Introduction to R and RStudio

IMMERSE Training Team

Updated: March 13, 2023

Contents

IMMERSE Project

The Institute of Mixture Modeling for Equity-Oriented Researchers, Scholars, and Educators (IMMERSE) is an IES funded training grant (R305B220021) to support Education scholars in integrating mixture modeling into their research.

- Please visit our website to learn more and apply for the year-long fellowship.
- Follow us on Twitter!

How to reference this walkthrough: *This work was supported by the IMMERSE Project* (IES - 305B220021)

Visit our GitHub account to download the materials needed for this walkthrough.

Introduction to R and RStudio

This walkthrough is presented by the IMMERSE team and will go through some common tasks carried out in R. There are many free resources available to get started with R and RStudio. One of our favorites is *R for Data Science*.

PART 1: Installation

Step 0: Install R, RStudio, and Mplus

Here you will find a guide to installing both R and R Studio. You can also install Mplus here.

Note: The installation of Mplus requires a paid license with the mixture add-on. IMMERSE fellows will be given their own copy of Mplus for use during the one year training.

PART 2: Set-up

Step 1: Create a new R-project in RStudio

R-projects help us organize our folders , filepaths, and scripts. To create a new R project:

- File –> New Project...

Click "New Directory" –> New Project –> Name your project

Step 2: Create an R-markdown document

An R-markdown file provides an authoring framework for data science that allows us to organize our reports using texts and code chunks. This document you are reading was made using R-markdown!

To create an R-markdown:

- File –> New File –> R Markdown...

In the window that pops up, give the R-markdown a title such as "**Introduction to R and RStudio**" Click "OK." You should see a new markdown with some example text and code chunks. We want a clean document to start off with so delete everything from line 10 down. Go ahead and save this document in your R Project folder.

Table 1: LCA Indicators

Name	Label	Values
leaid	District Identification Code	
ncessch	School Identification Code	
report_dis	Number of students harassed or bullied on the basis of disability	0 = No reported incidents, 1 = At least one reported incident
report_race	Number of students harassed or bullied on the basis of race, color, or national origin	0 = No reported incidents, 1 = At least one reported incident
report_sex	Number of students harassed or bullied on the basis of sex	0 = No reported incidents, 1 = At least one reported incident
counselors_fte	Number of full time equivalent counselors hired as school staff	0 = No staff present, 1 = At least one staff present
report_sex	Number of full time equivalent psychologists hired as school staff	0 = No staff present, 1 = At least one staff present
counselors_fte	Number of full time equivalent law enforcement officers hired as school staff	0 = No staff present, 1 = At least one staff present

Step 3: Load packages

Your first code chunk in any given markdown should be the packages you will be using. To insert a code chunk, etiher use the keyboard shortcut ctrl + alt + i or Code –> Insert Chunk or click the green box with the letter C on it. There are a few packages we want our markdown to read in:

```
library(psych) # describe()
library(here) #helps with filepaths
library(gt) # create tables
library(tidyverse) #collection of R packages designed for data science
```

As a reminder, if a function does not work and you receive an error like this: `could not find function "random_function"`; or if you try to load a package and you receive an error like this: `there is no package called `random_package`` , then you will need to install the package using `install.packages("random_package")` in the console (the bottom-left window in R studio). Once you have installed the package you will *never* need to install it again, however you must *always* load in the packages at the beginning of your R markdown using `library(random_package)`, as shown in this document.

The style of code and package we will be using is called `tidyverse` . Most functions are within the `tidyverse` package and if not, I've indicated the packages used in the code chunk above.

PART 3: Explore the data

Step 4: Read in data

To demonstrate mixture modeling in the training program and online resource components of the IES grant we utilize the *Civil Rights Data Collection (CRDC)* (CRDC) data repository. The CRDC is a federally mandated school-level data collection effort that occurs every other year. This public data is currently available for selected latent class indicators across 4 years (2011, 2013, 2015, 2017) and all US states. In this example, we use the Arizona state sample. We utilize six focal indicators which constitute the latent class model in our example; three variables which report on harassment/bullying in schools based on disability, race, or sex, and three variables on full-time equivalent school staff hires (counselor, psychologist, law enforcement). This data source also includes covariates on a variety of subjects and distal outcomes reported in 2018 such as math/reading assessments and graduation rates.

