MDCE Assignment C, Group 9

Quintijn de Leng - 6829376 Alec Noppe - 6947794 Guglielmo de Santis - 6664652
Anna Teixeira Rodeia - 6263747

20-4-2022

Introduction

In this report the attempt to predict active heart rate in an incomplete dataset, using multiple imputation. Our goal is to deduce whether the imputed dataset can produce similar regression results as the complete dataset, and that there is not a consistent flaw in the imputation. To do this, we compare the prediction with the complete dataset.

In Assignment A we saw that after a recalculation of cases in variable *weight* that were not actually missing, 16.8% of the data was missing. The missing data pattern was connected and not monotone.

Also referring back to Assignment A, we consider the Missing at Random-assumption to be plausible because of a non-significant Little's test and relatively close estimations of means and variances under listwise deletion.

Modeling Decisions

Imputation model

Predictive Mean Matching (pmm) is used for the numeric variables *active*, *height* and *weight*. For the binary variable *smoke*, logistic imputation regression is used.

BMI is a derived variable and thus, passive imputation was used to preserve the relationship between *height* and *weight* and to fill in the missing data of BMI.

Predictors

As there are relatively few variables in this dataset, all remaining variables are used to imputate any variable, with the exception of *BMI* on *weight* and *height* to prevent circularity. The loss of information in the imputations for the 49 cases where *BMI* is known, but not *height* and *weight* is not considered of any importance, since only *BMI* is used in the scientifically interesting model.

Imputation order

The default order for imputations is used, as there is no clear reason to specify this otherwise. The missing data are not clearly close to monotone and thus do not benefit from this order and derived variable *BMI* already follows *height* and *weight* and thus the updated value is already taken into account in the following iteration.

Number of iterations

5 iterations were chosen as the default to start with. We did not see a convergence of the model yet, so this was increased to 20; after which the model showed convergence (see Appendix, figure 1). This may be due to some obvious correlations between variables: *height* and *weight* are obviously related, and women are generally also shorter and have a lower weight.

Number of imputed datasets

For the amount of imputations we followed the guideline to set this amount equal to the percentage of missing values (16.8%), rounded up to 17. We tried some higher and lower imputations as well, but this did not make a big impact on the plots. Because of this, we choose to keep the number of imputations equal to the percentage of missing values.

Final imputation model

Taking all of the former into account, the final imputation model looks like this:

```
ini <- mice(dfinc, maxit = 0)
meth <- ini$method
meth["bmi"] <- "~ I(weight / (height / 100)^2)"
pred <- ini$predictorMatrix
pred[c("weight", "height"), "bmi"] <- 0

imp1 <- mice(dfinc,
             m = 17,
             maxit = 20,
             method = meth,
             predictorMatrix = pred,
             seed = 1337,
             print = FALSE)
```

Diagnostic plots of the imputations are offered in the Appendix. All point to a convergence of the model and realistic imputation outcomes.

Means

In this section, means of the imputed, complete and incomplete dataset are compared. Only imputed numeric variables shown. As you cannot simply average the column means for each imputation m, the following calculation is used to come to the means:

```
impmeans <- complete(imp1, "all") %>%
            lapply(function(x) select(x, where(is.numeric)) %>% colMeans()) %>%
            do.call(rbind, .) %>%
            colMeans()
#From: Lang (2022, 3 December)
```

Table 1: Table of means in imputed, complete and incomplete dataset

	imputed	complete	listwise
active	93.13	93.13	93.95
height	174.58	173.99	173.97
weight	74.23	73.58	72.99
bmi	24.08	24.06	24.02

The mean of active heart rate in the imputed dataset is equal to the mean of the complete data. For *height*, the imputed dataset is actually further off than the listwise deletion method, as is slightly the case with *weight* (but overerstimating it instead of underestimating). However, this is still a minor difference. The imputed mean of *bmi* is closer to the complete data than the listwise-method is, although differences are very minor.

Variances

In this section, the variance of the imputed, complete and incomplete dataset are compared. Only imputed numeric variables shown.

Variances for the imputed datasets are calculated as follows:

```
impdat <- complete(imp1, action="long", include = FALSE)
pool_var <- with(impdat, by(impdat, .imp, function(x) c(var(x$active),
                                                        var(x$height),
                                                        var(x$weight),
                                                        var(x$bmi))))
pool_var <- Reduce("+", pool_var) / length(pool_var)
#Adapted from: Heymans & Eekhout (2019)
```

Table 2: Table of variances for imputed, complete and incomplete data

	imputed	complete	listwise
active	389.88	378.04	394.94
height	108.74	105.29	107.97
weight	288.31	274.85	272.06
bmi	13.48	13.38	13.41

Unsurprisingly due to the added uncertainty of the missingness of the data, the variances are higher in the imputed datasets. However, the amount of added variance is relatively low. In the case of listwise deletion, an underestimation of the variance can even be seen in the event of *weight*.

