MISSING DATA ANALYSIS

ASSIGNMENT 1

Hephaes Chuen Chau (I54105029)

2022-04-06

Contents

Question 1 **1**
(a) 1
(b) 1
(c) 1
(d) 2

Question 2 **2**
(a) Missing rate in each variable 2
(d) Missing patterns 4

Question 3 **5**
(a) Mean Imputation 8
(b) Regression imputation and stepwise multiple regression 13
(c) Stochastic regression imputation and stepwise multiple regression 17
(d) Random number imputation and stepwise multiple regression 20

Question 4 **23**

Question 1

(a)

We assume that all for all variables, the missing occurs in the same cases. Therefore, the largest possible subsample is $100 - 100 \times 10\% = 90$

(b)

We assume that the missing data follow a monotone pattern, meaning that for each variable 10% of the data is deleted. This means the smallest possible subsample is 0.

(c)

Let X_i with support $\{0, 1\}$ denote the random variable indicating whether case i has at least one missing value in one of the 10 variables (if missing then $X_i = 1$). We know therefore by assuming MCAR that

$$\mathbb{P}(X_i = 0) = (0.9)^{10}$$

A case i is retained in listwise deletion method if and only if $X_i = 0$. Thus, the expected number of retained cases is $100 \times (0.9)^{10} = 34.8$. This means on average 34 cases are retained.

(d)

The available sample is larger. In listwise deletion technique, a case is removed as long as there is at least one variable that has missing value. Thus, while the available sample is 100, in (a) we assume all the missing values in different variables occur in a fixed set of cases, then even with this stringent assumption we still have to remove 10 cases from the data.

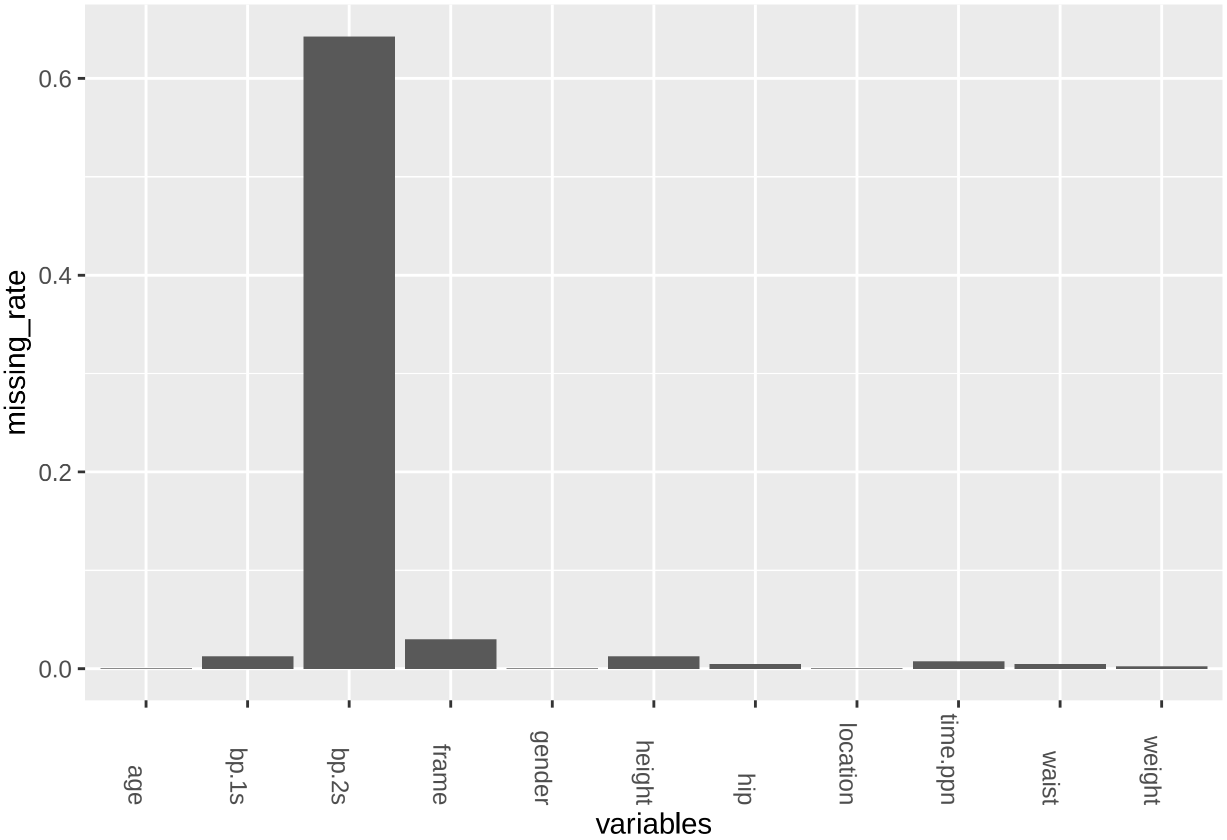
Question 2

(a) Missing rate in each variable

We first remove the unneeded variables and produce a plot

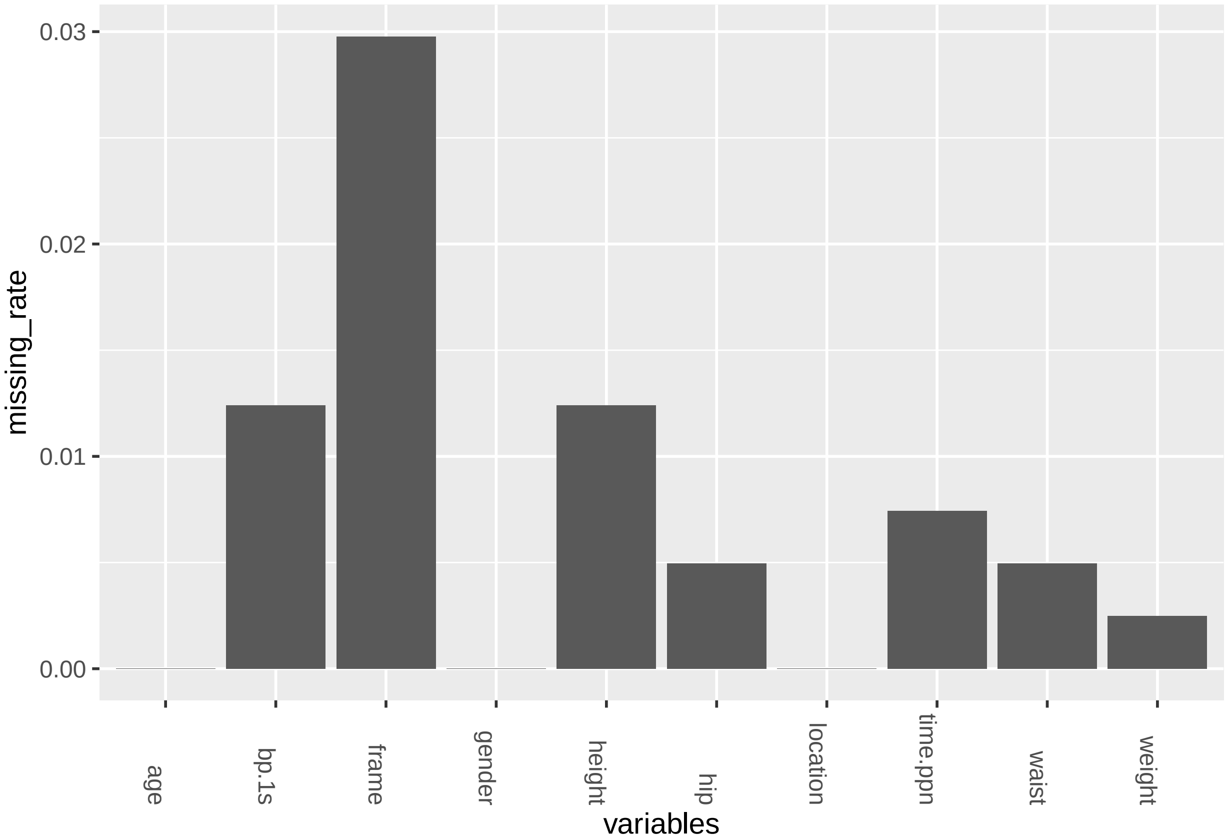
```
DATA <- read.csv("Diabetes.csv")
for (i in c("location", "gender", "frame")) {
  DATA[,i] <- as.factor(DATA[,i])
}
# Removing unnecessary variables
data <- DATA[,-which(colnames(DATA) %in% c("id", "chol", "stab.glu", "hdl", "ratio", "glyhb", "bp.2d",
missing_rate <- c()
for (i in seq_len(11)) {
  j <- colnames(data)[i]
  Nmissing <- sum(is.na(data[,j]))
  N <- length(data[,j])
  missing_rate[i] <- Nmissing/N
}
names(missing_rate) <- colnames(data)
missing_rate <- as.data.frame(missing_rate)
```

```
library(ggplot2)
missing_rate$variables <- rownames(missing_rate)
p <-ggplot(data=missing_rate, aes(x=variables, y=missing_rate)) +
  geom_bar(stat="identity") + theme(axis.text.x=element_text(angle=-90, hjust = 1))
p
```



