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:

```r
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.

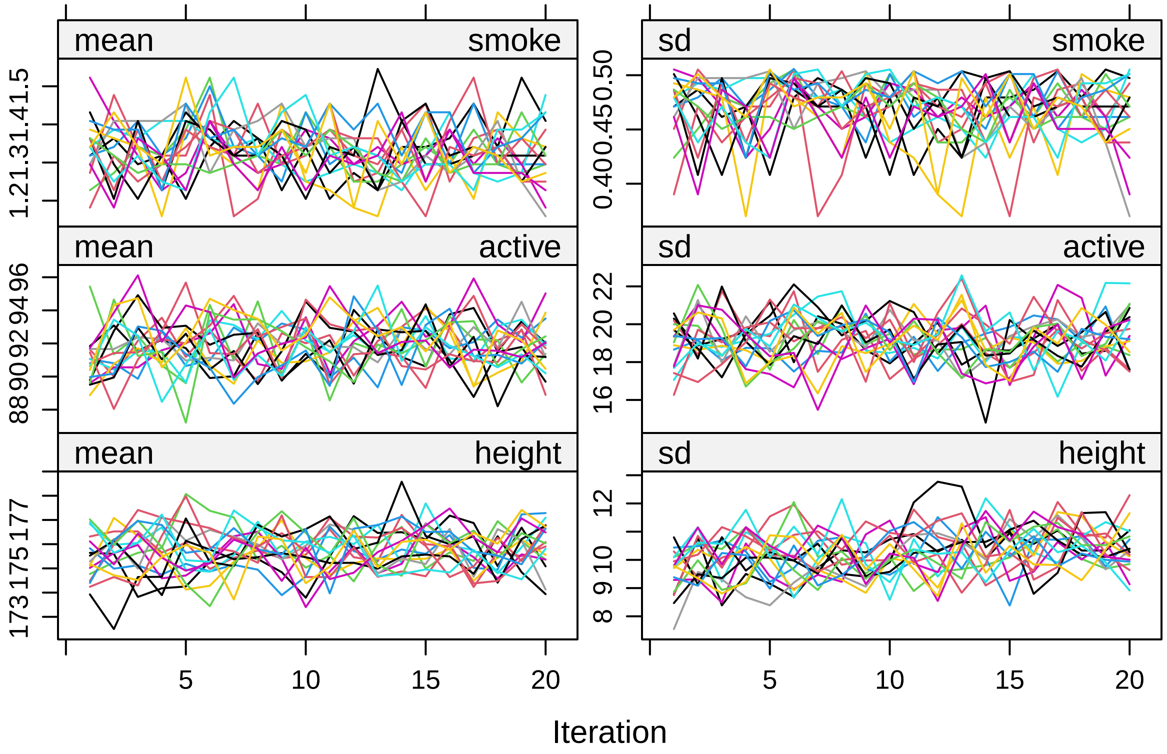
- For the binary variable *smoke*, logistic regression imputation was used.
- For the numeric variables *active*, *height* and *weight*, predictive mean matching (pmm) was used.
- *BMI* was imputed by passive imputation to preserve the relationsip between *height* and *weight*. To prevent any problems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

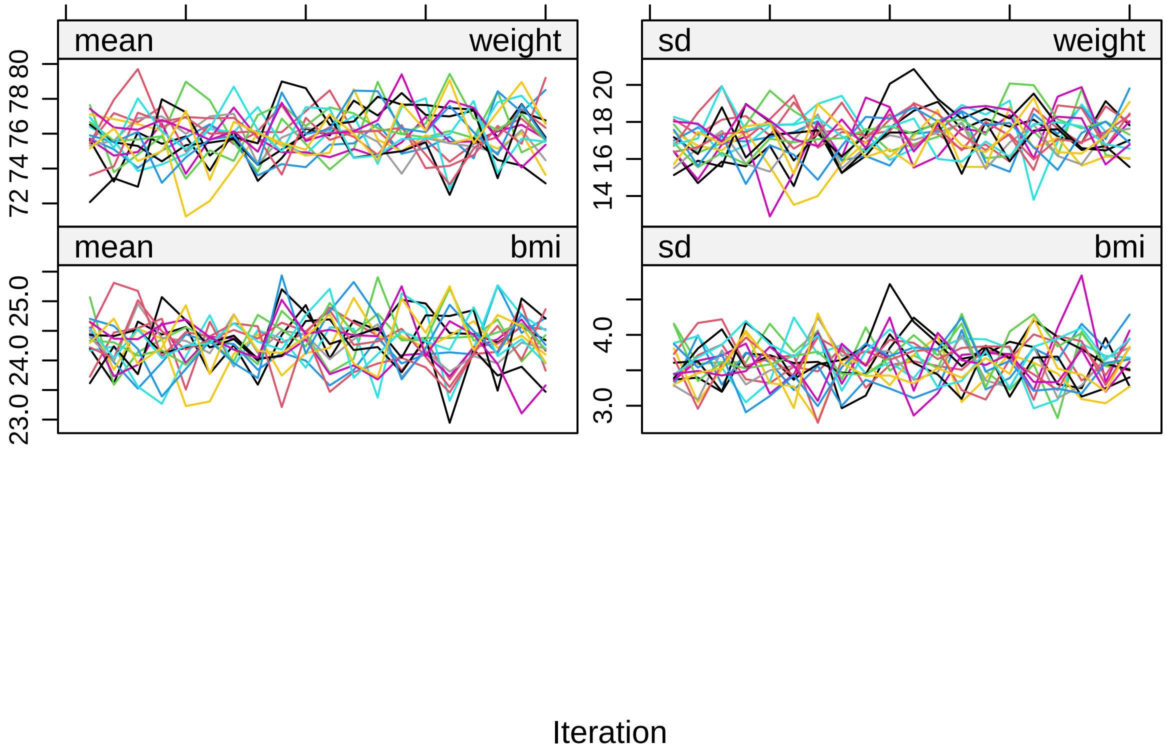
```r
ini <- mice(dfinc, maxit = 0)
meth <- ini$method
meth["bmi"] <- "~ I(weight / (height / 100)^2)"
pred <- ini$predictorMatrix
pred[c("weight", "height"), "bmi"] <- 0

