UPPSALA
UNIVERSITET

Methods for handling missing values

A simulation study comparing imputation methods for missing values on a Poisson distributed explanatory variable

By: Fanny Bengtsson and Klara Lindblad

Department of Statistics
Uppsala University

Supervisor: Philip Fowler

HT 2020

Abstract

Incomplete data that contains missing values is a common problem. This study examines how different methods, that handle missing values, affect a coefficient of a linear regression model from a dataset that contains both normal and Poisson distributed explanatory variables. By simulating data and generating missing values of different proportions for the two missingness mechanisms Missing Completely at Random and Missing at Random, the methods are examined under various circumstances. The methods used in the study are Listwise Deletion, Predictive Mean Matching and Poisson imputation. The methods are evaluated by comparing raw bias, coverage rate and average width for the generated coefficient estimates. The results show that for the contexts of this study, Listwise Deletion and Predictive Mean Matching performs best.

Keywords: Simulation study, Multiple imputation methods, Listwise Deletion, Missing Completely at random, Missing at Random, Linear regression

Table of Content

1 Introduction

Decision making based on information from data is highly dependent on the truthfulness of the data. For analysis of data being as accurate as possible, it follows that the data needs to be as accurate as possible. Accurate data imply that the data is complete since incomplete data increase the risk of weakening the validity. However, in the real world, data tends to be incomplete. In many cases, the incompleteness is due to the challenging problem of missing values. A missing value occurs when an observation does not have a collected value for a variable. With missing values, information about the population is missing which risks having data that does not reflect the population truthfully. This can have an effect on the conclusions drawn from the data. The largest concern regarding missing values is to what extent the missing information influences the result of a study (Allison 2002).

The amount of missingness provides a clue to what extent the missing values affect the results, as it is related to its impact on research conclusions. Generally, larger proportions of missing values tend to have a greater impact on statistical inference and generalizability since it indicates that more information about the population is missing. The sample data might reflect a bias as a lot of observed data gets deleted due to a lot of observations obtaining missing values, leading to biased parameter estimates and misleading statistical inference. Further, the cause of why the data is missing can have an effect on the legitimacy of the conclusions. Missing values can occur in different stages and from different sources such as participant recruitment, data collection, data entry, data analysis etc. Moreover, the pattern, amount and mechanism of missing values have important effects on the result of a study. Altogether, it is making the consequences of missing values seemingly countless (McKnight et al., 2007).

The different mechanisms of how missing values are missing are “Missing Completely at Random”, “Missing at Random” and “Missing Not at Random”. These reveal the pattern to which the missingness follows and disclose if there might be a specific reason to why missing values occur (Allison 2002).

The traditional method for managing missing values is Listwise Deletion which infers deleting an observation that does not have all values stored for all variables. However, since

data analysis gives the most valid result when the research is based on a lot of data, deleting observations will decrease the statistical power of the result (Scheff 2016). Therefore, other methods for managing missing values that do not imply deleting observations has been introduced. These methods involve imputing values where missing values occur. Imputation means that missing data, gets replaced with substituted values (van Buuren 2018).

According to van Buuren (2018) multiple imputations methods in general deal with incomplete data better than simple imputation methods. Simple imputation implies that the missing value is imputed one time, in contrast to multiple imputation that implies that the missing value is imputed several times and then pooled together to one value (Little, Ruben 2019). Several imputation methods produce imputations drawn from a normal distribution, even though data in practice many times are skewed, long-tailed, non-negative and rounded, to name some deviations from normality. There are different approaches to handle other types of data. One approach that historically has been used frequently is transforming the data into normally distributed data before the imputation. This approached have been noticed that it should be handle with caution, since it might not provide very good results (von Hippel 2013). To avoid that problem, there are methods that are more suitable for other types of data. For count data, which is data where the minimum number is zero and the values are integer, Predictive Mean Matching, Ordered Categorical imputation, Poisson imputation and Negative Binominal imputation are methods that can be used to impute missing values. Predictive Mean Matching is a well-known and used hot deck method, which means that it imputes missing values, by using values from the complete data matched with respect to some metric. Ordered Categorical imputation is a method that imputes missing data when the missing data are categorical, with help from a general linear model (van Buuren 2018). Poisson imputation and Negative Binominal imputation are two methods that can be used when the missing data are Poisson distributed respectively negative binominal distributed (Kleinke, Reinecke 2019).

1.1 Goals and objectives

Missingness is, as previously mentioned, a problem that occurs everywhere and has a great effect on statistical power and size. The purpose of this study is to examine which methods that best handle missing values on a Poisson distributed explanatory variable in a dataset with where the other explanatory variables are normally distributed. This research will examine how the different methods, Listwise Deletion, Predictive Mean Matching and Poisson imputation, perform in various situations, with different proportion of missing values. The data is simulated and contains missing values of both mechanisms Missing Completely at Random and Missing at Random.

When Poisson regression has been used in previous research, the missing data have occurred in the dependent variable, and the explanatory variables have been normally distributed (Kleinke, Reinecke 2019). In this study, one of the explanatory variables are Poisson distributed and the rest of the variables are normally distributed, and the goal is to see if Poisson imputation provides better imputation than Predictive Mean Matching and Listwise Deletion.

