Summer of NYTD, 2018

National Archive for Child Abuse and Neglect Bronfenbrenner Center for Translational Research Cornell University

Summer of NYTD Session 3

Session starts at 12pm EST

- Please turn your video off and mute your line
- This session is being recorded
- See ZOOM Help Center for connection issues: https://support.zoom.us/hc/en-us
  - If issues persist and solutions cannot be found through Zoom contact hl332@cornell.edu

Introduction

Summer schedule:

- August 8th - Introduction
- August 15th - Data Structure
- **August 22nd - Expert Presentation I**
- August 29th - Expert Presentation II
- September 5th - Linking to NCANDS & AFCARS
- September 12th - Research Presentation I
- September 19th - Research Presentation II

Today's Presentation:
Understanding and addressing missing data in NYTD

Presenters:
Michael Dineen (med39@cornell.edu) and Frank Edwards (fedwards@cornell.edu)

Agenda for today's webinar

- Develop a clear understanding of the design of the NYTD and the structure of the sample
- Discuss differences in the composition of state samples and methods states use to collect data
- Discuss sources of missing data and non-response
- Discuss the theories behind statistical approaches to missing data, with a focus on multiple imputation
- Discuss some practical strategies to address missing data in the NYTD

NYTD Design

Understanding the structure of the National Youth in Transition Database (NYTD)

- The user's guide and codebook are your friends
- The NYTD Outcomes Survey is ongoing, with new cohorts commencing every 3 years, starting with Federal Fiscal Year 2011.
  - Cohort 1 was 17 in 2011, Cohort 2 was 17 in 2014
- Each Cohort has three waves, with two years between surveys
  - Cohort 1 [2011, 2013, 2015], Cohort 2 [2014, 2016, 2018]

Who is in the cohort?

- Youth who:
  - Are in foster care at the time they took the survey
  - Answer at least one survey question on the baseline survey
  - Took the survey within 45 days of their 17th birthday
- Follow-up surveys are conducted during the six-month AFCARS reporting period that includes the youth's 19th and 21st birthdays.

State sampling

- States are permitted to sample the cohort for the age 19 and 21 follow-ups.
- Simple random sampling is required
- Sampling is done once, after the cohort is determined.
- The same sample is used for both the age 19 and age 21 surveys.

Sources of missing data in the NYTD

Sources of missing data: not-in-cohort

- Response in Wave 1 to voluntary questions is required to be selected for the cohort
  - Youth who do not respond to the baseline survey are not followed-up at subsequent waves, so all survey data for these cases are missing
- However, demographic data are present
- This means that the cohort is not a random or representative sample if choosing to respond is associated with any of the variables in the study.

Wave non-response

- Youth did not participate in a wave.
- All survey data for that wave will be missing for that row.
- Demographics will be present.

Reasons for non-response

- Youth declined: The State agency located the youth successfully and invited the youth's participation, but the youth declined to participate in the data collection.
- Parent declined: The State agency invited the youth's participation, but the youth's parent/guardian declined to grant permission.
    - This response may be used only when the youth has not reached the age of majority in the State and State law or policy requires a parent/guardian's permission for the youth to participate in information collection activities.

Reasons for non-response (continued)

- Incapacitated: The youth has a permanent or temporary mental or physical condition that prevents him or her from participating in the outcomes data collection.
- Incarcerated: The youth is unable to participate in the outcomes data collection because of his or her incarceration.
- Runaway/missing: A youth in foster care is known to have run away or be missing from his or her foster care placement.
- Unable to locate/invite: The State agency could not locate a youth who is not in foster care or otherwise invite such a youth's participation.
- Death: The youth died prior to his participation in the outcomes data collection.

Question non-response

- This is the easiest form of missing data to deal with, but rare in NYTD

Approaches to missing data 101

Why should we care?

- Most statistical software will conduct "complete-case analysis" by default
  - This uses only those observations where regression outcomes and all predictors are non-missing
- 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 parameter point 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. Missingness is determined by a coin flip/dice roll
- **Missing at random (MAR)**: The probability of a value being missing is *not* completely at random, depends only on available (observed) information. The probability of a value being missing is determined by other variables in the data
- **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)

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

- Deterministic imputation methods
  - Many examples: linear interpolation or last observed, regression imputation
  - This is generally a bad idea. Covariance estimates and standard errors are biased downward

Basic approaches to missing data

- Multiple imputation (MI)
  - Iterative modeling of all missing outcomes/predictors in model
  - Produces fake 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

What multiple imputation does:

