GOV51 Problem Set 1

Brian Jeon

2026-02-20

1 Project Set Up

Need immense help with Github.

2 Get to Know Your Data

1. See below.

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.6
v forcats   1.0.1     v stringr   1.6.0
v ggplot2   4.0.1     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.2.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(knitr)
library(kableExtra)
```

```
Attaching package: 'kableExtra'

The following object is masked from 'package:dplyr':

    group_rows
```

```
library(dplyr)
options(scipen = 999)
```

```
rm(list=ls())

acs2024 <- read.csv("data/raw/acs2024.csv")
```

2. Variables:
    - SAMPLE: IPUMS census identifier
    - SERIAL: unique household serial number
    - CBSERIAL: unique census bureau identification number
    - HHWT: household weight (how many households are represented by a given household)
    - CLUSTER: household cluster for variance estimation
    - STATEFIP: state which household was located using FIPS coding scheme
    - PERNUM: persons within each household
    - PERWT: how many persons in the US are represented by a given person in an IPUMUS sample
    - SEX: whether individual is male or female (1 = male, 2 = female, 9 = missing / blank)
    - AGE: age in years
    - EDUC: highest year of school or degree completed
    - EMPSTAT: whether the respondent was a part of the labor force (0 = N/A, 1 = employed, 2 = unemployed, 3 = not in labor force, 9 = unknown / illegible)
    - INCTOT: total pre-tax personal income or losses from all sources for the previous year
    - TRANTIME: total amount of time, in minutes, that it usually took the respondent to get from home to work last week (000 means not applicable aka no commuting time)
3. See below:

```r
acs2024_02 <- acs2024 %>%
  mutate(INCTOT = ifelse(INCTOT == 9999999 | INCTOT == 9999998, NA, INCTOT),
         female = ifelse(SEX == 2, 1, 0),
         less_than_hs = ifelse(EDUC <= 02, 1, 0),
         hs_only = ifelse(EDUC >= 03 & EDUC <=05, 1, 0),
         some_college = ifelse(EDUC >= 06 & EDUC <= 07, 1, 0),
         college_only = ifelse(EDUC >= 8 & EDUC <= 10, 1, 0),
         advanced_degree = ifelse(EDUC == 11, 1, 0),
         employed = ifelse(EMPSTAT == 1, 1, 0),
         unemployed = ifelse(EMPSTAT == 2, 1, 0),
         not_in_labor_force = ifelse(EMPSTAT == 3, 1, 0))
# making sure to create NA values since NA values are manually coded
# keeping NAs for the binary variables we create
```

4. If we include N/A codes as real zeros, our summary statistics may be biased measures of the true population parameter because those N/A values may reflect real values. As a result, our mean and median will be biased downward (since we are counting more 0s than actually exist), and our standard deviation will likely be greater (since more zeros likely means more distance from the mean since the mean is likely far from 0). For instance, the average income will be heavily biased downward because 0 is such a low number compared to the reported income values, so it'll really drag down our estimates.

5. See below

```r
# creating the table

summary_table <- acs2024_02 %>%
  summarise(across(
    c(AGE, female, less_than_hs, hs_only, some_college, college_only,
      advanced_degree, employed, unemployed, not_in_labor_force,
      TRANTIME, INCTOT),
                  list(
                    n = ~sum(!is.na(.)),
                    mean = ~mean(., na.rm = T),
                    sd = ~sd(., na.rm = T),
                    min = ~min(., na.rm = T),
                    max = ~max(., na.rm = T)),
            .names = "{.col}.{.fn}")) %>%
  pivot_longer(cols = everything(),
               names_to = c("variable", "stat"),
               names_sep = "\\.") %>%
```

```
  pivot_wider(names_from = stat,
              values_from = value) #combine the summary stat names
```

