Solutions - Missing data workshop

dr. Iris Eekhout

2022-05-03

Missing value analyses

Solution 1: amount of missing values

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

```
>
> Attaching package: 'mice'

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

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

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)
```

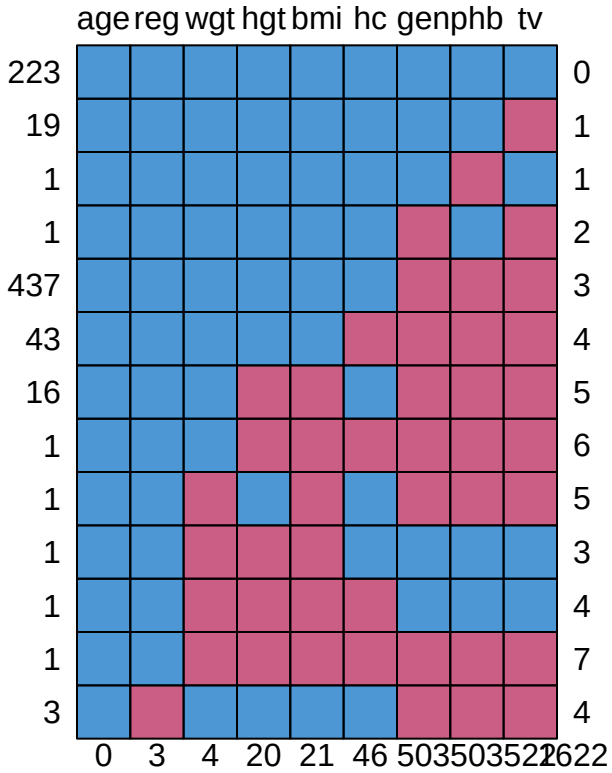
```
> [1] 0.2409388
```

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))
```

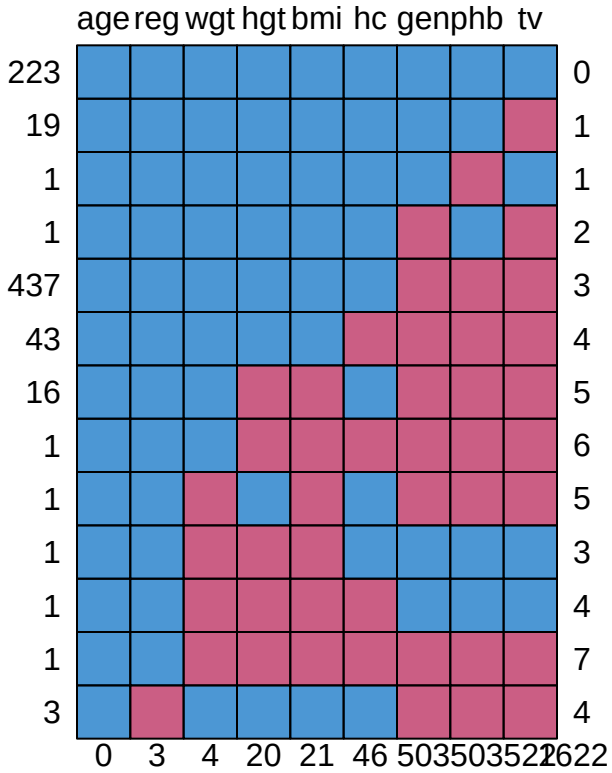



```
> [1] 14
```

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

The pattern with "gen", "phb" and "tv" missing, occurs 437 times.

```
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).

