e-Research Methods Festival 2023
Sensitivity analysis to missing data
Custody

S Geneletti and J Pina-Sanchez

Format of this tutorial

- You should have access to the data and the R code (as well as the Rmarkdown) associated with these slides.
- These can be found here https://github.com/sggi-76/Missing_data_sensitivity_tutorial
- If you haven't done so already, please download the data and the R code.
- I would like you to run the code alongside my slides.
- I will also have RStudio open so we can discuss any results.
- You are free to annotate or modify the code as you like.
- At the end you can change some values and run your own version of this sensitivity analysis.

Missing data Sensitivity analysis

- The aim of this tutorial is to show you how to use your complete case data to explore how sensitive your results are to missing data.
- Typically there are two outcomes of this sensitivity analysis:
  1. The first is that the results are *robust* to *plausible* missing data mechanisms and so we can trust the results of the complete case analysis.
  2. The second is that the results are *sensitive* to *plausible* missing data mechanisms and so we must view our results with scepticism.
- We draw on the paper "Now You See It, Now You Don't: A Simulation and Illustration of the Importance of Treating Incomplete Data in Estimating Race Effects in Sentencing" by B Stockton et al, 2023, Journal of Quantitative Criminology.

What we'll do

- In this tutorial, we will be looking at how to create the missing data mechanisms for a binary outcome.
- A separate file "ReMF_Missing_SL.pdf", available on the Github page for this course covers the case of the continuous outcome.
- We will have a quick look at the results for this outcome at the end of the tutorial but we won't cover it in depth.
- We will explore all three mechanisms covered in the Introduction, missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR).

Main steps

1. Assume your complete case data are the full data.
2. Use the complete case data to estimate your parameters of interest. These are typically coefficients (or interpretable transformations of coefficients) of covariates you are interested in. These are your benchmark.
3. Decide what missing data mechanisms are plausible in your data and re-create them in the analysis. This part is the one that takes the most thought and will need to be justified.
4. Use the missing data mechanisms to remove observations in your data.
5. Estimate the parameters again and compare them to the benchmark.

Problems

This process is not perfect.

- The complete case data could be biased and so your benchmark isn't the true estimate.
- But.. complete case data are the best data you have (as they most closely resemble the full data).
- Any missing data mechanism you attempt needs to be justified.
- Typically, you will attempt many plausible mechanisms and compare the results under these mechanisms to the benchmark values.
- The actual missing data patterns in your data can be used as a basis for your sensitivity analysis.
- There are a *lot* of moving parts

Data

- Based on the Sentencing Council data and specifically on the Assault offence.
- We have made some changes which include:
  1. changing some banded variables to continuous variables so that we have both categorical and continuous variables in the data
  2. adding `Ethnicity` which we use for our sensitivity analysis
  3. Removing some variables for simplicity

Table of variables

Variable	Definition	Type
Custody	Immediate custody	Binary
Sentence_Length	Length of sentence in days	Numeric
Age_cont	Age in years	Numeric
Ethnicity	Ethnicity, if 1 then non-white	Binary
Gender	Gender	Binary
Offence	6 sub-types of the Assault offence	Categorical
Prev_Convictions	Grouped into None, 1 to 3, 4 to 9 and 10 or more	Categorical
GP_First	The offender entered a guilty plea at the first occasion.	
AF_Injury	Aggravating factor: There was injury	Binary
AF_Vuln_Victim	Aggravating factor: The victim was vulnerable	Binary
AF_Rep_Assault	Aggravating factor: The victim was repeatedly assaulted	Binary
AF_Weapon	Aggravating factor: A weapon was used	Binary
MF_Remorse	Mitigating factor: The offender showed remorse	Binary
MF_Good_Char	Mitigating factor: The offender has good character	Binary

Custody and Ethnicity

- Our binary outcome is `Custody` which is 1 if an offender is incarcerated and 0 if they receive a different sentence.
- We are interested in the effect of `Ethnicity` on incarceration
- We know that people who are from ethnic minorities in the UK are less likely to report their ethnicity and therefore they will have missing entries.
- However, we also suspect that people from ethnic minorities are more likely to be incarcerated.
- It is therefore possible that the coefficients of ethnicity in our regressions are biased because ethnic minorities are under-represented in the complete case data.

