Sensitivity analysis to missing data: Sentence Length

S Geneletti and J Pina-Sanchez

Sentence length

Here we deal with `Sentence_Length` which is quite robust to the a mechanism where the data are MNAR. To simplify things, we remove `Custody` from the data. There is a companion file which looks at `Custody` only. Let's load the data and do some housekeeping.

```
ReMF.dat<- read.csv("ReMF_original.csv",header=T,stringsAsFactors = T)

#head(ReMF.dat)

ReMF.dat <- ReMF.dat %>% mutate(Prev_Convictions =
          fct_relevel(Prev_Convictions, c("None","1 to 3","4 to 9","10 or more")))
#some house-keeping.

#Only those with sentence length available
ReMF.dat <- subset(ReMF.dat,Custody==1)
ReMF.dat <- dplyr::select(ReMF.dat,-c(Custody))

#centre age
ReMF.dat <- ReMF.dat %>%
  mutate(Age_Cont = Age_Cont - mean(Age_Cont))
```

The benchmark coefficient of Ethnicity

We now obtain the coefficients for the baseline regression that we use as benchmarks for the missing data analyses. We focus on the coefficient of `Ethnicity` in this tutorial but it is worth checking that the coefficients of other predictors make sense

```
Sen_Len.benchmark <- lm(Sentence_Length~.,data=ReMF.dat)

#summary(Sen_Len.benchmark)
```

We will be comparing point estimates and confidence intervals for many models, so to expedite the process, I've created a function to do this quickly called `ci.tab`

```
ci.tab<-function(vec,Miss=1,SL=0,Ethn=0,Inter=0){
  #default values are miss=1, outcome=1 is custody,
  #Cust=Ethn=Inter=0 means there is no parameter for the missingness model
 ret.tab <- c(vec[1]-2*vec[2],vec[1],
      vec[1]+2*vec[2])
 ret.tab <-c(ret.tab,c(Miss, SL, Ethn, Inter))
 ret.tab <- as.data.frame(t(ret.tab))
  colnames(ret.tab) = c("Lower 95% CI", "coefficient","Upper 95% CI",
                        "Missingness","SL","Ethn","Inter")
  ret.tab
```

```
}
#Miss=1, no missing
#Miss=2, MCAR
#Miss=3, MAR
#Miss=4, MNAR
```

Let's look at the benchmark values for the coefficient of `Ethnicity`:

```
Ethn.SL.benchmark <- summary(Sen_Len.benchmark)$coefficients[4,1:2]
#the 4th row corresponds to ethnicity and the 1,2 columns are the estimate and its sd
Ethn.SL.benchmark.tab <- ci.tab(Ethn.SL.benchmark)
#Ethn.SL.benchmark.tab
```

MCAR

MCAR is random missingness. We will start with 30% missingness. Depending on how the data are missing, this can be a high or a low percentage of missingess.

```
#parameter that governs the percentage of missingness
per.miss <- 0.3

#generate the missing data indicator
MCAR <- rbinom(n=nrow(ReMF.dat),size=1,prob=c(per.miss))
#sum(MCAR)/nrow(ReMF.dat) #can use to check that the % of missing is approximately correct

#now add NAs to create a missingness pattern
MCAR_ReMF.dat <- ReMF.dat %>%
  mutate(Ethnicity=ifelse(MCAR==0,Ethnicity, NA))
```

The same MCAR can be used for both `Custody` and `Sentence_Length`. Let's see what the outcome

```
SL.MCAR <- summary(lm(Sentence_Length~.,data=MCAR_ReMF.dat))
Ethn.SL.MCAR <-SL.MCAR$coefficients[4,1:2]

Ethn.SL.MCAR.tab <- ci.tab(Ethn.SL.MCAR, Miss=2)

rbind(Ethn.SL.benchmark.tab,Ethn.SL.MCAR.tab)
```

```
##   Lower 95% CI coefficient Upper 95% CI Missingness SL Ethn Inter
## 1    0.7790444    2.059765     3.340485           1  0    0     0
## 2    0.4330785    1.955943     3.478807           2  0    0     0
```

We see that the coefficients are very similar. The width of the confidence interval is larger for the MCAR case, but otherwise, they are similar.

MAR

Now we can try MAR by making the probability of a missing data point depend on the covariates in the model. Let's make it depend on `Age_cont` and `MF_Remorse`. It is relatively straight-forward to extend this to all the covariates, but care must be taken when deciding the coefficients of each covariate in the missingness model. This is in fact the trickiest part of the process – although more so for the MNAR situation.

We'll set it up so that the older you are, the less likely the `Ethnicity` is missing. In other words the value of the missingness indicator is more likely to be 0 if you are older. Also, those who are remorseful are those who are less likely to have `Ethnicity` missing. In other words, the value of the missingness indicator is more likely to be 0 if you are `MF_Remorse`=1. This means that `Age_cont` has a negative coefficient and that `MF_Remorse` also has a negative coefficient.

