# MDCE Assignment B, Group 9

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

1-4-2022

## Preparing the dataset for imputation

As in the last assignment, we re-calculate some missing cases of *weight* to reflect the actual non-missingness of these cases. This time, we avoid recalculating all cases of *weight* using the following syntax:

```
prep_ini <- mice(dfinc, maxit = 0)
prep_ini_meth <- prep_ini$method
prep_ini_meth[c("smoke", "active", "height", "bmi")] <- ""
prep_ini_meth["weight"] <- "~ I(bmi * (height / 100)^2)"
imp_prep <- mice(dfinc,
                 m = 1,
                 maxit = 1,
                 method = prep_ini_meth,
                 print = FALSE)
dfinc <- complete(imp_prep)
```

## Imputation

To impute the missing data, we used the mice algorithm. 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.

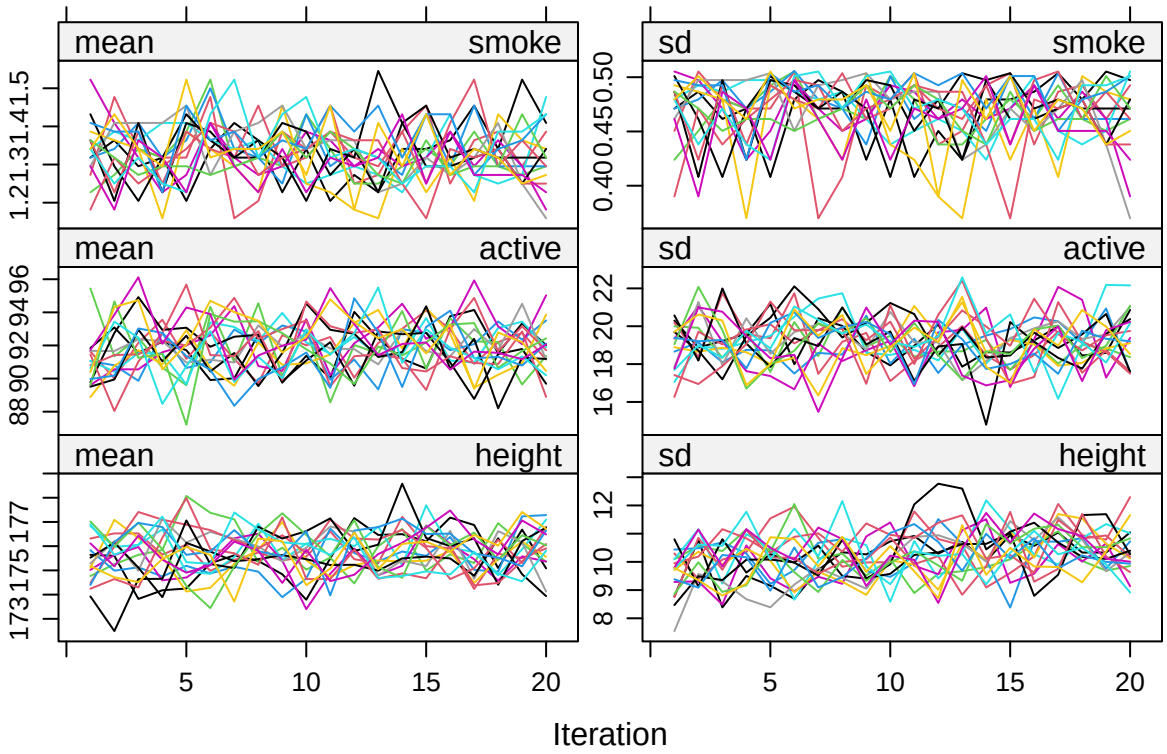
- For the binary variable *smoke*, logistic regression imputation was used.
- For the numeric variables *active*, *height* and *weight*, predictive mean matching (pmm) was used.
- *BMI* was imputed by passive imputation to preserve the relationsip between *height* and *weight*. To prevent any problems with circularity in the imputations, *bmi* is not used as a predictor for the imputations of *height* or *weight*.

