AssignmentB

Martijn Koster, William Schaafsma, Martijn van Dam, Victor Hovius

3/21/2022

Contents

1 Introduction

The matter of interest for this assignment will be the impact that incomplete data (**observed data**) have on our inferences compared to the inferences we make with complete data (**true data**). To investigate the effect that missing values have on model inferences, we will build a multiple regression model.

Firstly, we provide descriptive statistics and correlations. In table **??** we compare the head of the observed data and the true data. Additionally, in 3 the means and variances are compared. With regard to correlations, we present two correlations matrices; one for the observed data 6 and the other for the true data 7.

Secondly, we present our multiple regression model in table 8. Our model consists of the outcome variable: *active heart rate* and the predictors: *age* and *smoke*. We also included an interaction effect between *bmi* and *sex*. The first three columns reflect the observed data, whereas the latter reflect the true data.

The research question we try to answer in accordance with our model is: *What influence does a person's age, smoking habits, sex and bmi have on their heart rate during excercise?*

Thirdly, we start by inspecting the missing values. Then, we try to find out where the missing values occur. In 1 we begin by giving a global overview of the missingness. Then, in 2 we compare the distributions for the observed data and the missing values.

Lastly, we perform t-tests on the variables containing missing values to check the type of missingness, either MNAR, MAR, or MCAR. We also provide plots here to visualize where the missing values occur.

2 Observed vs True data

In this section we will compare the observed with the true data set.

Table 1: First Five Cases of Observed Data

age	smoke	sex	intensity	active	rest	height	weight	bmi
42	no	female	high	NA	75	NA	NA	22.4
31	NA	male	low	NA	62	NA	NA	23.8
36	no	male	low	109	76	182	78.0	23.5
31	no	female	low	78	62	164	53.9	20.0
42	no	male	low	NA	66	189	NA	23.4

Table 2: First Five Cases of True Data

age	smoke	sex	intensity	active	rest	height	weight	bmi
42	no	female	high	94	75	161	58.1	22.4
31	no	male	low	86	62	184	80.6	23.8
36	no	male	low	109	76	182	78.0	23.5
31	no	female	low	78	62	164	53.9	20.0
42	no	male	low	103	66	189	83.6	23.4

2.1 Descriptives

The means and the variance of the variables age and rest remain unchanged, as they have no missing values.

The means of the variables active, weight, and bmi are somewhat higher for the observed data set compared to the true data set. For height, the mean of the observed data set is somewhat smaller compared to the true data set.

Furthermore, the variance of the variable active in the observed data set is .01 lower than the true data set, and thus almost entirely unaffected. However the variables height, weight, and bmi have greater variance in the true data set than the observed data set. This implies that the missingness causes an underestimation of the variance.

Table 3: Means and variances in true and observed dataset

Variables	M obs	M true	var obs	var true	N obs	N true
Age	38.52	38.52	149.73	149.73	306	306
Active	92.58	93.13	383.05	378.04	183	306
Rest	69.83	69.83	120.78	120.78	306	306
Height	174.50	173.99	100.66	105.29	214	306
Weight	73.28	73.58	270.28	274.85	132	306
Bmi	24.11	24.06	12.91	13.38	213	306

Note.

obs = Observed Dataset, true = True Dataset

2.2 Categorical variables

The categorical variables in both data sets are *smoke*, *sex* and *intensity*. *Smoke* and *sex* both have two levels ("no" and "yes" for smoke and "male" and "female" for sex), while *intensity* has three levels ("low", "moderate", and "high"). Despite differences in the number of observed values between the data sets, differences between groups remain unchanged. For example, there are more males than females in both data sets and more non-smokers than smokers. Also, in both data sets, more males reported smoking than females. The most frequently reported workout intensity for both males and females in the two data sets is moderate, followed by low and high.

Table 4: proportion table of categorical variables in observed data

sex	smoke	intensity	Freq
male	no	high	0.040
female	no	high	0.060
male	yes	high	0.065
female	yes	high	0.056
male	no	moderate	0.113
female	no	moderate	0.149
male	yes	moderate	0.085
female	yes	moderate	0.073
male	no	low	0.165
female	no	low	0.129
male	yes	low	0.048
female	yes	low	0.016

Table 5: proportion table of categorical variables in true data

sex	smoke	intensity	Freq
male	no	high	0.046
female	no	high	0.059
male	yes	high	0.075
female	yes	high	0.046
male	no	moderate	0.108
female	no	moderate	0.147
male	yes	moderate	0.085

Table 7: Correlations of true data

	age	smoke	sex	intensity	active	rest	height	weight	bmi
age	1.00	-0.05	-0.17	0.21	-0.54	-0.39	0.20	0.23	0.20
smoke	-0.05	1.00	-0.11	-0.31	0.18	0.27	0.17	0.25	0.24
sex	-0.17	-0.11	1.00	-0.09	0.09	0.06	-0.72	-0.69	-0.47
intensity	0.21	-0.31	-0.09	1.00	-0.37	-0.55	0.12	0.06	0.01
active	-0.54	0.18	0.09	-0.37	1.00	0.61	-0.10	0.02	0.09
rest	-0.39	0.27	0.06	-0.55	0.61	1.00	-0.15	-0.04	0.05
height	0.20	0.17	-0.72	0.12	-0.10	-0.15	1.00	0.77	0.36
weight	0.23	0.25	-0.69	0.06	0.02	-0.04	0.77	1.00	0.87
bmi	0.20	0.24	-0.47	0.01	0.09	0.05	0.36	0.87	1.00

Table 5: proportion table of categorical variables in true data *(continued)*

sex	smoke	intensity	Freq
female	yes	moderate	0.062
male	no	low	0.190
female	no	low	0.124
male	yes	low	0.046
female	yes	low	0.013

2.3 Correlations

As shown in Table 6 and Table 7, the correlations between the variables of the observed data set are slightly different from the correlations between variables of the true data. Although most of the correlations are almost identical, a few correlations are negative in the observed data and positive in the true data. This effect also occurs vice versa. In example, the correlation between the variables smoke and age of the observed data set is positive (r = 0.01), albeit almost 0. In contrast, the correlation for these variables in the true data set is negative (r = -0.05). However, the impact of missing data on the correlations appears to be minor, as the difference in correlation coefficients between the two data sets is negligible. Although some correlations differ in valency between the data sets, the correlation coefficients remain close to 0 and thus, do not distort inferences made with the observed data set.