Correlations

In this section, matrices of the difference in correlations between the imputed and complete respectively the incomplete and complete datasets are shown.

Table 3: Differences in correlations between the imputed and complete dataset

	age	active	rest	height	weight	bmi
age	0.00	-0.01	0.00	0.04	-0.02	-0.04
active	-0.01	0.00	0.00	-0.01	-0.02	-0.03
rest	0.00	0.00	0.00	0.02	0.02	0.02
height	0.04	-0.01	0.02	0.00	0.01	0.00
weight	-0.02	-0.02	0.02	0.01	0.00	-0.08
bmi	-0.04	-0.03	0.02	0.00	-0.08	0.00

Table 4: Differences in correlations between the incomplete and complete dataset

	age	active	rest	height	weight	bmi
age	0.00	-0.01	0.07	-0.08	-0.12	-0.10

	age	active	rest	height	weight	bmi
active	-0.01	0.00	0.00	0.05	0.04	0.02
rest	0.07	0.00	0.00	0.07	0.10	0.09
height	-0.08	0.05	0.07	0.00	0.01	0.02
weight	-0.12	0.04	0.10	0.01	0.00	0.00
bmi	-0.10	0.02	0.09	0.02	0.00	0.00

Looking at the differences between correlations, we can see that the imputed dataset slightly underestimates the correlation between *weight* and *height*, perhaps because *bmi* was not used in their imputations. Overall however, the correlations are much more preserved in the imputed dataset over listwise deletion.

Smoke frequencies

In this section, the difference in frequency of the *smoke* variable are considered. Frequencies for the imputed dataset are calculated as follows:

```
pool_count <- with(impdat, by(impdat, .imp, function(x) summary(x$smoke)))
pool_count <- Reduce("+", pool_count) / length(pool_count)
#Adapted from: Heymans & Eekhout (2019)
```

Table 5: Relative frequencies for *smoke*

	imputed	complete	listwise
no	68.6	67.3	68.7
yes	31.4	32.7	31.3

In both listwise and MI the fraction of non-smokers is slightly overestimated. The differences are, however, very minor.

Scientifically Interesting Model

After having imputed the missing data, we want to investigate the performance of a linear model trained on the imputed data, as well as the model trained on the complete data. The goal of this section is to deduce whether the imputed dataset can produce similar regression results as the complete dataset, and that there is not a consistent flaw in the imputation.

Active heart rate is dependent on many factors. For this model, *active heart rate* is dependent on *age*, *BMI*, *resting heart rate*, *gender*, *smoking*, *intensity* and an interaction between *BMI* and *age*. These are all the variables included in the dataset, with an additional interaction between BMI and age. This is suggested to increase the active heartrate exponentially, as opposed to a linear addition (Watkinson et al., 2010). The most influential variables are intensity, gender, and age. These variables have the highest slopes, as shown in the table below. The table also shows that we have lost quite some information to the missing data. Especially the rest stage, respondents with the gender female and respondents that smoke show a high fmi.

	estimate	SE	statistic	df	p	fmi
(Intercept)	81.24	22.07	3.68	119.92	0.00	0.26
age	-1.46	0.54	-2.69	106.70	0.01	0.29

	estimate	SE	statistic	df	p	fmi
rest	0.73	0.12	6.07	68.27	0.00	0.40
bmi	-0.57	0.88	-0.65	112.09	0.52	0.28
sex (female)	3.71	2.33	1.59	64.13	0.12	0.42
smoke (yes)	0.75	2.37	0.31	69.90	0.75	0.39
intenstiy (moderate)	-4.80	2.68	-1.79	101.18	0.08	0.30
intenstiy (low)	-4.20	3.06	-1.37	99.03	0.17	0.31
age:bmi	0.03	0.02	1.56	108.20	0.12	0.28

To assess a model's performance, it is important to evaluate the accuracy of the model on seen values (training validation) as well as the accuracy of the model on new data (testing validatieon). The reason for this is that a model may perform well on data that has been seen before, but not on new data. The metrics used to determine the performance of the model are the sum of squares error *SSE*, relative standard error *RSE*, root mean square error *RMSE* and *R^2*. For the former three metrics, a lower score is ideal. Whereas for the R^2, a score closest to 1 is ideal.