- Has two effects on model uncertainty
  - Increases your N because we aren't deleting data (pushes standard errors downward)
  - Adds in appropriate noise due to uncertainty around where missing values are (pushes standard errors upward)
- If missingess is associated with observables, MI can correct bias in parameter 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

- Use available information
  - Borrow data from other observations when possible
  - Some variables are time-stable (age) and can be borrowed from prior observations - but remember cautions against deterministic imputation and inducing bias

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 NYTD: a very brief introduction

Some notes before starting

- This is a very brief introduction, 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) should be able to use similar techniques
- All code (and slides, but no data!) is available at https://github.com/f-edwards/nytd_missing_data_demo
- We are using NYTD Outcomes File, Cohort Age 17 in FY2011, Waves 1-3 (NDACAN Dataset 202).
- Submit data requests at https://www.ndacan.cornell.edu/datasets/request-dataset.cfm

Load in packages and data

```
### load required packages
library(tidyverse)
library(lubridate)
library(mice)
### read in tab separated data
nytd<-read_tsv("Outcomes_C11W3v2.tab")
```

Create cohort subset

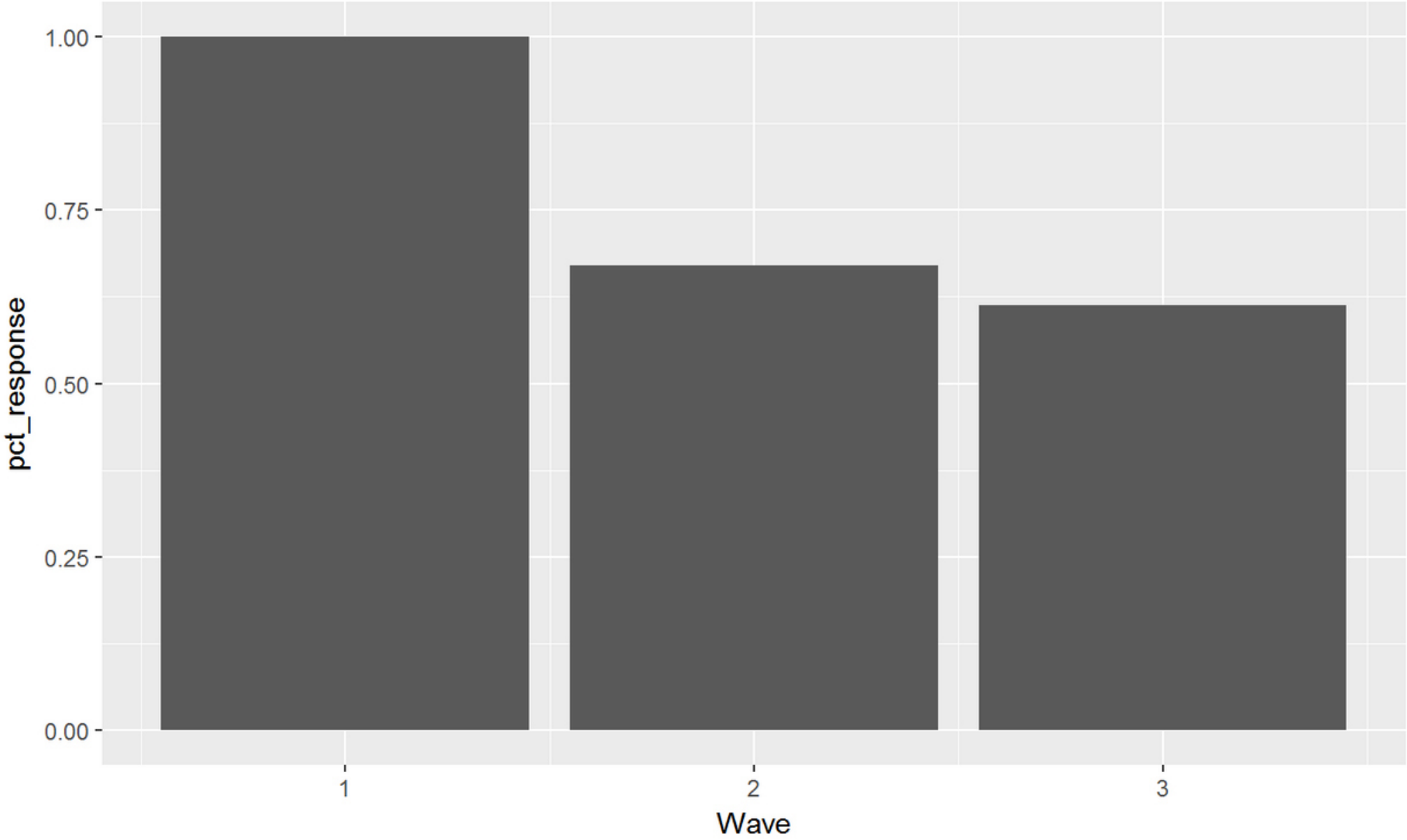
```
### count total population, cohort based on baseline
pop<-sum(nytd$Wave==1)
### subset on those in cohort
cohort<-nytd%>%
  filter(FY11Cohort==1)%>%
  filter(!(SampleState==1 & InSample==0))
```

