Who’s Underrepresented?
Modeling Undercount in the U.S. Census

Maria Tackett
Duke University

JSM
August 2020

🔗 bit.ly/jsm2020-teach

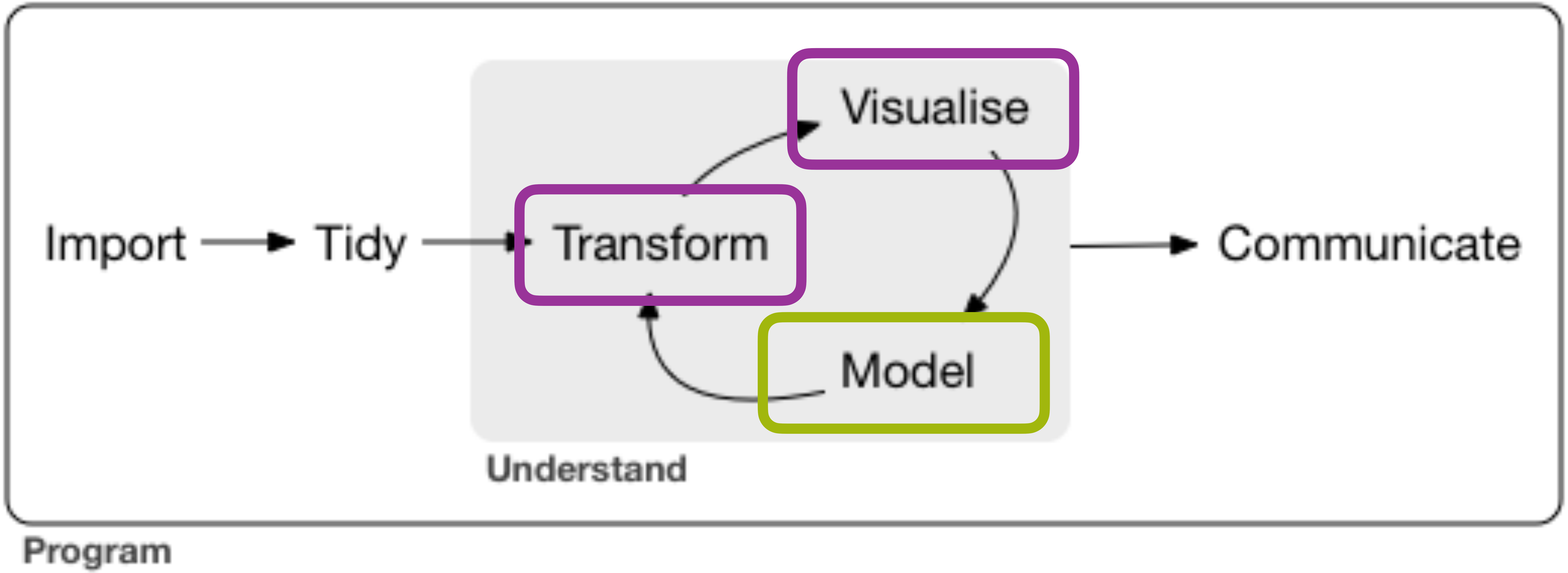
The course

Second semester undergraduate statistics course (~ 90 students)