Table 6: Correlations of observed data

	age	smoke	sex	intensity	active	rest	height	weight	bmi
age	1.00	0.01	-0.17	0.21	-0.49	-0.39	0.19	0.25	0.18
smoke	0.01	1.00	-0.09	-0.29	0.15	0.23	0.18	0.18	0.18
sex	-0.17	-0.09	1.00	-0.09	0.11	0.06	-0.73	-0.68	-0.42
intensity	0.21	-0.29	-0.09	1.00	-0.37	-0.55	0.13	0.12	0.02
active	-0.49	0.15	0.11	-0.37	1.00	0.56	0.00	0.01	0.05
rest	-0.39	0.23	0.06	-0.55	0.56	1.00	-0.20	-0.12	0.06
height	0.19	0.18	-0.73	0.13	0.00	-0.20	1.00	0.78	0.34
weight	0.25	0.18	-0.68	0.12	0.01	-0.12	0.78	1.00	0.88
bmi	0.18	0.18	-0.42	0.02	0.05	0.06	0.34	0.88	1.00

Table 8: Regression analysis of True (N=306) and Observed Data (N=155)

	Observed Data			True Data		
	b	SE	p	b	SE	p
(Intercept)	78.444	14.34	0.000	80.384	9.03	0.000
age	-0.809	0.11	0.000	-0.883	0.07	0.000
bmi	1.681	0.55	0.003	1.776	0.35	0.000
sexfemale	32.756	20.78	0.117	43.460	14.16	0.002
smokeyes	1.615	2.91	0.580	3.516	1.99	0.078
bmi:sexfemale	-1.131	0.88	0.199	-1.674	0.60	0.006

3 Regression

3.1 Answering the research question

When examining Table 8 *Regression analysis of True and Observed data* we observe several differences in the beta coefficients, standard error, and p-values. The table contains variables with missing values and an interaction effect. Although almost all beta coefficients are nearly equal, the beta coefficients of the observed data set are systematically underestimated. This underestimation is especially the case for *sexfemale*, as the difference between the beta coefficients is almost 9.0. Making inferences based on the observed data set would lead to underestimating the effect of sex on active hear rate. Regarding the standard errors, missing data caused these parameters of the observed data set to be systematically overestimated. Larger standard errors contribute to the possibility of making a type II error, as is the case in our data set. For example, the larger standard errors in the observed data set might have played a role in the variables *sexfemale* and the interaction *bmi:sexfemale* turning non-significant. These variables would wrongly be neglected when making inferences with the model based on the observed data. Concluding, the missing data causes the standard errors to be greater, resulting in less accurate beta coefficients. Moreover, some p-values turn out non-significant, caused by underestimated beta coefficients. The model based on observed data leads thus to inaccurate inferences.

4 Missingness

There are 540 missing values. 0 for age, 0 for sex, 0 for intensity, 0 for rest, 58 for smoke, 92 for height, 93 for bmi, 123 for active, and 174 for weight. Moreover, there are 132 completely observed rows, 15 rows with one missing value, 37 rows with two missing values, 52 rows with three missing values, 55 rows with four missing values, 15 rows with five missing values. The missingness in the data is non-monotone because the variable with the least missing values (*smoke*) has observed values for other variables with more missingness (e.g., *smoke* and *bmi*). The missingness would be monotone if the variable with the least missing values (*smoke*), would have missing values on all other variables with more missingness (e.g., *height*). Interestingly, a monotone pattern is only the case for *smoke* and *weight*.

4.1 Looking for the missingness

In this section, we will investigate whether the mean of the missing values differs significantly from the mean of the observed values. This will be done do by using a paired sampled t-test for the numeric variables. To compare the mean of the missing values with the true values, we computed a logical vector for each vector that has missing observations. The missingness vectors have the value `TRUE` for all missing entries and `FALSE` for all observed entries. These missingness vectors will be used as a grouping variable in the true data set to compare the missing values with the observed values. For smoke, which is a categorical variable, we will use

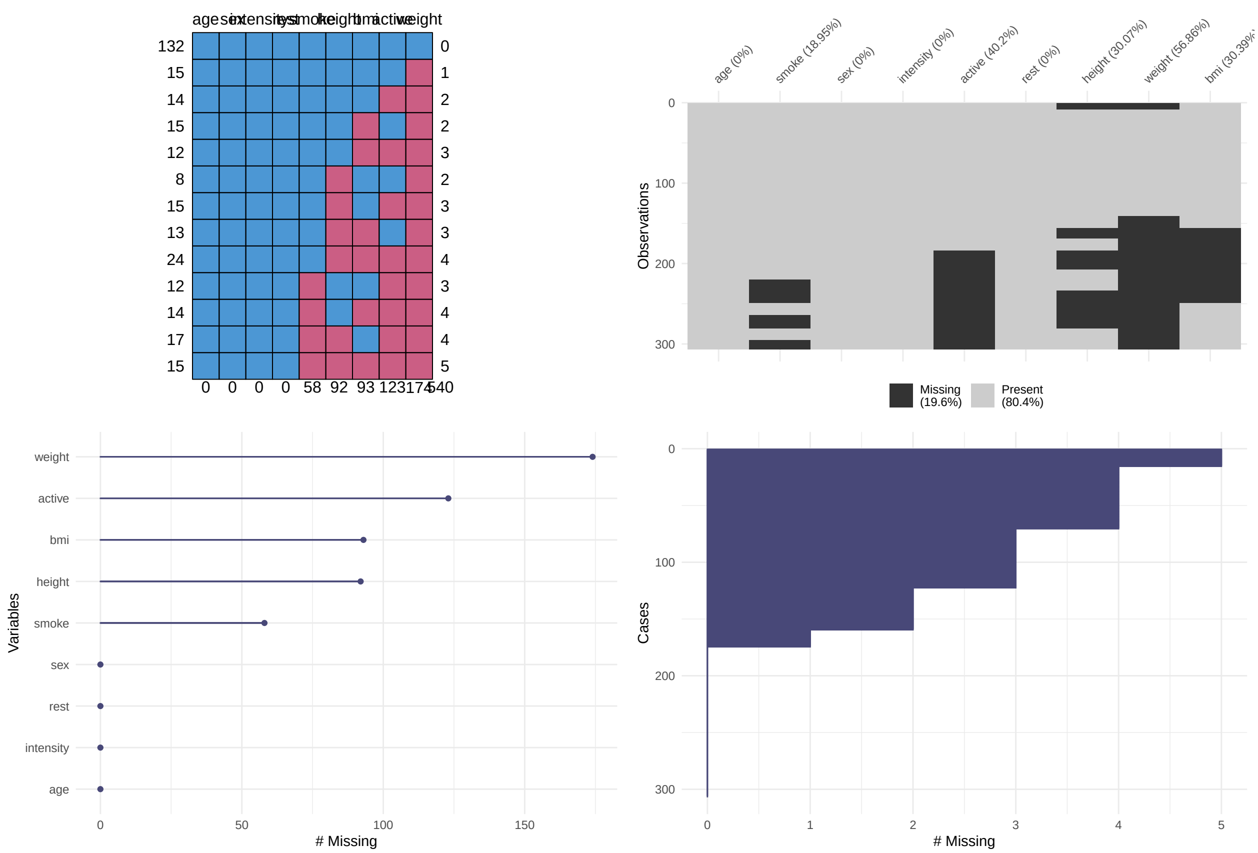


Figure 1: pattern of the missingness

Table 9: Difference in means of observed data and missing data

variables	M obs	M true	t	p
Weight	73.90	73.17	0.381	0.704
Height	174.50	172.83	1.271	0.205
Bmi	24.11	23.95	0.336	0.737
Active	95.58	93.95	-0.606	0.545

a x^2 test. For all variables, the missing values have a similar distribution as the observed values. However, the distribution for the variable smoke is not shown, as this is a categorical variable and does thus not have a distribution. The means of the variables from both data sets are marginally different, but the differences are non-significant, neither for smoke. Hence, the missing values are similar to the observed values.

