Incomplete Data Analysis

Assignment 1

Xuetao Liu s2257397

2022/2/2

Question 1

- (a) We should choose **(ii)0.3** as our answer. By assumption, ALQ is MCAR, that is to say the probability of ALQ is missing is completely at random. Therefore we have the following

$$P(\text{ALQ is missing}|\text{ALQ=Yes}) = P(\text{ALQ is missing}|\text{ALQ=No}) = P(\text{ALQ is missing}) = 0.3$$

- (b) We should choose **(ii)**. Since ALQ is MAR given gender means, so the probability of ALQ being missing is independent with the Yes/No value of the ALQ, but only depends on the gender.

- (c) We should choose **(iii)**. Since ALQ is MAR given gender, and we have the probability of ALQ being missing for men is 0.1, that is

$$P(\text{ALQ is missing}|\text{Men}) = 0.1.$$

  By definition, we have

$$P(\text{ALQ is missing}) = P(\text{ALQ is missing}|\text{Men}) + P(\text{ALQ is missing}|\text{Women}).$$

  But we do not know the probability of ALQ being missing, and thus we cannot know the probability of ALQ is missing given women.

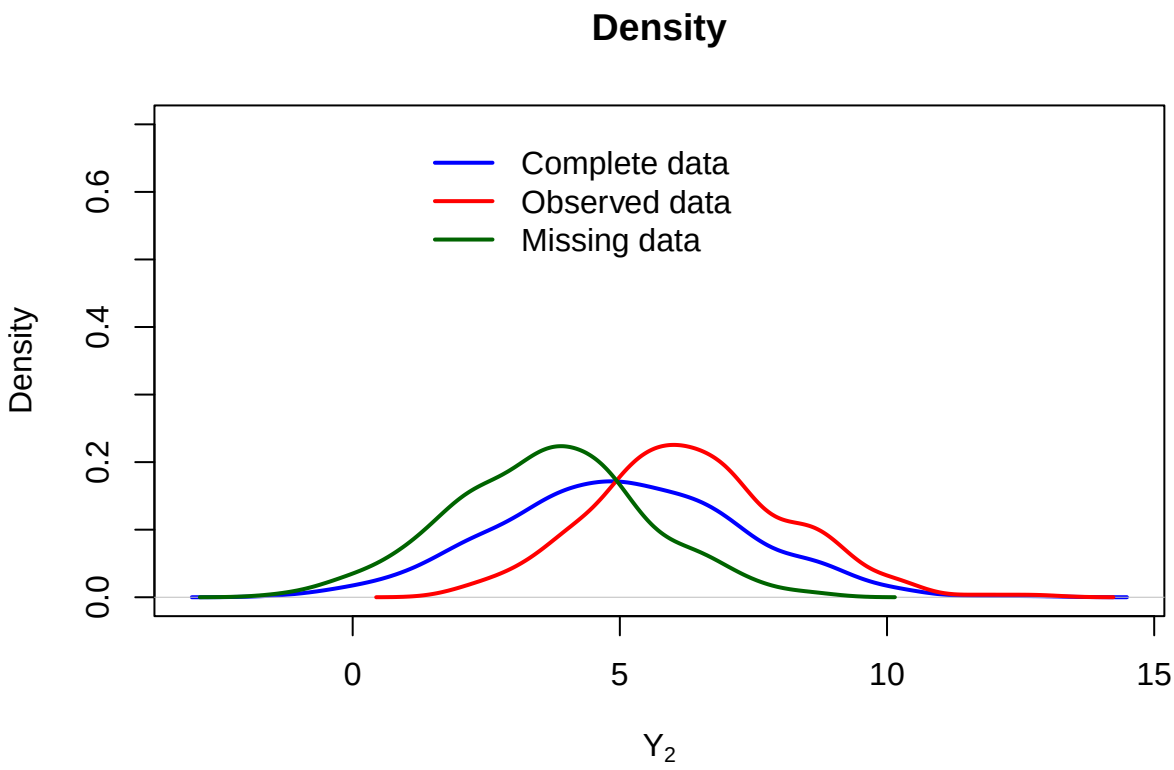
Question 2

- The largest possible subsample under a complete case analysis is that we have the 90% of the original data. In this case, the missing data are all in the same subjects, e.g. subjects 1-10 lost all the data for variable 1-10.
- The smallest subsample is that we have nothing left. In this case, the missing data has no intersection within subjects, e.g. subjects 1-10 lost variable 1, subjects 11-20 lost variable 2, ..., subjects 91-100 lost variable 10.

