Solutions - Missing data workshop

dr. Iris Eekhout

Zurich | 27 & 28 June 2022

Missing value analyses

Solution 1: amount of missing values

```
# load library mice to load boys data
library(mice)
```

a. How many variables have missing data?

All variables, except for age, have missing values, so in total 8 out of 9 variables have missing data.

```
# summary to see which variables have missing data
summary(boys)
```

```
>       age              hgt              wgt              bmi       
>  Min.   : 0.035   Min.   : 50.00   Min.   :  3.14   Min.   :11.77  
>  1st Qu.: 1.581   1st Qu.: 84.88   1st Qu.: 11.70   1st Qu.:15.90  
>  Median :10.505   Median :147.30   Median : 34.65   Median :17.45  
>  Mean   : 9.159   Mean   :132.15   Mean   : 37.15   Mean   :18.07  
>  3rd Qu.:15.267   3rd Qu.:175.22   3rd Qu.: 59.58   3rd Qu.:19.53  
>  Max.   :21.177   Max.   :198.00   Max.   :117.40   Max.   :31.74  
>                   NA's   :20       NA's   :4        NA's   :21     
>        hc          gen        phb            tv           reg     
>  Min.   :33.70   G1  : 56   P1  : 63   Min.   : 1.00   north: 81  
>  1st Qu.:48.12   G2  : 50   P2  : 40   1st Qu.: 4.00   east :161  
>  Median :53.00   G3  : 22   P3  : 19   Median :12.00   west :239  
>  Mean   :51.51   G4  : 42   P4  : 32   Mean   :11.89   south:191  
>  3rd Qu.:56.00   G5  : 75   P5  : 50   3rd Qu.:20.00   city : 73  
>  Max.   :65.00   NA's:503   P6  : 41   Max.   :25.00   NA's :  3  
>  NA's   :46                 NA's:503   NA's   :522                
```

b. How many rows in the data contain missing values?

In total 525 rows in the data have missing values, this is ~70%.

```
nic(boys)
```

```
> [1] 525
```

```
nic(boys)/nrow(boys)
```

```
> [1] 0.7018717
```

c. How many overall matrix entries are missing? And how many observed?

1622 matrix entries are missing and 5110 are observed; ~25% of the matrix entries are missing.

```
sum(is.na(boys))
```

```
> [1] 1622
```

```
sum(!is.na(boys))
```

```
> [1] 5110
```

```
1622/(1622+5110)
```

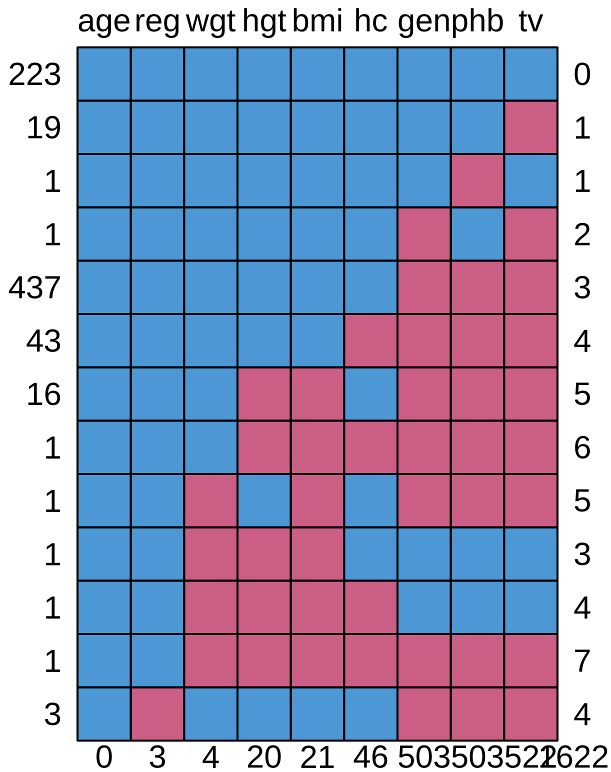
```
> [1] 0.2409388
```

Solution 2: missing data patterns

a. How many different missing data patterns occur in the data?

14 patterns

```
nrow(mice::md.pattern(boys, plot= T))
```

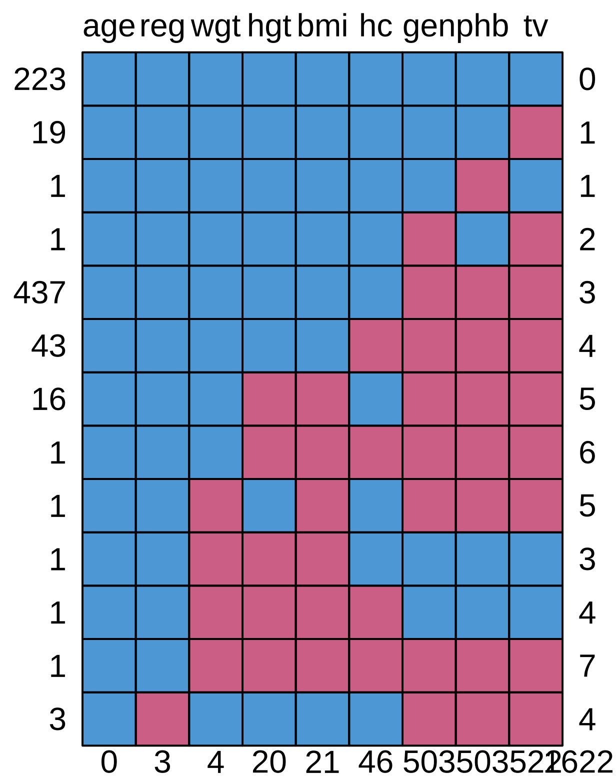



```
> [1] 14
```

b. What is the most frequently occurring pattern in the data?

The pattern with "gen", "phb" and "tv" missing, occurs 437 times, (i.e. 1,1,1,1,1,1,0,0,0).

```
mice::md.pattern(boys, plot= T)
```




```
>      age reg wgt hgt bmi hc gen phb  tv
> 223    1   1   1   1   1  1   1   1   1    0
> 19     1   1   1   1   1  1   1   1   0    1
> 1      1   1   1   1   1  1   1   0   1    1
> 1      1   1   1   1   1  1   0   1   0    2
> 437    1   1   1   1   1  1   0   0   0    3
> 43     1   1   1   1   1  0   0   0   0    4
> 16     1   1   1   0   0  1   0   0   0    5
> 1      1   1   1   0   0  0   0   0   0    6
> 1      1   1   0   1   0  1   0   0   0    5
> 1      1   1   0   0   0  1   1   1   1    3
> 1      1   1   0   0   0  0   1   1   1    4
> 1      1   1   0   0   0  0   0   0   0    7
> 3      1   0   1   1   1  1   0   0   0    4
>        0   3   4  20  21 46 503 503 522 1622
```

c. Looking at patterns that occur more than incidental (once or twice), which variables happen to be missing together often?

Variables that are most often missing at the same time are “gen”, “phb”, and “tv”. The patterns that occur more than once involve mostly all of these variables (pattern with “hc” and “gen”, “pbh”, “tv” 43 times and the pattern with “hgt”, “bmi” and “gen”, “phb”, “tv” 16 times, pattern “reg” and “gen”, “phb”, “tv”, 3 times, pattern with “tv” mising 19 times).

We can also see this in the `mm` matrix that is returned by the `md.pairs` function. The variables “gen”, “phb” and “tv” are missing together more than 500 times.

```
mice::md.pairs(boys)$mm
```

```
>     age hgt wgt bmi hc gen phb  tv reg
> age   0   0   0   0  0   0   0   0   0
> hgt   0  20   3  20  3  18  18  18   0
> wgt   0   3   4   4  2   2   2   2   0
> bmi   0  20   4  21  3  19  19  19   0
> hc    0   3   2   3 46  45  45  45   0
> gen   0  18   2  19 45 503 502 503   3
> phb   0  18   2  19 45 502 503 502   3
> tv    0  18   2  19 45 503 502 522   3
> reg   0   0   0   0  0   3   3   3   3
```

d. Inspect the missing data pairs. With what other variable(s) is height observed together with in more than half of the cases?

The answer can be found looking at the first matrix `rr`. In the row of "hgt", the column values that are higher than 374 indicate variables that are observed with hgt more than half of the time: "age", "wgt", "bmi", "hc", and "reg".

```
mice::md.pairs(boys)$rr
```

```
>     age hgt wgt bmi  hc gen phb  tv reg
> age 748 728 744 727 702 245 245 226 745
> hgt 728 728 727 727 685 243 243 224 725
> wgt 744 727 744 727 700 243 243 224 741
> bmi 727 727 727 727 684 243 243 224 724
> hc  702 685 700 684 702 244 244 225 699
> gen 245 243 243 243 244 245 244 226 245
> phb 245 243 243 243 244 244 245 225 245
> tv  226 224 224 224 225 226 225 226 226
> reg 745 725 741 724 699 245 245 226 745
```

Solution 3: understanding missing data mechanisms

a. What is the mean and standard deviation of knee pain score? And the association between BMI and knee pain (coefficient, standard error and p-value)?

Mean= 14.81, sd=3.21; The association is significant with coefficient=0.35; se=0.14.

b. What are the mean and standard deviation of the knee pain score? What is association between BMI and knee pain?

Mean= 14.70, sd=3.24; The association is not significant (coefficient=0.33; se=0.18).

c. How do these results compare to the complete data results?

The mean and standard deviation have not changed much, however the power for the association was decreased which caused the association between BMI and knee pain to not be significant anymore.

d. What happens to the association between BMI and knee pain? Explain differences with the previous answer (sample size 100).

At 0% missing: coefficient=0.57 and se=0.08 (significant); at 30% missing: coefficient=0.53 and se=0.10 (significant). The association does not change (much) and remains significant. The difference with answer c is explained by the change in sample size. Larger sample size, makes the analysis more robust against (MCAR) missing data.

e. What is the association between BMI and knee pain? How does this compare to the association when the data were MCAR?

There is a significant association with a coefficient of 0.49, se=0.11; the association is now less strong; the coefficient is lower (biased) (was 0.57 for 0% missing and 0.53 for 30% MCAR) and the standard error is similar to the standard error for MCAR (30% MCAR se = 0.10).

f. When there are 30% MAR missing data at sample size 250, at what BMI values do missing data on knee pain occur (inspect the scatterplot and the boxplots).

More missing values at higher BMI values. This also explains why the association between BMI and Knee pain in the MAR data is less strong.

g. Comparing the histograms, what knee pain values are mostly missing?

In the MNAR situation, mostly the higher values of knee pain are missing.

h. What happens with the association between BMI and knee pain?

- MCAR: coefficient= 0.46; se=0.10
- MAR: coefficient=0.37; se=0.11
- MNAR: coefficient=0.38; se=0.08
- 0% missing: coefficient=0.54; se=0.07.

The association becomes less strong when you change from MCAR to MAR to MNAR, so coefficients get more biased. Also for MCAR the missings are nicely distributed over the BMI values. In the MAR mechanism there are more missings at higher values of BMI but also lower Knee Pain scores are missing. In the MNAR mechanism, mostly higher values of Knee Pain scores are missing.

i What happens with the mean and standard deviation of the knee pain score?

- MCAR: mean=14.79; sd=2.98
- MAR: mean=14.01; sd=2.85
- MNAR: mean= 12.66; sd=2.21
- 0%missing: mean=14.77; sd=3.11

Both mean and standard deviation decrease when changing from MCAR to MAR to MNAR. The MNAR data results in a more larger decrease in mean and sd, compared with having MCAR or MAR data.

Solution 4: evaluating the missing data mechanism

a. Evaluate the missing data mechanism for the airquality data with univariate tests. What are your conclusions?

Evaluation using T-tests for continuous variables and the Chi-square for categorical variables.

First create the missing data indicators for each variable with missing data.

```
library(dplyr)
summary(airquality)
```

```
>      Ozone           Solar.R           Wind             Temp
>  Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00
>  1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:72.00
```

```
>  Median : 31.50   Median :205.0   Median : 9.700   Median :79.00
>  Mean   : 42.13   Mean   :185.9   Mean   : 9.958   Mean   :77.88
>  3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00
>  Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00
>  NA's   :37       NA's   :7
>      Month            Day
>  Min.   :5.000   Min.   : 1.0
>  1st Qu.:6.000   1st Qu.: 8.0
>  Median :7.000   Median :16.0
>  Mean   :6.993   Mean   :15.8
>  3rd Qu.:8.000   3rd Qu.:23.0
>  Max.   :9.000   Max.   :31.0
>
```

```
airqualitym <- airquality %>%
  #create missing data indicators
  mutate(ROzone = is.na(Ozone),
         RSolar.R = is.na(Solar.R))
```

Do a T-test for each missing data indicator with the continuous variables and a chi square test for the categorical variables. We investigate Month as categorical, but it can also be used as continuous.

