CEPR CENTER FOR ECONOMIC AND POLICY RESEARCH

Publicly Available Microdata:

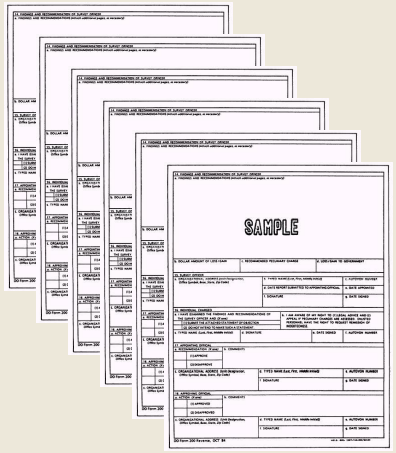
Why It's Useful and How to Use It in R

By Hayley Brown

Research Associate
Center for Economic and Policy Research

March 2021

What is microdata?



Summary data vs micro data

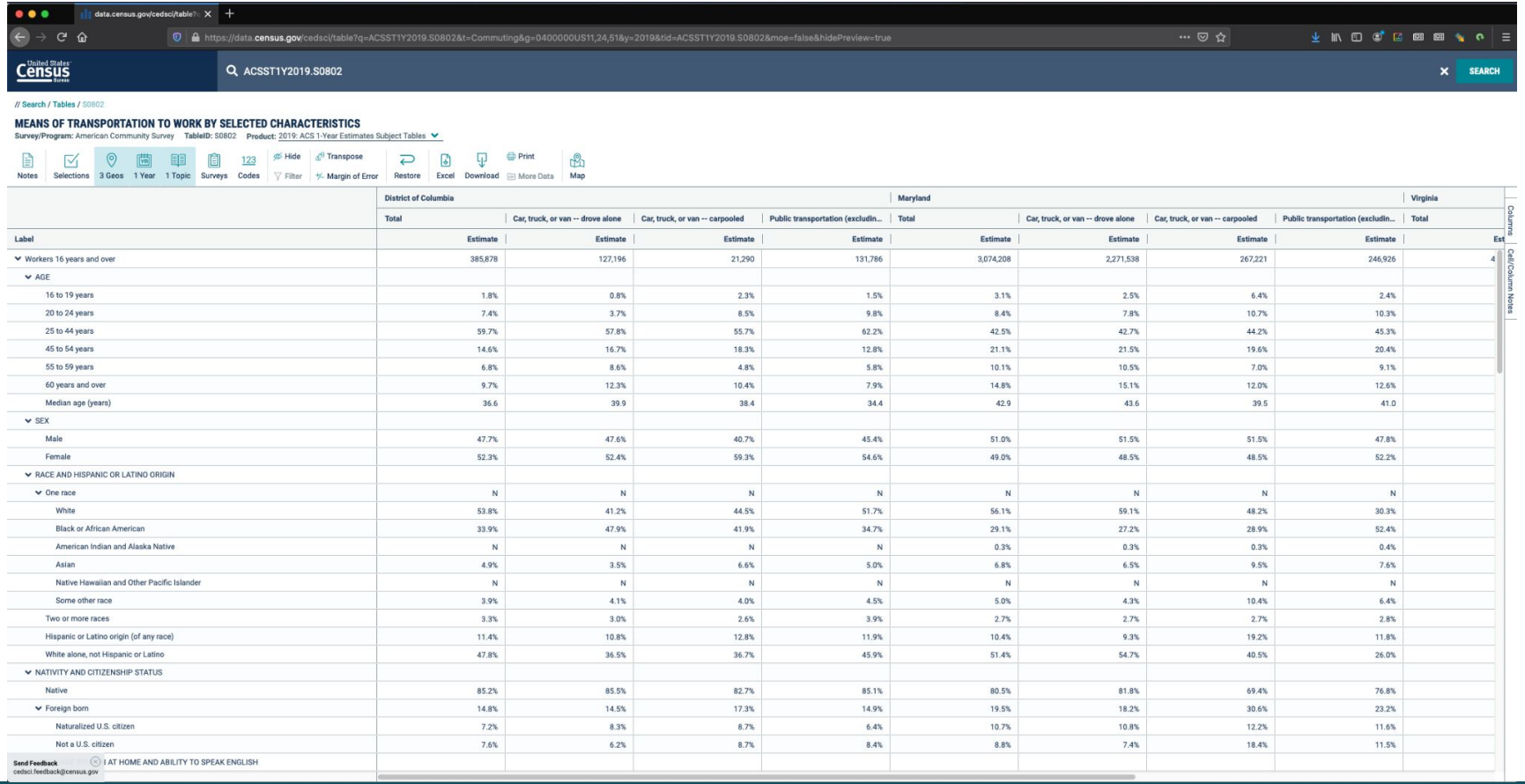
Aggregate data (summary data)