Multiple linear regression, logistic regression, ANOVA

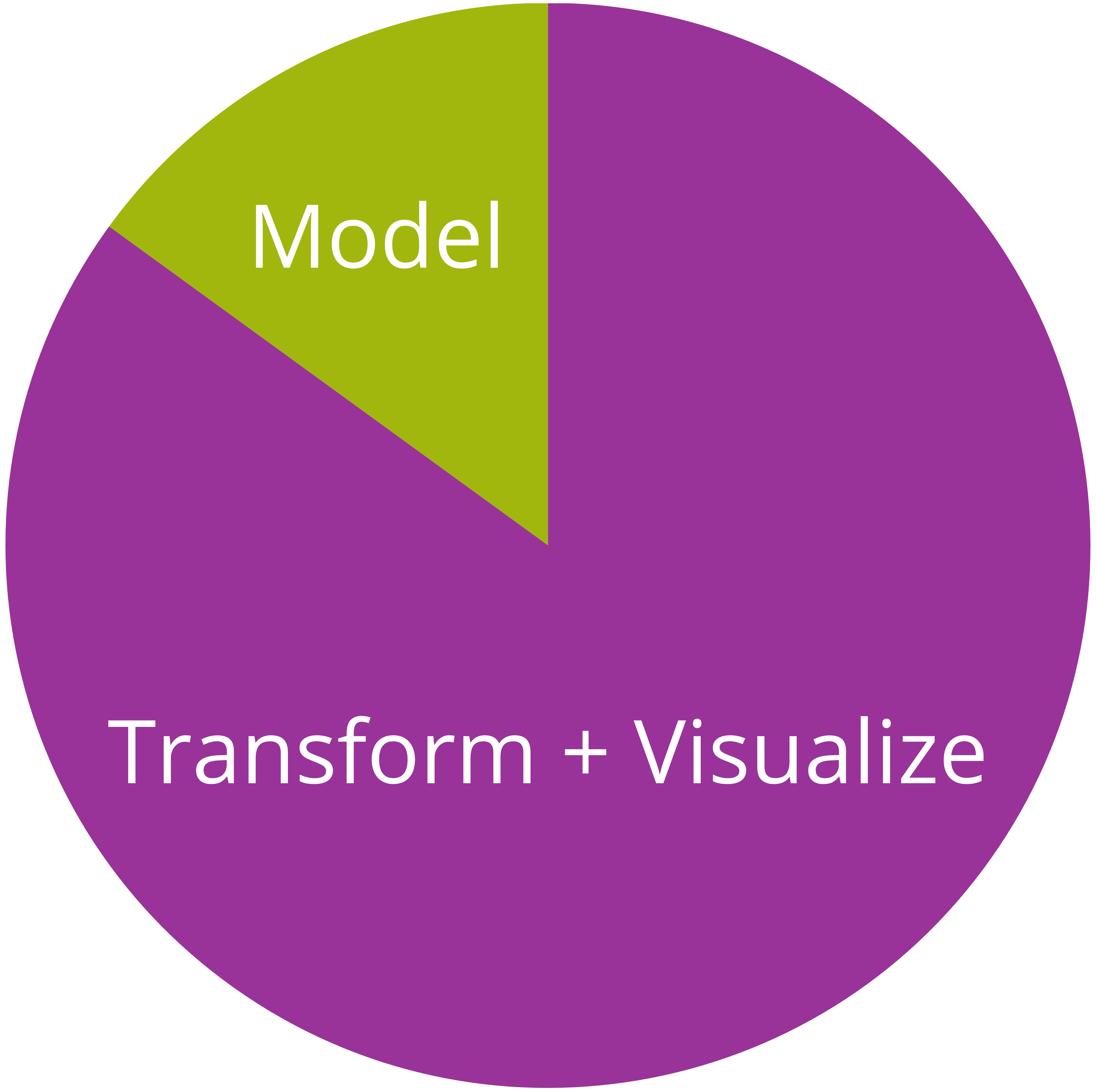
Computing using R and GitHub

Data science life cycle

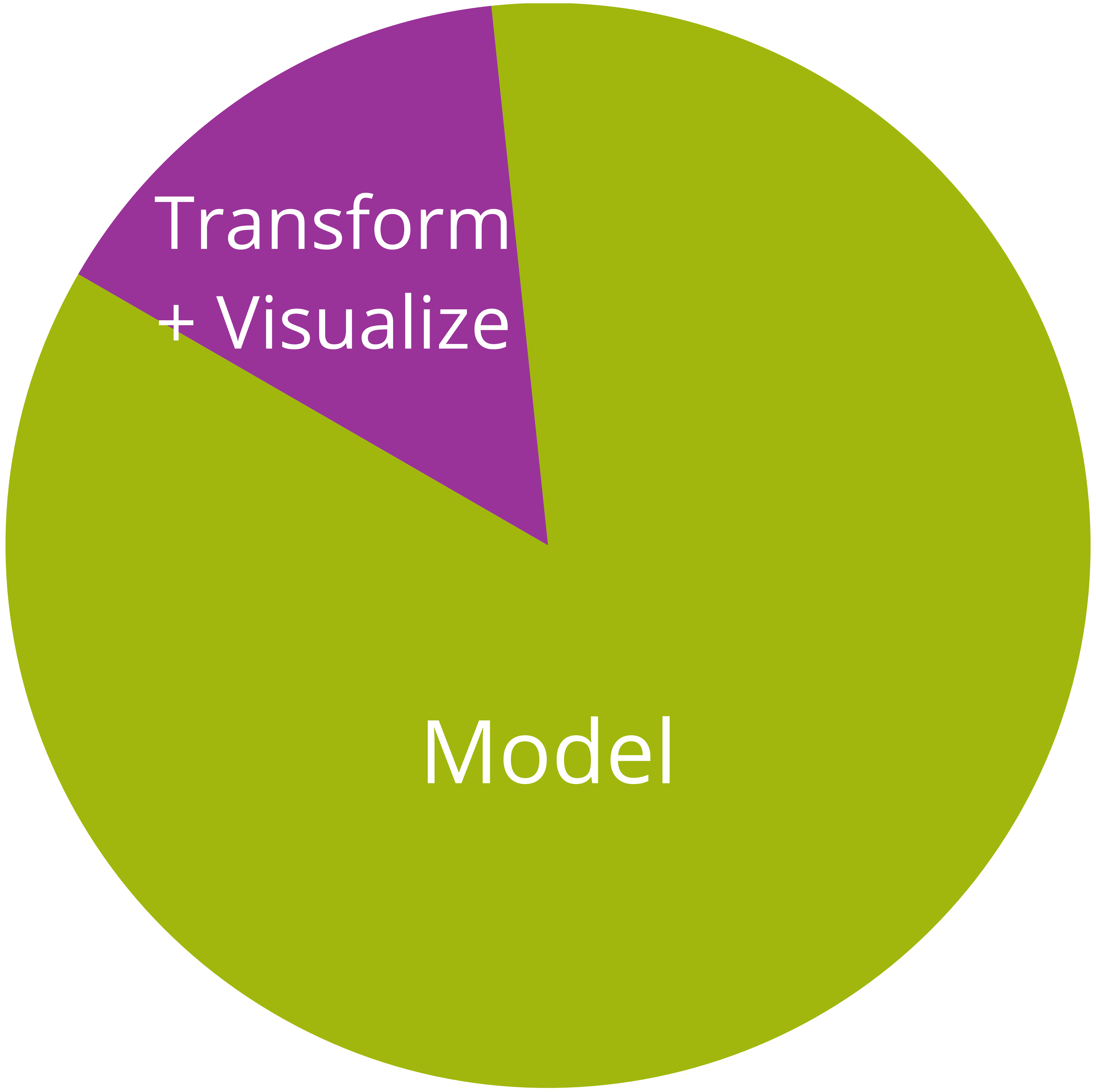


Data Science Life Cycle from R for Data Science

In practice



In class



Dealing with missing data

✓Identify different types of missingness

✓Use simple imputation methods to handle item nonresponse

✓***Think critically about unit nonresponse***

- ***Who is missing***
- ***Impact on analysis and conclusions***

Census 2020 (there's still time to fill it out!)

United States®
Census
2020

2020census.gov

- Headcount of every person living in the United States
- Occurs every 10 years
- Data is used to allocate…

  ✓seats in U.S. House of Representatives

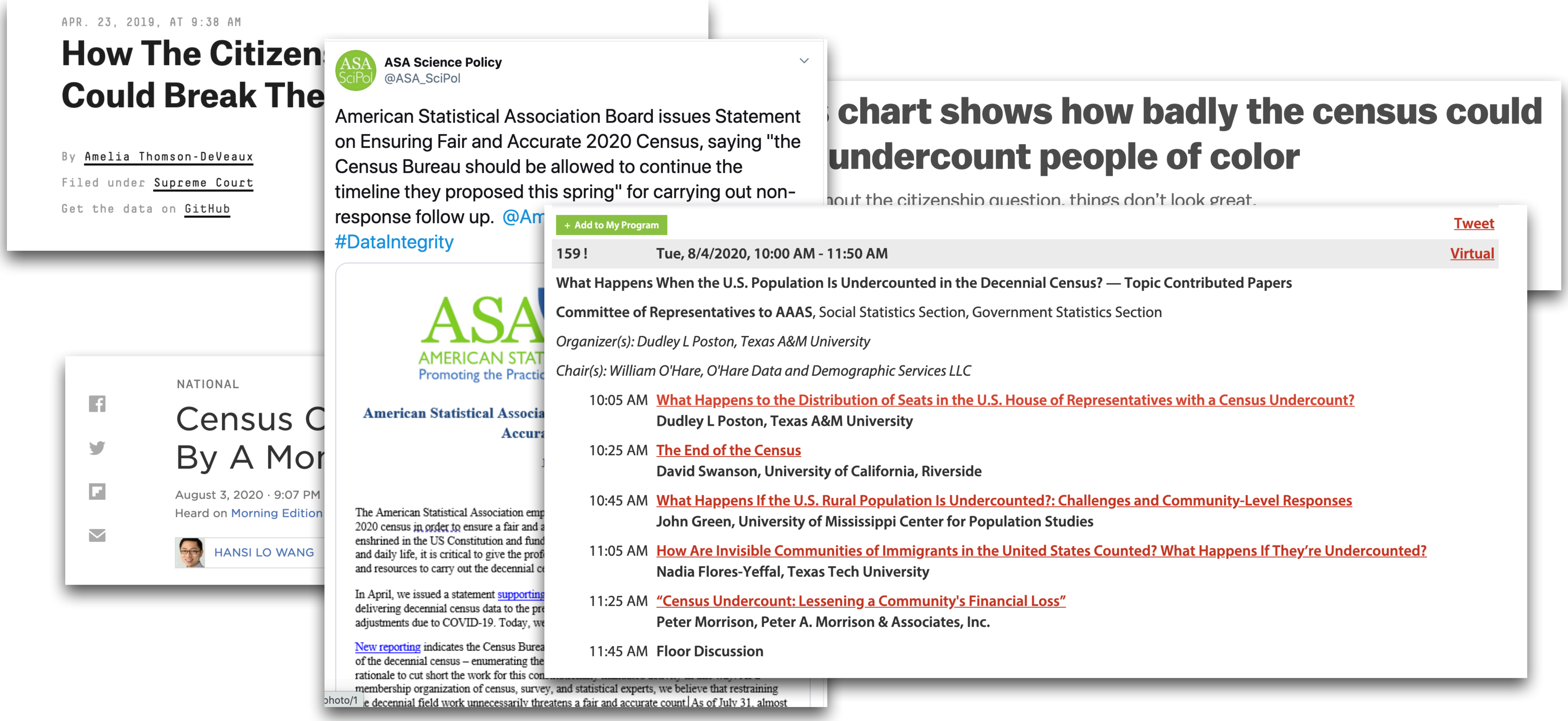
  ✓federal funding for public programs

Why use census data in class?

> *“Using **real data in context** is crucial in teaching and learning statistics, both to give students experience with analyzing genuine data and to illustrate the usefulness and fascination of our discipline.”*

2016 Guidelines for Assessment and Instruction in Statistics Education (GAISE)

People are talking about it!



APR. 23, 2019, AT 9:38 AM

How The Citizen

Could Break The

By Amelia Thomson-DeVeaux

Filed under Supreme Court

Get the data on GitHub

ASA Science Policy
@ASA_SciPol

American Statistical Association Board issues Statement on Ensuring Fair and Accurate 2020 Census, saying "the Census Bureau should be allowed to continue the timeline they proposed this spring" for carrying out non-response follow up. @Am

#DataIntegrity

ASA
AMERICAN STAT
Promoting the Practic

American Statistical Associa
Accura

The American Statistical Association emp
2020 census in order to ensure a fair and a
enshrined in the US Constitution and fund
and daily life, it is critical to give the prof
and resources to carry out the decennial ce

In April, we issued a statement supporting
delivering decennial census data to the pre
adjustments due to COVID-19. Today, we

New reporting indicates the Census Burea
of the decennial census – enumerating the
rationale to cut short the work for this con
membership organization of census, survey, and statistical experts, we believe that restraining
e decennial field work unnecessarily threatens a fair and accurate count. As of July 31, almost

photo/1

chart shows how badly the census could undercount people of color

out the citizenship question, things don't look great.

NATIONAL

Census C
By A Mor

August 3, 2020 · 9:07 PM
Heard on Morning Edition

HANSI LO WANG

+ Add to My Program

Tweet

159 ! Tue, 8/4/2020, 10:00 AM - 11:50 AM

Virtual

What Happens When the U.S. Population Is Undercounted in the Decennial Census? — Topic Contributed Papers

Committee of Representatives to AAAS, Social Statistics Section, Government Statistics Section

Organizer(s): Dudley L Poston, Texas A&M University

Chair(s): William O'Hare, O'Hare Data and Demographic Services LLC

10:05 AM What Happens to the Distribution of Seats in the U.S. House of Representatives with a Census Undercount?
Dudley L Poston, Texas A&M University

10:25 AM The End of the Census
David Swanson, University of California, Riverside

10:45 AM What Happens If the U.S. Rural Population Is Undercounted?: Challenges and Community-Level Responses
John Green, University of Mississippi Center for Population Studies

11:05 AM How Are Invisible Communities of Immigrants in the United States Counted? What Happens If They're Undercounted?
Nadia Flores-Yeffal, Texas Tech University

11:25 AM "Census Undercount: Lessening a Community's Financial Loss"
Peter Morrison, Peter A. Morrison & Associates, Inc.