We therefore see that `bp.2s` have particularly high missing rates (exceeding 60%). Next we compare the missing rates of remaining variables, by:

```
new_missing_rate <- missing_rate[-which(rownames(missing_rate) %in% c("bp.2s","bp.2d")), ]
p2 <-ggplot(data=new_missing_rate, aes(x=variables, y=missing_rate)) +
  geom_bar(stat="identity") + theme(axis.text.x=element_text(angle=-90, hjust = 1))
p2
```



Therefore, of the remaining variables, `frame` has particularly high missing rates (with the latter being close to 3%).

(d) Missing patterns

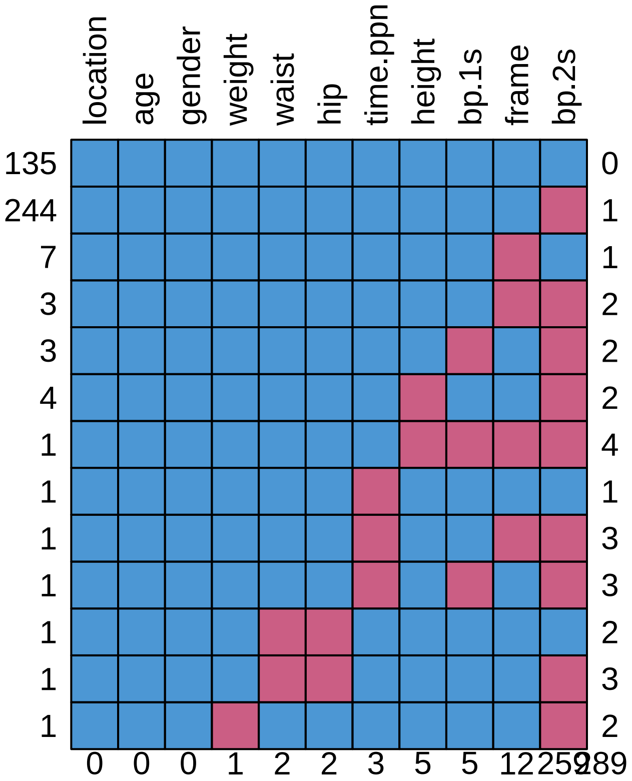
```
library(mice)
```

```
##
## Attaching package: 'mice'

## The following object is masked from 'package:stats':
##
##     filter

## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
md.pattern(data, rotate.names = TRUE)
```



```
##     location age gender weight waist hip time.ppn height bp.1s frame bp.2s
## 135        1   1      1      1     1   1        1      1     1     1     1   0
## 244        1   1      1      1     1   1        1      1     1     1     0   1
## 7          1   1      1      1     1   1        1      1     1     0     1   1
## 3          1   1      1      1     1   1        1      1     1     0     0   2
## 3          1   1      1      1     1   1        1      1     0     1     0   2
## 4          1   1      1      1     1   1        1      0     1     1     0   2
## 1          1   1      1      1     1   1        1      0     0     0     0   4
## 1          1   1      1      1     1   1        0      1     1     1     1   1
## 1          1   1      1      1     1   1        0      1     1     0     0   3
## 1          1   1      1      1     1   1        0      1     0     1     0   3
## 1          1   1      1      1     0   0        1      1     1     1     1   2
## 1          1   1      1      1     0   0        1      1     1     1     0   3
## 1          1   1      1      0     1   1        1      1     1     1     0   2
##            0   0      0      1     2   2        3      5     5    12   259 289
```

Therefore, from the above plot, we can identify a monotone pattern of missing data, because (i) there are some variables with no missing data, and (ii) there are more than one variables with missing data, and (iii) starting from the variable `chol`, all following variables have missing values in some cases.

The pattern is also unconnected because there are some observed data points that cannot be reached from other data points through vertical or horizontal moves.

Question 3

We select `bp.1s` as the response variable.

