

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 impute 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 A, 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 overestimating 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
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. The code is shown in Appendix B.

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.

Table 6: Estimates

	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
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
intensity (moderate)	-4.80	2.68	-1.79	101.18	0.08	0.30
intensity (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 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.

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

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

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

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

Training Validation

As seen below, the model trained on the imputed dataset produces more accurate results than the model trained on the complete dataset. The best imputed dataset model has an R^2 of 0.55, compared to the complete dataset model's 0.53. 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 on known data.

Table 7: Training results

	SSE	RSE	RMSE	R2
Average (Imputed)	43726.70	0.45	13.35	0.55
Best (Imputed)	38118.40	0.40	12.47	0.60
Worst (Imputed)	49672.59	0.49	14.24	0.51
Complete	42796.64	0.47	13.67	0.53

Testing Validation

The best imputed dataset model has a high testing accuracy than the complete dataset. The table below shows the chosen metrics for the imputed datasets, as well as the complete dataset. The best imputed dataset has an R^2 score of 0.49, whereas the complete dataset model has a R^2 of 0.46. The average testing accuracy of the imputed model is significantly lower than the complete dataset model. It has an average R^2 of 0.37. This shows there is great variance between the different imputed sets. We expect the accuracy of the imputed model to be slightly lower, but the results as shown through these metrics can be attributed to multiple things. The test validation set may be too small to produce consistent results. A train-test split of 0.8-0.2 was chosen for this model. There are a total of 306 rows, thus the test validation set is only 61 rows long. It is also possible that there is an omitted variable that systematically influences the other variables, that has a larger effect on the outcome. A possible omitted variable could be whether the patient has any respiratory issues, since this is proven to lead to increased active heart rate. But the results do suggest that the best imputed dataset model yields similar results to the complete dataset model. The only relevant model is the one that performs best, so if we had to predict more unknown heart rates, we would use the best imputed dataset (R^2 of 0.49). 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. The model shows significant signs of overfitting to the training data. The training R^2 of the best imputed dataset and complete dataset are 0.10 and 0.06 higher respectively. To improve the fitting of the model, a future implementation could use K-Fold cross validation. This allows for an easier approach to maximize the testing accuracy, since you have more subsets to compare.

Table 8: Testing results

	SSE	RSE	RMSE	R2
Average (Imputed)	13056.91	0.63	14.57	0.37
Best (Imputed)	10331.32	0.51	13.01	0.49
Worst (Imputed)	19101.37	0.86	17.70	0.14
Complete	12830.29	0.53	12.91	0.47

Analysis of Variance

Finally, an Analysis of variance *ANOVA* analysis is performed on both datasets to conclude whether there is a difference in relative variable importance between the imputed dataset model, and the complete dataset model. ANOVA is a statistical technique that is used to check if there are dominant variables in the chosen model. Each line represents a variable, and the metrics show to what extent it influences the outcome of the regression. In the imputed dataset, the most significant variables in terms of F-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). In the complete dataset, there are four variables under the significance level. These variables are: age, rest, gender, and the interaction between age & bmi. Since the imputed dataset only finds two significant variables, we conclude that some information is systematically misinterpreted by the imputations. We know that some information is lost, because the complete dataset has a lower residual sum of squares ($54764 < 56539$). However, the residual sum of squares is within 2% of each other. This shows that the imputed dataset model can produce similar results to the complete dataset model, regardless of the difference in variable importance. An explanation for this could be there is an omitted variable that influences the observed variables in a systematic way that is captured partially by the imputed dataset, and partially by the complete dataset. This would explain why there are different variable importances, but a similar residual sum of squares. An example used before is the presence of respiratory issues, which are not dependent on age or resting heart rate, but do inflate the active heart rate. The imputations may inflate a different variable to adjust for this unexplained variance, which results in the different variable importance shown below.

Table 9: ANOVA table for imputed data

	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

Table 10: ANOVA table for complete data

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
rest	1	42749.9276	42749.9276	232.626662	0.0000000
bmi	1	466.4327	466.4327	2.538126	0.1121867
sex	1	1032.6147	1032.6147	5.619044	0.0184027
intensity	2	471.0806	235.5403	1.281709	0.2790882
age	1	14710.5338	14710.5338	80.048378	0.0000000
bmi:age	1	1106.5605	1106.5605	6.021425	0.0147054
Residuals	298	54763.6214	183.7705	NA	NA

Sources

- Chunk 6 (imputation model means): Lang, K. (2022, 3 december). *Missing Data Theory & Causal Effects, practical 6, sec. 2.2* [R code].
- Chunks 8, 13 (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.
- Chunk 23 (Splitting the imputed models): Lang, K. (2022, 15 april). *Missing Data Theory & Causal Effects, miPredictionRoutines.R* [R code].