- Missing data in Ozone are related to Month, the probability for missing Ozone data is higher earlier in the year.
- The Chi-square test Solar.R and Month seems significant, but also throws a warning. So we cannot really be sure about the results. Probably also because there are only few missings in Solar.R. Tests for the other variables do not show a relation with the probability of missing data in Solar.R

Based on the univariate analyses we may conclude that the missing data are not Missing Completely at Random (not-MCAR). There are other measured variable related to the probability of missing data.

```
# Univariate tests for Ozone
t.test(Solar.R ~ ROzone, data = airqualitym)
```

```
>
>   Welch Two Sample t-test
>
> data:  Solar.R by ROzone
> t = -0.27457, df = 58.995, p-value = 0.7846
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>  -39.05621  29.63124
> sample estimates:
> mean in group FALSE  mean in group TRUE
>            184.8018            189.5143
```

```
t.test(Wind ~ ROzone, data = airqualitym)
```

```
>
>   Welch Two Sample t-test
>
```

```
> data:  Wind by ROzone
> t = -0.60911, df = 63.646, p-value = 0.5446
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>  -1.6893132  0.8999377
> sample estimates:
> mean in group FALSE  mean in group TRUE
>            9.862069           10.256757
```

```
t.test(Temp ~ ROzone, data = airqualitym)
```

```
>
>   Welch Two Sample t-test
>
> data:  Temp by ROzone
> t = -0.026831, df = 60.447, p-value = 0.9787
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>  -3.643306  3.546847
> sample estimates:
> mean in group FALSE  mean in group TRUE
>            77.87069            77.91892
```

```
t.test(Month ~ ROzone, data = airqualitym)
```

```
>
>   Welch Two Sample t-test
>
> data:  Month by ROzone
> t = 4.0092, df = 92.075, p-value = 0.0001236
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>  0.4273815 1.2664675
> sample estimates:
> mean in group FALSE  mean in group TRUE
>            7.198276            6.351351
```

```
t.test(Day ~ ROzone, data = airqualitym)
```

```
>
>   Welch Two Sample t-test
>
> data:  Day by ROzone
> t = -0.64426, df = 57.826, p-value = 0.522
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>  -4.576080  2.347749
> sample estimates:
> mean in group FALSE  mean in group TRUE
>            15.53448            16.64865
```

```r
chisq.test(airqualitym$ROzone, airqualitym$Month)
```

```
>
>   Pearson's Chi-squared test
>
> data:  airqualitym$ROzone and airqualitym$Month
> X-squared = 44.751, df = 4, p-value = 4.48e-09
```

```r
# Univariate tests for Solar.R
t.test(Ozone ~ RSolar.R, data = airqualitym)
```

```
>
>   Welch Two Sample t-test
>
> data:  Ozone by RSolar.R
> t = -0.052696, df = 4.4917, p-value = 0.9602
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>  -36.0892  34.6874
> sample estimates:
> mean in group FALSE  mean in group TRUE
>             42.0991             42.8000
```

```r
t.test(Wind ~ RSolar.R, data = airqualitym)
```

```
>
>   Welch Two Sample t-test
>
> data:  Wind by RSolar.R
> t = 0.65629, df = 6.4571, p-value = 0.5343
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>  -2.674488  4.681338
> sample estimates:
> mean in group FALSE  mean in group TRUE
>            10.00342             9.00000
```

```r
t.test(Temp ~ RSolar.R, data = airqualitym)
```

```
>
>   Welch Two Sample t-test
>
> data:  Temp by RSolar.R
> t = 0.98706, df = 6.2689, p-value = 0.3602
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>  -7.436381 17.669258
> sample estimates:
> mean in group FALSE  mean in group TRUE
>            78.11644            73.00000
```

```r
t.test(Month ~ RSolar.R, data = airqualitym)
```

```
>
>   Welch Two Sample t-test
>
> data:  Month by RSolar.R
> t = 1.2018, df = 6.4489, p-value = 0.2717
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>  -0.7432664  2.2266323
> sample estimates:
> mean in group FALSE  mean in group TRUE
>            7.027397            6.285714
```

```r
t.test(Day ~ RSolar.R, data = airqualitym)
```

```
>
>   Welch Two Sample t-test
>
> data:  Day by RSolar.R
> t = 2.1941, df = 6.6803, p-value = 0.06612
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>  -0.6161166 14.5769776
> sample estimates:
> mean in group FALSE  mean in group TRUE
>           16.123288            9.142857
```

```r
chisq.test(airqualitym$RSolar.R, airqualitym$Month)
```

```
> Warning in chisq.test(airqualitym$RSolar.R, airqualitym$Month): Chi-squared
> approximation may be incorrect
```

```
>
>   Pearson's Chi-squared test
>
> data:  airqualitym$RSolar.R and airqualitym$Month
> X-squared = 11.136, df = 4, p-value = 0.02507
```

b. Evaluate the missing data mechanism for the airquality data with a multivariate test. What are your conclusions?

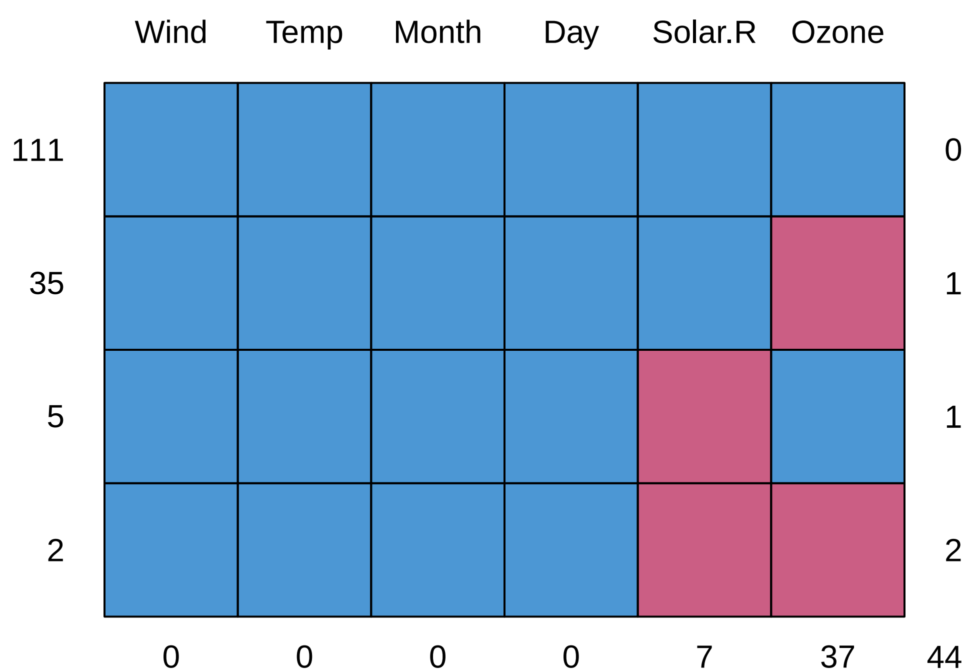
First, investigate which variables have missing data in the `airquality` data: Ozone and Solar.R

```r
summary(airquality)
```

```
>      Ozone           Solar.R           Wind             Temp
>  Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00
>  1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:72.00
>  Median : 31.50   Median :205.0   Median : 9.700   Median :79.00
>  Mean   : 42.13   Mean   :185.9   Mean   : 9.958   Mean   :77.88
```

```
>  3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00
>  Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00
>  NA's   :37       NA's   :7
>      Month            Day
>  Min.   :5.000   Min.   : 1.0
>  1st Qu.:6.000   1st Qu.: 8.0
>  Median :7.000   Median :16.0
>  Mean   :6.993   Mean   :15.8
>  3rd Qu.:8.000   3rd Qu.:23.0
>  Max.   :9.000   Max.   :31.0
>
```

```
mice::md.pattern(airquality)
```




```
>     Wind Temp Month Day Solar.R Ozone
> 111    1    1     1   1       1     1  0
> 35     1    1     1   1       1     0  1
> 5      1    1     1   1       0     1  1
> 2      1    1     1   1       0     0  2
>        0    0     0   0       7    37 44
```

Create missing data indicators for these two variables. Additional, we can also make one missing indicator for any missing values. Note that this is only useful if we have at least some variables that have no missing data.

```r
airqualitym <- airquality %>%
  mutate(ROzone = is.na(Ozone),
         RSolar.R = is.na(Solar.R),
         Rind = is.na(Ozone) | is.na(Solar.R))
```

Do a logistic regression analysis for each of the missing data indicators. Both Temp and Month seem to be related to the missing values in Ozone. There are no measured variables related to the missing values in Solar.R. Based on these results we can conclude that the missing values in the airquality dataset are not-MCAR.

```r
glm(ROzone ~ Solar.R + Wind + Temp + Month + Day, data = airqualitym, family = "binomial") %>% summary
```

```
>
> Call:
> glm(formula = ROzone ~ Solar.R + Wind + Temp + Month + Day, family = "binomial",
>     data = airqualitym)
>
> Deviance Residuals:
>     Min       1Q   Median       3Q      Max
> -1.4547  -0.8277  -0.5250  -0.2253   2.2683
>
> Coefficients:
>              Estimate Std. Error z value Pr(>|z|)
> (Intercept) -3.028609   2.316740  -1.307 0.191120
> Solar.R     -0.002155   0.002496  -0.864 0.387769
> Wind         0.057605   0.063930   0.901 0.367552
> Temp         0.081839   0.031363   2.609 0.009069 **
> Month       -0.726536   0.218087  -3.331 0.000864 ***
> Day          0.012030   0.023231   0.518 0.604555
> ---
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> (Dispersion parameter for binomial family taken to be 1)
>
>     Null deviance: 160.82  on 145  degrees of freedom
> Residual deviance: 144.39  on 140  degrees of freedom
>   (7 observations deleted due to missingness)
> AIC: 156.39
>
> Number of Fisher Scoring iterations: 5
```

```r
glm(RSolar.R ~ Ozone + Wind + Temp + Month + Day, data = airqualitym, family = "binomial") %>% summary
```

```
>
> Call:
> glm(formula = RSolar.R ~ Ozone + Wind + Temp + Month + Day, family = "binomial",
>     data = airqualitym)
>
> Deviance Residuals:
>      Min        1Q    Median        3Q       Max
> -0.84141  -0.27298  -0.15226  -0.06878   2.54340
>
```

```
> Coefficients:
>              Estimate Std. Error z value Pr(>|z|)
> (Intercept)  0.08394    6.41576   0.013   0.9896
> Ozone       -0.02426    0.02503  -0.969   0.3324
> Wind        -0.22019    0.21582  -1.020   0.3076
> Temp         0.06002    0.10130   0.592   0.5536
> Month       -0.46001    0.50740  -0.907   0.3646
> Day         -0.16317    0.08894  -1.835   0.0666 .
> ---
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> (Dispersion parameter for binomial family taken to be 1)
>
>     Null deviance: 41.223  on 115  degrees of freedom
> Residual deviance: 32.229  on 110  degrees of freedom
>   (37 observations deleted due to missingness)
> AIC: 44.229
>
> Number of Fisher Scoring iterations: 7
```

```
# analysis for the indicator of overall missing data
glm(Rind ~ Wind + Temp + Month + Day, data = airqualitym, family = "binomial") %>% summary
```

```
>
> Call:
> glm(formula = Rind ~ Wind + Temp + Month + Day, family = "binomial",
>     data = airqualitym)
>
> Deviance Residuals:
>     Min       1Q   Median       3Q      Max
> -1.3113  -0.8743  -0.5758   1.2091   2.2435
>
> Coefficients:
>              Estimate Std. Error z value Pr(>|z|)
> (Intercept) -0.665206   2.083304  -0.319 0.749496
> Wind         0.012608   0.059386   0.212 0.831873
> Temp         0.050067   0.025975   1.927 0.053922 .
> Month       -0.631889   0.186734  -3.384 0.000715 ***
> Day         -0.003769   0.021176  -0.178 0.858728
> ---
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> (Dispersion parameter for binomial family taken to be 1)
>
>     Null deviance: 179.83  on 152  degrees of freedom
> Residual deviance: 165.15  on 148  degrees of freedom
> AIC: 175.15
>
> Number of Fisher Scoring iterations: 4
```

Multiple imputation

Solution 5: multiple imputation in `mice`

a. How many imputed datasets are generated?