Before we apply any imputation method, we fit a linear model to the data with the incomplete cases deleted. The purpose of this step is to allow for a comparison of different imputation methods relative to complete case analysis. To conduct the complete case analysis, we fit a linear model on the data, ignoring any case with NA in any of the variables, and we conduct a stepwise regression model selection based on AIC.

```
library(MASS)
ccfit <- lm(data = na.omit(data[,!(colnames(data) %in% "bp.2s")]), bp.1s ~ location + age + gender + he
ccstep <- stepAIC(ccfit, direction = "both")
```

```
## Start:  AIC=2302.64
## bp.1s ~ location + age + gender + height + weight + frame + waist + 
##     hip + time.ppn
## 
##             Df Sum of Sq    RSS    AIC
## - height    1       3.5 155607 2300.7
## - location  1      16.2 155620 2300.7
## - gender    1     210.9 155814 2301.2
## - weight    1     265.4 155869 2301.3
## - frame     2    1169.4 156773 2301.5
## - hip       1     389.4 155993 2301.6
## - waist     1     427.5 156031 2301.7
## <none>                  155604 2302.6
## - time.ppn  1     982.5 156586 2303.0
## - age       1   22832.2 178436 2352.5
## 
## Step:  AIC=2300.65
## bp.1s ~ location + age + gender + weight + frame + waist + hip + 
##     time.ppn
## 
##             Df Sum of Sq    RSS    AIC
## - location  1      19.5 155627 2298.7
## - gender    1     268.8 155876 2299.3
## - weight    1     326.0 155933 2299.4
## - frame     2    1186.9 156794 2299.5
## - hip       1     413.4 156021 2299.7
## - waist     1     438.1 156045 2299.7
## <none>                  155607 2300.7
## - time.ppn  1     980.0 156587 2301.0
## + height    1       3.5 155604 2302.6
## - age       1   22908.4 178516 2350.7
## 
## Step:  AIC=2298.7
## bp.1s ~ age + gender + weight + frame + waist + hip + time.ppn
## 
##             Df Sum of Sq    RSS    AIC
## - gender    1     291.9 155919 2297.4
## - weight    1     341.6 155968 2297.5
## - frame     2    1207.5 156834 2297.6
## - waist     1     425.5 156052 2297.7
## - hip       1     464.7 156091 2297.8
## <none>                  155627 2298.7
## - time.ppn  1    1034.4 156661 2299.2
## + location  1      19.5 155607 2300.7
## + height    1       6.8 155620 2300.7
## - age       1   22950.1 178577 2348.8
## 
## Step:  AIC=2297.41
## bp.1s ~ age + weight + frame + waist + hip + time.ppn
```

```
##
##            Df Sum of Sq    RSS    AIC
## - weight    1     133.4 156052 2295.7
## - hip       1     202.8 156121 2295.9
## - frame     2    1209.7 157128 2296.3
## - waist     1     412.9 156332 2296.4
## <none>                  155919 2297.4
## - time.ppn  1    1048.3 156967 2297.9
## + gender    1     291.9 155627 2298.7
## + height    1      53.9 155865 2299.3
## + location  1      42.7 155876 2299.3
## - age       1   23850.9 179770 2349.4
##
## Step:  AIC=2295.73
## bp.1s ~ age + frame + waist + hip + time.ppn
##
##            Df Sum of Sq    RSS    AIC
## - hip       1     112.8 156165 2294.0
## - waist     1     280.9 156333 2294.4
## - frame     2    1174.4 157226 2294.6
## <none>                  156052 2295.7
## - time.ppn  1    1024.2 157076 2296.2
## + weight    1     133.4 155919 2297.4
## + gender    1      83.7 155968 2297.5
## + location  1      46.2 156006 2297.6
## + height    1       0.2 156052 2297.7
## - age       1   28655.2 184707 2357.6
##
## Step:  AIC=2294.01
## bp.1s ~ age + frame + waist + time.ppn
##
##            Df Sum of Sq    RSS    AIC
## - frame     2    1250.1 157415 2293.0
## <none>                  156165 2294.0
## - time.ppn  1    1075.9 157241 2294.6
## + hip       1     112.8 156052 2295.7
## + location  1      68.4 156096 2295.8
## + weight    1      43.4 156121 2295.9
## + gender    1      13.6 156151 2296.0
## + height    1      12.8 156152 2296.0
## - waist     1    1902.8 158068 2296.6
## - age       1   29138.1 185303 2356.8
##
## Step:  AIC=2293.03
## bp.1s ~ age + waist + time.ppn
##
##            Df Sum of Sq    RSS    AIC
## <none>                  157415 2293.0
## - time.ppn  1     958.0 158373 2293.3
## + frame     2    1250.1 156165 2294.0
## + hip       1     188.5 157226 2294.6
## + location  1     112.7 157302 2294.8
## + height    1      34.7 157380 2294.9
## + weight    1      17.3 157398 2295.0
## + gender    1       6.6 157408 2295.0
## - waist     1    3215.9 160631 2298.7
## - age       1   30356.7 187772 2357.9
```

```
ccstep$anova
```

```
## Stepwise Model Path 
## Analysis of Deviance Table
## 
## Initial Model:
## bp.1s ~ location + age + gender + height + weight + frame + waist + 
##     hip + time.ppn
## 
## Final Model:
## bp.1s ~ age + waist + time.ppn
## 
## 
##         Step Df    Deviance Resid. Df Resid. Dev      AIC
## 1                                 368   155603.6 2302.644
## 2   - height  1    3.534488       369   155607.2 2300.653
## 3 - location  1   19.526825       370   155626.7 2298.700
## 4   - gender  1  291.926970       371   155918.6 2297.411
## 5   - weight  1  133.368251       372   156052.0 2295.735
## 6      - hip  1  112.816380       373   156164.8 2294.009
## 7    - frame  2 1250.109769       375   157414.9 2293.030
```

```
summary(ccstep)
```

```
## 
## Call:
## lm(formula = bp.1s ~ age + waist + time.ppn, data = na.omit(data[, 
##     !(colnames(data) %in% "bp.2s")]))
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -51.599 -12.655  -2.261   9.420  94.806 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 92.913021   7.414631  12.531  < 2e-16 ***
## age          0.562444   0.066139   8.504 4.42e-16 ***
## waist        0.511407   0.184767   2.768  0.00592 ** 
## time.ppn    -0.005133   0.003398  -1.511  0.13172    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 20.49 on 375 degrees of freedom
## Multiple R-squared:  0.2004, Adjusted R-squared:  0.194 
## F-statistic: 31.33 on 3 and 375 DF,  p-value: < 2.2e-16
```

Therefore, our complete case analysis has settled for `bp.1s ~ age + waist time.ppn` as the final model to be selected based on AIC. The adjusted R-squared is very low (0.194).

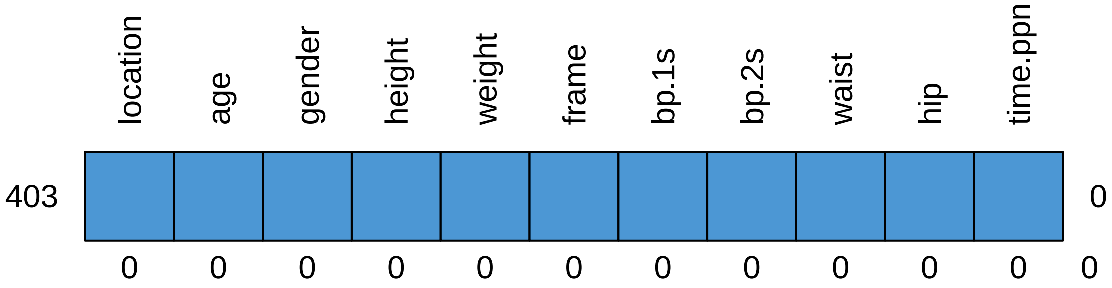
(a) Mean Imputation

(i) The imputation

We do mean imputation with the package `mice`, with predictive mean matching for categorical and mean imputation for numerical variables.

```
library(mice)
mean_imputated_data <- mice(data, defaultMethod = c("mean", "pmm", "pmm", "pmm"), m = 1, maxit = 20, pr
md.pattern(complete(mean_imputated_data), rotate.names = TRUE)
```

```
##  /\     /\
## {  `---'  }
## {  O   O  }
## ==>  V <==  No need for mice. This data set is completely observed.
##  \  \|/  /
##   `-----'
```



```
##     location age gender height weight frame bp.1s bp.2s waist hip time.ppn
## 403        1   1      1      1      1     1     1     1     1   1        1 0
##            0   0      0      0      0     0     0     0     0   0        0 0
```