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

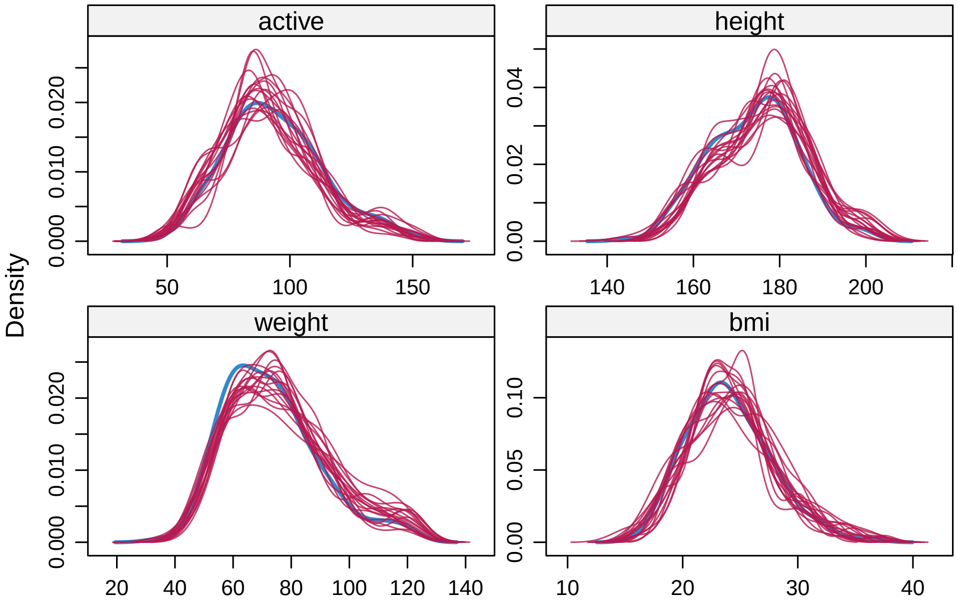
Using five iterations, we did not yet see a convergence in some variables. This may be because of the relatively high correlation between some variables, most notably (and obviously) *weight* and *height*. After 10 iterations, all variables showed convergence (see below).