smoke: $x^2 = 1.154$, $p = 0.283$

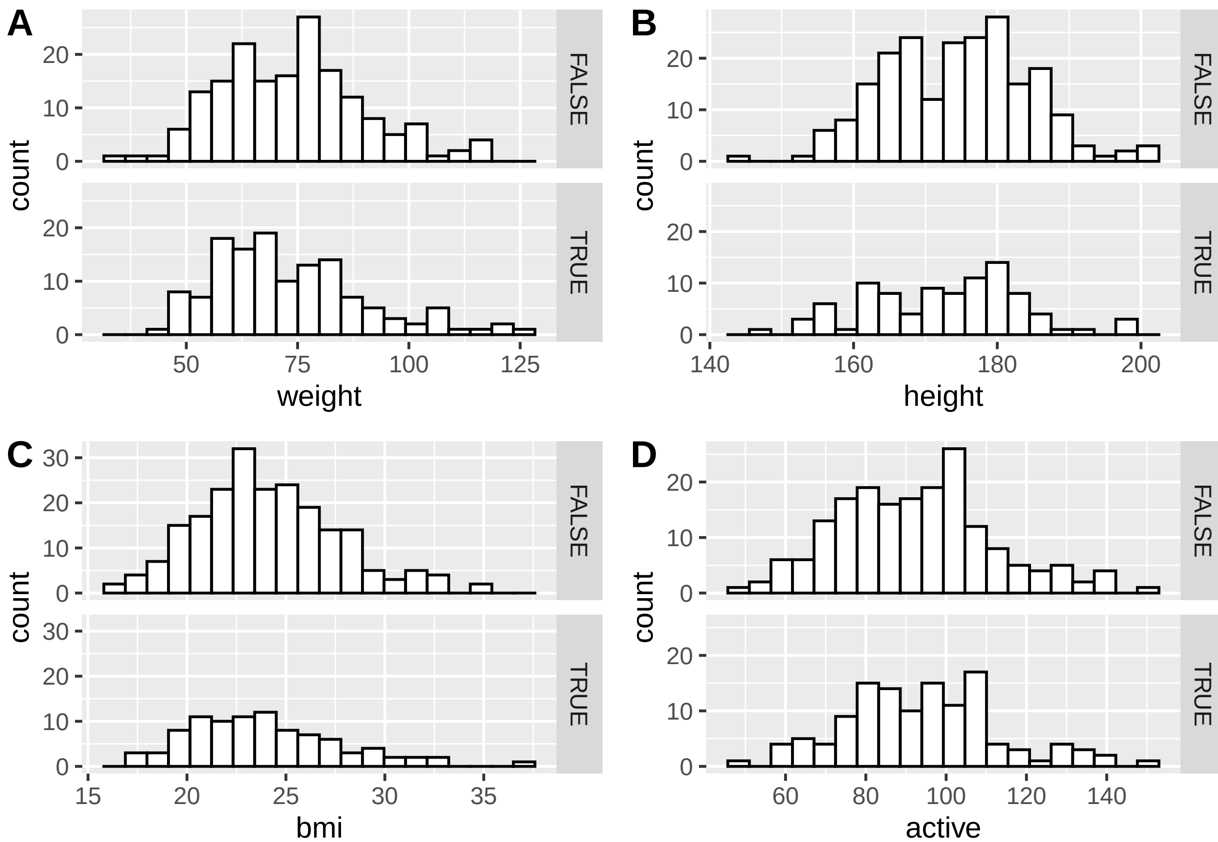


Figure 2: Comparing the distribution of the observed and true dataset

4.2 Missingness of weight

Looking at the differences of the missingness of weight in Table 10, no significant differences can be found in the means of the numeric variables in the data. However, the difference in mean of active is relatively high. It might be that this difference is not significant due to the low amount of observations of active when weight is missing ($N = 36$). Considering the categorical data, the missingness of weight on sex has no significant difference, where $x^2 = 0$, $p = 1$. For the missingness of weight on smoke no significant difference was found

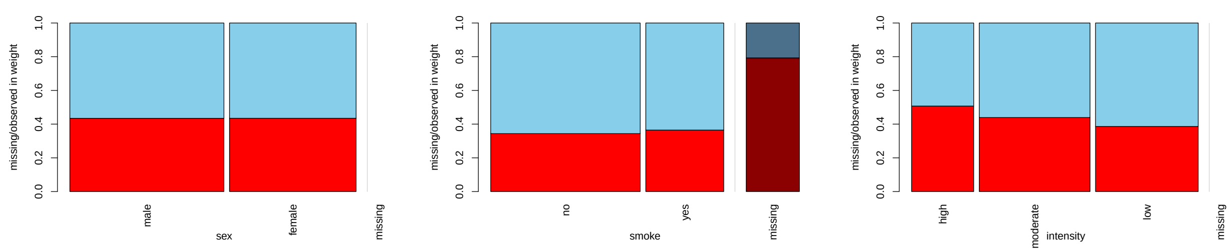
Table 10: Difference in means of missingness of Weight

variables	\$M obs	M true	t	p
rest	69.56	70.17	-0.482	0.630
age	38.16	38.98	-0.590	0.556
height	174.28	175.39	-0.639	0.525
bmi	24.11	24.12	-0.012	0.990
active	91.54	96.81	-1.440	0.156

also $x^2 = 0.036$, $p = 0.848$. Lastly, the missingness of weight on intensity is not significant: $x^2 = 2.589$, $p = 0.274$.
All results are non-significant; hence weight is not missing at random.

The three bar plots in Figure 3 show a visualization of how the missing data in the categorical columns is divided. The first plot shows us that there is almost no difference between missing values in weight for being a man or female in the sex column. The second plot also shows that there is almost no difference between missing values in the weight column for smokers and non-smokers in the smoke column. The third column shows that how lower the intensity is the less missing values in weight you can expect.

The five scatterplots in Figure 4 show a visualization of how the missing data is divided in the rest of the columns. In the first two plots between weight and rest or age is a clear trend where all the values with a low weight are missing, and everything above that is not. The two plots after that between weight and height or bmi show the same thing, but also a cluster of missing values when both columns have low values. The last column between weight and active shows a clear trend where low values for either column results in missing values with a cluster where both columns have low values.