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
>
> $rm
>     age hgt wgt bmi hc gen phb  tv reg
> age   0  20   4  21 46 503 503 522   3
> hgt   0   0   1   1 43 485 485 504   3
> wgt   0  17   0  17 44 501 501 520   3
> bmi   0   0   0   0 43 484 484 503   3
> hc    0  17   2  18  0 458 458 477   3
> gen   0   2   2   2  1   0   1  19   0
> phb   0   2   2   2  1   1   0  20   0
> tv    0   2   2   2  1   0   1   0   0
> reg   0  20   4  21 46 500 500 519   0
>
> $mr
>     age hgt wgt bmi  hc gen phb tv reg
> age   0   0   0   0   0   0   0  0   0
> hgt  20   0  17   0  17   2   2  2  20
> wgt   4   1   0   0   2   2   2  2   4
> bmi  21   1  17   0  18   2   2  2  21
> hc   46  43  44  43   0   1   1  1  46
> gen 503 485 501 484 458   0   1  0 500
> phb 503 485 501 484 458   1   0  1 500
> tv  522 504 520 503 477  19  20  0 519
> reg   3   3   3   3   3   0   0  0   0
>
> $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
```

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 the standard error has increased slightly (was 0.08 for 0% missing).

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.

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

In the MNAR situation, more missings are present in the higher values of knee pain.

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

- MCAR: coefficient= 0.46; se=0.10
- MAR: coefficient=0.37; se=0.11
- MNAR: coefficient=0.27; se=0.11
- 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= 13.34; sd=2.69
- 0%missing: mean=14.77; sd=3.11

Both mean and standard deviation decrease when changing from MCAR to MAR to MNAR.

Solution 4: evaluating the missing data mechanism

a. Evaluate the missing data mechanism for the boys 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)
```

```
>
> Attaching package: 'dplyr'

> The following objects are masked from 'package:stats':
>
>     filter, lag

> The following objects are masked from 'package:base':
>
>     intersect, setdiff, setequal, union
```

```
boysm <- boys %>%
  #create missing data indicators
  mutate(Rhgt = is.na(hgt),
         Rwgt = is.na(wgt),
         Rbmi = is.na(bmi),
         Rhc = is.na(hc),
         Rgen = is.na(gen),
         Rphb = is.na(phb),
         Rtv = is.na(tv),
         Rreg = is.na(reg))
```

Do a T-test for each missing data indicator with the continuous variables and a chi square test for the categorical variables.

- Missing data in hgt seem to be related to age, wgt and hc; the group with missing heights are younger, lower weights and smaller hc.
- Nothing found for missing data in wgt
- Missing data in bmi seem to be related to age and hc; the group with missing bmi are younger and have smaller head circumference.
- Missing data in hc seem to be related to age, hgt, wgt, bmi and reg; the group with missing hc are older, have higher heights, weights and bmi.
- Missing data in gen are related to age, hgt, wgt, bmi, hc and reg; the group with missing gen are younger, and have smaller heights, weights, bmi and hc.
- Missing data in phb are related to age, hgt, wgt, bmi, hc and reg; the group with missing gen are younger, and have smaller heights, weights, bmi and hc.
- Missing data in tv are related to age, hgt, wgt, bmi and hc; the group with missing gen are younger, and have smaller heights, weights, bmi and hc.
- Missing data in reg are related to age, wgt and bmi; the group with missing gen are younger, and have smaller weights and bmi.

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.

