Week 13 - Basic advanced techniques

Alex Lishinski

November 2, 2021

Welcome!

Welcome to *week 13*!

Record the meeting

Discussion!

- What is something about R you've learned working on your final project that goes beyond what we've covered in class?

A recap of last week (on functional programming)

- R has built in functions from packages, but you can write your own too!

- The purrr package offers tools to do multiple iterations of operations

Chunk options

For a single chunk:

- eval = FALSE
- echo = FALSE
- warning = FALSE
- message = FALSE

For the entire document:

- knitr::opts_chunk$set(eval = FALSE, echo = FALSE, warning = FALSE, message = FALSE)

Topics for today

Record the meeting

A. Continued discussion on purrr - Iteration (or, *applying* functions)

B. Introduction to missing data procedures with MICE

A. Iteration

Iteration is another name for your computer carrying out some step(s) multiple times.

Iteration is helpful when, even after writing a function, you find yourself copying and pasting the same code (with modifications)

A. Iteration

Iteration can be helpful when you want to apply a function multiple times.

The map functions can help with this. There are different ones based on the type of object returned.

The key to understanding (and writing) functions is understanding what kind of input they take and what kind of output they return.

`map()` - Applies function to every element of a list and return a list `map2()` - Applies function to every element from 2 lists and return a list

Lists are ordered collections of any other type of object. You can have a list of vectors, a list of data frames, and a list of different types of objects.

`map_chr()` - returns a char vector of objects created `map_dbl()` - returns a numeric vector of objects created `map_dfc()` - objects created become columns of data frame `map_dfr()` - objects created become rows of a data frame

A. Iteration

Example: COVID vaccine data

A. Iteration

Example: COVID vaccine data - Read in the data

One way to load data:

```
pfi <- read_csv(here("data", "COVID", "COVID-19_Vaccine_Distribution_Allocations_by_Jurisdiction_-_Pf
mod <- read_csv(here("data", "COVID", "COVID-19_Vaccine_Distribution_Allocations_by_Jurisdiction_-_Mo
jan <- read_csv(here("data", "COVID", "COVID-19_Vaccine_Distribution_Allocations_by_Jurisdiction_-_Ja
jan
```

A. Iteration

Example: COVID vaccine data - Read in the data

```r
vaccines <- c("Pfizer", "Moderna", "Janssen")
file_base <- "/COVID-19_Vaccine_Distribution_Allocations_by_Jurisdiction_-_"
file_ext <- ".csv"

file_names <- str_c(here("content", "data", "COVID"), file_base, vaccines, file_ext)

vax_files <- file_names %>% map(read_csv)
names(vax_files) <- vaccines
vax_files
```

A. Iteration

Example: COVID vaccine data - Rename the data

```r
# write my rename function
rename_vaccine_data <- function(dat){
  dat <- dat %>% rename("State" = "Jurisdiction",
                "Week" = "Week of Allocations",
                "First Dose" = "1st Dose Allocations",
                "Second Dose" = "2nd Dose Allocations")
  dat
}

# map it to every data set in the list
vax_files %>% map(rename_vaccine_data)
```

Error: Can't rename columns that don't exist. x Column `2nd Dose Allocations` doesn't exist.

A. Iteration

Example: COVID vaccine data - Rename the data

```r
rename_vaccine_data <- function(dat){
  if(ncol(dat) == 4){
    dat <- dat %>% rename("State" = "Jurisdiction",
                          "Week" = "Week of Allocations",
                          "First Dose" = "1st Dose Allocations",
                          "Second Dose" = "2nd Dose Allocations")
    dat
  } else {
    dat <- dat %>% rename("State" = "Jurisdiction",
                          "Week" = "Week of Allocations",
                          "First Dose" = "1st Dose Allocations")
    dat
  }

}

vax_files <- vax_files %>% map(rename_vaccine_data)
vax_files$Pfizer
```

B. Multiple Imputation of missing data with MICE

`lm()` default way of dealing with missing data: listwise deletion

Better way to deal with missing data: imputation

Basic idea:

If data missing, you can fill in values that let you estimate your model on whole data

Good: Mean imputation

Better: Model based imputation

Best: Multiple imputation

B. Multiple Imputation of missing data with MICE

The MICE package: Multiple Imputation using Chained Equations

Dealing with missing data is a big idea, and the nuances can be tricky

Our goal is to show you that you need not be intimidated by approaching this task in R

If you are doing inferential models in your analyses, MI is a tool that should be in your toolkit