Predictive Mean Matching has been used for Poisson distributed data with missingness before, and it generates acceptable results when missingness and skewness are not too high (Kleinke 2017). It is therefore relevant and interesting to use this method and examine if it works better than the other two. The goal with this study is to provide an increased understanding of how multiple imputation works on a variable that is not normally distributed. The aim is to examine and answer the following questions:

Which methods of multiple imputation best impute missing values of a simulated dataset with missing data on a Poisson distributed explanatory variable?

In comparison to Listwise Deletion, how do the methods of multiple imputation perform on a simulated dataset with missing data on a Poisson distributed explanatory variable?

1.2 Limitations

The time for this study is restricted and therefore limitations have been made. The study is based on a dataset containing four variables. This is a limited number of variables compared to other studies. However, this study will examine how well the methods impute missing values from a Poisson distributed explanatory variable, in a linear regression model. Using a limited number of variables will save time and reduce complexity in both calculations and simulations.

The missing values are limited to only one variable. A dataset with only one variable with missing values are rare in reality, but since the purpose is to examine how the methods impute values for a Poisson distributed explanatory variable, this limitation is made to save time and reduce complexity in the research. Further, a limitation involving the type of situation that is being examined has been done by choosing different proportions of missingness. The proportions of missingness has been set to 10%, 30% and 50%. Since Missing Completely at Random rarely occurs, this study observes imputation on data that is both Missing Completely at Random and Missing at Random.

Previous research has shown that multiple imputation performs better and more accurate results, than simple imputation. This because creating multiple imputations, reflect the uncertainty of the missing value (van Buuren 2018). Therefore, the study has been limited to only use multiple imputation methods and Listwise deletion. As time is restricted, two different multiple imputation methods are used in this study.

2 Theory

2.1 Missing data mechanism

Missing data is a problem that occurs in almost every real dataset. The definition of missing values is that some information about variables is missing. In general, the problem with missing values is that the analyses cannot be made correct based on the data and the conclusions drawn from a dataset with missing values might not be truthful (Little and Rubin 2019).

There are many reasons to why missing values occur. In survey research, nonresponse occurs when a respondent does not complete the whole survey. The reason could be that the respondent did not see the question, or the person did not know how to answer the question. For example, the topic could have been perceived as upsetting, or it could be that the person intended to come back to the skipped question but forgot etc. This is not only a problem in surveys, it also applies in other types of researches (Graham 2012).

A great understanding of how the missing values occur and which pattern they have is important in order to be able to handle them. Missingness patterns describes the missing values and observed values in the data matrix and Missing data mechanisms describes the relationship between missingness and the values of variables in the data matrix. Examining these gives a greater understanding of how to handle the missing values (Little and Rubin 2019).

There are different reasons to why missing values occur and can therefore be divided into three different mechanisms, Missing Completely at Random, Missing at Random and Missing Not at Random (Little and Rubin 2019). There are two assumptions that needs to be fulfilled for Missing Completely at Random. The first assumption is that there are no systematic differences in the observed variables between the variables with missing values. The second assumption that needs to be fulfilled is that there cannot be any relationship between the missing values on a specific variable and the values on that variable. When those assumptions are fulfilled the data is said to be Missing Completely at Random (Allison, 2002). One example of this mechanism is if a weighing scale run out of batteries. Some data will be

Missing Completely at Random since it is just a coincidence that the batteries run out and thus the missingness is not related to any of the variables (van Buuren 2018).

Missing at Random contains weaker assumptions than Missing Completely at Random. The definition for Missing at Random is that the missingness may, in comparisons to Missing Completely at Random, depend on the observations but it cannot depend on the missingness in a dataset. A classic example for Missing at Random is reading speed. If a long, self-administered study, has a limited amount of time for completion, only fast readers will complete the survey. However, reading speed can be measured early in the survey, where almost all respondents will provide data. By including the reading speed variable as in the missing data analysis model, biases with reading speed can be controlled (Graham, 2012).

Missing Not at Random means that the missingness depends on the missing values. Moreover, Missing Not at Random occurs when missingness is caused by other missing values or by some other related variables. The problem is that it is hard to measure how much the missingness depends on the other variables. This type of missingness is defined as inaccessible, since the cause of missingness has not been measured and can therefore not be analysed. An example for this kind of missingness can be the measure of income. In survey research, it is common for people with higher incomes to leave the income question blank (Graham, 2012).

2.2 Listwise Deletion

Listwise Deletion is a method that handles missing values. The method implies removing all the observations with missing values on any variable. Hence, the observations that contain all its values are the ones included in the analysis. The two advantages of Listwise Deletion are that it can be used for any kind of statistical analysis and no special computer methods are required (Allison, 2002). If the data is Missing Completely at Random the Listwise Deletion method has attractive statistical advantages. Since the missingness is completely random, the observations that will be deleted are randomly spread on the dataset and the method will therefore generate unbiased estimators and correct standard errors (Allison 2002).

Most statistical packages use Listwise Deletion by default since it is easy to use and calculate. The disadvantage with this method is that if the number of missing values is too large

compared to the sample size, even for a dataset that contains data Missing Completely at Random, the impact on the result can decrease the precision and statistical power (Myrtveit, Stensrud & Olsson 2001). Even though Listwise Deletion has these disadvantages it is not a bad method for handling missing values. Most of the time it gives valid inferences for data that is Missing Completely at Random, even if it does not use all of the available information (Allison, 2002).