Data prep and housekeeping

Let's load the data and look at the top 6 rows:

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

Some housekeeping:

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

# remove Sentence_Length
ReMF.dat <- dplyr::select(ReMF.dat,-c(Sentence_Length))
```

Let's centre the age so we can interpret the intercept more easily in later regressions.

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

The benchmark coefficient of Ethnicity

- First we obtain coefficients (as odds ratios) for the benchmark regression that we use as comparator 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.
- It is possible to investigate the sensitivity of multiple coefficients as well.

```
Cust.benchmark <- glm(Custody~.,data=ReMF.dat)

#summary(Cust.benchmark)
```

The equation looks like this $p(Custody = 1) = invlogit(\beta X)$ where β is the vector of coefficients and X is the design matrix

ci.tab

To compare values easily I've created a function that creates a small table to display results on the odds ratio scale called `ci.tab`

```
ci.tab<-function(vec,Miss=1,Cust=0,Ethn=0,Inter=0){
  #default values are miss=1,
  #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 <- exp(ret.tab)
 #If you want output on the log(OR) scale, comment the line above
 ret.tab <-c(ret.tab,c(Miss, Cust, Ethn, Inter))
 ret.tab <- as.data.frame(t(ret.tab))
  colnames(ret.tab) = c("Lower 95% CI", "Est","Upper 95% CI",
                        "Miss","Cust","Ethn","Inter")
  ret.tab
}
#Miss=1, no missing
#Miss=2, MCAR
#Miss=3, MAR
#Miss=4, MNAR
```

The benchmark coefficient of Ethnicity

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

Ethn.Cust.benchmark.tab <- ci.tab(Ethn.Cust.benchmark)
Ethn.Cust.benchmark.tab
```

```
##   Lower 95% CI      Est Upper 95% CI Miss Cust Ethn Inter
## 1     1.164303 1.175917     1.187648    1    0    0     0
```

As I have not `set(seed=xx)` when you run your R code, it won't be exactly the same as mine.

Missingness mechanisms

- There are three types of missingness mechanisms.
- Missing completely at random (MCAR) is equivalent to randomly removing data points.
  - The consequence is simply a less precise estimate of the benchmark coefficient.
- Missing at random (MAR) means that data points are removed randomly within the strata of observed covariates.
  - If we adjust for the correct covariates in a regression, the only impact this form of missingness has is on the precision of the estimates.
- Missing not at random (MNAR) means that the missingess is associated with the outcome (via unobserved confounders)
  - It is typically not possible to adjust for it by including only the observed confounders in the model.
  - A sensitivity analysis can provide a way to support the results of the complete case analysis or it can indicate that these are not reliable.

Missingness mechanisms on Ethnicity

- We will introduce missingness only in `Ethnicity`.
  1. It is a variable that is often missing in real data
  2. The way it is missing is typically not at random
- For crime data, there is likely to be structural racism
- Also ethnicity is often self-reported and white offenders are more likely to declare their ethnicity than ethnic minority offenders.

The missingness indicator

- In order to "implement" a missing data mechanism we create a binary vector of the same length as the data, such that:
  1. if its value is 1 then the corresponding value of `Ethnicity` is *missing* and
  2. if its value is 0, then the corresponding value of `Ethnicity` is *observed*.
- The pattern of the missingness indicator for MAR and MNAR is modelled using logistic regressions which allow us to include what factors we think govern the missingess.
- The easiest way to generate a binary variable is using a binomial distribution with the probability of missingess coming from the logistic regression for MAR and MNAR.

MCAR

- We will aim for 20% missingness.
- In a real analysis you should use the missingness actually present in your data.
- So if x% are missing in your data you should aim for x% missingness in all your sensitivity analyses.

```
#parameter that governs the percentage of missingness
per.miss <- 0.2

#generate the missing data indicator
MCAR <- rbinom(n=nrow(ReMF.dat),size=1,prob=c(per.miss))
#sum(MCAR)/nrow(ReMF.dat)
#should 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==1,NA,Ethnicity))
```

Results under MCAR

What is the effect of MCAR on the results?

```
Cust.MCAR <- summary(glm(Custody~.,data=MCAR_ReMF.dat))
Ethn.Cust.MCAR <- Cust.MCAR$coefficients[4,1:2]

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

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

