnames(jibu)<-list(1:4,c("Estimate", "std error",
"Llimit","Ulimit","LengthCI"))
col<-c("Estimate", "std error", "Llimit","Ulimit","LengthCI")
rows<-c("Intercept", "SEXE", "LEEFT","AVERP")
dimnames(jibu.impM301)<-list(rows,col)
m1<-c(fit.impM301[, "results"][1],fit.impM301[, "se"][1],fit.impM301[,
"(lower"][1],fit.impM301[, "upper)"][1],avCI.impM3011)
jibu.impM301[1,]<-m1
m2<-c(fit.impM301[, "results"][2],fit.impM301[, "se"][2],fit.impM301[,
"(lower"][2],fit.impM301[, "upper)"][2],avCI.impM3012)
jibu.impM301[2,]<-m2
m3<-c(fit.impM301[, "results"][3],fit.impM301[, "se"][3],fit.impM301[,
"(lower"][3],fit.impM301[, "upper)"][3],avCI.impM3013)
jibu.impM301[3,]<-m3
m4<-c(fit.impM301[, "results"][4],fit.impM301[, "se"][4],fit.impM301[,
"(lower"][4],fit.impM301[, "upper)"][4],avCI.impM3014)
jibu.impM301[4,]<-m4
##### END END END #####
```

Part 2: Single analysis: single missingness model

```
##  MCAR: 30% MISSINGNESS ###
set.seed(235); psi1<-(0.89); p1<-expit(psi1)
R1<-rbinom(n,1,1-p1); Pmiss<-sum(R1==1)/n

##  MCAR: 50% MISSINGNESS ###
set.seed(534); psi1<-(0.005); p1<-expit(psi1)
R1<-rbinom(n,1,1-p1); Pmiss<-sum(R1==1)/n

##  MAR: 30% MISSINGNESS ###
set.seed(3557); psi0<-2.95; psi1<-(-0.05);psi2<--0.005; psi3<-(-0.005)
p1<-expit(psi0+ psi1*LEEFT+ psi2*SEXE+ psi3*AVERP)
R1<-rbinom(n,1,1-p1); Pmiss<-sum(R1==1)/n

##  MAR: 50% MISSINGNESS ###
set.seed(5197); psi0<-1.97; psi1<-(-0.05);psi2<-(-0.055); psi3<-(0.005)
p1<-expit(psi0+ psi1*LEEFT+ psi2*SEXE+ psi3*AVERP)
R1<-rbinom(n,1,1-p1); Pmiss502<-sum(R1==1)/n

##  MNAR: 30% MISSINGNESS ###
set.seed(197); psi0<-1.5; psi1<-2.15; psi2<-2.05; psi3<-0.02; psi4<-(-1.0)
p1<-expit(psi0+ psi1*LEEFT+ psi2*SEXE+ psi3*AVERP+psi4*y)
R1<-rbinom(n,1,1-p1); Pmiss303<-sum(R1==1)/n

##  MNAR: 50% MISSINGNESS ###
set.seed(517); psi0<-1.5; psi1<-1.05; psi2<-2.05; psi3<-0.02; psi4<-(-1.0)
p1<-expit(psi0+ psi1*LEEFT+ psi2*SEXE+ psi3*AVERP+psi4*y)
R1<-rbinom(n,1,1-p1)
```

complete code with summary results: presented only for 30% MAR, the rest are similar

```
### Original data with Age by Missingness Indicator
#plot(LEEFT,y, xlab="Age",ylab="y",type="n",cex.main=0.9,
font.main=2,col.main="blue",main="Or.Data with Age by missingness Ind.")
#points(LEEFT[R1==1],y[R1==1],pch=1)
#points(LEEFT[R1==0],y[R1==0],pch=2,col="blue")

#win.graph()
#par(mfrow=c(1,2))
#plot(LEEFT,p1,type="n",xlab="Age",ylab="conditional missing
probability",ylim=range(0,1))
#points(LEEFT[SEXE==1],p1[SEXE==1],pch=1)
#points(LEEFT[SEXE==0],p1[SEXE==0],pch=4,col="blue")
#plot(AVERP,p1,type="n",xlab="Average trips",ylab="conditional missing
probability",ylim=range(0,1))
#points(AVERP[SEXE==1],p1[SEXE==1],pch=1)
#points(AVERP[SEXE==0],p1[SEXE==0],pch=4,col="blue")
##plot(LEEFT,p1, xlab="Age",ylab="conditional missing
probability",ylim=range(0,1))
##plot(AVERP,p1, xlab="Average trips",ylab="conditional missing
probability",ylim=range(0,1))

#plot(LEEFT,pSEXE, ylab="conditional missing probability")
#plot(SEXE,pSEXE, ylab="conditional missing probability")
##plot(SEXE,pLEEFT,  xlab="Sex", ylab="conditional missing probability")
#plot(AVERP,pSEXE, ylab="conditional missing probability")
##plot(AVERP,pLEEFT,  xlab="Average trips",ylab="conditional missing
probability")
##### [1] 0.3073152 == missingness

# creating dataset and generate missingness based on "r"
y.miss30b<-rep(NA,n)
for  (jj in 1:n) {
if (R1[jj]==0) y.miss30b[jj]<-y.star[,jj] else y.miss30b[jj]<-NA }
y.miss302<-y.miss30b

#### fit0: MODEL FROM ORIGINAL DATA
fit.od<-lm(y~SEXE+LEEFT+AVERP)
#summary(fit.od)
fitd.od<-cbind(fit.od$fitted.values)

#### fit1: MODEL FOR THE COMPLETE CASES cc for y.miss30b
fit.cc302<-lm(y.miss302~SEXE+LEEFT+AVERP)
summ.cc302<-summary(fit.cc302)
fitd.cc302<-cbind(fit.cc302$fitted.values)
ASE4302cc<-sum((cbind(fitd.cc302)-cbind(fitd.od[!is.na(y.miss302)]))^2) ##

##### SINGLE MEAN IMPUTATION
y.miss3021<-y.miss30b
## replace NA with mean of the available ones ##
rep.na<-function(y.miss3021, my.mean=TRUE) {
if (my.mean) {value<-mean(y.miss3021[!is.na(y.miss3021)])}
for (i in (1:length(y.miss3021))){if (is.na(y.miss3021[i])==TRUE)
{y.miss3021[i]<-value}}
y.miss3021<<-y.miss3021   }
(y.miss3021)
y.miss302.imp<-(rep.na(y.miss3021))

##### fit2: MODEL USING SINGLE MEAN IMPUTED DATA - 302
fit.sm302<-lm(y.miss302.imp~SEXE+LEEFT+AVERP)
summ.sm302<-summary(fit.sm302)
fitd.sm302<-cbind(fit.sm302$fitted.values)
ASE302sm<-sum((cbind(fitd.sm302)-cbind(fitd.od))^2)

##### CONDITIONAL MEANS IMPUTATION
y.miss30211<-y.miss30b
fit.cm302o<-lm(y.miss30211~SEXE+LEEFT+AVERP+Age2+Age3)
beta.CM302<-summary(fit.cm302o)$coefficients
#y.miss30211.imp<-y.miss30211
### replacing using fitted values
DD302<-
beta.CM302[1,1]+(beta.CM302[2,1]*SEXE)+(beta.CM302[3,1]*LEEFT)+(beta.
CM302[4,1]*AVERP)+(beta.CM302[5,1]*Age2)+(beta.CM302[6,1]*Age3)
```

```
y.miss30211.imp<- ifelse((is.na(y.miss30211)),DD302,y.miss30211)

### fit3: MODEL WITH CONDIONAL IMPUTED VALUES
fit.cm302<-lm(y.miss30211.imp~SEXE+LEEFT+AVERP)
#summ.cm302<-summary(fit.cm302)
fitd.cm302<-cbind(fit.cm302$fitted.values)
ASE302cm<-sum((cbind(fitd.cm302)-cbind(fitd.od))^2)

### MULTIPLE IMPUTATION - 302
### PartI-- Single imputation using PMM
dataCC302<-data.frame(y.miss302,SEXE,LEEFT,AVERP)
imp.CC302I<-mice(dataCC302,m=1,maxit=10, seed = 333)
imp.CC302I<-complete(imp.CC302I)
#complete(imp.CC302I)[1:10,1:4] # show some of completed data
MIfitsI<-lm(y.miss302 ~ SEXE+LEEFT+AVERP, imp.CC302I)
summ.multI302<-summary(MIfitsI)
fitd.multI302<-cbind(MIfitsI$fitted.values)
ASEM302I<-sum((cbind(fitd.multI302)-cbind(fitd.od))^2)

### PartII-- Multiple imputation using PMM
dataCC302<-data.frame(y.miss302,SEXE,LEEFT,AVERP)
imp.CC302II<- mice(dataCC302,m=5,maxit=10, seed = 333)
#complete(imp.CC302II)[1:10,1:4] # show some of completed data
MIfitsII<-lm.mids(y.miss302 ~ SEXE+LEEFT+AVERP, imp.CC302II)
summ.multII302<-summary(MIcombine(MIfitsII$analyses))

### To get fitted values for Multiple imputation from 5 models
fittd1<-cbind(MIfitsII$analyses[[1]]$fitted.values)
fittd2<-cbind(MIfitsII$analyses[[2]]$fitted.values)
fittd3<-cbind(MIfitsII$analyses[[3]]$fitted.values)
fittd4<-cbind(MIfitsII$analyses[[4]]$fitted.values)
fittd5<-cbind(MIfitsII$analyses[[5]]$fitted.values)
fittd.multII302<-cbind(rowSums(cbind(fittd1)+cbind(fittd2)+
cbind(fittd3)+cbind(fittd4)+cbind(fittd5))/5)

### To get data to calculate MASE1
dat1<-complete(imp.CC302II,1)[1];dat2<-complete(imp.CC302II,2)[1]
dat3<-complete(imp.CC302II,3)[1];dat4<-complete(imp.CC302II,4)[1]
dat5<-complete(imp.CC302II,5)[1]
y.mult<-cbind(rowSums(cbind(dat1)+cbind(dat2)+cbind(dat3)+
cbind(dat4)+cbind(dat5))/5)
ASEM302II<-sum((cbind(fittd.multII302)-cbind(fitd.od))^2)

#### GENERALIZED ADDITIVE MODEL==302
y.miss302.gam<-y.miss302
fit.sp302<-gam(y.miss302.gam~ s(LEEFT)+s(AVERP)+SEXE,
family="gaussian",fit=TRUE)
#,family=gaussian(link="identity")), fit=FALSE
#summary(fit.sp302)
#, data=dataCC302
hh302<-predict.gam(fit.sp302,newdata=data.frame(LEEFT,AVERP,SEXE),
type="response")
#hh302<-cbind(hh302)
y.smth302<-ifelse((is.na(y.miss302)),hh302,y.miss302)
sigmag302<-sd(y.smth302) ## sigmag302
fits.smth302<-lm(y.smth302~ SEXE+LEEFT+AVERP)
summary(fits.smth302)
fittd.smth302<-fits.smth302$fitted.values
ASEG302<-sum((cbind(fittd.smth302)-cbind(fitd.od))^2) ##
#ASEG302

