Exercise3

Miriam Fischer

12 10 2020

Creating Dataset

We first create our random variables, and the data of Y1 and Y2:

```
set.seed(40)
Z1 <- rnorm(500, 0, 1)
Z2 <- rnorm(500, 0, 1)
Z3 <- rnorm(500, 0, 1)

Y1 <- 1 + Z1
Y2 <- 5 + 2 * Z1 + Z2
```

We then set the function when Y2 should be missing:

```
a <- 2
b <- 0
v <- a * (Y1 - 1) + b * (Y2 - 5) + Z3
```

3a: Impose missingness on Y2

We impose missingness on Y2 for Exercise 3a

```
ind_a <- which(v < 0)
Y2_MAR_obs <- Y2[-ind_a]
Y2_MAR_mis <- Y2[ind_a]
```

Note that this is MAR, missing at random. The reason is that with a=2 and b=0, the function which determines missingness on Y2 is

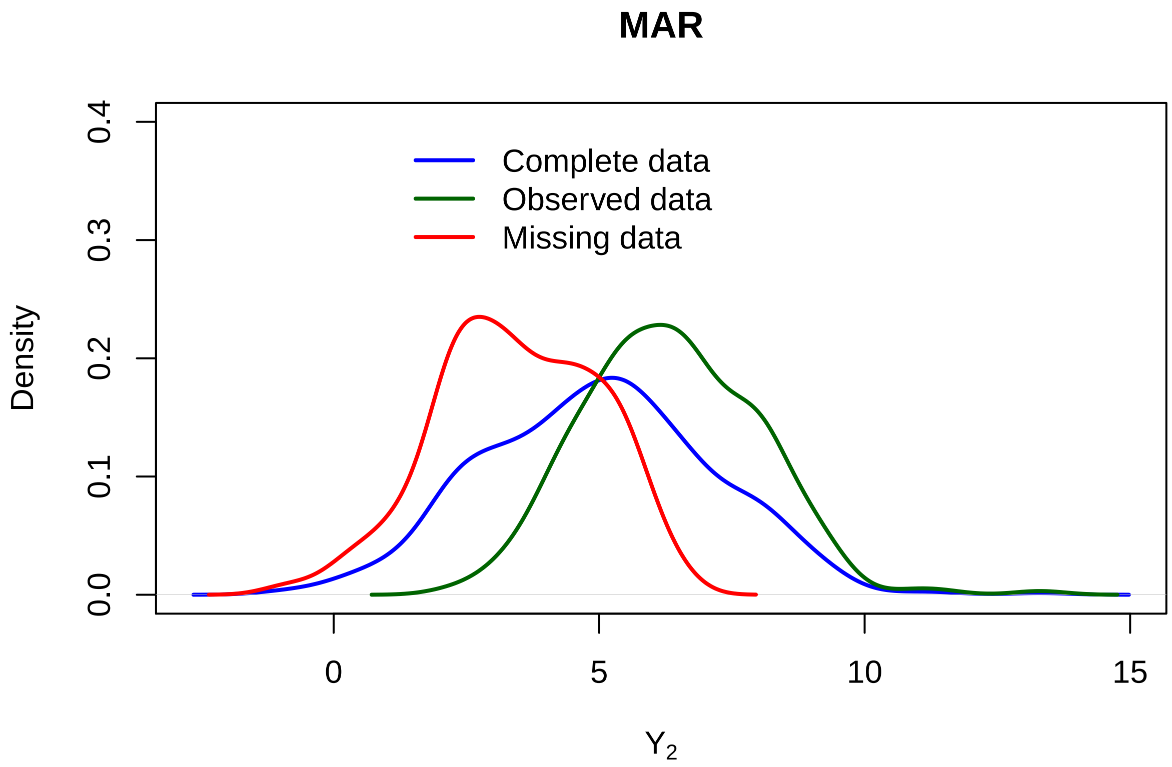
$$2(Y_1 - 1) + Z_3 < 0$$

This function clearly depends on Y1, but not on Y2. It also depends on a random variable Z3 (which is without connection to Y1 and Y2). Thus, missingness of Y2 depends on the value of Y1, making it MAR. However, it does not depend on Y2 itself. We further argue that the missingness depending on Z3 is not seen as missingness depending on another variable (of data which we do not have available), as this is solely random.

