University of Twente

Bachelorthesis

A comparison of multivariate and univariate models for pre-test post-test data concerning accuracy in the presence of missing data

Lukas Joscha Beinhauer
s1746863

20/06/2018

supervised by
Prof. Dr. Ir. Jean-Paul Fox
Prof. Dr. Job van der Palen

Contents

List of Tables

Abstract

Missing data has been posing a common problem to almost all scientific studies over the last decades. This study compares multivariate models to univariate models - the change score method and the regressor variable method - in terms of precision and accuracy over different conditions of missing data. In two simulation studies patterns of missing data were introduced, over degrees of missingness and covariance for pre-test post-test designs. The first study focuses on the MCAR pattern and its influence on the estimates of the different models. The second model makes use of a MAR pattern, where the likelihood of missing values is dependent on an observed variable. The results showed some differences between the models. The change score method appeared to have no particular advantages, the regressor variable method had the highest accuracy, while the multivariate method provided the best precision. Higher amounts of missing data had a negative impact on both accuracy and precision.

1 Introduction

Scholars all over the world from different sciences are using statistical models to evaluate and analyze their data. Making claims regarding the causality of something typically requires data from before and after the event supposedly causing the change, while being able to compare that information to data from conditions not exposed to the event. Adding a control group is necessary for the researcher to study the influence of the variable on a carefully controlled sample. A pre-test post-test design is widely used in different fields to generate data. In more practical fields like medicine or education, it is oftentimes ethically impossible to add a control to an experimental group. It would be questionable to withhold a possible treatment to half of the patients or a better education technique to half a school. Therefore, the design known as a *quasi-experimental design* is often lacking in validity due to its constraints. Due to its enforced popularity, different methods have emerged making the analysis of such a study design possible. The event supposedly causing the change shall be referred to as the *treatment* for the remainder of the paper, its effect accordingly being the *treatment effect*.

1.1 Univariate Methods

Two different univariate methods appear to empower the researcher's evaluation of hypotheses, the *change score method* (CS) and the *regressor variable method* (RV). While controversially discussed, it appears that both methods have their place in the scientific literature (Allison, 1990).

1.1.1 Change Score Method

This method uses the difference in scores between pre- and post-test as the treatment effect. Hereby, the score on the pre-test will serve as baseline, ergo the individual's score before the treatment took place. Accordingly, the relative difference between pre- and post-test will become relevant. The model can be expressed in a regression equation as follows:

$$(Y_{2_{ij}} - Y_{1_{ij}}) = \beta_0 + \delta T_{ij} + \epsilon_{ij}, \epsilon_{ij} \sim N(0, \sigma^2)$$

In this case $Y_{2_{ij}}$ represents person i's score in group j on the post-test, while person i's score from group j on the pre-test is represented by $Y_{1_{ij}}$. $(Y_{2_{ij}} - Y_{1_{ij}})$ accordingly can be considered the range of the difference from pre-test to post-test, or the measurement of change from the baseline for person i in group j. T_{ij} is the assignment of person i to group j, indicating the presence (1) or absence of treatment (0), with ϵ_{ij} being the error that cannot be regulated. This error is assumed to be normally distributed, with mean

0 and variance σ^2.

1.1.2 Regressor Variable Method

This second method uses the scores on the pre-test as a covariate for analysis. This model can be expressed in:

$$Y_{2_{ij}} = \beta_0 + \delta T_{ij} + \beta_1 Y_{1_{ij}} = \epsilon_{ij}, \epsilon_{ij} \sim N(0, \sigma^2)$$

$Y_{2_{ij}}$ hereby represents the score of person i from group j on the post-test, $Y_{1_{ij}}$ on the other hand person i's score from group j on the pre-test. T_{ij} is again considered to be the indiciator of belonging to a treatment or control group for person i of group j, ϵ_{ij} again representing the error that cannot be accounted for.

1.2 Multivariate Method

Lastly, there is the multivariate method (MV) that can be used in scenarios described above.

$$Y_{1_{ij}} = \beta_{01} + \epsilon_{1_{ij}} \tag{1}$$

$$Y_{2_{ij}} = \beta_{02} + \delta T_{ij} + \epsilon_{2_{ij}} \tag{2}$$

$$\begin{pmatrix} \epsilon_{1_{ij}} \\ \epsilon_{2_{ij}} \end{pmatrix} \sim MVN(0, \Sigma), \Sigma = \begin{pmatrix} \sigma^2, \rho \\ \rho, \sigma^2 \end{pmatrix} \tag{3}$$

Considering the MV method, both pre- and post-test scores would be treated as dependent variables. $Y_{1_{ij}}$ and $Y_{2_{ij}}$, similarly to the previously mentioned models, represent the scores of person i from group j on the pre- and the post-test, while T_{ij} indicates the presence or absence of treatment. However, the uncontrolled errors $\epsilon_{1_{ij}}$ and $\epsilon_{2_{ij}}$ are assumed to be normally distributed, centering around 0 with a variance of Σ.

Each of the three methods has advantages and shortcomings, depending on different scenarios. Kleine-Bardenhorst demonstrated in simulation studies, that the methods give partially similar results (2017). The RV method and the MV method appeared to be posing as equally viable methods for data analysis, while the CS method led to less reliable results.

1.3 Missing Data

However, researchers often encounter problems of missing data, resulting in a questionable validity. Rubin (1976) highlighted the different problems, examining multiple possible

sources of missing data and their implications on the studies' results. Missing data remained an issue to researchers from all scientific fields, numerous scholars tried to analyze this issue and find possible solutions throughout the years. In 2002, Schafer and Graham described different patterns of nonresponse and distributions of missing data, they emphasize the fact that current methods are unable to fully resolve the issue.

Missing data is problematic, since many statistical methods presume complete data with no missing values. Few missing values in the participant's observations can quickly lead to a strongly diminished data pool. This can lead to biased results, lower statistical power and less accurate p-values and confidence intervals, highlighting the need for viable methods to handle missing data. The amount of missing data in a data set is also known as its *missingness*.

