Incomplete Data Analysis

Assignment 1

Callum Abbott

Question 1

Let us suppose that on a (hypothetical) survey there is a question about alcohol consumption and that the variable ALQ records the respondent's answer to the following specific question: "In the past year, have you had at least 12 drinks of any type of alcoholic beverage?". The possible answers are 'Yes' or 'No'. Not all participants respond to this question, that is, the ALQ variable has some missing values. Further, and again hypothetically, suppose that we only have additional data on the gender of the participants in the survey (which is fully observed). For each of the following situations, choose, justifying, the correct answer.

Question (1a)

- **Description**: Suppose that ALQ is MCAR. The probability of ALQ being missing for those with ALQ=Yes is 0.3. What is the probability of ALQ being missing for those with ALQ=No?
- **Answer**: In our hypothetical survey about alcohol consumption two variables are recorded: ALQ and gender. We note that the ALQ variable contains some missing values whilst the gender variable is fully observed. We now suppose that ALQ is MCAR, indicating that the missingness of the ALQ variable is independent of any data. This implies that the probability of the ALQ being missing for those with ALQ=No must **also be 0.3**. If it were any other value, the probability of the data being missing would then have an implied dependence on the ALQ variable rendering the data to **not** be MCAR.

Question (1b)

- **Description**: ALQ being MAR given gender means:
- **Answer**: (ii) The probability of ALQ being missing is independent of the Yes/No value of ALQ after adjusting for gender.

Question (1c)

- **Description**: Suppose again that ALQ is MAR given gender, and that the probability of ALQ being missing for men is 0.1. What is the probability of ALQ being missing for women?
- **Answer**: (iii) It is impossible to conclude from the information given. I believe this because answer (i) of 0.1 would imply that the data would *actually* be MCAR. Meanwhile, answer (ii) seems to indicate that some conservation of probability must be present which is incorrect to assume. This leaves answer (iii).

Question 2

- **Description**: Suppose that a dataset consists of 100 subjects and 10 variables. Each variable contains 10% of missing values. What is the largest possible subsample under a complete case analysis? What is the smallest? Justify.
- **Answer - Largest CCA Subset:** Let us begin with first defining how our data is presented and stored. We can view the data in a table format with 100 rows and 10 columns. The question does not state the total number of missing values in our dataset of $100 \times 10 = 1000$ values, and so let us first assume a scenario where we have 10 missing values (out of 1000) in our dataset. In order to satisfy the criterion that "*each variable (column) contains 10% of missing values*", then we must have 1 missing value in each of our 10 variables. We know that in a complete case analysis that if any one of a subject's 10 variables contains a missing value, then that subject will not be included in the analysis. Therefore, if all our 10 missing values align along a single row in our dataset, then we would only have to discard a single subject from our complete case analysis leaving a total of **99 subjects**.
- **Answer - Smallest CCA Subset:** We now imagine the converse situation where we have a much greater number of missing values in our dataset. For argument's sake, let us propose that this number is now 100. This time, in order to satisfy the criterion that "*each variable (column) contains 10% of missing values*", we must have 10 missing values in each of our 10 variables. Now suppose that rows 1-10 were missing for variable 1, rows 11-20 were missing for variable 2, rows 21-30 were missing for variable 3 and so on up to variable 10. This kind of structure in our dataset would result in a missing value being present for every subject in our dataset leaving a total of **0 subjects** under a complete case analysis.

Question 3

- **Description**: Consider a two variable (Y_1, Y_2) problem, with each variable defined as follows:

  $$Y_1 = 1 + Z_1,\ Y_2 = 5 + 2 \times Z1 + Z2$$

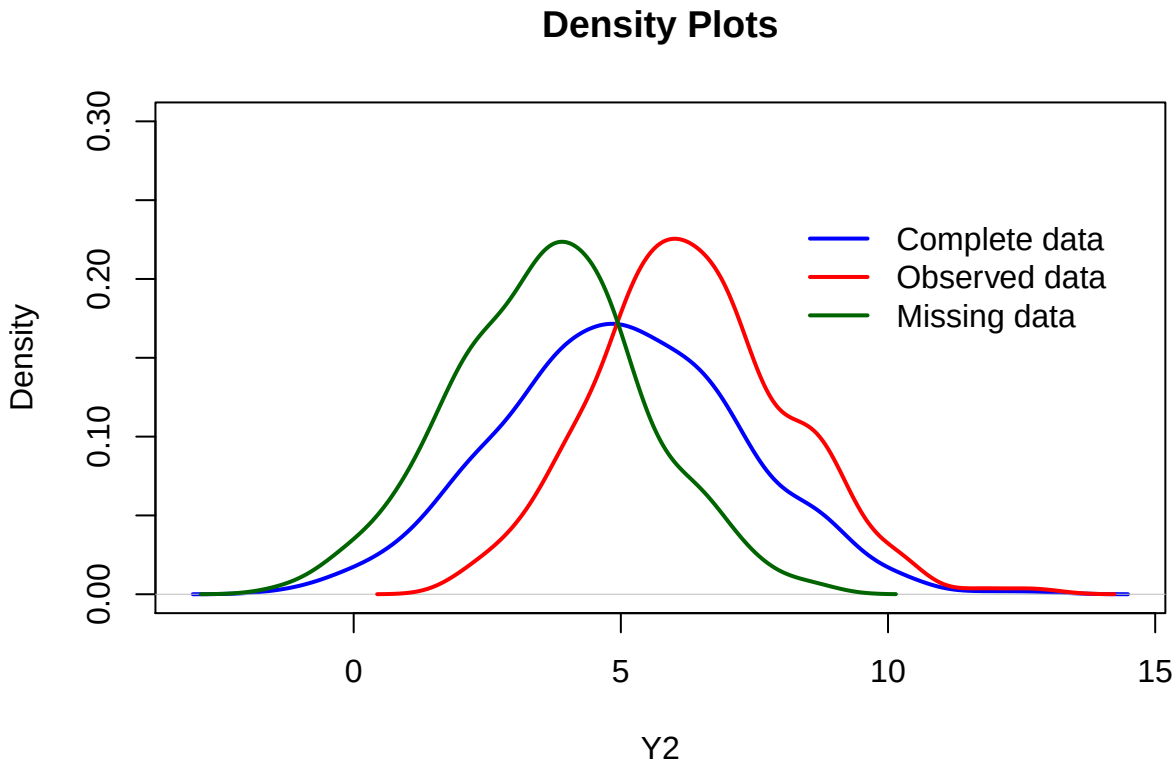
  where Y_1 is fully observed but Y_2 is subject to missingness. Further consider that $Y2$ is missing if $a \times (Y_1 - 1) + b \times (Y_2 - 5) + Z_3 < 0$, where Z_1, Z_2, and Z_3 follow independent standard normal (that is, mean 0 and variance 1) distributions.

Question (3a)

- **Description**: Start by simulating a (complete) dataset of size 500 on (Y_1, Y_2). Then, and considering $a = 2$ and $b = 0$, simulate the corresponding observed dataset (by imposing missingness on Y_2 as instructed above). Is this mechanism MCAR, MAR, or MNAR? Display the marginal distribution of Y_2 for the complete (as originally simulated) and observed (after imposing missingness) data. Comment.
- **Answer**: Given that we know Y_1 is completely observed and Y_2 is only partially observed, we see that our missingness indicator, r, for the parameters $a = 2$ and $b = 0$, is only dependent on Y_1. This, by definition, tells us that the missing data for Y_2 will be MAR. Our beliefs are then confirmed after observing that the distributions (density and box) of the complete, missing and observed data of $Y2$ are **not congruent** (see below).