Describe response rates

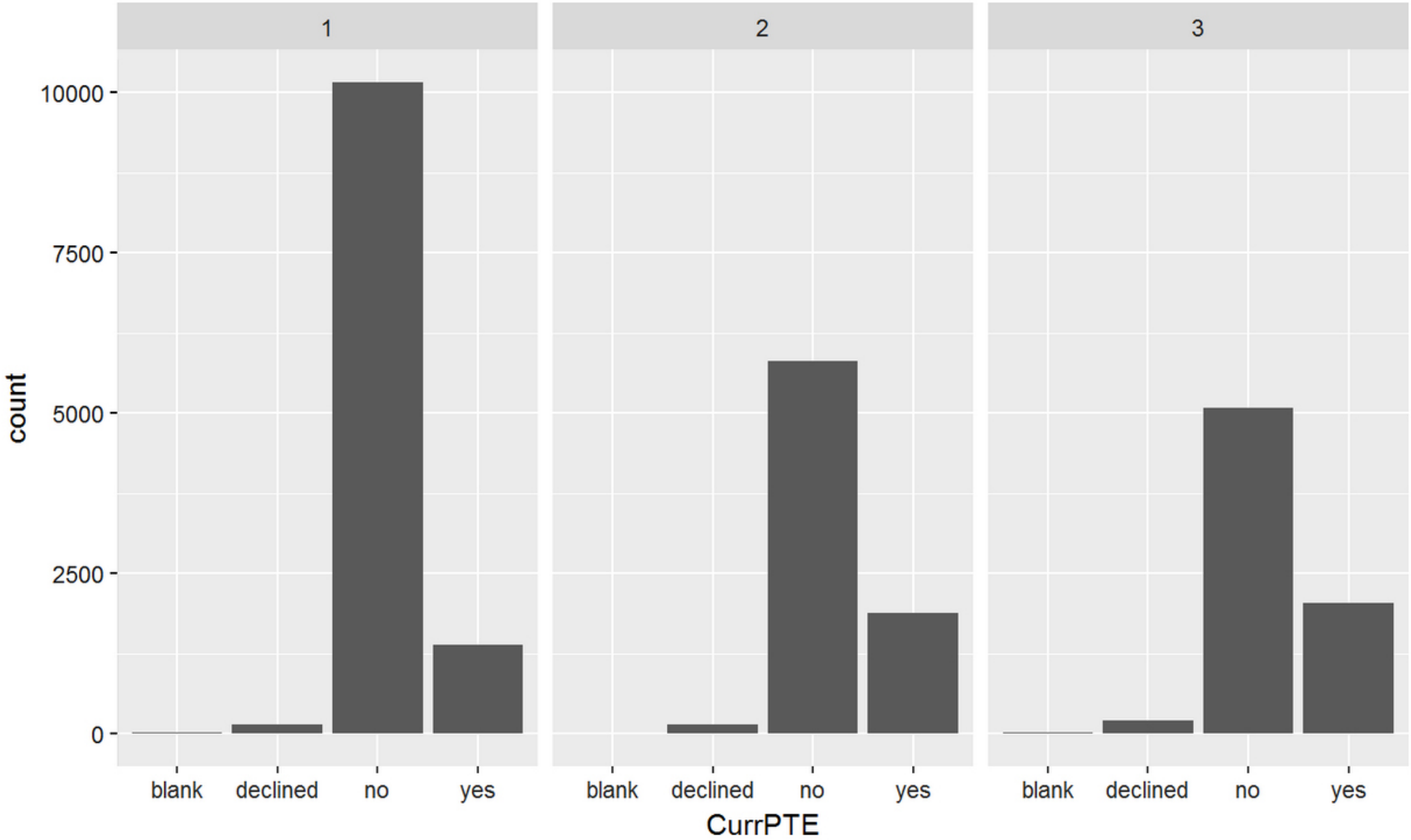
```
## response rate by wave
nytd%>%filter(FY11Cohort==1)%>%
  filter(Responded==1)%>%
  group_by(Wave)%>%
  summarise(baseline = pop, responses = n(), response_rate = n()/pop)
```

