Erasmus University Rotterdam

Tatev Karen Aslanyan

Topics in Advanced Statistics

Evaluatin of the performance of OLS and MM Estimations when using Single Imputation and Multiple Imputation

---

1 Introduction

Missing data is a widely-known issue in numerous fields of scientific research mainly because most of the statistical methods require complete data. Missing values in the data can have different reasons: respondents can mistakenly skip question resulting in nonresponse,the data might be combined from different surveys leading to incomplete information, failure in the network leading to the loss in the data and sometimes individuals consciously skip some questions which they might have found too personal, embarrassing or they simply didn't want to share that information. Especially when dealing with large data sets,very often observations that contain missing values are being simply removed from research to get complete data and perform the analysis. This might lead to biased results with lower statistical power. Therefore, it is important to know the reason for missingness in the data and it's effect on the analysis. Rubin and Little(2002) have introduced three types of missing data mechanisms: Missing At Random(MAR), Missing Completely At Random(MCAR) and Missing Not At Random(MNAR).

Numerous methods have been developed to handle this problem depending on the nature of the data and missingness pattern of data. Listwise Deletion(LD) is a widely-known method for handling missing data which simply removes cases with missing data and performs the analysis using remaining cases. This method is simple and works well when data is MCAR, leading to unbiased parameter results (Allison, 2001). But even in case of MCAR data, the loss of observations reduces the power by increasing standard errors. Moreover, when data is MNAR or MAR LD leads to biased results. Another way of handling missing data is by means of mean/median/mode imputations. This univariate type of imputations replaces the missing values with the mean/median(mode) of the observed values when the target variable is continuous(categorical) respectively. However, univariate imputation often ruins the multivariate structure of the data leading to inaccurate and biased results.

Troyanskaya et al.(2001), has introduced K-Nearest-Neighbor(KNN) single imputation method which aims to find for missing observation their K most similar neighbors(donors) in the observed data and use them to estimate the missing value. This method falls in the category of Donor-based methods since it uses K most similar donors to estimate the missing value. Another Donor-based method is Hot-deck imputation method, developed by US Census Bureau in 1950, which simply goes though the data sequentially and assigns to missing value the last observed value of that variable. One of it's major advantages is it's computation time. However, both of these methods might destroy the multivariate structure of the data leading to unreliable results.

More statistical method is model-based imputation which is based on regression model which estimates missing values. When we use imputed data in the regression analysis the estimation results are based on the assumption that the data has been fully observed and the uncertainty that missing values have been actually imputed are not taken into account. Therefore, the standard errors and the significant tests might be inaccurate (un-

derestimated). To avoid these problems of single imputations, Bootstrap and Multiple Imputation(MI) methods have been proposed to include this uncertainty. The idea behind the Bootstrap is to simulate the distribution of the estimator by resampling from the observed sample(Efron and Tibshirani,1993). However, this method works only when assumption of asymptotic normality of the estimator holds and it requires large amount of computation time. In this analysis we use both iterative model-based single imputation and multiple imputation to analyse the effect of imputation on the Ordinary Least Squares(OLS) and MM-robust regressions. We aim to answer following research questions:

• What is the impact of single imputation (SI) and multiple imputation on the accuracy of OLS and MM estimation results.

• Do model-based single imputation and multiple imputation techniques perform equally well when the data is MAR, MCAR and MNAR?

• Do outliers effect the estimation results?

The remainder of this report is structured as follows: Section 2 describes missing data mechanisms used for the simulation study and the methodology of iterative based single imputation and multiple imputation. Section 2 also describes the model selection technique that is used to select the model for the regression analysis. Section 3 presents the data used in the simulation study, the selected model and data generating process in the simulation study. Section 4 describes the simulation results and provide the answers to research questions.

2 Methodology

Let us denote complete data by $\mathbf{X}$, observed part of this data by $\mathbf{X_{obs}}$ and unobserved(missing) part of the data by $\mathbf{X_{miss}}$ such that $\mathbf{X_{obs}}$ and $\mathbf{X_{miss}}$ together form the entire data set. Then above mentioned three types of missing data mechanisms can be formulated as follows:

$$\begin{aligned} &MCAR : P(X_{miss}) = P(X_{miss}) \\ &MAR : P(X_{miss}) = P(X_{miss} \mid X_{obs}) \\ &MNAR : P(X_{miss}) = P(X_{miss} \mid X_{obs}, X_{miss}) \end{aligned} \tag{2.1}$$

Data is MCAR if the missingness in the data has not been caused by the observed and neither the unobserved parts of the data. Therefore, from equation 2.1 we can see that the probability of the missing data part $X_{miss}$is not dependent on the observed or the missing data. Data is MAR if the missingness has been caused by observed,but not unobserved, part of the data(Schafer and Graham, 2002). From equation 2.1 we see that probability of missingness conditional probability, conditioned on observed data part only. Finally, the data is MNAR if the probability of missingness is dependent on both observed and unobserved parts of the data. It is very difficult to investigate whether the data is MAR or MNAR when the variables are correlated.

2.1 Model Based Single Imputation

We use the iterative model-based single imputation as one method of imputing the missing values in the data. The idea behind it is to construct statistical model for each variable in

the data that contains missing values and to repeatably go through it until convergence. In each step of the iteration, one variable is used as a dependent variable and the rest of variables serve as independent variables. So, the entire multivariate information in data will be used for imputation in dependent variable, which means that the multivariate structure of data will stay intact. The algorithm IRMI, with corresponding function irmi(), which provides a software tool in R from package VIM will be used to perform iterative model-based imputation. One of the main reasons of using IRMI instead of for instance it's well-known alternative IVEWARE is its advantage with respect to robustness since we also want to study the impact of outliers on chosen estimation. IVEWARE and all the other alternatives of IRMI cannot adequately handle a data including outliers. The detailed description of iterative model-based algorithm can be found in Appendix.

2.2 Multiple Imputation

As it was mentioned earlier, single imputation used in regression analysis might lead to inaccurate results in terms of standard errors and significant tests. Ranjit Lall (2016) studied thisin his large scale analysis and has shown that when accounting for the uncertainty of missing data,key results in a large amount of research studies in the field of political science become insignificant. He also showed that MI can really make a large difference in terms of solving this problem. As the name of the method already suggests, the idea behind multiple imputation is to impute the same missing data for multiple times. By determining the within- and between imputation variability, estimation variability when using different imputation sets, and combining these variability's we expect that this will result in more accurate estimates of standard errors and significant tests.

