Assignment\_1

Peilin Long

S2038784

github:<https://github.com/lllucy-lpl/IDA_assignment1>

## Question 1

1. The answer is (ii) As ALQ is MCAR then probability of missing data of ALQ is unrelated to our observed data and also the missing value itself, therefore, ,the distribution of ALQ = NO is similar to distribution of ALQ = YES, so is the same as .
2. The answer is (ii)The probability of ALQ being missing is independent of the Yes/No value of ALQ after adjusting for gender. MAR means missing value is related to observed data but unrelated to missing data itself. So here missing value of ALQ is independent to its own value Yes/No, but is depended on given data gender.
3. The answer is (ii)It is impossibile to conclude from the information given. As ALQ is MAR given gender, the probability of ALQ is missing depends on gender, even though we konw the probability of ALQ missing for men is 0.1, we could not make further inference on the exact probability of ALQ missing for women.

## Question2

The largest subsample of this dataset would be 90 under complete case analysis: The data for any individual that has one or more is missing will be discarded, and when the 10% of missing valu in each variable is in the same row, so that the left 90 rows would remain in the dataset. The smallest subsample of this dataset would be 0 under complete case analysis: If the 10 missing values of each individuals lies in different rows and there is no overlap among the missing values position, in this case every individual contains a missing value so that all the data should be discarded, therefore, the subsample size turns to be 0.

## Question 3

# simulating a complete dataset of size = 500  
set.seed(111)  
n = 500  
z1 = rnorm(n)  
z2 = rnorm(n)  
z3 = rnorm(n)  
y1 = 1 + z1  
y2 = 5 + 2 \* z1 + z2  
df = data.frame(  
 Y1 = y1,  
 Y2 = y2  
)  
df[1:10,]

## Y1 Y2  
## 1 1.23522071 6.866263  
## 2 0.66926413 3.446542  
## 3 0.68837618 5.441782  
## 4 -1.30234566 1.196118  
## 5 0.82912396 4.290173  
## 6 1.14027823 5.553853  
## 7 -0.49742666 1.066547  
## 8 -0.01018842 1.739925  
## 9 0.05152440 3.952315  
## 10 0.50603778 3.208172

Consider y2 is missing if .

For a = 2 and b = 0 y2 is missing when The missing value is related to the observed value y1 but do not depend on missing value itself.

Therefore, it is a MAR mechanism.

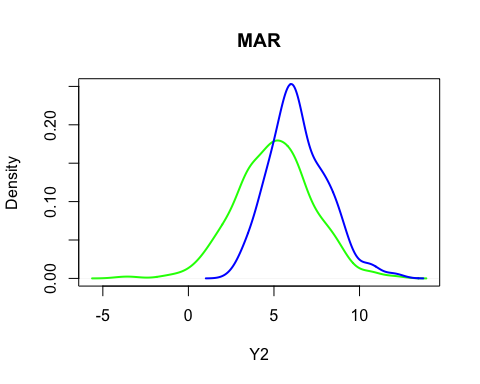
set.seed(111)  
Y2\_MAR\_obs = df$Y2[2\*(y1-1)+z3>=0]  
mean(Y2\_MAR\_obs)

## [1] 6.414957

sd(Y2\_MAR\_obs)

## [1] 1.715735

#Plot the densities of complete dataset and observed dataset  
  
plot(density(df$Y2), lwd = 2, col = "green", xlab = "Y2", main = "MAR", ylim = c(0,0.25))  
lines(density(Y2\_MAR\_obs), lwd = 2, col = "blue")  
legend(20, 0.25, legend = c("Complete data","Observed data"), col = c("green","blue"), lty = c(1,1), lwd = c(2,2), bty = "n")



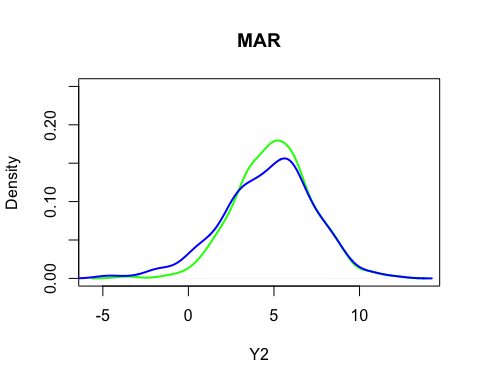
From the plot we can see that the distributions of observed data and complete data are quite different under data generated according to MAR.