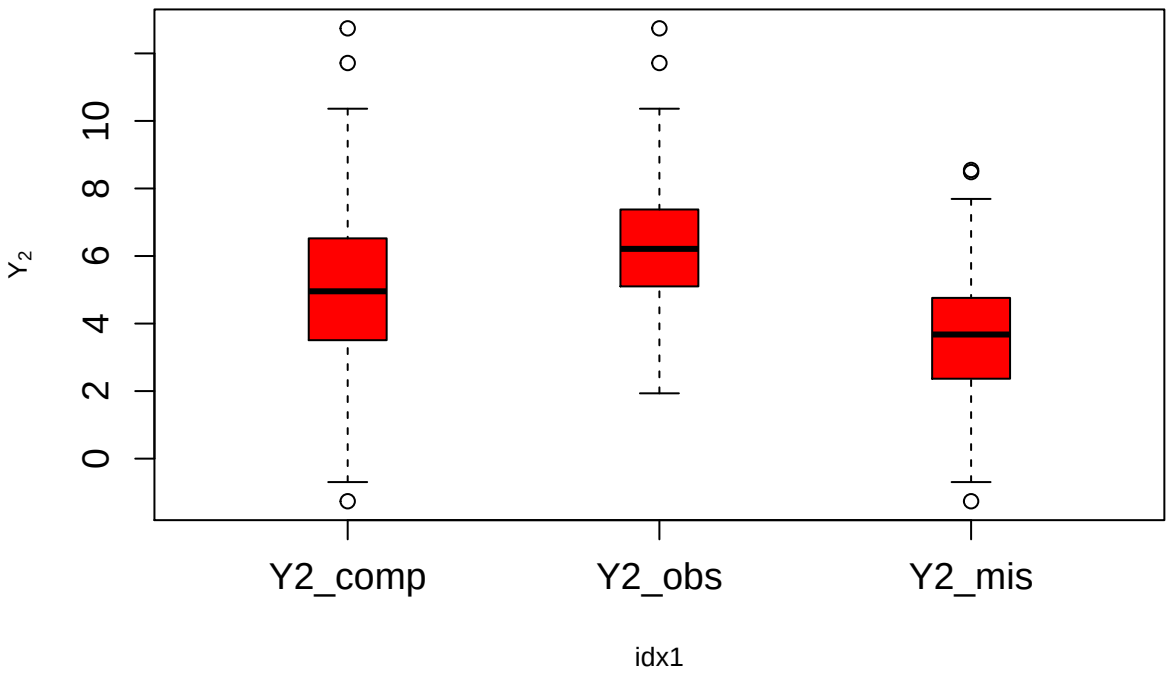
Density Plot

```r
set.seed(1)

get_dataset = function(nsim, mean, sd){
  # Simulating univariate Gaussians
  Z1 = rnorm(n = nsim, mean = mean, sd = sd)
  Z2 = rnorm(n = nsim, mean = mean, sd = sd)
  Z3 = rnorm(n = nsim, mean = mean, sd = sd)
  # Defining Ys in terms of Zs
  Y1 = 1 + Z1
  Y2 = 5 + 2 * Z1 + Z2
  return(data.frame(Y1, Y2, Z1, Z2, Z3))
}

# Generating dataset
nsim = 500
dataset = get_dataset(nsim, 0, 1)

# Missingness indicator params
a = 2; b = 0
# Defining missingness indicator
r = a * (dataset$Y1 - 1)  +  b * (dataset$Y1 - 5) + dataset$Z3

# Imposing missingness
mis_indxs = which(r < 0)
Y2_obs = dataset$Y2[-mis_indxs]
Y2_mis = dataset$Y2[mis_indxs]

# Plotting
plot(density(dataset$Y2), lwd = 2, col = "blue", xlab = expression(Y2),
     main = "Density Plots", ylim = c(0, 0.30))
lines(density(Y2_obs), lwd = 2, col = "red")
lines(density(Y2_mis), lwd = 2, col = "darkgreen")
legend(8, 0.25, 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")
```



Box Plot

```
# Get number of observed and missing variables.
n_obs = length(Y2_obs)
n_mis = length(Y2_mis)

# Ordering indices
idxs = rep("Y2_comp", nsim + n_obs + n_mis)
idxs[(nsim+1):(nsim+n_obs)] = "Y2_obs"
idxs[(nsim+n_obs+1):(nsim+n_obs+n_mis)] = "Y2_mis"
idx1 = factor(idxs, levels = c("Y2_comp","Y2_obs","Y2_mis"))

# Plotting box plots
Y2boxmcar = c(dataset$Y2, Y2_obs,Y2_mis)
boxplot(Y2boxmcar ~ idx1, boxwex = 0.25, col = "red", cex.lab = 0.8,
        cex.axis = 1.2, ylab = expression(Y[2]))
```

