

Incomplete Data Analysis

Assignment – **sketch** of the solutions

- MNAR. Data on income are missing because people with higher incomes are less likely to reveal them, thus MNAR.
 - MCAR. Missing observations are due to contamination and not related to the missing data themselves or any other factor, and thus missing data are MCAR.
 - MAR. Employment status of some Hispanic individuals is missing. Because race is a fully recorded variable, the missing data are MAR.
- The largest possible subsample under a complete case analysis is 475, corresponding to the case where the missing values occur in all the 10 variables exactly for the same subjects. The smallest possible subsample under a complete case analysis is 0, corresponding to the extreme case where missing values occur for ‘non overlapping sets of subjects’ (e.g., for the first variable, individuals 1–25 have missing values, for the second variable, individuals 26–50 have missing values, ..., for the 20th variable, individuals 475–500 have missing values).
- First, note that although not required, the induced distribution on (Y_1, Y_2) is a bivariate normal distribution with means $(1, 5)$, variances $(1, 5)$, and correlation $2\sqrt{5} = 0.89$. The missingness mechanism is MAR. Missingness on Y_2 depends on Y_1 , which is fully observed, and on Z_3 which is not in our dataset but is uncorrelated to Y_2 (by construction). I have fixed the seed (to 1) and simulated the data as follows.

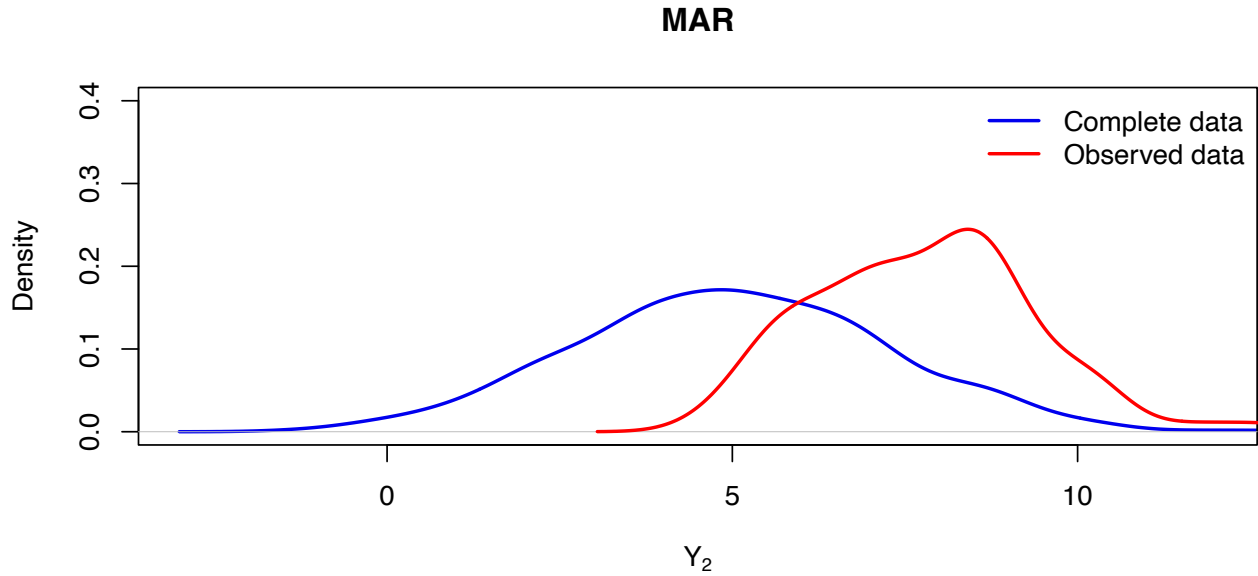
```
set.seed(1)
n <- 500
Z1 <- rnorm(n, 0, 1)
Z2 <- rnorm(n, 0, 1)
Z3 <- rnorm(n, 0, 1)

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

a <- 4; b <- 0
aux_MAR <- a*(Y1 - 2) + b*(Y2 - 6) + Z3

Y2_MAR <- ifelse(aux_MAR < 0, NA, Y2)
r_MAR <- which(is.na(Y2_MAR) == FALSE)
Y2_MAR_obs <- Y2[r_MAR]

plot(density(Y2), col = "blue2", xlim = c(-3, 12), ylim = c(0, 0.4),
     main = "MAR", xlab = expression(Y[2]), ylab = "Density", lwd = 2)
lines(density(Y2_MAR_obs), col = "red", lwd = 2)
legend("topright", legend = c("Complete data", "Observed data"),
     col = c("blue2", "red"), lty = c(1,1), lwd = c(2,2), bty = "n")
```



The observed and complete data distributions on Y_2 are distinct, which implies that if we were to use the observed data to make inferences that depend on Y_2 , those would be biased. In particular, we see that the mean of the observed data distribution is shifted upwards and the variability of the observed data is reduced compared to that of the complete data.