(ii) Stepwise multiple regression

After the imputation, we hope to conduct a multiple regression analysis of the data. We proceed by stepwise regression to select the best model based on AIC, as follows:

```
fit <- lm(data = complete(mean_imputated_data), bp.1s ~ location + age + gender + height + weight + fran
library(MASS)
step <- stepAIC(fit, direction = "both")
```

```
## Start:  AIC=2430.6
## bp.1s ~ location + age + gender + height + weight + frame + waist +
##     hip + time.ppn
##
##            Df Sum of Sq    RSS    AIC
## - location  1      12.7 158847 2428.6
## - height    1      38.9 158874 2428.7
## - weight    1     132.3 158967 2428.9
## - waist     1     244.1 159079 2429.2
## - gender    1     269.8 159104 2429.3
## - hip       1     400.8 159235 2429.6
## - frame     2    1283.0 160118 2429.8
## <none>                  158835 2430.6
## - time.ppn  1     955.0 159790 2431.0
## - age       1   26973.2 185808 2491.8
##
## Step:  AIC=2428.64
## bp.1s ~ age + gender + height + weight + frame + waist + hip +
##     time.ppn
##
##            Df Sum of Sq    RSS    AIC
## - height    1      33.4 158881 2426.7
## - weight    1     129.9 158977 2427.0
## - gender    1     257.5 159105 2427.3
## - waist     1     258.7 159106 2427.3
## - hip       1     388.2 159236 2427.6
## - frame     2    1276.3 160124 2427.9
## <none>                  158847 2428.6
## - time.ppn  1     942.7 159790 2429.0
## + location  1      12.7 158835 2430.6
## - age       1   26964.3 185812 2489.8
##
## Step:  AIC=2426.72
## bp.1s ~ age + gender + weight + frame + waist + hip + time.ppn
##
##            Df Sum of Sq    RSS    AIC
## - weight    1     204.7 159085 2425.2
## - gender    1     235.6 159116 2425.3
## - waist     1     269.8 159151 2425.4
## - hip       1     454.6 159335 2425.9
## - frame     2    1338.6 160219 2426.1
## <none>                  158881 2426.7
## - time.ppn  1     937.9 159819 2427.1
## + height    1      33.4 158847 2428.6
## + location  1       7.2 158874 2428.7
## - age       1   27144.6 186025 2488.3
##
## Step:  AIC=2425.24
## bp.1s ~ age + gender + frame + waist + hip + time.ppn
##
##            Df Sum of Sq    RSS    AIC
## - gender    1        91 159176 2423.5
## - waist     1       127 159212 2423.6
## - hip       1       260 159345 2423.9
## - frame     2      1284 160370 2424.5
## <none>                  159085 2425.2
## - time.ppn  1       923 160009 2425.6
## + weight    1       205 158881 2426.7
## + height    1       108 158977 2427.0
## + location  1         2 159083 2427.2
## - age       1     32837 191922 2498.9
##
## Step:  AIC=2423.47
## bp.1s ~ age + frame + waist + hip + time.ppn
##
##            Df Sum of Sq    RSS    AIC
## - hip       1       177 159354 2421.9
## - waist     1       206 159382 2422.0
## - frame     2      1321 160497 2422.8
## <none>                  159176 2423.5
## - time.ppn  1       944 160120 2423.8
## + gender    1        91 159085 2425.2
## + weight    1        60 159116 2425.3
## + height    1         2 159174 2425.5
## + location  1         0 159176 2425.5
## - age       1     32752 191928 2496.9
##
## Step:  AIC=2421.92
## bp.1s ~ age + frame + waist + time.ppn
##
##            Df Sum of Sq    RSS    AIC
## - frame     2      1409 160762 2421.5
## <none>                  159354 2421.9
## - time.ppn  1      1005 160359 2422.4
## + hip       1       177 159176 2423.5
## + height    1        29 159325 2423.8
## + gender    1         9 159345 2423.9
## + weight    1         5 159349 2423.9
## + location  1         2 159352 2423.9
## - waist     1      1893 161247 2424.7
## - age       1     33182 192536 2496.2
##
## Step:  AIC=2421.46
## bp.1s ~ age + waist + time.ppn
##
##            Df Sum of Sq    RSS    AIC
## <none>                  160762 2421.5
## - time.ppn  1       897 161659 2421.7
## + frame     2      1409 159354 2421.9
## + hip       1       265 160497 2422.8
## + height    1        55 160708 2423.3
## + gender    1        12 160751 2423.4
## + location  1        11 160751 2423.4
## + weight    1         3 160760 2423.5
## - waist     1      3270 164032 2427.6
## - age       1     35542 196305 2500.0
```

```
step$anova
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## bp.1s ~ location + age + gender + height + weight + frame + waist +
##     hip + time.ppn
##
## Final Model:
## bp.1s ~ age + waist + time.ppn
##
##
##         Step Df   Deviance Resid. Df Resid. Dev      AIC
## 1                                392   158834.6 2430.603
## 2 - location  1   12.66448       393   158847.3 2428.635
## 3   - height  1   33.38172       394   158880.7 2426.720
## 4   - weight  1  204.70564       395   159085.4 2425.239
## 5   - gender  1   90.99586       396   159176.4 2423.469
## 6      - hip  1  177.49041       397   159353.9 2421.918
## 7    - frame  2 1408.58096       399   160762.5 2421.465
```

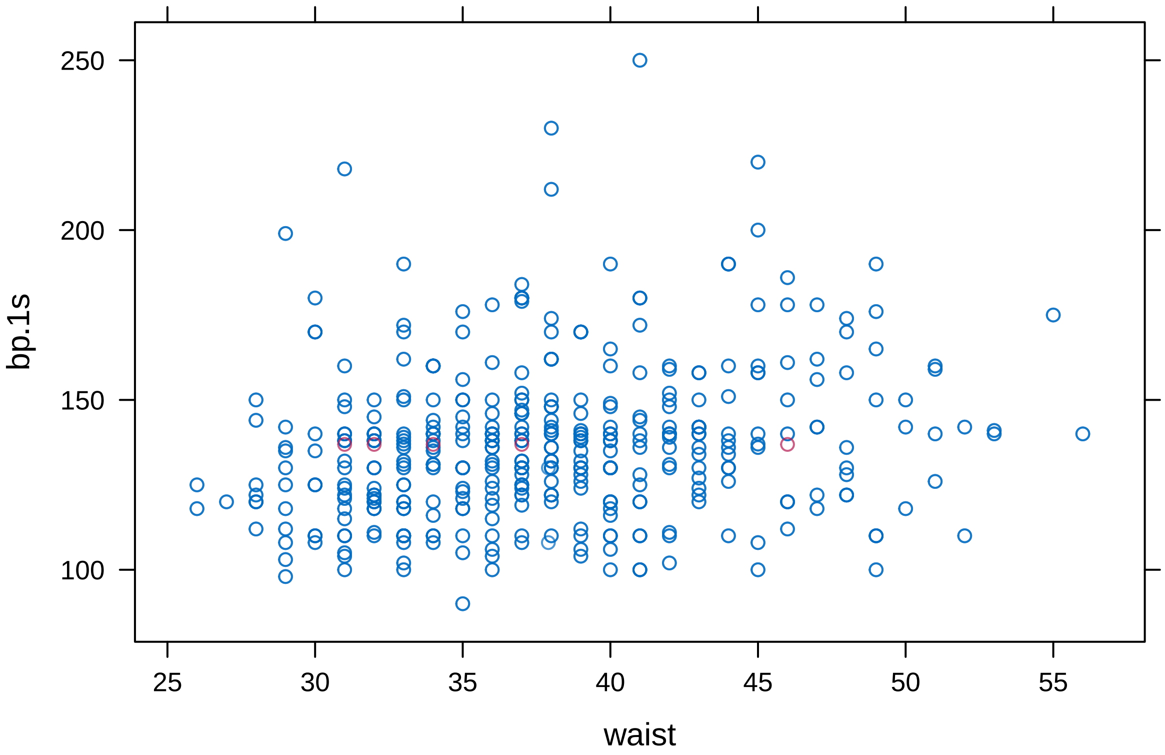
```
summary(step)
```

```
##
## Call:
## lm(formula = bp.1s ~ age + waist + time.ppn, data = complete(mean_imputated_data))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.137 -12.346  -1.625   9.558  94.331
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 92.078885   7.153653  12.872  < 2e-16 ***
## age          0.583164   0.062090   9.392  < 2e-16 ***
## waist        0.505549   0.177458   2.849  0.00462 **
## time.ppn    -0.004855   0.003254  -1.492  0.13647
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.07 on 399 degrees of freedom
## Multiple R-squared:  0.217,  Adjusted R-squared:  0.2111
## F-statistic: 36.85 on 3 and 399 DF,  p-value: < 2.2e-16
```

Therefore, with mean imputation, the stepwise model selection based on AIC applied on a family of simple multiple regression models has settled for a model with `age`, `waist` and `time.ppn` as the important explanatory variables. The summary of the final model reveals that only `age` and `waist` have effects significantly different from 0, and the effects of `time.ppn` is very weak. The adjusted R-squared is very low.