2.3 Multiple Imputation by Chained Equations

Multiple Imputation by Chained Equations (MICE) is an R package addressing missing values in multiple imputation (van Buuren, Broothuis-oudshoorn 2011). The idea with multiple imputation is to generate and replace the missing value multiple times. That means that more than one dataset is created from the imputations. The different datasets are pooled together before the different methods are compared (van Buuren 2018). The chained equation approach is very flexible and can handle variables of varying types and complexity. The assumption that has to be fulfilled for MICE is that the given datasets missing values are Missing Completely at Random or Missing at Random (Azur, Stuart, Frangakis, and Leaf 2011).

MICE can be divided into six steps. The first step is a simple imputation for every missing value in the dataset. All missing values obtains a temporary value that can be, for instance, a mean or median. These imputations can be thought of as “place holders”. The second step is to remove the place holder imputation for one observation, so it gets a missing value again. Step three is making a regression model from the observed and temporary values that can generate a value that replaces the missing value. The regression models run under the assumptions of a regression model. The fourth step is then replacing the missing value with the prediction from the regression model. The fifth step is to repeat step 2-4 for every missing value. One iteration is complete when all of the missing values are predicted and imputed. In step six the steps 2-5 are repeated for *m* number of cycles, which generates *m* number of datasets. The different datasets with imputed values are then pooled together (Azur, Stuart, Frangakis, and Leaf 2011). Different methods, also called algorithms, predict the missing values in different ways. The algorithms that will be compared are the Predictive Mean Matching and Poisson imputation. In Figure 2.3 the flow chart for MICE is visualized, starting with the incomplete dataset with missing values, followed by the different imputation

steps and several imputed datasets occurs. Figure 2.3 contains five different datasets, this is a number often chosen by default. The different datasets are analyzed before they are pooled together to one dataset.

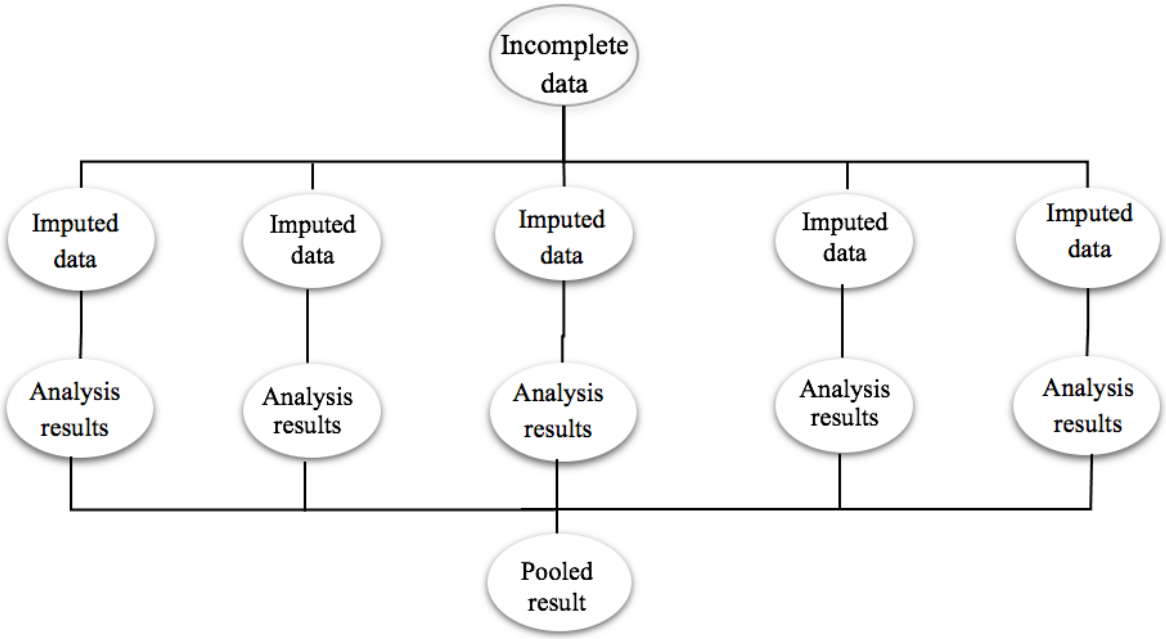

Figure 2.3. *Flow chart for MICE*

2.3.1 Predictive Mean Matching

Predictive Mean Matching calculates the predicted value for the variable with missing values, it is called a Hot Deck method. The missing values are imputed with help from the observations with no missingness, relying on a standard linear regression. The general idea with Predictive Mean Matching is that a linear regression model is estimated. For every missing value, the methods set up a small set of donors with no missingness. These donors come from the complete observations that possesses similar values on other variables.
One of the donors are randomly chosen, and the observed value of that donor is replacing the missing value. The assumption for Predictive Mean Matching is that the distribution of missing values is the same as for the observed data of the donors. It is an easy method to use and the imputed variables are realistic since they are based on observed data. The method will not give values outside of the observed data range (van Buuren, 2018). Predictive Mean Matching is a method that can be recommended for many different scenarios and can handle different types of data (Kleinke 2017).

2.3.2 Poisson imputation

The Poisson imputation method imputes missing values with Poisson regression that follows the regular assumptions of Poisson regression. Hence, this method works well on missing data from variables that are Poisson distributed. The general linear regression model is created from the observed and temporary values, which generates a value to replace the missing value with. The imputations are drawn from a Poisson distribution with mean (λ) that is predicted for each of the missing values. The Poisson imputation method follows the steps described for MICE in Section 2.3 (Kleinke, Reinecke 2019).