11:45 AM Floor Discussion

Larry the Cat Retweeted

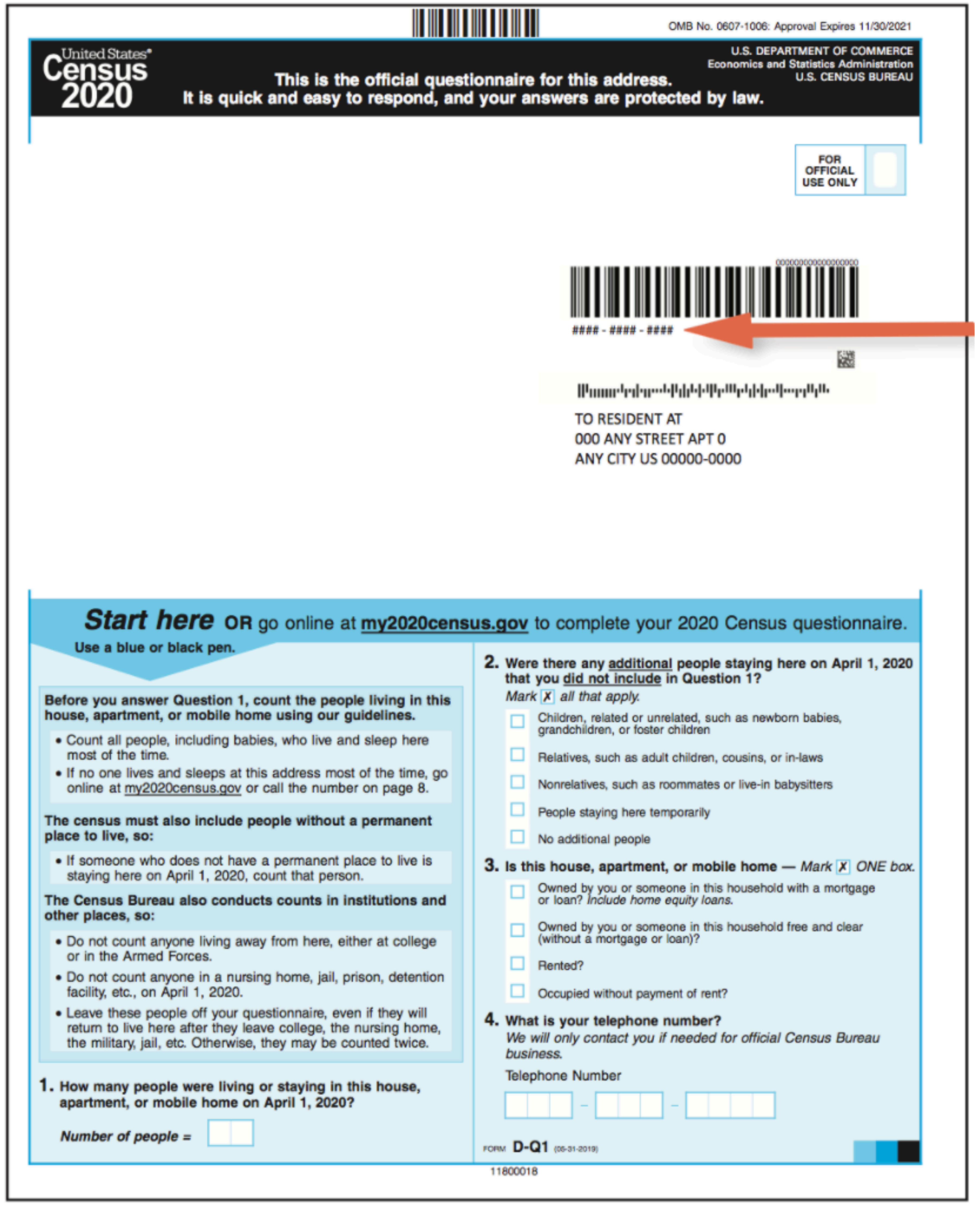
Larry the Cat
@Number10cat

My plans: 2020:

5:16 PM · May 19, 2020 · Twitter for iPhone

Opportunity to include more discussion and reflection in our classes!

Collecting data for the census



OMB No. 0607-1006: Approval Expires 11/30/2021

United States® Census 2020

This is the official questionnaire for this address.
It is quick and easy to respond, and your answers are protected by law.

U.S. DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. CENSUS BUREAU

FOR OFFICIAL USE ONLY

- #### -

TO RESIDENT AT
000 ANY STREET APT 0
ANY CITY US 00000-0000

Start here OR go online at my2020census.gov to complete your 2020 Census questionnaire.

Use a blue or black pen.

Before you answer Question 1, count the people living in this house, apartment, or mobile home using our guidelines.

- Count all people, including babies, who live and sleep here most of the time.
- If no one lives and sleeps at this address most of the time, go online at my2020census.gov or call the number on page 8.

The census must also include people without a permanent place to live, so:

- If someone who does not have a permanent place to live is staying here on April 1, 2020, count that person.

The Census Bureau also conducts counts in institutions and other places, so:

- Do not count anyone living away from here, either at college or in the Armed Forces.
- Do not count anyone in a nursing home, jail, prison, detention facility, etc., on April 1, 2020.
- Leave these people off your questionnaire, even if they will return to live here after they leave college, the nursing home, the military, jail, etc. Otherwise, they may be counted twice.

1. How many people were living or staying in this house, apartment, or mobile home on April 1, 2020?

Number of people = □□

2. Were there any additional people staying here on April 1, 2020 that you did not include in Question 1?
Mark ☒ all that apply.

□ Children, related or unrelated, such as newborn babies, grandchildren, or foster children
□ Relatives, such as adult children, cousins, or in-laws
□ Nonrelatives, such as roommates or live-in babysitters
□ People staying here temporarily
□ No additional people

3. Is this house, apartment, or mobile home — *Mark ☒ ONE box.*

□ Owned by you or someone in this household with a mortgage or loan? *Include home equity loans.*
□ Owned by you or someone in this household free and clear (without a mortgage or loan)?
□ Rented?
□ Occupied without payment of rent?

4. What is your telephone number?
We will only contact you if needed for official Census Bureau business.

Telephone Number

□□□ - □□□ - □□□□

FORM D-Q1 (05-31-2019)

11800018

- Invitation sent to households in March
- Respond by mail, phone, or online
- Door knocking effort in August to interview those who haven’t responded

Image source: https://my2020census.gov/static/letter.html?lang=en

Data collection discussion

- What populations are most likely to be *hard to count*? Why?