```
##   Lower 95% CI      Est Upper 95% CI Miss Cust Ethn Inter
## 1     1.164303 1.175917     1.187648    1    0    0     0
## 2     1.163818 1.176825     1.189976    2    0    0     0
```

I provide the odds ratio and 95% confidence intervals so that you can see that there is substantial overlap in the confidence intervals.

Results under MCAR

- Most of you will see that the odds ratios are very close.
- Your results will not be the same as mine as you will have generated a different missingness indicator.
- In some cases, by chance, you will have significantly different results.
- In a complete sensitivity analysis, you run each step a number of times and make sure the results are stable
- I do not show you how to do this here, however it is a very important step.
- In principle, if you ran the code above only once, you could by chance get a result that is significantly different. It is only by running the code many times (e.g. 100) that you can be confident of your results.
- This is true for all mechanisms we consider.

Important point

- Before moving onto MAR, it is worth pointing out that in this type of sensitivity analysis there are a *lot* of moving parts.
- The % missing, the values of parameters that govern the missingness mechanisms (MAR, MNAR) such as the intercept and the coefficients of `Ethnicity`
- This is not a quick fix, rather, it is way of gaining deeper understanding of your data and it may raise more questions than it answers
- Having said that, it is worth constraining some things early on (e.g. the % missing and some coefficients), justifying these values and then varying fewer parameters

MAR

- In MAR we make the probability of a missing data point depend on some covariates in the model.
- MAR stands for missing at random but more accurately it is missing at random within strata of observed covariates
- Let's make it depend on `Age_cont` and `MF_Remorse`.
- I choose these because they are are realistically associated with missingness and they are continuous and binary respectively
- In principle you can use all the covariates in the model.
- 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.

MAR

- We assume that the older you are, the less likely the `Ethnicity` is missing.
- i.e. the value of the missingness indicator is more likely to be 0 if you are older.
- Those who are remorseful are those who are less likely to have `Ethnicity` missing.
- i.e. the value of the missingness indicator is more likely to be 0 if you are `MF_Remorse`=1.
- This means that *both* `Age_cont` and `MF_Remorse` have *negative* coefficients.

MAR

- 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.

MAR

- When we choose the *intercept*, there are two things to consider
  1. The first is that we want to overall missingness to reflect the missingness we have in our data
  2. The intercept has to be interpretable.
- The intercept tells us about baseline missingness.
- In our case, it is the log odds of missingness for someone with mean age and no remorse.
- By trial and error we get -1.3 corresponds to approximately 20% missing for the data we have and 21% missing for those who have mean age and no remorse

The missingness equation then looks like this:

$$p(R = 1) = invlogit(-1.3 - 0.02Age\_cont - 0.2MF\_Remorse)$$

where $R = 1$ means a particular value is missing

MAR Coding the missingness indicator

```
MM.Age_Cont <- -0.02
MM.MF_Remorse <- -0.20



#vector of probabilities that are associated with Age and remors
pMAR <- with(ReMF.dat, invlogit(-1.3 + MM.Age_Cont*Age_Cont +
                        MM.MF_Remorse*MF_Remorse))

#missing data indicator
MAR <- rbinom(n = nrow(ReMF.dat), size = 1, prob = pMAR)
#sum(MAR)/nrow(ReMF.dat)
#check that this is approx 0.2

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

MAR Results

What is the effect of MAR on the estimates of the coefficient of `Ethnicity`?

```
Cust.MAR <- summary(glm(Custody~.,data=MAR_ReMF.dat))
Ethn.Cust.MAR <- Cust.MAR$coefficients[4,1:2]

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

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