- Higher level data that results from grouping individual responses in a certain way
- Must balance data integrity with privacy protection
- May be available for smaller geographies, but at the expense of dimensionality
- Reduced dimensionality impedes certain types of analysis

Disaggregate data (***microdata***)

- Information at the level of individual responses
- Must balance data integrity with privacy protection
- Data may not be available for smaller subgroups or geographies
- Increased dimensionality and flexibility, well-suited to exploratory analysis

An example of summary data - Table S0802 from the Census website



ACSST1Y2019.S0802

// Search / Tables / S0802

MEANS OF TRANSPORTATION TO WORK BY SELECTED CHARACTERISTICS

Survey/Program: American Community Survey TableID: S0802 Product: 2019: ACS 1-Year Estimates Subject Tables

Label	District of Columbia				Maryland				Virginia
	Total	Car, truck, or van -- drove alone	Car, truck, or van -- carpooled	Public transportation (excludin...	Total	Car, truck, or van -- drove alone	Car, truck, or van -- carpooled	Public transportation (excludin...	Total
	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate	Est
Workers 16 years and over	385,878	127,196	21,290	131,786	3,074,208	2,271,538	267,221	246,926	4
AGE									
16 to 19 years	1.8%	0.8%	2.3%	1.5%	3.1%	2.5%	6.4%	2.4%	
20 to 24 years	7.4%	3.7%	8.5%	9.8%	8.4%	7.8%	10.7%	10.3%	
25 to 44 years	59.7%	57.8%	55.7%	62.2%	42.5%	42.7%	44.2%	45.3%	
45 to 54 years	14.6%	16.7%	18.3%	12.8%	21.1%	21.5%	19.6%	20.4%	
55 to 59 years	6.8%	8.6%	4.8%	5.8%	10.1%	10.5%	7.0%	9.1%	
60 years and over	9.7%	12.3%	10.4%	7.9%	14.8%	15.1%	12.0%	12.6%	
Median age (years)	36.6	39.9	38.4	34.4	42.9	43.6	39.5	41.0	
SEX									
Male	47.7%	47.6%	40.7%	45.4%	51.0%	51.5%	51.5%	47.8%	
Female	52.3%	52.4%	59.3%	54.6%	49.0%	48.5%	48.5%	52.2%	
RACE AND HISPANIC OR LATINO ORIGIN									
One race	N	N	N	N	N	N	N	N	
White	53.8%	41.2%	44.5%	51.7%	56.1%	59.1%	48.2%	30.3%	
Black or African American	33.9%	47.9%	41.9%	34.7%	29.1%	27.2%	28.9%	52.4%	
American Indian and Alaska Native	N	N	N	N	0.3%	0.3%	0.3%	0.4%	
Asian	4.9%	3.5%	6.6%	5.0%	6.8%	6.5%	9.5%	7.6%	
Native Hawaiian and Other Pacific Islander	N	N	N	N	N	N	N	N	
Some other race	3.9%	4.1%	4.0%	4.5%	5.0%	4.3%	10.4%	6.4%	
Two or more races	3.3%	3.0%	2.6%	3.9%	2.7%	2.7%	2.7%	2.8%	
Hispanic or Latino origin (of any race)	11.4%	10.8%	12.8%	11.9%	10.4%	9.3%	19.2%	11.8%	
White alone, not Hispanic or Latino	47.8%	36.5%	36.7%	45.9%	51.4%	54.7%	40.5%	26.0%	
NATIVITY AND CITIZENSHIP STATUS									
Native	85.2%	85.5%	82.7%	85.1%	80.5%	81.8%	69.4%	76.8%	
Foreign born	14.8%	14.5%	17.3%	14.9%	19.5%	18.2%	30.6%	23.2%	
Naturalized U.S. citizen	7.2%	8.3%	8.7%	6.4%	10.7%	10.8%	12.2%	11.6%	
Not a U.S. citizen	7.6%	6.2%	8.7%	8.4%	8.8%	7.4%	18.4%	11.5%	
I AT HOME AND ABILITY TO SPEAK ENGLISH									

Summary data example - American Community Survey

MEANS OF TRANSPORTATION TO WORK BY SELECTED CHARACTERISTICS

Survey/Program: American Community Survey **TableID:** S0802 **Product:** 2019: ACS 1-Year Estimates Subject Tables

Notes | Selections | 3 Geos | 1 Year | 1 Topic | Surveys | Codes | Hide | Filter | Transpose | Margin of Error | Restore | Excel | Download | Print | More Data | Map