The theoretical motivation behind MI is Bayesian and it requires independent random draws from the posterior distribution $\mathrm{f}_{X_{mis}|X_{obs}}$ of the missing part of the data X_{mis} conditional on observed part X_{obs}. The MI algorithm is described in detailed in Appendix. Let us denote $\hat{W}^*$ the average within-imputation variance with U_r^* as estimated variance of T assuming that $\hat{\mathbf{X}}_\mathbf{r}^*$ is the true observed data[1]. Moreover, let us denote $\hat{B}^*$ the between-imputation variance and (m+1)/m is a correction factor for the cases when m, number of imputations, is small. MI works properly if and only if the conditional distribution $\hat{\mathbf{X}}_r^*$ is right. We assume that following conditions for defining conditional distribution as "proper" are satisfied.[2] $\bar{T}^*$ and $\hat{V}^*$ can be used to construct significance tests based on asymptotic normality. This holds when m is large but when m is small, the distribution of $\bar{T}^*$ should be adjusted for $\hat{V}^*$. We use inference based on t-test proposed by Rubin(1987):

$$v_m = \frac{m-1}{\gamma_m^2} \qquad \hat{\gamma} = \frac{m+1}{m}\frac{\hat{B}^*}{\hat{V}^*}$$

Where the v_m is the proposed df of the t-test and $\hat{\gamma}_m$ is the estimated fraction of relevant missing information. However, Bernard and Rubin(1999) have argued that the v_m can become larger than the complete data degrees of freedom. Therefore, to account for this

[1] It is equal to the squared standard error from the regression.

[2] $\bar{T}^*$ is consistent and asymptotically normal and $\hat{V}^*$,estimator of asymptotic variance, is weakly consistent (Rubin,1987).

they suggested following adjustments for the number of degrees of freedom:

$$\tilde{v}_m = \frac{v_m v_{obs}}{v_m + v_{obs}} \qquad v_{obs} = \frac{v_{comp}+1}{v_{comp}+3} v_{comp}(1-\hat{\gamma}_m)$$

v_{comp} is the degrees of freedom of the complete data equal to n - p - 1 where n is the number of observations in the entire data and p is the number of independent variables. This adjustment assures that the $\tilde{v}_m$,the harmonic total of two components, is monotonically increasing in v_{comp} while $\tilde{v}_m$ is smaller than v_{comp}.

2.3 Performance Measures

In order to evaluate the performance of single and multiple imputations on point and standard error we will use four regression based characteristics. For the accuracy of point estimates we use the bias which is the difference between expected value of the estimator and its true value being estimated which is defined as follows:

$$Bias_\beta(\hat{\beta}) = \mathbf{E}_{X|\beta}[\hat{\beta}] - \beta = \mathbf{E}_{X|\beta}[\hat{\beta} - \beta]$$

Another accuracy measure that we use to evaluate performance of regression estimates is the Mean Squared Error(MSE) which incorporates both variance and bias of the estimates.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Where the Y is the is the vector of observed values of the variable that has to be predicted, in our case it is the logMedValue. Moreover, the $\hat{Y}$ is the predicted value determined by the regression. We also calculate the coverage rate of the 95% confidence intervals which is defined as the amount of samples for which the known population parameter falls in the 95% confidence interval. CR is a combined performance measure for both coefficient estimates and standard errors. Note that the 95% CI of β is defined as follows $\hat{\beta} \pm t_{0.95}\sqrt{\hat{V}^*}$. It is worth to mention that we don't use the R^2 since they don't make since in robust regression.

3 Data and Simulation study

3.1 Data generating Process

In this analysis we use the Boston housing data introduced by Harrison and Rubinfeld (1978) later corrected by Gilley and Kelley Pace (1996). Detailed description of the variables used in this analysis can be found in the appendix. However, instead of using their proposed regression model we use the Bayesian Information Criterion (BIC) (Schwarz, 1978) to perform forward stepwise variable selection from robustness point of view. Assuming that the error terms are normally distributed, we define BIC as BIC = n*ln($\hat{\sigma}^2$) + k*ln(n) where $\hat{\sigma}^2$ is the maximum likelihood estimate of the residual standard deviation, k is number of estimated coefficients and n is number of observations. Using the lmrob() function from the robustbase package in R minimization of BIC leads to following regression model:

$$logMedValue \sim RoomsSq + Pre1940 + logDistance + PTRatio + Black$$
$$+ logStatus + CrimeRate + logHighways + Tax + NOxSq$$

For performing a model-based simulation we randomly draw first 100 then 400 observations from the entire data set. We then use these data sets and the selected regression method to perform OLS and MM estimation which we call as TRUE estimation results, which is based on the complete data before introducing missing values.

In order to investigate the impact of imputation on these two regression methods we perform following steps each simulation iteration:
Step 1: Randomly draw n observations from the Boston data set: $S_{complete}$
Step 2: Introduce missing observations with λ amount to all independent variables with equal proportions with missing data mechanism M: S_{NA}
Step 3: Using multimp() function impute data with both SI and MI : $S_{imputed}$
Step: Perform both OLS and MM estimation on the imputed data using created fit() function.
By using pool() function pool the regression results with corresponding pooled estimation results. In the algorithm, λ = {10%, 20%,30%, 50%}, M = {MCAR, MNAR, MAR} , n = {100,400} with equal proportions to all independent variables. The multimp() function is created in such a way that it is able to perform both single and multiple imputations and as default value m = 10. However, in case of amount of NA's in the input data that has to be imputed is larger than 10% it puts as value of m the closest integer equal to the percentage of the incomplete observations in the data, this is based on simple rule of thumb recommended by Bodner(2008) and White et al.(2011). We repeat above formulated by using "KNN" as initialisation method.[3].

In order to implement three different missing data mechanisms we use the ampute() function of package MICE. Under MCAR the probability of being missing is unrelated to any other variable in the dataset and the missingness is a random sample from the observed data we generate missing values completely at random. In this case missing values in data are then just a random subset of that data and probability of missingness is the same for all observations. Unlike MCAR in case of MAR missing values are added in such a way that the propensity for a data point to be missing is unrelated to the missing data but it is related to some of the observed data. For instance, for the independent variable PTRatio the missing values are added in such a way that PTRatio is missing with higher probability fro higher values of CrimeRate. So, the missingness probability depends only on the available information but not on the missing values. Finally, under MNAR we generate missing values in such a way that probability of missingness depends on missing values themselves, for instance probability of missing in PTRatio is higher for lower PTRatio values.