Appendix A: diagnostic figures

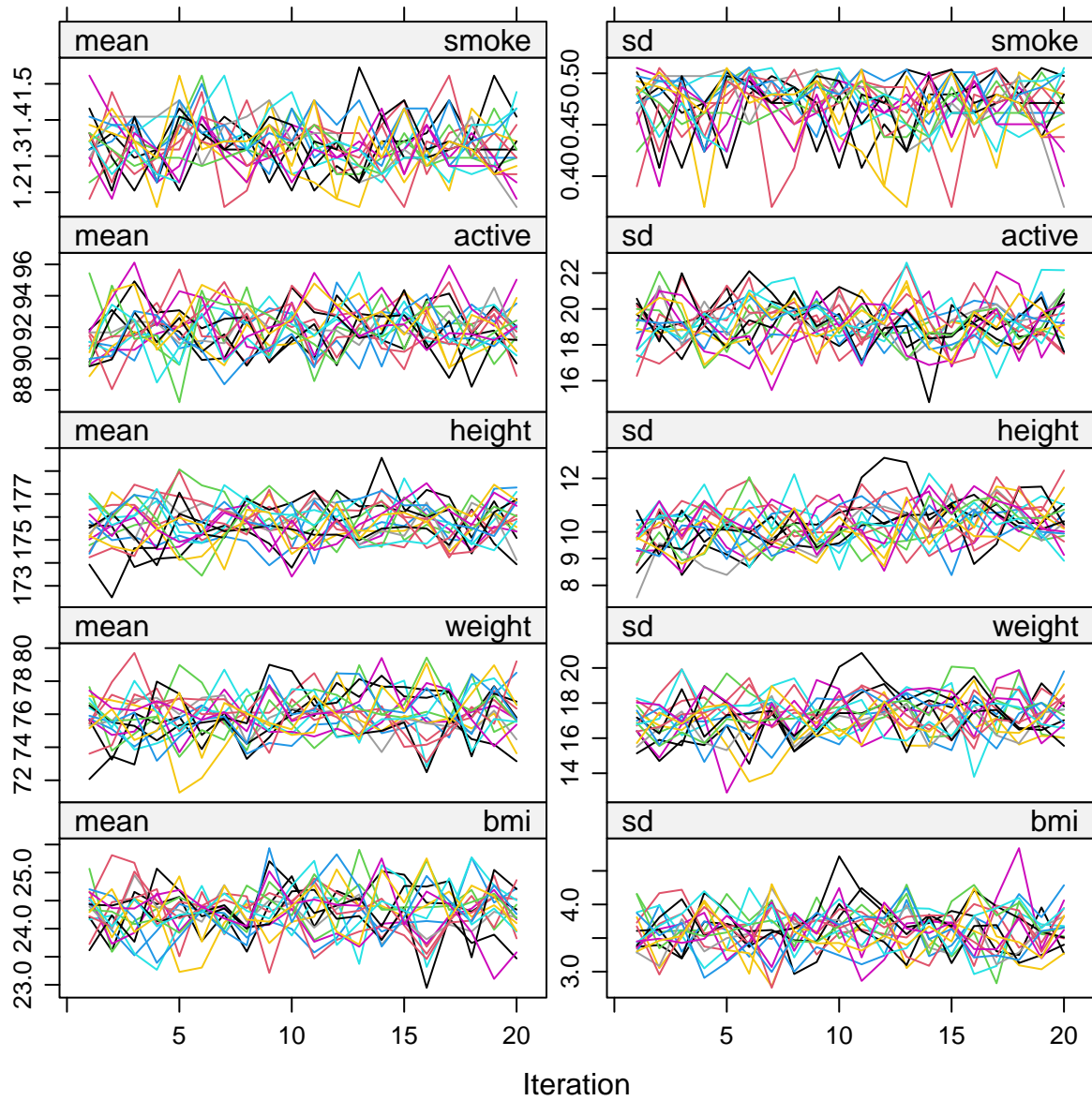


Figure 1: Convergence of imputed data

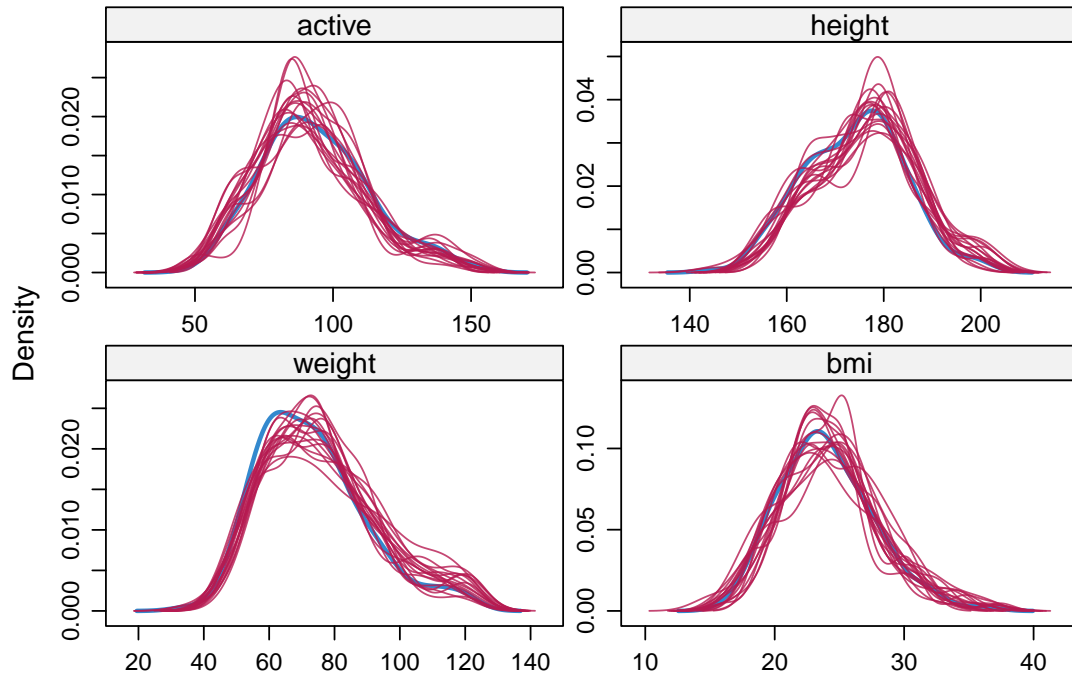


Figure 2: Density plot of imputed data

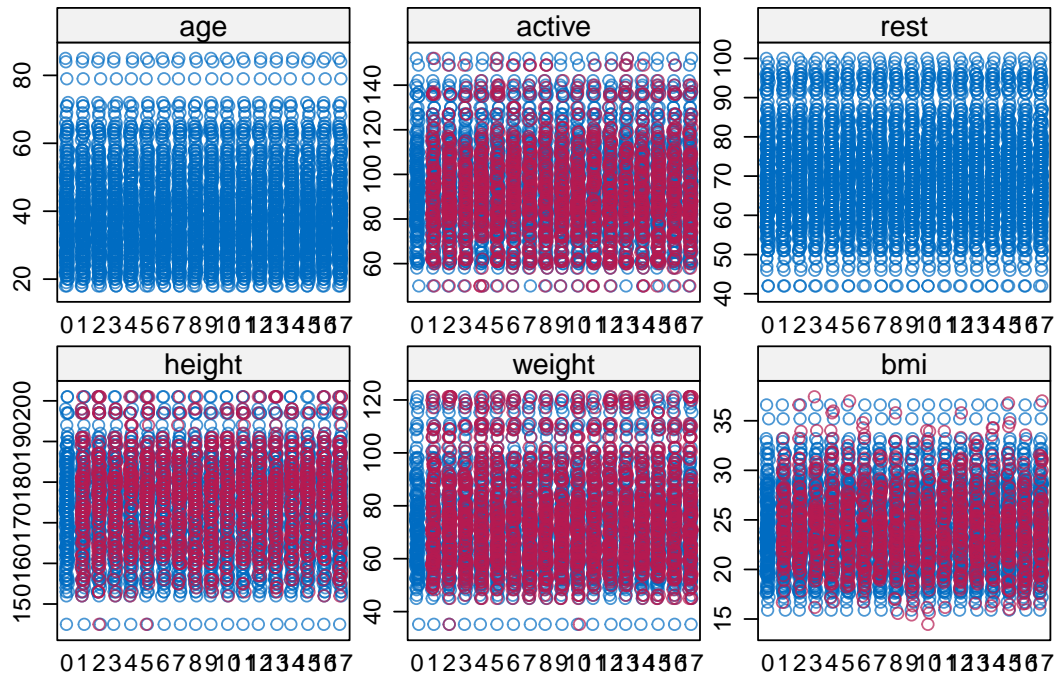


Figure 3: Stripplot of imputed data

Appendix B: Prediction Routine

```
model.com <- lm(active ~ age + rest + bmi + sex + smoke + intensity + bmi * age, data=train.com)
pred.com <- predict(model.com, newdata=test.com)

tActivity <- test.com$active
mActivity <- pred.com

test.com.active<- tActivity
pred.com.error <- tActivity - mActivity
pred.com.R2=1-sum(pred.com.error^2)/sum((test.com.active- mean(test.com.active))^2)
eval.com <- c(sse(tActivity, mActivity), rse(tActivity, mActivity), rmse(tActivity, mActivity), pred.com)
eval.p <- eval.com

splitX <- function(imps, index) {
  tmp <- lapply(imps, split, f = index)
  out <- list()
  for(i in 1 : length(tmp[[1]]))
    out[[i]] <- lapply(tmp, "[", x = i)
  names(out) <- names(tmp[[1]])
  out
}

modelmaker <- function(data, index) {
  a <- as.data.frame(data$one[index])
  colnames(a) <- c("age", "smoke", "sex", "intensity",
                  "active", "rest", "height", "weight", "bmi")
  b <- as.data.frame(data$two[index])
  colnames(b) <- c("age", "smoke", "sex", "intensity",
                  "active", "rest", "height", "weight", "bmi")
  c <- as.data.frame(data$three[index])