Diagnostics

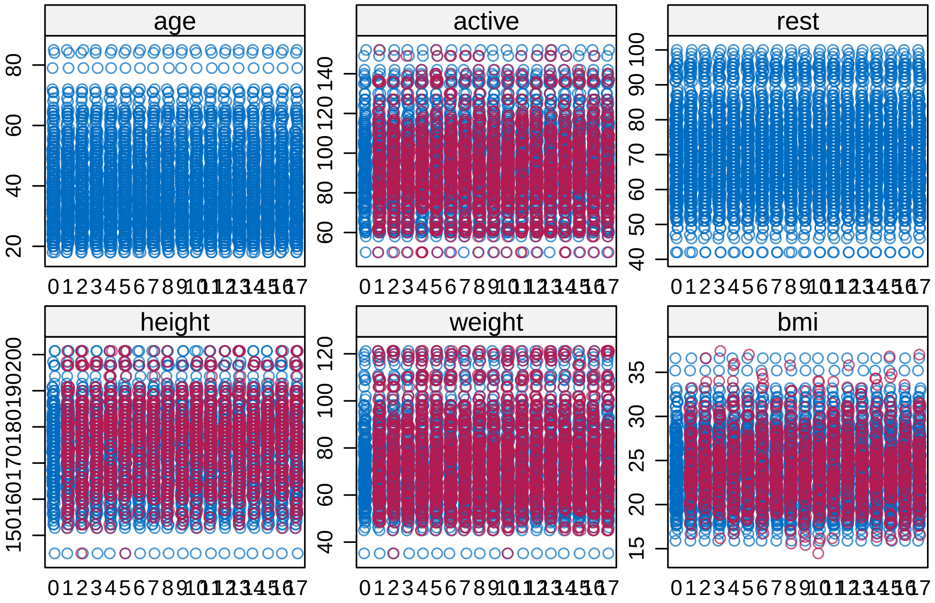




Density plot of imputed data:



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:

```r
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

In this section, the scientifically interesting model of predicting active heart rate by age, resting heart rate, bmi, sex, intensity of exercise and smoking is considered. The pooled imputed datasets, complete dataset and incomplete dataset are compared.

```
fit <- with(imp1, lm(active ~ age + rest + bmi + sex + smoke + intensity))
est <- pool(fit)
a <- as.data.frame(summary(est))
a[,2:6] <- round(a[,2:6], 2)
colnames(a) <- c("term", "estimate", "SE", "statistic", "df", "p")
rownames(a) <- c("(Intercept)", "age", "rest", "bmi", "sex (female)", "smoke (yes)", "intenstiy (modera
a[,2:6]
```

```
##                        estimate    SE statistic    df    p
## (Intercept)               51.53 13.02      3.96 67.27 0.00
## age                       -0.62  0.09     -6.81 70.42 0.00
## rest                       0.71  0.12      5.91 68.06 0.00
## bmi                        0.73  0.39      1.85 33.44 0.07
## sex (female)               3.80  2.33      1.63 65.07 0.11
## smoke (yes)                0.77  2.36      0.33 72.80 0.75
## intenstiy (moderate)      -5.04  2.71     -1.86 96.79 0.07
## intenstiy (low)           -4.38  3.08     -1.42 98.41 0.16
```

Running the regression analysis on the pooled imputed datasets, only *age* and *rest*, originally complete variables, are significant in the model.

```
compfit <- lm(active ~ age + rest + bmi + sex + smoke + intensity, data = dfcom)
summary(compfit)
```

```
## 
## Call:
## lm(formula = active ~ age + rest + bmi + sex + smoke + intensity, 
##     data = dfcom)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -46.913  -8.603  -1.169   8.048  53.358 
## 
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)    
## (Intercept)        45.01490   10.07505   4.468 1.12e-05 ***
## age                -0.63604    0.07182  -8.857  < 2e-16 ***
## rest                0.71625    0.09276   7.721 1.76e-13 ***
## bmi                 1.00892    0.25209   4.002 7.92e-05 ***
## sexfemale           3.48884    1.79999   1.938   0.0535 .  
## smokeyes           -0.15396    1.83029  -0.084   0.9330    
## intensitymoderate -4.86015    2.24370  -2.166   0.0311 *  
## intensitylow       -3.35048    2.56732  -1.305   0.1929    
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
```

```
## Residual standard error: 13.69 on 298 degrees of freedom
## Multiple R-squared:  0.5155, Adjusted R-squared:  0.5041
## F-statistic: 45.29 on 7 and 298 DF,  p-value: < 2.2e-16
```

Comparing to the complete model, we see that due to the added imputation uncertainty, the imputed dataset no longer recognizes *bmi* and *intensity (moderate)*

```
a <- miceadds::mi.anova(imp1, "active ~ age + rest + bmi + sex + smoke + intensity + bmi*age", type=2)
```

```
## 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
```

```
b <- summary(aov(active ~ rest + bmi + sex + intensity + bmi*age, dfcom))
a
```

```
## $r.squared
## [1] 0.5245324
##
## $anova.table
##                     SSQ df1       df2 F value   Pr(>F)     eta2 partial.eta2
## age          35729.2362   1  49.67451 81.3876 0.000000 0.300464     0.387230
## rest         22394.6552   1 120.40460 74.5202 0.000000 0.188327     0.283713
## bmi           1322.6014   1  48.71537  2.3922 0.128412 0.011122     0.022858
## sex            911.5515   1  98.79478  2.4133 0.123509 0.007666     0.015867
## smoke          165.5943   1 301.82964  0.4158 0.519518 0.001393     0.002920
## intensity     1142.7960   2 134.60633  1.6923 0.187988 0.009610     0.019812
## age:bmi        707.5556   1 203.74651  2.3603 0.126006 0.005950     0.012360
## Residual     56539.5310  NA        NA      NA       NA       NA           NA
##
## $type
## [1] 2
```

```
b
```

```
##               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
```

Code 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].