2.4 Evaluation measures

When imputing missing values, the intention is to obtain better estimates for the variable from the incomplete data. There are a lot of different measures that can be used when evaluating the results. For this study a linear regression model is created and evaluated. Following van Buuren (2018), the evaluation of the various methods is done by using three measures. These measures are raw bias, coverage rate and average width of the confidence interval. Van Buuren (2018) mentions another measure as well, root mean squared error, but clarifies that it

is not suitable to use when evaluating multiple imputation methods, therefore it is not used in this study.

The first measure, raw bias, measures the difference between the expected value of the estimate and the actual value for the estimated parameter. The raw bias should be close to zero (van Buuren 2018).

$$Raw\ Bias = E(\hat{\beta}) - \beta$$

The second evaluation measure is the coverage rate, that gives the proportion of confidence intervals that hold the true parameter value. The coverage rate gets affected by two things, the estimate and the confidence interval. The actual rate should be greater than, or equal to the nominal rate. If the coverage rate is smaller than the nominal rate, the method is too optimistic. If the nominal rate is 95% and the coverage rate is 90%, it is indicating that the methods give estimates of poor quality. Although a high coverage rate indicates a good result, a too high coverage rate, e.g., 99%, might also indicate that the confidence interval is too wide which leads to results showing an inefficient method (van Buuren 2018).

The third and last measure used is the average width. It gives the average width for the confidence interval and indicates how efficient the estimator is. The average width should be as low as possible without obtaining the coverage rate lower than the nominal rate. The average width also indicates how well the standard deviations are estimated for the coefficients (van Buuren 2018).

The best method is the one that has a raw bias closest to zero and a coverage rate near the nominal rate. The methods that fulfil those assumptions are called randomization valid. If more than one method is randomization valid, the method with a shorter confidence interval is more efficient (van Buuren 2018).

3 Methodology

Simulating data, from which missing values are generated, is a key procedure in the evaluation of the performance of missing data techniques. The amputation procedure, i.e., generating the missing values, is a critical part of the evaluation of the missing data methods as it defines the missing data problem. In simulation studies that are evaluating imputation methodology like these, there are generally four steps that the study follows (Schouten, Lugtig and Vink, 2017). The structure of this study is built from these four steps and is visualized in Figure 3.1. In order to obtain valid imputations, the simulation needs to be replicated many times. In this study, all simulations were replicated 1000 times. The study is implemented in R and the packages used are "mice" (van Buuren, Broothuis-oudshoorn 2011) and "countimp" (Kleinke 2020).

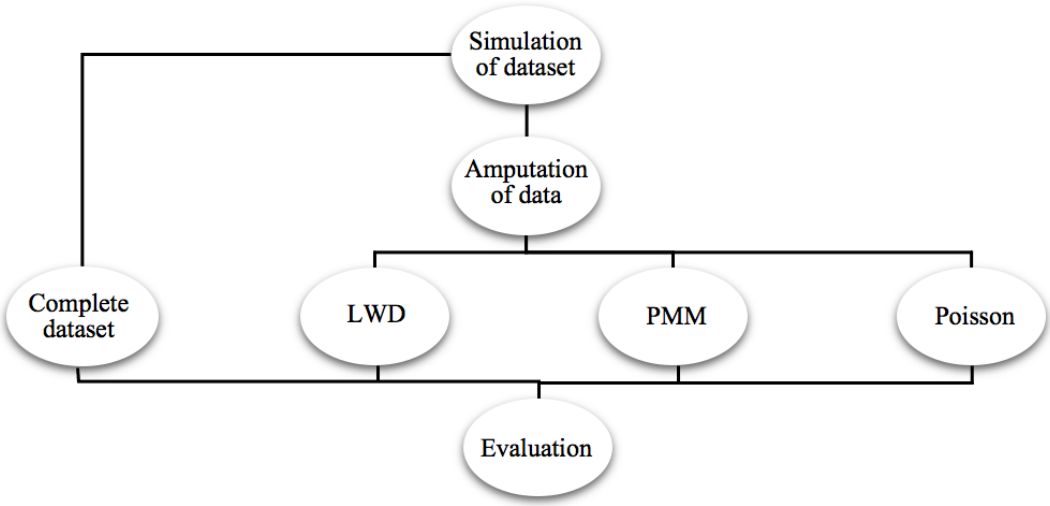

Figure 3.1 *Flow chart for the simulation study*

3.1 Data simulation

In a first step of simulating the data, a complete dataset, that can represent a sample of the distribution, is made. In this study, random variables were generated and considered to represent a population. Two explanatory variables, X_1, with $\mu = 10$ and $\sigma = 1$, and X_2, with $\mu = 5$ and $\sigma = 1$, were generated from a normal distribution and the explanatory variable X_3, with $\lambda = 4$, was generated from a Poisson distribution. Together, the variables created the dataset from which a linear regression model was generated. The dependent variable, Y_1, was created from a normal distribution based on the three explanatory variables in the regression. The dataset contained 1000 observations and the variables, Y_1, X_1, X_2, and X_3, were all numeric. The explanatory variables were generated as independent variables. They were all significantly correlated with Y_1. The model could constitute as an indicator when comparing the different outputs of imputation methods and how they affected the coefficient estimate. The linear regression model can be written as follows

$$Y_i = \beta_1 X_i + \beta_2 X_i + \beta_3 X_i + \varepsilon_i$$

where $\beta_1 = 1,\ \beta_2 = 1, \beta_3 = 1$ and ε_i is a normally distributed error term with $E(\varepsilon_i) = 0$ and $\sigma = 1$ and where the error terms are iid.

