WELCOME TO THE 2023 NDACAN SUMMER TRAINING SERIES!

- The session will begin at 12pm EST.
- Please submit questions to the Q&A box.
- This session is being recorded.

NDACAN SUMMER TRAINING SERIES

National Data Archive on Child Abuse and Neglect

Cornell University & Duke University

NATIONAL DATA ARCHIVE ON CHILD ABUSE AND NEGLECT

NDACAN

Children's Bureau
An Office of the Administration for Children & Families

NDACAN SUMMER TRAINING SERIES SCHEDULE 2023

- July 5 — Introduction to NDACAN and the Administrative Data Series
- July 12 — New Data Acquisition: CCOULD Data
- July 19 — Causal Inference Using Administrative Data
- July 26 — Evaluating and Dealing with Missing Data in R
- August 2 — Time Series Analysis in Stata
- August 9 — Data Visualization in R

SESSION AGENDA

- Understanding why data are missing
- Common approaches to missing data
- Multiple imputation with AFCARS/NCANDS and R
- All code and demo data is available at
- https://github.com/f-edwards/ndacan_workshops/tree/main

INTRODUCTION TO MISSING DATA IN R

WHY SHOULD WE CARE?

- Most statistical software will conduct "complete-case analysis" by default
- Depending on how much data is missing in the variables you've chosen, this may result in throwing away a lot of perfectly good information!
- This (at minimum) biases your standard errors, and may bias your coefficient estimates
- With a few assumptions, we can correct the problem

WHY ARE DATA MISSING?

- **Missing completely at random (MCAR)**: The probability of a value being missing is the same for all observations in the data
- **Missing at random (MAR)**: The probability of a value being missing is random, conditional on other observed variables
- **Non-random missing data (MNAR)**: The probability of a value being missing depends on either *A)* some unobserved variable or *B)* the value itself (censorship)

COMMON APPROACHES TO MISSING DATA

BASIC APPROACHES TO MISSING DATA

- Listwise deletion (complete case analysis)
  - Appropriate for data with very few missing observations, or when missingness is completely at random and missingness is rare (independent of all observed and unobservable variables)
- Using alternative information (e.g. borrowing observation of sex from prior survey wave)
- Nonresponse weighting
  - Becomes difficult when many variables are missing, sub-populations of interest differ

BASIC APPROACHES TO MISSING DATA

- Multiple imputation
  - Iterative modeling of all missing outcomes/predictors in model
  - Produces multiple possible random datasets, allows you to average over uncertainty generated by missing data
  - Does not recover "true" values
  - Under missing at random assumption, generates unbiased parameter and variance estimates

MY PREFERRED APPROACH

- Understand your data!
  - Read the documentation
  - Do plenty of exploratory data analysis (cross tabs, data visuals, descriptives, look at the raw data)
  - Develop an understanding of the mechanisms of missing data in each dataset you use
  - Test your ideas for mechanisms of missing data when feasible

MY PREFERRED APPROACH

- If MAR is a reasonable assumption (it often is), conduct multiple imputation
  - Because MAR is conditional on observables, including many variables in imputation models is often a good idea
- Apply preferred final model / analysis over each imputed dataset, combine with Rubin's rules, report revised estimates.

APPLYING MISSING DATA METHODS TO AFCARS/NCANDS: A BRIEF INTRODUCTION

SOME NOTES BEFORE STARTING

- More work will be required to get it right for your analysis
- I'm using R (and the mice package) for my demo, but all major statistical packages (Stata, SAS, SPSS) use similar techniques
- All code and demo data is available at
  - https://github.com/f-edwards/ndacan_workshops/tree/main
- Submit data requests at https://www.ndacan.acf.hhs.gov/datasets/request-dataset.cfm

SET UP

```
library(mice)
library(tidyverse)
```

THE DATA WE ARE WORKING WITH: AFCARS FOSTER CARE 2018

```
names(afcars)

## [1] "FY"          "FIPSCode" "Entered"  "RaceEthn"

length(unique(afcars$FIPSCode))

## [1] 115
```

TASK 1: IMPUTATION OF INDIVIDUAL-LEVEL RACE-ETHNICITY DATA

- This is computationally intensive, so we'll work with a single year of the data
- If available, try to use a remote server for this kind of work
- Multiple imputation benefits from having all relevant information included
- I'll use population composition here, but more variables = better imputations

JOIN AFCARS TO POPULATION DATA TO IMPROVE PREDICTION

```
### Data from NIH; https://seer.cancer.gov/popdata/download.html
pop<-read_fwf("~/Projects/cps_lifetables/data/us.1990_2018.singleages.adjusted.txt",
              fwf_widths(c(4, 2, 2, 3, 2, 1, 1, 1, 2, 8),
                         c("year", "state", "st_fips",
                           "cnty_fips", "reg", "race",
                           "hisp", "sex", "age", "pop")))


pop<-pop%>%
  mutate(age = as.numeric(age),
         pop = as.numeric(pop),
         FIPSCode = paste(st_fips, cnty_fips, sep = "")) %>%
  rename(FY = year)
```