1.3.1 Methods of Missing Data

However, its important to keep in mind, that differences between kinds of missing data have to be made. There are three different assumptions regarding mechanisms of missing data, originally made by Rubin (1976).

(a) Missing Completely at Random (MCAR)
The missing values of a variable Y will be considered MCAR, if the probability of missing values is unrelated to the variable Y and all other variables from the data set. This means that the value of any of the variables may not be correlated to the likelihood of missing a value of the variable Y. None of the observed or unobserved values of any variable in the model of interest may influence the distribution of missing data (Allison, 2001).

(b) Missing at Random (MAR)
Missing values of a variable Y may be considered to be MAR, if the likelihood of missing these values is only directly related to observed values. Values of Y, X or any other variable may be influential on the probability of data missing, as long as they are known and can be controlled for. MAR is a similar, but stronger assumption of MCAR (Allison, 2001).

(c) Missing Not at Random (MNAR)
If the missing values of variable Y are to be considered MNAR, they have to depend on values that could not have been observed. This means, that the values that are missing, are correlated to their own (unobserved) values. Basically, whenever the conditions for MCAR or MAR are violated, the missing values have to be treated as MNAR. The problematic factor of MNAR is its untestability, one may assume data to be M(C)AR,

but whether the unobserved values play a role can never be inferred from available data - per definition. (Allison, 2001)

1.3.2 Working with Missing Data

Complete Case Analysis

Currently, various methods are being used to work around missing data, traditional and advanced (Soley-Buri, 2013). However, almost all of these methods come down to three key mechanisms. Firstly, there is the complete case analysis (CCA). CCA is an elimination technique. When a researcher chooses to use CCA, he/she eliminates all incomplete cases. This can mean that, from a big participant pool, the data of those will be removed where data on pre-test, post-test or both is missing. However, this technique is highly limited. Should the missing data not be MCAR, the results of the study would likely be strongly biased.

Data Imputation

Secondly, there is the mechanism of data imputation. Methods making use of data imputation try to substitute the missing data by means of reasonable guessing. Different techniques exist, offering the opportunity to enrich the data pool. The researcher can use the mean of the relevant variable-score or compute conditional means using regression. This method is believed to handle non-random missing values better than CCA, while it still inhibits validity. Its quality is strongly dependent of the quality of imputation method. This imputation method cannot be validated for obvious reasons, as the bias between missing and replacement value can never be known. This technique generally leads to an underestimation of standard errors and accordingly an overestimation of test statistics. Since the data is computed using a model applied to the data, it cannot contain any error.

Observed Case Analysis

Lastly, observed case analysis (OCA) can be used. OCA is a technique using neither elimination or imputation of values. Instead, using regression models, all present cases will be used. A detailed description of this process will follow in a later section.

2 Experiment

A quantitative simulation study, concerning the differences between the different methods when analyzing data, should be helpful to advance our understanding of both missing data and the analysis of it. A simulation approach should prove beneficial, it offers various

advantages over the usage of real datasets. Simulating data firstly gives the researcher the option to manipulate the values however necessary. Covariances between pre-test and post-test values, the size of samples, the amount of drop-outs/missing data only scratch the surface of the possibilities. Data from the field typically does not offer these benefits, the researcher can only work with whatever the nature offers.

Therefore, data according to a onegroup pre-test post-test design will be simulated. Using programs for statistical coding and analysis, values are removed as desired by the researchers, to create a dataset containing missing data. Accordingly, datasets with and without missing data along varying conditions will be analyzed using the two univariate methods, as well as the MV method. The two univariate methods will directly analyze the data, using the standard of CCA. For the MV method on the other hand, a model is created, using Markov chain Monte Carlo methods (MCMC) to estimate the parameters. These allow for the data containing missing values to be analyzed under OCA. Accordingly, the MV method is expected to handle those datasets better and result in more accurate estimations.

Consequently, a research question is formulated:

"To what extent does the MV method offer more precise and accurate estimators of coefficients, compared to the univariate methods in different missing data scenarios?"

3 Methods

In order to reliably test the univariate methods against the MV methods, data has been simulated under varying conditions. The conditions were predefined in so far, that it was decided to manipulate the covariances of values between pre-test and post-test as well as the sample size. The sample size differed from pre-test and post-test, while the full data sets contained full samples, the incomplete data sets had varying amounts of missing values on pre-test and/or post-test. The specifics of these conditions were worked out "on the go", depending on the results from the previous data sets. Generally, the aim was to provide a broad picture of how the different methods behave when exposed to the different conditions. Two slightly different versions of simulation studies were done. Firstly, data was simulated with values missing in a MCAR pattern. A second version of data simulation was done, according to a MAR pattern. For this second study, the values of scores on an observed post-test variable is linked to their missingness, with higher values being less likely to occur than lower values. However, the models of the univariate

and MV methods would remain the same for both versions.

MCAR simulation study

Under all conditions the process was replicated 100 times, resulting in 100 different data sets with equal covariance and amount of missing data. For all data sets under all conditions, the RV model, the CS model, and lastly the MV model were fitted. The estimated treatment effect, standard deviations and variance of residuals were stored. Additionally, the 100 data sets per condition were stored as well. Also the variance of the differences of residuals between MV model and RV model or CS model were stored. The estimated variance of the residuals (sigma), of each of the different models was stored. Lastly, the calculated covariance (tau) of the MV method was stored.

For each replication, an overall sample size of $n = 500$ was used, however this number would deviate for the pre-test (n_1) or post-test (n_2), depending on the condition. The intercept and treatment effect varied over each replication under all conditions. In order to compare the estimations, the bias, mean square error (MSE), and standard error of the estimate were calculated.