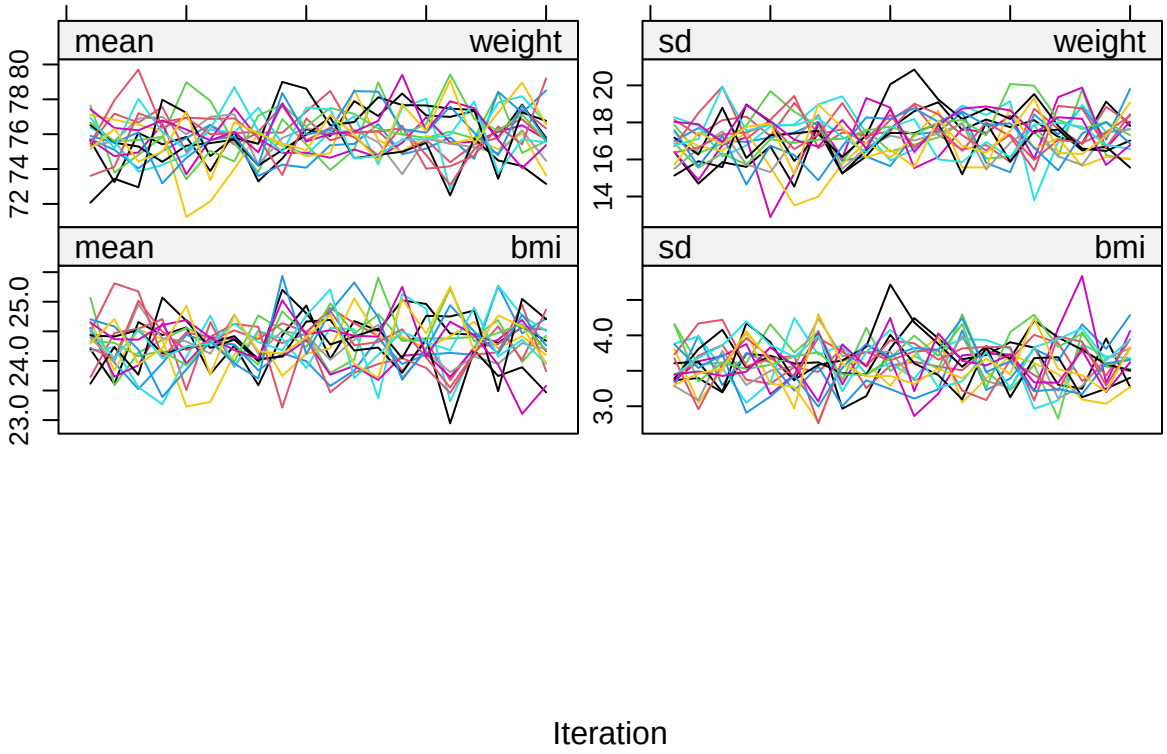
```
ini <- mice(dfinc, maxit = 0)
meth <- ini$method
meth["bmi"] <- "~ I(weight / (height / 100)^2)"
pred <- ini$predictorMatrix
pred[c("weight", "height"), "bmi"] <- 0

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

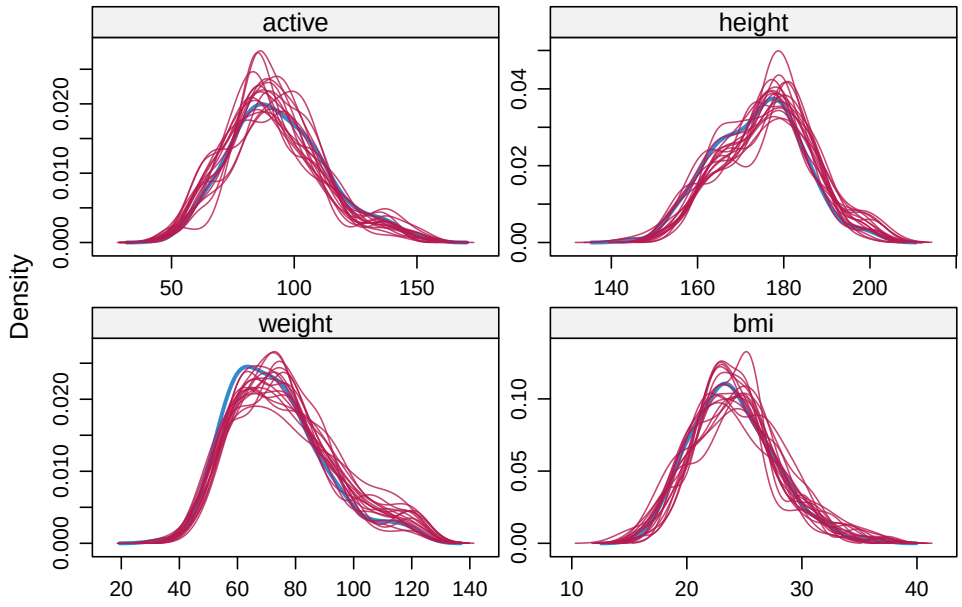
Using five iterations, we did not yet see a convergence in some variables. This may be because of the relatively high correlation between some variables, most notably (and obviously) *weight* and *height*. After 10 iterations, all variables showed convergence (see below).

## Diagnostics

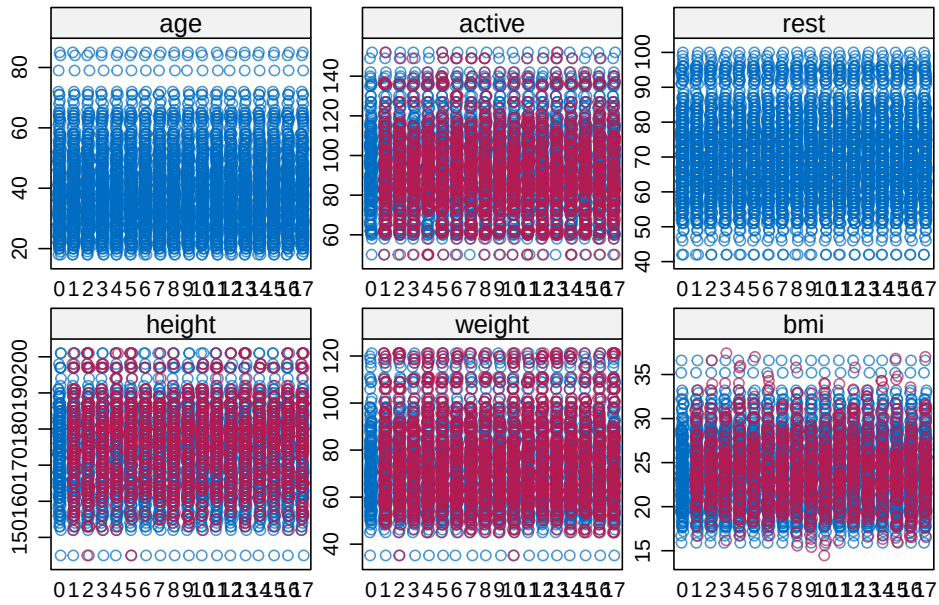


Density plot of imputed data:




Stripplot of imputed data:




These plots look good and show convergence. The stripplots show realistic values for the imputed data. *bmi* is the only variable that does not exactly match observed values due to the nature of passive imputation.

# Comparison of imputed and complete data

## 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 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 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.

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
```