Wave	baseline	responses	response_rate
<int>	<int>	<int>	<dbl>
1	29104	15597	0.536
2	29104	7897	0.271
3	29104	7470	0.257

Response rates for cohort

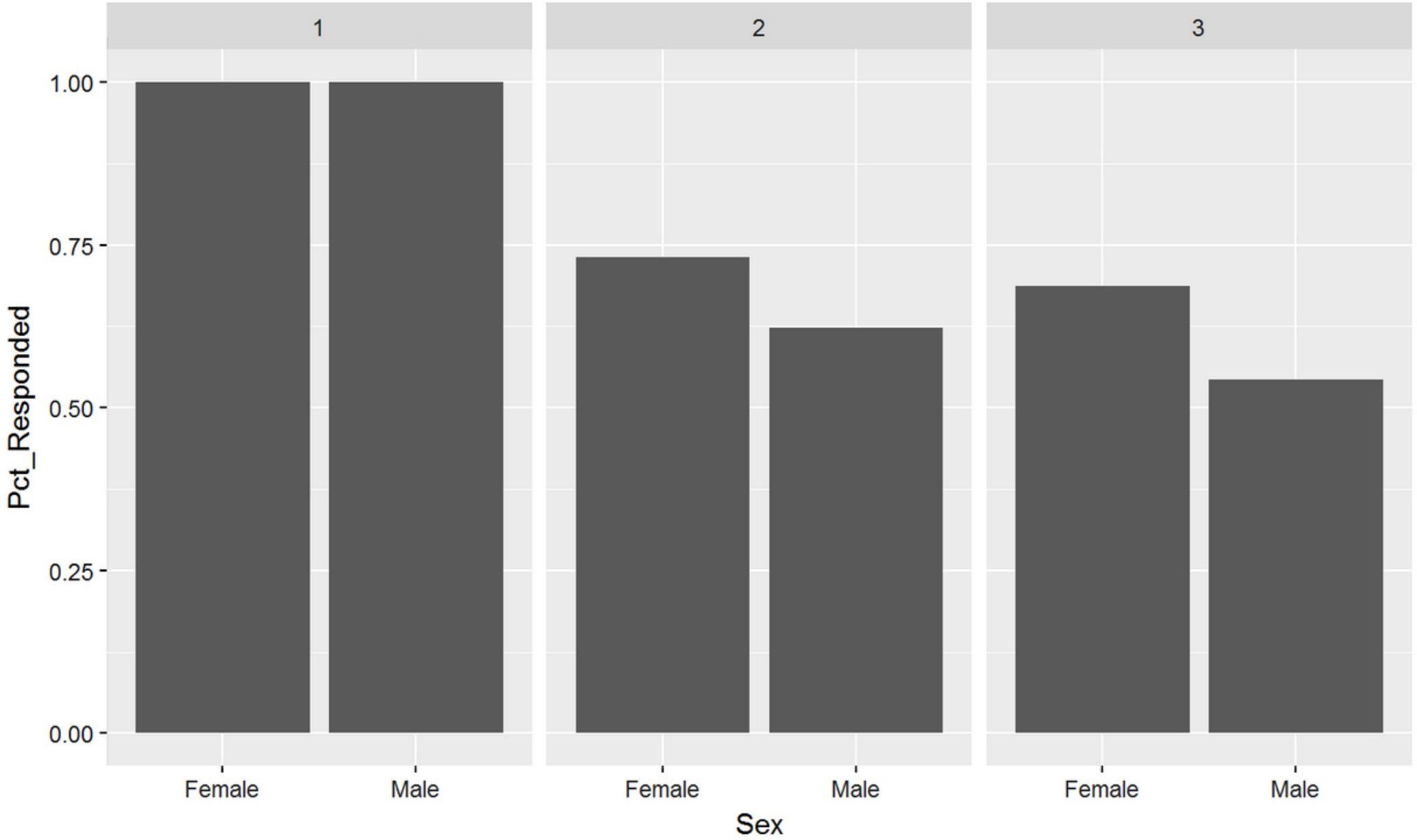


Question non-response

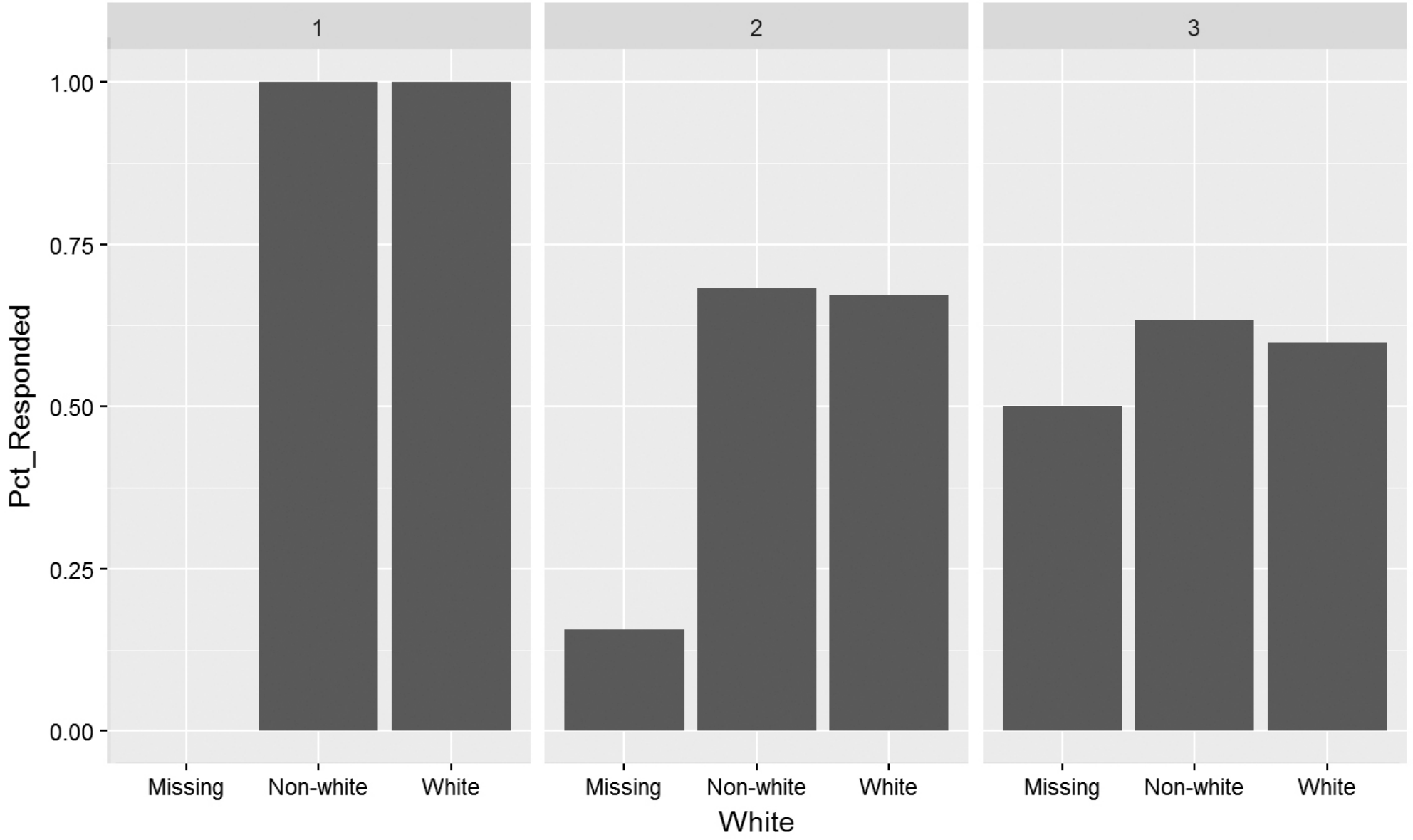


What drives non-response?

Non-response by gender



Non-response by race and wave (white / non-white)



Proceeding with Multiple Imputation

- For the purposes of demonstration, we'll assume that CurrPTE is MAR conditional on sex, race/ethnicity, and age (Wave).
- But a full imputation model should include *AS MANY* predictors as possible to maximize predictive performance and satisfy MAR assumption. This is just a demonstration of what we can do
- Note that models grow in complexity as new predictors (with their own missing values) increase computation time

Set up imputation dataset

```
to_impute<-cohort%>%
  select(Wave, Sex, CurrPTE, RaceEthn)%>%
  mutate(CurrPTE = ifelse(CurrPTE == 77, NA, CurrPTE))%>%
  mutate(Sex = factor(Sex),
         Wave = factor(Wave),
         CurrPTE = factor(CurrPTE),
         RaceEthn = factor(ifelse(RaceEthn==99, NA, RaceEthn)))

head(to_impute)
```