B. Multiple Imputation of missing data with MICE

Example: Regression model

```r
ug_data <- read_csv(here("content", "data", "undergrad_data.csv"))

vis_dat(ug_data)
```

B. Multiple Imputation of missing data with MICE

Example: Regression model

```
fit <- lm(PROJ_TOTAL ~ Prob_solve_pre + extraversion + agreeableness + conscient. + stability + openn
          data = ug_data)
```

B. Multiple Imputation of missing data with MICE

Example: Regression model

```
summary(fit)
```

```
##
## Call:
## lm(formula = PROJ_TOTAL ~ Prob_solve_pre + extraversion + agreeableness +
##     conscient. + stability + openness + Metacog. + SelfEfficacy +
##     IntGoalOrient + ExtGoalOrient, data = ug_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -334.72  -16.93   23.01   44.39   97.40
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    307.7820    35.8305   8.590  < 2e-16 ***
## Prob_solve_pre   3.3656     0.7639   4.406 1.34e-05 ***
## extraversion    -0.2442     0.3854  -0.634    0.527
## agreeableness   -0.1068     0.4867  -0.219    0.826
## conscient.       0.9207     0.4823   1.909    0.057 .
## stability        0.2047     0.3803   0.538    0.591
## openness         0.3604     0.5615   0.642    0.521
## Metacog.        -0.1107     0.5011  -0.221    0.825
## SelfEfficacy    -0.2373     0.5916  -0.401    0.689
## IntGoalOrient    0.1982     1.0673   0.186    0.853
## ExtGoalOrient    0.1893     0.8883   0.213    0.831
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 73.46 on 416 degrees of freedom
##   (187 observations deleted due to missingness)
## Multiple R-squared:  0.07321,    Adjusted R-squared:  0.05093
## F-statistic: 3.286 on 10 and 416 DF,  p-value: 0.0004114
```

B. Multiple Imputation of missing data with MICE

The MICE package: viewing missing data pattern

```r
library(mice)

md.pattern(ug_data)
```

B. Multiple Imputation of missing data with MICE

The MICE package: Generating imputed data sets with `mice()`

```
ug_data_imp <- mice(ug_data, m = 5)
```

```
##
## iter imp variable
## 1   1   Prob_solve_pre  Prob_solve_post  extraversion  agreeableness  conscient.  stability  openness  Met
## 1   2   Prob_solve_pre  Prob_solve_post  extraversion  agreeableness  conscient.  stability  openness  Met
## 1   3   Prob_solve_pre  Prob_solve_post  extraversion  agreeableness  conscient.  stability  openness  Met
## 1   4   Prob_solve_pre  Prob_solve_post  extraversion  agreeableness  conscient.  stability  openness  Met
## 1   5   Prob_solve_pre  Prob_solve_post  extraversion  agreeableness  conscient.  stability  openness  Met
## 2   1   Prob_solve_pre  Prob_solve_post  extraversion  agreeableness  conscient.  stability  openness  Met
## 2   2   Prob_solve_pre  Prob_solve_post  extraversion  agreeableness  conscient.  stability  openness  Met
## 2   3   Prob_solve_pre  Prob_solve_post  extraversion  agreeableness  conscient.  stability  openness  Met
## 2   4   Prob_solve_pre  Prob_solve_post  extraversion  agreeableness  conscient.  stability  openness  Met
## 2   5   Prob_solve_pre  Prob_solve_post  extraversion  agreeableness  conscient.  stability  openness  Met
## 3   1   Prob_solve_pre  Prob_solve_post  extraversion  agreeableness  conscient.  stability  openness  Met
## 3   2   Prob_solve_pre  Prob_solve_post  extraversion  agreeableness  conscient.  stability  openness  Met
## 3   3   Prob_solve_pre  Prob_solve_post  extraversion  agreeableness  conscient.  stability  openness  Met
## 3   4   Prob_solve_pre  Prob_solve_post  extraversion  agreeableness  conscient.  stability  openness  Met
## 3   5   Prob_solve_pre  Prob_solve_post  extraversion  agreeableness  conscient.  stability  openness  Met
## 4   1   Prob_solve_pre  Prob_solve_post  extraversion  agreeableness  conscient.  stability  openness  Met
## 4   2   Prob_solve_pre  Prob_solve_post  extraversion  agreeableness  conscient.  stability  openness  Met
## 4   3   Prob_solve_pre  Prob_solve_post  extraversion  agreeableness  conscient.  stability  openness  Met
## 4   4   Prob_solve_pre  Prob_solve_post  extraversion  agreeableness  conscient.  stability  openness  Met
## 4   5   Prob_solve_pre  Prob_solve_post  extraversion  agreeableness  conscient.  stability  openness  Met
## 5   1   Prob_solve_pre  Prob_solve_post  extraversion  agreeableness  conscient.  stability  openness  Met
## 5   2   Prob_solve_pre  Prob_solve_post  extraversion  agreeableness  conscient.  stability  openness  Met
## 5   3   Prob_solve_pre  Prob_solve_post  extraversion  agreeableness  conscient.  stability  openness  Met
## 5   4   Prob_solve_pre  Prob_solve_post  extraversion  agreeableness  conscient.  stability  openness  Met
## 5   5   Prob_solve_pre  Prob_solve_post  extraversion  agreeableness  conscient.  stability  openness  Met

## Warning: Number of logged events: 276
```