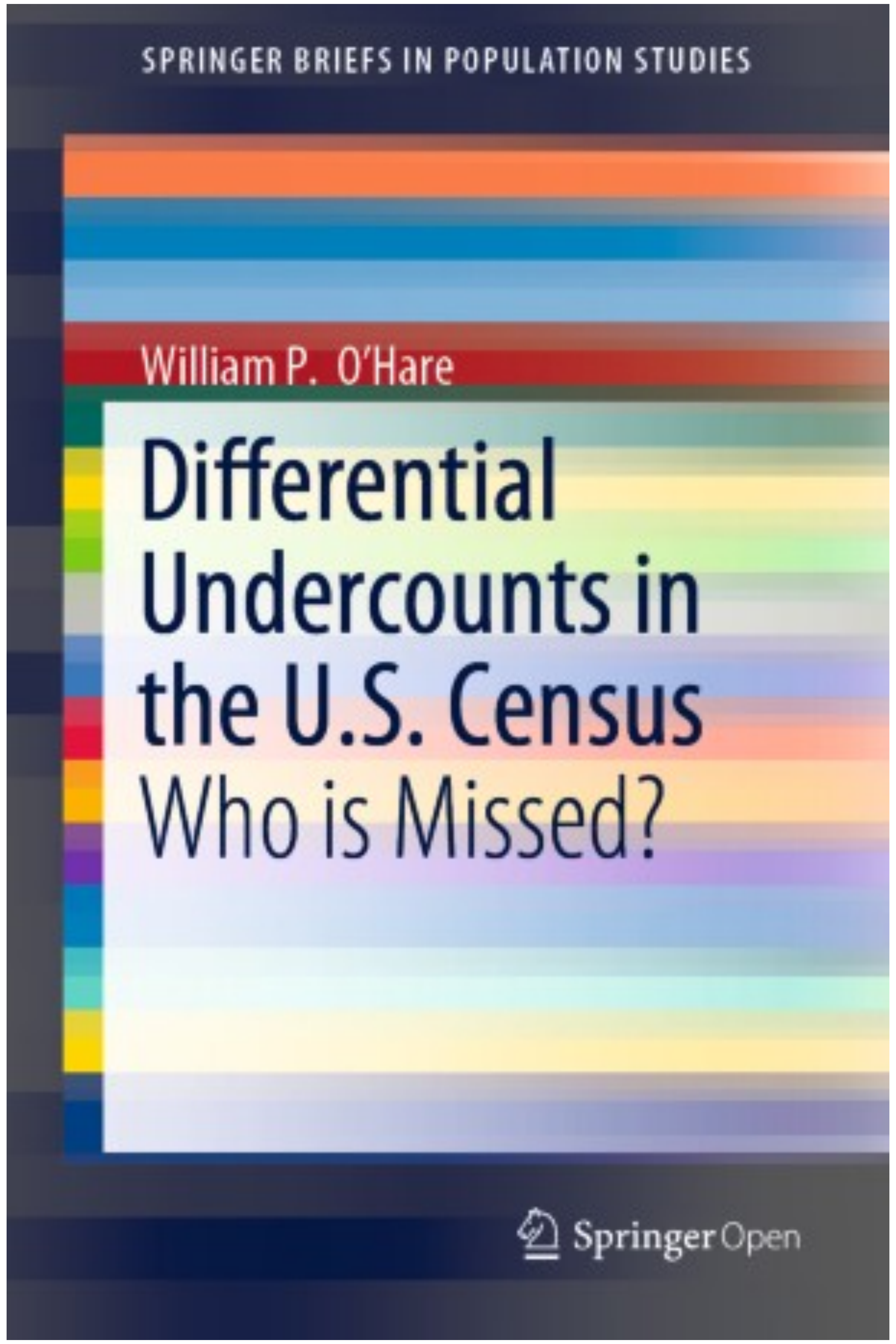
- What are the potential impacts of having underrepresented subgroups in the data when using census data to…
  - allocate funds or make other societal decisions?
  - conduct statistical analysis?

Measuring undercount

Demographic Analysis (DA): Compare census counts to a independent population estimate

$$Pop_{0-74} = Births - Deaths + NetMig$$

Duel System Estimates (DSE): Compare census counts to results from a Post-Enumeration Survey (PES)



SPRINGER BRIEFS IN POPULATION STUDIES

William P. O'Hare

Differential Undercounts in the U.S. Census

Who is Missed?

Springer Open

Differential Undercounts in the U.S. Census

Whole-person imputations

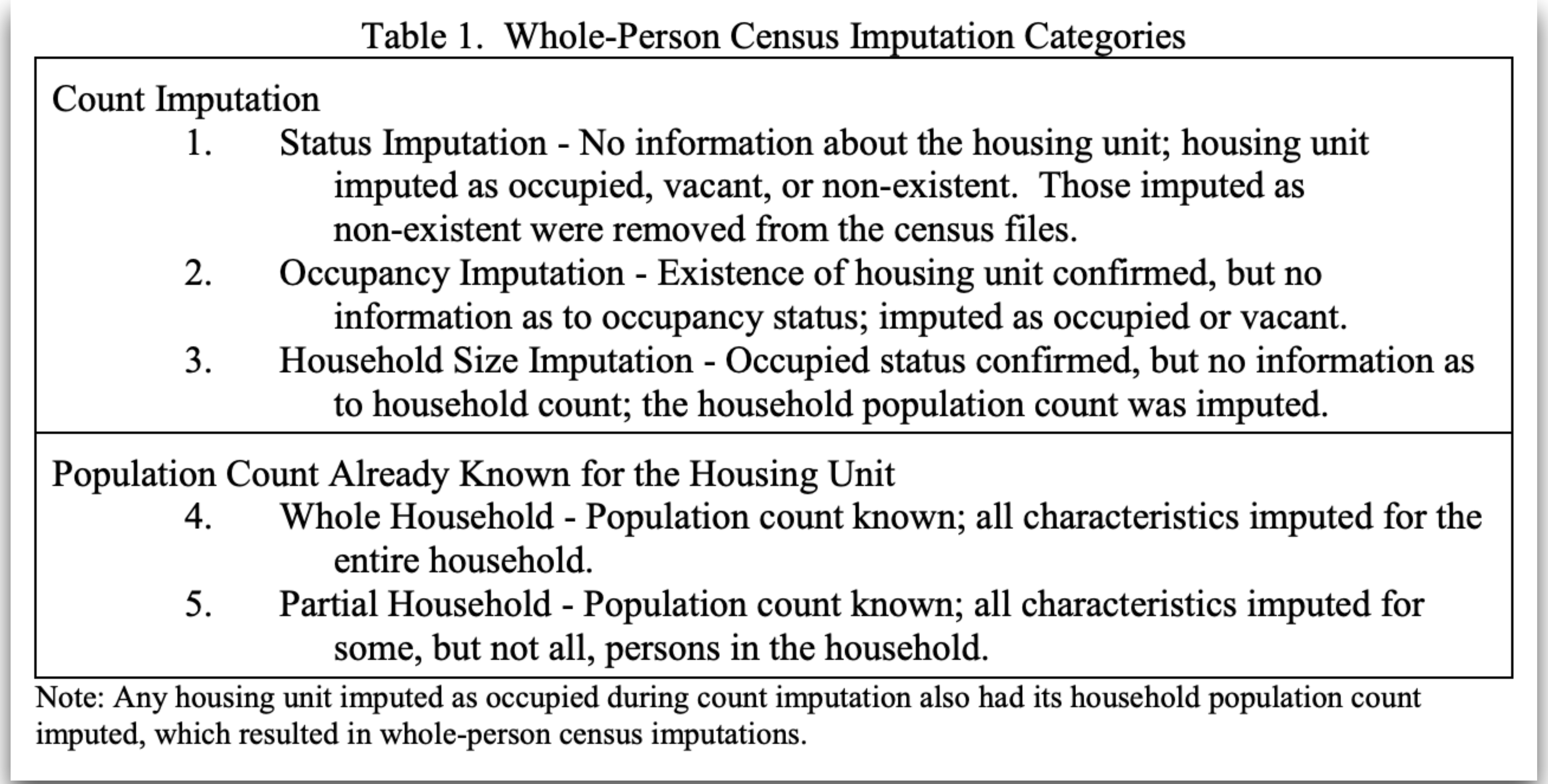


Table 1. Whole-Person Census Imputation Categories

Count Imputation

1. Status Imputation - No information about the housing unit; housing unit imputed as occupied, vacant, or non-existent. Those imputed as non-existent were removed from the census files.
2. Occupancy Imputation - Existence of housing unit confirmed, but no information as to occupancy status; imputed as occupied or vacant.
3. Household Size Imputation - Occupied status confirmed, but no information as to household count; the household population count was imputed.

Population Count Already Known for the Housing Unit

4. Whole Household - Population count known; all characteristics imputed for the entire household.
5. Partial Household - Population count known; all characteristics imputed for some, but not all, persons in the household.

Note: Any housing unit imputed as occupied during count imputation also had its household population count imputed, which resulted in whole-person census imputations.