[3]The Hot-deck imputation can't be use in irmi() function since it doesn't draw imputed values from distribution and therefore is not suitable for multiple imputation

It order to evaluate performance of OLS and MM with SI and MI when the data contains outliers we will add 25% contamination at each iteration. In this way we can compare the results of both imputation methods for both OLS and MM regressions to investigate their sensitiveness to outliers. We again randomly sample n observations from the Boston housing data but this time we replace a part of this sampled data by the sample with the size of n*0.25 generated from a multivariate distribution with mean μ_ε and covariance matrix Σ_ε such that the data used in the simulation becomes: (1-ε)Boston$_n$ + $\varepsilon N_p(\mu_\varepsilon,\Sigma_\varepsilon)$ where $\mu_\varepsilon = 100\mathbf{1}_{p+1}$ and $\Sigma_\varepsilon = 0.25^2\mathbf{I}_{p+1}$. We also keep the indices of added outliers and determine the percentage of these outliers that get weight 1 by MM estimation, the average of these percentages we report in the simulation results.
So, the purpose is to calculate all performance measures firstly on the complete data without missing, imputed and outlying observations and then by means of simulation add missing values/outliers to this complete sample to recalculate all performance measures. Once the performance measures are calculated for both complete and simulated data, these two sets of results can be compared per case.

3.3 Detection of outliers in multiple imputed datasets

We use the MM robust estimation as an alternative to linear regression OLS in order to handle contaminated data. MM estimation is a combination of M estimation (Huber, 1973) and S estimation(Rousseeuw and Yohai,1984). Different methods have been proposed to detect outliers. Often times even robust methods like MM-robust estimation don't correctly exclude outliers. It is worth to mention that there are two types of outliers rowwise and cellwise and according to Agostinelli et al.(2015) and Leung et al.(2016) these two should be handled seperatly. Baarda (1968), Cross et al. (1994) and Teunissen (1998) have proposed w-test for identifying outliers in data. The statistic w is normalised residual for observation i:

$$w_i = \frac{\mid h_i^T P v \mid}{\sigma_0\sqrt{h_i^T P Q_v P h_i}} \quad v = A\hat{X} - l \quad Q_v = P^{-1} - A(A^T P A)^{-1}A^T$$

Where h is a vector [0...1...0] with ith element equal to one,v is the matrix of residuals, A is the design matrix,$\hat{X}$ is matrix of parameters, l is measurement matrix and P is the weight matrix. If |w$_i$| is larger than $\mathbf{N}_{1-\alpha/2}$(0,1) is defined as an outlier. One can also use function lmrob() to perform MM regression which output object contains the weights and we use the function summarizeRobWeights() to extract this weights and return those where the absolute value of the weight is smaller than given cut-off value with argument eps. Other possible outlier detection rules are proposed in the Appendix.

3.2 Simulation Results

Table 1 presents the OLS and MM regression results applied on the entire data set of 506 observations with chosen model selected earlier. We observe that almost all variables except of PRE1940 are highly significant in OLS and all significant in MM case. Table 2 presents the mean MSE, mean Bias, mean of standard errors and mean of coverage rates

per regression type for different proportion of missingness for n = 100 and n = 400 settings for both single and multiple imputations where MCAR data generating process has been used. We see that increasing amount of missingness in data lead to increase in mean MSE for both OLS and MM. Moreover,we observe that mean bias is 0 in OLS regression for all missingness rates which is not the case in MM estimation.

What we also find is that when missingness is 10% the mean MSE and mean standard errors are very close to the their true values for both single and multiple imputations. However, when there is more than 10% missingness in the data mean SE is lower than the true mean SE for single imputation case when n = 100 while in MI the SE estimates are close to the true one for both low and high amount of missingness. So, when comparing the mean SE's for single and multiple imputation for n = 100 case we observe that for OLS the values in case $\lambda = 10\%$, SI and MI perform equally well but when = 20% or larger, the mean standard errors are much smaller then the true value in SI and almost the same as TRUE value in MI, even for 50% of missingness. We also see that MM doesn't perform better than OLS in terms of bias and estimation of standard errors. Comparing to the case when n = 400 we observe similar patters as discussed earlier. However, we observe that now SI OLS performs well also when $\lambda = 20\%$ but not for higher missingness values while MI provides estimates close to the TRUE values for all λ. In terms of coverage rates, we see that CR is close to 95% when the missingness amount is small and SI is applied while the CR rates are very close to 95% for all cases when MI is applied. Table 3 presents the results when MAR mechanism has been used. What we observe is that for n = 100 bias has been introduced for both OLS and MM estimations. OLS still outperforms MM estimation in terms of average bias, standard errors and MSE. Bias vanishes when MI is used instead of SI. We see that $\bar{SE}$ is almost two times less than the true $\bar{SE}$ for the small sample size with single imputation. However, for large sample size there is small bias in case of SI and no bias in case of MI. Coverage rates are far from 0.95 for n = 100 when SI is used and get closer to 0.95 when MI is used. Moreover, for n = 400 it holds that for small amount of missingness(10% and 20%) the mean SE in case of both SI and MI are very close to the true values while SI underestimates the standard errors when larger amount of missingness is introduced, in those cases MI outperforms SI. In terms of coverage rates MI leads to better results for all possible missingness amounts but still not very close to 0.95. Finally, what we also observe is that for $\lambda = 50\%$ for both SI and MI both OLS and MM perform poorly.

Table 5 and Table 6 present the results for the cases when beside of adding missing observations to randomly sampled data set also 25% missing observations has been added. What we observe is that when missing data mechanism is MCAR for almost all cases the mean bias is approximately 0. However, the estimates are not very close to the true estimates in single imputation, the coverage rate is close to 0.95 only for 10% missingness case. Mean SE is much lower than it's true value in case of SI and closer to it's true value in MI. OLS performs poorly in terms of mean MSE, mean SE and mean CR. MM outperform OLS in all cases for multiple imputation and for single imputation MM results in reliable estimates when missingness is small. Unlike MCAR, in case of MAR data estimates of

both OLS and MM are little biased even though this bias decreases in case of multiple imputation for MM regression but not for OLS. For almost all cases it holds that mean SE is lower in SI than in MI. For MI the MM estimation performs well when not more than 20% missingness has been added. Finally, case of MNAR both MI and SI result in biased and inaccurate estimates for both OLS and MM estimations even though MM still outperforms OLS but CR rates are very far from 0.95 in all cases.

Table 7 presents the MNAR results and we observe that compared to MAR and MCAR the mean bias is much larger in almost all cases for both OLS and MM for both single and multiple imputations. However when only small amount of missing data is introduced (10%) OLS regression lead to low almost no mean bias and also performs well in terms of mean SE and mean MSE. Moreover, in those cases especially for larger sample when MI is used the OLS perform well. When there is more than 10% missingness in the data for both SI and MI OLS and MM estimates are biased and perform poorly in terms of coverage rates and precision. Even though in case of both single and multiple imputations both OLS and MM lead to biased estimates the bias is much lower in case of MI than in SI.