To read in data in R:

```r
data <- read_csv(here("data", "crdc_lca_data.csv")) %>%
  mutate_if(is.character, as.numeric)

# Ways to view data in R:
# 1. click on the data in your Global Environment (upper right pane) or use...
View(data)
# 2. summary() gives basic summary statistics & shows number of NA values
# *great for checking that data has been read in correctly*
summary(data)
```

```
##      leaid           ncessch            report_dis      report_race
##  Min.   :400001   Min.   :4.000e+10   Min.   :0.0000   Min.   :0.000
##  1st Qu.:400804   1st Qu.:4.008e+10   1st Qu.:0.0000   1st Qu.:0.000
##  Median :403420   Median :4.034e+10   Median :0.0000   Median :0.000
##  Mean   :403865   Mean   :4.038e+10   Mean   :0.0425   Mean   :0.103
##  3rd Qu.:406330   3rd Qu.:4.063e+10   3rd Qu.:0.0000   3rd Qu.:0.000
##  Max.   :409734   Max.   :4.097e+10   Max.   :1.0000   Max.   :1.000
##  NA's   :20       NA's   :40          NA's   :27       NA's   :27
##    report_sex   counselors_fte     psych_fte         law_fte
##  Min.   :0.00   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.00   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.00   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.17   Mean   :0.4595   Mean   :0.4742   Mean   :0.1255
##  3rd Qu.:0.00   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1.00   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##  NA's   :27     NA's   :27       NA's   :30       NA's   :27
```

```r
# 3. names() provides a list of column names. Very useful if you don't have them memorized!
names(data)
```

```
## [1] "leaid"          "ncessch"        "report_dis"     "report_race"
## [5] "report_sex"     "counselors_fte" "psych_fte"      "law_fte"
```

```r
# 4. head() prints the top x rows of the dataframe
head(data)
```

```
## # A tibble: 6 x 8
##    leaid     ncessch report_dis report_race report_sex counselors_fte psych_fte
##    <dbl>       <dbl>      <dbl>       <dbl>      <dbl>          <dbl>     <dbl>
## 1 400001 40000100120          0           0          0              1         1
## 2 400001 40000100616          0           0          1              1         1
## 3 400001 40000101204          0           0          1              1         1
## 4 400001 40000101871          0           1          1              1         1
## 5 400001 40000101872          0           0          0              1         1
## 6 400001 40000102344          0           0          0              1         1
## # ... with 1 more variable: law_fte <dbl>
```

Step 5: Descriptive Statistics

Let's look at descriptive statistics for each variable. Because looking at the ID variables' (`leaid`) and (`necessch`) descriptives is unnecessary, we use `select()` to remove the variable by using the minus (`-`) sign:

```
data %>%
  select(-leaid, -ncessch) %>%
  summary()
```

```
##    report_dis      report_race      report_sex   counselors_fte
##  Min.   :0.0000   Min.   :0.000   Min.   :0.00   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.00   1st Qu.:0.0000
##  Median :0.0000   Median :0.000   Median :0.00   Median :0.0000
##  Mean   :0.0425   Mean   :0.103   Mean   :0.17   Mean   :0.4595
##  3rd Qu.:0.0000   3rd Qu.:0.000   3rd Qu.:0.00   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.000   Max.   :1.00   Max.   :1.0000
##  NA's   :27       NA's   :27      NA's   :27     NA's   :27
##    psych_fte         law_fte
##  Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000
##  Mean   :0.4742   Mean   :0.1255
##  3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000
##  NA's   :30       NA's   :27
```