The output returned from the mice functions states: "Number of multiple imputations: 5". *Note that in order to obtain exactly the same results as in this document, set* `seed = 1234`*.*

```r
imp <- mice(nhanes2, seed = 1234)
```

```
>
>  iter imp variable
>   1   1  bmi  hyp  chl
>   1   2  bmi  hyp  chl
>   1   3  bmi  hyp  chl
>   1   4  bmi  hyp  chl
>   1   5  bmi  hyp  chl
>   2   1  bmi  hyp  chl
>   2   2  bmi  hyp  chl
>   2   3  bmi  hyp  chl
>   2   4  bmi  hyp  chl
>   2   5  bmi  hyp  chl
>   3   1  bmi  hyp  chl
>   3   2  bmi  hyp  chl
>   3   3  bmi  hyp  chl
>   3   4  bmi  hyp  chl
>   3   5  bmi  hyp  chl
>   4   1  bmi  hyp  chl
>   4   2  bmi  hyp  chl
>   4   3  bmi  hyp  chl
>   4   4  bmi  hyp  chl
>   4   5  bmi  hyp  chl
>   5   1  bmi  hyp  chl
>   5   2  bmi  hyp  chl
>   5   3  bmi  hyp  chl
>   5   4  bmi  hyp  chl
>   5   5  bmi  hyp  chl
```

```r
imp
```

```
> Class: mids
> Number of multiple imputations:  5
> Imputation methods:
>      age      bmi      hyp      chl
>       ""    "pmm" "logreg"    "pmm"
> PredictorMatrix:
>     age bmi hyp chl
> age   0   1   1   1
> bmi   1   0   1   1
> hyp   1   1   0   1
> chl   1   1   1   0
```

b. How many sets of results are generated

The `with` function can be used to automatically analyze all imputed datasets. The analysis results are stored as `fit`. There are 5 sets of results generated.

```
fit <- with(imp, lm(bmi ~ age + hyp + chl))
fit
```

```
> call :
> with.mids(data = imp, expr = lm(bmi ~ age + hyp + chl))
>
> call1 :
> mice(data = nhanes2, seed = 1234)
>
> nmis :
> age bmi hyp chl
>   0   9   8  10
>
> analyses :
> [[1]]
>
> Call:
> lm(formula = bmi ~ age + hyp + chl)
>
> Coefficients:
> (Intercept)     age40-59     age60-99       hypyes          chl
>    19.20198     -6.19523     -7.58670      1.83049      0.05459
>
>
> [[2]]
>
> Call:
> lm(formula = bmi ~ age + hyp + chl)
>
> Coefficients:
> (Intercept)     age40-59     age60-99       hypyes          chl
>    20.85742     -4.53480     -7.92024      1.28448      0.04507
>
>
> [[3]]
>
> Call:
> lm(formula = bmi ~ age + hyp + chl)
>
> Coefficients:
> (Intercept)     age40-59     age60-99       hypyes          chl
>    21.41965     -4.86066     -6.28743      2.49869      0.04224
>
>
> [[4]]
>
> Call:
> lm(formula = bmi ~ age + hyp + chl)
>
```

```
> Coefficients:
> (Intercept)     age40-59     age60-99       hypyes          chl
>    18.54737     -4.71719     -6.35810      3.75288      0.05262
>
>
> [[5]]
>
> Call:
> lm(formula = bmi ~ age + hyp + chl)
>
> Coefficients:
> (Intercept)     age40-59     age60-99       hypyes          chl
>    18.35544     -5.28014     -6.43238      2.65880      0.05553
```

c. What are the final parameter estimates of the regression analysis?

The parameter estimates are summarized in the table below. Older respondents tend to have a lower BMI and cholesterol is positively related to BMI.

```r
combi <- pool(fit)
summary(combi) %>% knitr::kable(digits = 3)
```

term	estimate	std.error	statistic	df	p.value
(Intercept)	19.676	3.409	5.772	12.800	0.000
age40-59	-5.118	1.729	-2.960	13.476	0.011
age60-99	-6.917	2.054	-3.367	13.644	0.005
hypyes	2.405	1.857	1.295	9.863	0.225
chl	0.050	0.019	2.698	15.041	0.017

Solution 6: multiple imputation model and convergence

a. Adjust the predictor matrix so that the variables that have more than 50% missing values are excluded as predictors for the imputation.

First create the predictor matrix for the `boys` dataset with the `make.predictorMatrix()` function in `mice`.

```r
pred <- make.predictorMatrix(boys)
```

Inspect the boys dataset, and find out what variables have more than 50% missing values. These are “gen”, “phb”, and “tv”.

```r
colMeans(is.na(boys))
```

```
>         age         hgt         wgt         bmi          hc         gen
> 0.000000000 0.026737968 0.005347594 0.028074866 0.061497326 0.672459893
>         phb          tv         reg
> 0.672459893 0.697860963 0.004010695
```

Now, exclude variables “gen”, “phb”, and “tv” as predictors from this matrix. Note that predictors are in the columns.

```r
pred[,c("gen", "phb", "tv")] <- 0
pred
```

```
>     age hgt wgt bmi hc gen phb tv reg
> age   0   1   1   1  1   0   0  0   1
> hgt   1   0   1   1  1   0   0  0   1
> wgt   1   1   0   1  1   0   0  0   1
> bmi   1   1   1   0  1   0   0  0   1
> hc    1   1   1   1  0   0   0  0   1
> gen   1   1   1   1  1   0   0  0   1
> phb   1   1   1   1  1   0   0  0   1
> tv    1   1   1   1  1   0   0  0   1
> reg   1   1   1   1  1   0   0  0   0
```

Perform multiple imputation on the `boys` data with the predictor matrix designed at assignment 6a with 10 imputations and 10 iterations.

```r
imp <- mice(boys, m = 10, maxit = 10, predictorMatrix = pred)
```

```
> 
>  iter imp variable
>   1   1  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   1   2  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   1   3  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   1   4  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   1   5  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   1   6  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   1   7  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   1   8  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   1   9  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   1   10  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   2   1  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   2   2  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   2   3  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   2   4  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   2   5  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   2   6  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   2   7  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   2   8  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   2   9  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   2   10  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   3   1  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   3   2  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   3   3  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   3   4  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   3   5  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   3   6  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   3   7  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   3   8  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   3   9  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   3   10  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   4   1  hgt  wgt  bmi  hc  gen  phb  tv  reg
```

```
>   4   2  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   4   3  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   4   4  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   4   5  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   4   6  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   4   7  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   4   8  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   4   9  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   4   10  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   5   1  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   5   2  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   5   3  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   5   4  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   5   5  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   5   6  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   5   7  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   5   8  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   5   9  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   5   10  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   6   1  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   6   2  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   6   3  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   6   4  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   6   5  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   6   6  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   6   7  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   6   8  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   6   9  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   6   10  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   7   1  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   7   2  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   7   3  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   7   4  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   7   5  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   7   6  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   7   7  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   7   8  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   7   9  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   7   10  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   8   1  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   8   2  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   8   3  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   8   4  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   8   5  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   8   6  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   8   7  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   8   8  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   8   9  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   8   10  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   9   1  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   9   2  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   9   3  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   9   4  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   9   5  hgt  wgt  bmi  hc  gen  phb  tv  reg
```

```
>   9   6  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   9   7  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   9   8  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   9   9  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   9   10  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   10   1  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   10   2  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   10   3  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   10   4  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   10   5  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   10   6  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   10   7  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   10   8  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   10   9  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   10   10  hgt  wgt  bmi  hc  gen  phb  tv  reg
```

b. What methods are used for the imputation of each variable and explain why these are used.

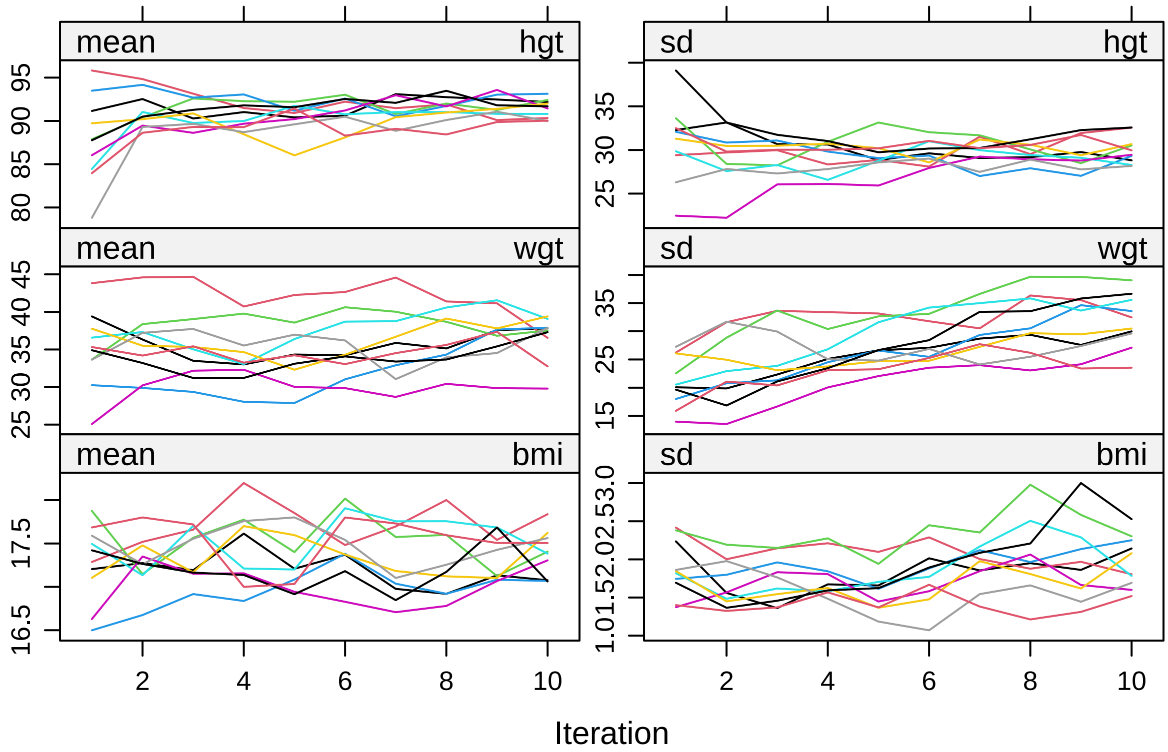
- For “hgt”, “wgt”, “bmi”, “hc” and “tv” the “pmm” (predictive mean matching) method is used. This is the default for continuous variables and the variables indicated are all continuous.
- For “gen” and “phb” the “polr” method is used, which is the default for ordinal variables. The imputation function is a proportional odds model.
- For “reg” the “polyreg” method is used. This method is de default for unordered nominal variables and is the polytomous logistic regression.