- (b) For stochastic regression imputation, the following model will be used to impute the missing values on Y_2 .

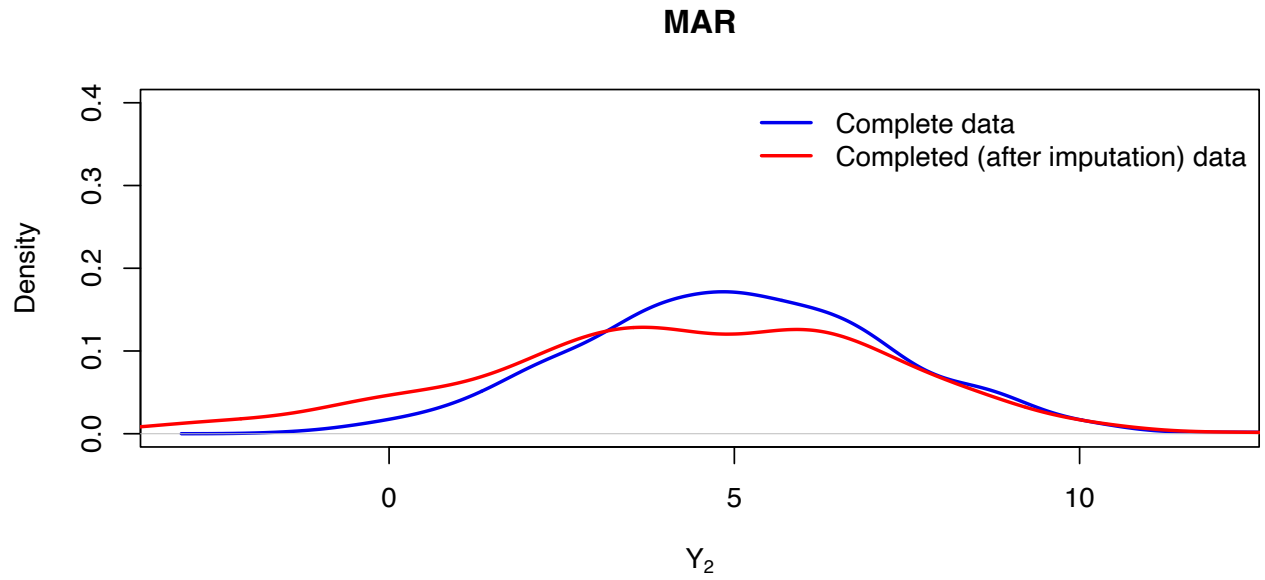
$$Y_2 = \beta_0 + \beta_1 Y_1 + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2),$$

$$\hat{Y}_2 = \hat{\beta}_0 + \hat{\beta}_1 Y_1 + U, \quad U \sim N(0, \hat{\sigma}^2).$$

I will not conduct any check of linearity, constant variance and normality, as by construction, we know that these assumptions are met.

```
data_MAR <- data.frame("Y1" = Y1, "Y2" = Y2_MAR)
fit_MAR <- lm(Y2 ~ Y1, data = data_MAR)
set.seed(1)
predict_MAR_sri <- predict(fit_MAR, newdata = data_MAR) + rnorm(n, 0, sigma(fit_MAR))
Y2_completed_MAR <- ifelse(is.na(Y2_MAR) == TRUE, predict_MAR_sri, Y2)

plot(density(Y2), col = "blue2", xlim = c(-3, 12), ylim = c(0, 0.4),
     main = "MAR", xlab = expression(Y[2]), ylab = "Density", lwd = 2)
lines(density(Y2_completed_MAR), col = "red", lwd = 2)
legend("topright", legend = c("Complete data", "Completed (after imputation) data"),
     col = c("blue2", "red"), lty = c(1,1), lwd = c(2,2), bty = "n")
```



By imputing the missing values using stochastic regression imputation, we are able to ‘reconstruct’ the distribution of the complete data very accurately. Note that here it really helps that Y_1 and Y_2 are highly correlated.