#### MULTIPLE IMPUTATION USING GAM == 302
y.miss302.Mgam<-y.miss302
fit.MG302<-gam(y.miss302.Mgam~ s(LEEFT)+s(AVERP)+SEXE,
family="gaussian",fit=TRUE)
hhMG302<-
predict.gam(fit.MG302,newdata=data.frame(LEEFT,AVERP,SEXE),
type="response")
### Data generation/sampling
YY<-matrix(0,n,5)
col<-c("Y.gam1", "Y.gam2", "Y.gam3","Y.gam4","Y.gam5")
rows<-seq(1:n)
dimnames(YY)<-list(rows,col)
sigmaMG302<-sqrt(fit.MG302$sig2)

set.seed(337)
for (bb in 1:n){ YY[bb,1]=rnorm(1,hhMG302[bb],sigmaMG302) }
set.seed(456)
for (bb in 1:n){ YY[bb,2]=rnorm(1,hhMG302[bb],sigmaMG302) }
set.seed(231)
for (bb in 1:n){ YY[bb,3]=rnorm(1,hhMG302[bb],sigmaMG302) }
set.seed(567)
for (bb in 1:n){ YY[bb,4]=rnorm(1,hhMG302[bb],sigmaMG302) }
set.seed(123)
for (bb in 1:n){ YY[bb,5]=rnorm(1,hhMG302[bb],sigmaMG302) }

Y.gam1<-YY[,1]; Y.gam2<-YY[,2]
Y.gam3<-YY[,3]; Y.gam4<-YY[,4]; Y.gam5<-YY[,5]

Y.gamdat<-cbind(rowSums(cbind(Y.gam1)+cbind(Y.gam2)+
cbind(Y.gam3)+cbind(Y.gam4)+cbind(Y.gam5))/5)

### fitting 5 GAM using the 5 datasets
g1.302<-lm(Y.gam1~ SEXE+LEEFT+AVERP)
g2.302<-lm(Y.gam2~ SEXE+LEEFT+AVERP)
g3.302<-lm(Y.gam3~ SEXE+LEEFT+AVERP)
g4.302<-lm(Y.gam4~ SEXE+LEEFT+AVERP)
g5.302<-lm(Y.gam5~ SEXE+LEEFT+AVERP)

####################
#### MULTIPLE IMPUTATION BY HAND FOR GAM
lmids.vals<-function(obj,param)
{ out.mat<-NULL
for(f in 1:obj$call1$m)
out.mat<-rbind(out.mat,summary.lm(obj$analyses[[f]])$coef[,param])
out.mat  }

### CREATE MATRICES FOR COEFs AND STDs
coef.all<-matrix(0,5,4)
dimnames(coef.all) <- list( c("[1,]","[2,]","[3,]","[4,]","[5,]") ,
c("Intercept","Sex","Age","Averp"))
std.all<-matrix(0,5,4)
dimnames(std.all) <- list( c("[1,]","[2,]","[3,]","[4,]","[5,]") ,
c("Intercept","Sex","Age","Averp"))

## CREATE VECTORS OF ALL 5 ESTIMATES 4@ COVARIATE
coef.all[,1]<-coef.int<-c(g1.302$coefficients[[1]],g2.302$coefficients[[1]],
g3.302$coefficients[[1]],g4.302$coefficients[[1]],g5.302$coefficients[[1]])
coef.all[,2]<-coef.sex<-c(g1.302$coefficients[[2]],g2.302$coefficients[[2]],
g3.302$coefficients[[2]],g4.302$coefficients[[2]],g5.302$coefficients[[2]])
coef.all[,3]<-coef.age<-c(g1.302$coefficients[[3]],g2.302$coefficients[[3]],
g3.302$coefficients[[3]],g4.302$coefficients[[3]],g5.302$coefficients[[3]])
coef.all[,4]<-coef.ave<-c(g1.302$coefficients[[4]],g2.302$coefficients[[4]],
g3.302$coefficients[[4]],g4.302$coefficients[[4]],g5.302$coefficients[[4]])

## CREATE VECTORS OF ALL 5 STD ERRORS FOR THE ESTIMATES
FOR EACH COVARIATE
std.all[,1]<-std.int<-
c(summary(g1.302)$coef[,2][[1]],summary(g2.302)$coef[,2][[1]],
summary(g3.302)$coef[,2][[1]],summary(g4.302)$coef[,2][[1]],summary(g5.30
2)$coef[,2][[1]])
std.all[,2]<-std.int<-c(summary(g1.302)$coef[,2][[2]],summary(g2.302)
$coef[,2][[2]],summary(g3.302)$coef[,2][[2]],summary(g4.302)
$coef[,2][[2]],summary(g5.302)$coef[,2][[2]])
std.all[,3]<-std.int<-c(summary(g1.302)$coef[,2][[3]],summary(g2.302)$
coef[,2][[3]],summary(g3.302)$coef[,2][[3]],summary(g4.302)$
coef[,2][[3]],summary(g5.302)$coef[,2][[3]])
std.all[,4]<-std.int<-c(summary(g1.302)$coef[,2][[4]],summary(g2.302)
$coef[,2][[4]],summary(g3.302)$coef[,2][[4]],summary(g4.302)
$coef[,2][[4]],summary(g5.302)$coef[,2][[4]])

### FUNCTION TO GET THE THREE REQUIRED VECTORS
impute.coef.vec<-apply(coef.all,2,mean)
between.var<-apply(coef.all,2,var)
within.var<-apply(std.all^2,2,mean)
### # COMPUTE THE STANDARD ERROR VECTOR
m <- 5
impute.se.vec <- sqrt(within.var + ((m+1)/m)*between.var)

# ADJUSTing DEGREES OF FREEDOM FOR THE T-STATISTIC .
# SEE LITTLE AND RUBIN (1987), PAGE 257
impute.df <- (m-1)*(1 + (1/(m+1)) * within.var/between.var)^2

# TO OBTAIN REGRESSION TABLE:
multG.302 <- round( cbind(
impute.coef.vec,impute.se.vec,impute.coef.vec/impute.se.vec,
1-pt(abs(impute.coef.vec/impute.se.vec),impute.df) ),4)
dimnames(multG.302) <- list( c("(Intercept)","Sex","Age","Averp"),
c("Estimate","Std. Error","t value","Pvalue") )
multG.302
```

```
### To get fitted values for MI for GAM
ftd1<-cbind(g1.302$fitted.values);ftd2<-cbind(g2.302$fitted.values)
ftd3<-cbind(g3.302$fitted.values);ftd4<-cbind(g4.302$fitted.values)
ftd5<-cbind(g5.302$fitted.values)
ftd.multG302<-cbind(rowSums(cbind(ftd1)+cbind(ftd2)+cbind(ftd3)+
cbind(ftd4)+cbind(ftd5))/5)
ASEMG302II<-sum((cbind(ftd.multG302)-cbind(fitd.od))^2)

######################################################
                    ### Summary for all models ###
################ for cc302 ################
cc302.coef<-fit.cc302$coefficients
cc302.std<-summary(fit.cc302)$coef[, 2]
CI.cc3021<-c(cc302.coef[1]-
1.96*cc302.std[1],cc302.coef[1]+1.96*cc302.std[1])
CI.cc3022<-c(cc302.coef[2]-
1.96*cc302.std[2],cc302.coef[2]+1.96*cc302.std[2])
CI.cc3023<-c(cc302.coef[3]-
1.96*cc302.std[3],cc302.coef[3]+1.96*cc302.std[3])
CI.cc3024<-c(cc302.coef[4]-
1.96*cc302.std[4],cc302.coef[4]+1.96*cc302.std[4])
LCI.cc3021<-((cc302.coef[1]+1.96*cc302.std[1])-(cc302.coef[1]-
1.96*cc302.std[1]))
LCI.cc3022<-((cc302.coef[2]+1.96*cc302.std[2])-(cc302.coef[2]-
1.96*cc302.std[2]))
LCI.cc3023<-((cc302.coef[3]+1.96*cc302.std[3])-(cc302.coef[3]-
1.96*cc302.std[3]))
LCI.cc3024<-((cc302.coef[4]+1.96*cc302.std[4])-(cc302.coef[4]-
1.96*cc302.std[4]))
LCI.cc302<-c(LCI.cc3021,LCI.cc3022,LCI.cc3023,LCI.cc3024)
## Jibu
jibu.cc302<-matrix(0,4,5)
col<-c("Estimate", "SE", "LL","UL","LCI")
rows<-c("Intercept", "Sex", "Age","AVERP")
dimnames(jibu.cc302)<-list(rows,col)
ll1<-c(cc302.coef[1],cc302.std[1],CI.cc3021,LCI.cc302[1])
jibu.cc302[1,]<-ll1
ll2<-c(cc302.coef[2],cc302.std[2],CI.cc3022,LCI.cc302[2])
jibu.cc302[2,]<-ll2
ll3<-c(cc302.coef[3],cc302.std[3],CI.cc3023,LCI.cc302[3])
jibu.cc302[3,]<-ll3
ll4<-c(cc302.coef[4],cc302.std[4],CI.cc3024,LCI.cc302[4])
jibu.cc302[4,]<-ll4

################ for sm302 ################
sm302.coef<-fit.sm302$coefficients
sm302.std<-summary(fit.sm302)$coef[, 2]
CI.sm3021<-c(sm302.coef[1]-
1.96*sm302.std[1],sm302.coef[1]+1.96*sm302.std[1])
CI.sm3022<-c(sm302.coef[2]-
1.96*sm302.std[2],sm302.coef[2]+1.96*sm302.std[2])
CI.sm3023<-c(sm302.coef[3]-
1.96*sm302.std[3],sm302.coef[3]+1.96*sm302.std[3])
CI.sm3024<-c(sm302.coef[4]-
1.96*sm302.std[4],sm302.coef[4]+1.96*sm302.std[4])
LCI.sm3021<-((sm302.coef[1]+1.96*sm302.std[1])-(sm302.coef[1]-
1.96*sm302.std[1]))
LCI.sm3022<-((sm302.coef[2]+1.96*sm302.std[2])-(sm302.coef[2]-
1.96*sm302.std[2]))
LCI.sm3023<-((sm302.coef[3]+1.96*sm302.std[3])-(sm302.coef[3]-
1.96*sm302.std[3]))
LCI.sm3024<-((sm302.coef[4]+1.96*sm302.std[4])-(sm302.coef[4]-
1.96*sm302.std[4]))
LCI.sm302<-c(LCI.sm3021,LCI.sm3022,LCI.sm3023,LCI.sm3024)

## Jibu
jibu.sm302<-matrix(0,4,5)
col<-c("Estimate", "SE", "LL","UL","LCI")
rows<-c("Intercept", "Sex", "Age","AVERP")
dimnames(jibu.sm302)<-list(rows,col)
ll1<-c(sm302.coef[1],sm302.std[1],CI.sm3021,LCI.sm302[1])
jibu.sm302[1,]<-ll1
ll2<-c(sm302.coef[2],sm302.std[2],CI.sm3022,LCI.sm302[2])
jibu.sm302[2,]<-ll2
ll3<-c(sm302.coef[3],sm302.std[3],CI.sm3023,LCI.sm302[3])
jibu.sm302[3,]<-ll3
ll4<-c(sm302.coef[4],sm302.std[4],CI.sm3024,LCI.sm302[4])
jibu.sm302[4,]<-ll4

################ for cm302 ################
cm302.coef<-fit.cm302$coefficients
cm302.std<-summary(fit.cm302)$coef[, 2]
CI.cm3021<-c(cm302.coef[1]-
1.96*cm302.std[1],cm302.coef[1]+1.96*cm302.std[1])
CI.cm3022<-c(cm302.coef[2]-
1.96*cm302.std[2],cm302.coef[2]+1.96*cm302.std[2])
CI.cm3023<-c(cm302.coef[3]-
1.96*cm302.std[3],cm302.coef[3]+1.96*cm302.std[3])
CI.cm3024<-c(cm302.coef[4]-
1.96*cm302.std[4],cm302.coef[4]+1.96*cm302.std[4])
LCI.cm3021<-((cm302.coef[1]+1.96*cm302.std[1])-(cm302.coef[1]-
1.96*cm302.std[1]))
LCI.cm3022<-((cm302.coef[2]+1.96*cm302.std[2])-(cm302.coef[2]-
1.96*cm302.std[2]))
LCI.cm3023<-((cm302.coef[3]+1.96*cm302.std[3])-(cm302.coef[3]-
1.96*cm302.std[3]))
LCI.cm3024<-((cm302.coef[4]+1.96*cm302.std[4])-(cm302.coef[4]-
1.96*cm302.std[4]))
LCI.cm302<-c(LCI.cm3021,LCI.cm3022,LCI.cm3023,LCI.cm3024)