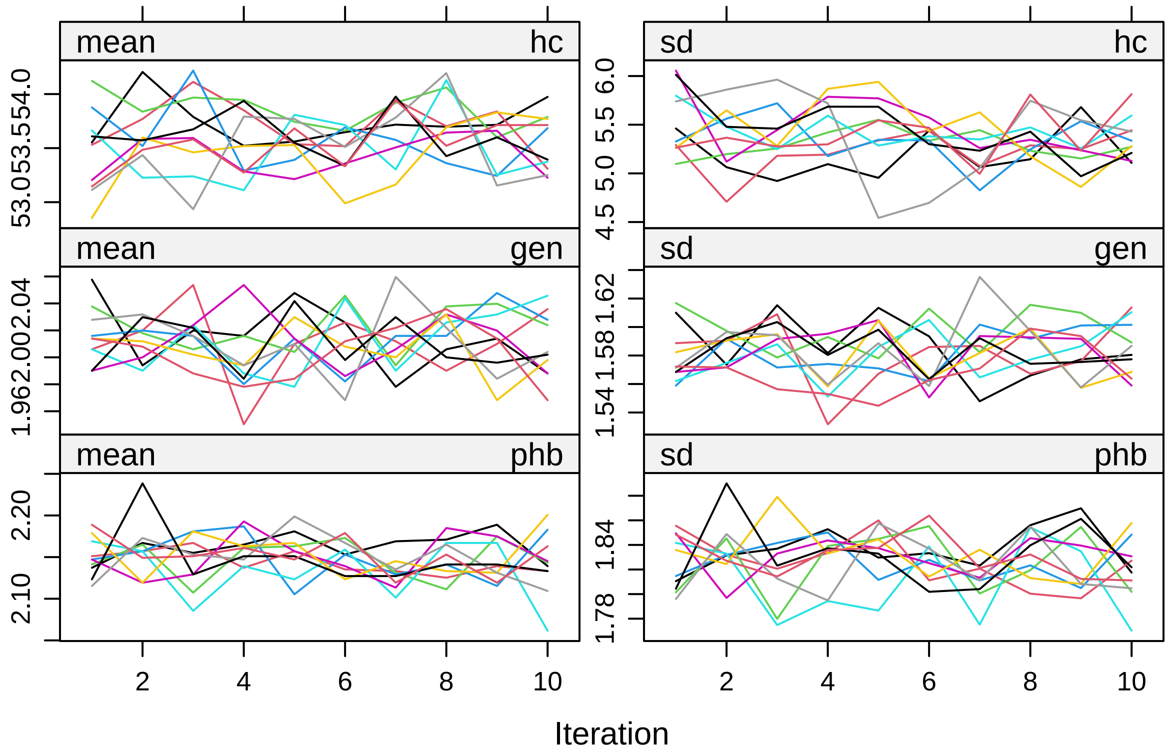
```
imp$method
```

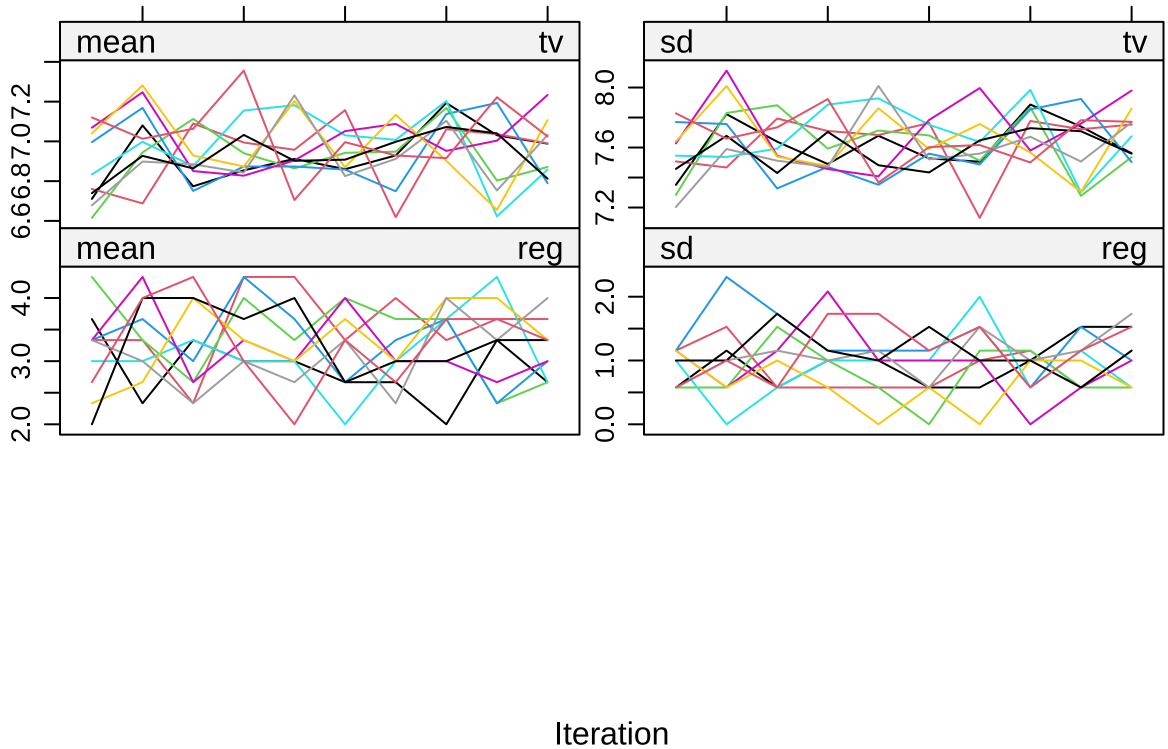
```
>       age       hgt       wgt       bmi        hc       gen       phb        tv
>        ""     "pmm"     "pmm"     "pmm"     "pmm"    "polr"    "polr"     "pmm"
>       reg
> "polyreg"
```

c. Inspect the iteration plots. What are your observations?

```
plot(imp)
```


mean
hgt
sd
hgt
mean
wgt
sd
wgt
mean
bmi
sd
bmi
80 85 90 95
25 30 35
25 30 35 40 45
15 25 35
16.5 17.5
1.0 1.5 2.0 2.5 3.0
2 4 6 8 10
Iteration


mean
hc
sd
hc
mean
gen
sd
gen
mean
phb
sd
phb
53.0
53.5
54.0
4.5
5.0
5.5
6.0
1.96
2.00
2.04
1.54
1.58
1.62
2.10
2.20
1.78
1.84
2
4
6
8
10
Iteration



The iteration plots for the variables “hc”, “gen”, “phb”, “tv”, and “reg” all show crossing interacting lines that are more or less centered around. For the variables “hgt”, “wgt” and “bmi” the lines to not cross as much and some lines are below the mean for all iterations, whereas other lines are above the mean for all iterations. So the iteration plots for “hgt”, “wgt” and “bmi” do not show a very good convergence.

d. Adjust the predictor matrix such that hgt and wgt are not imputed by bmi, or vice versa, and that hgt and wgt are not used as predictors together with bmi. Why do you think that these changes are needed?

Option 1:

First, we remove “hgt” and “wgt” as predictors for the imputation of “bmi”. Second, we remove “bmi” as a predictor for all variables, to ensure that “bmi” is not used as predictor in a model together with “wgt” and “hgt”.

```
pred1 <- pred
pred1["bmi", c("hgt", "wgt")] <- 0
pred1[,"bmi"] <- 0
pred1
```

```
>     age hgt wgt bmi hc gen phb tv reg
> age   0   1   1   0  1   0   0  0   1
> hgt   1   0   1   0  1   0   0  0   1
> wgt   1   1   0   0  1   0   0  0   1
> bmi   1   0   0   0  1   0   0  0   1
> hc    1   1   1   0  0   0   0  0   1
> gen   1   1   1   0  1   0   0  0   1
```

```
> phb    1    1    1    0  1    0    0  0    1
> tv     1    1    1    0  1    0    0  0    1
> reg    1    1    1    0  1    0    0  0    0
```

Option 2:

We remove "hgt" and "wgt" as predictors for all variables, that way they won't be together in the model with "bmi" and they will not be used as predictors for imputing "bmi".

```
pred2 <- pred
pred2[, c("hgt", "wgt")] <- 0
pred2[c("hgt", "wgt"), "bmi"] <- 0
pred2
```

```
>      age hgt wgt bmi hc gen phb tv reg
> age    0   0   0   1  1   0   0  0   1
> hgt    1   0   0   0  1   0   0  0   1
> wgt    1   0   0   0  1   0   0  0   1
> bmi    1   0   0   0  1   0   0  0   1
> hc     1   0   0   1  0   0   0  0   1
> gen    1   0   0   1  1   0   0  0   1
> phb    1   0   0   1  1   0   0  0   1
> tv     1   0   0   1  1   0   0  0   1
> reg    1   0   0   1  1   0   0  0   0
```

Maybe there are other options?

These changes are needed, because bmi is calculated directly from height and weight. So using these variables together results in multi-collinearity problems in the imputation-model.

e. Use the adjusted predictor matrix to impute the boys dataset again, with 10 imputations and 10 iterations. Inspect the iteration plots again, do you see improvements for hgt, wgt and bmi?

Yes, for both options the convergence for "hgt", "wgt" and "bmi" is much better than it was before.

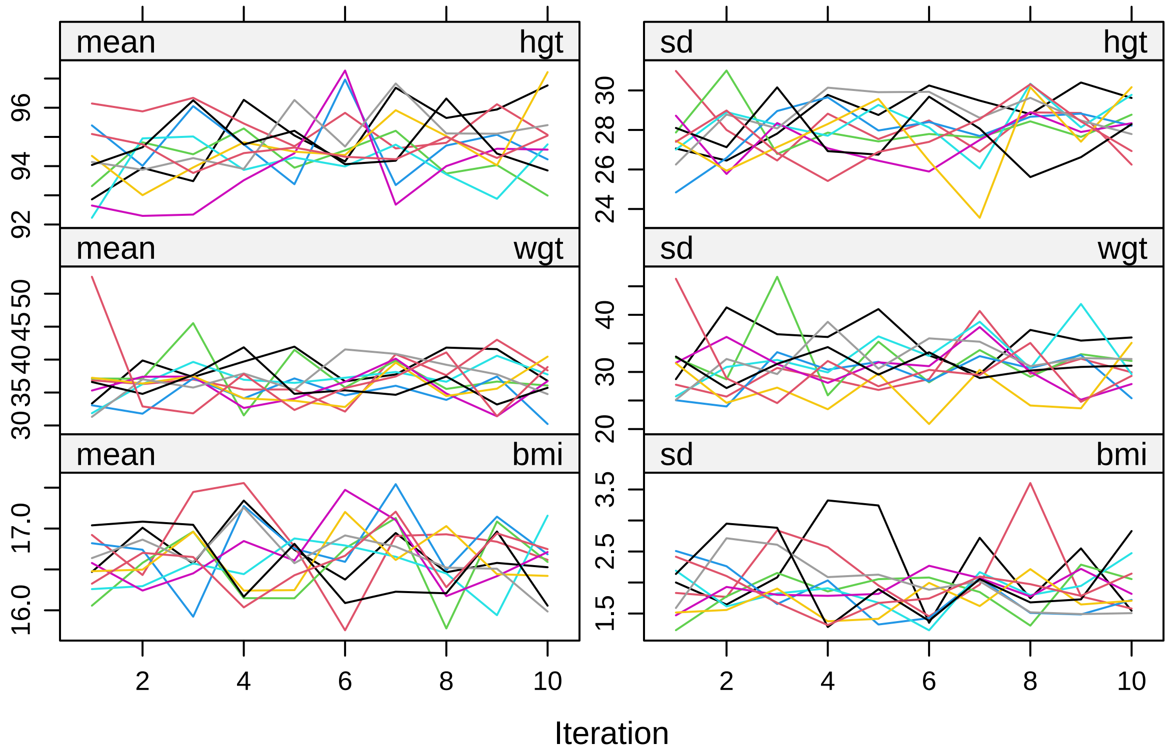
```
imp1 <- mice(boys, m = 10, maxit = 10, pred = pred1)
```

```
>
>  iter imp variable
>   1   1  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   1   2  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   1   3  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   1   4  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   1   5  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   1   6  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   1   7  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   1   8  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   1   9  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   1   10  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   2   1  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   2   2  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   2   3  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   2   4  hgt  wgt  bmi  hc  gen  phb  tv  reg
```

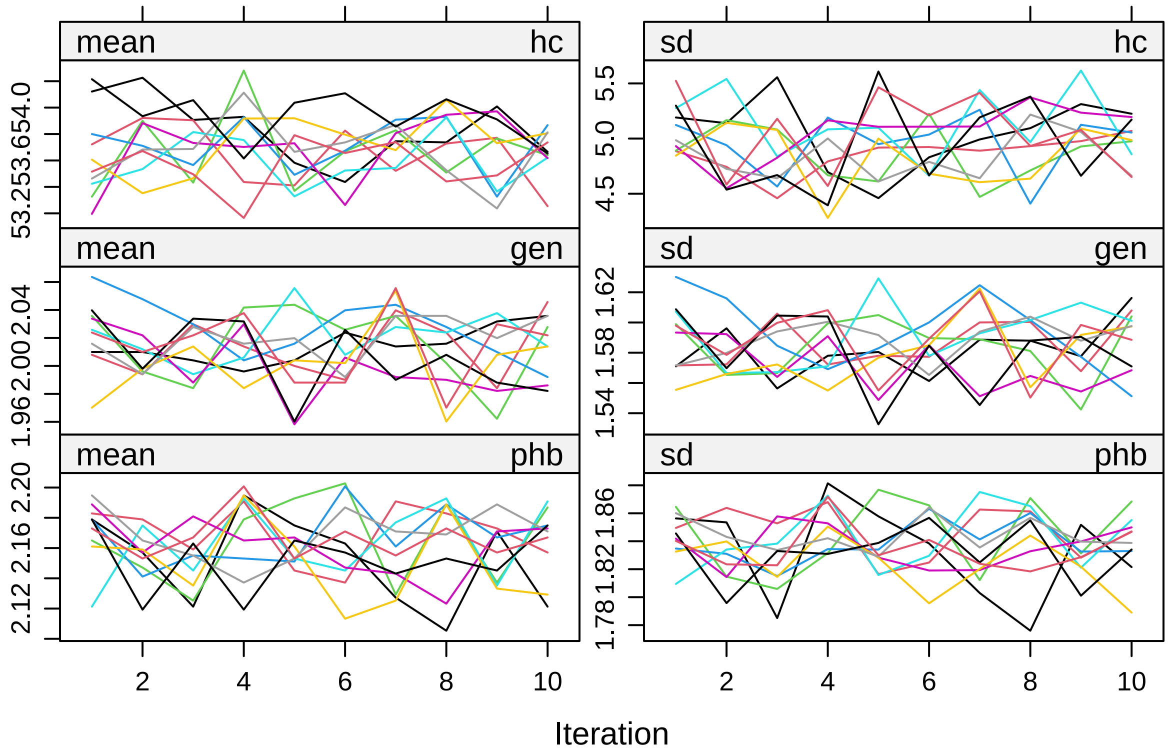
```
>   2   5  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   2   6  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   2   7  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   2   8  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   2   9  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   2   10  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   3   1  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   3   2  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   3   3  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   3   4  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   3   5  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   3   6  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   3   7  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   3   8  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   3   9  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   3   10  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   4   1  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   4   2  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   4   3  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   4   4  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   4   5  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   4   6  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   4   7  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   4   8  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   4   9  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   4   10  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   5   1  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   5   2  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   5   3  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   5   4  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   5   5  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   5   6  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   5   7  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   5   8  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   5   9  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   5   10  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   6   1  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   6   2  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   6   3  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   6   4  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   6   5  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   6   6  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   6   7  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   6   8  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   6   9  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   6   10  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   7   1  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   7   2  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   7   3  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   7   4  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   7   5  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   7   6  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   7   7  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   7   8  hgt  wgt  bmi  hc  gen  phb  tv  reg
```