Compared with the complete case analysis, `waist` is introduced as a new explanatory variable after mean imputation, and therefore we are interested in looking at the relationship between `waist` and `bp.1s`, as follows:

```
xyplot(mean_imputated_data, bp.1s ~ waist)
```



(b) Regression imputation and stepwise multiple regression

We use linear regression for numerical variables and polytomous logistic regression for categorical variables.

```
library(mice)
regr_imputed_data <- mice(data, defaultMethod = c("norm.predict", "polyreg", "polyreg", "polyreg"), m
fit <- lm(data = complete(regr_imputed_data), bp.1s ~ location + age + gender + height + weight + fra
library(MASS)
step <- stepAIC(fit, direction = "both")
```

```
## Start:  AIC=2430.14
## bp.1s ~ location + age + gender + height + weight + frame + waist +
##     hip + time.ppn
##
##            Df Sum of Sq    RSS    AIC
## - location  1       7.2 158660 2428.2
## - height    1      18.3 158671 2428.2
## - weight    1     170.8 158823 2428.6
## - frame     2     992.7 159645 2428.7
## - waist     1     280.6 158933 2428.8
## - gender    1     320.4 158973 2428.9
## - hip       1     521.9 159175 2429.5
## <none>                  158653 2430.1
## - time.ppn  1     945.6 159598 2430.5
## - age       1   27632.7 186285 2492.8
##
## Step:  AIC=2428.16
## bp.1s ~ age + gender + height + weight + frame + waist + hip +
##     time.ppn
##
```

```
##             Df Sum of Sq    RSS    AIC
## - height    1      15.4 158675 2426.2
## - weight    1     169.2 158829 2426.6
## - frame     2     986.9 159647 2426.7
## - waist     1     293.0 158953 2426.9
## - gender    1     313.4 158973 2426.9
## - hip       1     515.5 159175 2427.5
## <none>                  158660 2428.2
## - time.ppn  1     940.9 159601 2428.5
## + location  1       7.2 158653 2430.1
## - age       1   27627.6 186288 2490.8
##
## Step:  AIC=2426.2
## bp.1s ~ age + gender + weight + frame + waist + hip + time.ppn
##
##             Df Sum of Sq    RSS    AIC
## - frame     2    1019.1 159694 2424.8
## - weight    1     235.9 158911 2424.8
## - waist     1     302.7 158978 2425.0
## - gender    1     344.0 159019 2425.1
## - hip       1     576.0 159251 2425.7
## <none>                  158675 2426.2
## - time.ppn  1     936.5 159612 2426.6
## + height    1      15.4 158660 2428.2
## + location  1       4.3 158671 2428.2
## - age       1   27768.5 186444 2489.2
##
## Step:  AIC=2424.78
## bp.1s ~ age + gender + weight + waist + hip + time.ppn
##
##             Df Sum of Sq    RSS    AIC
## - weight    1     220.5 159915 2423.3
## - waist     1     306.1 160000 2423.6
## - gender    1     364.6 160059 2423.7
## - hip       1     687.6 160382 2424.5
## <none>                  159694 2424.8
## - time.ppn  1     853.9 160548 2424.9
## + frame     2    1019.1 158675 2426.2
## + height    1      47.6 159647 2426.7
## + location  1       0.1 159694 2426.8
## - age       1   29542.6 189237 2491.2
##
## Step:  AIC=2423.33
## bp.1s ~ age + gender + waist + hip + time.ppn
##
##             Df Sum of Sq    RSS    AIC
## - waist     1       140 160055 2421.7
## - gender    1       175 160090 2421.8
## - hip       1       467 160382 2422.5
## <none>                  159915 2423.3
## - time.ppn  1       855 160770 2423.5
## + weight    1       221 159694 2424.8
## + frame     2      1004 158911 2424.8
## + height    1       128 159787 2425.0
## + location  1         1 159914 2425.3
## - age       1     35458 195373 2502.0
##
## Step:  AIC=2421.69
## bp.1s ~ age + gender + hip + time.ppn
##
##            Df Sum of Sq    RSS    AIC
## - gender    1       338 160393 2420.5
## <none>                  160055 2421.7
## - time.ppn  1       841 160896 2421.8
## + frame     2      1006 159048 2423.2
## + waist     1       140 159915 2423.3
## + height    1       106 159949 2423.4
## + weight    1        54 160000 2423.6
## + location  1         1 160054 2423.7
## - hip       1      3959 164014 2429.5
## - age       1     39053 199107 2507.7
##
## Step:  AIC=2420.54
## bp.1s ~ age + hip + time.ppn
##
##            Df Sum of Sq    RSS    AIC
## <none>                  160393 2420.5
## - time.ppn  1       832 161225 2420.6
## + gender    1       338 160055 2421.7
## + waist     1       302 160090 2421.8
## + frame     2      1079 159314 2421.8
## + height    1        29 160364 2422.5
## + weight    1        27 160365 2422.5
## + location  1         1 160392 2422.5
## - hip       1      3623 164015 2427.5
## - age       1     39881 200273 2508.0
```

```
step$anova
```

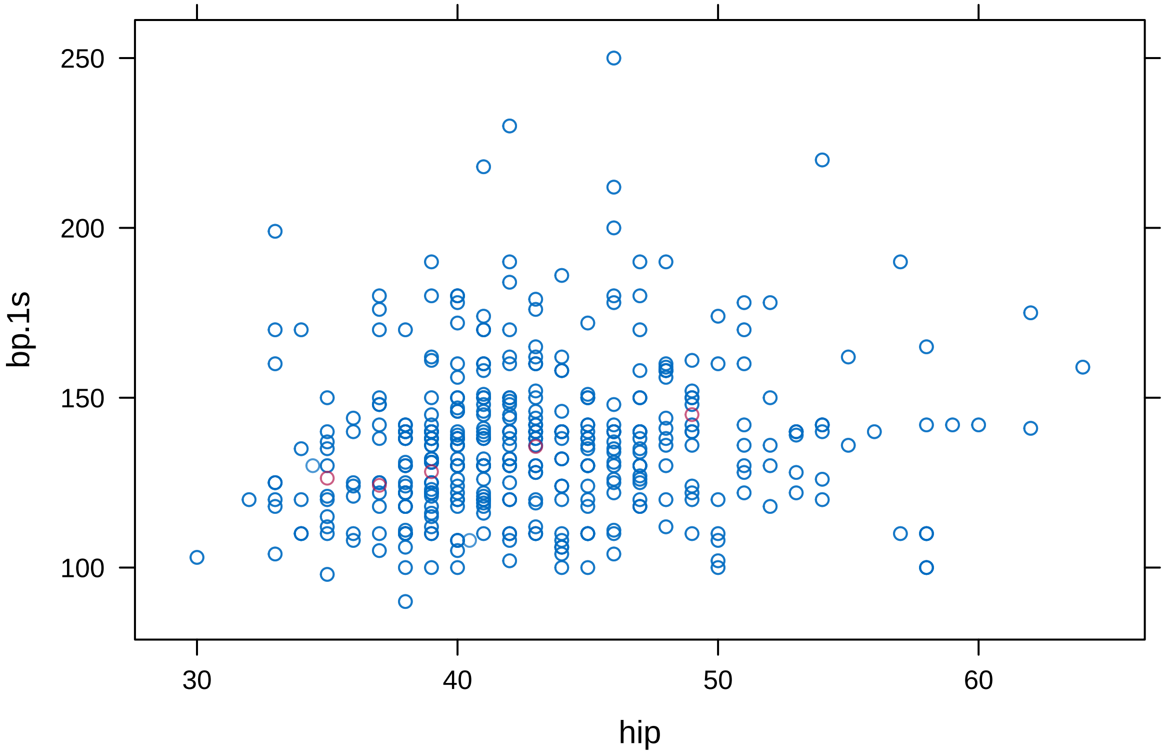
```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## bp.1s ~ location + age + gender + height + weight + frame + waist +
##     hip + time.ppn
##
## Final Model:
## bp.1s ~ age + hip + time.ppn
##
##
##         Step Df    Deviance Resid. Df Resid. Dev      AIC
## 1                                 392   158652.7 2430.141
## 2 - location  1    7.201231       393   158659.9 2428.159
## 3   - height  1   15.372829       394   158675.3 2426.198
## 4    - frame  2 1019.098057       396   159694.4 2424.778
## 5   - weight  1  220.534010       397   159914.9 2423.335
## 6    - waist  1  139.716392       398   160054.6 2421.687
## 7   - gender  1  337.933026       399   160392.5 2420.537
```

```
summary(step)
```

```
##
## Call:
## lm(formula = bp.1s ~ age + hip + time.ppn, data = complete(regr_imputated_data))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.124 -12.991  -1.033   8.747  93.661
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 86.927089   8.364881  10.392  < 2e-16 ***
## age          0.610989   0.061342   9.960  < 2e-16 ***
## hip          0.532166   0.177269   3.002  0.00285 **
## time.ppn    -0.004684   0.003255  -1.439  0.15094
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.05 on 399 degrees of freedom
## Multiple R-squared:  0.2204, Adjusted R-squared:  0.2145
## F-statistic: 37.59 on 3 and 399 DF,  p-value: < 2.2e-16
```