Row #	Wave	Sex	CurrPTE	RaceEthn
1	1	Female	0	1
2	1	Female	0	3
3	1	Female	1	1
4	1	Male	0	1
5	1	Female	0	3
6	1	Male	2	7

Look at missing data patterns

```
### install.packages("mice") if needed
### wonderful tutorials at https://stefvanbuuren.github.io/mice/
summary(to_impute)
```

Wave	Sex	CurrPTE	RaceEthn
1:11713	Female: 17109	0 : 21066	1 :15872
2:11712	Male: 18235	1 : 5307	2 :10431
3:11994	NA's: 75	2 : 529	7 : 6173
		NA's: 8517	6 : 1499
			3 : 605
			(Other): 377
			NA's : 462

Predictors and methods

```
imp<-mice(to_impute, maxit=0, seed = 123)
### show the predictor matrix
imp$predictorMatrix
```

	Wave	Sex	CurrPTE	RaceEthn
Wave	0	0	0	0
Sex	1	0	1	1
CurrPTE	1	1	0	1
RaceEthn	1	1	1	0

```
### show imputation methods, logistic and multinomial
imp$method
```

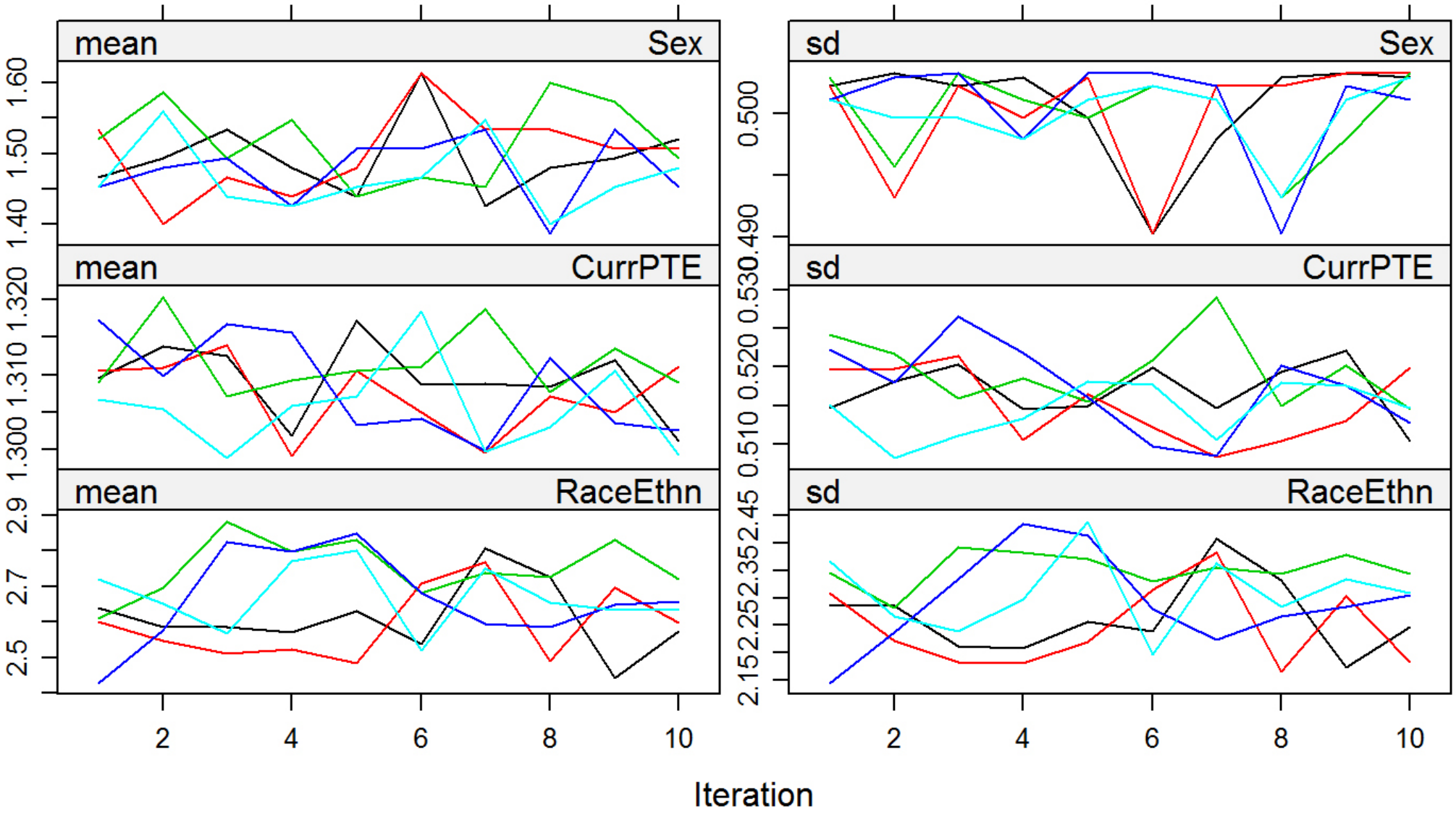
Wave	Sex	CurrPTE	RaceEthn
""	"logreg"	"polyreg"	"polyreg"