## Jibu
jibu.cm302<-matrix(0,4,5)
col<-c("Estimate", "SE", "LL","UL","LCI")
rows<-c("Intercept", "Sex", "Age","AVERP")
dimnames(jibu.cm302)<-list(rows,col)
ll1<-c(cm302.coef[1],cm302.std[1],CI.cm3021,LCI.cm302[1])
jibu.cm302[1,]<-ll1
ll2<-c(cm302.coef[2],cm302.std[2],CI.cm3022,LCI.cm302[2])
jibu.cm302[2,]<-ll2
ll3<-c(cm302.coef[3],cm302.std[3],CI.cm3023,LCI.cm302[3])
jibu.cm302[3,]<-ll3
ll4<-c(cm302.coef[4],cm302.std[4],CI.cm3024,LCI.cm302[4])
jibu.cm302[4,]<-ll4

################ for multI302 ################
multI302.coef<- MIfitsI $coefficients
multI302.std<-summary(MIfitsI)$coef[, 2]
CI.multI3021<-c(multI302.coef[1]-
1.96*multI302.std[1],multI302.coef[1]+1.96*multI302.std[1])
CI.multI3022<-c(multI302.coef[2]-
1.96*multI302.std[2],multI302.coef[2]+1.96*multI302.std[2])
CI.multI3023<-c(multI302.coef[3]-
1.96*multI302.std[3],multI302.coef[3]+1.96*multI302.std[3])
CI.multI3024<-c(multI302.coef[4]-
1.96*multI302.std[4],multI302.coef[4]+1.96*multI302.std[4])
LCI.multI3021<-((multI302.coef[1]+1.96*multI302.std[1])-(multI302.coef[1]-
1.96*multI302.std[1]))
LCI.multI3022<-((multI302.coef[2]+1.96*multI302.std[2])-(multI302.coef[2]-
1.96*multI302.std[2]))
LCI.multI3023<-((multI302.coef[3]+1.96*multI302.std[3])-(multI302.coef[3]-
1.96*multI302.std[3]))
LCI.multI3024<-((multI302.coef[4]+1.96*multI302.std[4])-(multI302.coef[4]-
1.96*multI302.std[4]))
LCI.multI302<-c(LCI.multI3021,LCI.multI3022,LCI.multI3023,LCI.multI3024)

## Jibu
jibu.multI302<-matrix(0,4,5)
col<-c("Estimate", "SE", "LL","UL","LCI")
rows<-c("Intercept", "Sex", "Age","AVERP")
dimnames(jibu.multI302)<-list(rows,col)
ll1<-c(multI302.coef[1],multI302.std[1],CI.multI3021,LCI.multI302[1])
jibu.multI302[1,]<-ll1
ll2<-c(multI302.coef[2],multI302.std[2],CI.multI3022,LCI.multI302[2])
jibu.multI302[2,]<-ll2
ll3<-c(multI302.coef[3],multI302.std[3],CI.multI3023,LCI.multI302[3])
jibu.multI302[3,]<-ll3
ll4<-c(multI302.coef[4],multI302.std[4],CI.multI3024,LCI.multI302[4])
jibu.multI302[4,]<-ll4

################ for multII302 ################
multII302.coef<- MIfitsII$coefficients
multII302.std<-summary(MIfitsI)$coef[, 2]
summ.multII302
CI.multII3021<-c(summ.multII302[1,3],summ.multII302[1,4])
CI.multII3022<-c(summ.multII302[2,3],summ.multII302[2,4])
CI.multII3023<-c(summ.multII302[3,3],summ.multII302[3,4])
CI.multII3024<-c(summ.multII302[4,3],summ.multII302[4,4])
LCI.multII3021<-(summ.multII302[1,4])-(summ.multII302[1,3])
LCI.multII3022<-(summ.multII302[2,4])-(summ.multII302[2,3])
```

```
LCI.multII3023<-(summ.multII302[3,4])-(summ.multII302[3,3])
LCI.multII3024<-(summ.multII302[4,4])-(summ.multII302[4,3])
LCI.multII302<-c(LCI.multI3021,LCI.multI3022,LCI.multI3023,LCI.multI3024)

## Jibu
jibu.multII302<-matrix(0,4,5)
col<-c("Estimate", "SE", "LL","UL","LCI")
rows<-c("Intercept", "Sex", "Age","AVERP")
dimnames(jibu.multII302)<-list(rows,col)
ml1<-c(summ.multII302[1,1],summ.multII302[1,2],
CI.multII3021,LCI.multII3021)
jibu.multII302[1,]<-ml1
ml2<-c(summ.multII302[2,1],summ.multII302[2,2],
CI.multII3022,LCI.multII3022)
jibu.multII302[2,]<-ml2
ml3<-c(summ.multII302[3,1],summ.multII302[3,2],
CI.multII3023,LCI.multII3023)
jibu.multII302[3,]<-ml3
ml4<-c(summ.multII302[4,1],summ.multII302[4,2],
CI.multII3024,LCI.multII3024)
jibu.multII302[4,]<-ml4

################ for GAM302  ################
smth302.coef<-fits.smth302$coefficients
smth302.std<-summary(fits.smth302)$coef[, 2]
CI.smth3021<-c(smth302.coef[1]-
1.96*smth302.std[1],smth302.coef[1]+1.96*smth302.std[1])
CI.smth3022<-c(smth302.coef[2]-
1.96*smth302.std[2],smth302.coef[2]+1.96*smth302.std[2])
CI.smth3023<-c(smth302.coef[3]-
1.96*smth302.std[3],smth302.coef[3]+1.96*smth302.std[3])
CI.smth3024<-c(smth302.coef[4]-
1.96*smth302.std[4],smth302.coef[4]+1.96*smth302.std[4])
LCI.smth3021<-((smth302.coef[1]+1.96*smth302.std[1])-(smth302.coef[1]-
1.96*smth302.std[1]))
LCI.smth3022<-((smth302.coef[2]+1.96*smth302.std[2])-(smth302.coef[2]-
1.96*smth302.std[2]))
LCI.smth3023<-((smth302.coef[3]+1.96*smth302.std[3])-(smth302.coef[3]-
1.96*smth302.std[3]))
LCI.smth3024<-((smth302.coef[4]+1.96*smth302.std[4])-(smth302.coef[4]-
1.96*smth302.std[4]))
LCI.smth302<-c(LCI.smth3021,LCI.smth3022,LCI.smth3023,LCI.smth3024)

## Jibu
jibu.smth302<-matrix(0,4,5)
col<-c("Estimate", "SE", "LL","UL","LCI")
rows<-c("Intercept", "Sex", "Age","AVERP")
dimnames(jibu.smth302)<-list(rows,col)
ll1<-c(smth302.coef[1],smth302.std[1],CI.smth3021,LCI.smth302[1])
jibu.smth302[1,]<-ll1
ll2<-c(smth302.coef[2],smth302.std[2],CI.smth3022,LCI.smth302[2])
jibu.smth302[2,]<-ll2
ll3<-c(smth302.coef[3],smth302.std[3],CI.smth3023,LCI.smth302[3])
jibu.smth302[3,]<-ll3
ll4<-c(smth302.coef[4],smth302.std[4],CI.smth3024,LCI.smth302[4])
jibu.smth302[4,]<-ll4

################ for GAM302  ################
CI.multG3021<-c(multG.302[1,1]-
1.96*multG.302[1,2],multG.302[1,1]+1.96*multG.302[1,2])
CI.multG3022<-c(multG.302[2,1]-
1.96*multG.302[2,2],multG.302[2,1]+1.96*multG.302[2,2])
CI.multG3023<-c(multG.302[3,1]-
1.96*multG.302[3,2],multG.302[3,1]+1.96*multG.302[3,2])
CI.multG3024<-c(multG.302[4,1]-
1.96*multG.302[4,2],multG.302[4,1]+1.96*multG.302[4,2])

LCI.multG3021<-(multG.302[1,1]+1.96*multG.302[1,2])-(multG.302[1,1]-
1.96*multG.302[1,2])
LCI.multG3022<-(multG.302[2,1]+1.96*multG.302[2,2])-(multG.302[2,1]-
1.96*multG.302[2,2])
LCI.multG3023<-(multG.302[3,1]+1.96*multG.302[3,2])-(multG.302[3,1]-
1.96*multG.302[3,2])
LCI.multG3024<-(multG.302[4,1]+1.96*multG.302[4,2])-(multG.302[4,1]-
1.96*multG.302[4,2])
LCI.multG302<-
c(LCI.multG3021,LCI.multG3022,LCI.multG3023,LCI.multG3024)

## Jibu
jibu.multG302<-matrix(0,4,5)
col<-c("Estimate", "SE", "LL","UL","LCI")
rows<-c("Intercept", "Sex", "Age","AVERP")
dimnames(jibu.multG302)<-list(rows,col)
lj1<-c(multG.302[1,1],multG.302[1,2],CI.multG3021,LCI.multG302[1])
jibu.multG302[1,]<-lj1
lj2<-c(multG.302[2,1],multG.302[2,2],CI.multG3022,LCI.multG302[2])
jibu.multG302[2,]<-lj2
lj3<-c(multG.302[3,1],multG.302[3,2],CI.multG3023,LCI.multG302[3])
jibu.multG302[3,]<-lj3
lj4<-c(multG.302[4,1],multG.302[4,2],CI.multG3024,LCI.multG302[4])
jibu.multG302[4,]<-lj4
################################################

### FINAL MAJIBU
jibu.cc302; jibu.sm302; jibu.cm302; jibu.multI302
jibu.multII302; jibu.smth302; jibu.multG302

#### SUMMARY FOR THE ASE VALUES
ASE.302<-
c(ASE4302cc,ASE302sm,ASE302cm,ASEM302I,ASEM302II,ASEG302,ASE
MG302II)

### PLOTS OF ASE AND MASE ### THIS is just one of the used codes,
similar codes for all analysis and for MASE also
pp<-c(0,0.2918429,0.4995192); asecc<-c(0,5856.839,9487.931)
asesmi<-c(0,1482.966,2426.333); asecmi<-c(0,12.71369,9219.44)
asemi1<-c(0,4875.289,9428.789); asemi2<-c(0,2509.473,4766.304)
asegam1<-c(0,2764.912,9116.295); asegam2<-c(0,1841.889,2461.796)
plot(pp,asecc, type="b", lty=1, xlab="Missingness Proportion",
cex.main=1.2, font.main=2,main="MASE values under
MCAR",ylab="MASE",ylim=range(0,10000),
xlim=range(0,0.6),lwd=1.5)
lines(pp, asesmi, type="b",col="purple", lty=3,lwd=2)
lines(pp, asecmi, type="b", col="blue",lty=5,lwd=1.5)
lines(pp, asemi1, type="b", col="yellow",lty=7,lwd=2.1)
lines(pp, asemi2, type="b", col="green",lty=9,lwd=2.1)
lines(pp, asegam1, type="b", col="red", lty=11,lwd=2)
lines(pp, asegam2, type="b", col="magenta", lty=13,lwd=2)
lgg<-c("CC","SM","CM","PMM-I","PMM-II","GAM-I","GAM-II")
legend(locator(1),legend=lgg ,lty=1:13, ncol=2, adj = c(0, 0.5),
col = c("black","purple","blue","yellow","green","red","magenta"),lwd=2)
#### END END END ###
```

Simulation code