3.2 Amputation

The second step involved generating the missing values and by doing so, making the dataset incomplete. For this study, the function "Amputation" in the package "mice" was used to create missing values in the dataset (Van Buuren, Groothuis-Oudshoorn 2011). Simulating a dataset with reliable missing data problems is difficult. The procedure of amputating data defines the missing data problem; hence, it is critical for the evaluation of the imputation methods.

The methods used in this study work best under the assumption that the missing mechanism is either Missing Completely at Random or Missing at Random, and therefore, those mechanisms were used in this study. In the procedure of amputation, the pattern of the missing data was first chosen. A missing data pattern is a combination of variables of missing values and values that remain complete. In this case that was only one pattern and indicated that missing values would only occur on variable X_3. Thereafter, the missingness mechanism

was chosen. For when data was Missing Completely at Random the next step involved the proportion of missingness but for when data was Missing at Random it involved setting weighted sum scores. Weighted sum scores are outcomes of a linear regression equation that indicates which variables the missingness depends on. The coefficients are determined by the user and for this study, the weights were set to "default". For the case of data Missing Completely at Random, weights are not used since it would infer the missingness not being completely random. The next step implicated the distribution of probabilities which are fixed and determined by the proportion of missingness for data Missing Completely at Random. For data Missing at Random, a candidate value obtained a probability of being missed based on its weighted sum score. Thereafter, the candidates were divided into two groups based on the proportion of missingness where one group received missingness pattern and the other group remained complete (Schouten, Lugtig and Vink, 2017). Further, an evaluation of how well the methods handle different proportions of missing values was included in the study. Thereby, three different sizes of the proportion were used. The proportions were 10% missing values, 30% missing values and 50% missing values.

3.3 Imputation

In a third step, the different methods were applied to the dataset with the generated missing values. The methods that use multiple imputation generated 5 imputed datasets each that individually were generated by 100 iterations. Linear regression models were generated from the datasets and combined together in a "pool". The imputed datasets, m, was set to 5 since that is a "default" for methods like the ones used in this study. Having a higher number of imputations is theoretically better but requires a lot of time which was limited for this study. It is convenient to set $m = 5$ during the stage of model building and raising the amount in the evaluation stage if it is needed (van Buren, 2018). Further, the pooled linear regression models generated coefficient estimates of β_3.

3.4 Evaluation

The last and fourth step in the study, following the implementation of the methods, is evaluation and comparison of inferences of estimates from the complete dataset and the generated datasets from the methods. The comparison generates an indication of the

performance of the different methods (Schouten, Lugtig and Vink, 2017). The new coefficient estimates of β_3 generated from the imputations were compared to the true coefficient value $\beta_3 = 1$. For evaluation, measures are done by comparing raw bias, coverage rate, and average width. These measures are compared for the three proportions of missing values in combination with the mechanisms of Missing Completely at Random and Missing at Random.

4 Results and Discussion

The results of this study are presented below with two tables that present the results for each missingness mechanisms. For each table, the estimates of β_3, raw bias, average width, and the coverage rate are included. For every method, there are three measures given in combination with every proportion of missingness. The proportion of missingness is given in column *p*. After that, the confidence intervals for the estimated coefficients of β_3 are shown. For every situation of missingness mechanism with a proportion of missingness 1000 estimates with confidence intervals were generated from the 1000 iterations. An average of these are the results presented in the tables below.

In Table 4.2 the results for data Missing Completely at Random is presented. The table shows that a different percentage of missingness gave similar results for Listwise Deletion and Predictive Mean Matching. When the proportion of missingness were higher, Listwise Deletion generated slightly better estimates based on the Raw bias compared to Predictive Mean Matching. The Poisson imputation method did not generate valid estimates when the data was Missing Completely at Random, as they were all below the true value.

Table 4.2 Results for data Missing Completely at Random

Method	**p**	**Estimate**	**Raw Bias**	**Coverage Rate**	**Average Width**
Listwise Deletion	0.1	1.001	0.001	0.949	0.066
	0.3	1.001	0.001	0.949	0.074
	0.5	1.000	0.000	0.950	0.088
Predictive Mean Matching	0.1	1.001	0.001	0.940	0.065
	0.3	1.002	0.002	0.928	0.072
	0.5	1.005	0.005	0.929	0.081
Poisson imputation	0.1	0.945	-0.055	0.334	0.094
	0.3	0.835	-0.165	0.000	0.126
	0.5	0.743	-0.257	0.000	0.133

The coverage rate indicates that Listwise Deletion performed better than the two imputation methods which can be seen in Table 4.2. The coverage rate for Listwise Deletion is almost 95% for all the different proportions of missingness. Predictive Mean Matching also gave acceptable results for the coverage rate, since it is not below 90%. The Poisson imputation method did not provide any good coverage rate. When the proportions of missingness were high, the true value was not covered at all, indicating a very bad result.

Shown in Table 4.2, the average width gives an indicator that the Predictive Mean Matching and Listwise Deletion, are the best methods since they generated the smallest widths for all different proportions of missingness. The Poisson imputation method did not perform as well. It gave a wider average width compared to the other methods, but it was not completely off.

The results for data Missing at Random are given in Table 4.3. For this missingness mechanism, Predictive Mean Matching generated the best bias results when the missingness was both 10% and 30%. When the missingness was 50%, Predictive Mean Matching and Listwise Deletion performed equally. Further, for this missingness mechanism as well, the Poisson imputation method performed poorly as it generates quite high bias.