Conclusion

In this analysis we have investigated the impact of single an multiple imputation of missing and contaminated data with missing data mechanisms MCAR, MNAR and MAR on the accuracy of OLS and MM estimates. Simulation study based on Boston Housing data showed that when data is MCAR both OLS and MM estimates are unbiased. Both SI and MI lead to accurate results when missingness proportion is small(e.g. 10%), but since MI requires much more computation time, SI is preferred for this reason. When missingness amount is larger SI underestimates the mean standard errors in OLS and in MM, while MI performs well therefore is preferred. When sample size is larger SI leads to accurate results for OLS even when 20% of data is missing. MAR data leads to biased estimates for small samples especially when single imputation has been used and biased decreases when MI is used instead, here also OLS outperforms MM. When n is large and MI is used OLS estimates are unbiased and accrate unless data contains 50% missing values. Under MNAR both SI and MI performs poorly for both OLS and MM except when only small amount of data is missing. When n is large MI leads to accurate and almost unbiased OLS estimates. So, SI performs well in terms of computation time for MAR and MCAR data mechanism for OLS regression when small amount of data is missing otherwise MI outperforms SI. However, when data is MNAR only for small missingness and large sample size MI leads to accurate results for OLS. Finally, we found that significant amount of outliers (25%) effect the OLS estimation results significantly for both SI and MI but when for MCAR and MAR data MI leads to accurate results for MM estimation.

4 References

Alfons, A, Templ, M, Filzmoser, P, 2009. On the influence of imputation methods on laeken indicators: Simulations and recommendations, 10.

Alfons A, Templ M, Filzmoser P 2010. "An Object-Oriented Framework for Statistical Simulation: The R Package simFrame." Journal of Statistical Software, 37(3), 1–36.

Efron B and Tibshirani R.J,1993: An introduction to the bootstrap. Chapman Hall, 436.

Gilley O.W and Kelley Pace R, 1996. On the Harrison and Rubinfeld data. Journal of Environmental Economics and Management, 31(3), 403-405.

Harrison D and Rubinfeld D.L, 1978. Hedonic housing prices and the demand for clean air, 5(1), 81-102.

Barnard J. and Donald B, Rubin B, 1999. Small-Sample Degrees of Freedom with Multiple Imputation, 86(4), 948-955.

Lall R., 2016. How Multiple Imputation Makes a Difference, 24(4), 414-433.

Peter J. Huber, 1973. The Annals of Statistics, 1(5), 799-821.

Raessler S, Munnic R.,2004. The impact of multiple imputation for DACSEIS, 2000-26057.

Rousseeuw, P.J. and Yohai, V., 1984. Robust Regression by Means of S Estimators in Robust and Nonlinear Time Series Analysis, 5(4), 256-274.

Rousseeuw P.J. and Van Zomeren B.C.,1990. Unmasking Multivariate Outliers and Leverage Points, American Statistical Association, 85(411), 635-636.

Rubin D.B.,1987. Multiple Imputation for Nonresponse in Surveys.

Schafer J.L., and Graham J.W., 2002. Missing Data: Our View of the State of the Art. Psychological Methods, 7, 147-177.

Schouten R.M, Peter Lugtig and Vink G.,2018. Generating missing values for simulation purposes: a multivariate amputation procedure, 88(15), 2909-2930.

Schopfhauser S, Templ M, Alfons A, Kowarik A, Prantner B, 2016. VIMGUI: Visualization and Imputation of Missing Values. R package version 0.10.0.

Schwarz G.,1978.Estimating the dimension of a model. The Annals of Statistics, 6(2), 461-464.

Templ M., Kowarikb A., Filzmosera P., 2011. Iterative stepwise regression imputation using standard and robust methods, 8(10)

Templ M., Alfons A., Kowarik, A., 2009. VIM: Visualization and Imputation of Missing Values.

Yohai V.J, 1987. High Breakdown-Point and High Efficiency Robust Estimates for Regression, 15(2).

5 Appendix

5.1 IRMI algorithm

IRMI algorithm can be summarized as follows:
Step 1: Initialization of missing values using imputataion techniques KNN or median imputataion.
Step 2: Sorting the varaibles according to their original amount of missingness(this will increase the computationl efficiency of the method).
Step 3: Estimate the statistical model for with x_j as dependent variable and remaining x_k's k≠j for all variables j = 1,...,p (column-wise) to obtain $\hat{\mathbf{X}}$
For each observation x_{ij} for i = 1,...,n if x_{ij} is missing in the original data, set it to predicted value from the estimated model of Step 3 and let it unchanged if it's value is observed in the original data. That is :

$$\hat{x}_{ij} = \begin{cases} g(X_{obs}) & \text{if } x_{ij} = NA \\ x_{ij} & \text{else} \end{cases}$$

Step 4: Repeat Step 2 and Step 3 until imputed values converge.
Step 5: Return imputed data matrix X.
We use the function irmi() in the created multimp() function where as default values for maximum number of iterations we use 200 and in the case where we use the robust version of the algorithm for the fully iterated best candidates for the fast-S algorithm we use as maximal 1000 refinement steps.

5.2 MI algorithm

The MI algorithm can be described as follows:
Step 1: For each repetition r = 1,...,m impute the missing values in X_{miss} by drawing from the estimated $\hat{\mathbf{f}}_{\mathbf{X_{mis}}|\mathbf{X_{obs}}}$ to obtain the complete imputed data matrix $\hat{\mathbf{X}}_r^*$.
Step 2: Compute the corresponding regression estimates $T_r^* = T(\hat{\mathbf{X}}_r^*)$.
Step 3: Compute Pooled estimate $\bar{T^*}$ and variance $\hat{V^*}$ as follows:

$$\bar{T^*} = \frac{1}{m}\sum_{r=1}^{m} T_r^* \qquad \hat{V^*} = \hat{W^*} + \frac{m+1}{m}\hat{B^*}$$

$$\hat{W^*} = \frac{1}{m}\sum_{r=1}^{m} U_r^* \qquad \hat{B^*} = \frac{1}{m-1}\sum_{r=1}^{m}(T_r^* - \bar{T^*})^2$$

