Handling missing data with MICE

Neil Yetz & Gemma Wallace

01/29/2020

Missing data

Missing data are observations that should be part of your data but aren't

ID	Y	X1	X2	X3
1	32	0	6	5
2	25	1	5	3
3	40	1	7	6
4	?	?	?	?
5	5	1	4	2
6	27	0	?	7

Methods for handling missing data

- There are several!
- R uses listwise deletion by default
  - Can lose power and/or bias results
- Multiple Imputation by Chained Equations (MICE)
  - Imputation = substituting missing data with estimated values

MICE Lab Demo

1) Run a simple linear regression using pairwise deletion, the default in R
2) Impute dataset’s missing vales using the mice package
3) Run a simple linear regression in the imputed data
4) Compare model estimates across missing data techniques

Load Libraries

```
---
title: "PSY 653 Module 1: Missing Data"
subtitle: "Jan 29, 2020"
output:
  html_document:
    df_print: paged
---

# Part 1: In class Demo

## Load Libraries
```{r,message=FALSE}
library(tidyverse)
library(mice)
library(olsrr)
```
```

Read in data

```
## Read in data
```{r,message=FALSE}
mice_data1 <- read_csv("mice_data1.csv")
```
```

This dataset has 2 simulated variables: X1 and X2

X1 has some missing values

X1	X2
1	3
2	0
3	3
1	0
2	0
3	4
NA	0
1	0
2	1
3	0
1	2
2	0

Use a Simple Linear Regression to regress X1 on X2

```
## Simple Linear regression model X1 ~ X2
Using pairwise deletion for missing data by default
```{r}
mod1 <- lm(X1 ~ X2, data = mice_data1)
ols_regress(mod1)
```
```

Simple linear regression output with pairwise deletion

```
                          Model Summary
---------------------------------------------------------------
R                         0.026       RMSE                0.714
R-Squared                 0.001       Coef. Var          39.353
Adj. R-Squared           -0.002       MSE                 0.510
Pred R-Squared           -0.009       MAE                 0.592
---------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                ANOVA
-------------------------------------------------------------------
              Sum of
             Squares         DF    Mean Square      F        Sig.
-------------------------------------------------------------------
Regression     0.137          1          0.137    0.27     0.6039
Residual     204.939        402          0.510
Total        205.077        403
-------------------------------------------------------------------

                                  Parameter Estimates
---------------------------------------------------------------------------------------
      model      Beta    Std. Error    Std. Beta      t        Sig      lower    upper
---------------------------------------------------------------------------------------
(Intercept)     1.828         0.044                41.152    0.000      1.741    1.916
         X2    -0.011         0.022       -0.026   -0.519    0.604     -0.055    0.032
---------------------------------------------------------------------------------------
```

We interpret this output as usual

In write-up, would specify “missing data were handled using pairwise deletion”

Impute the data with mice

```
## Impute the dataset 5 times (using mice)
```{r}
imputed_data <- mice(mice_data1, m=5, maxit = 50, method = 'pmm', seed = 500)
```
```

- **mice_data1** = name of the dataset you are imputing
- **m** = # of imputations (# of imputed *versions* of the dataset you will create)
- **maxit** = number of iterations for each imputation (default is 5, generally do more)
- **method = pmm** = “Predictive Mean Matching”
- **seed** = specifying a # will allow you to get the same results each time

What the imputation process look like

MICE uses all of the other variables to predict each missing value

```
iter imp variable
 1   1  X1
 1   2  X1
 1   3  X1
 1   4  X1
 1   5  X1
 2   1  X1
 2   2  X1
 2   3  X1
 2   4  X1
 2   5  X1
 3   1  X1
 3   2  X1
```

Run the same simple linear regression of X1 on X2, but this time use the imputed dataset

```
## Regress X1 on X2 on imputed dataset using the "with" function
```{r, results= hide}
mod.imp <- with(imputed_data, exp= lm(X1 ~ X2))
summary(mod.imp)
```
```

- **mod.imp** = model name
- **with()** = tells R to run the analysis in all imputations of the data
- **exp** = an expression with a formula object
- **lm(X1 ~ X2)** = the model you want to run. In this case, a simple linear regression
- **summary()** = use to view model output

Simple linear regression output with MICE for *all imputations*

```
mod.imp <- with(imputed_data, exp= lm(X1 ~ X2))
summary(mod.imp)
```

term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>
(Intercept)	1.843998789	0.04306806	42.81592551	1.230265e-158
X2	-0.019184199	0.02092467	-0.91682191	3.597405e-01
(Intercept)	1.829037385	0.04236086	43.17752920	6.251121e-160
X2	-0.001437869	0.02058108	-0.06986363	9.443341e-01
(Intercept)	1.835000757	0.04178940	43.91067814	1.559009e-162
X2	-0.019259876	0.02030343	-0.94860191	3.433462e-01
(Intercept)	1.832034206	0.04266410	42.94088903	4.385727e-159
X2	-0.020546390	0.02072841	-0.99121896	3.221259e-01
(Intercept)	1.791168458	0.04269351	41.95411954	1.589717e-155
X2	0.012751627	0.02074270	0.61475263	5.390373e-01

1-10 of 10 rows

Imputation 1 | Imputation 2 | Imputation 3 | Imputation 4 | Imputation 5

Pool model estimates from all the imputations

```
## Pool model estimates across imputed versions of the dataset
```{r}
combined_imp <- pool(mod.imp)
summary(combined_imp)
```
```

```
                estimate  std.error  statistic       df p.value
(Intercept)  1.822457999 0.04641326 39.2658866 100.5742 0.00000
X2          -0.006076888 0.02123498 -0.2861734 289.4252 0.77495
```

- **mod.imp** = model name for SLR
- **pool()** = tells R to combine model estimates across each imputation
- **summary()** = use to view model output

Compare model results between the missing data techniques

Pairwise Deletion:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.82821    0.04443  41.152   <2e-16 ***
X2          -0.01142    0.02200  -0.519    0.604
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Pooled across MICE datasets:

```
                estimate  std.error  statistic       df p.value
(Intercept)  1.822457999 0.04641326 39.2658866 100.5742 0.00000
X2          -0.006076888 0.02123498 -0.2861734 289.4252 0.77495
```

A few notes on the mice package

- The with() and pool() functions allow you to pool model estimates for many common analyses
- In general, you should examine missing data patterns before using mice
- Can take a lot of computational power and time to run in larger datasets
- Not currently compatible with machine learning and some multivariate analyses
    - Mplus has its own code for multiple imputation
- To read more on the mice package, view the vignette here:
    - https://cran.r-project.org/web/packages/mice/mice.pdf