Sensible values for logistic regression parameters are between -2 to 2, larger effects are rare. As `Age_cont` is continuous, we allocate a small positive effect to it. We start with -0.02. This means that the odds of having a missing `Ethnicity` decrease by 2% for every additional year. `MF_Remorse` is binary and we want a relatively strong association so we start with -0.20. This corresponds to an decrease in 18% in the odds of being missing for those who are remorseful.

We also need an intercept value which represents the log(odds) that a person of `Age_cont=0` and who does not exhibit remorse will have a missing value. Let's use the baseline MAR case for that `per.miss=0.2`. We want the overall probability of missing to be approximately 0.2 so we choose an intercept of 0.9 by trial and error to obtain 30% missing. You should check this is a believable number in your context.

```
MM.Age_Cont <- -0.02
MM.MF_Remorse <- -0.2

#vector of probabilities that are associated with Age and remorse
pMAR <- invlogit(-0.8 + MM.Age_Cont*ReMF.dat$Age_Cont + MM.MF_Remorse*ReMF.dat$MF_Remorse)

#missing data indicator
MAR <- rbinom(n = nrow(ReMF.dat), size = 1, prob = pMAR)
sum(MAR)/nrow(ReMF.dat)
```

```
## [1] 0.3021117
```

```
#check that this is approx 0.2

MAR_ReMF.dat <- ReMF.dat %>%
  mutate(Ethnicity=ifelse(MAR==0,Ethnicity,NA))
```

Now that we have missing data in ethnicity (about 20%), let's see what, if any effect this has on the estimates of the coefficient of `Ethnicity`.

```
Sen_Len.MAR <- summary(lm(Sentence_Length~.,data=MAR_ReMF.dat))
Ethn.SL.MAR <- Sen_Len.MAR$coefficients[4,1:2]

Ethn.SL.MAR.tab <- ci.tab(Ethn.SL.MAR, Miss=3)


rbind(Ethn.SL.benchmark.tab,Ethn.SL.MCAR.tab, Ethn.SL.MAR.tab)
```

```
##   Lower 95% CI coefficient Upper 95% CI Missingness SL Ethn Inter
## 1    0.7790444    2.059765     3.340485           1  0    0     0
## 2    0.4330785    1.955943     3.478807           2  0    0     0
## 3    0.5718262    2.107800     3.643773           3  0    0     0
```

Most of the time, we see that the width of the confidence interval increases again but that the value of the coefficient remains similar. However, sometimes, the coefficient decreases and becomes non-significant. This is always a possibility when randomly removing values.

MNAR

We can start off with a coefficient of 0.05 for `Ethnicity`. We associate a very small positive coefficient to `Sentence_Length` which implies that those who end up getting a longer custodial sentence are more likely to have a missing `Ethnicity`. We initially choose 0.005 which corresponds to an odds ratio of 1.005. We keep these values constant for the simulation study below, but these can and should be modified.

Finally we need to choose values for the interaction term between `Ethnicity` and `Sentence_Length`. We make this small again to reflect the continuous nature of `Sentence_Length`: 0.005. A positive coefficient

indicates that those of an ethnic minority who have longer sentences are even more likely to have missing `Ethnicity`

The percentage of missingness is *very* sensitive to the values of the coefficients.

```
MM.Age_Cont <- -0.01
MM.MF_Remorse <- -0.1
MM.Ethn <- 0.05
MM.SL <- 0.0025
MM.Ethn.SL <- 0.0025
MNAR_intercept <- -1.5

pMNAR <- with(ReMF.dat,
              invlogit(MNAR_intercept + MM.Age_Cont*Age_Cont + MM.MF_Remorse*MF_Remorse +
                    MM.SL*Sentence_Length + MM.Ethn*Ethnicity +
                      MM.Ethn.SL*Ethnicity*Sentence_Length))

#missing data indicator
MNAR <- rbinom(n = nrow(ReMF.dat), size = 1, prob = pMNAR)
sum(MNAR, na.rm=T)/nrow(subset(ReMF.dat))
```

```
## [1] 0.2716025
```

```
MNAR_ReMF.dat <- ReMF.dat %>%
  mutate(Ethnicity=ifelse(MNAR==0,Ethnicity,NA))
```