The Predictive Mean Matching generated best coverage rate when the missingness was 10% and 30%. When the proportion of missingness was larger, Listwise Deletion actually provides a better coverage rate. Poisson imputation generated unacceptable coverage rate, it got worse when the missingness proportion increased.

For all the different proportions of missingness, Predictive Mean Matching, followed by Listwise Deletion, performed best when it comes to average width. Poisson imputation generated wider intervals than the other methods, but they are acceptable.

Table 4.3 *Results for data Missing at Random*

Method	p	Estimate	Raw Bias	Coverage Rate	Average Width
Listwise Deletion	0.1	0.997	-0.003	0.928	0.066
	0.3	0.994	-0.006	0.931	0.077
	0.5	0.992	-0.008	0.939	0.093
Predictive Mean Matching	0.1	1.001	0.001	0.950	0.066
	0.3	1.005	0.005	0.933	0.075
	0.5	1.008	0.008	0.907	0.086
Poisson imputation	0.1	0.930	-0.070	0.241	0.111
	0.3	0.810	-0.190	0.000	0.141
	0.5	0.719	-0.281	0.000	0.143

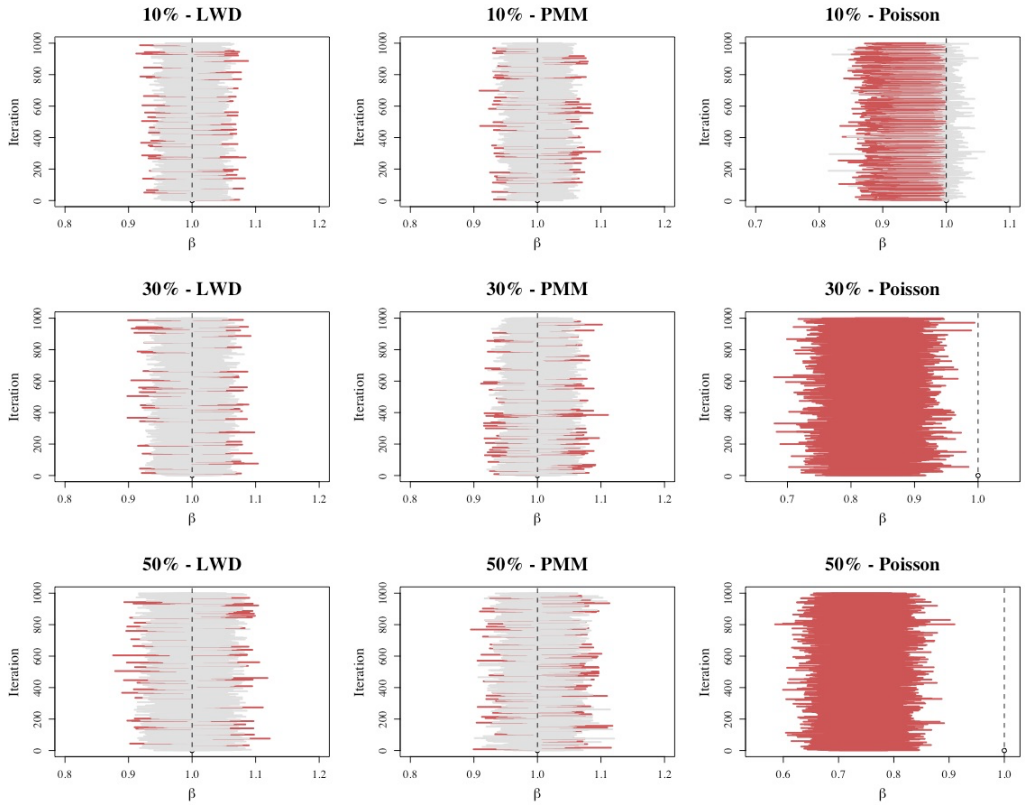

Figure 4.4 *Intervals for estimated parameter of data Missing Completely at Random*

In Figure 4.4, the confidence intervals for the estimated coefficients of β_3 for data Missing Completely at Random together with the true value of β_3 are visualized. The grey colour of the interval indicates that the true value is covered within the interval. When the confidence interval is red it indicates that the true value is not covered in the interval. Most of the time for Listwise Deletion and Predictive Mean Matching the true value is covered in the interval, which is an expected result. For Poisson imputation, the coverage of the true value is very low when the missingness proportion is 10%. As the proportion of missingness increases, the true value is not covered at all. This can be explained by the estimated coefficient of β_3 by Poisson imputation that was below the true value.

In Figure 4.5, where the confidence interval for the estimated coefficient of β_3 for data Missing at Random together with the true value of β_3 is shown, the results are quite similar to the results from when data are Missing Completely at Random. Both Listwise Deletion and Predictive Mean Matching gave good results. Poisson imputation did not provide very good results. As well as for data Missing Completely at Random, the estimated parameter is below the true value and is therefore not covered in the interval.