```
##   Lower 95% CI      Est Upper 95% CI Miss Cust Ethn Inter
## 1     1.164303 1.175917     1.187648    1    0    0     0
## 2     1.163818 1.176825     1.189976    2    0    0     0
## 3     1.161420 1.174421     1.187568    3    0    0     0
```

A different way of thinking about MAR

- It is possible to think about the coefficients in terms of probabilities
- This can be done in two ways:

1) We believe that the probability of being missing for those showing remorse is 0.2 and

- the probability of being missing for those not showing remorse is 0.6
- This corresponds to an odds ratio of $\frac{\frac{0.2}{1-0.2}}{\frac{0.6}{1-0.6}} = 0.17$
- which corresponds to log(OR) of -1.8 which would then be the value of the coefficient of `MF_Remorse` in the missingness logistic regression

A different way of thinking about MAR

2) The probability of being missing for those showing remorse is thought to be 0.2 and

- the probability of being missing for those not showing remorse is 50% more meaning it is 0.3
- This corresponds to an odds ratio of 0.37 and a log(OR) of -0.54
- If we look at the running example, it could correspond to a probability of incarceration for the remorseful of 0.2 and 0.235 for the non-remorseful.

MNAR

- What happens if we do not observe all the drivers of missingness as in MAR?
- What happens if the missingess is related to the outcome?
- For example, in many surveys, people are reluctant to disclose their incomes and sexual orientation etc.
- If the outcome of interest is related to the missing covariate e.g. voting behaviour or mental health, then we have missing not at random and our results may be biased.
- In our MNAR example, the missingess mechanism depends on the two covariates in the MAR case and *also* on the outcome and the `Ethnicity` directly.
- We govern the impact of the missingess via the interaction between `Ethnicity` and the outcome.
- The interaction allows us to vary the strength of the combined effect of `Ethnicity` and `Custody`.

MNAR Ethnicity

- We assume that being from an ethnic minority increases the chance of `Ethnicity` being missing
- i.e. the value of the missingness indicator is more likely to be 1 when `Ethnicity`=1).
- This is in line with observed data where people from ethnic minorities are more likely to refuse to report their ethnicity.
- This means that the coefficient of `Ethnicity` is positive.
- Let's start with 0.2 which corresponds to an increase in the odds of missing of 18%.

MNAR Custody

- We associate a small negative coefficient to `Custody`
- This implies that those who end up getting a custodial sentence are less likely to have a missing ethnicity.
- This reflects the idea that most people in custody are white and therefore we expect a small negative association (although we could debate this point and the results may well be different).
- We initially choose -0.05 which corresponds to an odds ratio of 0.95.
- We keep these values constant for the simulation study below.

MNAR Interaction

- We need to choose values for the interaction term between `Ethnicity` and `Custody`.
- For this example, we start with 0.9 to induce a strong association between `Ethnicity` and missingess.
- At the end of this tutorial, we run this for a number of different values of the interaction to see the overall impact.

```
MM.Ethn <- 0.2
MM.Cust <- -0.05
MM.Ethn.Cust <- 0.9
```

MNAR Intercept

- In the MNAR case, we choose an intercept of -1.55.

```
MNAR_intercept <- -1.55
```

- This means that an offender of mean age, no remorse, who is white and does not get custody has as 17% probability of being missing
- Is this a plausible value?
- Best thing is to check.

MNAR overivew

- We focus on those with no remorse and mean age.

```r
#Ethn=1, Cust=1
invlogit(MNAR_intercept+MM.Cust+MM.Ethn+MM.Ethn.Cust)
```

```
## [1] 0.3775407
```

```r
#Ethn=1, Cust=0
invlogit(MNAR_intercept+MM.Ethn)
```

```
## [1] 0.2058704
```

```r
#Ethn=0, Cust=1
invlogit(MNAR_intercept+MM.Cust)
```

```
## [1] 0.1679816
```

```r
#Ethn=0, Cust=0
invlogit(MNAR_intercept)
```

```
## [1] 0.1750863
```

MNAR overview

- Perhaps these values are not plausible.
- For instance we may feel that more than 38% of people from ethnic minorities in custody would be willing to divulge their ethnicity.
- You may have auxiliary data that can help with this.
- For example, you may have collected information about `Ethnicty` from a different source other than the self-reported values.
- This could help you choose the coefficient of `Ethnicity` in the missingness model

MNAR equation

$$p(R=1) = invlogit(-1.55 - 0.02 Age\_cont - 0.2 MF\_Remorse$$
$$+0.2 Ethnicity - 0.05 Custody + 0.9 Ethnicity \times Custody)$$

MNAR

- Recap: in this sensitivity analysis you vary the value of the interaction between `Ethnicty` and `Custody`
- And monitor the difference between the benchmark and the estimated coefficients of `Ethnicity` as well as the levels of missingness
- If a plausible value of interaction term coefficient is
  1. associated with unrealistic levels (%) of missingness or
  2. results in an implausible value of the coefficient of `Ethnicty` in the outcome equation,
- then you have to concede that your results are sensitive to missingness.
- If a plausible value of the interaction term coefficient is
  1. associated with the observed level (%) of missingbess
  2. results in plausible value of the coefficient of `Ethnicty` in the outcome equation,
- You can argue that the results are robust to missingness.

MNAR missingness indicator

Let's generate the missingness indicator and run the regression.

