AMS 315 Project 1

Zarif Ahmed

April 2025

Part A

Introduction

The objective of this assignment is to merge the given datasets and impute the missing values. Each file contains a column for subject ID and a column for either the dependent variable value or the independent variable value. We must sort and merge both datasets by subject ID, then count how many subjects are missing the independent variable, how many are missing the dependent variable, and how many are missing both. Then we need to impute the missing data using methods other than listwise deletion or mean/median imputation. The assignment simulates statistical processing work a newly hired statistician might be given.

Methods

All statistical computations and analyses mentioned were done in R. The code used is given in Figure 1. To start, we load and save both datasets in R using the read.csv() function. We then merge both datasets by ID using the merge() function. To quantify the missing data in each variable, we can pass the merged dataset to the function md.pattern() which gives a visual summary of the missing data pattern. Prior to imputing the values, we drop any row missing both values as there is no way to impute those values. We use the norm.boot method provided by the mice package in R to perform imputation. norm.boot is a linear regression using the bootstrap method and is effective when the values being imputed are normally distributed.

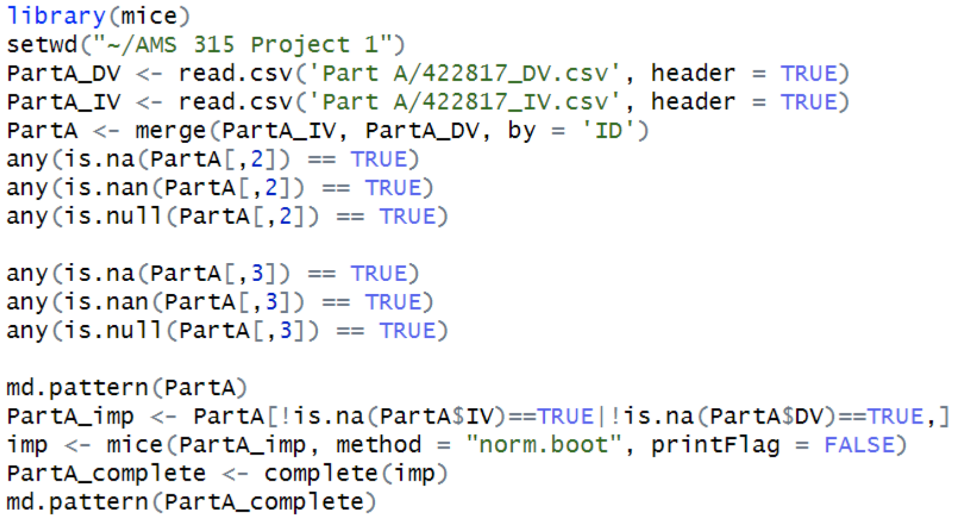
Results

We find that out of 584 subjects, 471 had both IV and DV, 66 subjects were missing just IV, 40 subjects were missing just DV, and 7 subjects were missing both IV and DV. After removing subjects missing both variables and performing imputation using bootstrap linear regression, we are returned a fully complete dataset with no missing values.

Conclusions and Discussion

The goal of this assignment was to merge two related datasets and impute the missing values. In the merged dataset we find that 113 out of 584 subjects were missing a value for at least one variable, approximately 20% of total subjects. To address this we use linear regression bootstrap model to impute the missing values. However, since the fraction of missing data is less than 30%, the choice of imputation method has little effect on the result. After imputation, the completed dataset has 66 subjects whose independent variable was imputed and 40 subjects whose dependent variable was imputed. This dataset is ready for further analysis without concerns of missing values.

Appendix



```r
library(mice)
setwd("~/AMS 315 Project 1")
PartA_DV <- read.csv('Part A/422817_DV.csv', header = TRUE)
PartA_IV <- read.csv('Part A/422817_IV.csv', header = TRUE)
PartA <- merge(PartA_IV, PartA_DV, by = 'ID')
any(is.na(PartA[,2]) == TRUE)
any(is.nan(PartA[,2]) == TRUE)
any(is.null(PartA[,2]) == TRUE)

any(is.na(PartA[,3]) == TRUE)
any(is.nan(PartA[,3]) == TRUE)
any(is.null(PartA[,3]) == TRUE)

md.pattern(PartA)
PartA_imp <- PartA[!is.na(PartA$IV)==TRUE|!is.na(PartA$DV)==TRUE,]
imp <- mice(PartA_imp, method = "norm.boot", printFlag = FALSE)
PartA_complete <- complete(imp)
md.pattern(PartA_complete)
```

Figure 1: R code used to perform all analyses