Notes: we are performing a lot of tests here, so we may want to adjust for multiple testing; Some of the chi-square tests do not have sufficient observations in the cells so be careful about conclusions for these tests; some tests could not be performed because variables are always missing at the same time.

```
# Univariate tests for hgt
t.test(age ~ Rhgt, data = boysm)
```

```
>
>   Welch Two Sample t-test
>
> data:  age by Rhgt
> t = 5.7892, df = 21.5, p-value = 8.723e-06
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>  3.865264 8.189489
> sample estimates:
> mean in group FALSE  mean in group TRUE
>            9.320026            3.292650
```

```
t.test(wgt ~ Rhgt, data = boysm)
```

```
>
>   Welch Two Sample t-test
>
> data:  wgt by Rhgt
> t = 22.827, df = 205.76, p-value < 2.2e-16
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>  23.35971 27.77627
> sample estimates:
> mean in group FALSE  mean in group TRUE
>            37.73740            12.16941
```

```
#t.test(bmi ~ Rhgt, data = boysm)  ~ cannot be performed, because when hgt is missing, bmi is also miss
t.test(hc ~ Rhgt, data = boysm)
```

```
>
>   Welch Two Sample t-test
>
> data:  hc by Rhgt
> t = 4.1497, df = 19.386, p-value = 0.0005246
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>  1.546859 4.686649
> sample estimates:
> mean in group FALSE  mean in group TRUE
>            51.58146            48.46471
```

```
t.test(tv ~ Rhgt, data = boysm)
```

```
>
>   Welch Two Sample t-test
>
> data:  tv by Rhgt
> t = 0.046652, df = 1.0079, p-value = 0.9703
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>  -105.8340  106.6287
> sample estimates:
> mean in group FALSE  mean in group TRUE
>            11.89732            11.50000
```

```
chisq.test(boysm$Rhgt, boysm$gen)
```

```
> Warning in chisq.test(boysm$Rhgt, boysm$gen): Chi-squared approximation may be
> incorrect
```

```
>
>   Pearson's Chi-squared test
>
> data:  boysm$Rhgt and boysm$gen
> X-squared = 1.8358, df = 4, p-value = 0.7659
```

```
chisq.test(boysm$Rhgt, boysm$phb)
```

```
> Warning in chisq.test(boysm$Rhgt, boysm$phb): Chi-squared approximation may be
> incorrect
```

```
>
>   Pearson's Chi-squared test
>
> data:  boysm$Rhgt and boysm$phb
> X-squared = 2.9564, df = 5, p-value = 0.7067
```

```
chisq.test(boysm$Rhgt, boysm$reg)
```

```
> Warning in chisq.test(boysm$Rhgt, boysm$reg): Chi-squared approximation may be
> incorrect
```

```
>
>   Pearson's Chi-squared test
>
> data:  boysm$Rhgt and boysm$reg
> X-squared = 2.3698, df = 4, p-value = 0.6681
```

```
# t-tests for wgt
t.test(age ~ Rwgt, data = boysm)
```

```
>
>   Welch Two Sample t-test
>
> data:  age by Rwgt
> t = -0.035363, df = 3.0224, p-value = 0.974
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>  -13.29059  12.99726
> sample estimates:
> mean in group FALSE  mean in group TRUE
>            9.158082            9.304750
```

```
#t.test(hgt ~ Rwgt, data = boysm) ~ cannot be performed not enough observed hgt when wgt is missing.
#t.test(bmi ~ Rwgt, data = boysm) ~ cannot be performed, because when wgt is missing, bmi is also missi
t.test(hc ~ Rwgt, data = boysm)
```

```
>
>   Welch Two Sample t-test
>
> data:  hc by Rwgt
> t = 0.61098, df = 1.0021, p-value = 0.6507
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>  -83.06856  91.50456
> sample estimates:
> mean in group FALSE  mean in group TRUE
>              51.518              47.300
```

```
t.test(tv ~ Rwgt, data = boysm)
```

```
>
>   Welch Two Sample t-test
>
> data:  tv by Rwgt
> t = 0.046652, df = 1.0079, p-value = 0.9703
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>  -105.8340  106.6287
> sample estimates:
> mean in group FALSE  mean in group TRUE
>            11.89732            11.50000
```

```
chisq.test(boysm$Rwgt, boysm$gen, simulate.p.value = TRUE)
```

```
>
>   Pearson's Chi-squared test with simulated p-value (based on 2000
>   replicates)
>
> data:  boysm$Rwgt and boysm$gen
> X-squared = 1.8358, df = NA, p-value = 1
```

```
chisq.test(boysm$Rwgt, boysm$phb, simulate.p.value = TRUE)
```

```
>
>   Pearson's Chi-squared test with simulated p-value (based on 2000
>   replicates)
>
> data:  boysm$Rwgt and boysm$phb
> X-squared = 2.9564, df = NA, p-value = 0.8831
```

```
chisq.test(boysm$Rwgt, boysm$reg, simulate.p.value = TRUE)
```

```
>
>   Pearson's Chi-squared test with simulated p-value (based on 2000
>   replicates)
>
> data:  boysm$Rwgt and boysm$reg
> X-squared = 8.5143, df = NA, p-value = 0.07546
```

```
# t-tests for bmi
t.test(age ~ Rbmi, data = boysm)
```

```
>