```
### Creating arrays for saving simulated data
nsample <-200; nmeasures <-14; nparam <-4
ccase<-simean<-sicmean<-PMM1<-PMM2<-GAM1<-GAM2<-array(data=NA,
dim=c(nsample,nmeasures,nparam),
dimnames=list(paste(1:nsample),c("Est","Std","LL","UL","LCI","Sigma","ASE0
","ASE1","ASE2","ASE3",
"ASE4","ASE5","ASE6","Miss"),c("Int","SEX","Age","Averp")))

nmeasures2 <-1; nparam2 <-4
minf<-array(data=NA, dim=c(nsample,nmeasures2,nparam2),
dimnames=list(paste(1:nsample),"Minf",c("Int","SEX","Age","Averp")))

##******* 30% MISSINGNESS : MCAR
for (i in 1:nsample){
set.seed(i)
psi1<-(0.89)
p1<-expit(psi1); R1<-rbinom(n,1,1-p1); Pmiss<-sum(R1==1)/n
```

same missingness models as ones used under single analysis with single missingness model were used so here only MCAR for 30% is presented

```
# # creating dataset and generate missingness based on "R" FOR Y
y.miss30a<-rep(NA,n)
for  (jj in 1:n) {
if (R1[jj]==0) y.miss30a[jj]<-y.star[,jj] else y.miss30a[jj]<-NA }
#y.miss30a[1:15]
y.miss301<-y.miss30a

# creating dataset and generate missingness based on "R" FOR AGE
# ag.miss30a<-rep(NA,n)
# for  (jj in 1:n) {
# i f (R1[jj]==0) ag.miss30a[jj]<-LEEFT[jj] else ag.miss30a[jj]<-NA }
# ag.miss30a[1:15]
# ag.miss301<-ag.miss30a

#### fit1: MODEL FOR THE COMPLETE CASES cc
fit.cc301<-lm(y.miss301~SEXE+LEEFT+AVERP)
```

```
# fit.cc301<-lm(y~SEXE+ag.miss301+AVERP) ### FOR MISSING IN AGE
summ.cc301<-summary(fit.cc301)
fitd.cc301<-cbind(fit.cc301$fitted.values)
ccase[i,"Est","Int"]<-summ.cc301$coef[1,1]
ccase[i,"Std","Int"]<-summ.cc301$coef[1,2]
ccase[i,"LL","Int"]<-ccase[i,"Est","Int"]-1.96*ccase[i,"Std","Int"]
ccase[i,"UL","Int"]<-ccase[i,"Est","Int"]+1.96*ccase[i,"Std","Int"]
ccase[i,"LCI","Int"]<-ccase[i,"UL","Int"]-ccase[i,"LL","Int"]
ccase[i,"Sigma","Int"]<-summ.cc301$sigma
ccase[i,"Miss",]<-Pmiss
ccase[i,"ASE0",]<-sum((cbind(fitd.od[!is.na(y.miss301)])-
cbind(fit.cc301$fitted.values))^2)
#ccase[i,"ASE0",]<-sum((cbind(fitd.od[!is.na(ag.miss301)])-
cbind(fit.cc301$fitted.values))^2) ### MISS IN AGE

ccase[i,"Est","SEX"]<-summ.cc301$coef[2,1]
ccase[i,"Std","SEX"]<-summ.cc301$coef[2,2]
ccase[i,"LL","SEX"]<-ccase[i,"Est","SEX"]-1.96*ccase[i,"Std","SEX"]
ccase[i,"UL","SEX"]<-ccase[i,"Est","SEX"]+1.96*ccase[i,"Std","SEX"]
ccase[i,"LCI","SEX"]<-ccase[i,"UL","SEX"]-ccase[i,"LL","SEX"]
ccase[i,"Sigma","SEX"]<-summ.cc301$sigma

ccase[i,"Est","Age"]<-summ.cc301$coef[3,1]
ccase[i,"Std","Age"]<-summ.cc301$coef[3,2]
ccase[i,"LL","Age"]<-ccase[i,"Est","Age"]-1.96*ccase[i,"Std","Age"]
ccase[i,"UL","Age"]<-ccase[i,"Est","Age"]+1.96*ccase[i,"Std","Age"]
ccase[i,"LCI","Age"]<-ccase[i,"UL","Age"]-ccase[i,"LL","Age"]
ccase[i,"Sigma","Age"]<-summ.cc301$sigma

ccase[i,"Est","Averp"]<-summ.cc301$coef[4,1]
ccase[i,"Std","Averp"]<-summ.cc301$coef[4,2]
ccase[i,"LL","Averp"]<-ccase[i,"Est","Averp"]-1.96*ccase[i,"Std","Averp"]
ccase[i,"UL","Averp"]<-ccase[i,"Est","Averp"]+1.96*ccase[i,"Std","Averp"]
ccase[i,"LCI","Averp"]<-ccase[i,"UL","Averp"]-ccase[i,"LL","Averp"]
ccase[i,"Sigma","Averp"]<-summ.cc301$sigma

##### SINGLE MEAN IMPUTATION
y.miss3011<-y.miss30a
## replace NA with mean of the available ones ##
rep.na<-function(y.miss3011, my.mean=TRUE) {
if (my.mean) {value<-mean(y.miss3011[!is.na(y.miss3011)])}
for (i in (1:length(y.miss3011))){if (is.na(y.miss3011[i])==TRUE)
{y.miss3011[i]<-value}}
y.miss3011<<-y.miss3011   }
(y.miss3011)
y.miss301.imp<-(rep.na(y.miss3011))

## FOR AGE
ag.miss3011<-ag.miss30a
## replace NA with mean of the available ones ##
#rep.na<-function(ag.miss3011, my.mean=TRUE) {
#if (my.mean) {value<-mean(ag.miss3011[!is.na(ag.miss3011)])}
#for (i in (1:length(ag.miss3011))){if (is.na(ag.miss3011[i])==TRUE)
{ag.miss3011[i]<-value}}
#ag.miss3011<<-ag.miss3011   }
# (ag.miss3011)
# ag.miss301.imp<-(rep.na(ag.miss3011))

##### fit2: MODEL USING SINGLE MEAN IMPUTED DATA - 301
fit.sm301<-lm(y.miss301.imp~SEXE+LEEFT+AVERP)
# fit.sm301<-lm(y~SEXE+ag.miss301.imp+AVERP)
summ.sm301<-summary(fit.sm301)
fitd.sm301<-cbind(fit.sm301$fitted.values)

simean[i,"Est","Int"]<-summ.sm301$coef[1,1]
simean[i,"Std","Int"]<-summ.sm301$coef[1,2]
simean[i,"LL","Int"]<-simean[i,"Est","Int"]-1.96*simean[i,"Std","Int"]
simean[i,"UL","Int"]<-simean[i,"Est","Int"]+1.96*simean[i,"Std","Int"]
simean[i,"LCI","Int"]<-simean[i,"UL","Int"]-simean[i,"LL","Int"]
simean[i,"Sigma","Int"]<-summ.sm301$sigma
simean[i,"Miss",]<-Pmiss
simean[i,"ASE1",]<-sum((cbind(fitd.sm301)-cbind(fitd.od))^2)
```

same for SEX, Age, Averp ## just change position in the matrix

```
##### CONDITIONAL MEANS IMPUTATION
y.miss30111<-y.miss30a
fit.cm301o<-lm(y.miss30111~SEXE+LEEFT+AVERP+Age2+Age3)
beta.CM301<-summary(fit.cm301o)$coefficients

### replacing using fitted values
DD301<-
beta.CM301[1,1]+(beta.CM301[2,1]*SEXE)+(beta.CM301[3,1]*LEEFT)+(beta.
CM301[4,1]*AVERP)+(beta.CM301[5,1]*Age2)+(beta.CM301[6,1]*Age3)
y.miss30111.imp<- ifelse((is.na(y.miss30111)),DD301,y.miss30111)

### fit3: MODEL WITH CONDITIONAL IMPUTED VALUES
fit.cm301<-lm(y.miss30111.imp~SEXE+LEEFT+AVERP)
summ.cm301<-summary(fit.cm301)
fitd.cm301<-cbind(fit.cm301$fitted.values)

## FOR AGE
#ag.miss30111<-ag.miss30a
#fit.cm301o<-lm(ag.miss30111~SEXE+y+AVERP+Age2+Age3)
#beta.CM301<-summary(fit.cm301o)$coefficients
#rrr<-fit.cm301o$fitted.values

### replacing using fitted values
#DD301<-
beta.CM301[1,1]+(beta.CM301[2,1]*SEXE)+(beta.CM301[3,1]*y)+(beta.CM30
1[4,1]*AVERP)+(beta.CM301[5,1]*Age2)+(beta.CM301[6,1]*Age3)
#ag.miss30111.imp<- ifelse((is.na(ag.miss30111)),DD301,ag.miss30111)

### fit3: MODEL WITH CONDITIONAL IMPUTED VALUES
#fit.cm301<-lm(y~SEXE+ag.miss30111.imp+AVERP)
#summ.cm301<-summary(fit.cm301)
#fitd.cm301<-cbind(fit.cm301$fitted.values)

sicmean[i,"Est","Int"]<-summ.cm301$coef[1,1]
sicmean[i,"Std","Int"]<-summ.cm301$coef[1,2]
sicmean[i,"LL","Int"]<-sicmean[i,"Est","Int"]-1.96*sicmean[i,"Std","Int"]
sicmean[i,"UL","Int"]<-sicmean[i,"Est","Int"]+1.96*sicmean[i,"Std","Int"]
sicmean[i,"LCI","Int"]<-sicmean[i,"UL","Int"]-sicmean[i,"LL","Int"]
sicmean[i,"Sigma","Int"]<-summ.cm301$sigma
sicmean[i,"Miss",]<-Pmiss
sicmean[i,"ASE2",]<-sum((cbind(fitd.cm301)-cbind(fitd.od))^2)
```

same for SEX, Age, Averp ## just change position in the matrix

```
### MULTIPLE IMPUTATION - 301
############## PartI-- Single imputation using PMM
dataCC301<-data.frame(y.miss301,SEXE,LEEFT,AVERP)
imp.CC301I<-mice(dataCC301,m=1,maxit=10, seed = 333)
imp.CC301I<-complete(imp.CC301I)
#complete(imp.CC301I)[1:10,1:4] # show some of completed data
MIfitsI<-lm(y.miss301 ~ SEXE+LEEFT+AVERP, imp.CC301I)
summ.multI301<-summary(MIfitsI)
fitd.multI301<-cbind(MIfitsI$fitted.values)

## FOR AGE
#dataCC301<-data.frame(y,SEXE,ag.miss301,AVERP)
#imp.CC301I<-mice(dataCC301,m=1,maxit=10, seed = 333)
#imp.CC301I<-complete(imp.CC301I)
#complete(imp.CC301I)[1:10,1:4] # show some of completed data
#MIfitsI<-lm(y~ SEXE+ag.miss301 +AVERP, imp.CC301I)
#summ.multI301<-summary(MIfitsI)
#fitd.multI301<-cbind(MIfitsI$fitted.values)

PMM1[i,"Est","Int"]<-summ.multI301$coef[1,1]
PMM1[i,"Std","Int"]<-summ.multI301$coef[1,2]
PMM1[i,"LL","Int"]<-PMM1[i,"Est","Int"]-1.96*PMM1[i,"Std","Int"]
PMM1[i,"UL","Int"]<-PMM1[i,"Est","Int"]+1.96*PMM1[i,"Std","Int"]
PMM1[i,"LCI","Int"]<-PMM1[i,"UL","Int"]-PMM1[i,"LL","Int"]
PMM1[i,"Sigma","Int"]<-summ.multI301$sigma
PMM1[i,"Miss",]<-Pmiss
PMM1[i,"ASE3",]<-sum((cbind(fitd.multI301)-cbind(fitd.od))^2)
```

same for SEX, Age, Averp ## just change position in the matrix