Source: DSSD 2010 CENSUS COVERAGE MEASUREMENT MEMORANDUM SERIES #2010-G-01

Overall percent net undercount

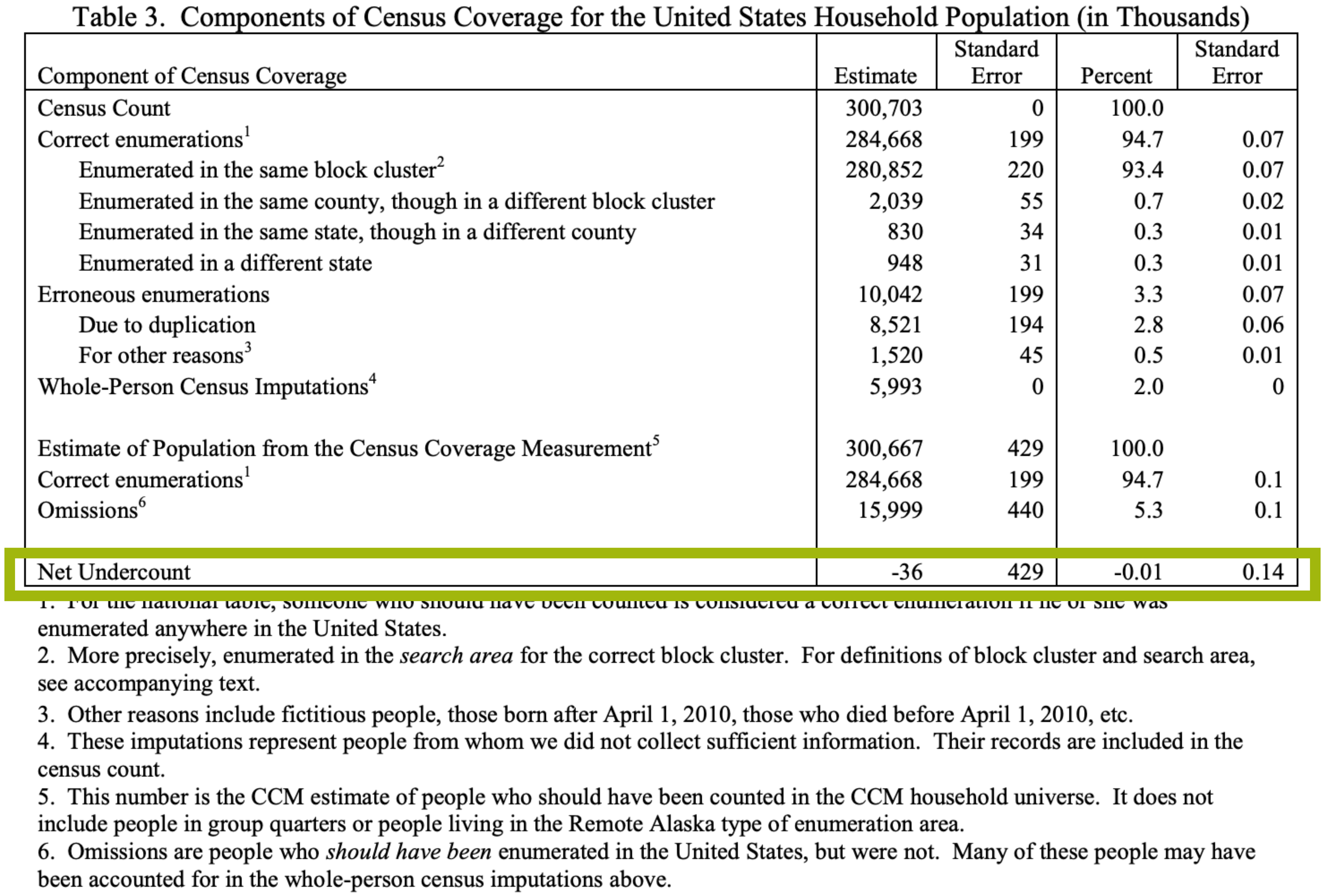


Table 3. Components of Census Coverage for the United States Household Population (in Thousands)

Component of Census Coverage	Estimate	Standard Error	Percent	Standard Error
Census Count	300,703	0	100.0	
Correct enumerations[1]	284,668	199	94.7	0.07
Enumerated in the same block cluster[2]	280,852	220	93.4	0.07
Enumerated in the same county, though in a different block cluster	2,039	55	0.7	0.02
Enumerated in the same state, though in a different county	830	34	0.3	0.01
Enumerated in a different state	948	31	0.3	0.01
Erroneous enumerations	10,042	199	3.3	0.07
Due to duplication	8,521	194	2.8	0.06
For other reasons[3]	1,520	45	0.5	0.01
Whole-Person Census Imputations[4]	5,993	0	2.0	0
Estimate of Population from the Census Coverage Measurement[5]	300,667	429	100.0	
Correct enumerations[1]	284,668	199	94.7	0.1
Omissions[6]	15,999	440	5.3	0.1
Net Undercount	-36	429	-0.01	0.14

1. For the national table, someone who should have been counted is considered a correct enumeration if he or she was enumerated anywhere in the United States.
2. More precisely, enumerated in the *search area* for the correct block cluster. For definitions of block cluster and search area, see accompanying text.
3. Other reasons include fictitious people, those born after April 1, 2010, those who died before April 1, 2010, etc.
4. These imputations represent people from whom we did not collect sufficient information. Their records are included in the census count.
5. This number is the CCM estimate of people who should have been counted in the CCM household universe. It does not include people in group quarters or people living in the Remote Alaska type of enumeration area.
6. Omissions are people who *should have been* enumerated in the United States, but were not. Many of these people may have been accounted for in the whole-person census imputations above.

$$\% \text{ Net Undercount} = \frac{\text{DSE} - \text{Census}}{\text{DSE}} \times 100$$

🙂

Overall things look great!

Source: DSSD 2010 CENSUS COVERAGE MEASUREMENT MEMORANDUM SERIES #2010-G-01

but subgroups matter….

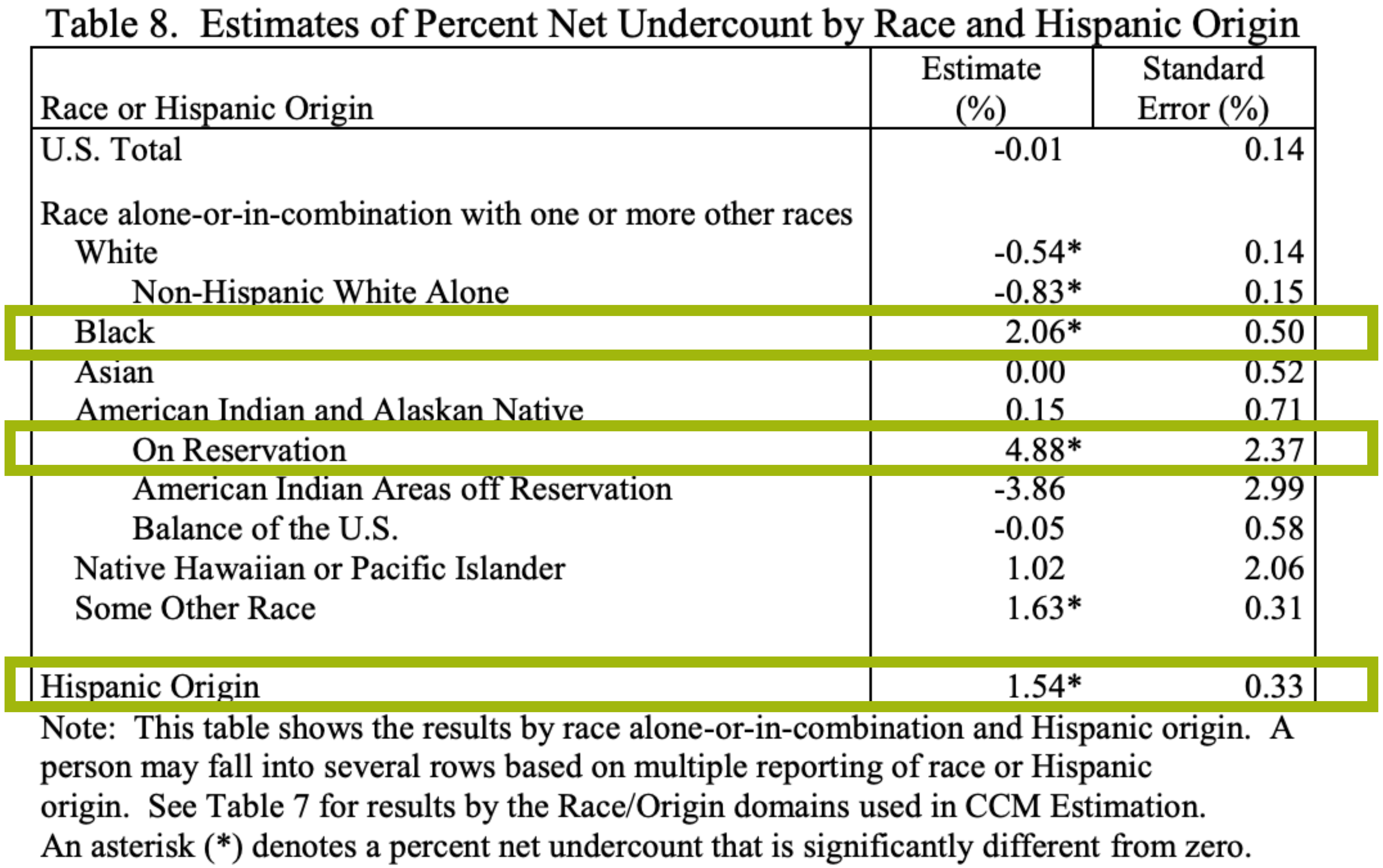


Table 8. Estimates of Percent Net Undercount by Race and Hispanic Origin

Race or Hispanic Origin	Estimate (%)	Standard Error (%)
U.S. Total	-0.01	0.14
Race alone-or-in-combination with one or more other races		
White	-0.54*	0.14
Non-Hispanic White Alone	-0.83*	0.15
Black	2.06*	0.50
Asian	0.00	0.52
American Indian and Alaskan Native	0.15	0.71
On Reservation	4.88*	2.37
American Indian Areas off Reservation	-3.86	2.99
Balance of the U.S.	-0.05	0.58
Native Hawaiian or Pacific Islander	1.02	2.06
Some Other Race	1.63*	0.31
Hispanic Origin	1.54*	0.33

Note: This table shows the results by race alone-or-in-combination and Hispanic origin. A person may fall into several rows based on multiple reporting of race or Hispanic origin. See Table 7 for results by the Race/Origin domains used in CCM Estimation. An asterisk (*) denotes a percent net undercount that is significantly different from zero.

$$\%\ \text{Net Undercount} = \frac{\text{DSE} - \text{Census}}{\text{DSE}} \times 100$$

😕

Not as great as I thought…

Source: DSSD 2010 CENSUS COVERAGE MEASUREMENT MEMORANDUM SERIES #2010-G-01

Discussion

Suppose you work for a major grocery store, and your team wants to use regression models to help determine how to stock shelves with goods that suit the local customers’ preferences.

