Missing data

STAT5003

Justin Wishart

Semester 2, 2022

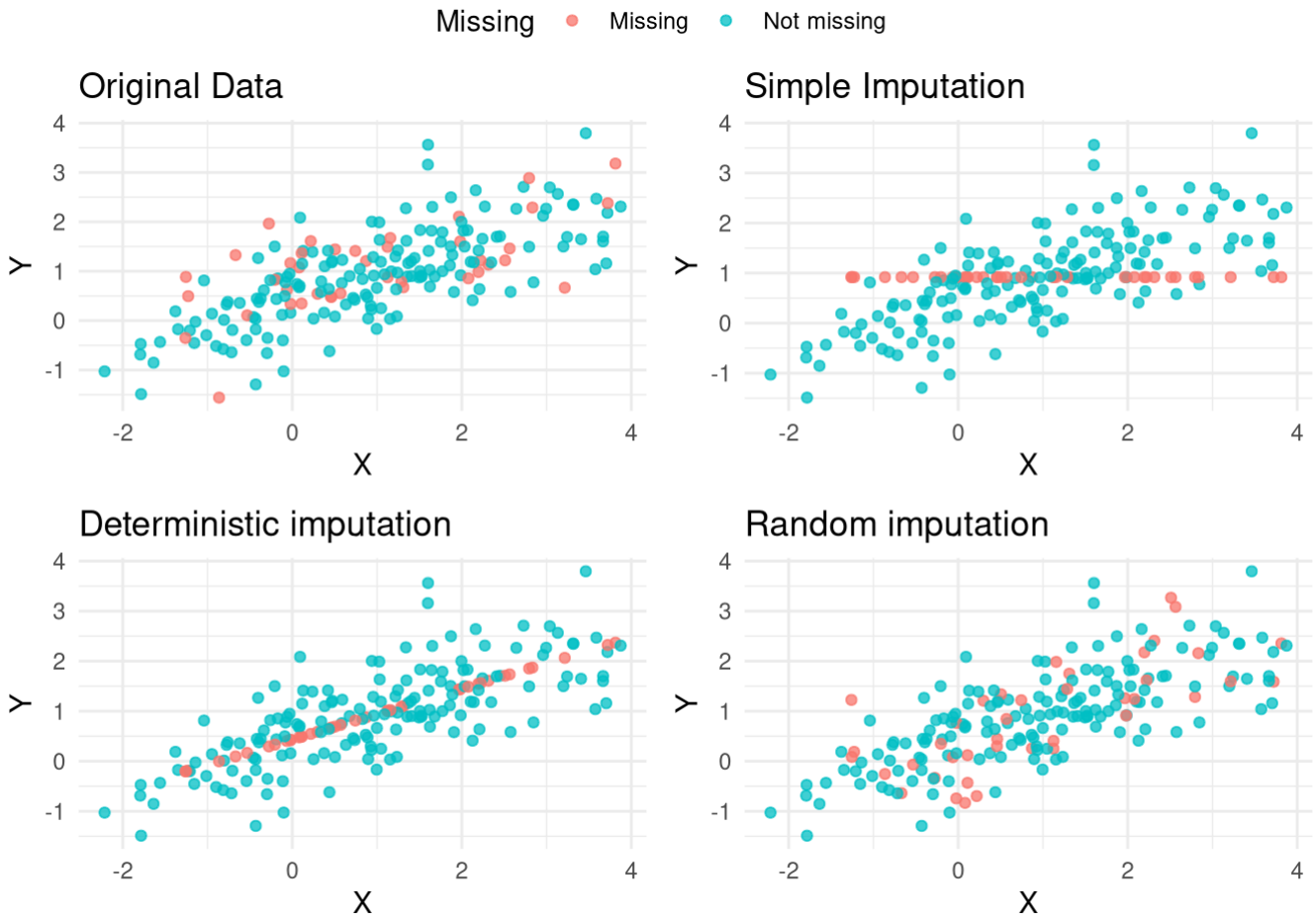
Libraries to load

```
library(mice)
library(dplyr)
library(VIM)
library(ggplot2)
```

