DA6213

Exercise #5

Missing Data

Kilger

Emily Bates

This exercise will cover some of the basics in dealing with missing data. The data set is presented to you in this exercise as the last pages.

1. Produce means and standard deviations in a table for the five variables in the data set.

| Var | Mean | Standard Deviation |

|:----------|:-------------:|:--------------------:|

|used\_car | 3.128788 | 1.213193 |

|safety | 3.366412 | 1.144929 |

|celebrity | 1.976562 | 1.090148 |

|foreign | 2.201550 | 1.134542 |

|meals | 2.178862 | 1.293315 |

1. Replace the missing values for all the variables in the data set using the mean substitution method. Place that table here…

| Var | Mean | Standard Deviation |

|:----------|:-------------:|:--------------------:|

|used\_car | 3.128788 | 1.141391 |

|safety | 3.366412 | 1.073049 |

|celebrity | 1.976562 | 1.009849 |

|foreign | 2.201550 | 1.055102 |

|meals | 2.178862 | 1.174231 |

1. Compare the means and standard deviations from the original data set table with the mean substitution table. What should happen to the means? What should happen to the standard deviations?

The means stay the same, while the standard deviation decreases with the mean substitution. This logically makes sense, because if more values are the same as the mean the overall standard deviation will decrease (there is less deviation from the mean).

1. Run one simple non-multiple imputation models (e.g. just a simple regression model) – for v3\_miss. Be sure to replace the missing values in v3\_miss with the predicted values from the regression. (I shortened this to just one variable – too much work otherwise I think).

| Var | Mean | Standard Deviation |

|:----------|:-------------:|:--------------------:|

|used\_car | NA | NA |

|safety | NA | NA |

|celebrity | 2.019074 | 1.083058 |

|foreign | NA | NA |

|meals | NA | NA |

1. Compare the table of means and standard deviations from the original data set to the one constructed in step 4 above. What do you see?

The mean has slightly increased compared to both tables above. The standard deviation has slightly decreased from original, but has not decreased as much as the last method.

1. Run a multiple imputation model on the five variables – you can pick from multiple imputation regression, mcmc or any other flavor.

| Var | Mean | Standard Deviation |

|:----------|:-------------:|:--------------------:|

|used\_car | 3.154362 | 1.212128 |

|safety | 3.328859 | 1.141614 |

|celebrity | 1.973154 | 1.077700 |

|foreign | 2.234899 | 1.158921 |

|meals | 2.194631 | 1.282301 |

1. Compare the table of means and standard deviations from step 6 with the original data set. What do you expect? What do you actually see?

You should expect to see similar means and standard deviations between the imputed data and the original data, as multiple imputation aims to retain the original data distribution as closely as possible. This appears to be what happened, as the means and standard deviations here are closer to the original table.

1. For an extra 8 points extra credit, produce a missing values matrix like the one shown in class. You can use SAS or name your poison in terms of platform as long as it produces a similar matrix.

|  |
| --- |
| **The SAS System** |

**The MI Procedure**

| **Missing Data Patterns** | | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Group** | **v1\_miss** | **v2\_miss** | **v3\_miss** | **v4\_miss** | **v5\_miss** | **Freq** | **Percent** | **Group Means** | | | | |
| **v1\_miss** | **v2\_miss** | **v3\_miss** | **v4\_miss** | **v5\_miss** |
| **1** | X | X | X | X | X | 86 | 57.33 | 3.023256 | 3.383721 | 1.976744 | 2.197674 | 2.127907 |
| **2** | X | X | X | X | . | 11 | 7.33 | 3.363636 | 3.363636 | 2.181818 | 2.090909 | . |
| **3** | X | X | X | . | X | 8 | 5.33 | 2.750000 | 4.000000 | 1.875000 | . | 2.625000 |
| **4** | X | X | X | . | . | 1 | 0.67 | 3.000000 | 3.000000 | 3.000000 | . | . |
| **5** | X | X | . | X | X | 8 | 5.33 | 3.375000 | 2.875000 | . | 1.875000 | 2.875000 |
| **6** | X | X | . | X | . | 4 | 2.67 | 3.500000 | 4.000000 | . | 2.750000 | . |
| **7** | X | X | . | . | X | 2 | 1.33 | 3.500000 | 3.500000 | . | . | 3.000000 |
| **8** | X | . | X | X | X | 6 | 4.00 | 3.500000 | . | 1.833333 | 2.000000 | 1.833333 |
| **9** | X | . | X | X | . | 3 | 2.00 | 3.666667 | . | 2.000000 | 1.666667 | . |
| **10** | X | . | X | . | . | 1 | 0.67 | 3.000000 | . | 3.000000 | . | . |
| **11** | X | . | . | X | X | 2 | 1.33 | 4.000000 | . | . | 1.500000 | 2.000000 |
| **12** | . | X | X | X | X | 7 | 4.67 | . | 3.000000 | 2.142857 | 2.857143 | 2.000000 |
| **13** | . | X | X | . | X | 1 | 0.67 | . | 1.000000 | 1.000000 | . | 1.000000 |
| **14** | . | X | X | . | . | 2 | 1.33 | . | 3.500000 | 1.500000 | . | . |
| **15** | . | X | . | X | X | 1 | 0.67 | . | 3.000000 | . | 3.000000 | 3.000000 |
| **16** | . | . | X | X | X | 1 | 0.67 | . | . | 1.000000 | 3.000000 | 1.000000 |
| **17** | . | . | X | . | X | 1 | 0.67 | . | . | 1.000000 | . | 1.000000 |
| **18** | O | O | O | O | O | 5 | 3.33 | . | . | . | . | . |