	District of Columbia			
	Total	Car, truck, or van -- drove alone	Car, truck, or van -- carpooled	Public transportation (excludin...
Label	Estimate	Estimate	Estimate	Estimate
Workers 16 years and over	385,878	127,196	21,290	131,786
AGE				
16 to 19 years	1.8%	0.8%	2.3%	1.5%
20 to 24 years	7.4%	3.7%	8.5%	9.8%
25 to 44 years	59.7%	57.8%	55.7%	62.2%
45 to 54 years	14.6%	16.7%	18.3%	12.8%
55 to 59 years	6.8%	8.6%	4.8%	5.8%
60 years and over	9.7%	12.3%	10.4%	7.9%
Median age (years)	36.6	39.9	38.4	34.4

Summary data example - American Community Survey

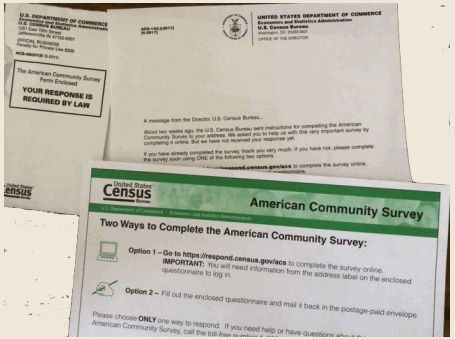
	District of Columbia			
	Total	Car, truck, or van -- drove alone	Car, truck, or van -- carpooled	Public transportation (excludin...
Label	Estimate	Estimate	Estimate	Estimate
Workers 16 years and over	385,878	127,196	21,290	131,786
AGE				
SEX				
RACE AND HISPANIC OR LATINO ORIGIN				
One race	N	N	N	N
White	53.8%	41.2%	44.5%	51.7%
Black or African American	33.9%	47.9%	41.9%	34.7%
American Indian and Alaska Native	N	N	N	N
Asian	4.9%	3.5%	6.6%	5.0%
Native Hawaiian and Other Pacific Islander	N	N	N	N
Some other race	3.9%	4.1%	4.0%	4.5%
Two or more races	3.3%	3.0%	2.6%	3.9%
Hispanic or Latino origin (of any race)	11.4%	10.8%	12.8%	11.9%
White alone, not Hispanic or Latino	47.8%	36.5%	36.7%	45.9%

Summary data example - American Community Survey

What if we want to know how many Black people age 50 and older take public transit to work?

We need to use microdata!

Microdata from publicly available surveys



Federally administered household surveys in the US

An excellent use of federal tax dollars

- Survey data is different from administrative data
- Substantial variation in scope, survey type, sample size, administration (frequency, means of contact)
- Typically include both household and person level items

Some examples of US household surveys

- **Current Population Survey (CPS):** *Primary source of US labor force statistics; basic survey conducted monthly, supplemental surveys administered less frequently*
- **American Community Survey (ACS):** *Annual survey that replaced the long-form census, largest household survey administered by the US Census Bureau*
- **National Health Interview Survey (NHIS):** *Annual survey encompassing a wide range of health status and health care utilization measures*
- **American Time Use Survey (ATUS):** *Annual survey measuring how much time respondents spend on various activities*

A few more examples of US household surveys

- **Household Pulse Survey (HPS):** *New longitudinal survey designed to measure socioeconomic impact of the COVID-19 pandemic; administered online*
- **Survey of Income and Program Participation (SIPP):** *Continuous national panel survey with detailed information on household income dynamics and participation in government transfer programs*
- **American Housing Survey (AHS):** *Largest regular national housing sample survey in the US, conducted every other year*
- **...and many more!**

Considerations when deciding which survey to use

- **Population of interest:** *Geography, age, labor force status, housing type, etc.*
- **Available variables:** *Does the survey contain the variables I need to perform my analysis?*
- **Sample size:** *Larger samples needed to study smaller subgroups and geographies, and to perform certain types of statistical analysis (e.g. matching)*
- **Survey type:** *How important is longitudinal data to your analysis?*
- **Availability and ease of use:** *Is there a cleaned up version available or will you need to clean the raw data yourself?*

Accessing Public-Use Microdata for Household Surveys

Raw data vs cleaned extracts

Raw public-use microdata

- Tends to require a lot of cleaning, especially to harmonize variables that undergo changes between sample years
- Some datasets are not available any other way
- The data dictionary is your best friend

Cleaned extracts

- Raw government data has been processed, cleaned, and harmonized by a third party
- May contain additional variables based on pre-existing ones; important to be aware of edits made to underlying data
- Typically much easier to use