```
########## PartII-- Multiple imputation using PMM
dataCC301<-data.frame(y.miss301,SEXE,LEEFT,AVERP)
imp.CC301II<- mice(dataCC301,m=5,maxit=10, seed = 333)
#complete(imp.CC301II)[1:10,1:4] # show some of completed data
MIfitsII<-lm.mids(y.miss301 ~ SEXE+LEEFT+AVERP, imp.CC301II)
summ.multII301<-summary(MIcombine(MIfitsII$analyses))

## FOR AGE
#dataCC301<-data.frame(y,SEXE,ag.miss301,AVERP)
#imp.CC301II<- mice(dataCC301,m=5,maxit=10, seed = 333)
#complete(imp.CC301II)[1:10,1:4] # show some of completed data
```

```
#MIfitsII<-lm.mids(y ~ SEXE+ag.miss301 +AVERP, imp.CC301II)
#summ.multII301<-summary(MIcombine(MIfitsII$analyses))

### To get fitted values for Multiple imputation from 5 models
fittd1<-cbind(MIfitsII$analyses[[1]]$fitted.values)
fittd2<-cbind(MIfitsII$analyses[[2]]$fitted.values)
fittd3<-cbind(MIfitsII$analyses[[3]]$fitted.values)
fittd4<-cbind(MIfitsII$analyses[[4]]$fitted.values)
fittd5<-cbind(MIfitsII$analyses[[5]]$fitted.values)
fitd.multII301<-
cbind(rowSums(cbind(fittd1)+cbind(fittd2)+cbind(fittd3)+cbind(fittd4)+cbind(fitt
d5))/5)

### To get data to calculate MASE1
dat1<-complete(imp.CC301II,1)[1]; dat2<-complete(imp.CC301II,2)[1]
dat3<-complete(imp.CC301II,3)[1]; dat4<-complete(imp.CC301II,4)[1]
dat5<-complete(imp.CC301II,5)[1]
ag.mult<-cbind(rowSums(cbind(dat1)+cbind(dat2)+cbind(dat3)+
cbind(dat4)+cbind(dat5))/5)


PMM2[i,"Est","Int"]<-summ.multII301[1,1]
PMM2[i,"Std","Int"]<-summ.multII301[1,2]
PMM2[i,"LL","Int"]<-PMM2[i,"Est","Int"]-1.96*PMM2[i,"Std","Int"]
PMM2[i,"UL","Int"]<-PMM2[i,"Est","Int"]+1.96*PMM2[i,"Std","Int"]
PMM2[i,"LCI","Int"]<-PMM2[i,"UL","Int"]-PMM2[i,"LL","Int"]
minf[i,"Minf","Int"]<-summ.multII301[1,5]
PMM2[i,"Sigma","Int"]<-sd(ag.mult)
PMM2[i,"Miss",]<-Pmiss
PMM2[i,"ASE4",]<-sum((cbind(fitd.multII301)-cbind(fitd.od))^2)

### same for SEX, Age, Averp  ## just change position in the matrix

#### GENERALIZED ADDITIVE MODEL==301
y.miss301.gam<-y.miss301
fit.sp301<-gam(y.miss301.gam~ s(LEEFT)+s(AVERP)+SEXE,
family="gaussian",fit=TRUE)
#summary(fit.sp301)
hh301<-predict.gam(fit.sp301,newdata=data.frame(LEEFT,AVERP,SEXE),
type="response")
y.smth301<-ifelse((is.na(y.miss301)),hh301,y.miss301)
sigmag301<-sd(y.smth301) ## sigmag301
fits.smth301<-lm(y.smth301~ SEXE+LEEFT+AVERP)
summ.smth301<-summary(fits.smth301)
fitd.smth301<-fits.smth301$fitted.values

### FOR AGE
#ag.miss301.gam<-ag.miss301
#fit.sp301<-gam(ag.miss301.gam~ s(y)+s(AVERP)+SEXE,
family="gaussian",fit=TRUE)
#summary(fit.sp301)
#hh301<-predict.gam(fit.sp301,newdata=data.frame(y,AVERP,SEXE),
type="response")
#ag.smth301<-ifelse((is.na(ag.miss301)),hh301,ag.miss301)
#sigmag301<-sd(y) ## sigmag301
#fits.smth301<-lm(y~ SEXE+ag.smth301+AVERP)
#summ.smth301<-summary(fits.smth301)
#fitd.smth301<-fits.smth301$fitted.values

GAM1[i,"Est","Int"]<-summ.smth301$coef[1,1]
GAM1[i,"Std","Int"]<-summ.smth301$coef[1,2]
GAM1[i,"LL","Int"]<-GAM1[i,"Est","Int"]-1.96*GAM1[i,"Std","Int"]
GAM1[i,"UL","Int"]<-GAM1[i,"Est","Int"]+1.96*GAM1[i,"Std","Int"]
GAM1[i,"LCI","Int"]<-GAM1[i,"UL","Int"]-GAM1[i,"LL","Int"]
GAM1[i,"Sigma","Int"]<-summ.smth301$sigma
GAM1[i,"Miss",]<-Pmiss
GAM1[i,"ASE5",]<-sum((cbind(fitd.smth301)-cbind(fitd.od))^2)

### same for SEX, Age, Averp  ## just change position in the matrix

#### MULTIPLE IMPUTATION USING GAM == 301
y.miss301.Mgam<-y.miss301
fit.MG301<-gam(y.miss301.Mgam~ s(LEEFT)+s(AVERP)+SEXE,
family="gaussian",fit=TRUE)
hhMG301<-
predict.gam(fit.MG301,newdata=data.frame(LEEFT,AVERP,SEXE),
type="response")
### Data generation/sampling
YY<-matrix(0,n,5)
col<-c("Y.gam1", "Y.gam2", "Y.gam3","Y.gam4","Y.gam5")
rows<-seq(1:n)
dimnames(YY)<-list(rows,col)
sigmaMG301<-sqrt(fit.MG301$sig2)

set.seed(337)
for (bb in 1:n){ YY[bb,1]=rnorm(1,hhMG301[bb],sigmaMG301) }
set.seed(456)
for (bb in 1:n){ YY[bb,2]=rnorm(1,hhMG301[bb],sigmaMG301) }
set.seed(231)
for (bb in 1:n){ YY[bb,3]=rnorm(1,hhMG301[bb],sigmaMG301) }
set.seed(567)
for (bb in 1:n){ YY[bb,4]=rnorm(1,hhMG301[bb],sigmaMG301) }
set.seed(123)
for (bb in 1:n){ YY[bb,5]=rnorm(1,hhMG301[bb],sigmaMG301) }

Y.gam1<-YY[,1]; Y.gam2<-YY[,2]; Y.gam3<-YY[,3]
Y.gam4<-YY[,4]; Y.gam5<-YY[,5]

Y.gamdat<-
cbind(rowSums(cbind(Y.gam1)+cbind(Y.gam2)+cbind(Y.gam3)+cbind(Y.gam
4)+cbind(Y.gam5))/5)

### fitting 5 GAM using the 5 datasets
g1.301<-lm(Y.gam1~ SEXE+LEEFT+AVERP)
g2.301<-lm(Y.gam2~ SEXE+LEEFT+AVERP)
g3.301<-lm(Y.gam3~ SEXE+LEEFT+AVERP)
g4.301<-lm(Y.gam4~ SEXE+LEEFT+AVERP)
g5.301<-lm(Y.gam5~ SEXE+LEEFT+AVERP)

####################
#### MULTIPLE IMPUTATION BY HAND FOR GAM
lmids.vals<-function(obj,param)
{ out.mat<-NULL
for(f in 1:obj$call1$m)
out.mat<-rbind(out.mat,summary.lm(obj$analyses[[f]])$coef[,param])
out.mat }

### CREATE MATRICES FOR COEFs AND STDs
coef.all<-matrix(0,5,4)
dimnames(coef.all) <- list( c("[1,]","[2,]","[3,]","[4,]","[5,]") ,
c("Intercept","Sex","Age","Averp"))
std.all<-matrix(0,5,4)
dimnames(std.all) <- list( c("[1,]","[2,]","[3,]","[4,]","[5,]") ,
c("Intercept","Sex","Age","Averp"))

## CREATE VECTORS OF ALL 5 ESTIMATES FOR EACH COVARIATE
coef.all[,1]<-coef.int<-c(g1.301$coefficients[[1]],g2.301$coefficients[[1]],
g3.301$coefficients[[1]],g4.301$coefficients[[1]],g5.301$coefficients[[1]])
coef.all[,2]<-coef.sex<-c(g1.301$coefficients[[2]],g2.301$coefficients[[2]],
 g3.301$coefficients[[2]],g4.301$coefficients[[2]],g5.301$coefficients[[2]])
coef.all[,3]<-coef.age<- c(g1.301$coefficients[[3]],g2.301$coefficients[[3]],
g3.301$coefficients[[3]],g4.301$coefficients[[3]],g5.301$coefficients[[3]])
coef.all[,4]<-coef.ave<-c(g1.301$coefficients[[4]],g2.301$coefficients[[4]],
g3.301$coefficients[[4]],g4.301$coefficients[[4]],g5.301$coefficients[[4]])

## CREATE VECTORS OF ALL 5 STD ERRORS FOR THE ESTIMATES
FOR EACH COVARIATE
std.all[,1]<-std.int<-
c(summary(g1.301)$coef[,2][[1]],summary(g2.301)$coef[,2][[1]],
summary(g3.301)$coef[,2][[1]],summary(g4.301)$coef[,2][[1]],summary(g5.30
1)$coef[,2][[1]])
std.all[,2]<-std.int<-c(summary(g1.301)$coef[,2][[2]],summary(g2.301)
$coef[,2][[2]], summary(g3.301)$coef[,2][[2]],summary
(g4.301)$coef[,2][[2]],summary(g5.301)$coef[,2][[2]])
std.all[,3]<-std.int<-c(summary(g1.301)$coef[,2][[3]],summary(g2.301)
$coef[,2][[3]],summary(g3.301)$coef[,2][[3]],summary(g4.301)
$coef[,2][[3]],summary(g5.301)$coef[,2][[3]])
std.all[,4]<-std.int<-c(summary(g1.301)$coef[,2][[4]],summary(g2.301)
$coef[,2][[4]],summary(g3.301)$coef[,2][[4]],summary(g4.301)
$coef[,2][[4]],summary(g5.301)$coef[,2][[4]])
### FUNCTION TO GET THE THREE REQUIRED VECTORS
impute.coef.vec<-apply(coef.all,2,mean)
between.var<-apply(coef.all,2,var)
within.var<-apply(std.all^2,2,mean)

### # COMPUTE THE STANDARD ERROR VECTOR
m <- 5
impute.se.vec <- sqrt(within.var + ((m+1)/m)*between.var)
```

```
# THE DEGREES OF FREEDOM FOR THE T-STATISTIC NEEDS TO BE
ADJUSTED. # SEE LITTLE AND RUBIN (1987), PAGE 257
impute.df <- (m-1)*(1 + (1/(m+1)) * within.var/between.var)^2

# TO OBTAIN REGRESSION TABLE:
multG.301 <- round( cbind(
impute.coef.vec,impute.se.vec,impute.coef.vec/impute.se.vec,
1-pt(abs(impute.coef.vec/impute.se.vec),impute.df) ),4)
dimnames(multG.301) <- list( c("(Intercept)","Sex","Age","Averp"),
c("Estimate","Std. Error","t value","Pvalue") )
summ.multG301<-multG.301