Impute data with 5 imputed data sets

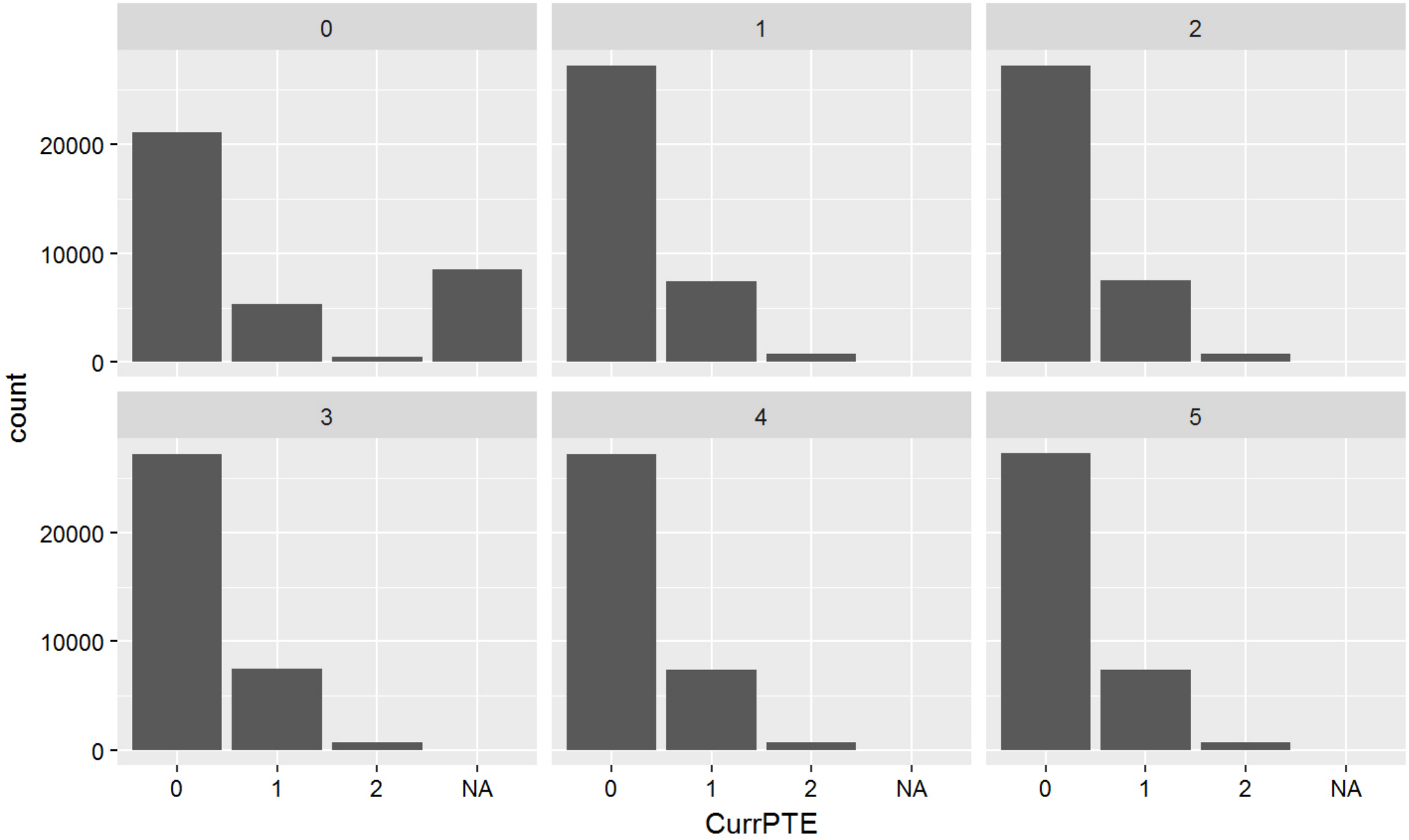
```
imp_out<-mice(to_impute,
              maxit = 10,
              m = 5,
              seed = 123,
              predictorMatrix = imp$predictorMatrix,
              method = imp$method)
```