(a) Sex (b) Smoke (c) Intensity

Figure 3: Looking whether the missingness of weight is MAR

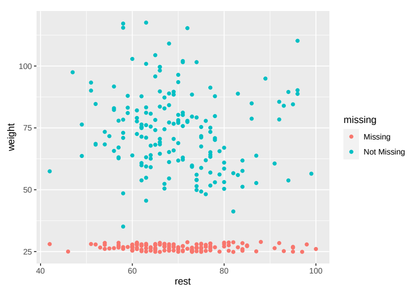
4.3 Missingness of height

Looking at the differences of the missingness of height in Table 11, no significant differences can be found in the means of the numeric variables in the data. Similar to the missingness of weight, the difference in mean of `active` is relatively high. Again, it might be that this difference is not significant due to the low amount of observations of `active` when height is missing ($N = 40$).

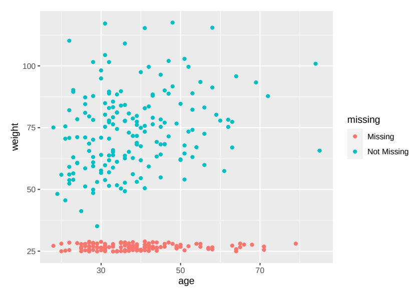
Considering the categorical data, the missingness of height on sex has no significant difference, where $x^2 = 0$, $p = 1$. For the missingness of height on smoke no significant difference was found also $x^2 = 0.111$, $p = 0.739$. Lastly, the missingness of weight on intensity is not significant: $x^2 = 3.563$, $p = 0.168$.

All results are insignificant, hence height is not missing at random.

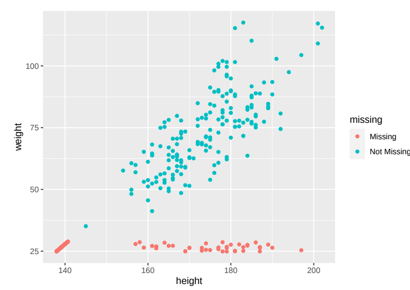
The three bar plots in Figure 5 show a visualization of how the missing data in the categorical columns is divided. The first plot shows us almost no difference between missing values in height for being a man



(a) Rest

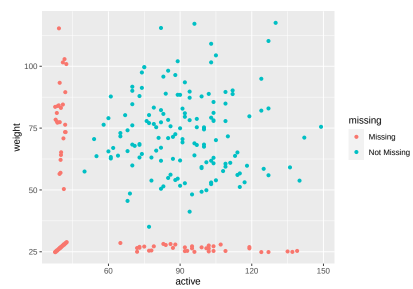


(b) Age



(c) Height

(d) Bmi



(e) Active

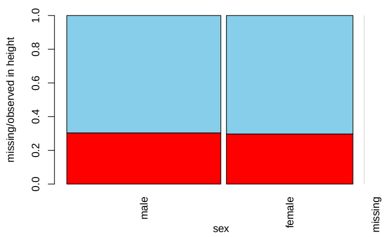
Figure 4: Looking whether the missingness of weight is MAR

Table 11: Difference in means of missingness of Height

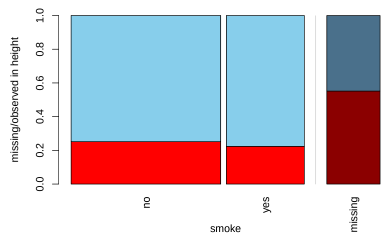
variables	$M obs	M true	t	p
rest	69.93	69.60	0.242	0.809
age	38.66	38.18	0.320	0.749
bmi	24.11	24.11	-0.012	0.990
active	91.74	99.05	-1.535	0.137

or female in the sex column. The second plot also shows virtually no difference between missing values in the height column for smokers and non-smokers in the smoke column. The third column shows almost no difference between missing values in the height column for high and moderate-intensity but less missing values in the low-intensity category.

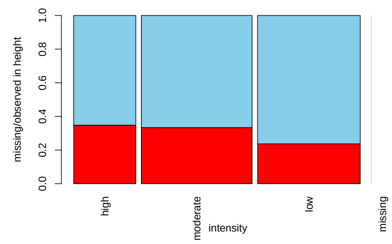
The four scatterplots in Figure 6 show how the missing data is divided into the rest of the columns. There is a clear trend in the first two plots between height and rest or age where all the values with a low height are missing and everything above that is not. The two plots after that between height and bmi or active show a clear trend where low values for either column result in missing values with a cluster where both columns have low values.



(a) Sex

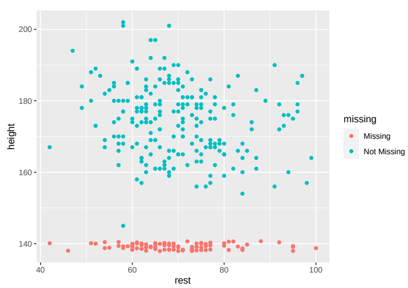


(b) Smoke



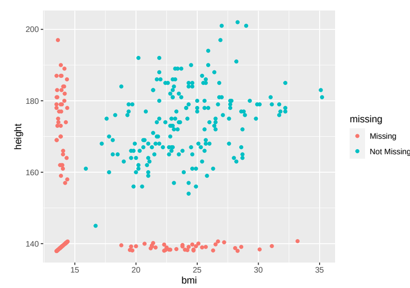
(c) Intensity

Figure 5: Looking whether the missingness of height is MAR



(a) Rest

(b) Age



(c) Bmi

(d) Active

Figure 6: Looking whether the missingness of height is MAR

4.4 Missingness of Active

Looking at the differences of the missingness of active in Table 12, no significant differences can be found in the means of the numeric variables in the data. However, both `bmi` ($p = 0.062$) and `weight` ($p = 0.059$) are

Table 12: Difference in means of missingness of active

variables	M obs	M true	t	p
rest	69.02	71.03	-1.558	0.120
age	37.96	39.35	-0.963	0.337
height	174.41	174.77	-0.232	0.817
bmi	23.83	24.85	-1.883	0.062
weight	72.94	79.38	-1.948	0.059

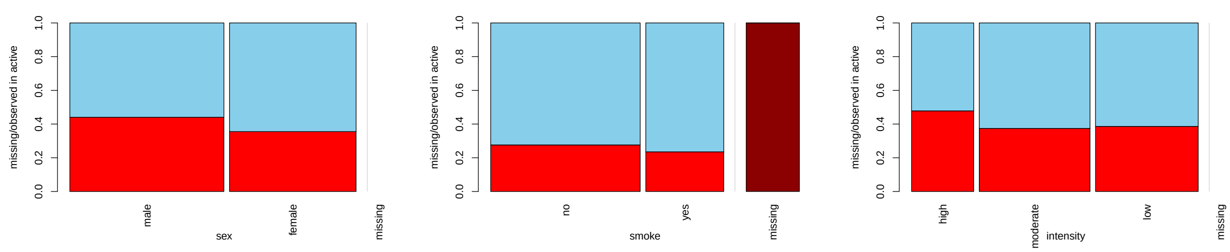
relatively close to the significance threshold of 0.05.