For stochastic regression we will add a noise value to the predictions obtained using the regression imputation method.

set.seed(111)  
ind = which(2\*(y1-1)+z3<0)  
y2\_b = y2  
y2\_b[ind] = NA  
df\_1 = data.frame(  
 Y1 = y1,  
 Y2 = y2\_b  
)  
fity = lm(Y2 ~ Y1, data = df\_1)  
summary(fity)

##   
## Call:  
## lm(formula = Y2 ~ Y1, data = df\_1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.3536 -0.6644 -0.0274 0.6463 2.6022   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.90768 0.15447 18.82 <2e-16 \*\*\*  
## Y1 2.08447 0.08492 24.55 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9297 on 249 degrees of freedom  
## (249 observations deleted due to missingness)  
## Multiple R-squared: 0.7076, Adjusted R-squared: 0.7064   
## F-statistic: 602.5 on 1 and 249 DF, p-value: < 2.2e-16

set.seed(111)  
predy2\_b = predict(fity,newdata = df\_1) + rnorm(n,0,sigma(fity))  
y2sri\_b = ifelse(is.na(df\_1$Y2),predy2\_b,df\_1$Y2)  
plot(density(df$Y2), lwd = 2, col = "green", xlab = "Y2", main = "MAR", ylim = c(0,0.25))  
lines(density(y2sri\_b), lwd = 2, col = "blue")  
legend(20, 0.25, legend = c("Complete data","Observed data"), col = c(zgreen","blue"), lty = c(1,1), lwd = c(2,2), bty = "n")



From the plot we can see that the distribution of oberved data after imputation is quite similar to the distribution of originally completed data. So it is make sense to use stochastic regression to deal with missing values under MAR mechanism.

With a = 0 and b = 2 Y2 will be missing if It is a MNAR mechanism as the missing values of Y2 are due to the value of y2, in another word, the missing value in Y2 depends on the missing values themselves.

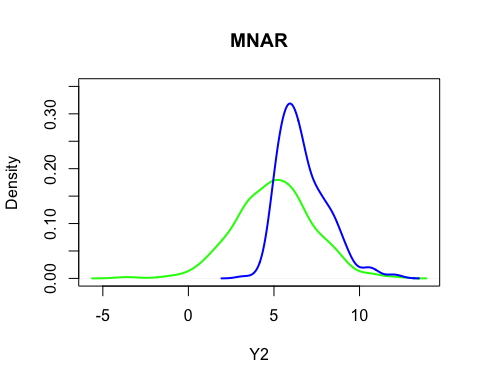
set.seed(111)  
Y2\_MANR\_obs = df$Y2[2\*(y2-5)+z3>=0]  
mean(Y2\_MANR\_obs)

## [1] 6.700805

sd(Y2\_MANR\_obs)

## [1] 1.442649

#Plot the densities of complete dataset and observed dataset  
  
plot(density(df$Y2), lwd = 2, col = "green", xlab = "Y2", main = "MNAR", ylim = c(0,0.35))  
lines(density(Y2\_MANR\_obs), lwd = 2, col = "blue")  
legend(20, 0.35, legend = c("Complete data","Observed data"), col = c("green","blue"), lty = c(1,1), lwd = c(2,2), bty = "n")

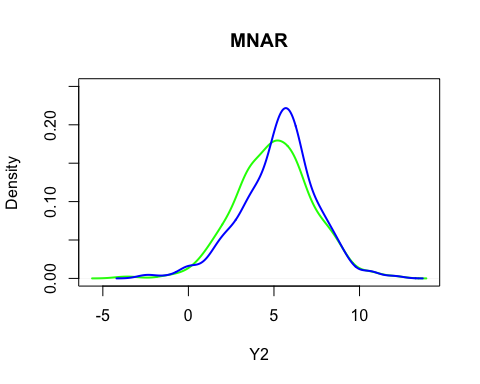


From the plot we can see that the distributions of observed data and complete data are more different under data generated according to MANR than that according to MAR.