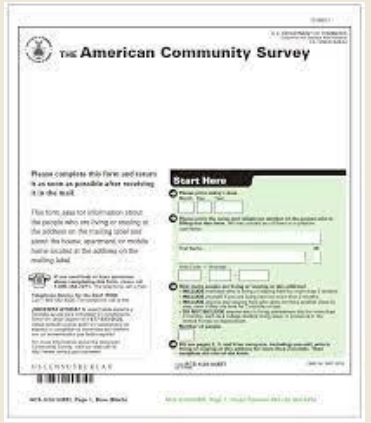
Accessing raw public-use microdata

- **Agency websites:** *You can typically find the public-use microdata online via the agency or agencies that administer/ sponsor the survey*
- **Limited API availability for now**
- **For some datasets and sample years, increased customization for downloads**
    - *This is especially useful for larger datasets like ACS*
    - *Some data is still only available via FTP, which means you have to download the whole thing before you can work with it*

Some sources of US household survey microdata extracts

- **IPUMS:** *Large-scale project that allows users to download customized extracts; incredibly comprehensive, with many additional useful variables, including geographic crosswalks. They also have their own R package, ipumsr.*
- **CEPRdata:** *Ongoing project of CEPR, limited number of datasets but working on improvements; program files used to create extracts are available but are in Stata*
- **EPI Microdata:** *Some variables similar to CEPRdata, has CPS Basic and ORG data updated regularly*

Using Survey Microdata in R



Revisiting this line of questioning:

What share of people age 50 and older who take public transit to work are Black? What share are Hispanic?

Using R to work with a subset of ACS PUMS

- **You can follow along on GitHub:** *https://github.com/hcbrown/TR_ACSexercise*
- **We will be working with a subset of the CEPRdata ACS PUMS extract.** *Specifically, we will be using the 2019 1-year sample for DC, MD, and VA. You can download the data here:* *https://ceprdata.org/wp-content/acs/data/cepr_acs_2019_dmv.csv.zip*
- **This exercise uses RStudio and the dplyr and pollster packages**

Using R to work with a subset of ACS PUMS

- **Importing the data:**
  - *Setting directories:*

    ```
    setwd("/your/path/to/your/data/")
    ```

  - *Reading without decompressing - will be important when working with larger files:*

    ```
    acs19dmv <- read.csv(unz("cepr_acs_2019_dmv.csv.zip", "cepr_acs_2019_dmv.csv"),
    header = TRUE, sep = ",")
    ```

Using R to work with a subset of ACS PUMS

- **Selecting only the variables we care about:**
  - *We need age, race/ethnicity, whether the person uses public transit to commute to work, and the person weight.*
  - *Note that two of these four variables are CEPRextract creations, rather than ACS originals.*

```
acs19dmv_sub <- acs19dmv %>%
  select(age,
     wbhao,
     pubtran,
     perwgt)
```

Using R to work with a subset of ACS PUMS

- **Create a new variable combining age and race/ethnicity:**
  - *Use mutate to create the new variable*
    - *We've restricted our new variable, agerace, to those in-universe for pubtran*

```
acs19dmv_sub <- acs19dmv_sub %>%
  mutate(agerace = case_when(age >= 50 & wbhao == 'White' & (pubtran == 0 | pubtran == 1) ~ 'White, age 50+',
                             age >= 50 & wbhao == 'Black' & (pubtran == 0 | pubtran == 1) ~ 'Black, age 50+',
                             age >= 50 & wbhao == 'Hispanic' & (pubtran == 0 | pubtran == 1) ~ 'Hispanic, age 50+',
                             age >= 50 & wbhao == 'AAPI' & (pubtran == 0 | pubtran == 1) ~ 'AAPI, age 50+',
                             age >= 50 & wbhao == 'Other' & (pubtran == 0 | pubtran == 1) ~ 'Other, age 50+'))
```

Using R to work with a subset of ACS PUMS

- **Running a crosstab:**
  - *Eliminate the rest of the observations with NA values for agerace, the variable we just defined*

    ```
    acs19dmv_sub <- acs19dmv_sub[!is.na(acs19dmv_sub$agerace),]
    ```

  - *Run crosstab for agerace and pubtran*

    ```
    shares <- crosstab(acs19dmv_sub, pubtran, agerace, w = perwgt)
    ```

    - *Note the use of perwgt. It is important to use the appropriate weights when tabulating with microdata.*

Our answered question

Pre-pandemic in the District of Columbia, Maryland, and Virginia: Black workers age 50 and older were more likely to commute to work via public transit than White workers age 50 and older.

Workers Age 50 and Older in DC, MD, and VA, by Race/Ethnicity and Use of Public Transit to Commute to Work

	Does not use public transit to commute to work	Uses public transit to commute to work	All
White, age 50+	64.9%	38.8%	63.3%
Black, age 50+	21.5%	44.7%	22.9%
Hispanic, age 50+	6.4%	8.6%	6.5%
AAPI, age 50+	6.6%	7.2%	6.7%
Other, age 50+	0.6%	0.7%	0.6%
Total	*100%*	*100%*	*100%*