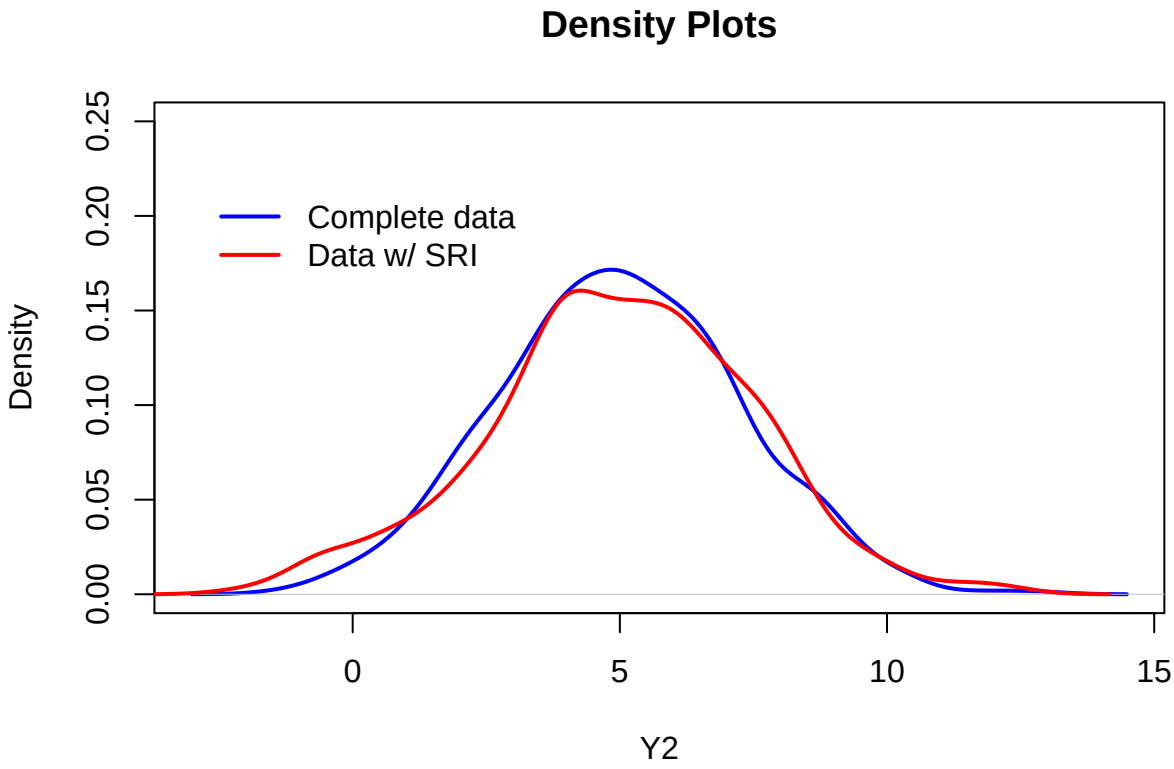


Question (3b)

- **Description**: For the observed dataset simulated in (a), impute the missing values using stochastic regression imputation. Display the marginal distribution of Y2 for the complete (as originally simulated) and completed (after imputation) data. Comment.
- **Answer**: We see that both the complete data density plot, and the density plot after imposing missingness and then imputing the missing values using stochastic regression imputation, are very similar. The only distinguishing difference between the two is that the "Data w/ SRI" density plot is more noisy and less smooth than the complete data density plot as one would expect. I have quantified the similarity between these distributions by computing the mean of both distributions for which I observe their means to both be ≈ 5 to two decimal places. I have also computed the standard deviations for both distributions and observe the standard deviation of `Y2_sri` to be slightly larger (2.47) than the complete `Y2` (2.24) distribution which is reflective of the stochastic noise we added to `Y2_sri`. Finally, I have computed the correlation coefficient between Y_1 and Y_2 for the complete dataset and the dataset that has SRI imputed values for the missing values of Y_2. We observe a slight increase from 0.8823041 to 0.896824 due to our imputed values lying on the regression line; effectively inducing an artificial positive correlation between the two variables. Overall, I believe SRI has done a good job at imputing the missing data values due to the relatively unbiased estimates of the mean, standard deviation and correlation coefficient.

```
set.seed(1)
# Generating dataset
nsim = 500
complete_dataset = get_dataset(nsim, 0, 1)
```

```r
dataset = get_dataset(nsim, 0, 1)

# Missingness indicator params
a = 2; b = 0
# Defining missingness indicator
r = a * (complete_dataset$Y1 - 1) + b * (complete_dataset$Y2 - 5) +
         complete_dataset$Z3

# Imposing missingness
mis_indxs = which(r < 0)
dataset$Y2[mis_indxs] = NA

# Regressing Y2 on Y1
fits = lm(Y2 ~ Y1, data = dataset[c(1, 2)])
coeffs = fits$coefficients

# Generating predictions
pred_sri = predict(fits, newdata = dataset[c(1, 2)]) + rnorm(nsim, 0, sigma(fits))
Y2_sri = ifelse(is.na(dataset$Y2) == TRUE, pred_sri, dataset$Y2)

# Plotting
plot(density(complete_dataset$Y2), lwd = 2, col = "blue", xlab = expression(Y2),
     main = "Density Plots",  ylim = c(0, 0.25))
lines(density(Y2_sri), lwd = 2, col = "red")
legend(-3, 0.22, legend = c("Complete data", "Data w/ SRI"),
       col = c("blue", "red"), lty = c(1,1,1), lwd = c(2,2,2), bty ="n")
```




```
mean_sri = mean(Y2_sri)
err_sri = sd(Y2_sri)/sqrt(nsim)
mean_comp = mean(complete_dataset$Y2)
sd_sri = sd(Y2_sri)
sd_comp = sd(complete_dataset$Y2)
cor = cor(complete_dataset$Y1, complete_dataset$Y2)
cor_sri = cor(dataset$Y1, Y2_sri)
mean_sri; err_sri; mean_comp; sd_sri; sd_comp; cor; cor_sri
```

```
## [1] 4.998138
```

```
## [1] 0.1103595
```

```
## [1] 4.999348
```

```
## [1] 2.467714
```