To assess a model's performance, it is important to evaluate the accuracy of the model on seen values (training validation) as well as the accuracy of the model on new data (testing validation). The reason for this is that a model may perform well on data that has been seen before, but not on new data. The metrics used to determine the performance of the model are the sum of squares error *SSE*, relative standard error *RSE*, root mean square error *RMSE* and *R^2*. For the former three metrics, a lower score is ideal. Whereas for the R^2, a score closest to 1 is ideal.

$$SSE = \sum_{i=1}^{n} (y_i - y_i')^2$$

$$RSE = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y_i})^2}$$

Training Validation

As seen below, the model trained on the imputed dataset produces less accurate results than the model trained on the complete dataset. However, both models have a R^2 score of 0.52 and 0.53 respectively. An R^2 of > 0.5, indicates quite good model fit. It shows that roughly 50% of the variation of *activity* is explained through the independent variables. This implies that there are some omitted variables that explain some of the activity variance. Nonetheless, the imputed dataset model performs similarly to the complete dataset model.

```
##                      SSE       RSE     RMSE        R2
## Imputed Data  755351.04 0.4825336 13.91510 0.5174664
## Complete Data  42796.64 0.4705243 13.67059 0.5294757
```

Testing Validation

The imputed dataset model has a slightly higher training accuracy than the complete dataset. It has a R^2 of 0.51, whereas the complete dataset model has a R^2 of 0.47. This difference is likely because the imputed data model uses far more training data than the complete dataset, as a result of the multiple imputations. Another possible explanation could be that the test set used to evaluate the complete model is too small to accurately indicate the testing accuracy. But these results do suggest that the imputed dataset model yields similar results to the complete dataset model. Thus, the multiple imputation as performed in this research successfully fills in the missing values to an extent that yields proper results for our scientifically interesting model.

```
##                      SSE       RSE     RMSE        R2
## Imputed Data  221928.72 0.4850111 13.06075 0.5149889
## Complete Data  12830.29 0.5346858 12.90842 0.4653142
```

Analysis of Variance

Finally, an Analysis of variance *ANOVA* analysis is performed on both datasets to conclude whether there is a significant difference in means between variables. ANOVA is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. In our dataset, the most significant variables in terms of mean scores are age and resting heart rate. These variables have the highest F-values and the lowest p-values. Other variables show noteworthy variance, but none have a p-score lower than the significance level (0.05). When comparing the two ANOVA regressions, the explained variance of the linear model trained on both datasets are remarkably similar. The residual variance is within 2% of each other, and thus we conclude that the multiple imputation produces a reasonable dataset.

ANOVA table for imputed data:

```
## Univariate ANOVA for Multiply Imputed Data (Type 2)
##
## lm Formula:  active ~ age + rest + bmi + sex + smoke + intensity + bmi*age
## R^2=0.5245
## ........................................................................
## ANOVA Table
##                     SSQ df1       df2 F value  Pr(>F)    eta2 partial.eta2
## age         35729.2362   1  49.67451 81.3876 0.00000 0.30046      0.38723
## rest        22394.6552   1 120.40460 74.5202 0.00000 0.18833      0.28371
## bmi          1322.6014   1  48.71537  2.3922 0.12841 0.01112      0.02286
## sex           911.5515   1  98.79478  2.4133 0.12351 0.00767      0.01587
## smoke         165.5943   1 301.82964  0.4158 0.51952 0.00139      0.00292
## intensity    1142.7960   2 134.60633  1.6923 0.18799 0.00961      0.01981
## age:bmi       707.5556   1 203.74651  2.3603 0.12601 0.00595      0.01236
## Residual    56539.5310  NA        NA      NA      NA      NA           NA
```

ANOVA table for complete data:

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## rest          1  42750   42750 232.627 <2e-16 ***
## bmi           1    466     466   2.538 0.1122
## sex           1   1033    1033   5.619 0.0184 *
## intensity     2    471     236   1.282 0.2791
## age           1  14711   14711  80.048 <2e-16 ***
## bmi:age       1   1107    1107   6.021 0.0147 *
```

```
## Residuals   298  54764     184
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Sources

- Chunk 9 (imputation model means): Lang, K. (2022, 3 december). *Missing Data Theory & Causal Effects, practical 6, sec. 2.2* [R code].
- Chunks 11, 16 (imputation model variances; *smoke* frequencies): Heymans, M.W. & Eekhout, I. (2019). *Applied Missing Data Analysis with SPSS and Rstudio*, section 5.2.2 [R code].
- Watkinson, C., van Sluijs, E. M., Sutton, S., Hardeman, W., Corder, K., & Griffin, S. J. (2010). Overestimation of physical activity level is associated with lower BMI: a cross-sectional analysis. *International journal of behavioral nutrition and physical activity, 7*(1), 1-9.