>   Welch Two Sample t-test
>
> data:  age by Rbmi
> t = 6.1608, df = 22.848, p-value = 2.845e-06
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>  4.109642 8.266991
> sample estimates:
> mean in group FALSE  mean in group TRUE
>            9.332602            3.144286
```

```
#t.test(hgt ~ Rbmi, data = boysm)
#t.test(wgt ~ Rbmi, data = boysm) ~ cannot be performed, because when hgt is missing, bmi is also missi
t.test(hc ~ Rbmi, data = boysm)
```

```
>
>   Welch Two Sample t-test
>
> data:  hc by Rbmi
> t = 4.2566, df = 19.774, p-value = 0.0003946
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>  1.824921 5.337360
> sample estimates:
> mean in group FALSE  mean in group TRUE
>            51.59781            48.01667
```

```
t.test(tv ~ Rbmi, data = boysm)
```

```
>
>   Welch Two Sample t-test
>
> data:  tv by Rbmi
> t = 0.046652, df = 1.0079, p-value = 0.9703
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>  -105.8340  106.6287
> sample estimates:
> mean in group FALSE  mean in group TRUE
>            11.89732            11.50000
```

```
chisq.test(boysm$Rbmi, boysm$gen)
```

```
> Warning in chisq.test(boysm$Rbmi, boysm$gen): Chi-squared approximation may be
> incorrect
```

```
>
>   Pearson's Chi-squared test
>
> data:  boysm$Rbmi and boysm$gen
> X-squared = 1.8358, df = 4, p-value = 0.7659
```

```
chisq.test(boysm$Rbmi, boysm$phb)
```

```
> Warning in chisq.test(boysm$Rbmi, boysm$phb): Chi-squared approximation may be
> incorrect
```

```
>
>   Pearson's Chi-squared test
>
> data:  boysm$Rbmi and boysm$phb
> X-squared = 2.9564, df = 5, p-value = 0.7067
```

```
chisq.test(boysm$Rbmi, boysm$reg)
```

```
> Warning in chisq.test(boysm$Rbmi, boysm$reg): Chi-squared approximation may be
> incorrect
```

```
>
>   Pearson's Chi-squared test
>
> data:  boysm$Rbmi and boysm$reg
> X-squared = 3.1532, df = 4, p-value = 0.5325
```

```
# t-tests for hc
t.test(age ~ Rhc, data = boysm)
```

```
>
>   Welch Two Sample t-test
>
> data:  age by Rhc
> t = -3.5923, df = 51.172, p-value = 0.0007352
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>  -5.791186 -1.639066
> sample estimates:
> mean in group FALSE  mean in group TRUE
>            8.930396           12.645522
```

```
t.test(hgt ~ Rhc, data = boysm)
```

```
>
>   Welch Two Sample t-test
>
> data:  hgt by Rhc
> t = -3.3584, df = 48.24, p-value = 0.001538
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>  -36.751657  -9.228048
> sample estimates:
> mean in group FALSE  mean in group TRUE
>            130.7939            153.7837
```

```
t.test(wgt ~ Rhc, data = boysm)
```

```
>
>   Welch Two Sample t-test
>
> data:  wgt by Rhc
> t = -3.5041, df = 48.815, p-value = 0.0009915
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>  -21.73409  -5.89033
> sample estimates:
> mean in group FALSE  mean in group TRUE
>            36.33634            50.14855
```

```
t.test(bmi ~ Rhc, data = boysm)#
```

```
>
>   Welch Two Sample t-test
>
> data:  bmi by Rhc
> t = -2.7846, df = 48.024, p-value = 0.007646
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>  -2.1934846 -0.3540361
> sample estimates:
> mean in group FALSE  mean in group TRUE
>            17.99322            19.26698
```

```
#t.test(tv ~ Rhc, data = boysm) ~ cannot be performed not enough observed tv when hc is missing.
chisq.test(boysm$Rhc, boysm$gen)
```

```
> Warning in chisq.test(boysm$Rhc, boysm$gen): Chi-squared approximation may be
> incorrect

>
>   Pearson's Chi-squared test
>
> data:  boysm$Rhc and boysm$gen
> X-squared = 2.276, df = 4, p-value = 0.6851
```

```
chisq.test(boysm$Rhc, boysm$phb)
```

```
> Warning in chisq.test(boysm$Rhc, boysm$phb): Chi-squared approximation may be
> incorrect

>
>   Pearson's Chi-squared test
>
> data:  boysm$Rhc and boysm$phb
> X-squared = 4.996, df = 5, p-value = 0.4164
```

```
chisq.test(boysm$Rhc, boysm$reg)#
```

```
> Warning in chisq.test(boysm$Rhc, boysm$reg): Chi-squared approximation may be
> incorrect

>
>   Pearson's Chi-squared test
>
> data:  boysm$Rhc and boysm$reg
> X-squared = 20.517, df = 4, p-value = 0.0003948
```

```
# t-tests for gen
t.test(age ~ Rgen, data = boysm)
```

```
>
>   Welch Two Sample t-test
>
> data:  age by Rgen
> t = 19.57, df = 742.01, p-value < 2.2e-16
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>  6.516194 7.969335
> sample estimates:
> mean in group FALSE  mean in group TRUE
>           14.029335            6.786571
```

```
t.test(hgt ~ Rgen, data = boysm)
```

```
>
>   Welch Two Sample t-test
>
> data:  hgt by Rgen
> t = 20.183, df = 681.73, p-value < 2.2e-16