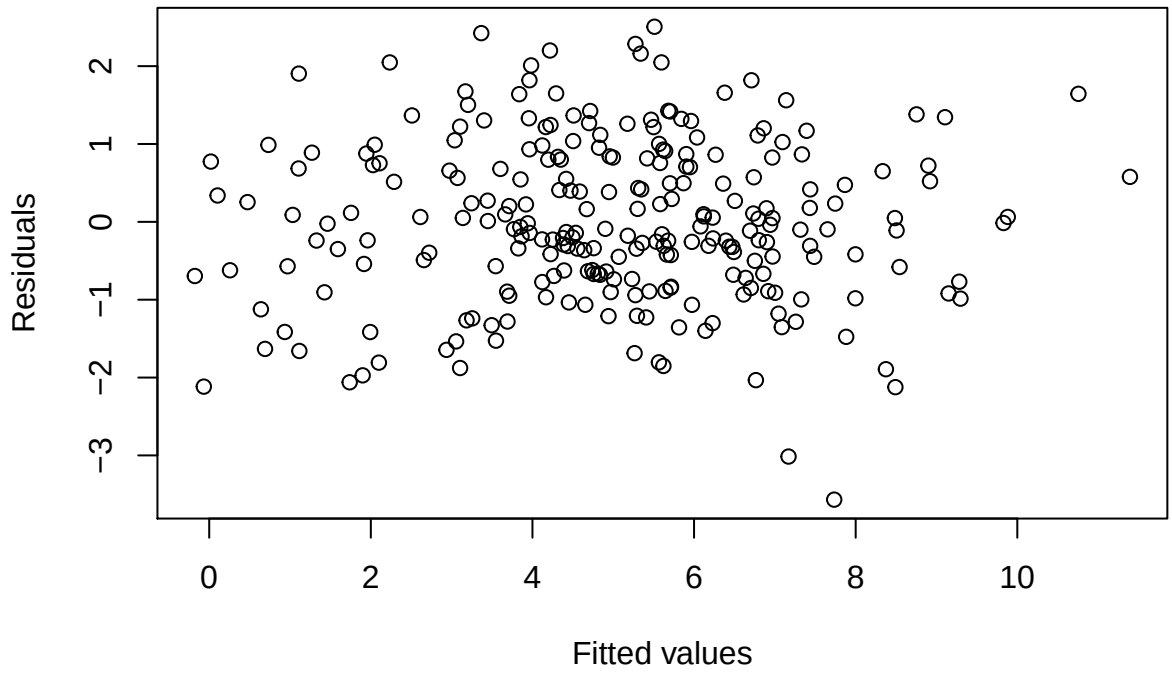
```
## [1] 2.244424
```

```
## [1] 0.8823041
```

```
## [1] 0.896824
```

We also check that our linear regression fit on Y2 with Y1 is valid before moving on by plotting the residuals. We see no correlation between the residuals and the fitted values, and hence, the linear regression fit was valid.

```
plot(fits$fitted.values, residuals(fits), xlab = "Fitted values", ylab = "Residuals")
```