### To get fitted values for MI for GAM
ftd1<-cbind(g1.301$fitted.values);ftd2<-cbind(g2.301$fitted.values)
ftd3<-cbind(g3.301$fitted.values);ftd4<-cbind(g4.301$fitted.values)
ftd5<-cbind(g5.301$fitted.values)
fitd.multG301<-cbind(rowSums(cbind(ftd1)+cbind(ftd2)+cbind(ftd3)+
cbind(ftd4)+cbind(ftd5))/5)
#ASEMG301II<-sum((cbind(fitd.multG301)-cbind(fitd.od))^2)

### FOR AGE
ag.miss301.Mgam<-ag.miss301
fit.MG301<-gam(ag.miss301.Mgam~ s(y)+s(AVERP)+SEXE,
family="gaussian",fit=TRUE)
hhMG301<-predict.gam(fit.MG301,newdata=data.frame(y,AVERP,SEXE),
type="response")
### Data generation/sampling
AG<-matrix(0,n,5)
col<-c("ag.gam1", "ag.gam2", "ag.gam3","ag.gam4","ag.gam5")
rows<-seq(1:n)
dimnames(AG)<-list(rows,col)
sigmaMG301<-sqrt(fit.MG301$sig2)

set.seed(337)
for (bb in 1:n){ AG[bb,1]=rnorm(1,hhMG301[bb],sigmaMG301) }
set.seed(456)
for (bb in 1:n){ AG[bb,2]=rnorm(1,hhMG301[bb],sigmaMG301) }
set.seed(231)
for (bb in 1:n){ AG[bb,3]=rnorm(1,hhMG301[bb],sigmaMG301) }
set.seed(567)
for (bb in 1:n){ AG[bb,4]=rnorm(1,hhMG301[bb],sigmaMG301) }
set.seed(123)
for (bb in 1:n){ AG[bb,5]=rnorm(1,hhMG301[bb],sigmaMG301) }

ag.gam1<-AG[,1]; ag.gam2<-AG[,2]; ag.gam3<-AG[,3]
ag.gam4<-AG[,4]; ag.gam5<-AG[,5]

ag.gamdat<-
cbind(rowSums(cbind(ag.gam1)+cbind(ag.gam2)+cbind(ag.gam3)+cbind(ag.g
am4)+cbind(ag.gam5))/5)

### fitting 5 GAM using the 5 datasets
g1.301<-lm(y~SEXE+ag.gam1+AVERP); g2.301<-
lm(y~SEXE+ag.gam2+AVERP); g3.301<-lm(y~SEXE+ag.gam3+AVERP)
g4.301<-lm(y~SEXE+ag.gam4+AVERP);
g5.301<-lm(y~SEXE+ag.gam5+AVERP)

####################
#### MULTIPLE IMPUTATION FOR GAM
lmids.vals<-function(obj,param)
{  out.mat<-NULL
for(f in 1:obj$call1$m)
out.mat<-rbind(out.mat,summary.lm(obj$analyses[[f]])$coef[,param])
out.mat  }

### CREATE MATRICES FOR COEFs AND STDs
coef.all<-matrix(0,5,4)
dimnames(coef.all) <- list( c("[1,]","[2,]","[3,]","[4,]","[5,]") ,
c("Intercept","Sex","Age","Averp"))
std.all<-matrix(0,5,4)
dimnames(std.all) <- list( c("[1,]","[2,]","[3,]","[4,]","[5,]") ,
c("Intercept","Sex","Age","Averp"))

## CREATE VECTORS OF ALL 5 ESTIMATES FOR EACH COVARIATE
coef.all[,1]<-coef.int<-c(g1.301$coefficients[[1]],g2.301$coefficients[[1]],
g3.301$coefficients[[1]],g4.301$coefficients[[1]],g5.301$coefficients[[1]])
coef.all[,2]<-coef.sex<-c(g1.301$coefficients[[2]],g2.301$coefficients[[2]],
g3.301$coefficients[[2]],g4.301$coefficients[[2]],g5.301$coefficients[[2]])
coef.all[,3]<-coef.age<-c(g1.301$coefficients[[3]],g2.301$coefficients[[3]],
g3.301$coefficients[[3]],g4.301$coefficients[[3]],g5.301$coefficients[[3]])
coef.all[,4]<-coef.ave<-c(g1.301$coefficients[[4]],g2.301$coefficients[[4]],
g3.301$coefficients[[4]],g4.301$coefficients[[4]],g5.301$coefficients[[4]])

## CREATE VECTORS OF ALL 5 STD ERRORS FOR THE ESTIMATES
FOR EACH COVARIATE
std.all[,1]<-std.int<-
c(summary(g1.301)$coef[,2][[1]],summary(g2.301)$coef[,2][[1]],
summary(g3.301)$coef[,2][[1]],summary(g4.301)$coef[,2][[1]],summary(g5.30
1)$coef[,2][[1]])
std.all[,2]<-std.int<-
c(summary(g1.301)$coef[,2][[2]],summary(g2.301)$coef[,2][[2]],
summary(g3.301)$coef[,2][[2]],summary(g4.301)$coef[,2][[2]],summary(g5.30
1)$coef[,2][[2]])
std.all[,3]<-std.int<-
c(summary(g1.301)$coef[,2][[3]],summary(g2.301)$coef[,2][[3]],
summary(g3.301)$coef[,2][[3]],summary(g4.301)$coef[,2][[3]],summary(g5.30
1)$coef[,2][[3]])
std.all[,4]<-std.int<-
c(summary(g1.301)$coef[,2][[4]],summary(g2.301)$coef[,2][[4]],
summary(g3.301)$coef[,2][[4]],summary(g4.301)$coef[,2][[4]],summary(g5.30
1)$coef[,2][[4]])

### FUNCTION TO GET THE THREE REQUIRED VECTORS
impute.coef.vec<-apply(coef.all,2,mean)
between.var<-apply(coef.all,2,var)
within.var<-apply(std.all^2,2,mean)
### # COMPUTE THE STANDARD ERROR VECTOR
m <- 5
impute.se.vec <- sqrt(within.var + ((m+1)/m)*between.var)

# THE DEGREES OF FREEDOM FOR THE T-STATISTIC NEEDS TO BE
ADJUSTED.
# SEE LITTLE AND RUBIN (1987), PAGE 257
impute.df <- (m-1)*(1 + (1/(m+1)) * within.var/between.var)^2

# TO OBTAIN REGRESSION TABLE:
multG.301 <- round( cbind(
impute.coef.vec,impute.se.vec,impute.coef.vec/impute.se.vec,
1-pt(abs(impute.coef.vec/impute.se.vec),impute.df) ),4)
dimnames(multG.301) <- list( c("(Intercept)","Sex","Age","Averp"),
c("Estimate","Std. Error","t value","Pvalue") )
summ.multG301<-multG.301

### To get fitted values for MI for GAM
ftd1<-cbind(g1.301$fitted.values); ftd2<-cbind(g2.301$fitted.values)
ftd3<-cbind(g3.301$fitted.values); ftd4<-cbind(g4.301$fitted.values)
ftd5<-cbind(g5.301$fitted.values)
fitd.multG301<-
cbind(rowSums(cbind(ftd1)+cbind(ftd2)+cbind(ftd3)+cbind(ftd4)+cbind(ftd5))/5
)
#ASEMG301II<-sum((cbind(fitd.multG301)-cbind(fitd.od))^2)
###################

##multG301.coef<-fits.multG301$coefficients
##multG301.std<-summary(fits.multG301)$coef[, 2]
GAM2[i,"Est","Int"]<-summ.multG301[1,1]
GAM2[i,"Std","Int"]<-summ.multG301[1,2]
GAM2[i,"LL","Int"]<-GAM2[i,"Est","Int"]-1.96*GAM2[i,"Std","Int"]
GAM2[i,"UL","Int"]<-GAM2[i,"Est","Int"]+1.96*GAM2[i,"Std","Int"]
GAM2[i,"LCI","Int"]<-GAM2[i,"UL","Int"]-GAM2[i,"LL","Int"]
GAM2[i,"Sigma","Int"]<-sd(y)
GAM2[i,"Miss",]<-Pmiss
GAM2[i,"ASE6",]<-sum((cbind(fitd.multG301)-cbind(fitd.od))^2)

### same for SEX, Age, Averp  ## just change position in the matrix
}

############### SUMMARY RESULTS ###############
# SUMMARY RESULTS FOR COMPLETE CASES
#MAJIBU
### Jibu1.Scc301 for 30a: FOR CC- Average of the 200
Jibu1.Scc301<-matrix(0,4,5)
col<-c("Estimate", "AvSE", "AvLlimit","AvUlimit","AvLengthCI")
rows<-c("Intercept", "SEXE", "LEEFT","AVERP")
dimnames(Jibu1.Scc301)<-list(rows,col)
b1<-
c(mean(as.data.frame(ccase[,,"Int"]))[1],mean(as.data.frame(ccase[,,"Int"]))[2]
,
mean(as.data.frame(ccase[,,"Int"]))[3],mean(as.data.frame(ccase[,,"Int"]))[4],
mean(as.data.frame(ccase[,,"Int"]))[5])
```

```
Jibu1.Scc301[1,]<-b1
b2<-c(mean(as.data.frame(ccase[,,"SEX"]))[1],
mean(as.data.frame(ccase[,,"SEX"]))[2],
mean(as.data.frame(ccase[,,"SEX"]))[3],mean(as.data.frame(ccase[,,"SEX"]))[
4], mean(as.data.frame(ccase[,,"SEX"]))[5])
Jibu1.Scc301[2,]<-b2
b3<-c(mean(as.data.frame(ccase[,,"Age"]))[1],
mean(as.data.frame(ccase[,,"Age"]))[2],
mean(as.data.frame(ccase[,,"Age"]))[3],mean(as.data.frame(ccase[,,"Age"]))[
4], mean(as.data.frame(ccase[,,"Age"]))[5])
Jibu1.Scc301[3,]<-b3
b4<-c(mean(as.data.frame(ccase[,,"Averp"]))[1],
mean(as.data.frame(ccase[,,"Averp"]))[2],
mean(as.data.frame(ccase[,,"Averp"]))[3],mean(as.data.frame(ccase[,,"Averp"
]))[4], mean(as.data.frame(ccase[,,"Averp"]))[5])
Jibu1.Scc301[4,]<-b4

### Jibu2.Scc301 for 30a: FOR CC- Jibu2.Scc301<-matrix(0,1,3)
col<-c("AvSigma","MASE0","AvPMiss")
rows<-c("Value")
dimnames(Jibu2.Scc301)<-list(rows,col)
kk<-c(mean(as.data.frame(ccase[,,"Int"]))[6],
mean(as.data.frame(ccase[,,"Int"]))[7],
mean(as.data.frame(ccase[,,"Int"]))[14])
Jibu2.Scc301[1,]<-kk

# SUMMARY RESULTS FOR SINGLE MEAN IMPUTATION
#MAJIBU
### Jibu1.SSI301 for 30a: FOR SI- Average of the 1000
Jibu1.SSI301<-matrix(0,4,5)
col<-c("Estimate", "AvSE", "AvLlimit","AvUlimit","AvLengthCI")
rows<-c("Intercept", "SEXE", "LEEFT","AVERP")
dimnames(Jibu1.SSI301)<-list(rows,col)
b1<-c(mean(as.data.frame(simean[,,"Int"]))[1],
mean(as.data.frame(simean[,,"Int"]))[2],
mean(as.data.frame(simean[,,"Int"]))[3],
mean(as.data.frame(simean[,,"Int"]))[4],
mean(as.data.frame(simean[,,"Int"]))[5])
Jibu1.SSI301[1,]<-b1
b2<-
c(mean(as.data.frame(simean[,,"SEX"]))[1],mean(as.data.frame(simean[,,"SE
X"]))[2],
mean(as.data.frame(simean[,,"SEX"]))[3],mean(as.data.frame(simean[,,"SEX"
]))[4],
mean(as.data.frame(simean[,,"SEX"]))[5])
Jibu1.SSI301[2,]<-b2
b3<-c(mean(as.data.frame(simean[,,"Age"]))[1]
,mean(as.data.frame(simean[,,"Age"]))[2],
mean(as.data.frame(simean[,,"Age"]))[3],mean(as.data.frame(simean[,,"Age"]
))[4],
mean(as.data.frame(simean[,,"Age"]))[5])
Jibu1.SSI301[3,]<-b3
b4<-c(mean(as.data.frame(simean[,,"Averp"]))[1],
mean(as.data.frame(simean[,,"Averp"]))[2],
mean(as.data.frame(simean[,,"Averp"]))[3],mean(as.data.frame(simean[,,"Ave
rp"]))[4],
mean(as.data.frame(simean[,,"Averp"]))[5])
Jibu1.SSI301[4,]<-b4