Question 3

(a)

```r
set.seed(1)
n = 500
mu = 0
sd = 1
a = 2
b = 0

## Generate the standard normal
z1 = rnorm(n,mu,sd)
z2 = rnorm(n,mu,sd)
z3 = rnorm(n,mu,sd)

y1 = 1 + z1
y2 = 5 + 2*z1 + z2

## According to the condition
r = a*(y1-1) + b*(y2-5) + z3

# Find the indicator
ind = which(r<0)
y2_obs = y2[-ind]
y2_mis = y2[ind]

#plotting the densities
plot(density(y2), lwd = 2, col = "blue", xlab = expression(Y[2]),
main = "Density", ylim = c(0, 0.7))
lines(density(y2_obs), lwd = 2, col = "red")
lines(density(y2_mis), lwd = 2, col = "darkgreen")
legend(1, 0.7, legend = c("Complete data", "Observed data", "Missing data"),
col = c("blue", "red", "darkgreen"), lty = c(1,1,1), lwd = c(2,2,2), bty ="n")
```



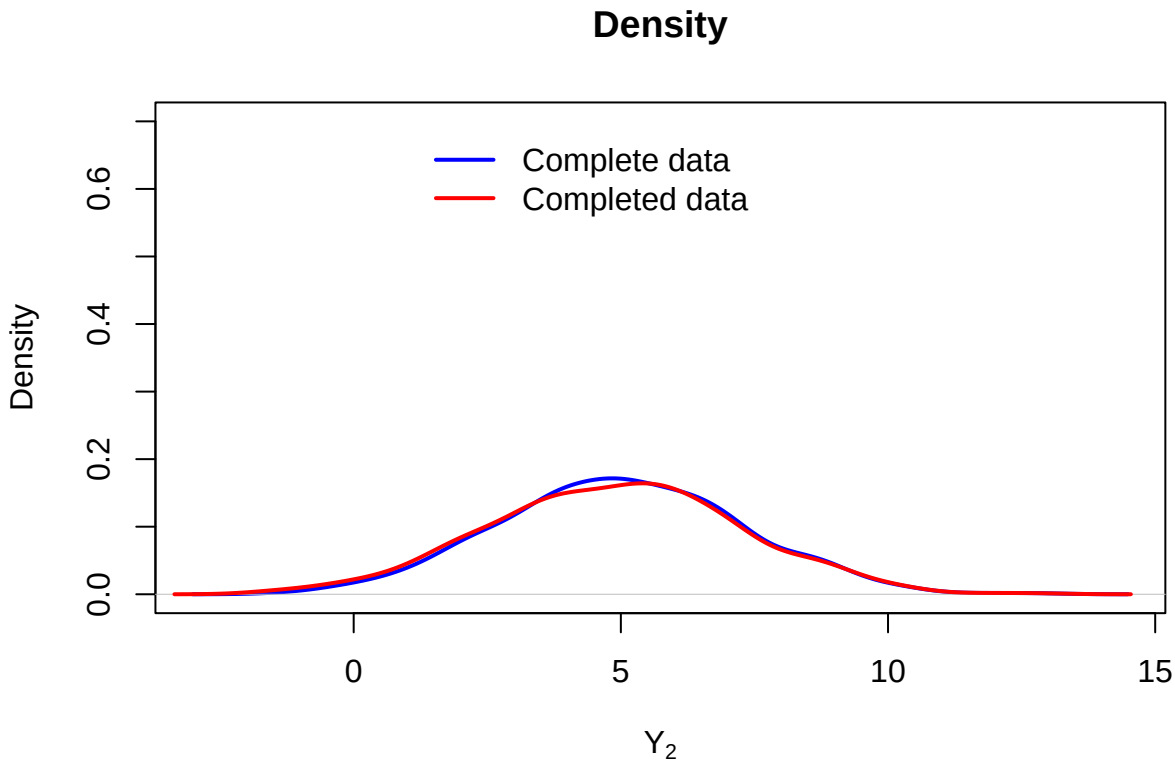
The above figure is the plot of density of complete data, observed data and missing data. By default, we set $a = 2$ and $b = 0$, which causes the missingness depends on the value of Y_1, which is **MAR**.