Now that the data are generated, let's look at the results.

```
Sen_Len.MNAR <- summary(lm(Sentence_Length~.,data=MNAR_ReMF.dat))
Ethn.SL.MNAR <- Sen_Len.MNAR$coefficients[4,1:2]

Ethn.SL.MNAR.tab <- ci.tab(Ethn.SL.MNAR,Miss=4, SL = MM.SL, Ethn = MM.Ethn, Inter = MM.Ethn.SL)

rbind(Ethn.SL.benchmark.tab,Ethn.SL.MCAR.tab,
      Ethn.SL.MAR.tab,Ethn.SL.MNAR.tab)
```

```
##   Lower 95% CI coefficient Upper 95% CI Missingness     SL Ethn  Inter
## 1    0.7790444    2.059765     3.340485           1 0.0000 0.00 0.0000
## 2    0.4330785    1.955943     3.478807           2 0.0000 0.00 0.0000
## 3    0.5718262    2.107800     3.643773           3 0.0000 0.00 0.0000
## 4    0.3957736    1.909647     3.423521           4 0.0025 0.05 0.0025
```

For `Sentence_Length` the effects are hard to see. This is partly because the estimate of `Ethnicity` in this regression is quite variable and partly because missing data increases it's variability further. Very high coefficients for `Sentence_Length` and it's interaction lead to very high missingness which also substantially changes the parameters.

Table of all results

```
Sen_Len.all.results <- rbind(Ethn.SL.benchmark.tab,Ethn.SL.MCAR.tab,Ethn.SL.MAR.tab,Ethn.SL.MNAR.tab)
rownames(Sen_Len.all.results) <- c("True","MCAR","MAR","MNAR")
Sen_Len.all.results
```

```
##      Lower 95% CI coefficient Upper 95% CI Missingness     SL Ethn  Inter
## True    0.7790444    2.059765     3.340485           1 0.0000 0.00 0.0000
## MCAR    0.4330785    1.955943     3.478807           2 0.0000 0.00 0.0000
## MAR     0.5718262    2.107800     3.643773           3 0.0000 0.00 0.0000
```

```
## MNAR    0.3957736    1.909647     3.423521           4 0.0025 0.05 0.0025
```

Some useful plots

To investigate the impact of the `MM.Ethn.SL` value, let's run the MNAR model with a range of "plausible" values. We will assume that the value is always positive which means being of an ethnic minority is always more likely to increase missingness.

```
MM.Ethn.SL.Vec <- seq(0.00,0.01,by=0.001)

#to store the percentage missing so this remains "plausible"
store.per.miss <- rep(NA, length(MM.Ethn.SL.Vec))

#to store the estimate of ethnicity
store.coeff.ethn <- c()

for(i in 1:length(MM.Ethn.SL.Vec)){
MM.Ethn.SL <- MM.Ethn.SL.Vec[i]
pMNAR <- with(ReMF.dat,
              invlogit(MNAR_intercept + MM.Age_Cont*Age_Cont + MM.MF_Remorse*MF_Remorse +
                    MM.SL*Sentence_Length + MM.Ethn*Ethnicity +
                      MM.Ethn.SL*Ethnicity*Sentence_Length))

#missing data indicator
MNAR <- rbinom(n = nrow(ReMF.dat), size = 1, prob = pMNAR)
store.per.miss[i]<-sum(MNAR, na.rm=T)/nrow(subset(ReMF.dat))

MNAR_ReMF.dat <- ReMF.dat %>%
  mutate(Ethnicity=ifelse(MNAR==0,Ethnicity,NA))

Sen_Len.MNAR <- summary(lm(Sentence_Length~.,data=MNAR_ReMF.dat))
Ethn.SL.MNAR <- Sen_Len.MNAR$coefficients[4,1:2]

store.coeff.ethn <- rbind(store.coeff.ethn,ci.tab(Ethn.SL.MNAR, Miss=4,
SL = MM.SL, Ethn = MM.Ethn, Inter = MM.Ethn.SL))

}
```