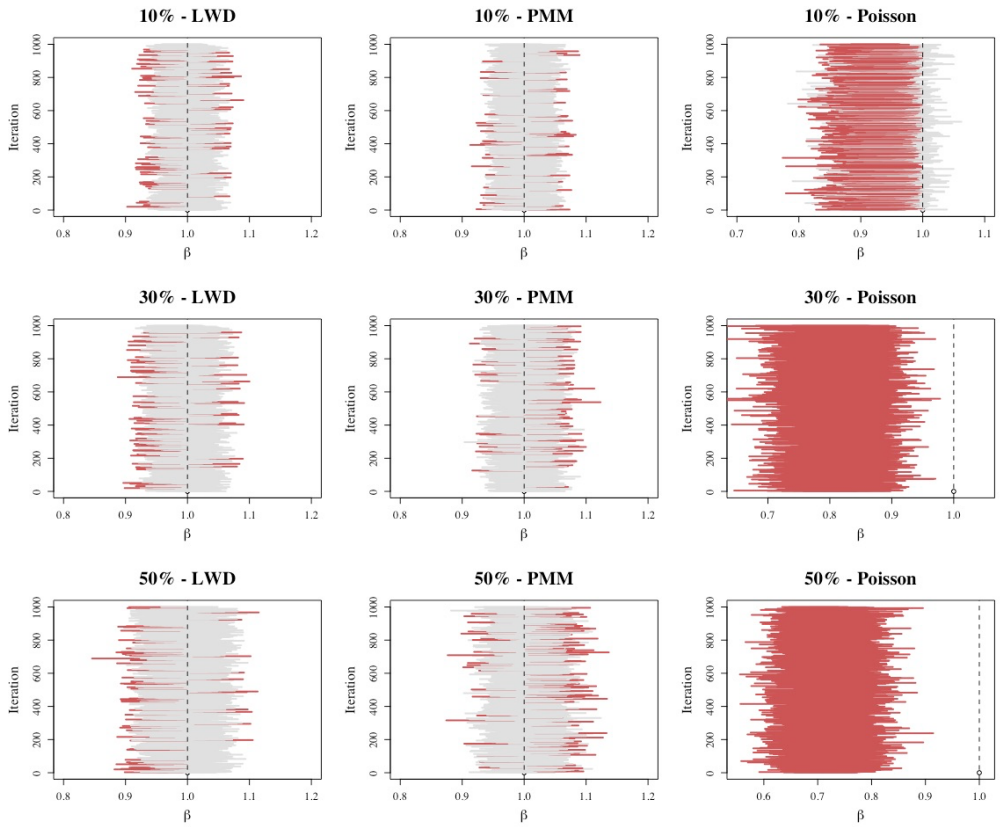

Figure 4.5 *Intervals for estimated parameter of data Missing at Random*

The Poisson imputation showed poor results with estimates that were below the true value for both missing mechanisms. When the coverage rate is 0, the probability of committing type 1 error is 100% for all shares of missingness. The average width is a little wider than for the other methods, but it is not completely off. That indicates that the estimate of β_3 is not very good, rather than that the standard deviations are bad. It is shown in both Table 4.1 and Table 4.2 that the estimate of β_3 is below the true value.

Listwise Deletion performed best when the data was Missing Completely at Random and when the proportion of missingness was not too high. When the data was Missing at Random the raw bias for Listwise Deletion increased as the proportion of missingness increased, however the increase was not as high as expected. When looking at the choices made for how the missing values were created in the situation of data Missing at Random it can be discussed whether the weights within the amputation function should have been set to "default" or not. When the weights were set to "default" all variables had the same influence on the missingness of variable X_3 which ignored the fact that there was higher correlation between X_3 and Y than there was between X_3 and the remaining variables. This might have led to a faint Missing at Random pattern which led to similar results as for when data was Missing Completely at Random.

Therefore, a small study on the same dataset but with adjusted weights for the amputation of data Missing at Random was implemented with 100 iterations to examine whether this theory might be correct. When the function "ampute" was used for when data was Missing at Random, the weights was not set to "default" but adjusted to the correlation between the variables. For this dataset, variable Y was given a higher weight than the other variables since the correlation obtained between variable Y and variable X_3 was higher. It shows that applying weights that take into account the correlation that exists between the variables generates missing values with a pattern that more clearly shows the characteristics for data Missing at Random. Given these findings, the results in the main study regarding the low bias of Listwise Deletion for when data is Missing at Random can be explained. Further, the small study was implemented for the Predictive Mean Matching and Poisson imputation as well. The results generated for Predictive Mean Matching was very similar to the results from the main study but for Poisson imputation they were slightly worse.

Further, in the main study, Poisson imputation gave estimates that were below the true value for both missingness mechanisms, which led to very poor coverage rate and high bias. These results imply that the Poisson imputation method does not work very well on the given dataset. In order to get a greater understanding of why the results were as they were, another small study with 100 iterations was implemented. This small study examined what effect the characteristics of the dataset had on the results for Poisson imputation. In contrast to the main study, where the explanatory variables were created independently, the dataset in the small study with two normally distributed explanatory variables and one Poisson distributed explanatory variable was generated dependent on each other. The explanatory variables obtained the same beta coefficients as in the main study. The package that was used for the generation of the dataset was "PoisNor" (Amatya, Demirtas and Gao 2020) whose data generation mechanism is a connection between Poisson and normal correlations that are pre-specified. The correlations between the explanatory variables were set to 0.2, 0.15, and 0.25. With regression, a dependent variable was generated based on the three explanatory variables. Beyond these stated changes, the small study was implemented on the same basis as the main study. The small study showed that when the explanatory variables were created dependent on each other the Poisson imputation method worked better. The results generated from the small study indicated low bias and high coverage rate.