### Jibu2.SSI301 for 30a: FOR SI- Average of the 1000- OTHER
STATISTICS
Jibu2.SSI301<-matrix(0,1,3)
col<-c("AvSigma","MASE1","AvPMiss")
rows<-c("Value")
dimnames(Jibu2.SSI301)<-list(rows,col)
kk<-
c(mean(as.data.frame(simean[,,"Int"]))[6],mean(as.data.frame(simean[,,"Int"]))
[8],
mean(as.data.frame(simean[,,"Int"]))[14])
Jibu2.SSI301[1,]<-kk


# SUMMARY RESULTS FOR CONDITIONAL MEAN IMPUTATION
#sicmean[,"ESS","Int"]

#MAJIBU
### Jibu1.SCM301 for 30a: FOR SI- Average of the 1000
Jibu1.SCM301<-matrix(0,4,5)
col<-c("Estimate", "AvSE", "AvLlimit","AvUlimit","AvLengthCI")
rows<-c("Intercept", "SEXE", "LEEFT","AVERP")
dimnames(Jibu1.SCM301)<-list(rows,col)
b1<-c(mean(as.data.frame(sicmean[,,"Int"]))[1],
mean(as.data.frame(sicmean[,,"Int"]))[2],
mean(as.data.frame(sicmean[,,"Int"]))[3],
mean(as.data.frame(sicmean[,,"Int"]))[4],
mean(as.data.frame(sicmean[,,"Int"]))[5])
Jibu1.SCM301[1,]<-b1
b2<-
c(mean(as.data.frame(sicmean[,,"SEX"]))[1],mean(as.data.frame(sicmean[,,"
SEX"]))[2], mean(as.data.frame(sicmean[,,"SEX"]))[3],
mean(as.data.frame(sicmean[,,"SEX"]))[4],
mean(as.data.frame(sicmean[,,"SEX"]))[5])
Jibu1.SCM301[2,]<-b2
b3<-
c(mean(as.data.frame(sicmean[,,"Age"]))[1],mean(as.data.frame(sicmean[,,"A
ge"]))[2], mean(as.data.frame(sicmean[,,"Age"]))[3],
mean(as.data.frame(sicmean[,,"Age"]))[4],
mean(as.data.frame(sicmean[,,"Age"]))[5])
Jibu1.SCM301[3,]<-b3
b4<-
c(mean(as.data.frame(sicmean[,,"Averp"]))[1],mean(as.data.frame(sicmean[,,
"Averp"]))[2], mean(as.data.frame(sicmean[,,"Averp"]))[3],
mean(as.data.frame(sicmean[,,"Averp"]))[4],mean(as.data.frame(sicmean[,,"A
verp"]))[5])
Jibu1.SCM301[4,]<-b4

### Jibu2.SCM301 for 30a: FOR SI- Average of the 1000- OTHER
STATISTICS
Jibu2.SCM301<-matrix(0,1,3)
col<-c("AvSigma","MASE2","AvPMiss")
rows<-c("Value")
dimnames(Jibu2.SCM301)<-list(rows,col)
HH<-c(mean(as.data.frame(sicmean[,,"Int"]))[6],
ean(as.data.frame(sicmean[,,"Int"]))[9],
mean(as.data.frame(sicmean[,,"Int"]))[14])
Jibu2.SCM301[1,]<-HH

# SUMMARY RESULTS FOR MULTIPLE IMPUTATION 1
#MAJIBU
### Jibu1.PMM1301 for 30a: FOR MI- Average of the 1000
Jibu1.PMM1301<-matrix(0,4,5)
col<-c("Estimate", "AvSE", "AvLlimit","AvUlimit","AvLengthCI")
rows<-c("Intercept", "SEXE", "LEEFT","AVERP")
dimnames(Jibu1.PMM1301)<-list(rows,col)
b1<-c(mean(as.data.frame(PMM1[,,"Int"]))[1],
mean(as.data.frame(PMM1[,,"Int"]))[2],
mean(as.data.frame(PMM1[,,"Int"]))[3],
mean(as.data.frame(PMM1[,,"Int"]))[4],
mean(as.data.frame(PMM1[,,"Int"]))[5])
Jibu1.PMM1301[1,]<-b1
b2<-c(mean(as.data.frame(PMM1[,,"SEX"]))[1],
mean(as.data.frame(PMM1[,,"SEX"]))[2],
mean(as.data.frame(PMM1[,,"SEX"]))[3],mean(as.data.frame(PMM1[,,"SEX"])
)[4],
mean(as.data.frame(PMM1[,,"SEX"]))[5])
Jibu1.PMM1301[2,]<-b2
b3<-c(mean(as.data.frame(PMM1[,,"Age"]))[1],
mean(as.data.frame(PMM1[,,"Age"]))[2],
mean(as.data.frame(PMM1[,,"Age"]))[3],mean(as.data.frame(PMM1[,,"Age"]))
[4],
mean(as.data.frame(PMM1[,,"Age"]))[5])
Jibu1.PMM1301[3,]<-b3
b4<-c(mean(as.data.frame(PMM1[,,"Averp"]))[1],
mean(as.data.frame(PMM1[,,"Averp"]))[2],
mean(as.data.frame(PMM1[,,"Averp"]))[3],mean(as.data.frame(PMM1[,,"Aver
p"]))[4],
mean(as.data.frame(PMM1[,,"Averp"]))[5])
Jibu1.PMM1301[4,]<-b4

### Jibu2.PMM1301 for 30a: FOR MI- Average of the 1000- OTHER
STATISTICS
Jibu2.PMM1301<-matrix(0,1,3)
col<-c("AvSigma","MASE3","AvPMiss")
rows<-c("Value")
dimnames(Jibu2.PMM1301)<-list(rows,col)
FF<-
c(mean(as.data.frame(PMM1[,,"Int"]))[6],mean(as.data.frame(PMM1[,,"Int"]))[
10],
mean(as.data.frame(PMM1[,,"Int"]))[14])
```

```
Jibu2.PMM1301[1,]<-FF

# SUMMARY RESULTS FOR MULTIPLE IMPUTATION 2
### Jibu1.PMM2301 for 30a: FOR MI- Average of the 1000
Jibu1.PMM2301<-matrix(0,4,5)
col<-c("Estimate", "AvSE", "AvLlimit","AvUlimit","AvLengthCI")
rows<-c("Intercept", "SEXE", "LEEFT","AVERP")
dimnames(Jibu1.PMM2301)<-list(rows,col)
b1<-c(mean(as.data.frame(PMM2[,,"Int"]))[1],
mean(as.data.frame(PMM2[,,"Int"]))[2],
mean(as.data.frame(PMM2[,,"Int"]))[3],
mean(as.data.frame(PMM2[,,"Int"]))[4],
mean(as.data.frame(PMM2[,,"Int"]))[5])
Jibu1.PMM2301[1,]<-b1
b2<-c(mean(as.data.frame(PMM2[,,"SEX"]))[1],
mean(as.data.frame(PMM2[,,"SEX"]))[2],
mean(as.data.frame(PMM2[,,"SEX"]))[3],mean(as.data.frame(PMM2[,,"SEX"])
)[4],
mean(as.data.frame(PMM2[,,"SEX"]))[5])
Jibu1.PMM2301[2,]<-b2
b3<-c(mean(as.data.frame(PMM2[,,"Age"]))[1],
mean(as.data.frame(PMM2[,,"Age"]))[2],
mean(as.data.frame(PMM2[,,"Age"]))[3],mean(as.data.frame(PMM2[,,"Age"]))
[4],
mean(as.data.frame(PMM2[,,"Age"]))[5])
Jibu1.PMM2301[3,]<-b3
b4<-c(mean(as.data.frame(PMM2[,,"Averp"]))[1],
mean(as.data.frame(PMM2[,,"Averp"]))[2],
mean(as.data.frame(PMM2[,,"Averp"]))[3],mean(as.data.frame(PMM2[,,"Aver
p"]))[4],
mean(as.data.frame(PMM2[,,"Averp"]))[5])
Jibu1.PMM2301[4,]<-b4

### Jibu2.PMM2301 for 30a: FOR MI- Average of the 1000- OTHER
STATISTICS
Jibu2.PMM2301<-matrix(0,1,3)
col<-c("AvSigma","MASE4","AvPMiss")
rows<-c("Value")
dimnames(Jibu2.PMM2301)<-list(rows,col)
FF<-
c(mean(as.data.frame(PMM2[,,"Int"]))[6],mean(as.data.frame(PMM2[,,"Int"]))[
11],
mean(as.data.frame(PMM2[,,"Int"]))[14])
Jibu2.PMM2301[1,]<-FF

# SUMMARY RESULTS FOR GAM 1
#MAJIBU
### Jibu1.GAM1301 for 30a: FOR MI- Average of the 1000
Jibu1.GAM1301<-matrix(0,4,5)
col<-c("Estimate", "AvSE", "AvLlimit","AvUlimit","AvLengthCI")
rows<-c("Intercept", "SEXE", "LEEFT","AVERP")
dimnames(Jibu1.GAM1301)<-list(rows,col)
b1<-c(mean(as.data.frame(GAM1[,,"Int"]))[1],
mean(as.data.frame(GAM1[,,"Int"]))[2],
mean(as.data.frame(GAM1[,,"Int"]))[3],mean(as.data.frame(GAM1[,,"Int"]))[4],
mean(as.data.frame(GAM1[,,"Int"]))[5])
Jibu1.GAM1301[1,]<-b1
b2<-c(mean(as.data.frame(GAM1[,,"SEX"]))[1],
mean(as.data.frame(GAM1[,,"SEX"]))[2],
mean(as.data.frame(GAM1[,,"SEX"]))[3],mean(as.data.frame(GAM1[,,"SEX"])
)[4],
mean(as.data.frame(GAM1[,,"SEX"]))[5])
Jibu1.GAM1301[2,]<-b2
b3<-c(mean(as.data.frame(GAM1[,,"Age"]))[1],
mean(as.data.frame(GAM1[,,"Age"]))[2],
mean(as.data.frame(GAM1[,,"Age"]))[3],
mean(as.data.frame(GAM1[,,"Age"]))[4],
mean(as.data.frame(GAM1[,,"Age"]))[5])
Jibu1.GAM1301[3,]<-b3
b4<-c(mean(as.data.frame(GAM1[,,"Averp"]))[1],
mean(as.data.frame(GAM1[,,"Averp"]))[2],
mean(as.data.frame(GAM1[,,"Averp"]))[3],
mean(as.data.frame(GAM1[,,"Averp"]))[4],
mean(as.data.frame(GAM1[,,"Averp"]))[5])
Jibu1.GAM1301[4,]<-b4