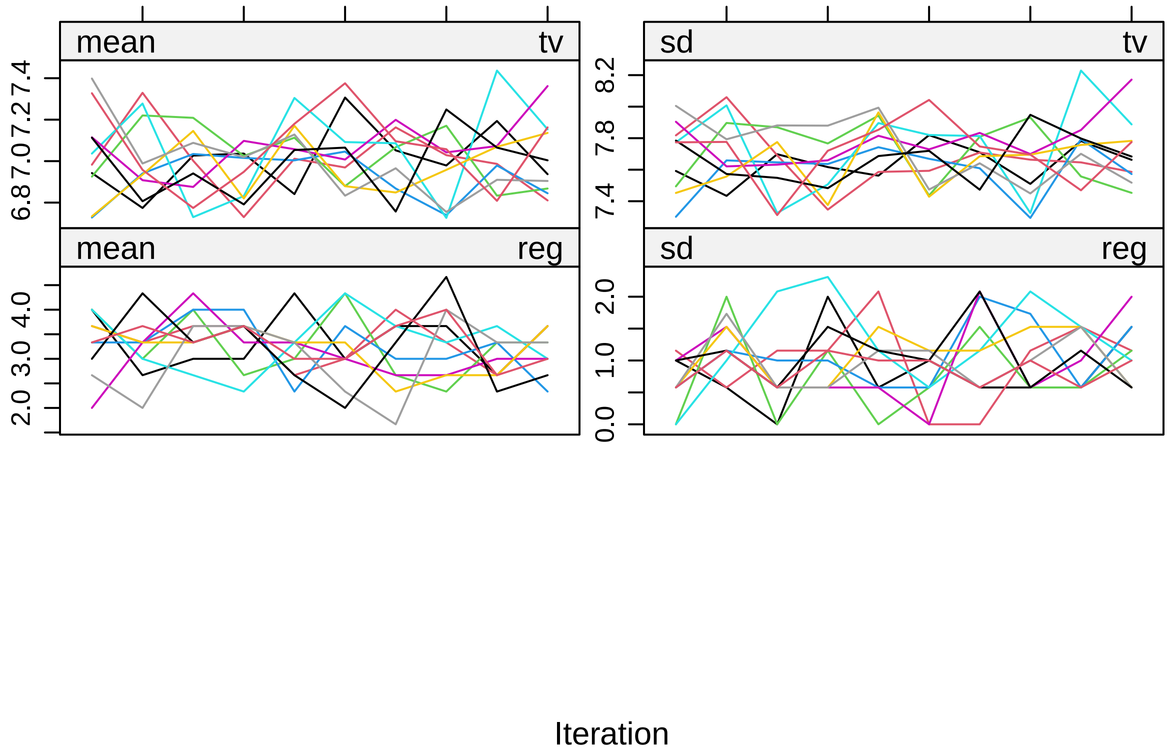
```
>   7   9  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   7   10  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   8   1  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   8   2  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   8   3  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   8   4  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   8   5  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   8   6  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   8   7  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   8   8  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   8   9  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   8   10  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   9   1  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   9   2  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   9   3  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   9   4  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   9   5  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   9   6  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   9   7  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   9   8  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   9   9  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   9   10  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   10   1  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   10   2  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   10   3  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   10   4  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   10   5  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   10   6  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   10   7  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   10   8  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   10   9  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   10   10  hgt  wgt  bmi  hc  gen  phb  tv  reg
```

```
plot(imp1)
```



mean
hgt
sd
hgt
mean
wgt
sd
wgt
mean
bmi
sd
bmi
92
94
96
24 26 28 30
30 35 40 45 50
20
30
40
16.0
17.0
1.5
2.5
3.5
2
4
6
8
10
Iteration


mean
hc
sd
hc
mean
gen
sd
gen
mean
phb
sd
phb
53.2
53.6
54.0
4.5
5.0
5.5
1.96
2.00
2.04
1.54
1.58
1.62
2.12
2.16
2.20
1.78
1.82
1.86
2
4
6
8
10
Iteration



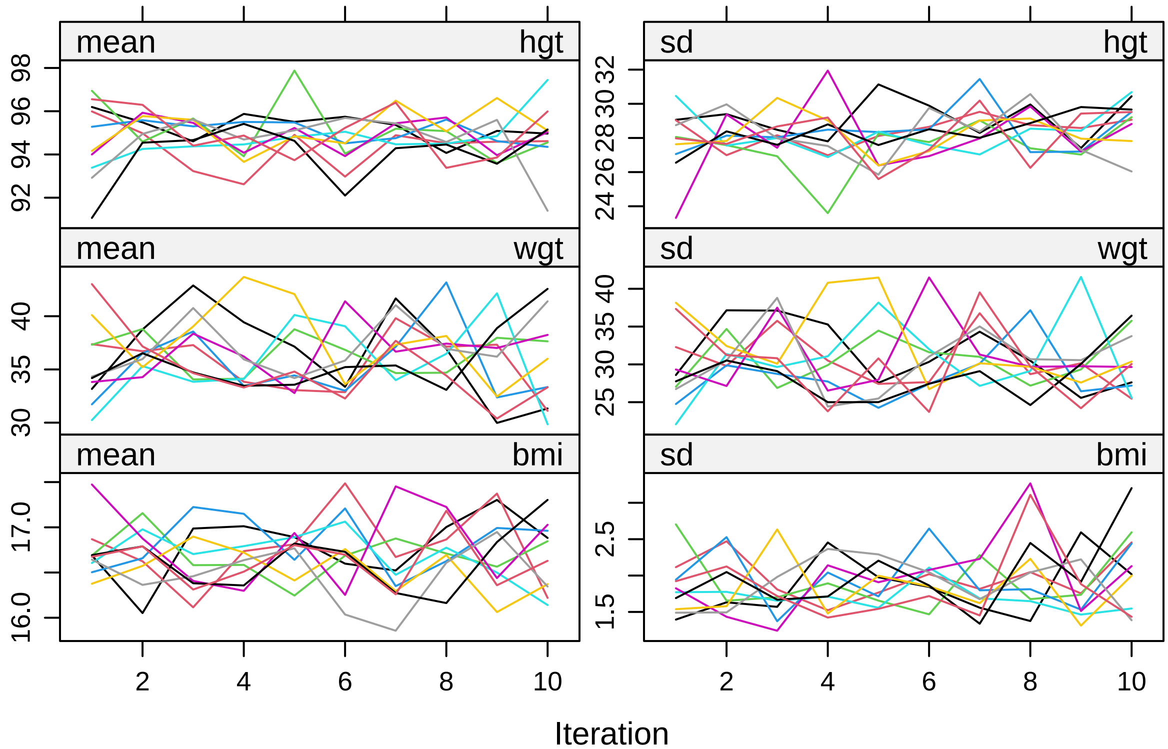

```r
imp2 <- mice(boys, m = 10, maxit = 10, pred = pred2)
```

```
>
> iter imp variable
>   1   1  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   1   2  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   1   3  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   1   4  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   1   5  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   1   6  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   1   7  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   1   8  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   1   9  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   1   10  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   2   1  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   2   2  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   2   3  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   2   4  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   2   5  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   2   6  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   2   7  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   2   8  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   2   9  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   2   10  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   3   1  hgt  wgt  bmi  hc  gen  phb  tv  reg
```

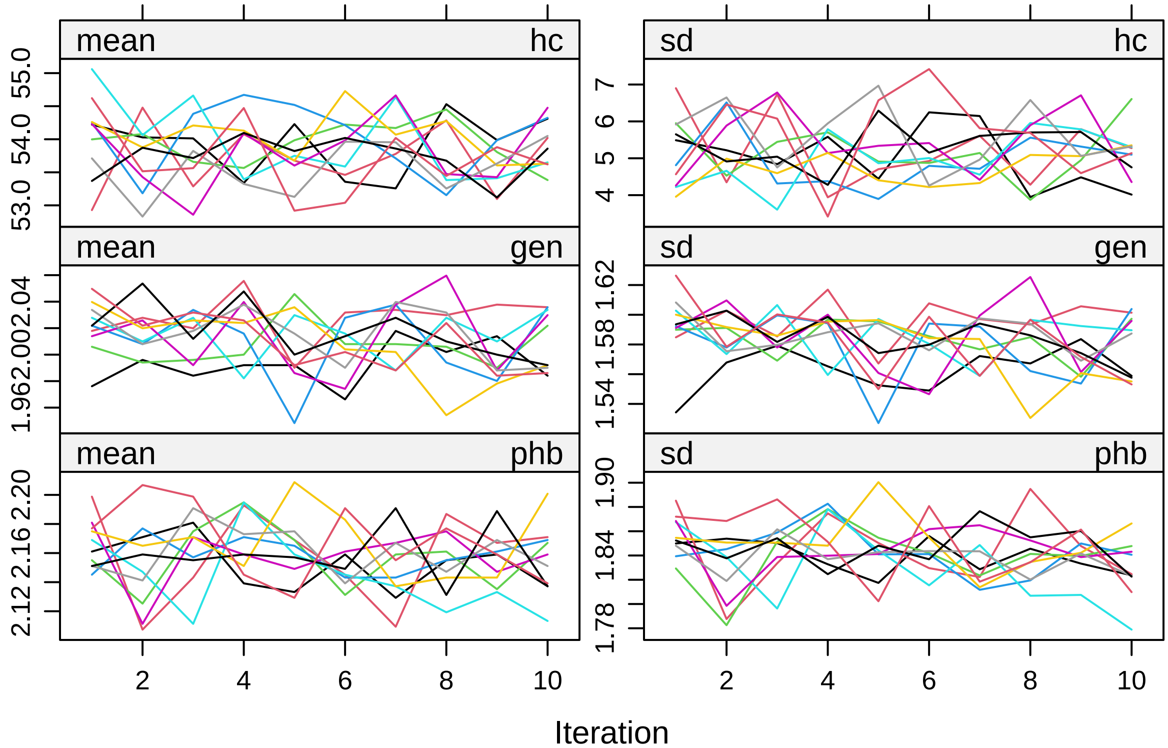
```
>   3   2  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   3   3  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   3   4  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   3   5  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   3   6  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   3   7  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   3   8  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   3   9  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   3   10  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   4   1  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   4   2  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   4   3  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   4   4  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   4   5  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   4   6  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   4   7  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   4   8  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   4   9  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   4   10  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   5   1  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   5   2  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   5   3  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   5   4  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   5   5  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   5   6  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   5   7  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   5   8  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   5   9  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   5   10  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   6   1  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   6   2  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   6   3  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   6   4  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   6   5  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   6   6  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   6   7  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   6   8  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   6   9  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   6   10  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   7   1  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   7   2  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   7   3  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   7   4  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   7   5  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   7   6  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   7   7  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   7   8  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   7   9  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   7   10  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   8   1  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   8   2  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   8   3  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   8   4  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   8   5  hgt  wgt  bmi  hc  gen  phb  tv  reg
```

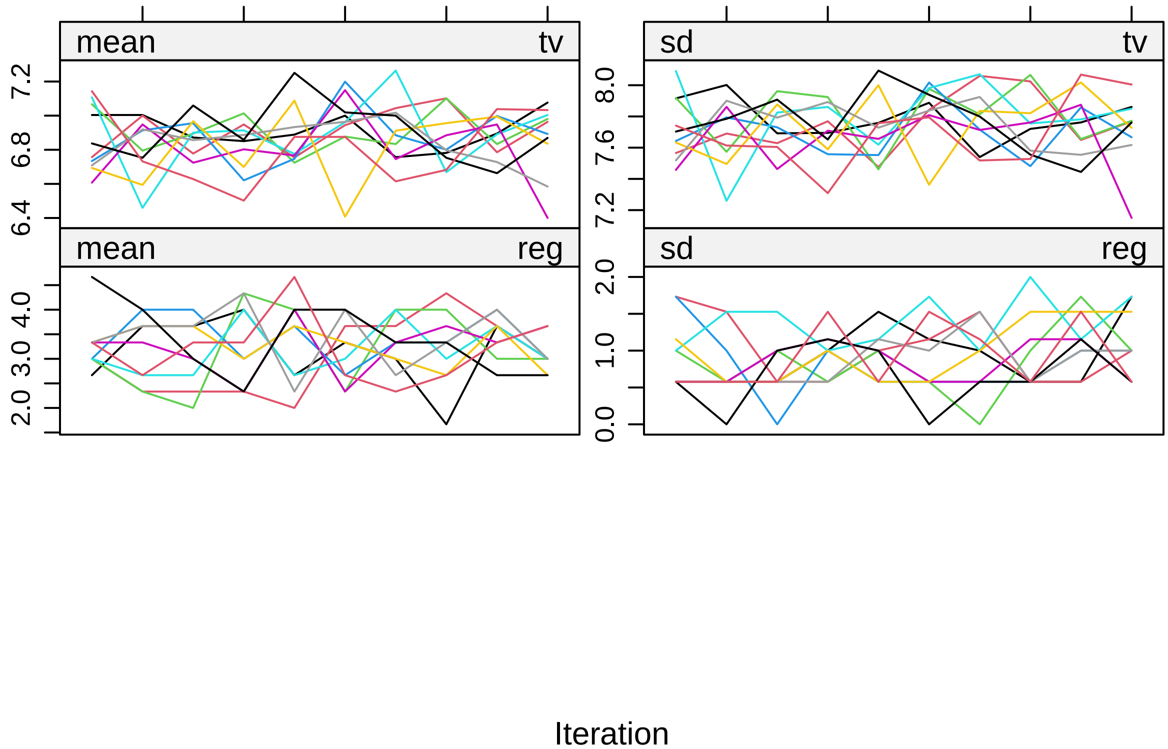
```
>    8    6  hgt  wgt  bmi  hc  gen  phb  tv  reg
>    8    7  hgt  wgt  bmi  hc  gen  phb  tv  reg
>    8    8  hgt  wgt  bmi  hc  gen  phb  tv  reg
>    8    9  hgt  wgt  bmi  hc  gen  phb  tv  reg
>    8    10  hgt  wgt  bmi  hc  gen  phb  tv  reg
>    9    1  hgt  wgt  bmi  hc  gen  phb  tv  reg
>    9    2  hgt  wgt  bmi  hc  gen  phb  tv  reg
>    9    3  hgt  wgt  bmi  hc  gen  phb  tv  reg
>    9    4  hgt  wgt  bmi  hc  gen  phb  tv  reg
>    9    5  hgt  wgt  bmi  hc  gen  phb  tv  reg
>    9    6  hgt  wgt  bmi  hc  gen  phb  tv  reg
>    9    7  hgt  wgt  bmi  hc  gen  phb  tv  reg
>    9    8  hgt  wgt  bmi  hc  gen  phb  tv  reg
>    9    9  hgt  wgt  bmi  hc  gen  phb  tv  reg
>    9    10  hgt  wgt  bmi  hc  gen  phb  tv  reg
>    10    1  hgt  wgt  bmi  hc  gen  phb  tv  reg
>    10    2  hgt  wgt  bmi  hc  gen  phb  tv  reg
>    10    3  hgt  wgt  bmi  hc  gen  phb  tv  reg
>    10    4  hgt  wgt  bmi  hc  gen  phb  tv  reg
>    10    5  hgt  wgt  bmi  hc  gen  phb  tv  reg
>    10    6  hgt  wgt  bmi  hc  gen  phb  tv  reg
>    10    7  hgt  wgt  bmi  hc  gen  phb  tv  reg
>    10    8  hgt  wgt  bmi  hc  gen  phb  tv  reg
>    10    9  hgt  wgt  bmi  hc  gen  phb  tv  reg
>    10    10  hgt  wgt  bmi  hc  gen  phb  tv  reg
```

```
plot(imp2)
```