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
## 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)
```

Frequency table for *smoke*

```
##       imputed complete listwise
## no        210      206      180
## yes        96      100       82
## total     306      306      262
```

The binary variable *smoke* is represented in a much better way than in the case of listwise deletion.

## 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.

```
##                       estimate    SE statistic     df    p
## (Intercept)              81.24 22.07      3.68 119.92 0.00
## age                      -1.46  0.54     -2.69 106.70 0.01
## rest                      0.73  0.12      6.07  68.27 0.00
## bmi                      -0.57  0.88     -0.65 112.09 0.52
## sex (female)              3.71  2.33      1.59  64.13 0.12
## smoke (yes)               0.75  2.37      0.31  69.90 0.75
## intenstiy (moderate)     -4.80  2.68     -1.79 101.18 0.08
## intenstiy (low)          -4.20  3.06     -1.37  99.03 0.17
## age:bmi                   0.03  0.02      1.56 108.20 0.12
```

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.

### 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 only 0.51 and 0.53 respectively. This means that only roughly 50% of the variation of *activity* is explained through the independent variables. This implies that there are latent 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  44405.75 0.4882155 13.92521 0.5117845
## 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 0.468, whereas the complete dataset model has a R^2 of 0.465. The difference is too small to make any definitive claims on the relative testing accuracy between the two models. But this does 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. A final remark regarding the training and test scores is that both models show some signs of being overfitted to the training data. The R^2 of the models validated on training accuracy are up to 5% greater than the models validated on testing accuracy. This is a sign of overfitting.

```
##                     SSE       RSE     RMSE        R2
## Imputed Data  12751.41 0.5313985 12.86867 0.4686015
## 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.