  colnames(c) <- c("age", "smoke", "sex", "intensity",
                  "active", "rest", "height", "weight", "bmi")
  d <- as.data.frame(data$four[index])
  colnames(d) <- c("age", "smoke", "sex", "intensity",
                  "active", "rest", "height", "weight", "bmi")
  e <- as.data.frame(data$five[index])
  colnames(e) <- c("age", "smoke", "sex", "intensity",
                  "active", "rest", "height", "weight", "bmi")
  t1 <- rbind(a,b)
  t2 <- rbind(t1,c)
  df <- rbind(t2,d)
  out <- lm(active ~ age + rest + bmi + sex + smoke + intensity + bmi*age, data=df)
  val <- list(predict(out, e))
  return(val)
}

trainModelmaker <- function(data, index) {
  a <- as.data.frame(data$one[index])
  colnames(a) <- c("age", "smoke", "sex", "intensity",
                  "active", "rest", "height", "weight", "bmi")

```

```

b <- as.data.frame(data$two[index])
colnames(b) <- c("age", "smoke", "sex", "intensity",
               "active", "rest", "height", "weight", "bmi")
c <- as.data.frame(data$three[index])
colnames(c) <- c("age", "smoke", "sex", "intensity",
               "active", "rest", "height", "weight", "bmi")
d <- as.data.frame(data$four[index])
colnames(d) <- c("age", "smoke", "sex", "intensity",
               "active", "rest", "height", "weight", "bmi")
e <- as.data.frame(data$five[index])
colnames(e) <- c("age", "smoke", "sex", "intensity",
               "active", "rest", "height", "weight", "bmi")