Considering the categorical data, the missingness of active on sex has no significant difference, where $x^2 = 1.957$, $p = 0.162$. For the missingness of active on smoke no significant difference was found also $x^2 = 0.293$, $p = 0.589$. Lastly, the missingness of weight on intensity is not significant: $x^2 = 2.193$, $p = 0.334$.

All results are insignificant, hence active is not considered to be missing at random.

The three bar plots in Figure 7 show a visualization of how the missing data in the categorical columns is divided. The first plot shows us that the female category in sex has less missing values in the active column than the male category. The second column shows that smokers have fewer missing values than non-smokers in the active column. The third column shows almost no difference between missing values in the active column for the moderate and low-intensity category but more missing values in the high-intensity category.

The five scatterplots in 8 show a visualization of how the missing data is divided into the rest of the columns. In the first two plots between active and rest or age, there is a clear trend where all the values with a low active are missing, and everything above that is not. The three plots after that between active and height, bmi, or weight show a clear trend where low values for either column result in missing values with a cluster where both columns have low values.



(a) Sex (b) Smoke (c) Intensity

Figure 7: Looking whether the missingness of active is MAR

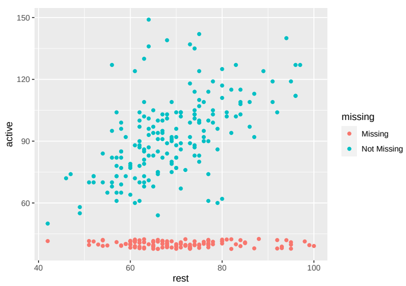
4.5 Missingness of Bmi

Looking at the differences of the missingness of bmi in Table 13, no significant differences can be found in the means of the numeric variables in the data.

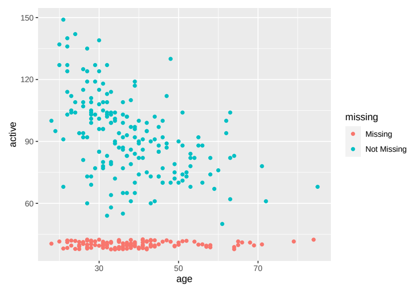
Considering the categorical data, the missingness of bmi on sex has no significant difference, where $x^2 = 0.019$, $p = 0.889$. For the missingness of bmi on smoke no significant difference was found also $x^2 = 0$, $p = 1$. Lastly, the missingness of bmi on intensity is not significant: $x^2 = 1.476$, $p = 0.478$.

All results are insignificant, hence bmi is considered not to be missing at random.

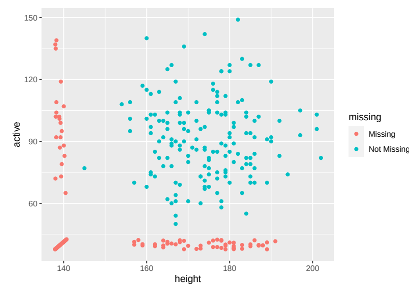
The three bar plots in Figure 9 show a visualization of how the missing data in the categorical columns is divided. The first plot shows us almost no difference between missing values in bmi for being a man or



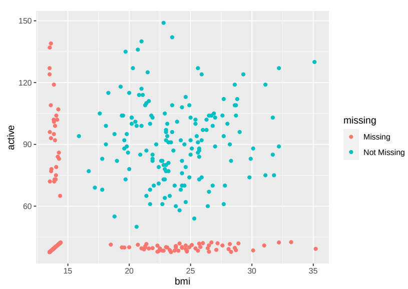
(a) Rest



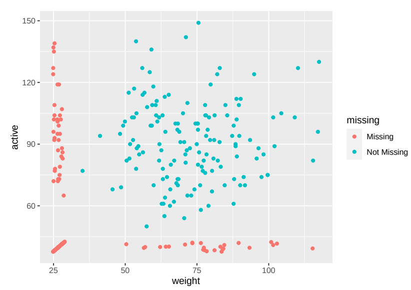
(b) Age



(c) Height



(d) Bmi



(e) Weight

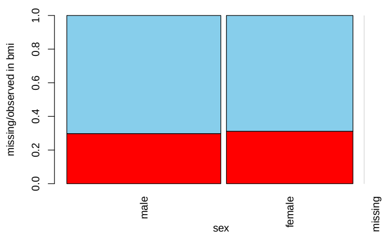
Figure 8: Looking whether the missingness of active is MAR

Table 13: Difference in means of missingness of Bmi

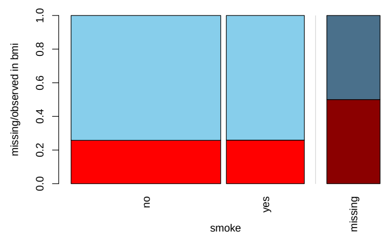
variables	$M obs	M true	t	p
rest	69.84	69.81	0.021	0.983
age	38.35	38.90	-0.368	0.713
height	174.28	175.39	-0.639	0.525
active	92.12	95.11	-0.717	0.478

female in the sex column. The second plot also indicates practically no difference between missing values in the bmi column for smokers and non-smokers in the smoke column. The third column shows almost no difference between missing values in the bmi column for the moderate and low-intensity category, but more missing values in the high-intensity category.

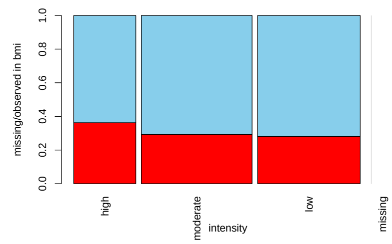
The four scatterplots in Figure 10 show a visualization of how the missing data is divided into the rest of the columns. In the first two plots between bmi and rest or age there is a clear trend where all the values with a low bmi are missing and everything above that is not. The two plots after that between bmi and height or active show a clear trend where low values for either column result in missing values with a cluster where both columns have low values.



(a) Sex

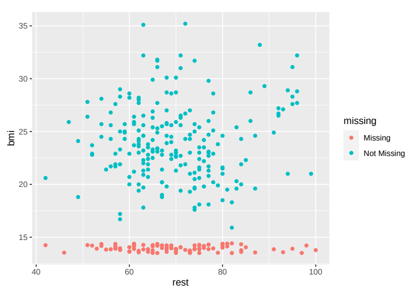


(b) Smoke

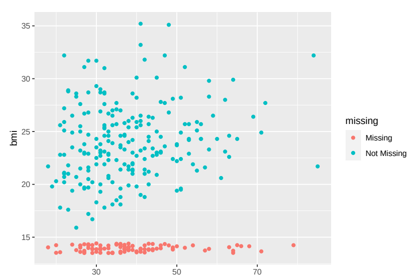


(c) Intensity

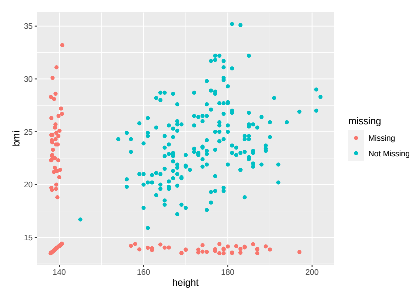
Figure 9: Looking whether the missingness of bmi is MAR



(a) Rest



(b) Age



(c) Height

(d) Active

Figure 10: Looking whether the missingness of bmi is MAR

4.6 Missingness of Smoke

Looking at the differences of the missingness of smoke in Table 14, no significant differences can be found in the means of the numeric variables in the data.