5.3 Detection of Outliers

M estimation is based on IRLS [4] algorithm as generalisation of Maximum Likelihood Estimation (MLE) which is similar to Weighted Least Squares but it defines the weights in such a way that they depend on data points. However, it might lead to non-robust estimators if there are bad-leverage points in the data. The S estimation (Rousseeuw

[4] IRLS: Iteratively Reweighted Least Squares

and Yohai,1984) which is generalization of OLS based so called I-steps do result in robust estimators but those are inefficient. The MM combines these two to get robust and efficient estimates. There are different ways to detect outliers in the data. The classical approach is based on the z-scores of the observations determined by z_i = (x_i - $\bar{x}$)/s) where $\bar{x}$ is the mean and s is the standard deviation of x. The observations whose absolute value of z_i is larger than some threshold value (e.g $|z_i| \geq 2.5$) are defined as outliers. However, this approach would lead to incorrect results because the formula is based on nonrobust estimates of mean $\bar{x}$ and standard deviation s drawn toward the outlier. Therefore, to make this rule more accurate we use robust estimators of location (median) and scale (MAD) instead to get accurate outlier detection rule as follows:

$$d_i = \begin{cases} 1 & \text{if } | \dfrac{(x_i - median(x_i)_{j=1,...,n})}{MAD} | \geq 2.5 \\ 0 & \text{else} \end{cases}$$

Another, widely-known way of detecting outlier is by means of Mahalanobis distances, however this method is known by its masking effect, when group of points can mask the impact of each other. To solve this problem Robust Distances were introduced which proposed to use the MVE robust estimates for location and scatter.
In case of regression analysis where X_i is a p-dimensional vector and Y_i one-dimensional response Rousseeuw and Van Zomeren(1990) the method of standardised residuals which can be used to detect outliers in MM robust estimation. Firstly, MM estimation has to be used to fit the model and get all estimates of regression model $Y_i = \beta X_i + \varepsilon_i$. Then we calculate the residuals using high-breakdown estimation method Least Median of Squares(LMS) which is: $\min_{\hat{\beta}} \mathbf{median}_{i=1,...,n} \epsilon_i^2$ where ϵ_i is the residual of i_{th} observation. Then we standardise the LMS residuals SR_i for some positive constant k as follows:

$$SR_i = \frac{\epsilon_i}{\hat{\sigma}_{LMS}} \quad \hat{\sigma}_{LMS} = k\sqrt{median_{i=1,...,n}\epsilon_i^2} \quad for\ i = 1, ..., n$$

Using the fitted values from MM regression, the corresponding standardized residuals and cutoff value 2.5 we define the observation i as outlier as follows:

$$d_i = \begin{cases} 1 & \text{if } | SR_i | \leq 2.5 \\ 0 & \text{else} \end{cases} \qquad for\ i = 1, ..., n$$

The corresponding regression diagnostic plot displaying these standardized residuals versus robust distances can be plotted using plot(...,which = "rdiag") from robustbase() package in R.

5.4 OLS and MM regression results with the entire Boston data

Full Data Estimation,n = 506				
OLS	**Estimates**	**Std. Error**	**t value**	**Pr($\succ$|t|)**
Intercept	4.5872	0.1510	30.3830	2e-16***
RoomsSq	0.0063	0.0013	4.8850	1.40e-06 ***
Pre1940	0.0001	0.0005	0.2940	0.7691
logDistance	-0.1989	0.0301	-6.6200	0.0000 ***
PTRatio	-0.0306	0.0047	-6.5380	1.55e-10***
Black	0.0004	0.0001	3.6560	0.0003 ***
logStatus	-0.3807	0.0246	-15.469	2e-16 ***
CrimeRate	-0.0119	0.0012	-9.8200	2e-16 ***
logHighways	0.0928	0.0180	5.1511	3.74e-07 ***
Tax	-0.0004	0.0001	-4.1273	4.31e-05 ***
NOxSq	-0.6110	0.1095	-5.5791	3.99e-08 ***
MM	**Estimates**	**Std. Error**	**t value**	**Pr($\succ$|t|)**
Intercept	3.4230	0.1152	29.7190	2e-16***
RoomsSq	0.0167	0.0103	16.3050	2e-16***
Pre1940	-0.0152	0.0004	-4.2600	0.0000 ***
logDistance	-1.4680	0.0207	-7.1020	4.31e-12 ***
PTRatio	-0.2664	0.0031	-8.7270	2e-16***
Black	0.0007	0.0008	8.8320	2e-16***
logStatus	-0.1794	0.0197	-9.1260	2e-16 ***
CrimeRate	-0.0292	0.0019	-15.2930	2e-16 ***
logHighways	0.0073	0.0123	5.9100	6.37e-09 ***
Tax	-0.0002	0.0001	-2.7340	0.0065 **
NOxSq	-0.2570	0.0761	-3.3760	0.0008 ***

Table 1: OLS and MM estimation results on the entire Boston data set

Single Imputation, n = 100, MCAR					
λ	**Method**	**$\bar{\text{B}}$ias**	**M$\bar{\text{S}}$E**	**$\bar{\text{S}}$E**	**CR**
TRUE	OLS	0.0000	0.0214	0.0682	0.9500
	MM	0.0000	0.0244	0.0531	0.9500
10%	OLS	0.0000	0.0216	0.0678	0.9399
	MM	-0.0007	0.0272	0.0615	0.9389
20%	OLS	0.0000	0.0317	0.0573	0.8990
	MM	-0.0003	0.0270	0.0619	0.8900
30%	OLS	0.0000	0.0321	0.0539	0.8357
	MM	-0.0005	0.0271	0.0654	0.8300
50%	OLS	0.0000	0.0319	0.0502	0.8301
	MM	-0.0006	0.0290	0.0672	0.8400
Multiple Imputation, n = 100, MCAR					
λ	**Method**	**$\bar{\text{B}}$ias**	**M$\bar{\text{S}}$E**	**$\bar{\text{S}}$E**	**CR**
TRUE	OLS	0.0000	0.0214	0.0682	0.9500
	MM	0.0000	0.0244	0.0531	0.9500
10%	OLS	0.0000	0.0215	0.0680	0.9400
	MM	-0.0007	0.0471	0.0830	0.9496
20%	OLS	0.0000	0.0218	0.0681	0.9399
	MM	-0.0003	0.0468	0.0859	0.9300
30%	OLS	0.0000	0.0219	0.0683	0.9395
	MM	-0.0007	0.0200	0.0843	0.9306
50%	OLS	0.0000	0.0220	0.0681	0.9391
	MM	-0.0005	0.0261	0.0850	0.9301
Single Imputation, n = 400, MCAR					
λ	**Method**	**$\bar{\text{B}}$ias**	**M$\bar{\text{S}}$E**	**$\bar{\text{S}}$E**	**CR**
TRUE	OLS	0.0000	0.0337	0.0355	0.9500
	MM	0.0000	0.0429	0.0266	0.9500
10%	OLS	0.0000	0.0331	0.0355	0.9480
	MM	0.0000	0.0541	0.0220	0.9488
20%	OLS	0.0000	0.0333	0.0357	0.9408
	MM	-0.0006	0.0489	0.0238	0.9490
30%	OLS	0.0000	0.0329	0.0258	0.9001
	MM	-0.0003	0.0449	0.0266	0.8510
50%	OLS	0.0000	0.0330	0.0264	0.8015
	MM	-0.0007	0.0411	0.0251	0.8999
Multiple Imputation, n = 400, MCAR					
λ	**Method**	**$\bar{\text{B}}$ias**	**M$\bar{\text{S}}$E**	**$\bar{\text{S}}$E**	**CR**
TRUE	OLS	0.0000	0.0329	0.0355	0.9500
	MM	0.0000	0.0429	0.0266	0.9500
10%	OLS	0.0000	0.0331	0.0358	0.9497
	MM	0.0000	0.0533	0.0344	0.9356
20%	OLS	0.0000	0.0332	0.0364	0.9318
	MM	-0.0006	0.0473	0.0376	0.9280
30%	OLS	0.0000	0.0329	0.0366	0.9299
	MM	-0.0008	0.0440	0.0381	0.9101
50%	OLS	0.0000	0.0331	0.0376	0.9302
	MM	-0.0007	0.0406	0.0406	0.9200