set.seed(111)  
ind = which(2\*(y2-5)+z3<0)  
y2\_d = y2  
y2\_d[ind] = NA  
df\_2 = data.frame(  
 Y1 = y1,  
 Y2 = y2\_d  
)  
fity = lm(Y2 ~ Y1, data = df\_2)  
summary(fity)

##   
## Call:  
## lm(formula = Y2 ~ Y1, data = df\_2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.7136 -0.5605 0.0278 0.5832 2.3313   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.96671 0.13700 28.95 <2e-16 \*\*\*  
## Y1 1.64407 0.07584 21.68 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8562 on 254 degrees of freedom  
## (244 observations deleted due to missingness)  
## Multiple R-squared: 0.6492, Adjusted R-squared: 0.6478   
## F-statistic: 470 on 1 and 254 DF, p-value: < 2.2e-16

set.seed(111)  
predy2\_d = predict(fity,newdata = df\_2) + rnorm(n,0,sigma(fity))  
y2sri\_d = ifelse(is.na(df\_2$Y2),predy2\_d,df\_2$Y2)  
plot(density(df$Y2), lwd = 2, col = "green", xlab = "Y2", main = "MNAR", ylim = c(0,0.25))  
lines(density(y2sri\_d), lwd = 2, col = "blue")  
legend(20, 0.25, legend = c("Complete data","Observed data"), col = c("green","blue"), lty = c(1,1), lwd = c(2,2), bty = "n")



From the plot we can see that the distribution of oberved data after imputation is more close to the distribution of originally completed data, even though not similar to. So it is make sense to use stochastic regression to deal with missing values under MANR mechanism.

## Question 4

Using complete case analysis, we get the mean and standard error of recovery time is 19.27273, 2.603013.

load("databp.Rdata")  
ind = which(is.na(databp$recovtime) == FALSE) #indices of subjects of observed recovtime  
mean\_cca\_recovtime = mean(databp$recovtime, na.rm = TRUE)  
se\_cca\_recovtime = sd(databp$recovtime, na.rm = TRUE)/sqrt(length(ind))  
mean\_cca\_recovtime

## [1] 19.27273

se\_cca\_recovtime

## [1] 2.603013

Further, asked by question we obtain the correlations between recovery time and the dose is 0.2391256, and the correlations between recovery time and blood pressure is -0.01952862.

cor(databp$recovtime,databp$logdose,use = "complete")

## [1] 0.2391256

cor(databp$recovtime,databp$bloodp,use = "complete")

## [1] -0.01952862

Under mean imputation the mean of recovery time does not change which is 19.27273, but the standard error would decrease to 2.284135.

recovtime\_mi = ifelse(is.na(databp$recovtime),mean\_cca\_recovtime,databp$recovtime)  
n = length(recovtime\_mi)  
mean\_mi\_recovtime = mean(recovtime\_mi)  
se\_mi\_recovtime = sd(recovtime\_mi)/sqrt(n)  
mean\_mi\_recovtime

## [1] 19.27273

se\_mi\_recovtime

## [1] 2.284135

The correlation between recovery time and dose is 0.21506121, and the correlation between recovery time and blood pressure is -0.01934126.The correlations drops when doing mean imputation than complete data analysis.

cor(recovtime\_mi,databp$logdose)

## [1] 0.2150612

cor(recovtime\_mi,databp$bloodp)

## [1] -0.01934126

To do mean regression imputation, first we fit a regression model on the complete cases, using the recovery time as the response variable and using dose and blood pressure as the covariates:

set.seed(666)  
fit = lm(recovtime ~ logdose + bloodp, data = databp)  
summary(fit)

##   
## Call:  
## lm(formula = recovtime ~ logdose + bloodp, data = databp)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.768 -10.250 -0.770 3.546 37.394   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 15.2159 19.8203 0.768 0.452  
## logdose 11.4290 8.4178 1.358 0.190  
## bloodp -0.2769 0.3411 -0.812 0.427  
##   
## Residual standard error: 12.25 on 19 degrees of freedom  
## (3 observations deleted due to missingness)  
## Multiple R-squared: 0.08879, Adjusted R-squared: -0.007129   
## F-statistic: 0.9257 on 2 and 19 DF, p-value: 0.4134