With regression imputation, `hip` is introduced as an important explanatory variable. Therefore we are interested in analysing the relationship between `bp.1s` and `hip`:

```
xyplot(regr_imputated_data, bp.1s ~ hip)
```



(c) Stochastic regression imputation and stepwise multiple regression

We use stochastic linear regression for numerical variables and polytomous logistic regression for categorical variables.

```
library(mice)
stregr_imputed_data <- mice(data, defaultMethod = c("norm.nob", "polyreg", "polyreg", "polyreg"), m =
fit <- lm(data = complete(stregr_imputed_data), bp.1s ~ location + age + gender + height + weight + fr
library(MASS)
step <- stepAIC(fit, direction = "both")
```

```
## Start:  AIC=2442.01
## bp.1s ~ location + age + gender + height + weight + frame + waist + 
##     hip + time.ppn
## 
##            Df Sum of Sq    RSS    AIC
## - height    1       0.1 163397 2440.0
## - location  1       0.4 163397 2440.0
## - gender    1     248.2 163645 2440.6
## - weight    1     383.3 163780 2441.0
## - waist     1     485.2 163882 2441.2
## - hip       1     555.3 163952 2441.4
## <none>                  163397 2442.0
## - frame     2    1630.3 165027 2442.0
## - time.ppn  1    1360.2 164757 2443.4
## - age       1   25957.1 189354 2499.4
## 
## Step:  AIC=2440.01
## bp.1s ~ location + age + gender + weight + frame + waist + hip + 
##     time.ppn
## 
##            Df Sum of Sq    RSS    AIC
## - location  1       0.3 163397 2438.0
## - gender    1     349.7 163746 2438.9
## - weight    1     442.1 163839 2439.1
## - waist     1     489.1 163886 2439.2
## - hip       1     576.8 163974 2439.4
## <none>                  163397 2440.0
## - frame     2    1660.7 165057 2440.1
## - time.ppn  1    1360.9 164758 2441.4
## + height    1       0.1 163397 2442.0
## - age       1   26004.6 189401 2497.5
## 
## Step:  AIC=2438.02
## bp.1s ~ age + gender + weight + frame + waist + hip + time.ppn
## 
##            Df Sum of Sq    RSS    AIC
## - gender    1     352.2 163749 2436.9
## - weight    1     443.2 163840 2437.1
## - waist     1     496.5 163894 2437.2
## - hip       1     594.3 163991 2437.5
## <none>                  163397 2438.0
## - frame     2    1661.0 165058 2438.1
## - time.ppn  1    1377.7 164775 2439.4
## + location  1       0.3 163397 2440.0
## + height    1       0.1 163397 2440.0
## - age       1   26012.1 189409 2495.6
##
## Step:  AIC=2436.88
## bp.1s ~ age + weight + frame + waist + hip + time.ppn
##
##            Df Sum of Sq    RSS    AIC
## - weight    1     183.9 163933 2435.3
## - hip       1     273.8 164023 2435.6
## - waist     1     481.8 164231 2436.1
## <none>                  163749 2436.9
## - frame     2    1691.9 165441 2437.0
## + gender    1     352.2 163397 2438.0
## - time.ppn  1    1409.6 165159 2438.3
## + height    1      99.2 163650 2438.6
## + location  1       2.8 163746 2438.9
## - age       1   27183.8 190933 2496.8
##
## Step:  AIC=2435.34
## bp.1s ~ age + frame + waist + hip + time.ppn
##
##            Df Sum of Sq    RSS    AIC
## - hip       1       152 164085 2433.7
## - waist     1       305 164238 2434.1
## - frame     2      1627 165560 2435.3
## <none>                  163933 2435.3
## - time.ppn  1      1378 165311 2436.7
## + weight    1       184 163749 2436.9
## + gender    1        93 163840 2437.1
## + location  1         4 163929 2437.3
## + height    1         0 163933 2437.3
## - age       1     32785 196718 2506.8
##
## Step:  AIC=2433.71
## bp.1s ~ age + frame + waist + time.ppn
##
##            Df Sum of Sq    RSS    AIC
## <none>                  164085 2433.7
## - frame     2      1730 165815 2433.9
## - time.ppn  1      1444 165529 2435.2
## + hip       1       152 163933 2435.3
## + weight    1        62 164023 2435.6
## + location  1        16 164069 2435.7
## + gender    1        12 164073 2435.7
## + height    1         9 164075 2435.7
## - waist     1      2277 166362 2437.3
## - age       1     33298 197383 2506.2
```

```
step$anova
```