Table 2: Simulation results with data mechanism MCAR

Single Imputation, n = 100, MAR					
λ	**Method**	**$\bar{\text{B}}$ias**	**$\bar{\text{MSE}}$**	**$\bar{\text{SE}}$**	**CR**
TRUE	OLS	0.0000	0.0214	0.0682	0.9500
	MM	0.0000	0.0244	0.0531	0.9500
10%	OLS	-0.0001	0.0218	0.0672	0.9101
	MM	-0.0007	0.0251	0.0339	0.8100
20%	OLS	-0.0002	0.0331	0.0675	0.7709
	MM	-0.0093	0.0343	0.0354	0.8019
30%	OLS	-0.0009	0.0437	0.0380	0.7512
	MM	-0.0005	0.0426	0.0357	0.7500
50%	OLS	-0.0012	0.0550	0.0711	0.7203
	MM	-0.0190	0.0404	0.0759	0.7109
Multiple Imputation, n = 100, MAR					
λ	**Method**	**$\bar{\text{B}}$ias**	**$\bar{\text{MSE}}$**	**$\bar{\text{SE}}$**	**CR**
TRUE	OLS	0.0000	0.0214	0.0682	0.9500
	MM	0.0000	0.0244	0.0531	0.9500
10%	OLS	0.0000	0.0207	0.0689	0.9209
	MM	-0.0007	0.0226	0.0810	0.8991
20%	OLS	0.0000	0.0297	0.0684	0.9102
	MM	-0.0001	0.0402	0.0892	0.8990
30%	OLS	0.0000	0.0325	0.0671	0.9100
	MM	-0.0015	0.0432	0.0812	0.8080
50%	OLS	-0.0051	0.0559	0.0759	0.8702
	MM	-0.0035	0.0039	0.0907	0.8011
Single Imputation, n = 400, MAR					
λ	**Method**	**$\bar{\text{B}}$ias**	**$\bar{\text{MSE}}$**	**$\bar{\text{SE}}$**	**CR**
TRUE	OLS	0.0000	0.0337	0.0355	0.9500
	MM	0.0000	0.0429	0.0266	0.9500
10%	OLS	-0.0002	0.0332	0.0353	0.9100
	MM	-0.0030	0.0333	0.0328	0.9009
20%	OLS	-0.0004	0.0334	0.0321	0.9108
	MM	-0.0030	0.0422	0.0328	0.9101
30%	OLS	-0.0005	0.0336	0.0257	0.0899
	MM	-0.0046	0.0396	0.0335	0.8010
50%	OLS	-0.0060	0.0476	0.0241	0.0890
	MM	-0.0057	0.0087	0.0341	0.7997
Multiple Imputation, n = 400, MAR					
λ	**Method**	**$\bar{\text{B}}$ias**	**$\bar{\text{MSE}}$**	**$\bar{\text{SE}}$**	**CR**
TRUE	OLS	0.0000	0.0329	0.0355	0.9500
	MM	0.0000	0.0429	0.0266	0.9500
10%	OLS	0.0000	0.0330	0.0352	0.9299
	MM	-0.0002	0.0423	0.0341	0.8910
20%	OLS	0.0000	0.0475	0.0355	0.9210
	MM	-0.0005	0.0412	0.0345	0.9102
30%	OLS	0.0000	0.0336	0.0359	0.9209
	MM	-0.0002	0.0393	0.0907	0.8919
50%	OLS	-0.0090	0.0339	0.0797	0.8901
	MM	-0.0066	0.0393	0.1348	0.8910