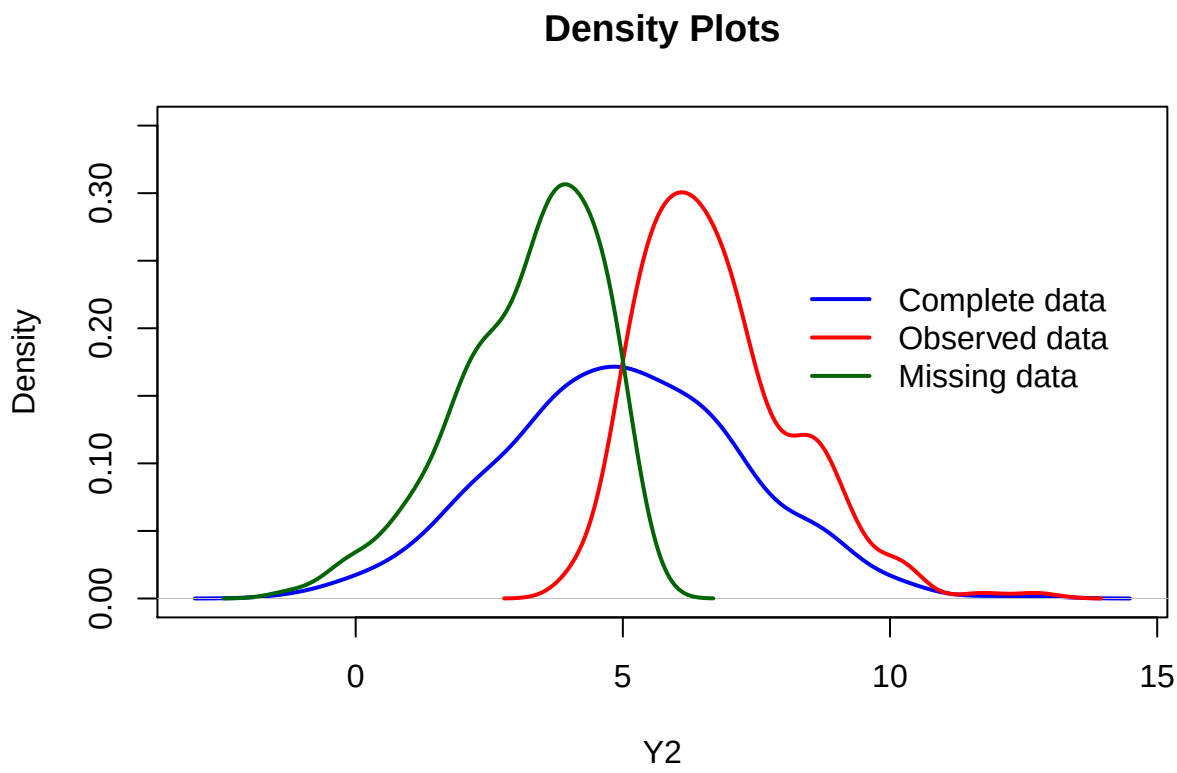


Question (3c)

- **Description**: Using the complete dataset simulated in (a), now impose missingness on Y2 by considering $a = 0$ and $b = 2$. Is this mechanism MCAR, MAR, or MNAR? Display the marginal distribution of Y2 for the complete (as originally simulated) and observed (after imposing missingness) data. Comment.
- **Answer**: Given that we know Y_1 is completely observed and Y_2 is only partially observed, we see that our missingness indicator, r, for the parameters $a = 0$ and $b = 2$, is only dependent on Y_2. This, by definition, tells us that the missing data for Y_2 will be MNAR since the missingess depends on the missing variable itself. Our beliefs are then confirmed after observing that the distributions (density and box) of the complete, missing and observed data of $Y2$ are **even less congruent** than the plots seen in 3(a) (see below).