Appendix: diagnostic figures

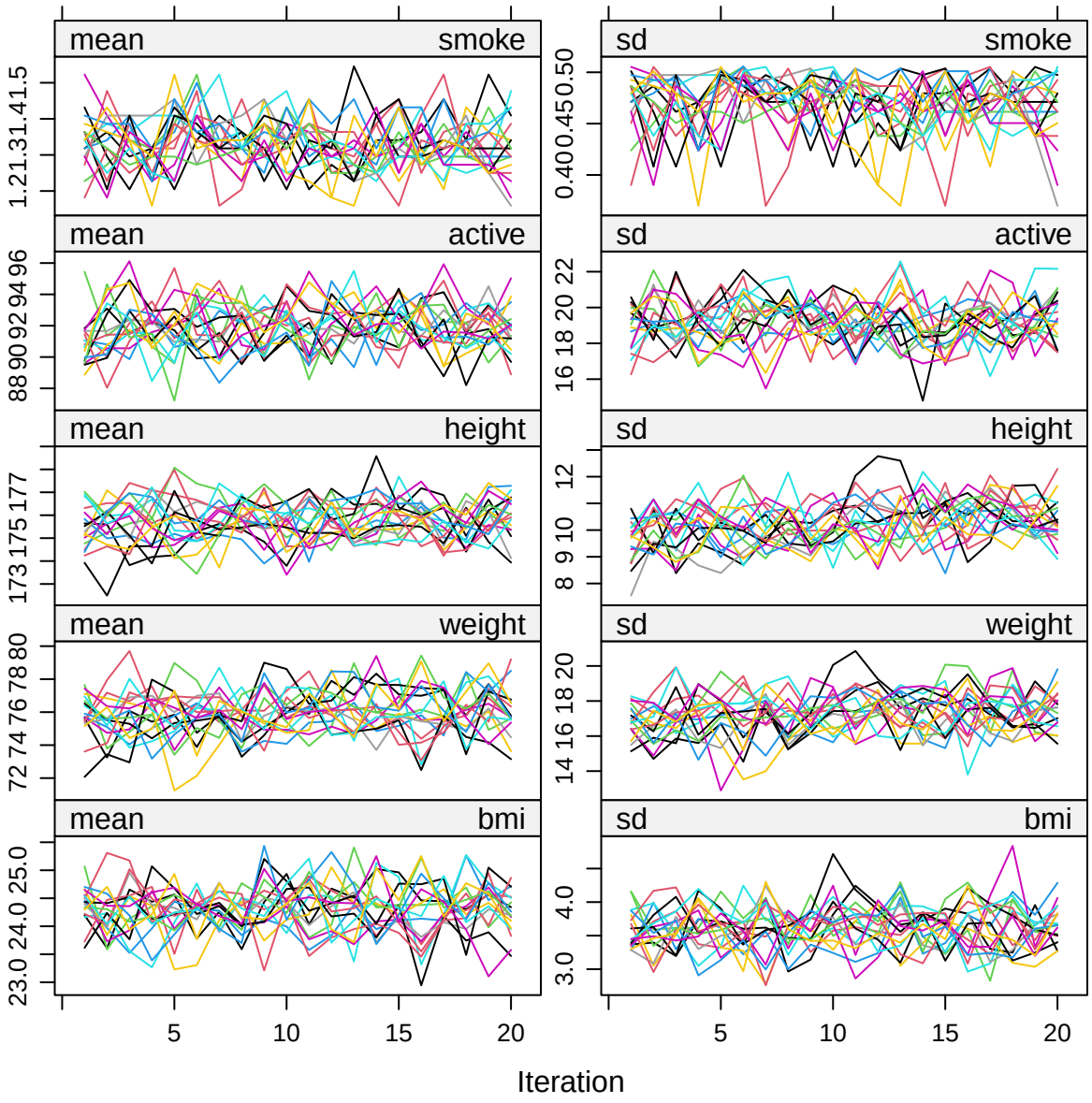

Figure 1: Convergence of imputed data

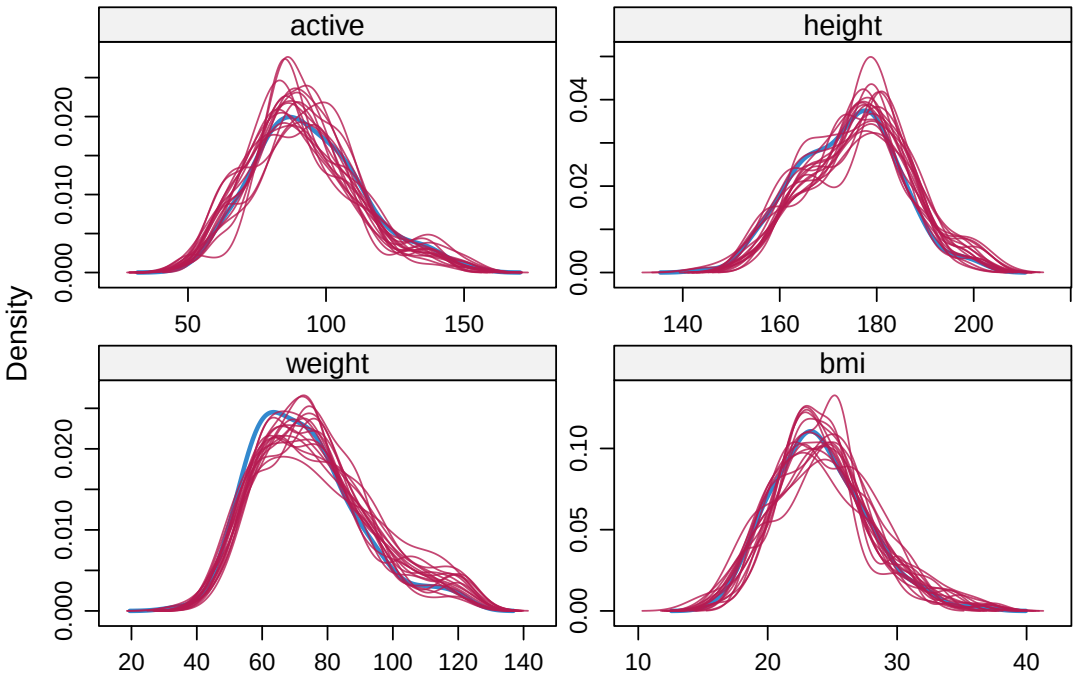