```r
summary_table %>%
  mutate(variable = recode(variable,
                           AGE = "Age",
                                female = "Female",
                                less_than_hs = "Less than High School",
                                hs_only = "High School Only",
                                some_college = "Some College",
                                college_only = "College Only",
                                advanced_degree = "Advanced Degree",
                                employed = "Employed",
                                unemployed = "Unemployed",
                                not_in_labor_force = "Not in Labor Force",
                                TRANTIME = "Commute Time (mins)",
                                INCTOT = "Total Income ($)")) %>%
  mutate(across(c(mean, sd, min, max), ~round(., 2))) %>%
  mutate(across(
    c(n, mean, sd, min, max),
    ~format(., big.mark = ",", scientific = F, trim = TRUE))) %>% # add commas
  kbl(caption = "Summary Statistics for 2024 ACS Sample",
    col.names = c("Variable", "N", "Mean", "Std. Dev.", "Min", "Max"),
    # format the left side column
    align = "lccccc",
    # center everything except first column
    booktabs = TRUE,
    linesep = "") # add the lines
```

6. The maximum commute time, 195 minutes or 3 hours and 15 minutes, while possible, implies that someone spends 6 hours a day total commuting. While it may not make sense for a *daily* commute (unless all that person does is work, commute, and sleep), it is a possible commute time (I know professors who live in New Haven and have a similar commute length to Harvard).

Table 1: Summary Statistics for 2024 ACS Sample

Variable	N	Mean	Std. Dev.	Min	Max
Age	3,422,888	43.39	24.03	0	96
Female	3,422,888	0.51	0.50	0	1
Less than High School	3,422,888	0.18	0.39	0	1
High School Only	3,422,888	0.06	0.24	0	1
Some College	3,422,888	0.40	0.49	0	1
College Only	3,422,888	0.24	0.43	0	1
Advanced Degree	3,422,888	0.11	0.32	0	1
Employed	3,422,888	0.47	0.50	0	1
Unemployed	3,422,888	0.02	0.14	0	1
Not in Labor Force	3,422,888	0.35	0.48	0	1
Commute Time (mins)	3,422,888	10.81	19.83	0	195
Total Income ($)	2,912,790	54,654.71	80,080.40	-11,500	1,945,000

3 Who Should Be in Your Analysis

1. See below:
    - 2,062,945 people have $TRANTIME == 0$. This is 60.27% of the census.
    - Because 000 represents N/A for TRANTIME, they might be individuals who are self-employed, are part of the gig economy (uber, doordash, etc.), do not commute (remote workers), or just don't work since one wouldn't have a commute time then.

```
acs2024_02 %>%
  filter(TRANTIME == 0) %>%
  summarize(n = n())
```

```
        n
1 2062945
```

```
2062945/3422888
```

```
[1] 0.6026914
```

```
# acs2024_02 %>% (verifying that 000 and 0 are treated identically)
  # filter(TRANTIME == 000) %>%
  # summarize(n = n())
```

2. Filter criterion:

    - I chose to remove individuals with no commute time since they don't actually commute to work or have invalid data.
    - I also chose to remove individuals who were not employed (since you can't commute to work if you don't work!).
    - I had no age restrictions since it's plausible that a young person commutes and works (minimum age of removing the above is 16 which is reasonable).
    - I had no income restrictions since it's plausible that someone who commutes to work can have negative or even zero income if they make poor financial decisions.
    - I had no education restriction since someone with no or lots of education can both commute to work.
    - I had no sex restriction since both women and men can commute to work.

```
commuters <- acs2024_02 %>%
  filter(TRANTIME != 0,
         employed == 1)

commuters %>%
  summarize(min_age = min(AGE, na.rm = T),
            min_inctot = min(INCTOT, na.rm = T))
```

```
  min_age min_inctot
1      16     -11500
```

3. See below

    - 1,359,943 observations remain.
    - Minimum commute time is 1 minute (makes sense since this person must live real close by).

```
commuters %>%
  summarize(n = n(),
    min_TRANTIME = min(TRANTIME))
```

```
        n min_TRANTIME
1 1359943            1
```

4. See below:

```r
# creating the table

summary_table02 <- commuters %>%
  summarise(across(c(AGE, female, less_than_hs, hs_only, some_college,
                     college_only, advanced_degree, employed, unemployed,
                     not_in_labor_force, TRANTIME, INCTOT),
      list(n = ~sum(!is.na(.)),
           mean = ~mean(., na.rm = T),
           sd = ~sd(., na.rm = T),
           min = ~min(., na.rm = T),
           max = ~max(., na.rm = T)),
      .names = "{.col}.{.fn}")) %>%
  mutate(AGE.n = sum(!is.na(commuters$AGE)),
         TRANTIME.n = sum(!is.na(commuters$TRANTIME)),
         INCTOT.n = sum(!is.na(commuters$INCTOT))) %>%
  # we can't sum these for n because they are non binaries
  pivot_longer(cols = everything(),
               names_to = c("variable", "stat"),
               names_sep = "\\.") %>%
  pivot_wider(names_from = stat,
              values_from = value) #combine the summary stat names
```