3a: Plots

We now plot the missing data for 3a.

```
plot(density(Y2), lwd = 2, col = "blue", xlab = expression(Y[2]), main = "MAR", ylim = c(0,
    0.4))
lines(density(Y2_MAR_obs), lwd = 2, col = "darkgreen")
lines(density(Y2_MAR_mis), lwd = 2, col = "red")
legend(1, 0.4, legend = c("Complete data", "Observed data", "Missing data"), col = c("blue",
    "darkgreen", "red"), lty = c(1, 1, 1), lwd = c(2, 2, 2), bty = "n")
```



3b: Regression imputation

```
Y2_MAR_na <- ifelse(v < 0, NA, Y2)
data_b <- data.frame(Y1_reg_b = Y1, Y2_reg_b = Y2_MAR_na)
reg_fit_b <- lm(Y2_reg_b ~ Y1_reg_b, data <- data_b)
```

The regression is:

```
reg_fit_b$coefficients
```

```
## (Intercept)    Y1_reg_b
##    2.989912    1.942356
```

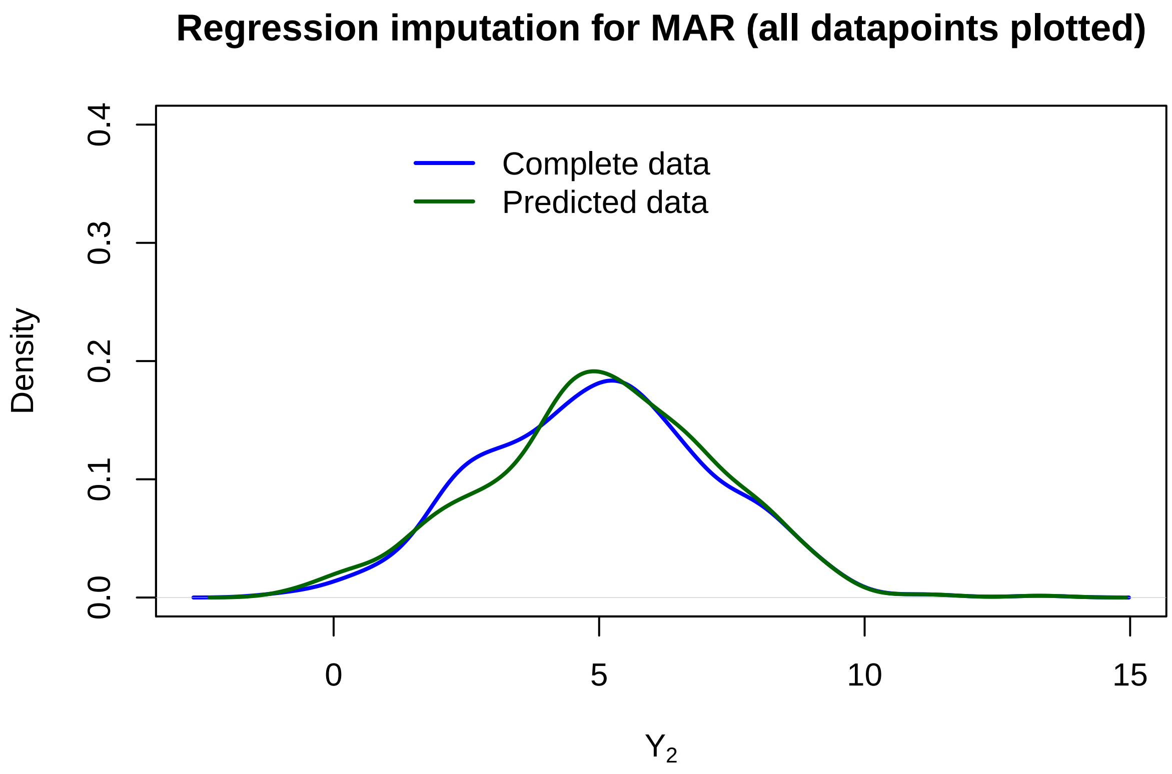
Our predicted dataset will have the known values of Y2 for the values where we did not impose missingness, and will use the values predicted with our regression for the values which are missing.

```
predicted_b <- predict(reg_fit_b, newdata = data_b) + rnorm(nrow(data_b), 0, sigma(reg_fit_b))
Y2_MAR_pre <- ifelse(is.na(data_b$Y2_reg_b), predicted_b, Y2)
```