What are the advantages of using data from the U.S. Census to build these models? What are the limitations?

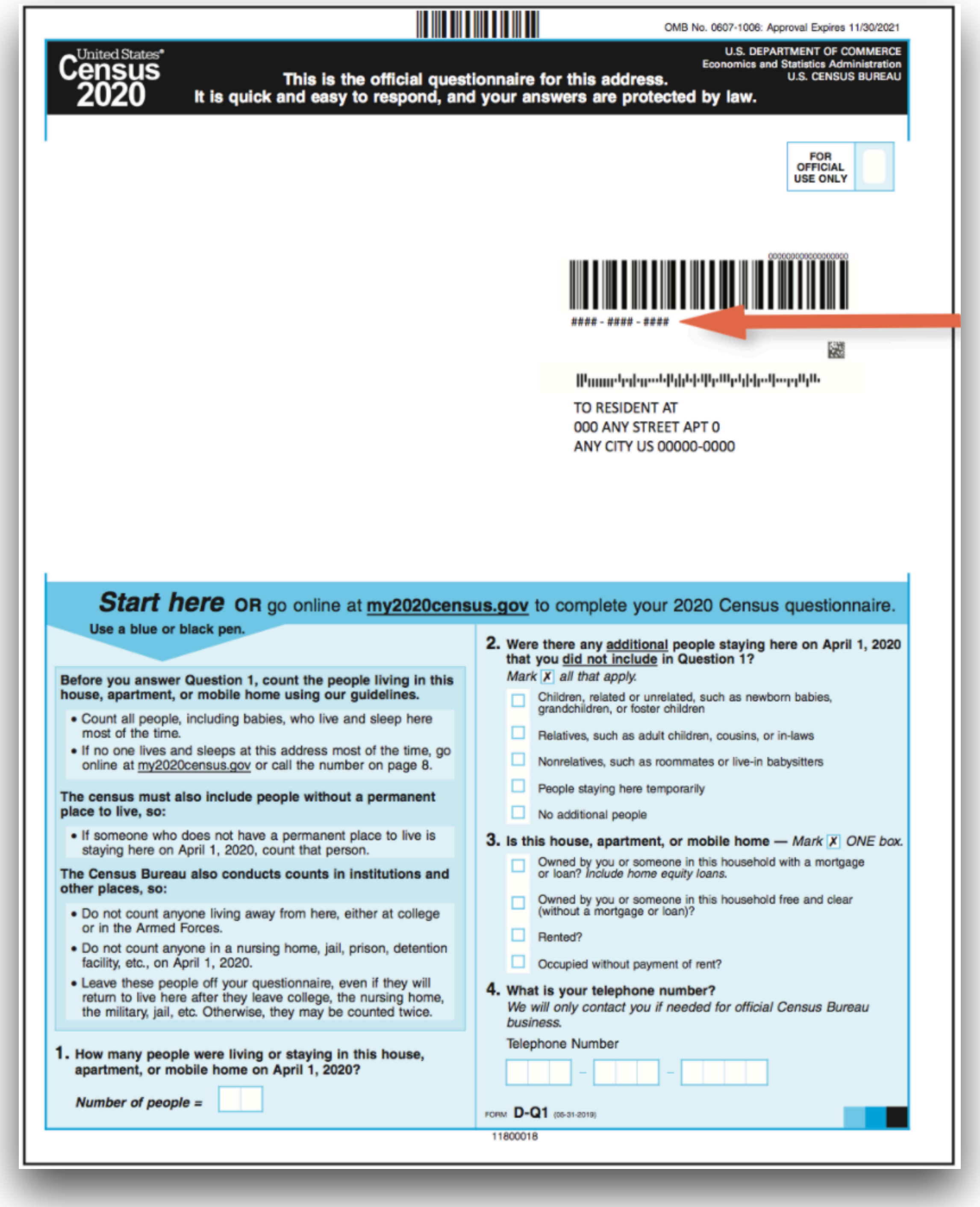
Discussion

Suppose you work for the Department of Education, and your team wants to use regression models to determine where to invest funding in new education initiatives.

What are the advantages of using data from the U.S. Census to build these models? What are the limitations?

What we've learned!

Data we think we're using



OMB No. 0607-1006: Approval Expires 11/30/2021

United States® Census 2020

U.S. DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. CENSUS BUREAU

This is the official questionnaire for this address.
It is quick and easy to respond, and your answers are protected by law.

FOR OFFICIAL USE ONLY

#### - #### - ####

TO RESIDENT AT
000 ANY STREET APT 0
ANY CITY US 00000-0000

Start here OR go online at my2020census.gov to complete your 2020 Census questionnaire.

Use a blue or black pen.

Before you answer Question 1, count the people living in this house, apartment, or mobile home using our guidelines.

- Count all people, including babies, who live and sleep here most of the time.
- If no one lives and sleeps at this address most of the time, go online at my2020census.gov or call the number on page 8.

The census must also include people without a permanent place to live, so:

- If someone who does not have a permanent place to live is staying here on April 1, 2020, count that person.

The Census Bureau also conducts counts in institutions and other places, so:

- Do not count anyone living away from here, either at college or in the Armed Forces.
- Do not count anyone in a nursing home, jail, prison, detention facility, etc., on April 1, 2020.
- Leave these people off your questionnaire, even if they will return to live here after they leave college, the nursing home, the military, jail, etc. Otherwise, they may be counted twice.

1. **How many people were living or staying in this house, apartment, or mobile home on April 1, 2020?**

Number of people = □□

2. **Were there any additional people staying here on April 1, 2020 that you did not include in Question 1?**
Mark ☒ all that apply.
□ Children, related or unrelated, such as newborn babies, grandchildren, or foster children
□ Relatives, such as adult children, cousins, or in-laws
□ Nonrelatives, such as roommates or live-in babysitters
□ People staying here temporarily
□ No additional people

3. **Is this house, apartment, or mobile home** — *Mark ☒ ONE box.*
□ Owned by you or someone in this household with a mortgage or loan? *Include home equity loans.*
□ Owned by you or someone in this household free and clear (without a mortgage or loan)?
□ Rented?
□ Occupied without payment of rent?