B. Multiple Imputation of missing data with MICE

The MICE package: Examining missing data

```
ug_data_imp$imp$SelfEfficacy
```

```
##       1  2  3  4  5
## 3    40 36 48 32 44
## 18   40 33 31 46 37
## 35   50 41 51 56 31
## 42   36 53 52 29 37
## 90   31 47 54 28 40
## 94   21 40 44 32 43
## 100  26 31 30 26 46
## 111  51 46 56 42 48
## 124  32 35 53 26 41
## 134  26 56 45 46 52
## 137  56 40 33 43 43
## 148  37 53 56 47 46
## 166  26 29 18 48 49
## 171  34 49 50 33 31
## 176  43 39 48 27 48
## 190  39 43 31 26 32
## 193  48 46 48 36 40
## 208  28 40 30 47 21
## 225  28 46 16 42 30
## 226  32 39 16 46 34
## 236  47 35 31 56 42
## 237  41 46 45 33 48
## 238  55 20 26 32 45
## 242  41 44 50 50 32
## 246  32 36 40 27 53
## 250  49 33 49 51 46
## 257  18 41 53 56 39
## 259  48 45 48 16 44
## 260  44 38 45 39 27
## 267  40 45 18 49 56
## 283  40 45 40 44 43
## 287  56 51 53 54 50
## 299  42 49 33 43 41
## 301  50 42 48 38 42
```

B. Multiple Imputation of missing data with MICE

The MICE package: Extracting data sets with `complete()`

```
comp_data <- complete(ug_data_imp, 2)
comp_data
```

```
##      ...1    Stud_ID PROJ_TOTAL Proj01 Proj02 Proj03 Proj04 Proj05 Proj06 Proj07
## 1       1 e75ab8254        347     14     19     13     25     27     35     40
## 2       2 07ee7a263        427     14     20     20     38     45     42     41
## 3       3 9c02eb000        393     14     18     16     37     42     44     45
## 4       4 52fb2d1ca        397     14     18     19      0     43     43     50
## 5       5 a4d76f570        316     14     20     12     36      8     34      0
## 6       6 94ea75daa        435     14     20     19     36     45     45     50
## 7       7 5e9c38d44        375     13     20     20     36     43     38     46
## 8       8 e5a5608e8        398     14     17     19     39     45     41     49
## 9       9 f2abdba22        350     14     18     18     37     44      0     45
## 10     10 21d036526        354     13     18     17     32     40     33     16
## 11     11 3fc8237b5        295     13     20     20     34     43     41     50
## 12     12 72ccca5a0        440     14     20     20     40     45     45     50
## 13     13 ca5073861         52     14     12      0     26      0      0      0
## 14     14 7e2674224        441     14     20     20     38     44     45     50
## 15     15 bc0032340        385     14     20     10     27     38     34     50
## 16     16 56bb07c2e        370     10     18     18     37     29     43     40
## 17     17 74afe8807        332     14     18     18     35     43      0     43
## 18     18 70653b198        279     14     20     18     32     43      0      0
## 19     19 c19bd5d09        406     14     20     16     37     41     45     50
## 20     20 758e0f2e4        340     15     20     20     40     32     43     44
## 21     21 2e3be0621        372     14     13     15     28     44     43     50
## 22     22 f7fc7c540        410     15     20     20     39     44     45     50
## 23     23 8a78da3d7        448     15     20     20     40     45     45     50
## 24     24 d032ddcba        408     15     20     10     25     45     45     50
## 25     25 f5fbfaa43        437     15     20     20     40     45     40     48
## 26     26 897c10248        348     15     18     14      0     37     42     44
## 27     27 006a7e31c        119     15     19     10     40      4     31      0
## 28     28 d7ebdcc64        299     13     20     20     40     39     45     14
## 29     29 2f9655115        392     13     20     17     33     39     45     49
## 30     30 1158a4bb2        446     15     20     20     38     43     45     50
## 31     31 da8783e45        438     15     20      9     40     45     45     50
## 32     32 34d07adfb        172      0     12     14     28     42     25      0
## 33     33 df6e080db        102     14      5      8     25     45      5      0
```