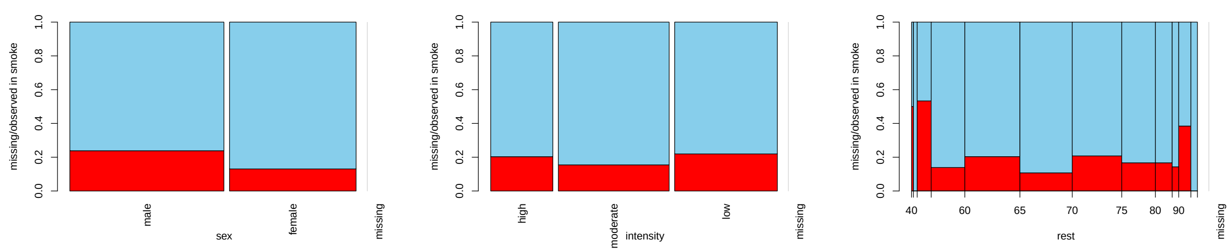
Table 14: Difference in means of missingness of Smoke

variables	M obs	M true	t	p
rest	70.08	68.72	0.779	0.438
age	38.03	40.57	-1.271	0.208
height	174.41	175.12	-0.347	0.731
bmi	24.00	24.85	-1.338	0.188
weight	73.74	76.13	-0.785	0.444

Considering the categorical data, the missingness of weight on sex has a significant difference, where $x^2 = 5.037$, $p = 0.025$. The missingness of smoke on intensity is not significant: $x^2 = 1.722$, $p = 0.423$.

The results indicate that missingness of smoke is missing at random in relation with sex.

The seven-bar plots in Figures 11 and 12 show how the missing data of smoke is divided into the other columns. The first plot shows us that the female category in sex has fewer missing values in the active column than the male category. The second plot shows almost no difference between missing values in the intensity column for the high and low category, but fewer missing values in the moderate-intensity category. The third plot shows that the missingness of smoke on rest is equally divided with two spikes where rest is lower than 55 and higher than 90. There are no more missing values after these spikes, except for one more spike where the rest is 40. The fourth plot shows that the missingness of smoke on age is equally divided with a spike where age is higher than 60. The fifth plot shows how smaller or higher (than a height of 170-175) the height gets, the more missing values there are in the smoke column, except for when the height is around 205. Then there are close to no missing values. The sixth plot shows that the missingness of smoke on bmi is equally divided except for when bmi is at its lowest or highest. Then there are almost no missing values. The seventh plot shows that the missingness of smoke on weight is equally divided, with one spike in the middle between weight of 70 and 80. There are almost no missing values when weight is at its lowest or highest.

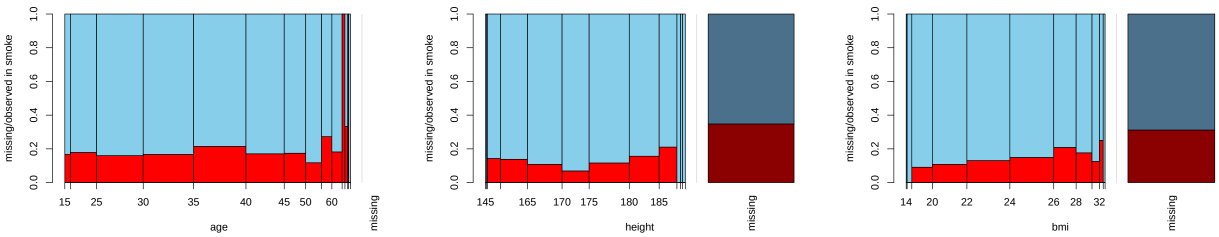


(a) Sex (b) Intensity (c) Rest

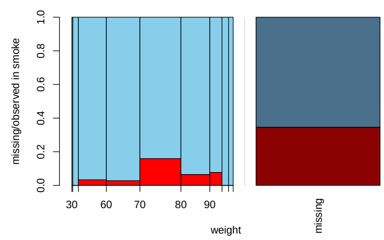
Figure 11: Looking whether the missingness of smoking is MAR

5 Imputations

After observing the data for missing values, we now try to solve the missingness problem by doing multiple imputation with the package `mice`. In order to answer the research question with the imputed data we will follow the main steps in multiple imputation following van Buuren, 2018, shown in Figure 13. In first instance we will use the default settings to impute the missingness, this will be further elaborated in the Default Imputations section. After the the default imputations we will evaluate the quality of the imputations by examining multiple plots about the convergence and the distribution. This will be done in the Evaluating Default Imputations section. After evaluating the default imputations, we will make an outline of how



(a) Age (b) Height (c) Bmi



(d) Weight

Figure 12: Looking whether the missingness of smoking is MAR

Table 15: Predictor Matrix of Default Imputations

	age	smoke	sex	intensity	active	rest	height	weight	bmi
age	0	1	1	1	1	1	1	1	1
smoke	1	0	1	1	1	1	1	1	1
sex	1	1	0	1	1	1	1	1	1
intensity	1	1	1	0	1	1	1	1	1
active	1	1	1	1	0	1	1	1	1
rest	1	1	1	1	1	0	1	1	1
height	1	1	1	1	1	1	0	1	1
weight	1	1	1	1	1	1	1	0	1
bmi	1	1	1	1	1	1	1	1	0

the imputations can be improved, by using more sophisticated imputations. This will be discussed in the Improving the Imputations section.

5.1 Default Imputations

As mentioned before we will first use the default settings of `mice` to impute the missingness. In this section we will describe them. By default, `mice` will produce 5 imputations and 5 iterations. **Say here something about the imputations and iterations (what do they do and why do we use them?)**

Say something about the predictor matrix:

Table 15 gives an overview of the predictor matrix for the imputations.