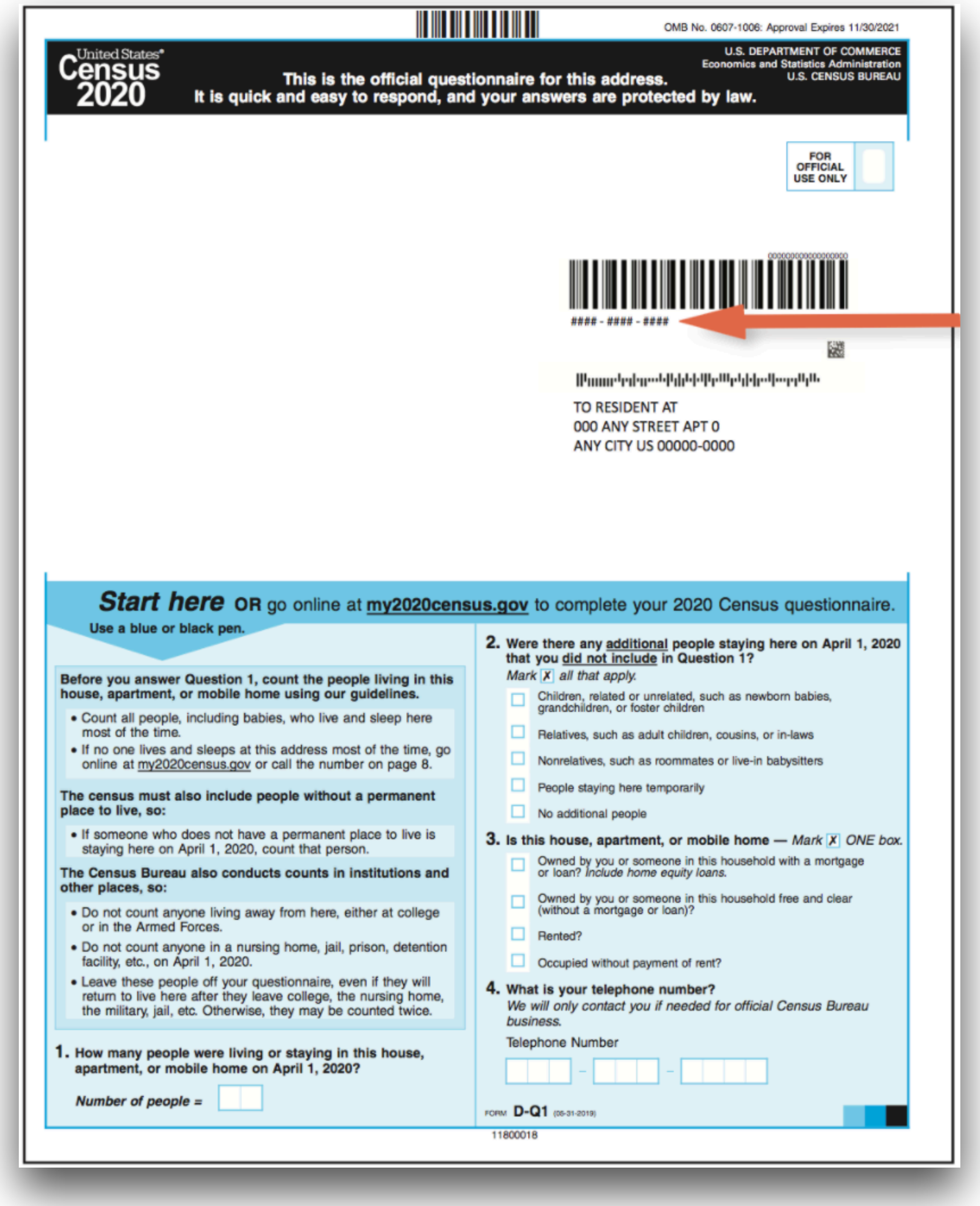
4. **What is your telephone number?**
We will only contact you if needed for official Census Bureau business.
Telephone Number
□□□ - □□□ - □□□□

FORM D-Q1 (05-31-2019)

11800018

It's a census!
Everyone is counted!

Data we're actually using



OMB No. 0607-1006: Approval Expires 11/30/2021

United States® Census 2020

U.S. DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. CENSUS BUREAU

This is the official questionnaire for this address.
It is quick and easy to respond, and your answers are protected by law.

FOR OFFICIAL USE ONLY

#### - #### - ####

TO RESIDENT AT
000 ANY STREET APT 0
ANY CITY US 00000-0000

Start here OR go online at my2020census.gov to complete your 2020 Census questionnaire.

Use a blue or black pen.

Before you answer Question 1, count the people living in this house, apartment, or mobile home using our guidelines.

- Count all people, including babies, who live and sleep here most of the time.
- If no one lives and sleeps at this address most of the time, go online at my2020census.gov or call the number on page 8.

The census must also include people without a permanent place to live, so:

- If someone who does not have a permanent place to live is staying here on April 1, 2020, count that person.

The Census Bureau also conducts counts in institutions and other places, so:

- Do not count anyone living away from here, either at college or in the Armed Forces.
- Do not count anyone in a nursing home, jail, prison, detention facility, etc., on April 1, 2020.
- Leave these people off your questionnaire, even if they will return to live here after they leave college, the nursing home, the military, jail, etc. Otherwise, they may be counted twice.

1. **How many people were living or staying in this house, apartment, or mobile home on April 1, 2020?**

Number of people = □□

2. **Were there any additional people staying here on April 1, 2020 that you did not include in Question 1?**
Mark ☒ all that apply.
□ Children, related or unrelated, such as newborn babies, grandchildren, or foster children
□ Relatives, such as adult children, cousins, or in-laws
□ Nonrelatives, such as roommates or live-in babysitters
□ People staying here temporarily
□ No additional people

3. **Is this house, apartment, or mobile home** — *Mark ☒ ONE box.*
□ Owned by you or someone in this household with a mortgage or loan? *Include home equity loans.*
□ Owned by you or someone in this household free and clear (without a mortgage or loan)?
□ Rented?
□ Occupied without payment of rent?

4. **What is your telephone number?**
We will only contact you if needed for official Census Bureau business.
Telephone Number
□□□ - □□□ - □□□□

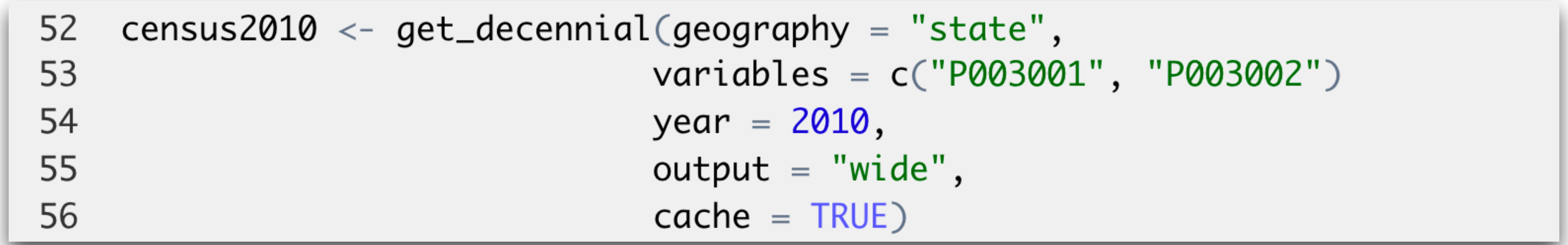
FORM D-Q1 (05-31-2019)

11800018

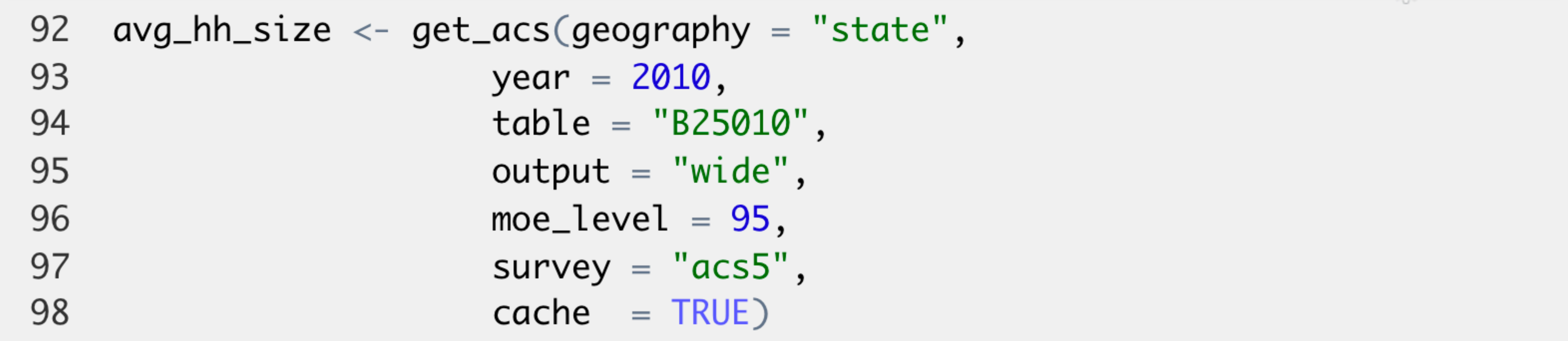
+ Imputations

Who is underrepresented?

tidycensus R package



```r
census2010 <- get_decennial(geography = "state",
                            variables = c("P003001", "P003002"),
                            year = 2010,
                            output = "wide",
                            cache = TRUE)
```