```r
summary_table02 %>%
  mutate(variable = recode(variable,
                           AGE = "Age",
                                female = "Female",
                                less_than_hs = "Less than High School",
                                hs_only = "High School Only",
                                some_college = "Some College",
                                college_only = "College Only",
                                advanced_degree = "Advanced Degree",
                                employed = "Employed",
                                unemployed = "Unemployed",
                                not_in_labor_force = "Not in Labor Force",
                                TRANTIME = "Commute Time (mins)",
                                INCTOT = "Total Income ($)")) %>%
  mutate(across(c(mean, sd, min, max), ~round(., 2))) %>%
  mutate(across(
    c(n, mean, sd, min, max),
    ~format(., big.mark = ",", scientific = F, trim = TRUE))) %>% # add commas
  kbl(caption = "Summary Statistics for 2024 ACS Sample for Commuters",
    col.names = c("Variable", "N", "Mean", "Std. Dev.", "Min", "Max"),
    # format the left side column
    align = "lccccc",
    # center everything except first column
    booktabs = TRUE,
    linesep = "") # add the lines
```

5.

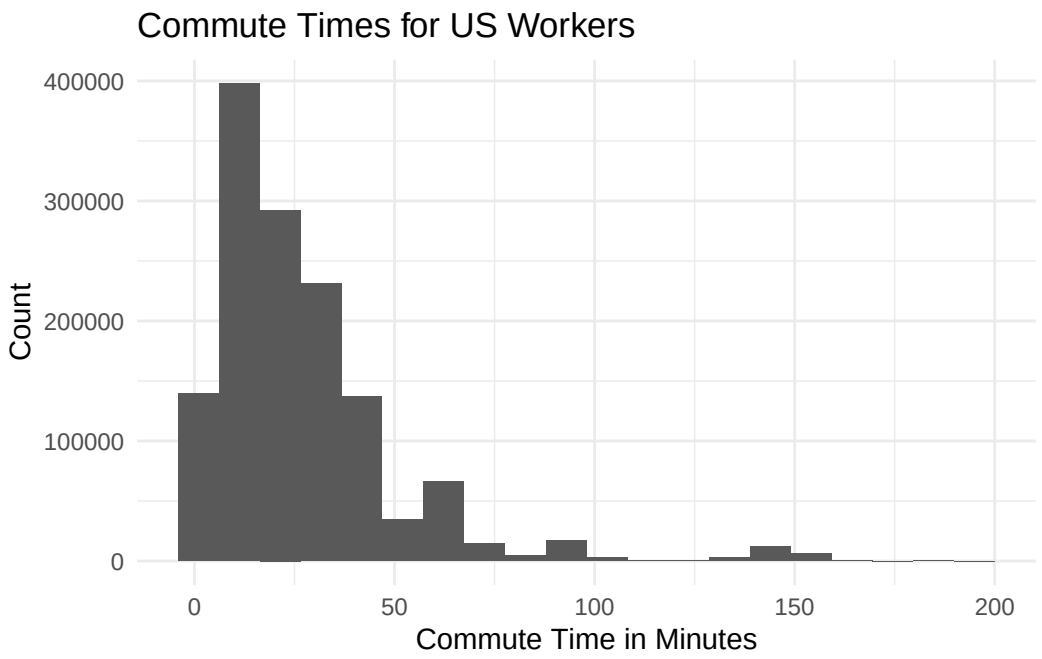
Table 2: Summary Statistics for 2024 ACS Sample for Commuters

Variable	N	Mean	Std. Dev.	Min	Max
Age	1,359,943	43.39	15.30	16	96
Female	1,359,943	0.47	0.50	0	1
Less than High School	1,359,943	0.03	0.17	0	1
High School Only	1,359,943	0.04	0.18	0	1
Some College	1,359,943	0.46	0.50	0	1
College Only	1,359,943	0.32	0.47	0	1
Advanced Degree	1,359,943	0.15	0.36	0	1
Employed	1,359,943	1.00	0.00	1	1
Unemployed	1,359,943	0.00	0.00	0	0
Not in Labor Force	1,359,943	0.00	0.00	0	0
Commute Time (mins)	1,359,943	27.22	23.31	1	195
Total Income ($)	1,359,943	72,992.65	88,726.05	-11,500	1,945,000

4 Visualize and Interpret

1. See below:

```
commuters %>%
  ggplot(mapping = aes(x = TRANTIME)) +
  geom_histogram(bins = 20) +
  labs(title = "Commute Times for US Workers",
       x = "Commute Time in Minutes",
       y = "Count") +
  theme_minimal()
```



2. The histogram is very skewed to the right. The center appears to be around the 10-30 minute spike. The 10-30 minute bucket has a major spike and the ~60 minute bucket has a minor spike.

3. The median (20) is less than the mean (27.22) which implies the shape of the distribution is skewed to the right since the outliers / upper tail drags the mean to the right while the median, an estimate less sensitive to extreme values is not as dragged). This matches what we see in the histogram.

```
# part 3 didn't ask us to calculate the median so

commuters %>%
  summarize(median = median(TRANTIME))
```

```
  median
1     20
```

4. See below:

```
ggsave("output/commute_histogram.png")
```

```
Saving 5.5 x 3.5 in image
```

5.

5 The Weight of Evidence

1. See below:

```
weighted.mean(commuters$TRANTIME, commuters$PERWT)
```

```
[1] 27.19112
```

2. The weighted mean (27.19) is very very very slightly lower than the mean in Section 3 (27.22).

3. The weighted mean represents the average where data points that are more frequently occurring are weighed more. In this case, because PERWT represents the number of people who have a certain row of values, the weighted average gives a better estimate of a person-level average versus the non-weighted mean which simply calculating the average commuting time flatly across unique individuals which doesn't take into account the proportion of different demographics. These values may differ if many people are represented by a single set of data but there are many unique sets of data that do not represent many people (ex: imagine commuting times for 90% of the population is 10 min and all the other data align for these 90% people, the weighted average would take that into account but the flat average would treat that 10 minute data as just one person).