For the MCAR simulation study, a total of 20 different conditions has been simulated, arranged in a 5x4 grid - see Table 1. Datasets with the covariances .75, .50, .25, and .10 were generated, alongside sample sizes of $n_1 = 350$ $n_2 = 380$, $n_1 = 450$ $n_2 = 290$, $n_1 = 500$ $n_2 = 100$, and $n_1 = 250$ $n_2 = 500$, with a complete set of $n_1 = n_2 = 500$. These conditions were chosen to offer a broad view of how amount of missing data and varying covariance can affect the reliability of the different measurement techniques.

MAR simulation study

Again, the process was replicated 100 times for all conditions, resulting in 100 different data sets with equal covariance and amount of missing data per condition, fitting the three models to all data sets. The same data was stored for all conditions as in the MCAR version of the simulation study. After evaluation of the two versions apart, they are to be compared to see whether the differences in patterns influence the method's behaviour differently.

For each replication, an overall sample size of $n = 500$ was used, for this version, only the missingness on the n_2 was manipulated. The intercept and treatment effect still varied over each replication under all conditions. To compare the three method's estimates, bias, MSE, and standard error of the estimate were used again.

In the MAR simulation study, 12 different conditions were simulated, arranged in a 3x4 grid as in Table 2. The covariances .10, .40 and .70 were simulated. The n_1 for all conditions were set to 500, while the amount of observations ranged from $n_2 = 100$ to

$n_2 = 400$ in steps of a hundred. These conditions are supposed to offer a broad view on how the MAR pattern of missingness affects the behaviour of the methods on different numbers of missing values, over varying covariances.

3.1 Materials

The statistical programming language R provided the best tool to both simulate and analyze the data. Furthermore, the packages "car", "MASS", and "sampling" were used. For all conditions, the amount of coefficients on pre-test and post-test as well as of the common coefficients remained the same. Data was simulated with one predictor at the pre-test, two at the post-test and one common predictor - the intercept. Accordingly, the models were developed. In order to simulate the data, full data sets of 500 observations on the pre- and on the post-test were created. For the conditions of missing values on either side, certain amounts of values would be removed afterwards - along random patterns for the MCAR simulation study and based on their own value for the MAR simulation study.

The accuracy of an estimation describes the estimator's performance, characterized as the overall difference between the estimated values and the true values (Walther & Moore, 2005). To evaluate the overall accuracy, the MSE is used. The MSE is the mean of squared differences, representing the proximity of estimates and true values, calculated with $MSE = \frac{1}{n}\sum_{j=1}^{n}(E_j - A_j)^2$. E_j here representing the estimated coefficient and A_j the true value of sample j. Precision on the other hand is not related to the true values of a sample, typically defined as the absence of random error (Walther & Moore, 2005). Therefore precision descibes the variability or spread of values surrounding their own mean, high precision is equal to a low variability. The precision of an estimator is often evaluated using the standard error of the estimate, also known as the estimator's standard deviation (SD). It is calculated using the formula $SD = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(E_j - \bar{E})^2}$ and indicates how widespread the estimates are around the sample mean. $\bar{E}$ indicating the mean coefficient-estimate. Lastly, bias is regarded as the distance between the sample mean and the true value (Walther & Moore, 2005). To measure the bias, the mean error (ME) is calculated, using the formula $ME = \frac{1}{n}\sum_{j=1}^{n}(E_j - A_j)$. It represents the mean of all differences between the true and the estimated values, accordingly indicating whether values were consistently over- or underestimated for each model. The MSE, while regarded as a measure of overall accuracy, borrows concepts of both precision and bias, as it is equal to the variance of estimates (closely related to SD) plus the squared ME (Casella & Berger, 1990).

Scholars have a tendency to see unbiased estimators as more desirable, bias appears to carry a negative connotation with it. However, it is important to keep in mind, that

an unbiased estimator does not necessarily lead to more *accurate* results. An unbiased estimator can only be seen as more accurate, if its estimates are reasonably precise. It is possible and not uncommon that a method may lead to slightly biased results on one hand, but on the other hand shows widespread and accordingly imprecise estimates. In order for a method to work as a truly accurate estimator, it needs to show both no bias and precision in its results.

Models

Firstly, the RV model was fitted to the data, leading to the model:

$$Y_{2_{ij}} = \beta_0 + \beta_1 T^*_{1_{ij}} + \beta_2 T^*_{2_{ij}} + \beta_3 Y_{1_{ij}} + \epsilon_{ij}, \epsilon_{ij} \sim N(0, \sigma^2)$$

With $T^*_{1_{ij}}$ indicating the presence of treatment effect nr 1 and $T^*_{2_{ij}}$ the presence of treatment effect nr 2 in the post-test scores of person i belonging to group j. The treatment effects change with every iteration, randomly sampled from a multivariate normal distribution. The RV method makes use of the pre-test score $Y_{1_{ij}}$ by including it in the regression equation model as a covariate.

Secondly, the CS method was fitted to the date, with the following model:

$$(Y_{2_{ij}} - Y_{1_{ij}}) = \beta_0 + \beta_1 T^*_{1_{ij}} + \beta_2 T^*_{2_{ij}} + \epsilon_{ij}, \epsilon_{ij} \sim N(0, \sigma^2)$$

For the CS method, the $Y_{1_{ij}}$ is not treated as a covariate, but included in the equation on the lefthand side. Accordingly, the CS method evaluates the change from pre-test to post-test.

Lastly, the MV model is given by:

$$Y_{1_{ij}} = \beta_{01} + \beta_{11} T_{1_{ij}} + \epsilon_{1_{ij}} \tag{4}$$

$$Y_{2_{ij}} = \beta_{02} + \beta_{12} T^*_{1_{ij}} + \beta_{22} T^*_{2_{ij}} + \epsilon_{2_{ij}} \tag{5}$$

$$\begin{pmatrix} \epsilon_{1_{ij}} \\ \epsilon_{2_{ij}} \end{pmatrix} \sim MVN(0, \Sigma), \Sigma = \begin{pmatrix} \sigma^2, \rho \\ \rho, \sigma^2 \end{pmatrix} \tag{6}$$

The error terms of $\epsilon_{1_{ij}}$ and $\epsilon_{2_{ij}}$ are correlated and normally distributed along a MV distribution with a mean of 0 and covariance of Σ.