iter	imp	variable
1	1	Sex CurrPTE RaceEthn
1	2	Sex CurrPTE RaceEthn
1	3	Sex CurrPTE RaceEthn
1	4	Sex CurrPTE RaceEthn
1	5	Sex CurrPTE RaceEthn
2	1	Sex CurrPTE RaceEthn
2	2	Sex CurrPTE RaceEthn
2	3	Sex CurrPTE RaceEthn
2	4	Sex CurrPTE RaceEthn
2	5	Sex CurrPTE RaceEthn

Check out convergence



Check out effects of imputation on CurrPTE



Compare imputed to original data

.imp	pct_PTE	pct_nonPTE
<fct>	<dbl>	<dbl>
0	0.150	0.595
1	0.210	0.769
2	0.211	0.767
3	0.212	0.767
4	0.210	0.769
5	0.209	0.770

Only among non-missings

.imp	pct_PTE	pct_nonPTE
<fct>	<dbl>	<dbl>
0	0.197	0.783
1	0.210	0.769
2	0.211	0.767
3	0.212	0.767
4	0.210	0.769
5	0.209	0.770

Conduct a pooled analysis

```
### fit logistic regression on each imputed data set
fit_imp<-with(imp_out, glm(CurrPTE == "1" ~ (Sex=="Male") +
                             Wave + RaceEthn,
                           family = "binomial"))
## Pool results with Rubin's rules
pooled<-pool(fit_imp)
### just with observed data
fit<-with(to_impute, glm(CurrPTE == "1" ~ (Sex=="Male") +
                           Wave + RaceEthn,
                         family = "binomial"))
```

Compare models

```
library(broom)
# with only observed
tidy(fit)[1:2, 1:3]
```

term	estimate	std.error
(Intercept)	-1.8688112	0.03678118
Sex == "Male"TRUE	-0.1768677	0.03153402

```
# with imputed data
summary(pooled)[1:2, 1:2]
```

term	estimate	std.error
(Intercept)	-1.8656265	0.03539415
Sex == "Male"TRUE	-0.1830675	0.02716801

Going deeper

What we accomplished

- We adjusted our models for non-response bias between waves
  - We imputed 5 complete datasets for the cohort, averaged over uncertainty in our models
- This appears to matter for our estimates of Employment ~ Gender

This approach is incomplete

- We haven't dealt with selection into the cohort
- We haven't fully explored the mechanisms of missing data
- Our model is too simplistic
  - Incorporate documented reason for non-response into models
  - Extend to focal variables for your analysis
  - Think carefully about why your variable is missing, what other observed variables in the data may help you estimate uncertainty
- Bonus points: go multilevel

Possible extensions of this method

- Theoretically, the method could be used to estimate uncertainty intervals for parameters for the full NYTD-eligible population Though this would be computationally intensive and
- Stretch the MAR assumption perhaps too far (if participation in the cohort is conditional on unobservables)

Final notes

- Approaches to missing data are not one-size fits all.
- Think hard about why your data are missing
- If they are MAR conditional on observables, MI may be appropriate

Further reading

- Rubin, "Multiple Imputation for Nonresponse in Surveys"
- Gelman and Hill, "Data Analysis Using Regression and Multilevel/Hierarchical Models"
- van Buuren, "mice: Multivariate Imputation by Chained Equations in R"

Questions?

- Please use the chat feature in Zoom to submit questions during the Q and A
- Code and slides available at:
- https://github.com/f-edwards/nytd_missing_data_demo
- Frank Edwards: fedwards@cornell.edu
  Michael Dineen: med39@cornell.edu

Chat Questions:

1. **Is the non-response reason captured even if they don't participate or are they dropped completely from the survey?**
2. **Of the 5 imputations, which imputation do you use to report results.**
3. **Could you describe uncertainty intervals and how you use them?**
4. **Have you tried other algorithms for imputation such as KNN, K nearest neighbor?**
5. **What about multiple imputation for simpler metrics such as medians?**

Next Week

- Date: Wednesday August 29th from 12pm-1pm
- Presenter: Michael Dineen, BCTR at Cornell University
- Topic: Expert Presentation II-Developing and using sample weights, and other common questions we get from data users.