mean
hgt
sd
hgt
mean
wgt
sd
wgt
mean
bmi
sd
bmi
92 94 96 98
24 26 28 30 32
30 35 40
25 30 35 40
16.0 17.0
1.5 2.5
2 4 6 8 10
Iteration


mean
hc
sd
hc
mean
gen
sd
gen
mean
phb
sd
phb
53.0 54.0 55.0
4 5 6 7
1.96 2.00 2.04
1.54 1.58 1.62
2.12 2.16 2.20
1.78 1.84 1.90
2 4 6 8 10
Iteration



f. Another way to deal with the relation between hgt and wgt with bmi is to use "passive imputation" to impute the bmi variable. Adjust the method and predictor matrix in such a way that the bmi variable is passively imputed by hgt and wgt. Inspect the iteration plots again, do you see improvements for hgt, wgt and bmi?

The formula for bmi is $bmi = \frac{weight}{height^2}$, with height in meters. This formula is added as imputation method to compute missing bmi values from the imputed weight and height values. We use the second option of the predictor matrix (option 1 is also possible), so that the updated bmi variable (from imputed hgt and wgt) is used as predictor for other variables. Option 2 might be preferred when bmi is used in the substantial analyses.

The iteration plots show that the convergence for hgt, wgt and bmi is good (better than in the first model).

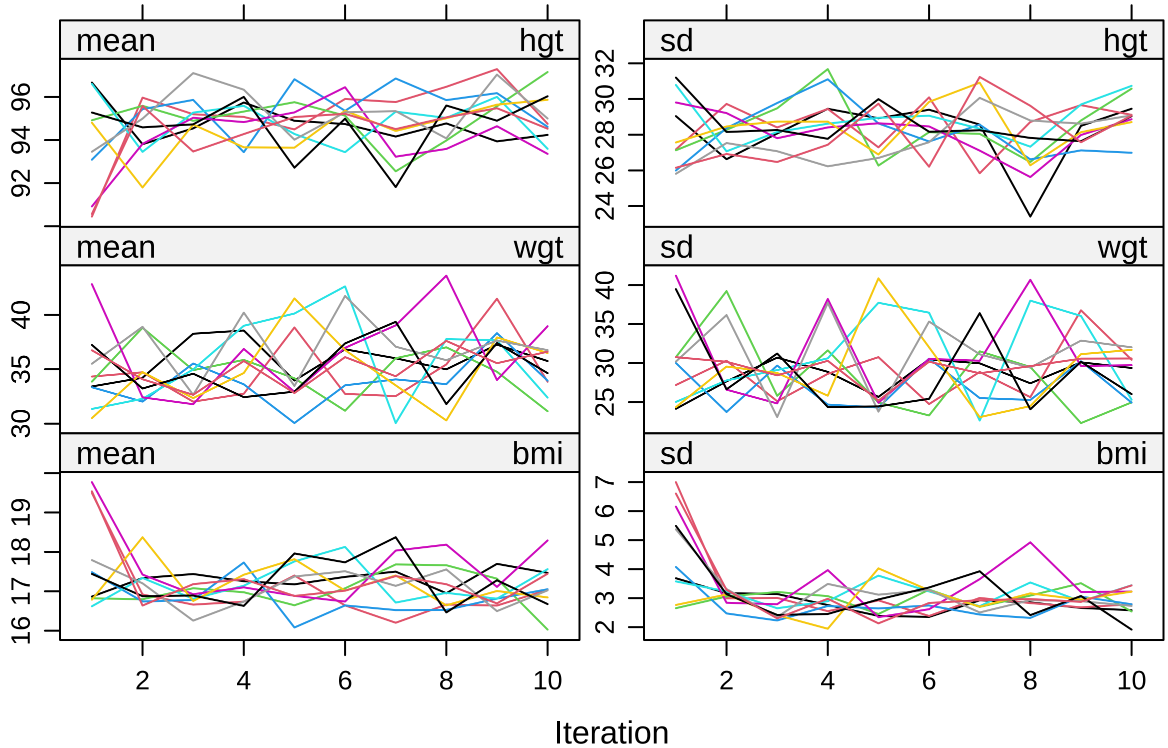
```
method <- imp$method
method["bmi"] <- "~I(wgt / (hgt/100)^2)"

imp <- mice(boys, m = 10, maxit = 10, pred = pred2, method = method)
```

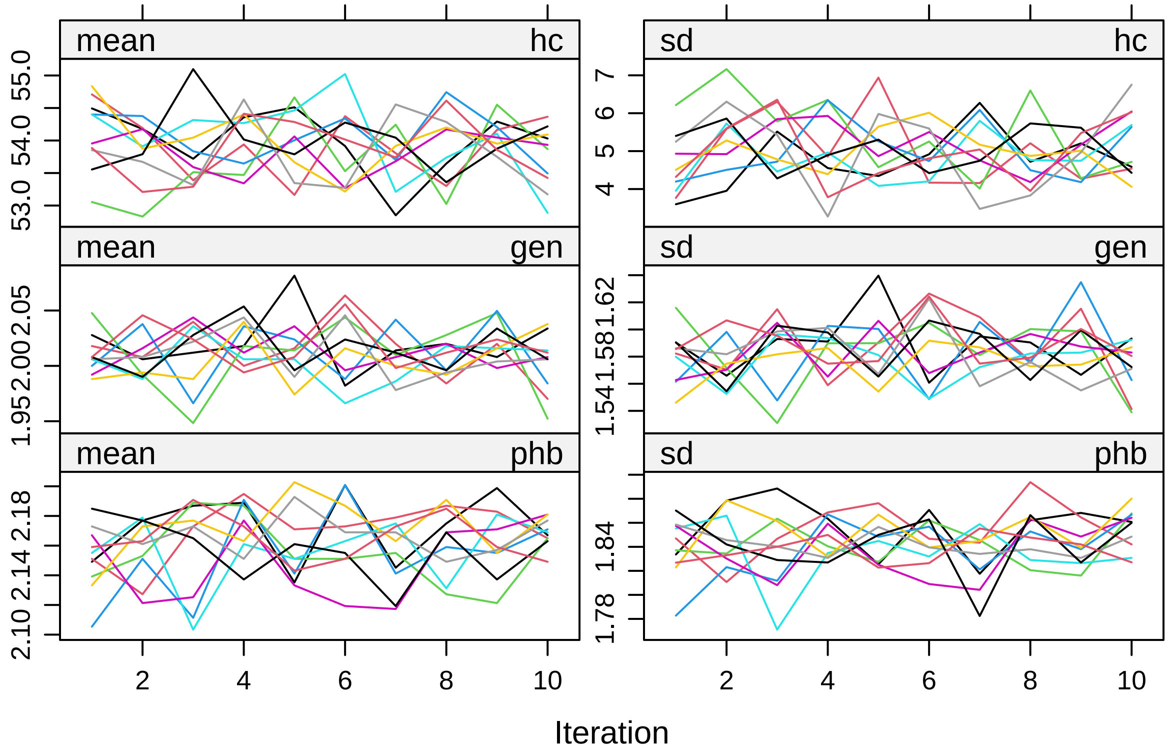
```
>
>  iter imp variable
>   1   1  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   1   2  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   1   3  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   1   4  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   1   5  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   1   6  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   1   7  hgt  wgt  bmi  hc  gen  phb  tv  reg
```

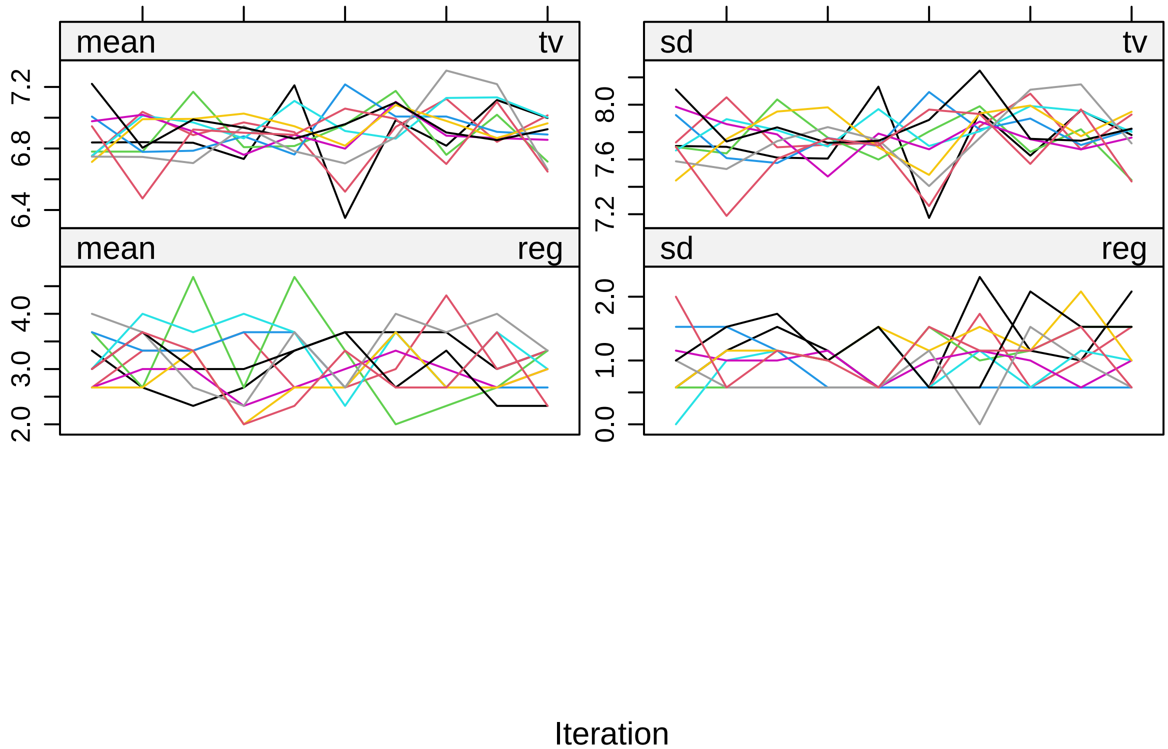
```
> 1 8 hgt wgt bmi hc gen phb tv reg
> 1 9 hgt wgt bmi hc gen phb tv reg
> 1 10 hgt wgt bmi hc gen phb tv reg
> 2 1 hgt wgt bmi hc gen phb tv reg
> 2 2 hgt wgt bmi hc gen phb tv reg
> 2 3 hgt wgt bmi hc gen phb tv reg
> 2 4 hgt wgt bmi hc gen phb tv reg
> 2 5 hgt wgt bmi hc gen phb tv reg
> 2 6 hgt wgt bmi hc gen phb tv reg
> 2 7 hgt wgt bmi hc gen phb tv reg
> 2 8 hgt wgt bmi hc gen phb tv reg
> 2 9 hgt wgt bmi hc gen phb tv reg
> 2 10 hgt wgt bmi hc gen phb tv reg
> 3 1 hgt wgt bmi hc gen phb tv reg
> 3 2 hgt wgt bmi hc gen phb tv reg
> 3 3 hgt wgt bmi hc gen phb tv reg
> 3 4 hgt wgt bmi hc gen phb tv reg
> 3 5 hgt wgt bmi hc gen phb tv reg
> 3 6 hgt wgt bmi hc gen phb tv reg
> 3 7 hgt wgt bmi hc gen phb tv reg
> 3 8 hgt wgt bmi hc gen phb tv reg
> 3 9 hgt wgt bmi hc gen phb tv reg
> 3 10 hgt wgt bmi hc gen phb tv reg
> 4 1 hgt wgt bmi hc gen phb tv reg
> 4 2 hgt wgt bmi hc gen phb tv reg
> 4 3 hgt wgt bmi hc gen phb tv reg
> 4 4 hgt wgt bmi hc gen phb tv reg
> 4 5 hgt wgt bmi hc gen phb tv reg
> 4 6 hgt wgt bmi hc gen phb tv reg
> 4 7 hgt wgt bmi hc gen phb tv reg
> 4 8 hgt wgt bmi hc gen phb tv reg
> 4 9 hgt wgt bmi hc gen phb tv reg
> 4 10 hgt wgt bmi hc gen phb tv reg
> 5 1 hgt wgt bmi hc gen phb tv reg
> 5 2 hgt wgt bmi hc gen phb tv reg
> 5 3 hgt wgt bmi hc gen phb tv reg
> 5 4 hgt wgt bmi hc gen phb tv reg
> 5 5 hgt wgt bmi hc gen phb tv reg
> 5 6 hgt wgt bmi hc gen phb tv reg
> 5 7 hgt wgt bmi hc gen phb tv reg
> 5 8 hgt wgt bmi hc gen phb tv reg
> 5 9 hgt wgt bmi hc gen phb tv reg
> 5 10 hgt wgt bmi hc gen phb tv reg
> 6 1 hgt wgt bmi hc gen phb tv reg
> 6 2 hgt wgt bmi hc gen phb tv reg
> 6 3 hgt wgt bmi hc gen phb tv reg
> 6 4 hgt wgt bmi hc gen phb tv reg
> 6 5 hgt wgt bmi hc gen phb tv reg
> 6 6 hgt wgt bmi hc gen phb tv reg
> 6 7 hgt wgt bmi hc gen phb tv reg
> 6 8 hgt wgt bmi hc gen phb tv reg
> 6 9 hgt wgt bmi hc gen phb tv reg
> 6 10 hgt wgt bmi hc gen phb tv reg
> 7 1 hgt wgt bmi hc gen phb tv reg
```

```
>   7   2  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   7   3  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   7   4  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   7   5  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   7   6  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   7   7  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   7   8  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   7   9  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   7   10  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   8   1  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   8   2  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   8   3  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   8   4  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   8   5  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   8   6  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   8   7  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   8   8  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   8   9  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   8   10  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   9   1  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   9   2  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   9   3  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   9   4  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   9   5  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   9   6  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   9   7  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   9   8  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   9   9  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   9   10  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   10   1  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   10   2  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   10   3  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   10   4  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   10   5  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   10   6  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   10   7  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   10   8  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   10   9  hgt  wgt  bmi  hc  gen  phb  tv  reg
>   10   10  hgt  wgt  bmi  hc  gen  phb  tv  reg
```

```
plot(imp)
```