As previously mentioned, the MV method makes use of MCMC methods. The code for the MCMC methods was run with 5,000 iterations per simulated dataset. Accordingly, based on the 5,000 iterations, the method provides estimates of treatment effects, variance, and covariance of parameters.

4 Results

4.1 MCAR simulation study

The MCAR version of the simulation studies led to the results in Table 1. The conditions of varying covariance and sample sizes on pre- and post-test influenced the data in different ways. For all approaches the ME-values seemed to center around zero, with no change ocurring based on the varying conditons.

The data suggests, that a rising covariance may lead to a rising MSE-value, especially for the multi-variate approach. However, this is not apparent for all conditions, as seen in Table 1. Furthermore, MSE-values generally seem lowest for the RV method, highest for the CS method and between those for the MV one. However, that does not seem the case for the conditions, of $n_1 = 250$ and $n_2 = 500$. Visualized in Figure 1, the MV method leads to lower MSE-values than the RV method. The MSE-values of the respective coefficient-estimates also seem to drop with lower amounts of missing data. Figure 2 shows that the lower the number of missing values, the more the MSE-value seems to drop.

Table 1: *MCAR Simulation Study - Results of All Conditions*

	Cov.	0.1			0.25			0.5			0.75		
n_1 n_2		ME	MSE	SD	Me	MSE	SD	Me	MSE	SD	Me	MSE	SD
	MV_1	.0041	.0028	.0444	.0014	.0031	.0448	-.0025	.0040	.0449	.0022	.0042	.0446
	$MV_{2.1}$	-.0025	.0030	.0446	-.0018	.0039	.0446	.0034	.0040	.0448	-.0104	.0041	.0448
	$MV_{2.2}$	.0007	.0024	.0449	-.0027	.0027	.0449	-.0034	.0040	.0451	-.0026	.0034	.0447
500 500	$RV_{2.1}$	-.0009	.0023	.0466	-.0025	.0025	.0490	.0041	.0031	.0523	-.0052	.0026	.0544
	$RV_{2.2}$	-.0027	.0019	.0469	-.0020	.0022	.0494	-.0026	.0028	.0526	.0020	.0026	.0542
	$CS_{2.1}$	-.0061	.0035	.0671	-.0007	.0040	.0674	-.0012	.0054	.0675	-.0019	.0037	.0677
	$CS_{2.2}$	-.0098	.0043	.0675	-.0087	.0049	.0679	-.0001	.0048	.0678	.0026	.0051	.0675
	MV_1	.0022	.0050	.0536	-.0057	.0042	.0537	-.0007	.0044	.0534	-.0040	.0064	.0534
	$MV_{2.1}$	.0018	.0041	.0512	-.0019	.0037	.0512	.0069	.0041	.0512	.0024	.0052	.0514
	$MV_{2.2}$	.0033	.0028	.0515	-.0006	.0036	.0517	-.0003	.0037	.0514	-.0014	.0055	.0514
380 350	$RV_{2.1}$	-.0085	.0041	.0645	.0050	.0042	.0673	.0058	.0057	.0713	.0066	.0066	.0745
	$RV_{2.2}$	.0011	.0025	.0648	.0008	.0057	.0680	.0058	.0051	.0721	.0014	.0046	.0742
	$CS_{2.1}$	-.0114	.0076	.0917	.0153	.0086	.0921	-.0064	.0074	.0913	.0046	.0099	.0917
	$CS_{2.2}$	-.0005	.0063	.0922	.0054	.0099	.0932	.0072	.0096	.0922	.0099	.0071	.0914
	MV_17	.0011	.0046	.0631	-.0015	.0057	.0637	.0094	.0070	.0638	-.0134	.0072	.0628
	$MV_{2.1}$	-.0098	.0024	.0448	-.0032	.0033	.0446	.0011	.0028	.0449	-.0065	.0042	.0447
	$MV_{2.2}$	.0026	.0024	.0451	.0084	.0031	.0449	-.0015	.0031	.0447	-.0047	.0038	.0446
250 500	$RV_{2.1}$	-.0069	.0036	.0667	-.0063	.0054	.0705	.0099	.0053	.0741	-.0007	.0069	.0769
	$RV_{2.2}$	.0060	.0044	.0670	.0134	.0041	.0707	.0022	.0051	.0737	.0055	.0055	.0761
	$CS_{2.1}$	.0030	.0093	.0952	-.0201	.0098	.0971	.0057	.0102	.0951	.0095	.0103	.0956
	$CS_{2.2}$	.0093	.0109	.0956	.0232	.0090	.0973	.0087	.0086	.0946	.0126	.0082	.0946
	MV_1	.0050	.0029	.0472	.0092	.0039	.0471	.0040	.0042	.0473	.0017	.0041	.0469
	$MV_{2.1}$	-.0023	.0041	.0588	.0029	.0045	.0592	.0075	.0058	.0587	.0007	.0077	.0592
	$MV_{2.2}$	-.0040	.0040	.0589	-.0021	.0060	.0591	.0070	.0057	.0589	.0105	.0071	.0584
450 290	$RV_{2.1}$	-.0015	.0038	.0651	.0036	.0042	.0694	.0085	.0061	.0717	-.0056	.0067	.0755
	$RV_{2.2}$	.0041	.0042	.0652	.0011	.0053	.0691	.0044	.0053	.0721	.0092	.0062	.0745
	$CS_{2.1}$	-.0096	.0089	.0923	.0088	.0067	.0950	.0051	.0105	.0925	-.0085	.0109	.0941
	$CS_{2.2}$	.0018	.0088	.0924	.0051	.0088	.0945	.0169	.0095	.0930	.0108	.0091	.0928
	MV_1	.0049	.0024	.0447	.0046	.0031	.0447	-.0115	.0038	.0448	.0054	.0038	.0451
	$MV_{2.1}$	-.0059	.0121	.1013	.0128	.0168	.0999	.0063	.0248	.1009	-.0112	.0201	.1001
	$MV_{2.2}$	-.0036	.0174	.1009	.0134	.0186	.1013	-.0290	.0180	.1022	.0081	.0200	.1007
500 100	$RV_{2.1}$	-.0109	.0089	.1076	.0020	.0125	.1103	-.0048	.0206	.1182	-.0107	.0147	.1238
	$RV_{2.2}$	-.0092	.0155	.1074	.0131	.0137	.1119	-.0159	.0149	.1198	-.0040	.0135	.1243
	$CS_{2.1}$	-.0068	.0206	.1543	-.0089	.0206	.1480	-.0028	.0291	.1535	-.0152	.0230	.1528
	$CS_{2.2}$	-.0221	.0284	.1539	.0005	.0230	.1501	.0001	.0262	.1556	-.0098	.0189	.1537