(b)

Here we use stochastic regression imputation, first we come to fit the regression model using the observed data

```
## ind is the index of those missing
## We can first create the data frame for regression
data = data.frame(y1, y2)
data$y2[ind] = NA # Impose the NA

fit = lm(y2 ~ y1, data = data)

# The predict values for the NAs
predicted_sri <- predict(fit, newdata = data) + rnorm(nrow(data), 0, sigma(fit))

# Impute the SRI
y2_sri <- ifelse(is.na(data$y2), predicted_sri, data$y2)

#plotting the densities
plot(density(y2), lwd = 2, col = "blue", xlab = expression(Y[2]),
main = "Density", ylim = c(0, 0.7))
lines(density(y2_sri), lwd = 2, col = "red")
```

```
legend(1, 0.7, legend = c("Complete data", "Completed data"),
col = c("blue", "red"), lty = c(1,1), lwd = c(2,2), bty ="n")
```

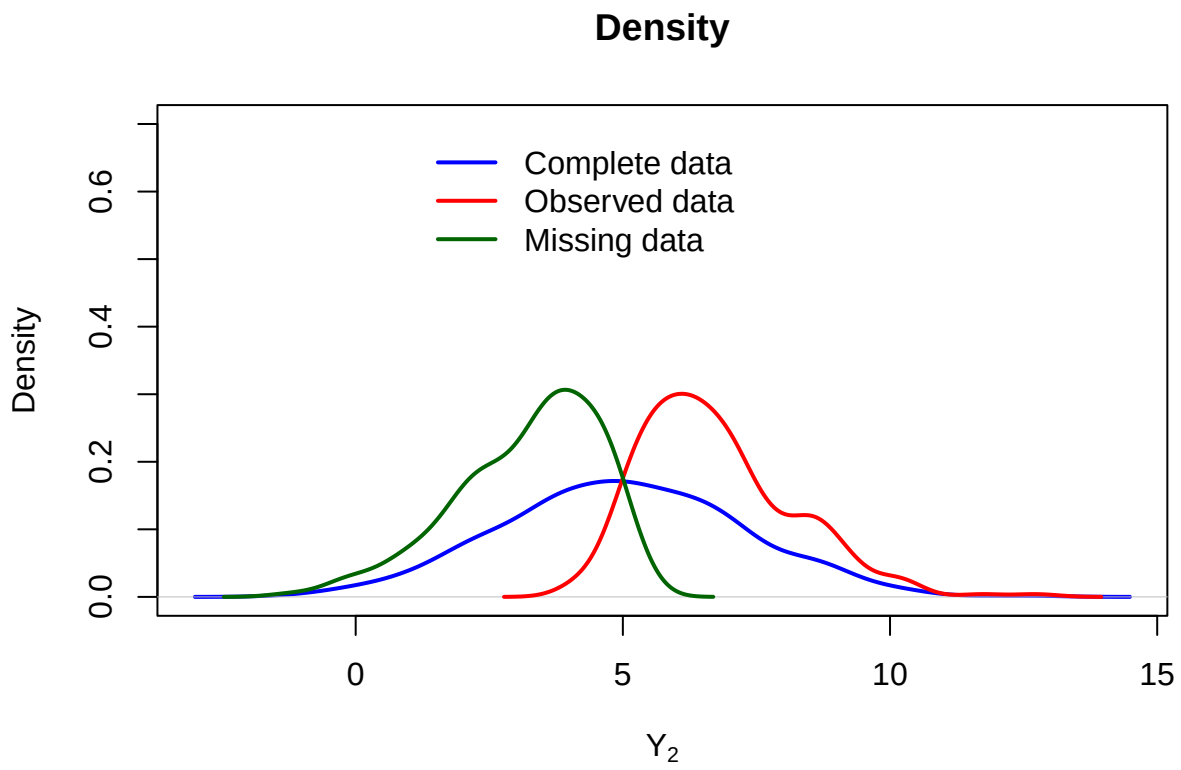


(c)

Here we take $a = 0$ and $b = 2$, which makes the missingness become **MNAR**

```
set.seed(1)
n = 500
mu = 0
sd = 1
a = 0
b = 2

## Generate the standard normal
z1 = rnorm(n,mu,sd)
z2 = rnorm(n,mu,sd)
z3 = rnorm(n,mu,sd)

y1 = 1 + z1
y2 = 5 + 2*z1 + z2

## According to the condition
r = a*(y1-1) + b*(y2-5) + z3
```

```r
# Find the indicator
ind = which(r<0)
y2_obs = y2[-ind]
y2_mis = y2[ind]

#plotting the densities
plot(density(y2), lwd = 2, col = "blue", xlab = expression(Y[2]),
main = "Density", ylim = c(0, 0.7))
lines(density(y2_obs), lwd = 2, col = "red")
lines(density(y2_mis), lwd = 2, col = "darkgreen")
legend(1, 0.7, legend = c("Complete data", "Observed data", "Missing data"),
col = c("blue", "red", "darkgreen"), lty = c(1,1,1), lwd = c(2,2,2), bty ="n")
```