Create simulation dataset

For the demonstration this week, we will first create a simulation dataset. Using `rnorm()`, we draw random features from a Gaussian distribution. We will simulate a dataset with two features - each feature will have a different mean but the same standard deviation for the two simulated samples.

```
set.seed(5003)
# Simulate some linearly related data
dat <- as.data.frame(MASS::mvrnorm(n = 200, mu = c(1, 1), Sigma = matrix(c(2, 1, 1, 1
), ncol = 2)))
names(dat) <- c("X", "Y")
dat.missing <- dat
attr(dat, "name") <- "Original Data"
# Replace 20% of the Y values as missing data
missing.inds <- sort(sample(seq(nrow(dat)), size = 0.2 * nrow(dat)))
dat.missing[["Y"]][missing.inds] <- NA
# Add a factor variable that specifies if there is missing data for that case (useful
for plotting)
dat[["Missing"]] <- dat.missing[["Missing"]] <- c("Not missing", "Missing")[as.numeri
c(is.na(dat.missing[["Y"]])) + 1]
mn <- mean(dat.missing[["Y"]], na.rm = TRUE)
# Do a simple single variable mean imputation
dat.simple <- dat.missing
dat.simple[["Y"]][is.na(dat.simple[["Y"]])] <- mn
dat.simple[["Missing"]] <- dat[["Missing"]]
attr(dat.simple, "name") <- "Simple Imputation"
# Do a simple linear regression imputation
simp.lm <- lm(Y ~ X, dat = dat.missing)
obs.X <- dat.missing %>% filter(is.na(Y)) %>% select(X)
simple.preds <- predict(simp.lm, newdata = obs.X)
dat.determ <- dat.missing
dat.determ[["Y"]][missing.inds] <- simple.preds %>% unname
attr(dat.determ, "name") <- "Deterministic imputation"
# Do a random imputation based on the estimated variability in the regression model
est.sigma <- sigma(simp.lm)
rand.preds <- simple.preds  + rnorm(length(simple.preds), sd = est.sigma)
dat.rand <- dat.missing
dat.rand[["Y"]][missing.inds] <- rand.preds %>% unname
attr(dat.rand, "name") <- "Random imputation"
# Create the plots
plots <- lapply(list(dat, dat.simple, dat.determ, dat.rand), function(dat) {
  ggplot(dat) +
      geom_point(aes(x = X, y = Y, colour = Missing), alpha = 0.75) +
      theme_minimal() +
      ggtitle(attr(dat, "name"))
})
library(ggpubr)
ggarrange(plotlist = plots, common.legend = TRUE)
```

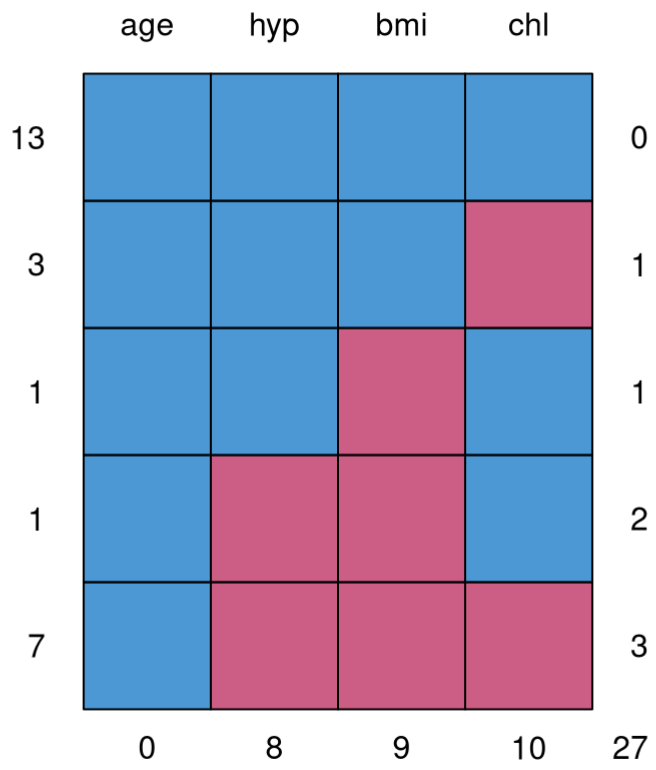


Visualize missing data

```
# Load the "caret" package. This package is used to partition data, training and test classification models etc.
data(nhanes, package = "mice")
```