```
pMNAR <- with(ReMF.dat,
             invlogit(MNAR_intercept + MM.Age_Cont*Age_Cont +
                   MM.MF_Remorse*MF_Remorse +
                   MM.Cust*Custody +
                   MM.Ethn*Ethnicity +
                   MM.Ethn.Cust*Custody*Ethnicity))

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

```
## [1] 0.2046871
```

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

From the output we see that about 20% of the data are missing.

MNAR results

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

```
Cust.MNAR <- summary(glm(Custody~.,data=MNAR_ReMF.dat))
Ethn.Cust.MNAR <- Cust.MNAR$coefficients[4,1:2]

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

Ethn.Cust.MNAR.tab
```

```
##   Lower 95% CI      Est Upper 95% CI Miss  Cust Ethn Inter
## 1     1.114996 1.127942     1.141038    4 -0.05  0.2   0.9
```

MNAR results

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

```
##   Lower 95% CI      Est Upper 95% CI Miss  Cust Ethn Inter
## 1     1.164303 1.175917     1.187648    1  0.00  0.0   0.0
## 2     1.163818 1.176825     1.189976    2  0.00  0.0   0.0
## 3     1.161420 1.174421     1.187568    3  0.00  0.0   0.0
## 4     1.114996 1.127942     1.141038    4 -0.05  0.2   0.9
```

- We can see that although there is a difference in the odds ratio coefficient estimates, the direction and significance of the result is similar.
- Higher values of `MM.Ethn.Cust` move this interval further down showing that the more the missingness in `Ethnicity` is linked with `Custody`, the more we underestimate the effect of Ethnicity on `Custody`.
- In this case, we can see that the results are robust! Yay.

Table of all results

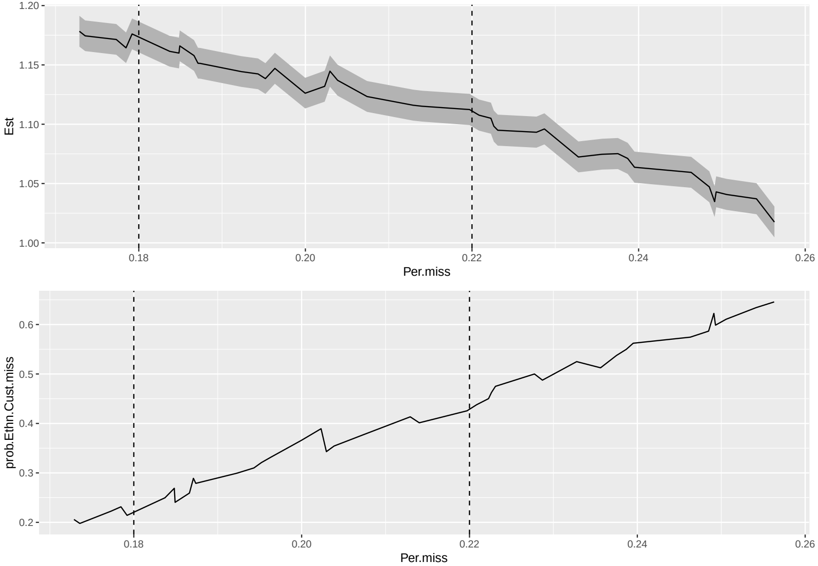
```
Cust.all.results <- rbind(Ethn.Cust.benchmark.tab,Ethn.Cust.MCAR.tab,
                          Ethn.Cust.MAR.tab,Ethn.Cust.MNAR.tab)
rownames(Cust.all.results) <- c("True","MCAR","MAR","MNAR")
Cust.all.results
```

```
##      Lower 95% CI      Est Upper 95% CI Miss  Cust Ethn Inter
## True     1.164303 1.175917     1.187648    1  0.00  0.0   0.0
## MCAR     1.163818 1.176825     1.189976    2  0.00  0.0   0.0
## MAR      1.161420 1.174421     1.187568    3  0.00  0.0   0.0
## MNAR     1.114996 1.127942     1.141038    4 -0.05  0.2   0.9
```