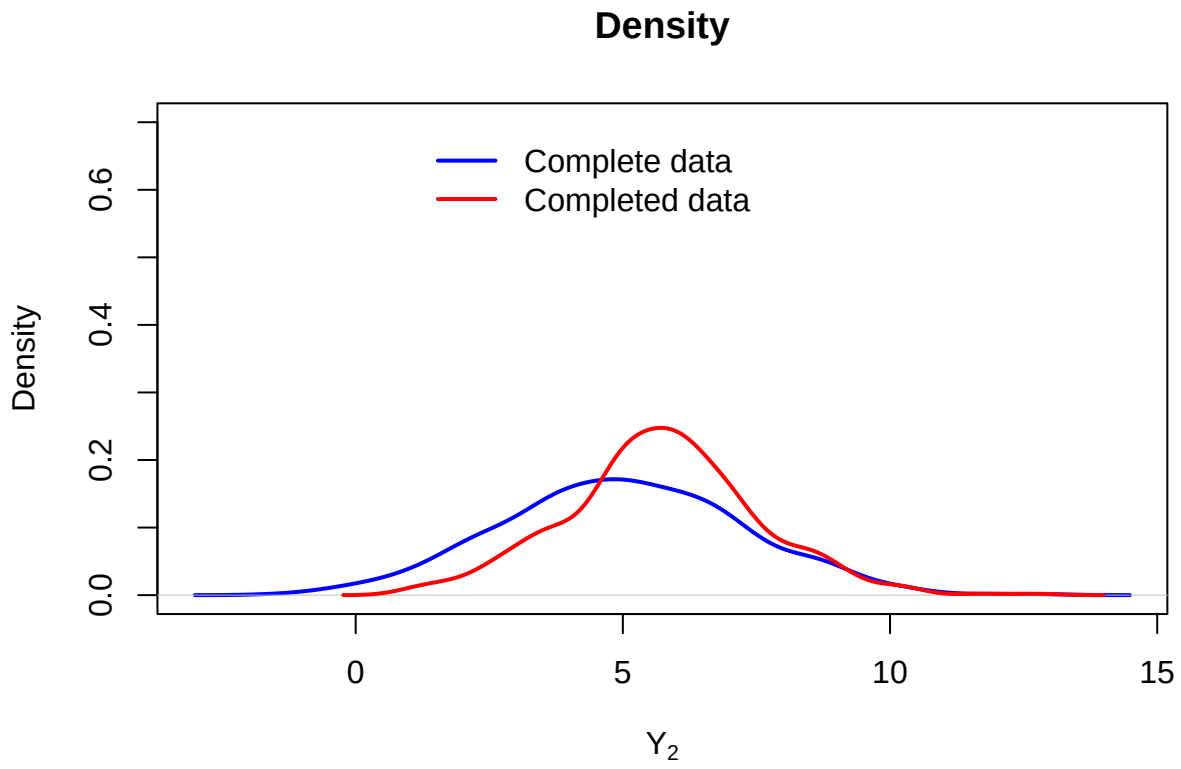


(d)

Like (b), we first create the dataframe, then fit the regression and then use the predicted value to conduct the stochastic regression imputation.

```r
## ind is the index of those missing
## We can first create the data frame for regression
data = data.frame(y1, y2)
data$y2[ind] = NA # Impose the NA

fit = lm(y2 ~ y1, data = data)
```

```r
# The predict values for the NAs
predicted_sri <- predict(fit, newdata = data) + rnorm(nrow(data), 0, sigma(fit))

# Impute the SRI
y2_sri <- ifelse(is.na(data$y2), predicted_sri, data$y2)

#plotting the densities
plot(density(y2), lwd = 2, col = "blue", xlab = expression(Y[2]),
main = "Density", ylim = c(0, 0.7))
lines(density(y2_sri), lwd = 2, col = "red")
legend(1, 0.7, legend = c("Complete data", "Completed data"),
col = c("blue", "red"), lty = c(1,1), lwd = c(2,2), bty ="n")
```



Question 4

(a)

First, we load the data

```r
load("databp.RData")
```

The mean of the recovery time after removing the NAs is given as follow:

```r
# The mean of recovery time after removing the NAs
mean(databp$recovtime, na.rm = T) ##19.27273
```

```
## [1] 19.27273
```

and the associated standard error of recovery time is

```r
sd(databp$recovtime, na.rm = T) ##12.20922
```

```
## [1] 12.20922
```

The Pearson correlations between the recovery time and the dose is given as

```r
cor(databp$recovtime, databp$logdose, use = "complete", method = "pearson") ##0.2391256
```

```
## [1] 0.2391256
```

The Pearson correlations between the recovery time and the blood pressure is given as