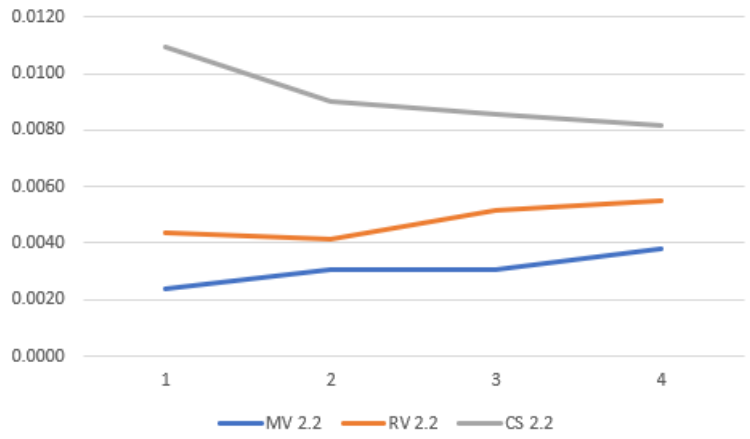

Figure 1: MSE-values in the case of $n_1 < n_2$ ($n_1 = 250, n_2 = 500; 1 - Cov. = .1, 2 - Cov. = .25, 3 - Cov. = .5, 4 - Cov. = 75$)

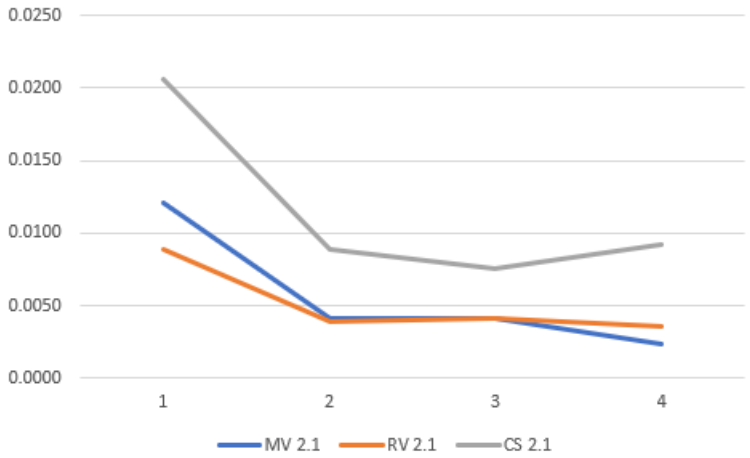

Figure 2: MSE-values falling with less missing values ($Cov. = .1; 1 - n_2 = 100, 2 - n_2 = 290, 3 - n_3 = 380, 4 - n_2 = 500$)

Concerning the standard error of estimates, the simulation studies showed that the varying covariance appears to be of no influence for the MV method and the CS method. For the RV method, the SD-values seem to rise with a higher covariance. Generally, the SD-values appear to differentiate between the three methods. However, for a high covariance, the MV method leads to the lowest values, with the RV method leading to values between those and the high values of the CS method. Figure 3 shows a representation of that behaviour. This relationship appears to be consistent throughout all conditions. Furthermore, it appears that the values of standard error of the estimates drop with more complete sample sizes. As seen in Figure 3 as well, the lower the amount of missing data, the lower the SD-value.

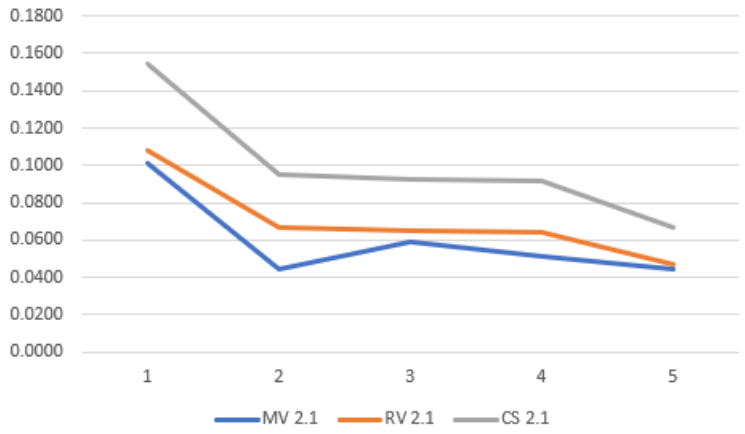

Figure 3: SD-values of the three methods with falling missingness ($Cov. = .1; 1 - n_1 = 500, n_2 = 100; 2 - n_1 = 250, n_2 = 500, 3 - n_1 = 450, n_2 = 290, 4 - n_1 = 350, n_2 = 380, 5 - n_1 = 500, n_2 = 500$)

4.2 MAR simulation study

Similarly, the results of the MAR version of simulated data can be found in Table 2. Generally, similar trends can be found in this second part of the study. Concerning the measurement of ME, all values seemed to center around zero. However, it became apparent that in the MAR simulation study a high amount of missingness, as in the $n_2 = 100$ condition, the ME-values fall systematically below zero. This behaviour was noted for all three methods.