HARMONIZE RACE/ETHNICITY LABELS, AGGREGATE BY AGE

```
pop<-pop %>%
  mutate(race_ethn =
           case_when(
             race==1 & hisp ==0 ~ "White",
             race==2 ~ "Black",
             race==3 ~ "AIAN",
             race==4 ~ "AsianPI",
             hisp==1 ~ "Hispanic"))
```

RESHAPE DATA TO MAKE COMPOSITION VARIABLES

```
pop<-pop %>%
  filter(age<=18) %>%
  group_by(FY, FIPSCode, race_ethn) %>%
  summarise(pop = sum(pop)) %>%
  pivot_wider(names_from = race_ethn,
              values_from = pop)

head(pop)

## # A tibble: 6 x 7
## # Groups:   FY, FIPSCode [6]
##        FY FIPSCode  AIAN AsianPI Black Hispanic White
##     <dbl> <chr>    <dbl>   <dbl> <dbl>    <dbl> <dbl>
## 1   1990 01001        24      32  2607       68  7852
## 2   1990 01003       185      64  4952      352 21249
## 3   1990 01005         8      12  4235       26  3507
## 4   1990 01007        NA      NA  1392        8  3623
## 5   1990 01009        41      10   175       93 10263
## 6   1990 01011        NA       2  2965        4   555
```

MAKE COMPOSITION VARIABLES

```
pop<-pop %>%
  mutate(tot = AIAN + AsianPI + Black + Hispanic + White,
         pct_AIAN = AIAN/tot,
         pct_AsianPI = AsianPI/tot,
         pct_Black = Black/tot,
         pct_Hispanic = Hispanic/tot) %>%
  select(FY, FIPSCode, pct_AIAN, pct_AsianPI, pct_Black,
pct_Hispanic)
```

JOIN

```
afcars<-afcars %>%
  left_join(pop)
```

WHAT WE'LL IMPUTE

```
table(is.na(afcars$RaceEthn))
##
##  FALSE   TRUE
## 295572   6819
```

THE IMPUTATION MODEL

- Will build a multinomial regression for race/ethnicity
- FC Entry, FY, and county population composition will be predictors

BUILDING AN IMPUTATION MODEL IN R

```
afcars_imps<-mice(afcars)

##

##  iter imp variable

##   1   1  RaceEthn

##   1   2  RaceEthn

##   1   3  RaceEthn

##   1   4  RaceEthn

##   1   5  RaceEthn

##   2   1  RaceEthn

##   2   2  RaceEthn

##   2   3  RaceEthn

##   2   4  RaceEthn

##   2   5  RaceEthn

##   3   1  RaceEthn

##   3   2  RaceEthn

##   3   3  RaceEthn

##   3   4  RaceEthn

##   3   5  RaceEthn

##   4   1  RaceEthn

##   4   2  RaceEthn

##   4   3  RaceEthn

##   4   4  RaceEthn

##   4   5  RaceEthn

##   5   1  RaceEthn

##   5   2  RaceEthn

##   5   3  RaceEthn

##   5   4  RaceEthn

##   5   5  RaceEthn

## Warning: Number of logged events: 2
```

EVALUATING IMPUTATIONS

```
## # A tibble: 7 x 7
##   RaceEthn   `0`   `1`   `2`   `3`   `4`   `5`
##   <fct>    <int> <int> <int> <int> <int> <int>
## 1 1        82621 84400 84387 84410 84381 84334
## 2 2        94897 96795 96734 96763 96779 96756
## 3 3         5139  5243  5240  5238  5239  5244
## 4 4         2460  2515  2520  2522  2517  2518
## 5 5         1000  1020  1024  1020  1013  1019
## 6 6        21121 21501 21537 21526 21551 21523
## 7 7        88334 90917 90949 90912 90911 90997
```

EXTENDING TO OTHER DATASETS / VARIABLES

- These methods extend relatively simply to other variables
- But pay attention to the meaning of variables and relative share of missingness
- Some variables are simply not reported in particular states/years
- These present additional challenges - think carefully about why data might be missing
- If you can meet the MAR assumptions, MI is a good approach
- More imputations = more precision for uncertainty estimates

WORKSHOP7_26_23.RMD R CODE PAGE 1 OF 3

```
### read pop data and harmonize variable names to afcars names
pop<-read_csv("./data/pop_demo.csv") %>%
  rename(St = state,
         FY = year)
```

Let's explore the AgeAtStart measure

```{r}
table(dat$AgeAtStart)

### explicitly recode missings
dat<-dat %>%
  mutate(AgeAtStart =
           case_when(
             AgeAtStart >= 99 ~ NA,
             T ~ AgeAtStart
           ))
```