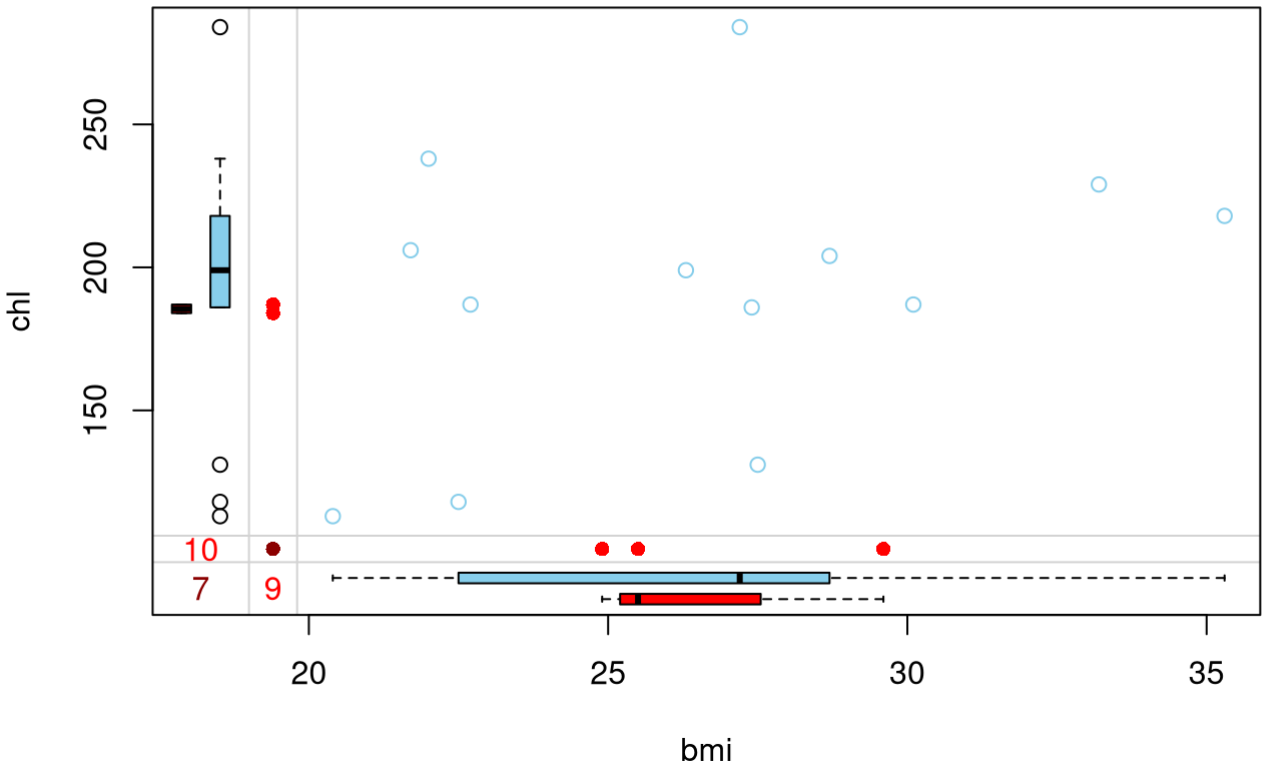
Check the missing data pattern in the nhanes data

```
md.pattern(nhanes)
```




```
##    age hyp bmi chl
## 13   1   1   1   1  0
## 3    1   1   1   0  1
## 1    1   1   0   1  1
## 1    1   0   0   1  2
## 7    1   0   0   0  3
##      0   8   9  10 27
```

```
marginplot(nhanes[, c("bmi", "chl")])
```


chl
250
200
150
10
7
9
20
25
30
35
bmi