method weight = pmm

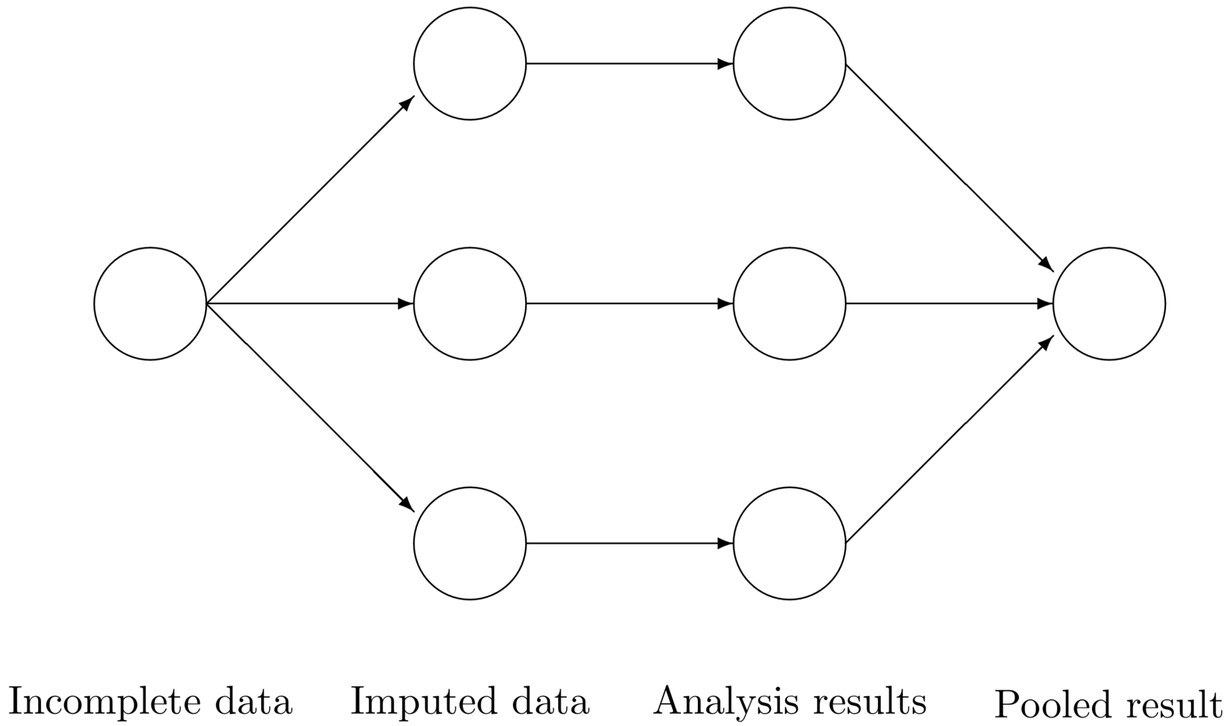
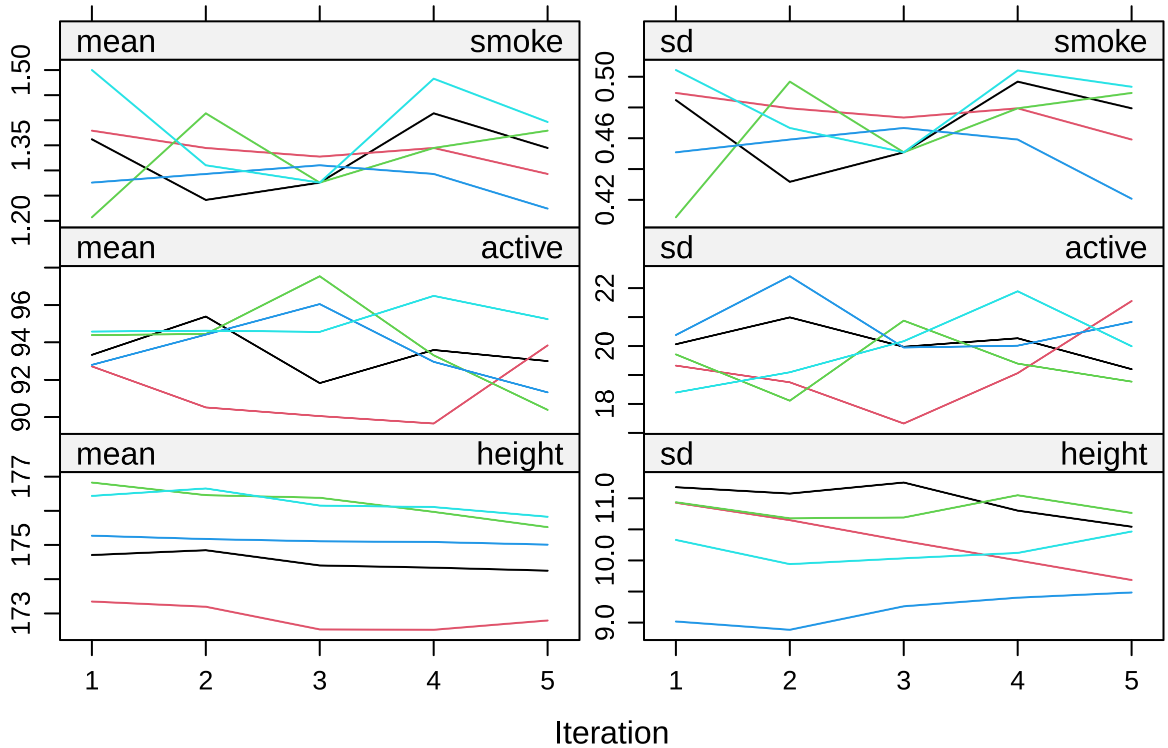
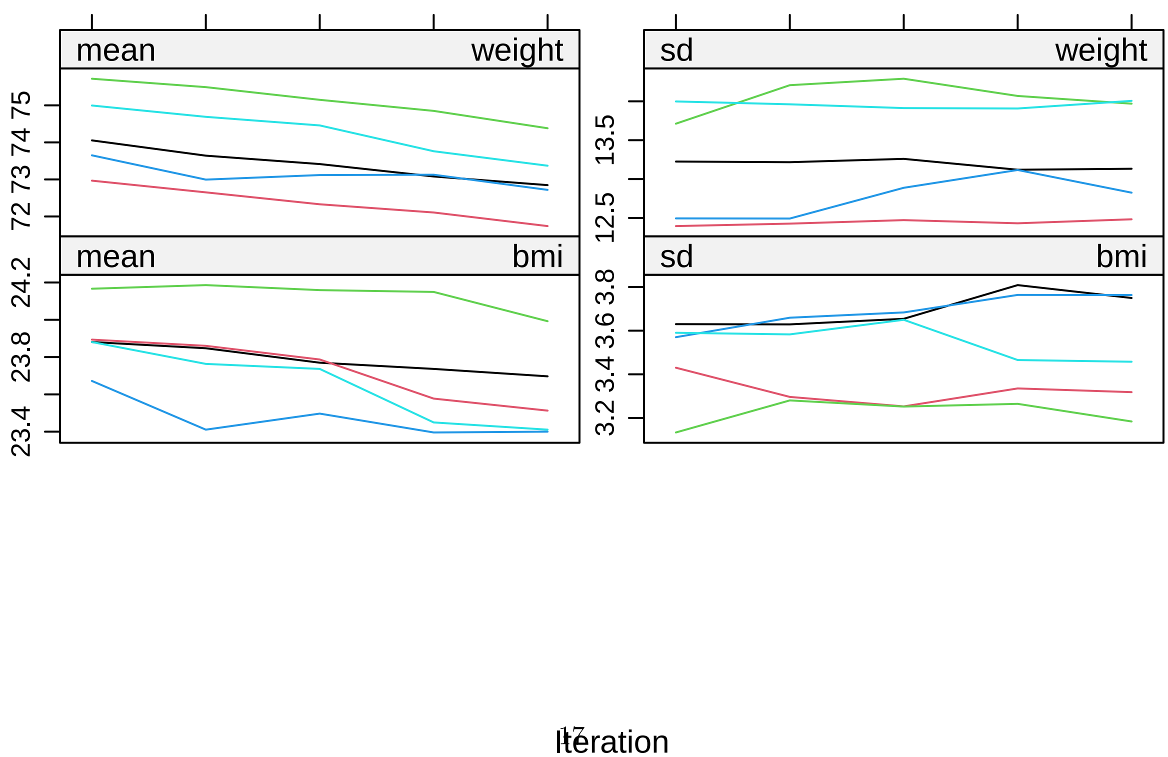