Table 3: Simulation results with data mechanism MAR

Single Imputation, n = 100, MNAR					
λ	**Method**	**$\bar{\text{B}}$ias**	**$\bar{\text{MSE}}$**	**$\bar{\text{SE}}$**	**CR**
TRUE	OLS	0.0000	0.0214	0.0682	0.9500
	MM	0.0000	0.0244	0.0531	0.9500
10%	OLS	-0.0009	0.0211	0.1133	0.8900
	MM	-0.0110	0.0261	0.1144	0.8010
20%	OLS	-0.0119	0.0209	0.1132	0.8012
	MM	-0.0117	0.0263	0.1148	0.7999
30%	OLS	-0.0201	0.0277	0.0680	0.7091
	MM	-0.0266	0.0283	0.0805	0.7601
50%	OLS	-0.0209	0.0276	0.0678	0.7080
	MM	-0.0234	0.0258	0.0819	0.7008
Multiple Imputation, n = 100, MNAR					
λ	**Method**	**$\bar{\text{B}}$ias**	**$\bar{\text{MSE}}$**	**$\bar{\text{SE}}$**	**CR**
TRUE	OLS	0.0000	0.0214	0.0682	0.9500
	MM	0.0000	0.0244	0.0531	0.9500
10%	OLS	-0.0005	0.0211	0.1150	0.9201
	MM	-0.0111	0.0262	0.0341	0.9204
20%	OLS	-0.0037	0.0209	0.1162	0.9091
	MM	-0.0112	0.0262	0.0345	0.9081
30%	OLS	-0.0098	0.0177	0.0709	0.8004
	MM	-0.0170	0.0279	0.0911	0.8097
50%	OLS	-0.0117	0.0176	0.0731	0.8097
	MM	-0.0141	0.0249	0.0945	0.7018
Single Imputation, n = 400, MNAR					
λ	**Method**	**$\bar{\text{B}}$ias**	**$\bar{\text{MSE}}$**	**$\bar{\text{SE}}$**	**CR**
TRUE	OLS	0.0000	0.0337	0.0355	0.9500
	MM	0.0000	0.0429	0.0266	0.9500
10%	OLS	-0.0007	0.0332	0.0345	0.9309
	MM	-0.0164	0.0551	0.0312	0.8979
20%	OLS	-0.0110	0.0332	0.0357	0.9340
	MM	-0.0142	0.0510	0.0334	0.8120
30%	OLS	-0.0110	0.0318	0.0359	0.9301
	MM	-0.0168	0.0390	0.0318	0.8220
50%	OLS	-0.0196	0.0311	0.0354	0.9209
	MM	-0.0147	0.0353	0.0326	0.7804
Multiple Imputation, n = 400, MNAR					
λ	**Method**	**$\bar{\text{B}}$ias**	**$\bar{\text{MSE}}$**	**$\bar{\text{SE}}$**	**CR**
TRUE	OLS	0.0000	0.0329	0.0355	0.9500
	MM	0.0000	0.0429	0.0266	0.9500
10%	OLS	-0.0003	0.0332	0.0360	0.9403
	MM	-0.0163	0.0546	0.0341	0.9302
20%	OLS	-0.0029	0.0331	0.0364	0.9102
	MM	-0.0134	0.0496	0.0345	0.9231
30%	OLS	-0.0080	0.0317	0.0370	0.9120
	MM	-0.0072	0.0389	0.0341	0.9001
50%	OLS	-0.0180	0.0312	0.0375	0.8995
	MM	-0.0151	0.0354	0.0348	0.7995

Table 4: Simulation results with data mechanism MNAR

Single Imputation, n = 100, MCAR, $\varepsilon = 0.25$					
λ	**Method**	**$\bar{\text{B}}$ias**	**$\bar{\text{MSE}}$**	**$\bar{\text{SE}}$**	**CR**
TRUE	OLS	0.0000	0.0214	0.0682	0.9500
	MM	0.0000	0.0244	0.0531	0.9500
10%	OLS	0.0000	0.0693	0.0999	0.8633
	MM	0.0000	0.0493	0.0553	0.9301
20%	OLS	0.0000	0.0766	0.0991	0.8601
	MM	-0.0001	0.0421	0.0489	0.9293
30%	OLS	-0.0010	0.0777	0.0989	0.8509
	MM	0.0000	0.0701	0.0491	0.9210
50%	OLS	-0.0007	0.0804	0.1036	0.8007
	MM	0.0000	0.0759	0.0496	0.8619
Multiple Imputation, n = 100, MCAR, $\varepsilon = 0.25$					
λ	**Method**	**$\bar{\text{B}}$ias**	**$\bar{\text{MSE}}$**	**$\bar{\text{SE}}$**	**CR**
TRUE	OLS	0.0000	0.0214	0.0682	0.9500
	MM	0.0000	0.0244	0.0531	0.9500
10%	OLS	0.0000	0.0641	0.0945	0.8791
	MM	0.0000	0.0249	0.0541	0.9402
20%	OLS	0.0000	0.0798	0.0966	0.8380
	MM	-0.0001	0.0234	0.0557	0.9304
30%	OLS	0.0000	0.0891	0.1067	0.8301
	MM	0.0000	0.0278	0.0587	0.9291
50%	OLS	0.0000	0.0813	0.1558	0.8305
	MM	0.0000	0.0491	0.0574	0.8980
Single Imputation, n = 400, MCAR, $\varepsilon = 0.25$					
λ	**Method**	**$\bar{\text{B}}$ias**	**$\bar{\text{MSE}}$**	**$\bar{\text{SE}}$**	**CR**
TRUE	OLS	0.0000	0.0337	0.0355	0.9500
	MM	0.0000	0.0429	0.0266	0.9500
10%	OLS	0.0000	0.0396	0.0421	0.9230
	MM	0.0000	0.0420	0.0263	0.9398
20%	OLS	0.0000	0.0412	0.0521	0.9108
	MM	0.0000	0.0435	0.0278	0.9301
30%	OLS	0.0000	0.0419	0.0582	0.8801
	MM	0.0000	0.0501	0.0280	0.9111
50%	OLS	-0.0001	0.0430	0.0544	0.7913
	MM	-0.0012	0.0513	0.0391	0.8871
Multiple Imputation, n = 400, MCAR,$\varepsilon = 0.25$					
λ	**Method**	**$\bar{\text{B}}$ias**	**$\bar{\text{MSE}}$**	**$\bar{\text{SE}}$**	**CR**
TRUE	OLS	0.0000	0.0329	0.0355	0.9500
	MM	0.0000	0.0429	0.0266	0.9500
10%	OLS	0.0000	0.0331	0.0392	0.8463
	MM	0.0000	0.0430	0.0269	0.9469
20%	OLS	0.0000	0.0332	0.0420	0.8232
	MM	0.0000	0.0441	0.0291	0.9309
30%	OLS	0.0000	0.0329	0.0429	0.8491
	MM	-0.0002	0.0439	0.0389	0.9306
50%	OLS	0.0000	0.0341	0.0432	0.8221
	MM	-0.0020	0.0496	0.0495	0.9099