t1 <- rbind(a,b)
t2 <- rbind(t1,c)
df <- rbind(t2,d)
out <- lm(active ~ age + rest + bmi + sex + smoke + intensity + bmi*age, data=df)
val <- list(predict(out, df))
return(val)
}

fetchTrainData <- function(data, index) {
  a <- as.data.frame(data$one[index])
  colnames(a) <- c("age", "smoke", "sex", "intensity",
                 "active", "rest", "height", "weight", "bmi")
  b <- as.data.frame(data$two[index])
  colnames(b) <- c("age", "smoke", "sex", "intensity",
                 "active", "rest", "height", "weight", "bmi")
  c <- as.data.frame(data$three[index])
  colnames(c) <- c("age", "smoke", "sex", "intensity",
                 "active", "rest", "height", "weight", "bmi")
  d <- as.data.frame(data$four[index])
  colnames(d) <- c("age", "smoke", "sex", "intensity",
                 "active", "rest", "height", "weight", "bmi")

  t1 <- rbind(a,b)
  t2 <- rbind(t1,c)
  df <- rbind(t2,d)
  df
}

fetchData <- function(data, index) {
  a <- as.data.frame(data$five[m])
  colnames(a) <- c("age", "smoke", "sex", "intensity",
                 "active", "rest", "height", "weight", "bmi")
  a
}

trainValidator <- function(data, pred, index, sigh) {
  a <- as.data.frame(data$one[index])
  colnames(a) <- c("age", "smoke", "sex", "intensity",
                 "active", "rest", "height", "weight", "bmi")
  b <- as.data.frame(data$two[index])
  colnames(b) <- c("age", "smoke", "sex", "intensity",