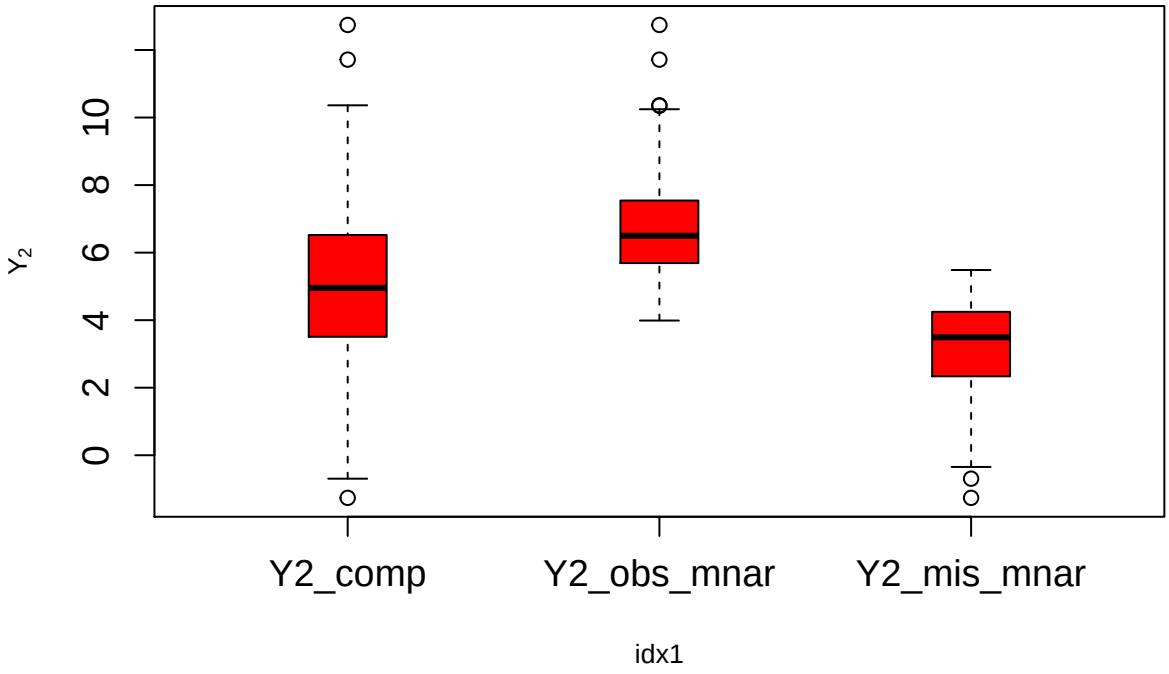
```
set.seed(1)
complete_dataset = get_dataset(nsim=500, mean=0, sd=1)

# Missingness indicator params
a = 0; b = 2
```

```r
# Defining missingness indicator
r = a * (complete_dataset$Y1 - 1) + b * (complete_dataset$Y2 - 5) + complete_dataset$Z3

# Imposing missingness
mis_indxs = which(r < 0)
Y2_obs_mnar = complete_dataset$Y2[-mis_indxs]
Y2_mis_mnar = complete_dataset$Y2[mis_indxs]

# Plotting
plot(density(complete_dataset$Y2), lwd = 2, col = "blue", xlab = expression(Y2), main = "Density Plots"
     ylim = c(0, 0.35))
lines(density(Y2_obs_mnar), lwd = 2, col = "red")
lines(density(Y2_mis_mnar), lwd = 2, col = "darkgreen")
legend(8, 0.25, 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")
```




```r
# Get number of observed and missing variables.
n_obs = length(Y2_obs_mnar)
n_mis = length(Y2_mis_mnar)

# Ordering indices
idxs = rep("Y2_comp", nsim + n_obs + n_mis)
idxs[(nsim+1):(nsim+n_obs)] = "Y2_obs_mnar"
idxs[(nsim+n_obs+1):(nsim+n_obs+n_mis)] = "Y2_mis_mnar"
idx1 = factor(idxs, levels = c("Y2_comp","Y2_obs_mnar","Y2_mis_mnar"))
```

```r
# Plotting box plots
Y2boxmcar = c(complete_dataset$Y2, Y2_obs_mnar,Y2_mis_mnar)
boxplot(Y2boxmcar ~ idx1, boxwex = 0.25, col = "red", cex.lab = 0.8,
        cex.axis = 1.2, ylab = expression(Y[2]))
```