Figure 2: Density plot of imputed data

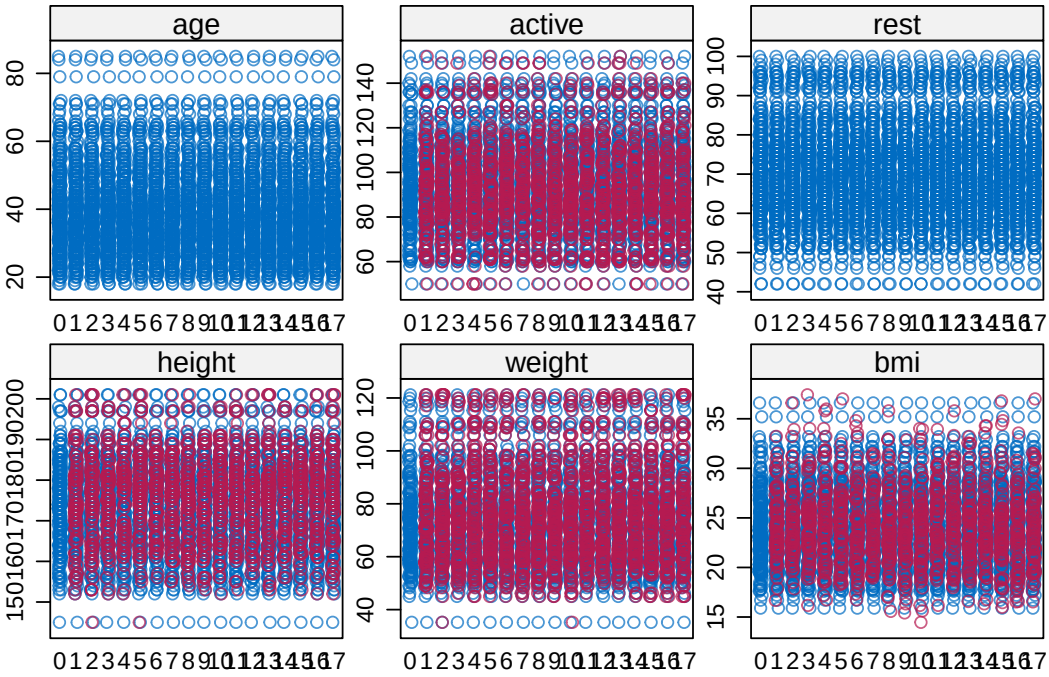

Figure 3: Stripplot of imputed data