**Sas code for extra credit:**

/\* Define library location \*/

libname myfiles 'P:\Semester 2';

/\* Import the data \*/

data ex5;

infile 'P:\Semester 2\exercise5data.csv' dsd;

input v1\_miss v2\_miss v3\_miss v4\_miss v5\_miss;

run;

proc mi data=ex5 nimpute=0 ;

var v1\_miss v2\_miss v3\_miss v4\_miss v5\_miss;

ods select misspattern;

run;

**R code for other questions:**

#######################################################################

############################ EXERCISE 5 #############################

#######################################################################

# Emily Bates

#install.packages('mice')

library(mice)

setwd("/Users/emilybates/Documents/Documents - Emily’s MacBook Air/MSDA - Semester 2/Data Driven Decision Making and Design")

df <- read.csv('exercise5data.csv')

str(df)

sum(is.na(df)) #102 NA values

# Rename variables in df

names(df)[names(df) == "v1\_miss"] <- "used\_car"

names(df)[names(df) == "v2\_miss"] <- "safety"

names(df)[names(df) == "v3\_miss"] <- "celebrity"

names(df)[names(df) == "v4\_miss"] <- "foreign"

names(df)[names(df) == "v5\_miss"] <- "meals"

################

## Question 1 ##

################

original\_means <- colMeans(df, na.rm = TRUE)

original\_stddevs <- apply(df, 2, sd, na.rm = TRUE)

original\_summary <- data.frame(

Variable = names(original\_means),

Mean = original\_means,

'Standard Deviation'= original\_stddevs)

# Set row names to variable names

rownames(original\_summary) <- original\_summary$Variable

# Remove the 'Variable' column

original\_summary$Variable <- NULL

print("Original Dataset Summary:")

print(original\_summary)

print(knitr::kable(original\_summary, align = "c"))

################

## Question 2 ##

################

mean\_sub\_df = df

# Replacing missing values with mean values

for (col in names(mean\_sub\_df)) {

mean\_sub\_df[is.na(mean\_sub\_df[, col]), col] <- original\_means[col]}

sum(is.na(mean\_sub\_df)) #zero NAs now

# summary table

mean\_sub\_means <- colMeans(mean\_sub\_df)

mean\_sub\_stddevs <- apply(mean\_sub\_df, 2, sd)

mean\_sub\_summary <- data.frame(

Variable = names(mean\_sub\_means),

Mean = mean\_sub\_means,

'Standard Deviation'= mean\_sub\_stddevs)

# Set row names to variable names

rownames(mean\_sub\_summary) <- mean\_sub\_summary$Variable

# Remove the 'Variable' column

mean\_sub\_summary$Variable <- NULL

print("Mean Substitution Dataset Summary:")

print(mean\_sub\_summary)

print(knitr::kable(mean\_sub\_summary, align = "c"))

################

## Question 4 ##

################

df2 =df

impute\_model <- mice(df2[, c("celebrity", "used\_car", "safety", "foreign", "meals")], method = "norm.nob")

# Extract the imputed values for v3\_miss

imputed\_data <- complete(impute\_model, 1)

# Replace missing values in v3\_miss with the imputed values

df2$celebrity <- imputed\_data$celebrity

####################

# summary table

df2\_means <- colMeans(df2)

df2\_stddevs <- apply(df2, 2, sd)

df2\_summary <- data.frame(

Variable = names(df2\_means),

Mean = df2\_means,

'Standard Deviation'= df2\_stddevs)

# Set row names to variable names

rownames(df2\_summary) <- df2\_summary$Variable

# Remove the 'Variable' column

df2\_summary$Variable <- NULL

print("Mean Substitution Dataset Summary:")

print(df2\_summary)

print(knitr::kable(df2\_summary, align = "c"))

################

## Question 6 ##

################

df3 = df

mice\_mod\_multiple <- mice(df3, method = "pmm", m = 5)

df\_imputed\_multiple <- complete(mice\_mod\_multiple)

# summary table

df3\_means <- colMeans(df\_imputed\_multiple)

df3\_stddevs <- apply(df\_imputed\_multiple, 2, sd)

df3\_summary <- data.frame(

Variable = names(df3\_means),

Mean = df3\_means,

'Standard Deviation'= df3\_stddevs)

# Set row names to variable names

rownames(df3\_summary) <- df3\_summary$Variable

# Remove the 'Variable' column

df3\_summary$Variable <- NULL

print("Mean Substitution Dataset Summary:")

print(df3\_summary)

print(knitr::kable(df3\_summary, align = "c"))