From the summary report of fit, we obtain and and . Then we compute the predictions using function . Then we can get the mean of recovery time at 19.44428 and associated standard error at 2.312845 using regression imputation case.

set.seed(666)  
predicted\_recovtime = predict(fit,newdata = databp)  
recovtime\_ri = ifelse(is.na(databp$recovtime), predicted\_recovtime, databp$recovtime)  
mean\_recovtime\_ri = mean(recovtime\_ri)  
se\_recovtime\_ri = sd(recovtime\_ri)/sqrt(length(recovtime\_ri))  
mean\_recovtime\_ri

## [1] 19.44428

se\_recovtime\_ri

## [1] 2.312845

And the correlation between recovery time and dose is 0.2801835, correlation between recovery time and blood pressure is -0.0111364.

cor(recovtime\_ri,databp$logdose)

## [1] 0.2801835

cor(recovtime\_ri,databp$bloodp)

## [1] -0.0111364

For stochastic regression imputation, we add noise to the prediction obtainted using the regression imputation method, The mean of recovery time at 19.83789 and associated standard error at 2.440745 using stochastic regression imputaion case.

set.seed(666)  
predicted\_recovtime\_sri = predict(fit, newdata = databp) + rnorm(nrow(databp), 0, sigma(fit))  
recovtime\_sri = ifelse(is.na(databp$recovtime), predicted\_recovtime\_sri,databp$recovtime)  
mean\_recovtime\_sri = mean(recovtime\_sri)  
se\_recovtime\_sri = sd(recovtime\_sri)/sqrt(length(recovtime\_sri))  
mean\_recovtime\_sri

## [1] 19.83789

se\_recovtime\_sri

## [1] 2.440745

And we obtain the correlation between recovery time and dose is 0.08690395, also the correlation between recovery time and blood pressure is -0.01827952.

cor(recovtime\_sri,databp$logdose)

## [1] 0.08690395

cor(recovtime\_sri,databp$bloodp)

## [1] -0.01827952

As for stochastic regression imputation here we need to check the the assumption of linearity and of constant variance, also we need to check the assumption of normality of the noise value.

indrecovtime4 = which(is.na(databp$recovtime) == FALSE & databp$logdose >1.4 & databp$logdose <1.6 & databp$bloodp >50 &databp$bloodp <70)  
indrecovtime4

## [1] 25

donor4 = indrecovtime4 #in this case only one individual with similar value   
donor4

## [1] 25

indrecovtime10 = which(is.na(databp$recovtime) == FALSE & databp$logdose >2.2 & databp$logdose <2.5 & databp$bloodp >60 &databp$bloodp <80)  
indrecovtime10

## [1] 1 8 20

donor10 = sample(indrecovtime10,1,replace = TRUE)  
donor10

## [1] 20

indrecovtime22 = which(is.na(databp$recovtime) == FALSE & databp$logdose >2.5 & databp$bloodp >60 )  
indrecovtime22

## [1] 7 17 19

donor22 = sample(indrecovtime22,1,replace = TRUE)  
donor22

## [1] 17

We consequently create a new recovery time variable with imputed values(the predicted recovery time from regression model) from the hot dec procedure. Here we get the mean of recovery time is 19.54081, and associated standard error is 2.30679.

set.seed(666)  
recovhd = c(databp$recovtime[is.na(databp$recovtime) == FALSE], predicted\_recovtime[donor4], predicted\_recovtime[donor10], predicted\_recovtime[donor22])  
mhd = mean(recovhd)  
sehd = sd(recovhd)/sqrt(length(recovhd))  
mhd;sehd

## [1] 19.54275

## [1] 2.306897

And we obtain the correlation between recovery time and dose is -0.2509738, also the correlation between recovery time and blood pressure is -0.2794184.

cor(recovhd,databp$logdose)

## [1] -0.2510463

cor(recovhd,databp$bloodp)

## [1] -0.279549

The predictive mean imputation generally does not attenuate the variability of the imputed data as stochastic regression imputation. However, predictive mean imputation can produce substantially biased estimates of correlations, which we can see in answer(d); Besides, it also underestimates standard errors.