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>  44.53797 54.13741
> sample estimates:
> mean in group FALSE  mean in group TRUE
>            165.0210            115.6833
```

```
t.test(wgt ~ Rgen, data = boysm)
```

```
>
>   Welch Two Sample t-test
>
> data:  wgt by Rgen
> t = 15.011, df = 632.89, p-value < 2.2e-16
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>  21.37916 27.81476
> sample estimates:
> mean in group FALSE  mean in group TRUE
>            53.71646            29.11950
```

```
t.test(bmi ~ Rgen, data = boysm)
```

```
>
>   Welch Two Sample t-test
>
> data:  bmi by Rgen
> t = 6.064, df = 408.37, p-value = 3.026e-09
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>  1.022829 2.004063
> sample estimates:
> mean in group FALSE  mean in group TRUE
>            19.07613            17.56269
```

```
t.test(hc ~ Rgen, data = boysm)
```

```
>
>   Welch Two Sample t-test
>
> data:  hc by Rgen
> t = 17.591, df = 658.59, p-value < 2.2e-16
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
```

```
>  5.165614 6.463697
> sample estimates:
> mean in group FALSE  mean in group TRUE
>            55.29959            49.48493
```

```
#t.test(tv ~ Rgen, data = boysm)
#chisq.test(boysm$Rgen, boysm$phb)
chisq.test(boysm$Rgen, boysm$reg)
```

```
>
>   Pearson's Chi-squared test
>
> data:  boysm$Rgen and boysm$reg
> X-squared = 10.669, df = 4, p-value = 0.03055
```

```
# t-tests for phb
t.test(age ~ Rphb, data = boysm)
```

```
>
>   Welch Two Sample t-test
>
> data:  age by Rphb
> t = 19.548, df = 742.09, p-value < 2.2e-16
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>  6.510761 7.964473
> sample estimates:
> mean in group FALSE  mean in group TRUE
>           14.025873            6.788256
```

```
t.test(hgt ~ Rphb, data = boysm)
```

```
>
>   Welch Two Sample t-test
>
> data:  hgt by Rphb
> t = 20.179, df = 681.77, p-value < 2.2e-16
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>  44.53146 54.13156
> sample estimates:
> mean in group FALSE  mean in group TRUE
>            165.0169            115.6854
```

```
t.test(wgt ~ Rphb, data = boysm)
```

```
>
>   Welch Two Sample t-test
>
> data:  wgt by Rphb
> t = 15.016, df = 632.98, p-value < 2.2e-16
```

```
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>  21.38501 27.81990
> sample estimates:
> mean in group FALSE  mean in group TRUE
>            53.72016            29.11771
```

```
t.test(bmi ~ Rphb, data = boysm)
```

```
>
>   Welch Two Sample t-test
>
> data:  bmi by Rphb
> t = 6.082, df = 408.66, p-value = 2.729e-09
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>  1.026874 2.007682
> sample estimates:
> mean in group FALSE  mean in group TRUE
>            19.07868            17.56140
```

```
t.test(hc ~ Rphb, data = boysm)
```

```
>
>   Welch Two Sample t-test
>
> data:  hc by Rphb
> t = 17.607, df = 658.52, p-value < 2.2e-16
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>  5.169534 6.467316
> sample estimates:
> mean in group FALSE  mean in group TRUE
>            55.30205            49.48362
```

```
#t.test(tv ~ Rphb, data = boysm)
chisq.test(boysm$Rphb, boysm$gen)
```

```
> Warning in chisq.test(boysm$Rphb, boysm$gen): Chi-squared approximation may be
> incorrect
```

```
>
>   Pearson's Chi-squared test
>
> data:  boysm$Rphb and boysm$gen
> X-squared = 3.3888, df = 4, p-value = 0.495
```

```
chisq.test(boysm$Rphb, boysm$reg)
```

```
>
>   Pearson's Chi-squared test
```

```
>
> data:  boysm$Rphb and boysm$reg
> X-squared = 10.156, df = 4, p-value = 0.03788
```

```
# t-tests for tv
t.test(age ~ Rtv, data = boysm)
```

```
>
>   Welch Two Sample t-test
>
> data:  age by Rtv
> t = 19.076, df = 745, p-value < 2.2e-16
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>  6.336659 7.790496
> sample estimates:
> mean in group FALSE  mean in group TRUE
>           14.088261            7.024684
```

```
t.test(hgt ~ Rtv, data = boysm)
```

```
>
>   Welch Two Sample t-test
>
> data:  hgt by Rtv
> t = 19.79, df = 708.07, p-value < 2.2e-16
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>  43.29211 52.82792
> sample estimates:
> mean in group FALSE  mean in group TRUE
>            165.4241            117.3641
```

```
t.test(wgt ~ Rtv, data = boysm)
```

```
>
>   Welch Two Sample t-test
>
> data:  wgt by Rtv
> t = 14.479, df = 571.26, p-value < 2.2e-16
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>  20.84685 27.39034
> sample estimates:
> mean in group FALSE  mean in group TRUE
>            54.01027            29.89167
```

```
t.test(bmi ~ Rtv, data = boysm)
```