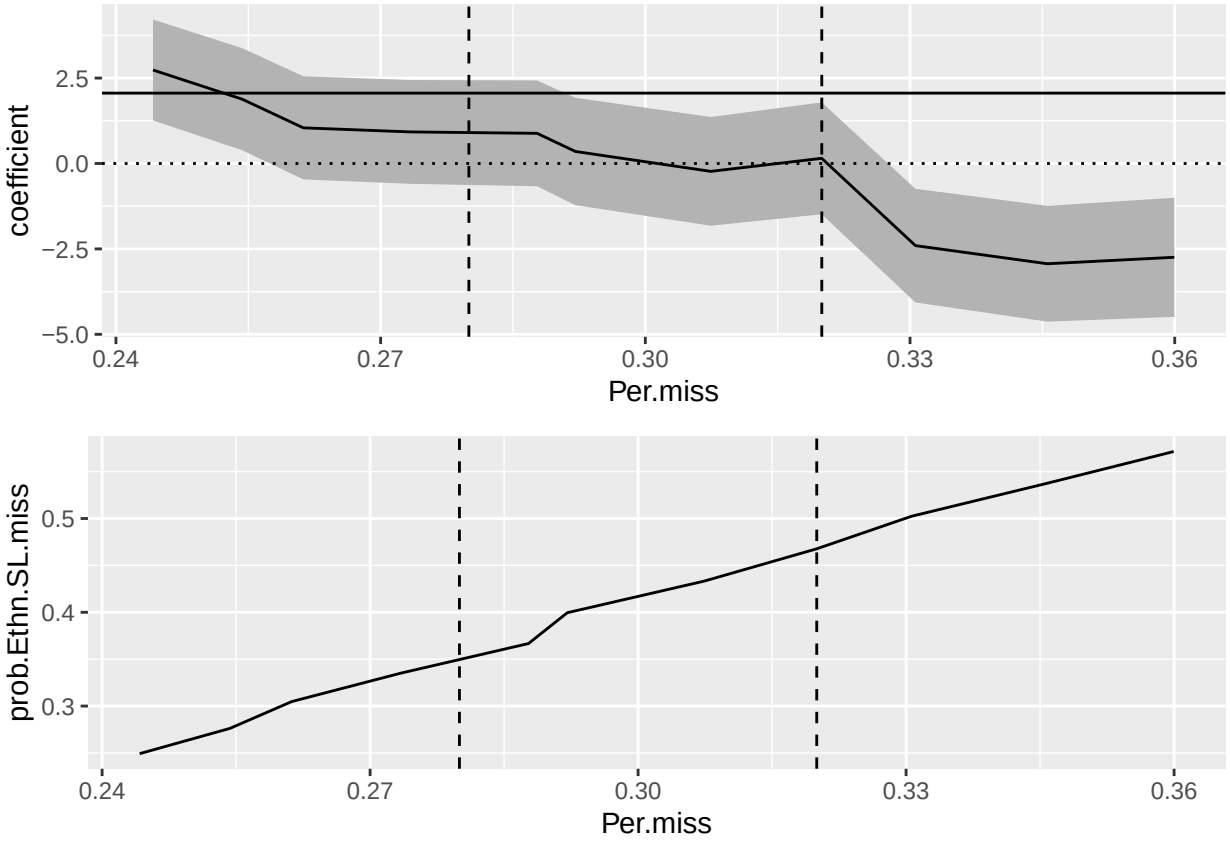
Now that we've produced all the coefficients, let's look at the values plotted. Typically, you would run this multiple times for the same values and plot the means. Ideally we want to monitor 3 things.

1. The coefficient of `Ethnicity` – does it stay significant?
2. The overall proportion of missingess – is this close to what we observe?
3. The proportion missing for the median (and other quantiles) for those who are non-white – is this sensible?

We produce 3 plots that tell us about these relationships

```
Quartiles.SL <- with(ReMF.dat, quantile(Sentence_Length)[2:4]) #25th, median and 75%

plot.coeff.miss <- ggplot(data=data.frame(Per.miss=store.per.miss,store.coeff.ethn),
                          aes(x=Per.miss)) +
                  geom_ribbon(aes(ymin=Lower.95..CI,
                                  ymax=Upper.95..CI), fill = "grey70")+
  geom_line(aes(y=coefficient)) +
  geom_hline(yintercept=Ethn.SL.benchmark.tab$coefficient) +
```

```
  geom_hline(yintercept=0, linetype="dotted") +
  geom_vline(xintercept=c(0.28,0.32),linetype="dashed")


prob.Ethn.SL.miss = with(store.coeff.ethn,
                         invlogit(MNAR_intercept+MM.SL*Quartiles.SL[2]+
                                    MM.Ethn+Inter*Quartiles.SL[2]))


plot.per.miss <- ggplot(data=data.frame(prob.Ethn.SL.miss=prob.Ethn.SL.miss,
                                        Per.miss=store.per.miss),
                        aes(y=prob.Ethn.SL.miss,x=Per.miss))+geom_line() +
                        geom_vline(xintercept=c(0.28,0.32),linetype="dashed")

grid.arrange(plot.coeff.miss, plot.per.miss, nrow = 2)
```



These results contrast with those of the example with custodial sentences as the coefficient of `Sentence_Length` is more sensitive to missing not at random mechanisms. The top figure plots the overall percentage missing against the coefficient. The solid horizontal line is the true value. In the range of missingness we are interested in, the estimate of the coefficient of `Ethnicity` is non-significant. In the bottom figure, we plot the overall percentage missing against the percentage missing for those who have median sentence length and are non-white. These are quite high, between 30-55%. Given the level of overall missingness at 30% is this plausible? If not, then we need to tweak parameters in the MNAR model.

Similar plots can be produced for the 25th and 75th quantiles using the code above.