Plots

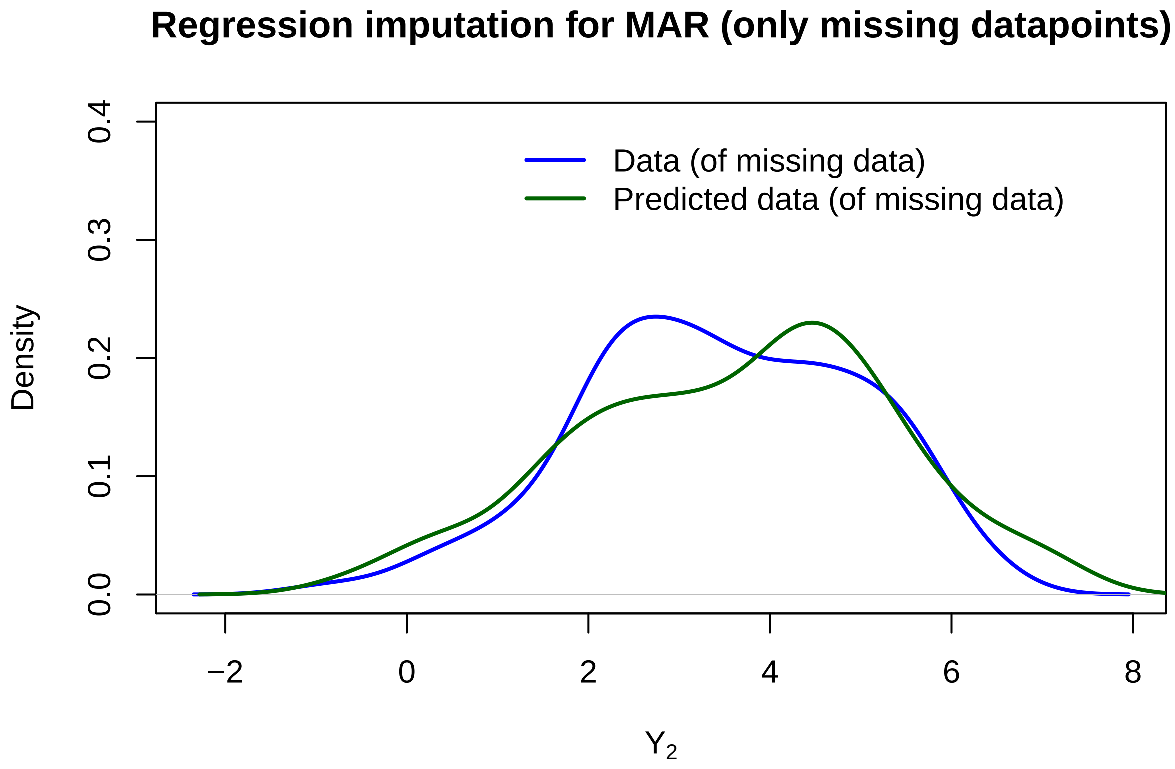
```
# plot
plot.new()
frame()
plot(density(Y2), lwd = 2, col = "blue", xlab = expression(Y[2]), main = "Regression imputation for MAR
    ylim = c(0, 0.4))
```

```
lines(density(Y2_MAR_pre), lwd = 2, col = "darkgreen")
legend(1, 0.4, legend = c("Complete data", "Predicted data"), col = c("blue", "darkgreen"),
    lty = c(1, 1), lwd = c(2, 2), bty = "n")
```



We note that for datapoints where Y2 is not missing, the "predicted" value and the correct value are equal (as, looking at our code, we our predicted data does not predict the values for given datapoints, but just takes their known values). We therefore use a second graph, which only has the data which we actually predict.

```
plot.new()
frame()
Y2_MAR_onlypre <- Y2_MAR_pre[ind_a]
plot(density(Y2_MAR_mis), lwd = 2, col = "blue", xlab = expression(Y[2]), main = "Regression imputation
    ylim = c(0, 0.4))
lines(density(Y2_MAR_onlypre), lwd = 2, col = "darkgreen")
legend(1, 0.4, legend = c("Data (of missing data)", "Predicted data (of missing data)"),
    col = c("blue", "darkgreen"), lty = c(1, 1), lwd = c(2, 2), bty = "n")
```