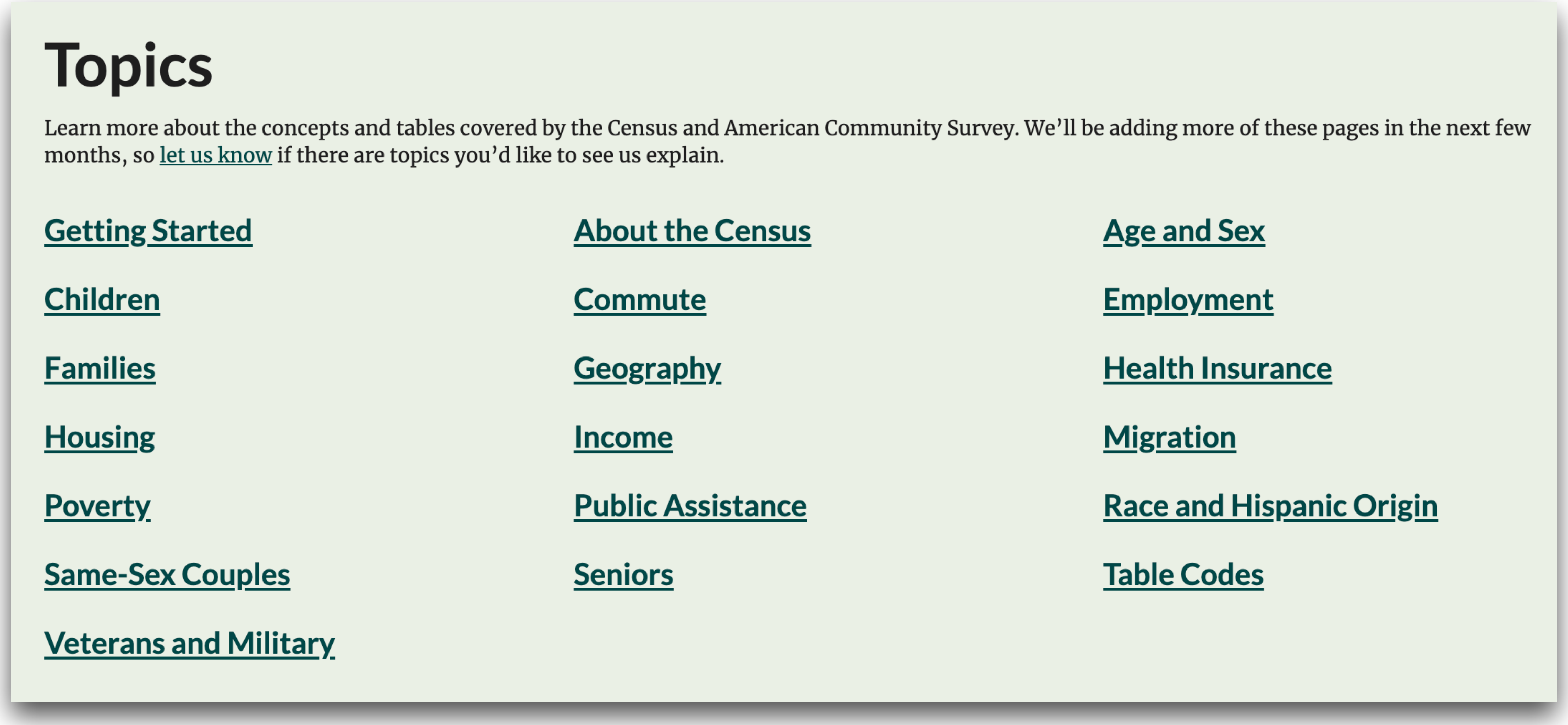


```r
avg_hh_size <- get_acs(geography = "state",
                       year = 2010,
                       table = "B25010",
                       output = "wide",
                       moe_level = 95,
                       survey = "acs5",
                       cache  = TRUE)
```

walker-data.com/tidycensus

Find data from census and ACS

censusreporter.org



Topics

Learn more about the concepts and tables covered by the Census and American Community Survey. We'll be adding more of these pages in the next few months, so let us know if there are topics you'd like to see us explain.

Getting Started

About the Census

Age and Sex

Children

Commute

Employment

Families

Geography

Health Insurance

Housing

Income

Migration

Poverty

Public Assistance

Race and Hispanic Origin

Same-Sex Couples

Seniors

Table Codes

Veterans and Military

💡 ***Provide data to students for short-term assignments.***

Modeling exercise

Suppose you're part of an organization whose goal is to reach people in hard-to-reach populations and encourage them to fill out the Census.

The organization has limited resources, so you will use data to help determine how to prioritize your time and effort.

Fit a regression model that you can use to describe how to prioritize your outreach efforts.

Response variable

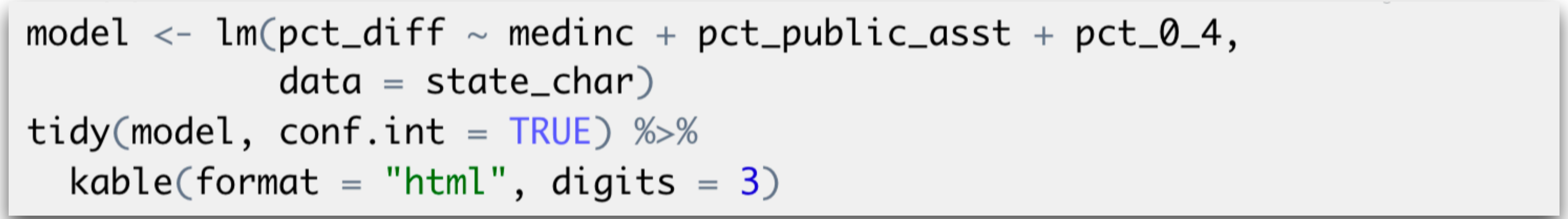
Estimate of population in 2010

$$Pop_{2010} = Pop_{2009} + Births_{2010} - Deaths_{2010} + NetMigration_{2010}$$

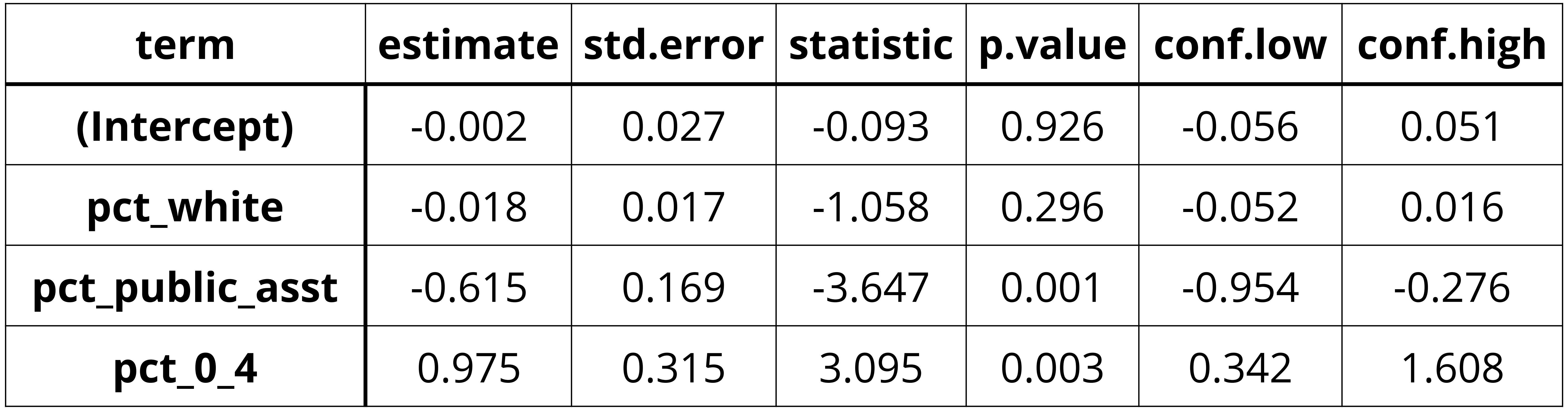
Use Pop_{2010} and the population from the 2010 Census to define a response variable.

Use data from the American Community Survey (ACS) for the explanatory variables.

Model + conclusion



```
model <- lm(pct_diff ~ medinc + pct_public_asst + pct_0_4,
            data = state_char)
tidy(model, conf.int = TRUE) %>%
  kable(format = "html", digits = 3)
```



term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.002	0.027	-0.093	0.926	-0.056	0.051
pct_white	-0.018	0.017	-1.058	0.296	-0.052	0.016
pct_public_asst	-0.615	0.169	-3.647	0.001	-0.954	-0.276
pct_0_4	0.975	0.315	3.095	0.003	0.342	1.608

Based on your model, describe how you will prioritize your efforts to encourage people to respond to the U.S. Census.

Reflection questions

- What is one observation from your model about undercount in the census? How does it compare to the results from the DA and DSE methods?
- Briefly explain why it is important to consider which subgroups are underrepresented in data used to build statistical models.
- What is one remaining question you have about the U.S. Census?
- What is one question you still have about missing data?

Thank You!

✉ maria.tackett@duke.edu

🐦 @MT_statistics