The influence of the covariance still seems to be ambiguous, but it becomes apparent that a rising covariance leads to a rising MSE-value for the respective MV estimate and less so for the univariate approaches, Figure 4 shows an example of this. The MSE-values are lowest in the RV method and highest for the CS method, the MV MSE-values lie in between those. Similarly to the MCAR version, in the MAR version the MSE-values also tend to fall with less missingness. Furthermore, in figure 4 it appears, that the MSE-values of the CS model drop with a rising covariance. However, this behaviour can not be identified in all conditions or all CS-parameters. For example, in the same condition ($n_1 = 500, n_2 = 300$) the MSE-values of the second coefficient estimate do not drop, but rise with a higher covariance - see Table 2.

The standard error of estimates was not influenced by varying the covariance, for the MV method and the CS method. The RV leads to higher SD-values with a rising covariance, Figure 5 shows that relationship, similarly to the MCAR results. Just like for the MCAR version, the MV method led to the lowest SD-values, the RV to higher and the CS method to the highest values, as seen in Figures 5 and 6. Furthermore, the SD-values seem related to the amount of missingness. The more missing values, the higher the SD-values, for all methods respectively. See Figure 6 for an example of this.

Table 2: *MAR Simulation Study - Results of All Conditions*

	Cov.	0.1			0.4			0.7		
n_1 n_2		ME	MSE	SD	ME	MSE	SD	ME	MSE	SD
	MV_1	-.0015	.0026	.0448	-.0047	.0037	.0446	-.0023	.0036	.0445
	$MV_{2.1}$	.0043	.0125	.1015	-.0041	.0211	.0999	-.0027	.0215	.1008
	$MV_{2.2}$	-.0035	.0146	.1018	.0133	.0182	.1015	-.0052	.0196	.1007
500 100	$RV_{2.1}$	.0004	.0111	.1089	-.0026	.0152	.1154	-.0097	.0146	.1221
	$RV_{2.2}$	-.0024	.0120	.1090	.0165	.0157	.1173	-.0042	.0149	.1218
	$CS_{2.1}$	.0092	.0221	.1551	-.0027	.0204	.1509	-.0229	.0235	.1533
	$CS_{2.2}$	-.0128	.0263	.1553	.0287	.0267	.1534	-.0110	.0211	.1530
	MV_1	-.0057	.0028	.0449	-.0039	.0043	.0447	-.0081	.0037	.0445
	$MV_{2.1}$	-.0050	.0067	.0716	-.0028	.0101	.0710	-.0154	.0097	.0709
	$MV_{2.2}$	.0137	.0071	.0703	.0007	.0069	.0708	-.0041	.0095	.0714
500 200	$RV_{2.1}$	-.0086	.0050	.0758	.0002	.0069	.0808	-.0091	.0069	.0859
	$RV_{2.2}$	.0121	.0054	.0742	.0011	.0061	.0805	-.0080	.0068	.0863
	$CS_{2.1}$	-.0143	.0119	.1080	-.0040	.0106	.1059	-.0024	.0132	.1065
	$CS_{2.2}$	.0141	.0100	.1058	-.0151	.0100	.1055	-.0031	.0111	.1070
	MV_1	-.0023	.0026	.0449	-.0062	.0037	.0449	-.0133	.0042	.0450
	$MV_{2.1}$	-.0038	.0046	.0578	.0048	.0061	.0580	.0027	.0077	.0580
	$MV_{2.2}$	.0040	..0057	.0581	-.0053	.0053	.0580	.0001	.0071	.0577
500 300	$RV_{2.1}$	.0007	.0037	.0611	.0079	.0039	.0667	.0034	.0039	..0701
	$RV_{2.2}$	.0036	.0043	.0612	-.0038	.0036	.0664	-.0054	.0051	.0696
	$CS_{2.1}$	-.0074	.0073	.0864	.0132	.0080	.0873	.0050	.0055	.0878
	$CS_{2.2}$	-.0018	.0076	.0866	-.0004	.0062	.0870	-.0113	.0071	.0872
	MV_1	.0006	.0023	.0446	.0069	.0029	.0447	-.0104	.0040	.0446
	$MV_{2.1}$	.0126	.0061	.0505	-.0078	.0070	.0500	.0005	.0051	.0497
	$MV_{2.2}$	-.0147	.0032	.0503	.0035	.0041	.0498	-.0016	.0042	.0503
500 400	$RV_{2.1}$	.0026	.0029	.0535	-.0059	.0033	.0570	-.0001	.0030	.0600
	$RV_{2.2}$	-.0118	.0022	.0532	.0034	.0029	.0568	-.0033	.0035	.0603
	$CS_{2.1}$	-.0067	.0055	.0766	-.0127	.0058	.0761	.0042	.0063	.0750
	$CS_{2.2}$	-.0043	.0041	.0762	-.0032	.0070	.0757	-.0083	.0057	.0754

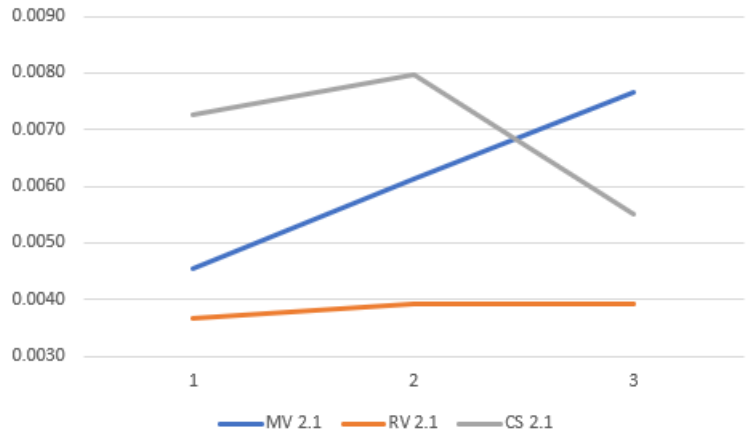

Figure 4: MSE-values of all three methods over the covariances ($n_1 = 500, n_2 = 300; 1 - Cov. = .1, 2 - Cov. = .4, 3 - Cov. = .7$)