3c: Impose missingness

We change the parameters when Y2 should be missing:

```
a <- 0
b <- 2
vv <- a * (Y1 - 1) + b * (Y2 - 5) + Z3
```

We impose missingness on Y2 for Exercise 3c

```
ind_c <- which(vv < 0)
Y2_MNAR_obs <- Y2[-ind_c]
Y2_MNAR_mis <- Y2[ind_c]
```

Note that this is MNAR, missing not at random. The reason is that with a=0 and b=2, the function which determines missingness on Y2 is

$$2(Y_2 - 5) + Z_3 < 0$$

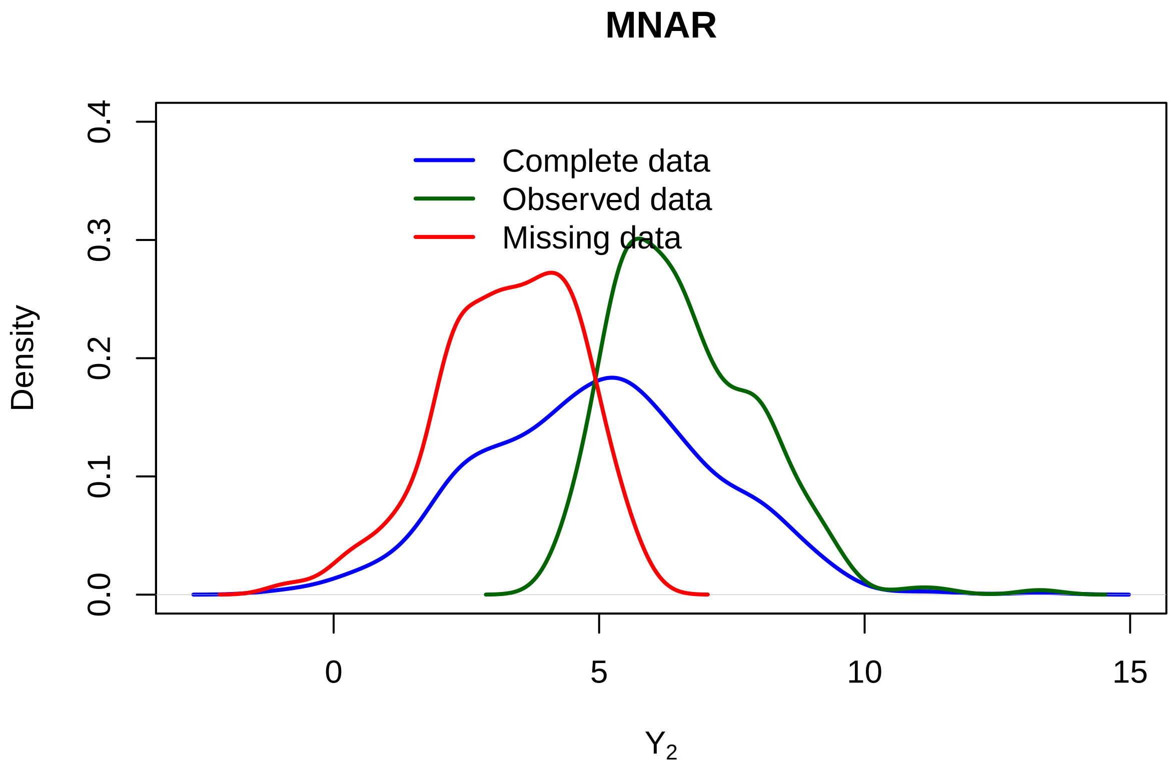
This function clearly depends on Y2, but not on Y1. Thus, must crucially, missingness of Y2 depends on the value of Y2 itself, making it MNAR.