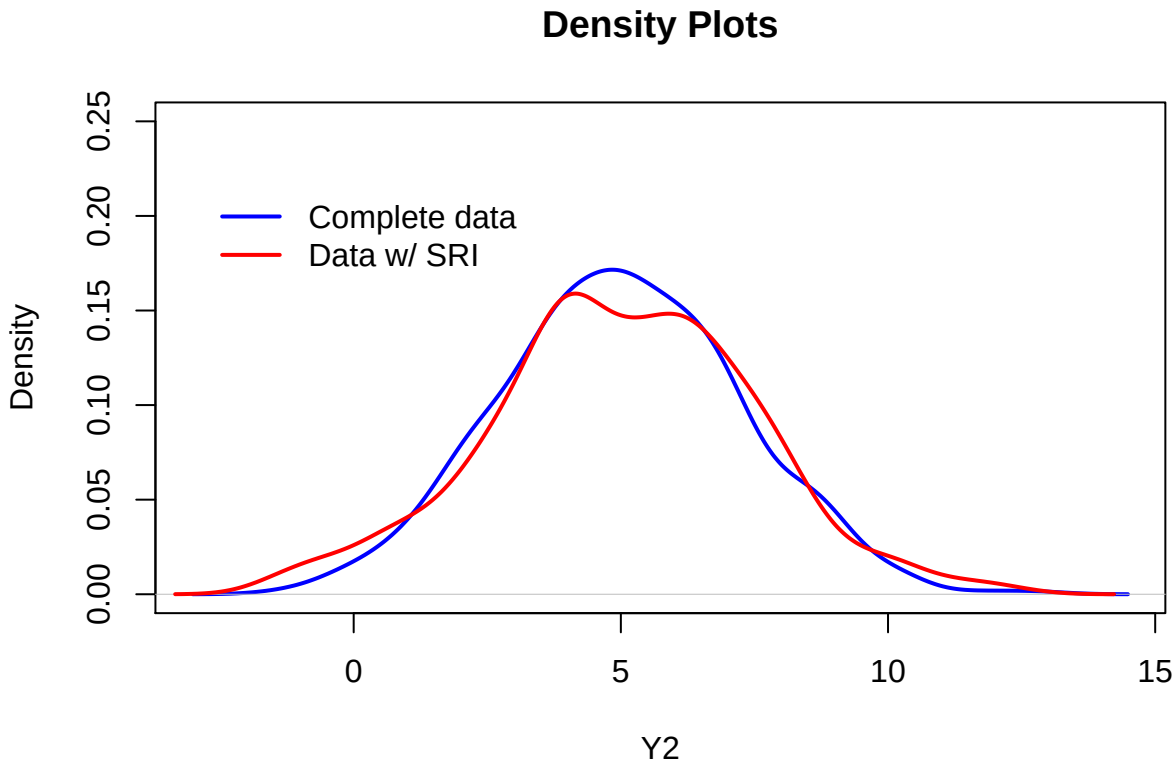


Question (3d)

- **Description:** The same as in (b) but for the observed data generated in (c).
- **Answer**: We see that both the complete data density plot, and the density plot after imposing missingness and then imputing the missing values using stochastic regression imputation, are **still** very similar. I have quantified the similarity between these distributions by computing the mean of both distributions for which I observe their means to both be ≈ 5 now only to **one** decimal place. I have also computed the standard deviations for both distributions and observe the standard deviation of `Y2_sri` to be slightly larger than observed in (b) (2.49 from 2.47). Finally, I have computed the correlation coefficient between Y_1 and Y_2 for the complete dataset and the dataset that has SRI imputed values for Y_2. The increase is now from 0.8823041 to 0.9031965 indicating SRI has more influence on MNAR data than MAR. Overall, **in this specific context**, SRI still provides relatively unbiased estimates for the relevant quanities however this is not the case in general.

```r
set.seed(1)
# Generating dataset
nsim = 500
complete_dataset = get_dataset(nsim, 0, 1)
```

```r
dataset = get_dataset(nsim, 0, 1)

# Missingness indicator params
a = 0; b = 2
# Defining missingness indicator
r = a * (complete_dataset$Y1 - 1) + b * (complete_dataset$Y2 - 5) +
         complete_dataset$Z3

# Imposing missingness
mis_indxs = which(r < 0)
dataset$Y2[mis_indxs] = NA

# Regressing Y2 on Y1
fits = lm(Y2 ~ Y1, data = dataset[c(1, 2)])
coeffs = fits$coefficients

# Generating predictions
pred_sri = predict(fits, newdata = dataset[c(1, 2)]) + rnorm(nsim, 0, sigma(fits))
Y2_sri = ifelse(is.na(dataset$Y2) == TRUE, pred_sri, dataset$Y2)

# Plotting
plot(density(complete_dataset$Y2), lwd = 2, col = "blue", xlab = expression(Y2),
     main = "Density Plots",  ylim = c(0, 0.25))
lines(density(Y2_sri), lwd = 2, col = "red")
legend(-3, 0.22, legend = c("Complete data", "Data w/ SRI"),
       col = c("blue", "red"), lty = c(1,1,1), lwd = c(2,2,2), bty ="n")
```




```
mean_sri = mean(Y2_sri)
err_sri = sd(Y2_sri)/sqrt(nsim)
mean_comp = mean(complete_dataset$Y2)
sd_sri = sd(Y2_sri)
sd_comp = sd(complete_dataset$Y2)
cor = cor(complete_dataset$Y1, complete_dataset$Y2)
cor_sri = cor(dataset$Y1, Y2_sri)
mean_sri; err_sri; mean_comp; sd_sri; sd_comp; cor; cor_sri
```

```
## [1] 5.010719
```

```
## [1] 0.1112771
```

```
## [1] 4.999348
```

```
## [1] 2.488232
```

```
## [1] 2.244424
```

```
## [1] 0.8823041
```

```
## [1] 0.9031965
```