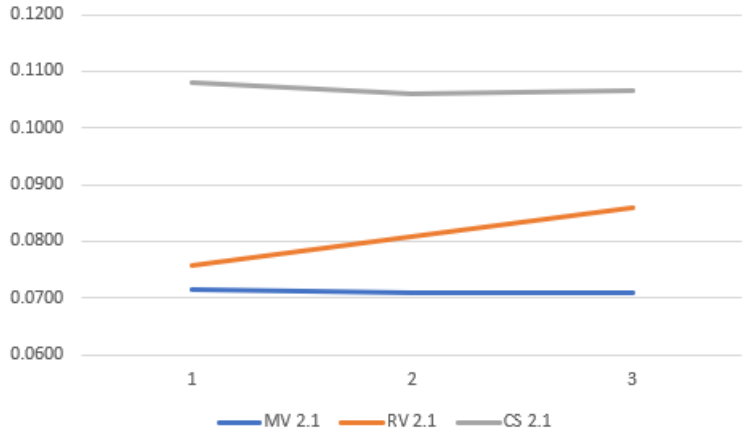

Figure 5: SD-values of all three methods over covariance ($n_1 = 500, n_2 = 300; 1 - Cov. = .1, 2 - Cov. = .4, 3 - Cov. = .7$)

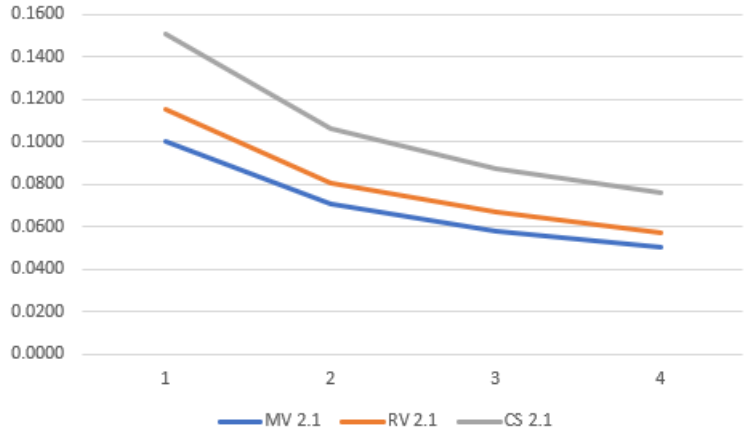

Figure 6: SD-values falling with less missing values over all three methods ($Cov. = .4; 1 - n_2 = 100, 2 - n_2 = 200, 3 - n_3 = 300, 4 - n_2 = 400$)

4.3 Comparing MCAR and MAR

Comparing the results of the MCAR version to the MAR version shows their similarities. Introducing a MAR-pattern to the missing of values on the post-test appears to have no big influence on the values received from all 3 methods. Comparing the $n1 = 500$, $n2 = 100$, covariance=.1 conditions in the MAR and MCAR studies shows barely differing values. Comparing the $n1 = 500$, $n2 = 100$, covariance=.75 condition of the MCAR version to the similar $n1 = 500$, $n2 = 100$, covariance=.70 condition of the MAR version shows that this behaviour appears consistent over the conditions.

5 Discussion

The two simulation studies resulted in some notable differences concerning the MV and univariate methods. The ME-measurement was used in order to test, whether any of the methods consistently under- or overestimate the coefficients. While this generally seemed not to be the case, all methods appeared to slightly but systematically underestimate the true value of the coefficients in the MAR study under the presence of a large missingness paired with high covariance. The MAR conditions introduced a pattern of more likely selection of lower scores to remain on the post-test. Accordingly, a distribution further on the left-hand side of the Y-axis was available for analysis. The more values would be missing from the data set, the bigger the impact on the distribution, moving it further away toward the left-hand side from the true values. This effect appears to be accentuated by the presence of a high covariance between the pre- and the post-test scores. Egbewale, Lewis and Sim (2014) found similar results of less accurate results gained from analyses of variance or CS, in the case of strong covariance.

In both versions of the test, a higher covariance between the pre- and the post-test was associated with rising MSE-values, for the MV method more evident than for the two univariate methods. The MSE-measurement was used to assess the methods overall accuracy, as described above. A positive relationship of covariance and MSE suggests that the accuracy of all methods is lower, the more strongly scores on pre- and post-test correlate. Specifically, the MV method appears to lead to more widespread estimates, influenced by that pattern. MV methods typically make use of information from modelling a correlation of the error terms on both a pre- and the post-test, referring to the model used for the two simulation studies. The MV method can therefore make use of information from both of these measurement points. However, as the scores on the two measurement points correlate more strongly, information gained from evaluating both declines as it starts to overlap. Accordingly, the MV method has less information to its disposal, random error increases leading to less accurate estimates. As the univariate

methods only model a single error term per observation, higher or lower covariance should have less influence on their estimations.

The MSE also appeared to be consistently lowest for the RV method, a bit higher for the MV approach and highest for the CS method. This suggests, that the methods accuracy can be evaluated in the same order; the RV method has the highest accuracy, MV only medium and CS method the lowest accuracy. However, in the rare case that the sample contains substantially more missing observations on the pre-test than on the post-test, the MV approach leads to more accurate results. This is likely due to the employment of MCMC methods. In the current literature, MCMC is widely used as a tool for data imputation, specifically multiple imputation (Nakai & Ke, 2011; Takahashi, 2017; Rezvan, Lee, & Simpson, 2015). The advantage of filling gaps with the generated pseudo-numbers, likely offers benefits in the condition of less observations on the pre-test. As the univariate methods automatically make use of CCA, in cases of large missingness, their estimations grow less accurate. While the employment of MCMC generally does not improve the accuracy of estimates of the 100 data sets, it appears to lead to more accurate results in the specific case of less observations on the pre-test than on the post-test.