Alternatively, we can use the `psych::describe()` function to give more information:

```
data %>%
  select(-leaid, -ncessch) %>%
  describe()
```

```
##                vars    n mean   sd median trimmed mad min max range skew
## report_dis        1 2000 0.04 0.20      0    0.00   0   0   1     1 4.53
## report_race       2 2000 0.10 0.30      0    0.00   0   0   1     1 2.61
## report_sex        3 2000 0.17 0.38      0    0.09   0   0   1     1 1.76
## counselors_fte    4 2000 0.46 0.50      0    0.45   0   0   1     1 0.16
## psych_fte         5 1997 0.47 0.50      0    0.47   0   0   1     1 0.10
## law_fte           6 2000 0.13 0.33      0    0.03   0   0   1     1 2.26
##                kurtosis   se
## report_dis        18.55 0.00
## report_race        4.82 0.01
## report_sex         1.08 0.01
## counselors_fte    -1.97 0.01
## psych_fte         -1.99 0.01
## law_fte            3.11 0.01
```

What if we want to look at a subset of the data? For example, what if we want to subset the data to observe a specific school district? (`leaid`) We can use `tidyverse::filter()` to subset the data using certain criteria.

```
data %>%
  filter(leaid == 408800) %>%
  describe()
```

```
##               vars  n      mean     sd    median   trimmed  mad       min
## leaid            1 86 4.088e+05   0.00 4.088e+05 4.088e+05  0.0 4.088e+05
```

```
## ncessch           2 86 4.088e+10 493.16 4.088e+10 4.088e+10 89.7 4.088e+10
## report_dis        3 86 5.000e-02   0.21 0.000e+00 0.000e+00  0.0 0.000e+00
## report_race       4 86 1.500e-01   0.36 0.000e+00 7.000e-02  0.0 0.000e+00
## report_sex        5 86 1.900e-01   0.39 0.000e+00 1.100e-01  0.0 0.000e+00
## counselors_fte    6 86 9.500e-01   0.21 1.000e+00 1.000e+00  0.0 0.000e+00
## psych_fte         7 86 1.900e-01   0.39 0.000e+00 1.100e-01  0.0 0.000e+00
## law_fte           8 86 1.400e-01   0.35 0.000e+00 6.000e-02  0.0 0.000e+00
##                      max range  skew kurtosis    se
## leaid          4.088e+05     0   NaN      NaN  0.00
## ncessch        4.088e+10  2597  2.58     7.77 53.18
## report_dis     1.000e+00     1  4.23    16.10  0.02
## report_race    1.000e+00     1  1.91     1.68  0.04
## report_sex     1.000e+00     1  1.59     0.52  0.04
## counselors_fte 1.000e+00     1 -4.23    16.10  0.02
## psych_fte      1.000e+00     1  1.59     0.52  0.04
## law_fte        1.000e+00     1  2.04     2.21  0.04
```

```
#You can use any operator to filter: >, <, ==, >=, etc.
```

Since we have binary data (0,1), it would be helpful to look at the proportions:

```
data %>%
  drop_na() %>%
  pivot_longer(report_dis:law_fte, names_to = "variable") %>%
  group_by(variable) %>%
  summarise(prop = sum(value)/n(),
            n = n()) %>%
  arrange(desc(prop))
```

```
## # A tibble: 6 x 3
##   variable         prop     n
##   <chr>           <dbl> <int>
## 1 psych_fte      0.481   1970
## 2 counselors_fte 0.459   1970
## 3 report_sex     0.173   1970
## 4 law_fte        0.127   1970
## 5 report_race    0.105   1970
## 6 report_dis     0.0431  1970
```

References

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R Package for Facilitating Large-Scale Latent Variable Analyses in Mplus. Structural equation modeling: a multidisciplinary journal, 25(4), 621-638.

Muthén, L.K. and Muthén, B.O. (1998-2017). Mplus User's Guide. Eighth Edition. Los Angeles, CA: Muthén & Muthén

US Department of Education Office for Civil Rights. (2014). Civil rights data collection data snapshot: School discipline. Issue brief no. 1.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

UC SANTA BARBARA