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

```
>
> data:  bmi by Rtv
> t = 5.7028, df = 364.59, p-value = 2.438e-08
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>  0.9610743 1.9727289
> sample estimates:
> mean in group FALSE  mean in group TRUE
>            19.08348            17.61658
```

```
t.test(hc ~ Rtv, data = boysm)
```

```
>
>   Welch Two Sample t-test
>
> data:  hc by Rtv
> t = 16.884, df = 685.56, p-value < 2.2e-16
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>  4.913502 6.206699
> sample estimates:
> mean in group FALSE  mean in group TRUE
>             55.2840             49.7239
```

```
chisq.test(boysm$Rtv, boysm$gen)
```

```
> Warning in chisq.test(boysm$Rtv, boysm$gen): Chi-squared approximation may be
> incorrect

>
>   Pearson's Chi-squared test
>
> data:  boysm$Rtv and boysm$gen
> X-squared = 3.7221, df = 4, p-value = 0.4449
```

```
chisq.test(boysm$Rtv, boysm$phb)
```

```
> Warning in chisq.test(boysm$Rtv, boysm$phb): Chi-squared approximation may be
> incorrect

>
>   Pearson's Chi-squared test
>
> data:  boysm$Rtv and boysm$phb
> X-squared = 6.6654, df = 5, p-value = 0.2467
```

```
chisq.test(boysm$Rtv, boysm$reg)
```

```
>
>   Pearson's Chi-squared test
>
> data:  boysm$Rtv and boysm$reg
> X-squared = 6.1038, df = 4, p-value = 0.1915
```

```r
# t-tests for reg
t.test(age ~ Rreg, data = boysm)
```

```
>
>   Welch Two Sample t-test
>
> data:  age by Rreg
> t = 6.3667, df = 2.1784, p-value = 0.01917
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>   2.943289 12.773479
> sample estimates:
> mean in group FALSE  mean in group TRUE
>            9.190384            1.332000
```

```r
t.test(hgt ~ Rreg, data = boysm)
```

```
>
>   Welch Two Sample t-test
>
> data:  hgt by Rreg
> t = 3.498, df = 2.0418, p-value = 0.07076
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>  -12.25484 131.11489
> sample estimates:
> mean in group FALSE  mean in group TRUE
>           132.39669            72.96667
```

```r
t.test(wgt ~ Rreg, data = boysm)
```

```
>
>   Welch Two Sample t-test
>
> data:  wgt by Rreg
> t = 5.8668, df = 2.1706, p-value = 0.02294
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>   8.937872 47.067664
> sample estimates:
> mean in group FALSE  mean in group TRUE
>           37.266101            9.263333
```

```r
t.test(bmi ~ Rreg, data = boysm)
```

```
>
>   Welch Two Sample t-test
>
> data:  bmi by Rreg
> t = 4.3643, df = 2.1196, p-value = 0.04382
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
```

```
> 95 percent confidence interval:
>  0.1919178 5.6626863
> sample estimates:
> mean in group FALSE  mean in group TRUE
>            18.08064            15.15333
```

```
t.test(hc ~ Rreg, data = boysm)
```

```
>
>   Welch Two Sample t-test
>
> data:  hc by Rreg
> t = 2.3032, df = 2.0127, p-value = 0.147
> alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
> 95 percent confidence interval:
>  -7.807393 26.030569
> sample estimates:
> mean in group FALSE  mean in group TRUE
>            51.54492            42.43333
```

```
#t.test(tv ~ Rreg, data = boysm)
#chisq.test(boysm$Rreg, boysm$gen)
#chisq.test(boysm$Rreg, boysm$phb)
```