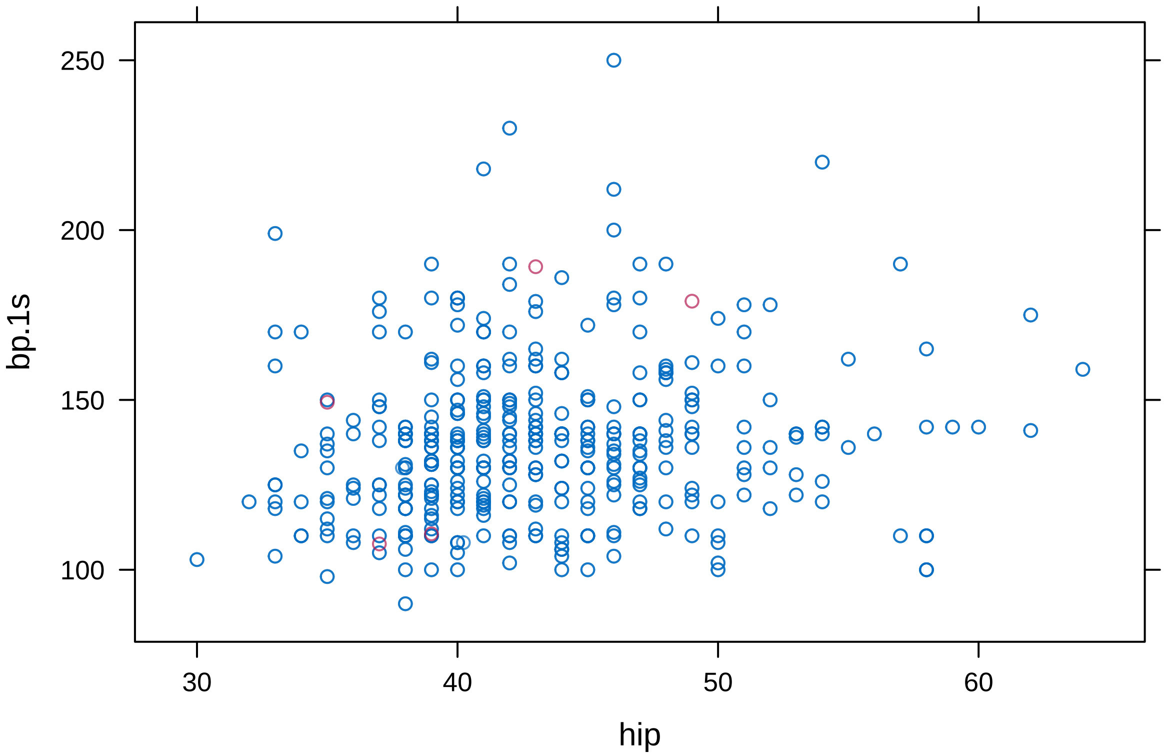
```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## bp.1s ~ location + age + gender + height + weight + frame + waist +
##     hip + time.ppn
##
## Final Model:
## bp.1s ~ age + frame + waist + time.ppn
##
##
##         Step Df    Deviance Resid. Df Resid. Dev      AIC
## 1                                 392   163396.6 2442.015
## 2   - height  1   0.1331173       393   163396.7 2440.015
## 3 - location  1   0.2975766       394   163397.0 2438.016
## 4   - gender  1 352.2250868       395   163749.2 2436.883
## 5   - weight  1 183.9384877       396   163933.2 2435.336
## 6      - hip  1 151.7257693       397   164084.9 2433.709
```

```
summary(step)
```

```
##
## Call:
## lm(formula = bp.1s ~ age + frame + waist + time.ppn, data = complete(stregr_imputated_data))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.067 -12.903  -0.900   9.015  92.084
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 93.564137   9.379442   9.975   <2e-16 ***
## age          0.576970   0.064281   8.976   <2e-16 ***
## framemedium  2.550779   2.607262   0.978   0.3285
## framesmall  -2.560083   3.255708  -0.786   0.4321
## waist        0.475020   0.202367   2.347   0.0194 *
## time.ppn    -0.006175   0.003303  -1.869   0.0623 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.33 on 397 degrees of freedom
## Multiple R-squared:  0.2243, Adjusted R-squared:  0.2145
## F-statistic: 22.96 on 5 and 397 DF,  p-value: < 2.2e-16
```

With stochastic regression imputation, `hip` is again introduced as an important explanatory variable. Therefore we are interested in analysing the relationship between `bp.1s` and `hip`:

```
xyplot(stregr_imputated_data, bp.1s ~ hip)
```



(d) Random number imputation and stepwise multiple regression

```
library(mice)
rand_imputed_data <- mice(data, defaultMethod = c("sample", "sample", "sample", "sample"), m = 1, max
fit <- lm(data = complete(rand_imputed_data), bp.1s ~ location + age + gender + height + weight + fra
library(MASS)
step <- stepAIC(fit, direction = "both")
```

```
## Start:  AIC=2436.37
## bp.1s ~ location + age + gender + height + weight + frame + waist + 
##     hip + time.ppn
## 
##            Df Sum of Sq    RSS    AIC
## - location  1       8.8 161133 2434.4
## - weight    1      39.1 161163 2434.5
## - height    1      56.2 161180 2434.5
## - gender    1     189.0 161313 2434.8
## - hip       1     196.4 161320 2434.9
## - frame     2    1006.1 162130 2434.9
## - waist     1     291.1 161415 2435.1
## <none>                  161124 2436.4
## - time.ppn  1     846.3 161970 2436.5
## - age       1   27695.3 188819 2498.3
## 
## Step:  AIC=2434.39
## bp.1s ~ age + gender + height + weight + frame + waist + hip + 
##     time.ppn
## 
##            Df Sum of Sq    RSS    AIC
## - weight    1      38.0 161171 2432.5
## - height    1      51.1 161184 2432.5
## - gender    1     180.6 161313 2432.8
## - hip       1     187.9 161320 2432.9
## - frame     2     998.8 162131 2432.9
## - waist     1     306.2 161439 2433.2
## <none>                  161133 2434.4
## - time.ppn  1     838.3 161971 2434.5
## + location  1       8.8 161124 2436.4
## - age       1   27689.6 188822 2496.3
##
## Step:  AIC=2432.49
## bp.1s ~ age + gender + height + frame + waist + hip + time.ppn
##
##            Df Sum of Sq    RSS    AIC
## - height    1      96.8 161267 2430.7
## - hip       1     153.9 161324 2430.9
## - gender    1     154.5 161325 2430.9
## - frame     2     981.0 162152 2430.9
## - waist     1     273.3 161444 2431.2
## <none>                  161171 2432.5
## - time.ppn  1     835.6 162006 2432.6
## + weight    1      38.0 161133 2434.4
## + location  1       7.6 161163 2434.5
## - age       1   31515.6 192686 2502.5
##
## Step:  AIC=2430.73
## bp.1s ~ age + gender + frame + waist + hip + time.ppn
##
##            Df Sum of Sq    RSS    AIC
## - gender    1        63 161330 2428.9
## - hip       1       161 161428 2429.1
## - frame     2      1017 162284 2429.3
## - waist     1       237 161505 2429.3
## <none>                  161267 2430.7
## - time.ppn  1       821 162089 2430.8
## + height    1        97 161171 2432.5
## + weight    1        84 161184 2432.5
## + location  1         1 161266 2432.7
## - age       1     33484 194751 2504.8
##
## Step:  AIC=2428.89
## bp.1s ~ age + frame + waist + hip + time.ppn
##
##            Df Sum of Sq    RSS    AIC
## - hip       1       106 161436 2427.2
## - frame     2      1024 162355 2427.4
## - waist     1       329 161659 2427.7
## <none>                  161330 2428.9
## - time.ppn  1       839 162169 2429.0
## + gender    1        63 161267 2430.7
## + weight    1        17 161313 2430.8
## + height    1         5 161325 2430.9
## + location  1         0 161330 2430.9
## - age       1     33428 194759 2502.8
##
## Step:  AIC=2427.15
## bp.1s ~ age + frame + waist + time.ppn
##
##            Df Sum of Sq    RSS    AIC
## - frame     2      1080 162517 2425.8
## <none>                  161436 2427.2
## - time.ppn  1       887 162323 2427.4
## + hip       1       106 161330 2428.9
## + height    1        29 161408 2429.1
## + gender    1         8 161428 2429.1
## + location  1         2 161435 2429.2
## + weight    1         0 161436 2429.2
## - waist     1      2024 163461 2430.2
## - age       1     34155 195592 2502.5
##
## Step:  AIC=2425.84
## bp.1s ~ age + waist + time.ppn
##
##            Df Sum of Sq    RSS    AIC
## - time.ppn  1       796 163313 2425.8
## <none>                  162517 2425.8
## + frame     2      1080 161436 2427.2
## + hip       1       162 162355 2427.4
## + height    1        51 162466 2427.7
## + location  1        12 162505 2427.8
## + gender    1         6 162511 2427.8
## + weight    1         5 162512 2427.8
## - waist     1      3156 165673 2431.6
## - age       1     35676 198193 2503.8
##
## Step:  AIC=2425.81
## bp.1s ~ age + waist
##
##            Df Sum of Sq    RSS    AIC
## <none>                  163313 2425.8
## + time.ppn  1       796 162517 2425.8
## + hip       1       216 163097 2427.3
## + frame     2       990 162323 2427.4
## + location  1        57 163256 2427.7
## + height    1        46 163267 2427.7
## + weight    1         9 163304 2427.8
## + gender    1         3 163310 2427.8
## - waist     1      3345 166658 2432.0
## - age       1     35940 199253 2504.0
```

```
step$anova
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## bp.1s ~ location + age + gender + height + weight + frame + waist +
##     hip + time.ppn
```

```
##
## Final Model:
## bp.1s ~ age + waist
##
##
##         Step Df   Deviance Resid. Df Resid. Dev      AIC
## 1                                 392   161123.8 2436.370
## 2 - location  1    8.75957       393   161132.5 2434.392
## 3   - weight  1   38.02755       394   161170.6 2432.487
## 4   - height  1   96.76294       395   161267.3 2430.729
## 5   - gender  1   63.00171       396   161330.3 2428.886
## 6      - hip  1  106.16272       397   161436.5 2427.151
## 7    - frame  2 1080.22080       399   162516.7 2425.839
## 8 - time.ppn  1  796.25496       400   163313.0 2425.808
```

```
summary(step)
```

```
##
## Call:
## lm(formula = bp.1s ~ age + waist, data = complete(rand_imputated_data))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.867 -12.639  -1.689   9.943  95.690
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  90.0276     7.0026  12.856  < 2e-16 ***
## age           0.5864     0.0625   9.382  < 2e-16 ***
## waist         0.5094     0.1780   2.862  0.00443 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.21 on 400 degrees of freedom
## Multiple R-squared:  0.2108, Adjusted R-squared:  0.2068
## F-statistic: 53.41 on 2 and 400 DF,  p-value: < 2.2e-16
```