WORKSHOP7_26_23.RMD R CODE PAGE 1 OF 3

PIVOTING TO THE MICE DEMO WITH BUILT IN DATA

````
```{r}
head(nhanes)

summary(nhanes)

imps<-mice(nhanes)

```


```{r}
m0<-lm(chl ~ age, data = nhanes)
m1<-lm(chl ~ age + bmi + hyp, data = nhanes)
```
````

MAKE_POP_DATA.R CODE

```
library(tidyverse)

pop<-read_fwf("./data/us.1990_2020.singleages.adjusted.txt",
              fwf_widths(c(4, 2, 2, 3, 2,
                           1, 1, 1, 2, 8),
                         c("year", "state", "st_fips",
                           "cnty_fips", "reg", "race",
                           "hisp", "sex", "age", "pop")))

pop_demo<-pop %>%
  filter(year==2019) %>%
  select(year, state, sex, age, pop) %>%
  mutate(age = as.numeric(age),
         pop = as.numeric(pop))

write_csv(pop_demo, "./data/pop_demo.csv")
```

MAKE_SAMPLE_DATA.R R CODE

```
######## make sample data for workshop
####### read in and deidentify admin data for geo / time join

library(data.table)
library(tidyverse)

ncands<-fread("~/Projects/ndacan_data/ncands/CF2019v1.tab")

afcars<-fread("~/Projects/ndacan_data/afcars/FC2019v1.tab")

###### select variables for join
ncands_demo<-ncands %>%
  select(subyr, StaTerr, ChAge)

afcars_demo<-afcars %>%
  select(FY, STATE, St, AgeAtStart)

write_csv(ncands_demo,
          "./data/ncands_demo.csv")

write_csv(afcars_demo,
          "./data/afcars_demo.csv")
```

READ_NDACAN_DATA.R R CODE PAGE 1 OF 3

```
### this script joins ndacan tables to SEER pop data
### load libraries

library(tidyverse)

### read in the demo files
ncands<-read_csv("./data/ncands_demo.csv")
afcars<-read_csv("./data/afcars_demo.csv")
pop<-read_csv("./data/pop_demo.csv")

### harmonize the names in ncands and pop

ncands<-ncands %>%
  rename(year = subyr,
         state = StaTerr,
         age = ChAge)
unique(ncands$age)

### note that 77 and 99 have special meaning
### recode 77 -> 0; 99 -> NA

ncands<-ncands %>%
  mutate(age = ifelse(age==77, 0,
                      ifelse(age==99, NA,
                             age)))
```

READ_NDACAN_DATA.R R CODE PAGE 2 OF 3

```
### collapse NCANDS to state - year, collapse pop to state - year

ncands_st<-ncands %>%
  group_by(year, state, age) %>%
  summarize(child_investigation = n())

pop_st<-pop %>%
  filter(age<18) %>%
  group_by(year, state, age) %>%
  summarize(pop = sum(pop))

##### join them together

ncands_pop<-ncands_st %>%
  left_join(pop_st)

### super cool!
### now let's do afcars

afcars<-afcars %>%
  rename(year = FY,
         state = St,
         age = AgeAtStart) %>%
  mutate(age = ifelse(age<0, 0, age),
         age = ifelse(age==99, NA, age)) %>%
  select(-STATE)
```

READ_NDACAN_DATA.R R CODE PAGE 3 OF 3

```
### collapse to state level
afcars_st<-afcars %>%
  group_by(year, state, age) %>%
  summarize(fc = n())

### now join to ncands_pop
ncands_afcars_pop<-ncands_pop %>%
  left_join(afcars_st)

### compute per capita rates
ncands_afcars_pop<-ncands_afcars_pop %>%
  mutate(investigation_rate = child_investigation / pop * 1000,
         fc_rate = fc / pop * 1000)

### quick visuals
ggplot(ncands_afcars_pop,
       aes(x = age, y = investigation_rate)) +
  geom_line() +
  facet_wrap(~state)

ggplot(ncands_afcars_pop,
       aes(x = age, y = fc_rate)) +
  geom_line() +
  facet_wrap(~state)

library(geofacet)

ggplot(ncands_afcars_pop,
       aes(x = age, y = fc_rate)) +
  geom_line() +
  facet_geo(~state)
```

WORKSHOP7_26_23.RMD R CODE PAGE 1 OF 3

```
---
title: "Handling missing data in AFCARS"
output: html_notebook
editor_options:
  chunk_output_type: inline
---
```

Load in the needed packages

```
```{r}
library(tidyverse)
library(mice)
```
```

First let's load in the de-identified AFCARS data and state population data

```
```{r}
dat<-read_csv("./data/afcars_demo.csv")
```
```

QUESTIONS?

FRANK EDWARDS

FRANK.EDWARDS@RUTGERS.EDU

NEXT WEEK…

August 2nd, 2023

Presenter:

Alexander F. Roehrkasse, Ph.D., Butler University

Topic:

Time Series Analysis in Stata