HW12

Animesh Sengupta

2023-05-24

3

a

MNAR - We can attribute the death and severe side effects of the participants due to some unobserved factors or variables. Hence their missingness can be concerning hence it is type of MNAR

b

MAR - Since higher income group participants refuses to answer the survey on income we can say that it is MAR. Since the missingness can be attributed to the observed income variable

c

MCAR - Running out of batteries can be a very random phenomenon and can happen to any equipment hence it is completely at random

d

MCAR - The extreme windy day can happen any random day hence it can also be attributed to completely at random. and since the wind speed can be unrelated to the obs and unobs data

1

a

As the name suggests, the observed data is the type of data which can be observed and collected and has its own real world distribution. Throughout the code , the variables related to the observed data is subscripted by "_obs". These set of observed data would be used to perform inferences and analysis and estimate model parameters

The unobserved data refers to the latent variable which cant be measured directly from the observable distribution. Hence, throughout the code, these variables are subscripted as "_unobs". These type of data assists in building efficient models and perform inferences.

Looking at the observed data variables one would be able to see the true data or the real world data meanwhile the variables corresponding to the unobserved data would showcase the latent data and variables. Lines 51-61 are where these variables are set and passed to the model

b

Variable W is some perturbed variation of the true data X. This perturbation refers to the induction of measurement error into the data. Hence , W is the data with measurement error, while X is the corresponding data devoid of error. The structure in this case is, that , X is drawn from normal distribution with mean as 1/2 and sd as 1/6. The W is then perturbed by drawing from a normal distribution centered around X as mean. This pertains to some adding measurement error the true data X. Further both of these are segregated into observed and unobserved data. The lines from 110 to 115 defines the distribution of these variables.

c

As per line 45 it is 10%.

d

There were two warning messages that indicated that the chains have not been mixed and hence wasnt long enough. Acc to the warning messages , there were 1000 iterations (a very high number) that were divergent and had exceed max treedepth. the second warning was that the Rhat was na indicating the chains werent mixed. Although the acf shows that the parameters eventually decays to 0 , but not so soon. It means, that the chains can be made longer to have accurate estimates. THe trace plots also shows that the chain werent enough as it osicallates around a value. THe chains were long enough but could be extended to give more accurate details

e

The green region is the pointwise 95% credible interval sets corresponding to the confidence interval of the smoothed curve. The dark green curve corresponds to the estimate mean function of y estimated from measurement error data and the red curve is the mean function of y estimated from true data. The darkblue points corresponds to the true x values while the lightblue corresponds to the values of W which is data with measurement error.

f

Sigma x, epsilon and W along with the three quartiles of the mean functions were traced in the final MCMC summary plots. These predictors were traced to identify their corresponding posterior ditributions for bayesian inference of the mean function of Y. The bi-modality means there exists two regions in the pdf which has high densities of probable parameter values. This can mean that the parameter values can take two values with high probability. This can maybe hint that the parameter maybe part of some mixture models with multiple peaks from each mixture. In my case , only the 3rd quanitle had some effect of bi-modalty rest for evey other parameter estimate it was single peaked.

g

There can be a lot of places where such measurement erros can take place. There can be IoT devices which measures health monitoring metrics of certain physical devices which would be susceptible to such measurement errros due to unmeasurable environmental factors. In such cases it is very important to identify the root cause of whats causing the measurement error and then to proceed from there. This kind of modelling can be efficient in controlled environment places like factories where chances of measurement errors related to environment are less. But this would fail if such an assumption doesnt hold.

2

a

MCAR - (Missing data completey at random) missing data mechanism relates to the phenomenon when the data is missing completely at random without having any influence of any observed or unobserved factors and thus haves little to no implication in the analysis and inference of the model building

MAR - (Missing at Random) missing data mechanism relates to the phenomenon when the data is missing depended only on the observed variables and not on the unobserved factors. Hence , the missingness can be explained.

MNAR - (Missing Not at Random) is the missing data mechanism when the missingness can be related to some unoberved or unrelated phenomenon. This makes the missingness introduce significant bias and irregularities in the model inference.

b

MCAR: line 37 to 44 MAR: line 43 and 44 MNAR: line 68 and 69

c

MCAR: lines 95 and 96 MAR: Line 125 MNAR: Line 144

d

From the mean function and the scatter plots we can check the pattern of missingness for the three cases. In the case of MCAR , we can see that the missing points are scatter are completely at random with no apparent pattern . In the case of MAR the missing points are scattered in a way which is specific to a certain range and closely relates to the estimated mean function . In the case of MNAR, the missing data completely follows the pattern as of the observed data inferring a dirrection relation between missingness and y values.

e

The most concerning MCMC summary plots are for the case of MCAR. In this case the trace plots have high regions of oscillations showing that the chains not converging. The acf plots also show no decay to 0 values further proving non convergence. for the other cases of MAR and MNAR chains are properly converging.

f

I wouldnt be too worried about the bimodality of the W unobserved due to the characteristic of these data. the parameter may denote bimodality due to the higher dimensional function of the mean function hence corresponding to them , multiple higher density regions in pdf correspondingly.