3c: Plots

We now plot the missing data for 3c.

```
plot(density(Y2), lwd = 2, col = "blue", xlab = expression(Y[2]), main = "MNAR",
    ylim = c(0, 0.4))
lines(density(Y2_MNAR_obs), lwd = 2, col = "darkgreen")
lines(density(Y2_MNAR_mis), lwd = 2, col = "red")
```

```
legend(1, 0.4, legend = c("Complete data", "Observed data", "Missing data"), col = c("blue",
    "darkgreen", "red"), lty = c(1, 1, 1), lwd = c(2, 2, 2), bty = "n")
```



3d: Regression imputation

```
Y2_MNAR_na <- ifelse(vv < 0, NA, Y2)
data_d <- data.frame(Y1_reg_d = Y1, Y2_reg_d = Y2_MNAR_na)
reg_fit_d <- lm(Y2_reg_d ~ Y1_reg_d, data <- data_d)
```

The regression is:

```
reg_fit_d$coefficients
```

```
## (Intercept)    Y1_reg_d
##    4.018839    1.500059
```

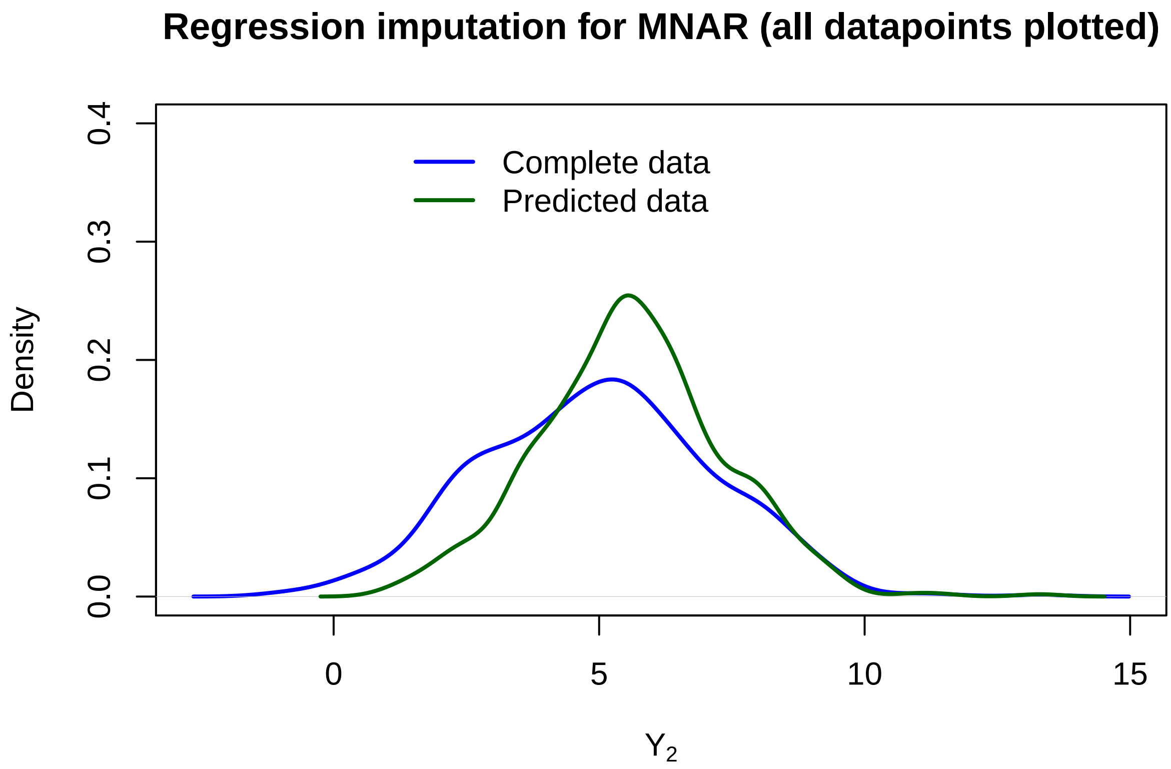
Our predicted dataset will have the known values of Y2 for the values where we did not impose missingness, and will use the values predicted with our regression for the values which are missing.

```
predicted_d <- predict(reg_fit_d, newdata = data_d) + rnorm(nrow(data_d), 0, sigma(reg_fit_d))
Y2_MNAR_pre <- ifelse(is.na(data_d$Y2_reg_d), predicted_d, Y2)
```