mean
hgt
sd
hgt
mean
wgt
sd
wgt
mean
bmi
sd
bmi
92 94 96
24 26 28 30 32
30 35 40
25 30 35 40
16 17 18 19
2 3 4 5 6 7
2 4 6 8 10
2 4 6 8 10
Iteration


mean
hc
sd
hc
mean
gen
sd
gen
mean
phb
sd
phb
Iteration



Solution 7: FIML

a. Repeat the example from the lecture on FIML and estimate the descriptive means for the `airquality` variables Ozone, Solar.R, Wind and Temp using FIML estimation.

```
library("lavaan")
```

First specify the variables in the `airquality` data with the intercept code. It is possible to copy the code used in the Lecture, both give the same result.

The means are as follows: Ozone = 42.129; Solar.R = 185.932; Wind = 9.958; Temp = 77.882

```
model <- '
  #variance
  Ozone ~1
  Solar.R ~1
  Wind ~1
  Temp ~1
  '
fit <- sem(model, data = airquality, missing = "fiml", meanstructure = TRUE)
summary(fit)
```

```
> lavaan 0.6-11 ended normally after 21 iterations
>
```

```
>   Estimator                                         ML
>   Optimization method                           NLMINB
>   Number of model parameters                         8
>
>   Number of observations                           153
>   Number of missing patterns                         4
>
> Model Test User Model:
>
>   Test statistic                               152.868
>   Degrees of freedom                                 6
>   P-value (Chi-square)                           0.000
>
> Parameter Estimates:
>
>   Standard errors                             Standard
>   Information                                 Observed
>   Observed information based on                Hessian
>
> Intercepts:
>                    Estimate  Std.Err  z-value  P(>|z|)
>     Ozone            42.129    3.050   13.815    0.000
>     Solar.R         185.932    7.428   25.032    0.000
>     Wind              9.958    0.284   35.076    0.000
>     Temp             77.882    0.763  102.112    0.000
>
> Variances:
>                    Estimate  Std.Err  z-value  P(>|z|)
>     Ozone          1078.817  141.655    7.616    0.000
>     Solar.R        8054.991  942.768    8.544    0.000
>     Wind             12.330    1.410    8.746    0.000
>     Temp             89.006   10.176    8.746    0.000
```

```
model <- '
  #variance
  Ozone ~~ Ozone
  Solar.R ~~ Solar.R
  Wind ~~ Wind
  Temp ~~ Temp
  '
fit <- sem(model, data = airquality, missing = "fiml", meanstructure = TRUE)
summary(fit)
```

```
> lavaan 0.6-11 ended normally after 21 iterations
>
>   Estimator                                         ML
>   Optimization method                           NLMINB
>   Number of model parameters                         8
>
>   Number of observations                           153
>   Number of missing patterns                         4
>
> Model Test User Model:
>
```

```
>   Test statistic                                 152.868
>   Degrees of freedom                                   6
>   P-value (Chi-square)                             0.000
>
> Parameter Estimates:
>
>   Standard errors                               Standard
>   Information                                   Observed
>   Observed information based on                  Hessian
>
> Intercepts:
>                    Estimate  Std.Err  z-value  P(>|z|)
>     Ozone            42.129    3.050   13.815    0.000
>     Solar.R         185.932    7.428   25.032    0.000
>     Wind              9.958    0.284   35.076    0.000
>     Temp             77.882    0.763  102.112    0.000
>
> Variances:
>                    Estimate  Std.Err  z-value  P(>|z|)
>     Ozone          1078.817  141.655    7.616    0.000
>     Solar.R        8054.991  942.768    8.544    0.000
>     Wind             12.330    1.410    8.746    0.000
>     Temp             89.006   10.176    8.746    0.000
```

b. Now estimate a regression between `Ozone` and `Temp` using FIML estimation (Ozone as dependent variable), what is the coefficient for Temp?

The coefficient for Temp = 2.429 (SE = 0.231).

```
model <- '
  Ozone ~ Temp
  '
fit_fiml <- sem(model, data = airquality, missing = "fiml")
summary(fit_fiml, header = FALSE, fmi = TRUE)
```

```
>
> Parameter Estimates:
>
>   Standard errors                               Standard
>   Information                                   Observed
>   Observed information based on                  Hessian
>
> Regressions:
>                    Estimate  Std.Err  z-value  P(>|z|)      FMI
>   Ozone ~
>     Temp              2.429    0.231   10.509    0.000    0.240
>
> Intercepts:
>                    Estimate  Std.Err  z-value  P(>|z|)      FMI
>    .Ozone          -146.995   18.129   -8.108    0.000    0.240
>
> Variances:
>                    Estimate  Std.Err  z-value  P(>|z|)      FMI
>    .Ozone           552.671   72.569    7.616    0.000    0.242
```

c. What is the Fraction of missing information in this regression analysis?

The fraction of missing information reported for this analysis is shown in the last column and is equal to 0.24.

d. Add the other variables in the data as auxiliary variables to estimate the regression between Ozone and Temp. Did the coefficient for Temp change?

The coefficient for Temp changed to 2.354 (SE = 0.225).

```
model_aux <- '
  Ozone ~ Temp
  Temp ~~ Wind + Solar.R + Month + Day
  Ozone ~~ Wind + Solar.R + Month + Day
  Wind ~~ Solar.R + Month + Day
  Solar.R ~~ Month + Day
  Month ~~ Day
  '
auxfit <- sem(model = model_aux, missing = "fiml", data = airquality)

summary(auxfit)
```

```
> lavaan 0.6-11 ended normally after 225 iterations
>
>   Estimator                                         ML
>   Optimization method                           NLMINB
>   Number of model parameters                        27
>
>   Number of observations                           153
>   Number of missing patterns                         4
>
> Model Test User Model:
>
>   Test statistic                                 0.000
>   Degrees of freedom                                 0
>
> Parameter Estimates:
>
>   Standard errors                             Standard
>   Information                                 Observed
>   Observed information based on                Hessian
>
> Regressions:
>                    Estimate  Std.Err  z-value  P(>|z|)
>   Ozone ~
>     Temp              2.354    0.225   10.470    0.000
>
> Covariances:
>                    Estimate  Std.Err  z-value  P(>|z|)
>   Temp ~~
>     Wind            -15.172    2.946   -5.151    0.000
>     Solar.R         232.954   73.899    3.152    0.002
>     Month             5.607    1.168    4.799    0.000
>     Day             -10.886    6.796   -1.602    0.109
>  .Ozone ~~
```

```
>     Wind             -29.527    7.140   -4.136    0.000
>     Solar.R          350.094  187.764    1.865    0.062
>     Month             -5.774    2.665   -2.166    0.030
>     Day               18.848   18.842    1.000    0.317
>   Wind ~~
>     Solar.R          -17.069   26.123   -0.653    0.514
>     Month             -0.884    0.407   -2.171    0.030
>     Day                0.843    2.509    0.336    0.737
>   Solar.R ~~
>     Month             -7.321   10.557   -0.693    0.488
>     Day             -119.301   66.532   -1.793    0.073
>   Month ~~
>     Day               -0.099    1.009   -0.098    0.922
>
> Intercepts:
>                    Estimate  Std.Err  z-value  P(>|z|)
>    .Ozone          -140.782   17.629   -7.986    0.000
>     Temp             77.882    0.763  102.112    0.000
>     Wind              9.958    0.284   35.076    0.000
>     Solar.R         185.534    7.410   25.038    0.000
>     Month             6.993    0.114   61.269    0.000
>     Day              15.804    0.714   22.125    0.000
>
> Variances:
>                    Estimate  Std.Err  z-value  P(>|z|)
>    .Ozone           550.649   71.707    7.679    0.000
>     Temp             89.006   10.176    8.746    0.000
>     Wind             12.330    1.410    8.746    0.000
>     Solar.R        8050.793  941.698    8.549    0.000
>     Month             1.993    0.228    8.746    0.000
>     Day              78.066    8.925    8.746    0.000
```

Solution 8: Longitudinal missing data

a. Which variables have missing data and how many?

First select the variables: “id”, “trt”, “age”, “sex”, “cbcl1”, “yc1”, “yc2” and “yc3” from the `fdd` data for this assignment.

```
fdd1 <- fdd[,c("id", "trt", "age", "sex", "cbcl1", "yc1", "yc2", "yc3")]
summary(fdd1)
```

```
>        id         trt         age         sex         cbcl1            yc1
>  Min.   : 1.00   E:26   Min.   : 4.0   M:29   Min.   : 1.00   Min.   : 6.00
>  1st Qu.:15.75   C:26   1st Qu.: 7.0   F:23   1st Qu.:36.00   1st Qu.:25.00
>  Median :31.50          Median : 9.0          Median :61.00   Median :30.00
>  Mean   :30.71          Mean   :10.1          Mean   :56.78   Mean   :31.11
>  3rd Qu.:46.25          3rd Qu.:14.0          3rd Qu.:78.00   3rd Qu.:39.00
>  Max.   :59.00          Max.   :18.0          Max.   :97.00   Max.   :64.00
>                                               NA's   :11      NA's   :16
>       yc2             yc3
>  Min.   : 1.0    Min.   : 2.00
>  1st Qu.: 8.5    1st Qu.: 6.75
```

```
>  Median :13.5   Median :12.50
>  Mean   :16.0   Mean   :15.46
>  3rd Qu.:22.5   3rd Qu.:26.00
>  Max.   :42.0   Max.   :35.00
>  NA's   :22     NA's   :24
```

The fdd1 data only contains the selected variables for this assignment. The summary for this dataset shows that cbcl1 has 11 missing values, yc1 16, yc2 22 and yc3 24.

b. Analyze the data with a longitudinal model (yc1, yc2 and yc3 are the dependent repeated measurements), using time, age, sex and cbcl1 as predictor. What can you conclude about the ptsd scores? How many study participants are used in the analysis?