Jibu2.GAM1301<-matrix(0,1,3)
col<-c("AvSigma","MASE5","AvPMiss")
rows<-c("Value")
dimnames(Jibu2.GAM1301)<-list(rows,col)
FF<-
c(mean(as.data.frame(GAM1[,,"Int"]))[6],mean(as.data.frame(GAM1[,,"Int"]))[1
2],
mean(as.data.frame(GAM1[,,"Int"]))[14])
Jibu2.GAM1301[1,]<-FF

# SUMMARY RESULTS FOR GAM 2
### Jibu1.GAM2301 for 30a: FOR MI- Average of the 1000
Jibu1.GAM2301<-matrix(0,4,5)
col<-c("Estimate", "AvSE", "AvLlimit","AvUlimit","AvLengthCI")
rows<-c("Intercept", "SEXE", "LEEFT","AVERP")
dimnames(Jibu1.GAM2301)<-list(rows,col)
b1<-c(mean(as.data.frame(GAM2[,,"Int"]))[1],mean(as.data.frame
(GAM2[,,"Int"]))[2], mean(as.data.frame(GAM2[,,"Int"]))[3],
mean(as.data.frame(GAM2[,,"Int"]))[4],
mean(as.data.frame(GAM2[,,"Int"]))[5])
Jibu1.GAM2301[1,]<-b1
b2<-c(mean(as.data.frame(GAM2[,,"SEX"]))[1],mean(as.data.frame
(GAM2[,,"SEX"]))[2], mean(as.data.frame(GAM2[,,"SEX"]))[3],mean
(as.data.frame(GAM2[,,"SEX"]))[4],
mean(as.data.frame(GAM2[,,"SEX"]))[5])
Jibu1.GAM2301[2,]<-b2
b3<-c(mean(as.data.frame(GAM2[,,"Age"]))[1],mean(as.data.frame
(GAM2[,,"Age"]))[2], mean(as.data.frame(GAM2[,,"Age"]))[3],
mean(as.data.frame(GAM2[,,"Age"]))[4],
mean(as.data.frame(GAM2[,,"Age"]))[5])
Jibu1.GAM2301[3,]<-b3
b4<-c(mean(as.data.frame(GAM2[,,"Averp"]))[1],mean(as.data.frame
(GAM2[,,"Averp"]))[2], mean(as.data.frame(GAM2[,,"Averp"]))[3],
mean(as.data.frame(GAM2[,,"Averp"]))[4],
mean(as.data.frame(GAM2[,,"Averp"]))[5])
Jibu1.GAM2301[4,]<-b4

Jibu2.GAM2301<-matrix(0,1,3)
col<-c("AvSigma","MASE6","AvPMiss")
rows<-c("Value")
dimnames(Jibu2.GAM2301)<-list(rows,col)
FF<-c(mean(as.data.frame(GAM2[,,"Int"]))[6],mean(as.data.frame
(GAM2[,,"Int"]))[13], mean(as.data.frame(GAM2[,,"Int"]))[14])
Jibu2.GAM2301[1,]<-FF


#### FINAL RESULTS === 301
#Complete cases; Jibu1.Scc301; Jibu2.Scc301
#Single mean Imputation; Jibu1.SSI301; Jibu2.SSI301
#Conditional mean Imputation; Jibu1.SCM301; Jibu2.SCM301
#PMM 1 ; Jibu1.PMM1301; Jibu2.PMM1301
#PMM 2 ; Jibu1.PMM2301; Jibu2.PMM2301
#GAM 1 ; Jibu1.GAM1301; Jibu2.GAM1301
#GAM 2 ; Jibu1.GAM2301; Jibu2.GAM2301
############             END HERE ###############
###### BOXPLOTS FOR THE ESTIMATES AND SE FOR ALL MODELS
par(mfrow=c(3,1))
### for Sex
boxplot(ccase[,"Est","SEX"],simean[,"Est","SEX"],sicmean[,"Est","SEX"],
PMM1[,"Est","SEX"],GAM1[,"Est","SEX"],PMM2[,"Est","SEX"],GAM2[,"Est","S
EX"],
main="Distribution of Estimates for Sex",cex.main=1.2, xlab="Method",
names=c("CC","SM","CM","PMM-I","GAM-I","PMM-II","GAM-II"))

boxplot(ccase[,"Std","SEX"],simean[,"Std","SEX"],sicmean[,"Std","SEX"],
PMM1[,"Std","SEX"],GAM1[,"Std","SEX"],PMM2[,"Std","SEX"],GAM2[,"Std","S
EX"],
main="Distribution of SE for Sex",cex.main=1.2, xlab="Method",
names=c("CC","SM","CM","PMM-I","GAM-I","PMM-II","GAM-II"))
### for Age
boxplot(ccase[,"Est","Age"],simean[,"Est","Age"],sicmean[,"Est","Age"],
PMM1[,"Est","Age"],GAM1[,"Est","Age"],PMM2[,"Est","Age"],
GAM2[,"Est","Age"], main="Distribution of Estimates for Age",cex.main=1.2,
xlab="Method",
names=c("CC","SM","CM","PMM-I","GAM-I","PMM-II","GAM-II"))

boxplot(ccase[,"Std","Age"],simean[,"Std","Age"],sicmean[,"Std","Age"],
PMM1[,"Std","Age"],GAM1[,"Std","Age"],
PMM2[,"Std","Age"],GAM2[,"Std","Age"],
main="Distribution of SE for Age",cex.main=1.2, xlab="Method",
names=c("CC","SM","CM","PMM-I","GAM-I","PMM-II","GAM-II"))
### for Averp
```

```
boxplot(ccase[,"Est","Averp"],simean[,"Est","Averp"],sicmean[,"Est","Averp"],
PMM1[,"Est","Averp"],GAM1[,"Est","Averp"],
PMM2[,"Est","Averp"],GAM2[,"Est","Averp"],
main="Distribution of Estimates for Av. Trips",cex.main=1.2, xlab="Method",
names=c("CC","SM","CM","PMM-I","GAM-I","PMM-II","GAM-II"))

boxplot(ccase[,"Std","Averp"],simean[,"Std","Averp"],sicmean[,"Std","Averp"],
PMM1[,"Std","Averp"],GAM1[,"Std","Averp"],PMM2[,"Std","Averp"],GAM2[,"St
d","Averp"],
main="Distribution of SE for Av. Trips",cex.main=1.2, xlab="Method",
names=c("CC","SM","CM","PMM-I","GAM-I","PMM-II","GAM-II"))


###### BOXPLOTS FOR MASE FOR ALL MODELS
par(mfrow=c(3,1))
boxplot(ccase[,"ASE0",],simean[,"ASE1",],sicmean[,"ASE2",],
PMM1[,"ASE3",],GAM1[,"ASE5",],PMM2[,"ASE4",],GAM2[,"ASE6",],
main="Distribution of ASE values, MCAR-30%",cex.main=1.2, xlab="Method",
names=c("CC","SM","CM","PMM1","GAM1","PMM2","GAM2"))

### PLOTS OF MASE
pp<-c(0,0.2918429,0.4995192); asecc<-c(0,5856.839,9487.931)
asesmi<-c(0,1482.966,2426.333); asecmi<-c(0,12.71369,9219.44)
asemi1<-c(0,4875.289,9428.789); asemi2<-c(0,2509.473,4766.304)
asegam1<-c(0,2764.912,9116.295)
asegam2<-c(0,1841.889,2461.796)
plot(pp,asecc, type="b", lty=1, xlab="Missingness Proportion",
cex.main=1.2, font.main=2,main="MASE values under
MCAR",ylab="MASE",ylim=range(0,10000),
xlim=range(0,0.6),lwd=1.5)
lines(pp, asesmi, type="b",col="purple", lty=3,lwd=2)
lines(pp, asecmi, type="b", col="blue",lty=5,lwd=1.5)
lines(pp, asemi1, type="b", col="yellow",lty=7,lwd=2.1)
lines(pp, asemi2, type="b", col="green",lty=9,lwd=2.1)
lines(pp, asegam1, type="b", col="red", lty=11,lwd=2)
lines(pp, asegam2, type="b", col="magenta", lty=13,lwd=2)
lgg<-c("CC","SM","CM","PMM-I","PMM-II","GAM-I","GAM-II")
legend(locator(1),legend=lgg ,lty=1:13, ncol=2, adj = c(0, 0.5),
col = c("black","purple","blue","yellow","green","red","magenta"),lwd=2)
```

```
#### EVALUATING EFFECT OF FRACTION OF COEFFICIENTS ###

#### FOR FIXED PSI
beta.nf<-jibu.od[3,1]*c(0.4,0.5,0.6,0.7,0.8, 0.9, 1.0, 1.1, 1.2,1.3)
f.psi<-psi1/beta.nf  ### fixed psi
ppp<-rep(0,10)
ind<-matrix(0,length(y),10)
for (h in 1:length(f.psi)){
pp<-expit(psi0+(f.psi[h])*(fitd.od-(jibu.od[1,1])
-(jibu.od[2,1])*SEXE-(jibu.od[4,1])*AVERP)+psi2*SEXE+psi3*AVERP)
ppp[h]<-sum(cbind(pp),na.rm=TRUE)/length(y)
ind[,h]<-pp

##### FOR FIXED BETA
#psi.nf<-psi1*c(0.4,0.5,0.6,0.7,0.8, 0.9, 1.0, 1.1, 1.2,1.3)
psi.nf<-psi1*c(0.8, 0.9, 1.0,1.1, 1.2,1.3,1.4,1.5,1.6,1.7)
f.beta<-psi.nf/jibu.od[3,1]   ### fixed beta
BBB<-rep(0,10)
ind<-matrix(0,length(y),10)
for (h in 1:length(f.beta)){
BB<-expit(psi0+(f.beta[h])*(fitd.od-(jibu.od[1,1])
-(jibu.od[2,1])*SEXE-(jibu.od[4,1])*AVERP)+psi2*SEXE+psi2*AVERP)
BBB[h]<-sum(cbind(BB),na.rm=TRUE)/length(y)
ind[,h]<-BB
}

### plots
par(mfrow=c(1,2))
plot(f.psi,ppp,xlab="psi/beta.Age",ylab="P(R=1)-MAR",
cex.main=0.9, main="For fixed psi.Age-1st case",ylim=range(0,1))
plot(f.beta,BBB,xlab="psi/beta.Age",ylab="P(R=1)-MAR",
cex.main=0.9, main="For fixed beta.Age-1st case",ylim=range(0,1))


win.graph()
par(mfrow=c(3,2))
plot(LEEFT,ind[,1],ylim=range(0,1))
#title(locator(1),main="Probability of missingness with age when Psi is fixed")
plot(LEEFT,ind[,2],ylim=range(0,1))
plot(LEEFT,ind[,3],ylim=range(0,1))
plot(LEEFT,ind[,4],ylim=range(0,1))
plot(LEEFT,ind[,5],ylim=range(0,1))

win.graph()
par(mfrow=c(3,2))
plot(LEEFT,ind[,6],ylim=range(0,1))
plot(LEEFT,ind[,7],ylim=range(0,1))
plot(LEEFT,ind[,8],ylim=range(0,1))
plot(LEEFT,ind[,9],ylim=range(0,1))
plot(LEEFT,ind[,10],ylim=range(0,1))
```

Auteursrechterlijke overeenkomst

Opdat de Universiteit Hasselt uw eindverhandeling wereldwijd kan reproduceren, vertalen en distribueren is uw akkoord voor deze overeenkomst noodzakelijk. Gelieve de tijd te nemen om deze overeenkomst door te nemen, de gevraagde informatie in te vullen (en de overeenkomst te ondertekenen en af te geven).

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Dealing with Missing Data in Cross Sectional Data on Transport
Richting: **Master of Science in Biostatistics** Jaar: **2007**
in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Ik ga akkoord,

Susan Fred Rumisha

Datum: **28.08.2007**