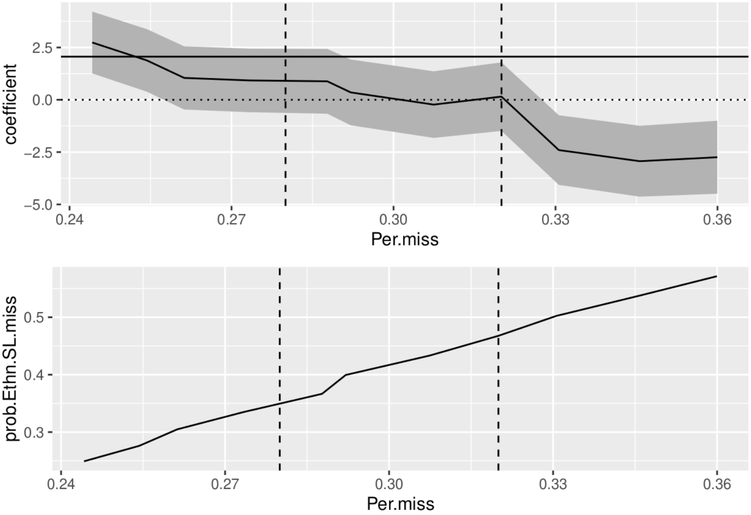
Some useful plots

- To investigate the impact of the `MM.Ethn.Cust` value, let's run the MNAR model for a range of "plausible" values, say between 0.1 and 2.
- We will assume that the value is always positive which means being of an ethnic minority is always more likely to increase missingness.

Some useful plots



Sentence Length

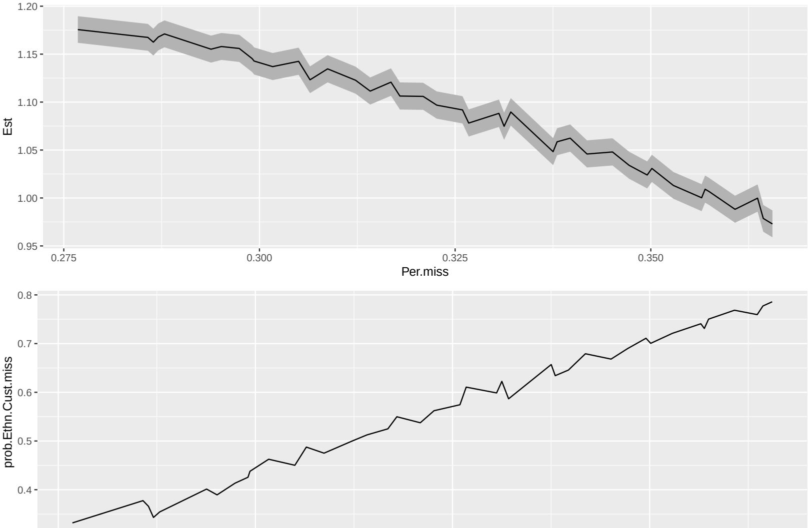


DIY

- Try different values of the various parameters.
- For example, how do things change if you play around with the coefficients of `Age_Cont` or `MF_Remorse`?
- Or if you change the parameters in the MNAR model?
- Can you keep the % missingness close to 20%?
- Try a different % missingness
- For this context I found the coefficient of ethnicity to be very robust
- The continuous example with Sentence length as the outcome is much more sensitive

Some values you could try (see the code)

- Try different values for some or all of your predictors.
- Even with % 30+ missingness, the value of the OR of `Ethnicity` is very stable



Results

```
##      Lower 95% CI      Est Upper 95% CI Miss  Cust Ethn Inter
## True     1.164303 1.175917     1.187648    1  0.00  0.0   0.0
## MCAR     1.163818 1.176825     1.189976    2  0.00  0.0   0.0
## MAR      1.161420 1.174421     1.187568    3  0.00  0.0   0.0
## MNAR     1.114996 1.127942     1.141038    4 -0.05  0.2   0.9
```