Table 5: Simulation results with data mechanism MCAR and contamination

Single Imputation, n = 100, MAR, $\varepsilon = 0.25$					
λ	**Method**	**Bias**	**MSE**	**SE**	**CR**
TRUE	OLS	0.0000	0.0214	0.0682	0.9500
	MM	0.0000	0.0244	0.0531	0.9500
10%	OLS	-0.0005	0.0345	0.0982	0.8301
	MM	-0.0001	0.0258	0.0614	0.9329
20%	OLS	-0.0015	0.0589	0.0631	0.8299
	MM	-0.0004	0.0249	0.0721	0.9309
30%	OLS	-0.0062	0.0789	0.0878	0.8305
	MM	-0.0006	0.0255	0.0836	0.9202
50%	OLS	-0.0064	0.1145	0.0970	0.8289
	MM	-0.0062	0.0396	0.0942	0.8901
Multiple Imputation, n = 100, MAR, $\varepsilon = 0.25$					
λ	**Method**	**Bias**	**MSE**	**SE**	**CR**
TRUE	OLS	0.0000	0.0214	0.0682	0.9500
	MM	0.0000	0.0244	0.0531	0.9500
10%	OLS	-0.0001	0.0345	0.0875	0.8412
	MM	0.0000	0.0249	0.0540	0.9351
20%	OLS	-0.0005	0.0478	0.0899	0.8313
	MM	0.0000	0.0256	0.0542	0.9310
30%	OLS	-0.0009	0.0701	0.1001	0.8221
	MM	-0.0003	0.0291	0.0550	0.9208
50%	OLS	-0.0013	0.0826	0.1104	0.8201
	MM	-0.0009	0.0301	0.0634	0.9005
Single Imputation, n = 400, MAR, $\varepsilon = 0.25$					
λ	**Method**	**Bias**	**MSE**	**SE**	**CR**
TRUE	OLS	0.0000	0.0337	0.0355	0.9500
	MM	0.0000	0.0429	0.0266	0.9500
10%	OLS	-0.0010	0.0401	0.0581	0.8693
	MM	-0.0009	0.0431	0.0259	0.9398
20%	OLS	-0.0019	0.0419	0.0559	0.8676
	MM	-0.0001	0.0434	0.0275	0.9292
30%	OLS	-0.0035	0.0379	0.0591	0.8751
	MM	-0.0002	0.0439	0.0324	0.8909
50%	OLS	-0.0008	0.0539	0.0712	0.0879
	MM	-0.0038	0.0521	0.0352	0.8221
Multiple Imputation, n = 400, MAR, $\varepsilon = 0.25$					
λ	**Method**	**Bias**	**MSE**	**SE**	**CR**
TRUE	OLS	0.0000	0.0329	0.0355	0.9500
	MM	0.0000	0.0429	0.0266	0.9500
10%	OLS	-0.0003	0.0420	0.0421	0.8391
	MM	-0.0001	0.0428	0.0272	0.9493
20%	OLS	-0.0004	0.0464	0.0439	0.8298
	MM	-0.0003	0.0435	0.0273	0.9389
30%	OLS	-0.0019	0.0471	0.0432	0.8210
	MM	-0.0008	0.0431	0.0271	0.9212
50%	OLS	-0.0078	0.0489	0.0420	0.8109
	MM	-0.0046	0.0442	0.0321	0.9129

Table 6: Simulation results with data mechanism MAR and contamination

Single Imputation, n = 100, MNAR, $\varepsilon = 0.25$					
λ	**Method**	**$\bar{\text{B}}$ias**	**$\bar{\text{MSE}}$**	**$\bar{\text{SE}}$**	**CR**
TRUE	OLS	0.0000	0.0214	0.0682	0.9500
	MM	0.0000	0.0244	0.0531	0.9500
10%	OLS	-0.0009	0.1313	0.0826	0.7901
	MM	-0.0008	0.1132	0.0738	0.8300
20%	OLS	-0.0128	0.1401	0.0832	0.7800
	MM	-0.0129	0.1187	0.0858	0.7760
30%	OLS	-0.0199	0.1405	0.0826	0.7607
	MM	-0.0197	0.1102	0.0793	0.7615
50%	OLS	-0.0209	0.1364	0.0829	0.7569
	MM	-0.0214	0.1679	0.1280	0.7810
Multiple Imputation, n = 100, MNAR, $\varepsilon = 0.25$					
λ	**Method**	**$\bar{\text{B}}$ias**	**$\bar{\text{MSE}}$**	**$\bar{\text{SE}}$**	**CR**
TRUE	OLS	0.0000	0.0214	0.0682	0.9500
	MM	0.0000	0.0244	0.0531	0.9500
10%	OLS	-0.0006	0.1317	0.0832	0.7981
	MM	-0.0007	0.0902	0.0746	0.7999
20%	OLS	-0.0120	0.1394	0.0839	0.7970
	MM	-0.0117	0.0982	0.0840	0.7969
30%	OLS	-0.0198	0.1304	0.0828	0.7891
	MM	-0.0187	0.1198	0.1125	0.7970
50%	OLS	-0.0202	0.0806	0.0866	0.7618
	MM	-0.0205	0.1125	0.1114	0.7809
Single Imputation, n = 400, MNAR, $\varepsilon = 0.25$					
λ	**Method**	**$\bar{\text{B}}$ias**	**$\bar{\text{MSE}}$**	**$\bar{\text{SE}}$**	**CR**
TRUE	OLS	0.0000	0.0214	0.0682	0.9500
	MM	0.0000	0.0244	0.0531	0.9500
10%	OLS	-0.0012	0.0989	0.0921	0.8109
	MM	-0.0008	0.0876	0.0781	0.8456
20%	OLS	-0.0094	0.1290	0.0930	0.8105
	MM	-0.0063	0.0919	0.0799	0.8239
30%	OLS	-0.0110	0.1293	0.0943	0.7991
	MM	-0.0099	0.1102	0.0821	0.8189
50%	OLS	-0.0189	0.1299	0.0975	0.7981
	MM	-0.0162	0.1379	0.0989	0.8091
Multiple Imputation, n = 400, MNAR, $\varepsilon = 0.25$					
λ	**Method**	**$\bar{\text{B}}$ias**	**$\bar{\text{MSE}}$**	**$\bar{\text{SE}}$**	**CR**
TRUE	OLS	0.0000	0.0214	0.0682	0.9500
	MM	0.0000	0.0244	0.0531	0.9500
10%	OLS	-0.0007	0.1134	0.0867	0.8239
	MM	-0.0004	0.0786	0.0775	0.8690
20%	OLS	-0.0089	0.1197	0.0890	0.8198
	MM	-0.0078	0.0975	0.0796	0.8539
30%	OLS	-0.0118	0.0902	0.0920	0.8103
	MM	-0.0097	0.0976	0.0891	0.8458
50%	OLS	-0.0134	0.0998	0.0998	0.8097
	MM	-0.0120	0.0956	0.1103	0.8379

Table 7: Simulation results with data mechanism MNAR and contamination

n = 100, ε = 0.25			
λ	**MCAR**	**MAR**	**MNAR**
10%		-0.0012	0.0989
20%		-0.0094	0.1290
30%		-0.0110	0.1293
50%		-0.0110	0.1293
n = 400, ε = 0.25			
λ	**MCAR**	**MAR**	**MNAR**
10%		-0.0012	0.0989
20%		-0.0094	0.1290
30%		-0.0110	0.1293
50%		-0.0110	0.1293

Table 8