B. Multiple Imputation of missing data with MICE

The MICE package: fitting models with `with()`

```
fit_imp <- with(data = ug_data_imp, exp = lm( PROJ_TOTAL ~ Prob_solve_pre + extraversion + agreeablen
fit_imp
```

```
## call :
## with.mids(data = ug_data_imp, expr = lm(PROJ_TOTAL ~ Prob_solve_pre +
##     extraversion + agreeableness + conscient. + stability + openness +
##     Metacog. + SelfEfficacy + IntGoalOrient + ExtGoalOrient))
##
## call1 :
## mice(data = ug_data, m = 5)
##
## nmis :
##            ...1          Stud_ID       PROJ_TOTAL           Proj01           Proj02
##               0                0                0                0                0
##          Proj03           Proj04           Proj05           Proj06           Proj07
##               0                0                0                0                0
##          Proj08           Proj09           Proj10           Proj11        Total_Pts
##               0                0                0                0                0
##           Exam1            Exam2            Exam3   Prob_solve_pre  Prob_solve_post
##               0                0                0               43               63
##    extraversion    agreeableness       conscient.        stability         openness
##              74               90               88              110               85
##        Metacog.     SelfEfficacy    IntGoalOrient    ExtGoalOrient
##              67               67               67               67
##
## analyses :
## [[1]]
##
## Call:
## lm(formula = PROJ_TOTAL ~ Prob_solve_pre + extraversion + agreeableness +
##     conscient. + stability + openness + Metacog. + SelfEfficacy +
##     IntGoalOrient + ExtGoalOrient)
##
## Coefficients:
##     (Intercept)  Prob_solve_pre     extraversion    agreeableness       conscient.
##        232.0382          4.0817          -0.2252           0.9618           1.9892
##        stability        openness         Metacog.     SelfEfficacy    IntGoalOrient
```

B. Multiple Imputation of missing data with MICE

The MICE package: pooling output with `pool()`

```
summary(pool(fit_imp))
```

```
##                 term        estimate  std.error        statistic          df
## 1       (Intercept) 248.112986534 46.4815456    5.3378815844   39.666268
## 2     Prob_solve_pre   4.120263217  0.8674578    4.7498137549  214.980177
## 3      extraversion   0.000150478  0.5393323    0.0002790079   24.786566
## 4      agreeableness   0.415878402  0.7081347    0.5872871669   16.778688
## 5        conscient.   1.703344479  0.6151632    2.7689311737   37.228709
## 6          stability  -0.043612071  0.6076974   -0.0717660996   13.378748
## 7           openness  -0.579983283  0.9709956   -0.5973078148    9.694167
## 8           Metacog.  -1.112383493  0.5618775   -1.9797615080  129.706777
## 9        SelfEfficacy   1.832012714  0.8734123    2.0975347940   13.528703
## 10   IntGoalOrient  -0.147924066  1.2137433   -0.1218742570  118.055904
## 11   ExtGoalOrient  -0.372668612  1.1583485   -0.3217240898   40.612093
##           p.value
## 1   4.102423e-06
## 2   3.719334e-06
## 3   9.997796e-01
## 4   5.648310e-01
## 5   8.716692e-03
## 6   9.438498e-01
## 7   5.639919e-01
## 8   4.984741e-02
## 9   5.525172e-02
## 10 9.032058e-01
## 11 7.493119e-01
```

```
broom::tidy(fit)
```

```
## # A tibble: 11 × 5
##     term          estimate std.error statistic   p.value
##     <chr>            <dbl>     <dbl>      <dbl>     <dbl>
##  1 (Intercept)      308.       35.8       8.59  1.77e-16
##  2 Prob_solve_pre     3.37      0.764      4.41  1.34e- 5
##  3 extraversion      -0.244     0.385     -0.634 5.27e- 1
##  4 agreeableness     -0.107     0.487     -0.219 8.26e- 1
```

Logistics

This week

- Homework 9: Available Friday; **Due by Tuesday, 11/9**

- Readings: https://r4ds.had.co.nz/iteration.html#the-map-functions

  `https://www.statisticssolutions.com/multiple-imputation-for-missing-data/`

Wrapping up

In your base group's Slack channel:

- What is one thing you learned today?
- What is something you want to learn more about?
- Share your feelings in GIF form!