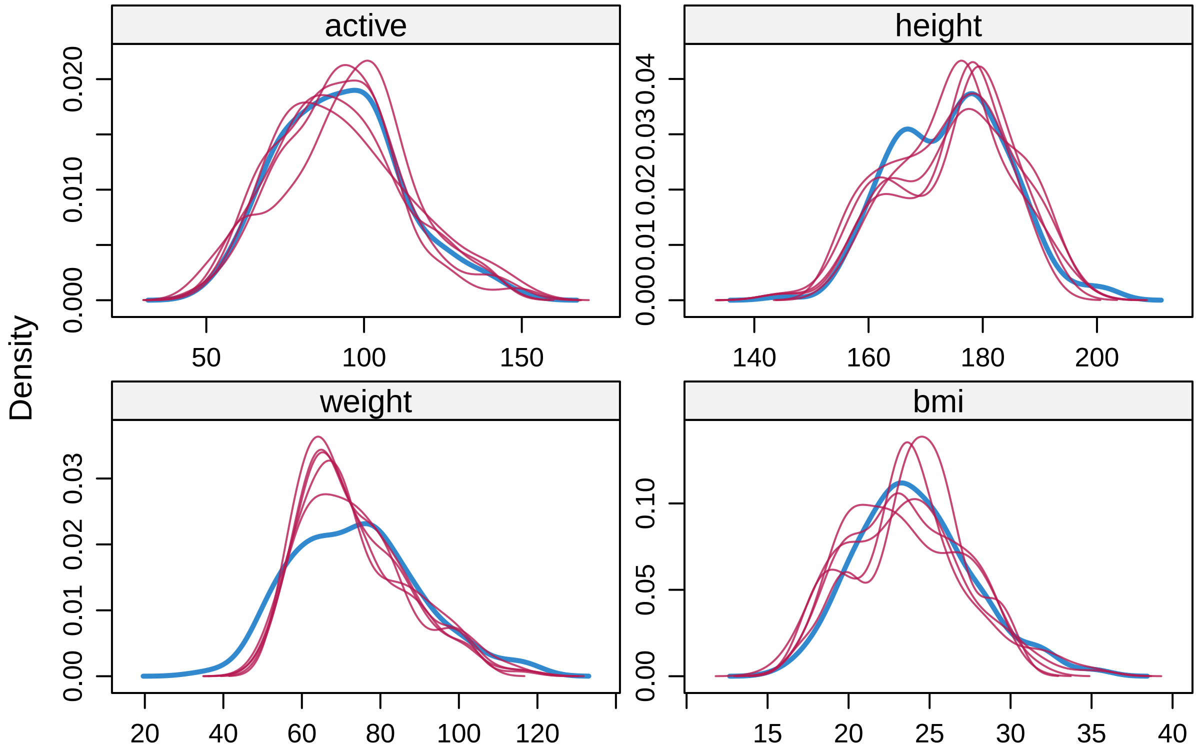
Figure 13: Scheme of main steps in multiple imputation

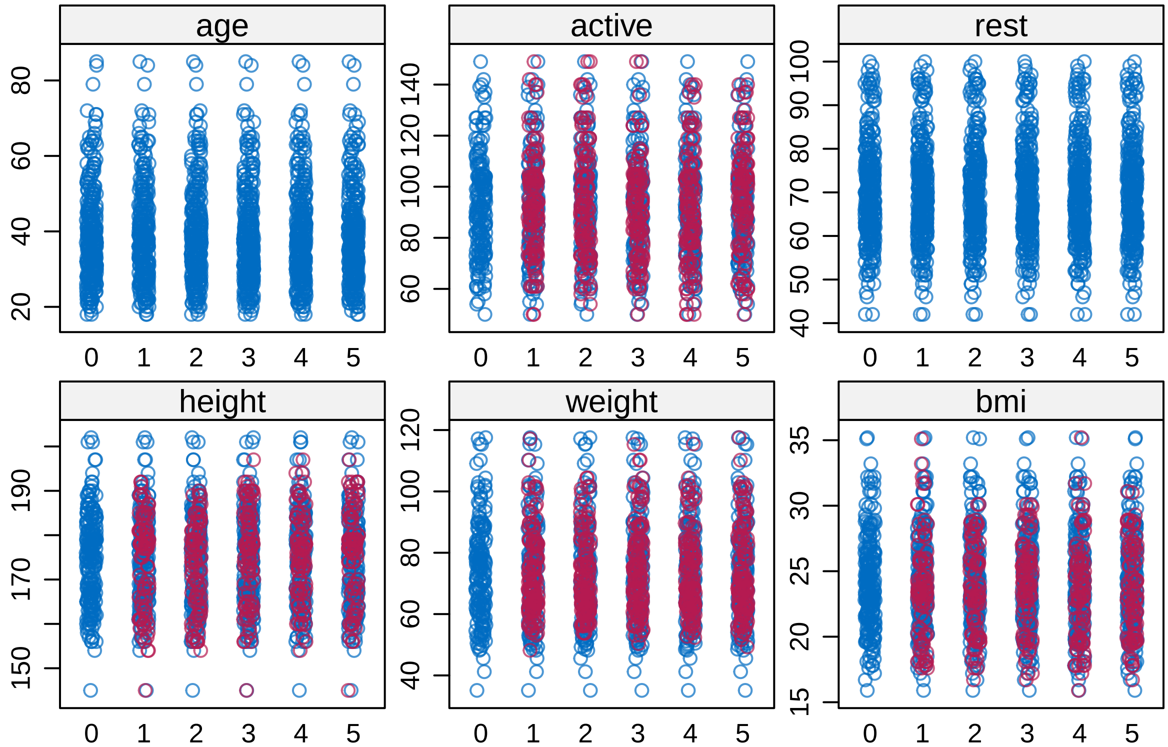
5.2 Evaluating Default Imputations

5.3 Improving the Imputations






active
height
weight
bmi
Density
0.000
0.010
0.020
50
100
150
0.000
0.010
0.020
0.030
0.04
140
160
180
200
0.00
0.01
0.02
0.03
20
40
60
80
100
120
0.00
0.05
0.10
15
20
25
30
35
40


age
active
rest
height
weight
bmi