```r
cor(databp$recovtime, databp$bloodp, use = "complete", method = "pearson") ##-0.01952862
```

```
## [1] -0.01952862
```

(b)

To conduct mean imputation, we first need to calculate the mean value of recovery time

```r
m_recov = mean(databp$recovtime, na.rm = TRUE)
m_recov
```

```
## [1] 19.27273
```

Then we substitute the NAs by the mean

```r
recov_mi <- ifelse(is.na(databp$recovtime), m_recov, databp$recovtime)
```

Since we just substitute the NAs by its mean, which would not change the global mean, so the mean after imputation remains the same

```r
mean(recov_mi)
```

```
## [1] 19.27273
```

```r
sd(recov_mi)
```

```
## [1] 11.42068
```

We can see that the standard error after imputations is **11.42068**, which is lower than the original **12.20922**, which is the same as our conclusion that mean imputation would lower the standard error. The Pearson correlations between the recovery time after imputation and the dose is given by

```r
cor(recov_mi, databp$logdose, use = "complete", method = "pearson") ##0.2150612
```

```
## [1] 0.2150612
```

We can see that the correlation also decreases from **0.2391256** to **0.2150612**. The Pearson correlations between the recovery time after imputation and the blood pressure is given as

```r
cor(recov_mi, databp$bloodp, use = "complete", method = "pearson") ##-0.01952862
```

```
## [1] -0.01934126
```

The correlation also decreases from **|-0.01952862|** to **|-0.01934126|**, since we only care about the magnitude.

(c)

(d)

(e)

(f)