To analyze the data with a longitudinal multilevel model (preferred, however repeated measurements model can be used too), we need to restructure the data into a long format. The repeated measurement indicator is the follow-up "time", and the outcome variable is a post-traumatic stress ("ptsd") score.

```
library(tidyr)

fdd1_long <- pivot_longer(fdd1, cols = c("yc1", "yc2", "yc3"), names_to = "time",
                          values_to = "ptsd", names_prefix = "yc")

#note by using names_prefix, the prefix "yc" is removed from the time variable,
#so that it can be easily transformed to numeric if needed.
```

Apply the longitudinal model: use a linear multilevel regression with a random intercept for "id".

```
library(lme4)
library(lmerTest)

model <- lmer(ptsd ~ time + age + sex + cbcl1+ (1|id), data = fdd1_long)
summary(model)
```

```
> Linear mixed model fit by REML. t-tests use Satterthwaite's method [
> lmerModLmerTest]
> Formula: ptsd ~ time + age + sex + cbcl1 + (1 | id)
>    Data: fdd1_long
>
> REML criterion at convergence: 594.5
>
> Scaled residuals:
>      Min       1Q   Median       3Q      Max
> -2.10950 -0.54501 -0.07779  0.51082  1.86986
>
> Random effects:
>  Groups   Name        Variance Std.Dev.
>  id       (Intercept) 59.22    7.696
>  Residual             55.73    7.465
> Number of obs: 83, groups:  id, 31
>
> Fixed effects:
>               Estimate Std. Error       df t value Pr(>|t|)
> (Intercept)  27.25975    9.18225  27.74893   2.969   0.0061 **
> time2       -13.81680    1.99108  52.89672  -6.939 5.70e-09 ***
```

```
> time3       -14.79972    2.07490  53.47572  -7.133 2.64e-09 ***
> age           0.08242    0.51695  27.25885   0.159   0.8745
> sexF         -0.89282    3.73239  27.23882  -0.239   0.8127
> cbcl1         0.02882    0.06531  26.36683   0.441   0.6626
> ---
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> Correlation of Fixed Effects:
>       (Intr) time2  time3  age    sexF
> time2 -0.095
> time3 -0.055  0.472
> age   -0.873 -0.011 -0.054
> sexF  -0.515 -0.004 -0.002  0.270
> cbcl1 -0.740  0.000 -0.020  0.413  0.448
```

In total 33 persons are included in the analyses (out of the in total 52 participants in the data). Over time, the ptsd scores decrease.

c. Perform a multilevel multiple imputation with only a random intercept. Use 20 imputations and 10 iterations for the imputations. Repeat the longitudinal analysis and compare the parameter estimates with the estimates from the previous answer. Describe your findings.

The multilevel imputation should be performed on the long data format, so we use the restructured dataset to make a predictor matrix. In the predictor matrix we change the 1's in the "id" column to -2 to indicate that this is the level variable.

```r
pred <- make.predictorMatrix(fdd1_long)

#id should be indicated as the level variable, and the column should have a -2.
pred[2:7, "id"] <- -2

pred
```

```
>       id trt age sex cbcl1 time ptsd
> id     0   1   1   1     1    1    1
> trt   -2   0   1   1     1    1    1
> age   -2   1   0   1     1    1    1
> sex   -2   1   1   0     1    1    1
> cbcl1 -2   1   1   1     0    1    1
> time  -2   1   1   1     1    0    1
> ptsd  -2   1   1   1     1    1    0
```

Initialize the imputation model to change the method to "2l.pmm" to use a multilevel imputation method for the variables with missing data (cbcl1 and ptsd).

```r
library(miceadds)

ini <- mice(fdd1_long, m = 1, maxit = 0, pred = pred)
```

```
> Warning: Number of logged events: 1
```

```r
long_meth <- ini$method
long_meth[c("cbcl1", "ptsd")] <- "2l.pmm"
long_meth
```

```
>       id      trt      age      sex    cbcl1     time     ptsd
>       ""       ""       ""       "" "2l.pmm"       "" "2l.pmm"
```

Run the imputation, with 20 imputations and 10 iterations using the predictor matrix with -2 for "id", and the "2l.pmm" methods for the variables that have missing values.

```r
imp_long <- mice(fddl_long, pred = pred, meth = long_meth, m = 20, maxit = 10, seed = 1234)
```

```
>
>  iter imp variable
>   1   1  cbcl1  ptsd
>   1   2  cbcl1  ptsd
>   1   3  cbcl1  ptsd
>   1   4  cbcl1  ptsd
>   1   5  cbcl1  ptsd
>   1   6  cbcl1  ptsd
>   1   7  cbcl1  ptsd
>   1   8  cbcl1  ptsd
>   1   9  cbcl1  ptsd
>   1   10  cbcl1  ptsd
>   1   11  cbcl1  ptsd
>   1   12  cbcl1  ptsd
>   1   13  cbcl1  ptsd
>   1   14  cbcl1  ptsd
>   1   15  cbcl1  ptsd
>   1   16  cbcl1  ptsd
>   1   17  cbcl1  ptsd
>   1   18  cbcl1  ptsd
>   1   19  cbcl1  ptsd
>   1   20  cbcl1  ptsd
>   2   1  cbcl1  ptsd
>   2   2  cbcl1  ptsd
>   2   3  cbcl1  ptsd
>   2   4  cbcl1  ptsd
>   2   5  cbcl1  ptsd
>   2   6  cbcl1  ptsd
>   2   7  cbcl1  ptsd
>   2   8  cbcl1  ptsd
>   2   9  cbcl1  ptsd
>   2   10  cbcl1  ptsd
>   2   11  cbcl1  ptsd
>   2   12  cbcl1  ptsd
>   2   13  cbcl1  ptsd
>   2   14  cbcl1  ptsd
>   2   15  cbcl1  ptsd
>   2   16  cbcl1  ptsd
>   2   17  cbcl1  ptsd
>   2   18  cbcl1  ptsd
```

```
>   2   19  cbcl1  ptsd
>   2   20  cbcl1  ptsd
>   3   1  cbcl1  ptsd
>   3   2  cbcl1  ptsd
>   3   3  cbcl1  ptsd
>   3   4  cbcl1  ptsd
>   3   5  cbcl1  ptsd
>   3   6  cbcl1  ptsd
>   3   7  cbcl1  ptsd
>   3   8  cbcl1  ptsd
>   3   9  cbcl1  ptsd
>   3   10  cbcl1  ptsd
>   3   11  cbcl1  ptsd
>   3   12  cbcl1  ptsd
>   3   13  cbcl1  ptsd
>   3   14  cbcl1  ptsd
>   3   15  cbcl1  ptsd
>   3   16  cbcl1  ptsd
>   3   17  cbcl1  ptsd
>   3   18  cbcl1  ptsd
>   3   19  cbcl1  ptsd
>   3   20  cbcl1  ptsd
>   4   1  cbcl1  ptsd
>   4   2  cbcl1  ptsd
>   4   3  cbcl1  ptsd
>   4   4  cbcl1  ptsd
>   4   5  cbcl1  ptsd
>   4   6  cbcl1  ptsd
>   4   7  cbcl1  ptsd
>   4   8  cbcl1  ptsd
>   4   9  cbcl1  ptsd
>   4   10  cbcl1  ptsd
>   4   11  cbcl1  ptsd
>   4   12  cbcl1  ptsd
>   4   13  cbcl1  ptsd
>   4   14  cbcl1  ptsd
>   4   15  cbcl1  ptsd
>   4   16  cbcl1  ptsd
>   4   17  cbcl1  ptsd
>   4   18  cbcl1  ptsd
>   4   19  cbcl1  ptsd
>   4   20  cbcl1  ptsd
>   5   1  cbcl1  ptsd
>   5   2  cbcl1  ptsd
>   5   3  cbcl1  ptsd
>   5   4  cbcl1  ptsd
>   5   5  cbcl1  ptsd
>   5   6  cbcl1  ptsd
>   5   7  cbcl1  ptsd
>   5   8  cbcl1  ptsd
>   5   9  cbcl1  ptsd
>   5   10  cbcl1  ptsd
>   5   11  cbcl1  ptsd
>   5   12  cbcl1  ptsd
```

```
>   5   13  cbcl1  ptsd
>   5   14  cbcl1  ptsd
>   5   15  cbcl1  ptsd
>   5   16  cbcl1  ptsd
>   5   17  cbcl1  ptsd
>   5   18  cbcl1  ptsd
>   5   19  cbcl1  ptsd
>   5   20  cbcl1  ptsd
>   6   1  cbcl1  ptsd
>   6   2  cbcl1  ptsd
>   6   3  cbcl1  ptsd
>   6   4  cbcl1  ptsd
>   6   5  cbcl1  ptsd
>   6   6  cbcl1  ptsd
>   6   7  cbcl1  ptsd
>   6   8  cbcl1  ptsd
>   6   9  cbcl1  ptsd
>   6   10  cbcl1  ptsd
>   6   11  cbcl1  ptsd
>   6   12  cbcl1  ptsd
>   6   13  cbcl1  ptsd
>   6   14  cbcl1  ptsd
>   6   15  cbcl1  ptsd
>   6   16  cbcl1  ptsd
>   6   17  cbcl1  ptsd
>   6   18  cbcl1  ptsd
>   6   19  cbcl1  ptsd
>   6   20  cbcl1  ptsd
>   7   1  cbcl1  ptsd
>   7   2  cbcl1  ptsd
>   7   3  cbcl1  ptsd
>   7   4  cbcl1  ptsd
>   7   5  cbcl1  ptsd
>   7   6  cbcl1  ptsd
>   7   7  cbcl1  ptsd
>   7   8  cbcl1  ptsd
>   7   9  cbcl1  ptsd
>   7   10  cbcl1  ptsd
>   7   11  cbcl1  ptsd
>   7   12  cbcl1  ptsd
>   7   13  cbcl1  ptsd
>   7   14  cbcl1  ptsd
>   7   15  cbcl1  ptsd
>   7   16  cbcl1  ptsd
>   7   17  cbcl1  ptsd
>   7   18  cbcl1  ptsd
>   7   19  cbcl1  ptsd
>   7   20  cbcl1  ptsd
>   8   1  cbcl1  ptsd
>   8   2  cbcl1  ptsd
>   8   3  cbcl1  ptsd
>   8   4  cbcl1  ptsd
>   8   5  cbcl1  ptsd
>   8   6  cbcl1  ptsd
```

```
>   8   7  cbcl1  ptsd
>   8   8  cbcl1  ptsd
>   8   9  cbcl1  ptsd
>   8   10  cbcl1  ptsd
>   8   11  cbcl1  ptsd
>   8   12  cbcl1  ptsd
>   8   13  cbcl1  ptsd
>   8   14  cbcl1  ptsd
>   8   15  cbcl1  ptsd
>   8   16  cbcl1  ptsd
>   8   17  cbcl1  ptsd
>   8   18  cbcl1  ptsd
>   8   19  cbcl1  ptsd
>   8   20  cbcl1  ptsd
>   9   1  cbcl1  ptsd
>   9   2  cbcl1  ptsd
>   9   3  cbcl1  ptsd
>   9   4  cbcl1  ptsd
>   9   5  cbcl1  ptsd
>   9   6  cbcl1  ptsd
>   9   7  cbcl1  ptsd
>   9   8  cbcl1  ptsd
>   9   9  cbcl1  ptsd
>   9   10  cbcl1  ptsd
>   9   11  cbcl1  ptsd
>   9   12  cbcl1  ptsd
>   9   13  cbcl1  ptsd
>   9   14  cbcl1  ptsd
>   9   15  cbcl1  ptsd
>   9   16  cbcl1  ptsd
>   9   17  cbcl1  ptsd
>   9   18  cbcl1  ptsd
>   9   19  cbcl1  ptsd
>   9   20  cbcl1  ptsd
>   10   1  cbcl1  ptsd
>   10   2  cbcl1  ptsd
>   10   3  cbcl1  ptsd
>   10   4  cbcl1  ptsd
>   10   5  cbcl1  ptsd
>   10   6  cbcl1  ptsd
>   10   7  cbcl1  ptsd
>   10   8  cbcl1  ptsd
>   10   9  cbcl1  ptsd
>   10   10  cbcl1  ptsd
>   10   11  cbcl1  ptsd
>   10   12  cbcl1  ptsd
>   10   13  cbcl1  ptsd
>   10   14  cbcl1  ptsd
>   10   15  cbcl1  ptsd
>   10   16  cbcl1  ptsd
>   10   17  cbcl1  ptsd
>   10   18  cbcl1  ptsd
>   10   19  cbcl1  ptsd
>   10   20  cbcl1  ptsd
```

Fit the model to each imputed dataset and pool the results.

```r
# load the broom.mixed library to pool results from a multilevel model.
library(broom.mixed)

fit <- with(imp_long, lmer(ptsd ~  time + age + sex + (1|id) ))

summary(pool(fit))
```

```
>          term   estimate std.error  statistic       df      p.value
> 1 (Intercept) 31.9896789 6.0316947  5.3035972 29.71136 1.016497e-05
> 2       time2 -9.3365385 2.4654132 -3.7870076 69.67944 3.200219e-04
> 3       time3 -9.4798077 2.3036814 -4.1150690 95.49597 8.211735e-05
> 4         age -0.2938997 0.4758073 -0.6176864 32.78874 5.410487e-01
> 5        sexF -0.7180854 3.0003062 -0.2393374 80.80889 8.114500e-01
```

The results show a similar conclusion as for the previous answer; the ptsd scores decrease over time for all participants. However the coefficients seem less strong after imputation. One possible reason might be that the intention to treat analysis (analysis with imputations) reveals that mostly participants with less progress (less decrease in ptsd score) drop-out. Further investigations into the missing data and other information in the data may give us more insight into this. Note that now all 52 participants are included in the analyses (by calling `fit`).