Plots

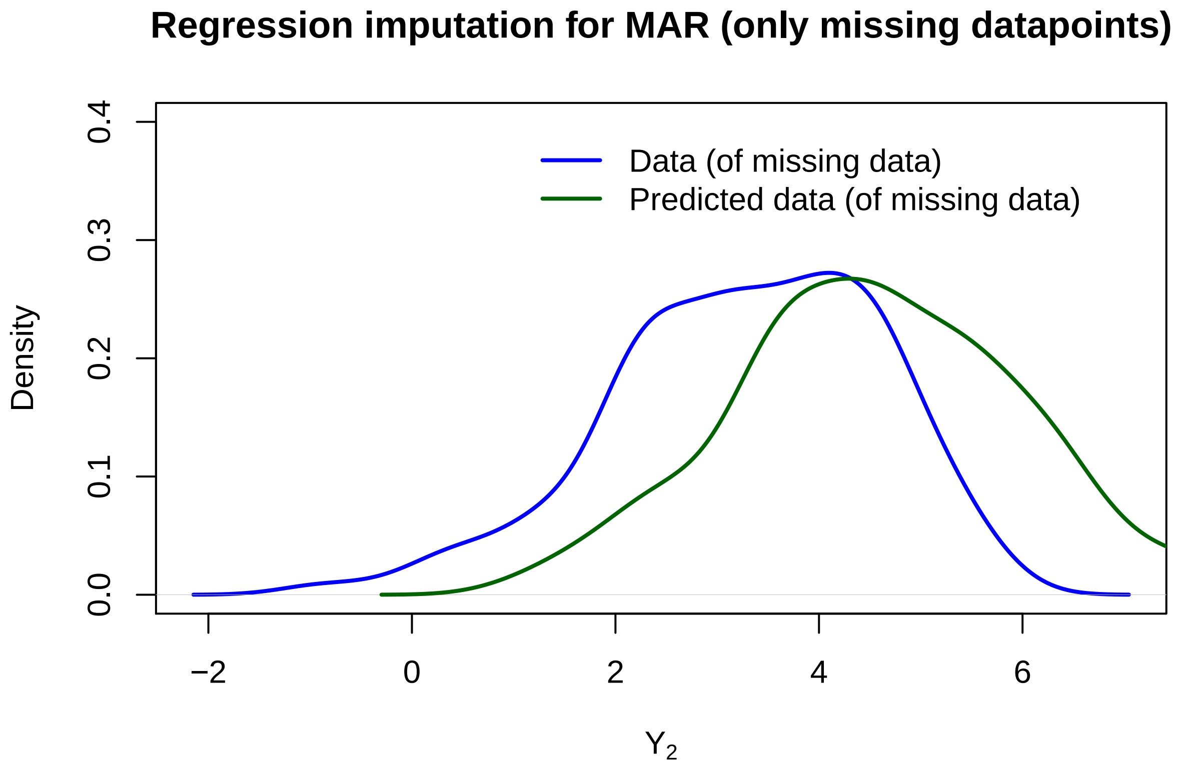
```
# plot
plot.new()
```

```
frame()
plot(density(Y2), lwd = 2, col = "blue", xlab = expression(Y[2]), main = "Regression imputation for MNA
    ylim = c(0, 0.4))
lines(density(Y2_MNAR_pre), lwd = 2, col = "darkgreen")
legend(1, 0.4, legend = c("Complete data", "Predicted data"), col = c("blue", "darkgreen"),
    lty = c(1, 1), lwd = c(2, 2), bty = "n")
```



We note that for datapoints where Y2 is not missing, the "predicted" value and the correct value are equal (as, looking at our code, we our predicted data does not predict the values for given datapoints, but just takes their known values). We therefore use a second graph, which only has the data which we actually predict.

```
plot.new()
frame()
Y2_MNAR_onlypre <- Y2_MNAR_pre[ind_c]
plot(density(Y2_MNAR_mis), lwd = 2, col = "blue", xlab = expression(Y[2]), main = "Regression imputatio
    ylim = c(0, 0.4))
lines(density(Y2_MNAR_onlypre), lwd = 2, col = "darkgreen")
legend(1, 0.4, legend = c("Data (of missing data)", "Predicted data (of missing data)"),
    col = c("blue", "darkgreen"), lty = c(1, 1), lwd = c(2, 2), bty = "n")
```



Particularly if we plot only the data of our missing data, we see that our prediction gives very poor results in predicting Y2. This is not a surprise: Our regression tries to predict Y2 based on Y1, but for our model, missingness on Y2 is not based on Y1, but on Y2! It is no surprise that the prediction is not good.