5 Conclusions

Choosing an imputation method for a situation of missing values might not always be a simple task. There are limitations for deciding on what is best suitable for a situation. An advantage of Listwise Deletion is that it is a simpler method to use. Predictive Mean Matching and Poisson imputation require both more knowledge and time to implement. However, Listwise Deletion entails other disadvantages as loss of information and power, especially in situations where the proportion of missingness is large or the sample size is small.

This study has compared how three different methods, Listwise Deletion, Predictive Mean Matching and Poisson imputation, handle missing values. The evaluation has been made by using linear regression and comparing the different estimates of a coefficient. The goal with this research was to examine which method that performs best when handling missing values, for when missing values were applied to an explanatory variable that is Poisson distributed and the remaining variables in the dataset are normally distributed.

The conclusions that can be drawn from this study is that when data is Missing Completely at Random, Listwise Deletion is a good method to handle missing values. For all three proportions of missingness, the method generates almost zero or very low bias, the coverage rate is good, and the average width is not too wide. For when the data was Missing at Random, Listwise Deletion presented similar results. However, the small study indicates that the results were influenced by the amputation procedure for data Missing at Random.

Further, imputation with Predictive Mean Matching performs similar for both missingness mechanisms. For when data is Missing Completely at Random, Predictive Mean Matching imputes values that generate acceptable results, even though the bias increases when the proportion of missingness increases. However, it does not perform better than Listwise Deletion. Therefore, when data is Missing Completely at Random it would be better to use Listwise Deletion for the given conditions of this study. When the data is Missing at Random, Predictive Mean Matching performs somewhat better than Listwise Deletion and is therefore considered to be the method that should be used.

Poisson imputation did not perform well in any of the different situations and proportions of missingness. This study shows that the method is not suitable to use for imputation when missing values occur on a Poisson distributed explanatory variable and the remaining explanatory variables are normally distributed. It shows that using an imputation method that uses an assumption of normally distributed data is more efficient in this situation. However, as discussed in the second small study, when the explanatory variables from different distributions were generated dependent on each other Poisson imputation performed better.

This is a simulation study, and it is impossible to cover all different situations that can occur when it comes to different missingness proportions, missing values on more than one variable, different sample sizes etc. This study has been limited to examine how the methods perform on a dataset with missingness applied on one variable, with data Missing Completely at Random and Missing at Random. Further research would be interesting to do on a dataset with missingness on more than one variable, different size on the sample and on a dataset with missingness Not at Random. A sample size of 1000 observations were chosen for this study as a limitation had to be made. However, it would be interesting to study how the methods would perform if the sample size were smaller or larger.

For further research, it would be interesting to examine more thoroughly a situation where the process of generating missing values for data Missing at Random, applies weights for variables that should have a greater influence on the missingness. Moreover, it would be interesting to see how the correlations between explanatory variables with different distributions have an effect on multiple imputation. As a predominant part of the variables in this study are normally distributed, it leaves room for exploring whether the imputation method would generate different results if more explanatory variables were Poisson distributed.

References

Allison, Paul D. 2002. *Missing data*. Thousand Oaks: Sage Publications, Inc.

Amatya, Anup; Demirtas, Hakan and Gao, Ran. 2020. PoisNor. Simultaneous Generation of Multivariate Data with Poisson and Normal Marginals. R package version 1.3.2.

Azur, Melissa; Stuart, Elizabeth; Frangakis, Constantine and Leaf, Philip. 2011. Multiple imputation by chained equations: what is it and how does it work?. *International Journal of Methods in Psychiatric Research.* 20(1): 40-49.

Graham, Johan W. 2012. *Missing Data: Analysis and Design*. New York: Springer science + business media.

Kleinke Kristian. 2017. Multiple imputation under violated distributional assumptions: Asystematic evaluation of the assumed robustness of predictive mean matching. *Journal of educational and behavioral statistics*. 42(4): 371-404.

Kleinke, Kristian; Reinecke, Jost. 2019. *Countimp version 2*. Available at: https://countimp.kkleinke.com/ (18/12/2020).

Kleinke, Kristian. 2020. countimp: Multiple Imputation of incomplete count data. R package version 2.0.7. https://www.kkleinke.com/#software

Little, Roderick & Rubin, Donald. 2019. *Statistical analysis with missing data*. Third edition. New Jersey: John Wiley & Sons Inc.

McKnight, Patrick E; McKnight, Katherine M; Sidani, Souraya; Figueredo, Aurelio José. 2007. *Missing data: A gentle introduction*. New York: The Guilford Publications

Myrtveit, Ingunn; Stensrud, Erik and Olsson, Ulf H. 2001. Analyzing data sets with missing data: an empirical evaluation of imputation methods and likelihood-based methods. *IEEE Transactions on Software Engineering*. 27(11): 999-1013.

R Core Team. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Austria.

Scheff, Stephen W. 2016. *Fundamental Statistical Principles for the Neurobiologist*. Academic Press.

Schouten, Rianne Margaretha; Lugtig, Peter and Vink, Gerko. 2018. Generating missing values for simulation purposes: a multivariate amputation procedure. *Journal of Statistical Computation and Simulation*. 88(15): 2909-2930.

Van Buuren, Stef. 2018. *Flexible Imputation of Missing Data.* Second edition. Chapman & Hall/CRC.

Van Buuren, Stef; Groothuis-Oudshoorn, Karin. 2011. Mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software.* 45(3): 1-55.

Von Hipple, Paul T. 2013. Should a Normal Imputation Model Be Modified to Impute Skewed Variables?. *Sociological Methods & Research.* 42(1): 105–38.