In order to describe the precision of the three methods, the SD-measure was used. Concerning the precision of estimating the coefficients, the MV-method clearly gave better result. Overall, the estimates of the MV-method were most precise, estimates of the RV-method less so and the estimates of the CS-method proved to be the least precise ones. However, it is also important to note, that the RV-method was the only method to show a positive relationship between the covariance and its precision. The more closely the observations on the pre-test seemed to be related to the ones on the post-test, the less precise the estimates seemed to become. The amount of missing data seemed to have a negative effect on the precision of all used methods. The more values were missing on an observation-timespot, the more imprecise the estimation of coefficients became. This is a natural result of missing data. As less values are available for calculation, the estimations will be more widespread surrounding the true value, for all used methods. The introduction of a MAR-pattern to the missingness however, seemed to have no influence on the precision of estimates, even for high amounts of missing values.

The standard error of estimates is relevant for the testing of treatment effects. The higher the value of the standard error of estimates is, the more widespread are the estimates around the sample mean. The standard deviation, which is crucial to the calculation of various test-statistics, is closely related to the standard error as $\sigma \approx \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$. σ here is used to estimate the standard error, using x_i as an estimator over all datasets. Models leading to higher SD-values accordingly lead to higher standard deviations as well, which is used to assess the significance of group differences, correlational effects and

most importantly - treatment effects. Confidence intervals or p-values are used most often to assess a treatment effect's statistical significance. Confidence intervals are computed using the formula ($\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$), which means that the confidence interval will grow broader with a rising standard deviation. Broader confidence intervals are more likely to contain a zero, which would lead the researcher to evaluate the effect as non-significant. The same would happen with p-values, which also rise with standard deviation. This means, that the RV- and CS-methods, given the same data, would lead to an underestimation of treatment effects and statistical significance. The MV-method on the other hand seems less prone to that, as its more precise estimates.

6 Conclusion

The simulation studies showed, that both the MV approach, as well as the RV method have their clear space in data analysis. The CS method on the other hand seemed to lead to the least accurate results in all regards. Generally, it appears for simple pre-test post-test procedure without control group, that the RV method poses a viable option leading to fairly accurate results. While the researcher has to keep in mind that its estimates appear to be less precise than those of a MV method, the estimators still result in accurate estimates of true values. In certain conditions, the MV method leads to more accurate results, specifically in cases of high missingness on the pre-test. Furthermore, the MV approach consistently led to more precise estimates of coefficients. Whenever a researcher is in need of high precision for more valid or reliable results, a MV approach is recommended, for research of similar conditions concerning factors on pre- or post-test. While the accuracy estimation may lead researchers to come to different conclusions, the estimation of precision might be more crucial to everyday research. As pre-test post-test studies are typically aiming to test treatment effects, the RV-method appears to have a tendency to underestimate their significance, possibly missing it. The MV-method therefore is to be preferred, as its more precise estimates make up for a generally seemingly lower accuracy.

The effects of missing data can have further negative implications for univariate methods, if CCA is used. Not only can the diminished precision lead to an underestimation of treatment effects, the opposite could happen as well. As data is removed from the sample due to missing values, less error terms can be modelled by the univariate methods, compared to the MV method making use of OCA. As less error terms are modelled, it might happen that bigger errors are neglected. The standard error estimation would therefore fall and treatment effects are more likely deemed to be significant. This however would make type-I errors, or false-positives, more likely. Accordingly, it appears that employing

a MV method offers various benefits and more reliable results in significance testing over the univariate methods.

This study only compared the immediate results concerning accuracy, bias and prediction. Schafer and Olson (1998) pointed out that employing a MV approach to analyse missing data typically calls for post-analysis procedures, accounting for the effects of missing data. For the future analysing these post-analysis procedures might lead to very different results. For further research, more varying conditions of coefficients of pre- and post-test should be included. Seeing the effects more predictors in the different models can have on the data, should prove beneficial. It is important to note, that all data in this study stems from carefully controlled simulation. For a measurement of accuracy, a different method than the MSE might prove beneficial to use. MSE-values may not be ideal, as they are not standardized, which makes the comparison between methods difficult. Real life data often behaves differently under analysis. Including data from field studies, survey or questionnaires in future research might bring crucial new information, that could not be discovered here. Also varying MAR pattern might prove crucial. Lastly, this study made a distinction between univariate methods employing CCA and MV method employing OCA. Different ways of data imputation to fill the gaps left by missing values might have various effects on the univariate and MV methods' accuracy and precision. It is clear that handling missing data still requires much looking into, in order to ensure the most valid results of everyday research.

References

Allison, P. D. (1990). Change scores as dependent variables in regression analysis. *Sociological methodology*, 93–114.

Allison, P. D. (1999). *Missing data*. Sage Thousand Oaks, CA.

Casella, G., & Berger, R. L. (1990). Statistical inference vol. 70.

Egbewale, B. E., Lewis, M., & Sim, J. (2014). Bias, precision and statistical power of analysis of covariance in the analysis of randomized trials with baseline imbalance: a simulation study. *BMC medical research methodology*, *14*(1), 49.

Kleine Bardenhorst, S. (2017). Multivariate models for pretest posttest data and a comparison to univariate models.

Nakai, M., & Ke, W. (2011). Review of the methods for handling missing data in longitudinal data analysis. *International Journal of Mathematical Analysis*, *5*(1), 1–13.

Rezvan, P. H., Lee, K. J., & Simpson, J. A. (2015). The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. *BMC medical research methodology*, *15*(1), 30.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods*, *7*(2), 147.

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

Soley-Bori, M. (2013). Dealing with missing data: Key assumptions and methods for applied analysis. *Boston University*.

Takahashi, M. (2017). Statistical inference in missing data by mcmc and non-mcmc multiple imputation algorithms: Assessing the effects of between-imputation iterations. *Data Science Journal*, *16*.

Walther, B. A., & Moore, J. L. (2005). The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography*, *28*(6), 815–829.