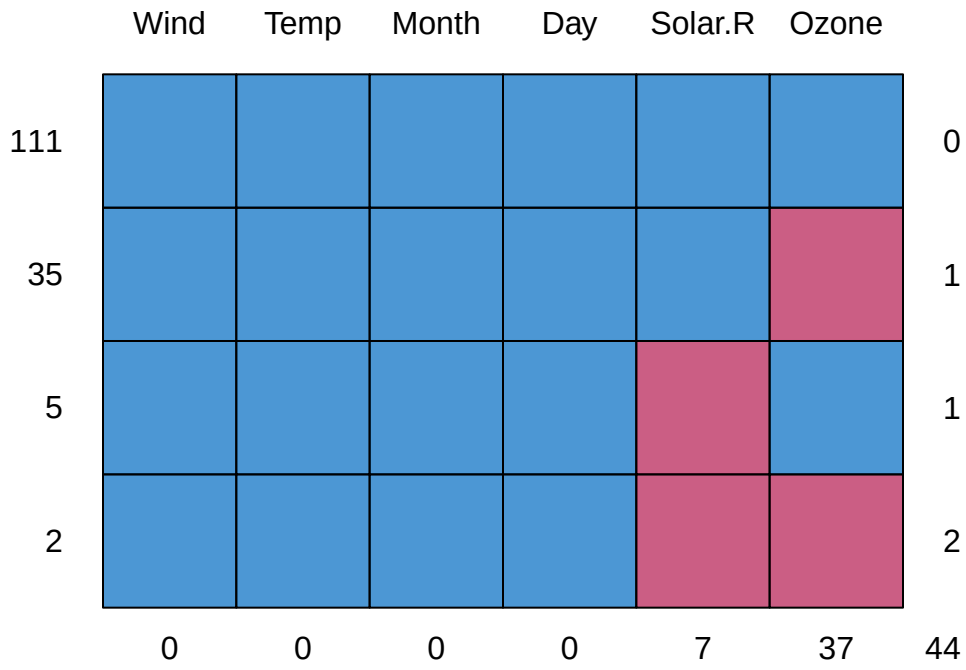
b. Evaluate the missing data mechanism for the nhanes data with a multivariate test. What are your conclusions?

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

```
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.

```
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
```

```
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`

Solution 6: multiple imputation predictor matrix

Solution 7: multiple imputation methods