```

```

        "active", "rest", "height", "weight", "bmi")
c <- as.data.frame(data$three[index])
colnames(c) <- c("age", "smoke", "sex", "intensity",
               "active", "rest", "height", "weight", "bmi")
d <- as.data.frame(data$four[index])
colnames(d) <- c("age", "smoke", "sex", "intensity",
               "active", "rest", "height", "weight", "bmi")
t1 <- rbind(a,b)
t2 <- rbind(t1,c)
e <- rbind(t2,d)
e

f <- as.data.frame(pred)
colnames(f) <- c("active")
tActivity <- e$active
mActivity <- f$active
test.com.active<- tActivity
pred.com.error <- tActivity - mActivity
pred.com.R2=1-sum(pred.com.error^2)/sum((test.com.active- mean(test.com.active))^2)
if (sigh=="r2")
  return(pred.com.R2)
else if(sigh=="sse")
  return(sse(tActivity, mActivity))
else if(sigh=="rse")
  return(rse(tActivity, mActivity))
else if(sigh=="rmse")
  return(rmse(tActivity, mActivity))
}

validator <- function(data, pred, index, sigh) {
  e <- as.data.frame(data$five[index])
  colnames(e) <- c("age", "smoke", "sex", "intensity",
                 "active", "rest", "height", "weight", "bmi")
  f <- as.data.frame(pred)
  colnames(f) <- c("active")
  tActivity <- e$active
  mActivity <- f$active
  test.com.active<- tActivity
  pred.com.error <- tActivity - mActivity
  pred.com.R2=1-sum(pred.com.error^2)/sum((test.com.active- mean(test.com.active))^2)
  if (sigh=="r2")
    return(pred.com.R2)
  else if(sigh=="sse")
    return(sse(tActivity, mActivity))
  else if(sigh=="rse")
    return(rse(tActivity, mActivity))
  else if(sigh=="rmse")
    return(rmse(tActivity, mActivity))
}

fixer <- function(data) {
  df <- t(as.data.frame(data))
  colnames(df) <- c("Metric")

```

```

    mean(df)
}

findX <- function(data, type) {
  df <- t(as.data.frame(data))
  colnames(df) <- c("Metric")
  if(type=="max")
    return(max(df))
  else if(type=="min")
    return(min(df))
}

d <- splitX(complete(imp1, "all"), c("one","two","three","four","five"))

```

```

predictions <- list()

for (m in 1 : length(d$one))
  predictions[m] <- modelmaker(d, m)

predSSE <- list()
predRSE <- list()
predRMSE <- list()
predR2 <- list()
for (m in 1: length(predictions))
  predSSE[m] <- validator(d, predictions[m], m, "sse")

for (m in 1: length(predictions))
  predRSE[m] <- validator(d, predictions[m], m, "rse")

for (m in 1: length(predictions))
  predRMSE[m] <- validator(d, predictions[m], m, "rmse")

for (m in 1: length(predictions))
  predR2[m] <- validator(d, predictions[m], m, "r2")

eval.com <- list()

eval.com[1] <- fixer(predSSE)
eval.com[2] <- fixer(predRSE)
eval.com[3] <- fixer(predRMSE)
eval.com[4] <- fixer(predR2)

eval.best <- list()
eval.best[1] <- findX(predSSE, "min")
eval.best[2] <- findX(predRSE, "min")
eval.best[3] <- findX(predRMSE, "min")
eval.best[4] <- findX(predR2, "max")

eval.worst <- list()
eval.worst[1] <- findX(predSSE, "max")
eval.worst[2] <- findX(predRSE, "max")
eval.worst[3] <- findX(predRMSE, "max")
eval.worst[4] <- findX(predR2, "min")

```

```
eval1 <- as.data.frame(eval.com)
eval2 <- rbind(eval1,eval.best)
eval3 <- rbind(eval2,eval.worst)
eval <- rbind(eval3, eval.p)
rownames(eval) <- c("Average (Imputed)", "Best (Imputed)", "Worst (Imputed)", "Complete")
colnames(eval) <- c("SSE", "RSE", "RMSE", "R2")
```