The final model selected is the same as in the complete case analysis.

Question 4

We select `bp.2s` as the response variable.

As in question 3, we fit a linear model to the data with the incomplete cases deleted. The purpose of this step is to allow for a comparison of different imputation methods relative to complete case analysis. To conduct the complete case analysis, we fit a linear model on the data, ignoring any case with `NA` in any of the variables, and we conduct a stepwise regression model selection based on AIC.

```
library(MASS)
ccfit <- lm(data = na.omit(data[,!(colnames(data) %in% "bp.1s")]), bp.2s ~ location + age + gender + hei
ccstep <- stepAIC(ccfit, direction = "both")
```

```
## Start:  AIC=829.79
## bp.2s ~ location + age + gender + height + weight + frame + waist +
##     hip + time.ppn
##
```

```
##            Df Sum of Sq   RSS    AIC
## - frame     2      76.7 53657 825.99
## - gender    1       0.4 53581 827.80
## - location  1       1.8 53582 827.80
## - hip       1       6.4 53587 827.81
## - height    1      21.3 53602 827.85
## - weight    1     461.2 54041 828.95
## <none>                  53580 829.79
## - waist     1    1146.3 54727 830.65
## - time.ppn  1    2891.9 56472 834.89
## - age       1    6954.1 60534 844.27
##
## Step:  AIC=825.99
## bp.2s ~ location + age + gender + height + weight + waist + hip +
##     time.ppn
##
##            Df Sum of Sq   RSS    AIC
## - location  1       0.8 53658 823.99
## - hip       1       4.5 53661 824.00
## - gender    1       6.4 53663 824.00
## - height    1      13.8 53671 824.02
## - weight    1     422.0 54079 825.04
## <none>                  53657 825.99
## - waist     1    1178.9 54836 826.92
## + frame     2      76.7 53580 829.79
## - time.ppn  1    2825.0 56482 830.91
## - age       1    7194.4 60851 840.97
##
## Step:  AIC=823.99
## bp.2s ~ age + gender + height + weight + waist + hip + time.ppn
##
##            Df Sum of Sq   RSS    AIC
## - hip       1       4.0 53662 822.00
## - gender    1       6.7 53664 822.01
## - height    1      14.0 53672 822.02
## - weight    1     429.3 54087 823.07
## <none>                  53658 823.99
## - waist     1    1199.5 54857 824.97
## + location  1       0.8 53657 825.99
## + frame     2      75.7 53582 827.80
## - time.ppn  1    2849.4 56507 828.97
## - age       1    7211.4 60869 839.01
##
## Step:  AIC=822
## bp.2s ~ age + gender + height + weight + waist + time.ppn
##
##            Df Sum of Sq   RSS    AIC
## - gender    1      13.4 53675 820.03
## - height    1      22.1 53684 820.05
## <none>                  53662 822.00
## - weight    1     934.4 54596 822.33
## - waist     1    1284.5 54946 823.19
## + hip       1       4.0 53658 823.99
## + location  1       0.3 53661 824.00
## + frame     2      74.2 53588 825.81
## - time.ppn  1    2877.9 56540 827.05
## - age       1    7208.7 60870 837.02
##
## Step:  AIC=820.03
## bp.2s ~ age + height + weight + waist + time.ppn
##
##            Df Sum of Sq   RSS    AIC
## - height    1      90.3 53765 818.26
## <none>                  53675 820.03
## - weight    1     924.0 54599 820.34
## - waist     1    1279.9 54955 821.21
## + gender    1      13.4 53662 822.00
## + hip       1      10.6 53664 822.01
## + location  1       0.3 53675 822.03
## + frame     2      84.5 53591 823.82
## - time.ppn  1    2867.8 56543 825.06
## - age       1    7216.3 60891 835.06
##
## Step:  AIC=818.26
## bp.2s ~ age + weight + waist + time.ppn
##
##            Df Sum of Sq   RSS    AIC
## <none>                  53765 818.26
## - weight    1     834.0 54599 818.34
## - waist     1    1199.4 54965 819.24
## + height    1      90.3 53675 820.03
## + gender    1      81.6 53684 820.05
## + hip       1      71.2 53694 820.08
## + location  1       0.0 53765 820.26
## + frame     2      87.4 53678 822.04
## - time.ppn  1    2823.5 56589 823.17
## - age       1    7158.4 60924 833.13
```

```
ccstep$anova
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## bp.2s ~ location + age + gender + height + weight + frame + waist +
##     hip + time.ppn
##
## Final Model:
## bp.2s ~ age + weight + waist + time.ppn
##
##
##         Step Df  Deviance Resid. Df Resid. Dev      AIC
## 1                               124   53580.24 829.7942
## 2    - frame  2 76.705085       126   53656.94 825.9873
## 3 - location  1  0.796784       127   53657.74 823.9893
## 4      - hip  1  4.018433       128   53661.76 821.9994
## 5   - gender  1 13.371912       129   53675.13 820.0331
## 6   - height  1 90.266852       130   53765.40 818.2599
```

```
summary(ccstep)
```

```
##
## Call:
## lm(formula = bp.2s ~ age + weight + waist + time.ppn, data = na.omit(data[,
##     !(colnames(data) %in% "bp.1s")]))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.726 -12.211  -4.242   9.884  75.102
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 112.147651  13.780699   8.138 2.82e-13 ***
## age           0.531195   0.127681   4.160 5.74e-05 ***
## weight       -0.117268   0.082579  -1.420    0.158
## waist         1.015679   0.596411   1.703    0.091 .
## time.ppn     -0.015539   0.005947  -2.613    0.010 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.34 on 130 degrees of freedom
## Multiple R-squared:  0.2126, Adjusted R-squared:  0.1883
## F-statistic: 8.774 on 4 and 130 DF,  p-value: 2.658e-06
```

Therefore, in complete case analysis, the model selected is `bp.2s ~ age + weight + waist + time.ppn`