560 Group Project: Descriptive Statistics

Brittany Jones, Tess Abrahamson-Richards, Natalie Turner, Hanyi Wang

2023-03-14

Population of Interest

Filter by women over 50 years old in the SIPP study, there are 163727 entries, and 13322 unique participants in our dataset.

Key Variables

Outcome: Monthly social service utilization

(The following variables are all binary variables with levels "Yes" or "No". We count each "Yes" as 1 point and calculate the monthly social service utilization for each participant.)

- `RGA_MNYN`: Received GA benefits this month
- `RSNAP_MNYN`: Received SNAP benefits this month
- `RSSI_MNYN`: Received SSI benefits this month
- `RTANF_MNYN`: Received TANF benefits this month
- `RWIC_MNYN`: Received WIC benefits this month
- `ECHLD_MNYN`: Received child care service assistance to go to work or school training this month (1-12)
- `ECVDMNYN`: Received the EIP(s), or stimulus check(s) this month (1-12)
- `EOTHAS_MNYN`: Received other (clothing or housing) assistance this month (1-12)
- `ETRANS_MNYN`: Received transportation assistance this month (1-12)
- `EUC1MNYN`: Received regular, government-provided Unemployment Compensation payments in this month of the reference period (1- 12)
- `EUC2MNYN`: Received supplemental, employer-provided Unemployment Compensation payments in this month of the reference period (1- 12)
- `EUC3MNYN`: Received other Unemployment Compensation payments, including union benefits, in this month of the reference period (1- 12)

	Overall
	(N=163727)
ECHLD_MNYN	
No	3786 (2.3%)
Yes	93 (0.1%)
Missing	159848 (97.6%)
ECVDMNYN	
No	85708 (52.3%)
Yes	13833 (8.4%)
Missing	64186 (39.2%)
EUC1MNYN	
No	5454 (3.3%)
Yes	3393 (2.1%)
Missing	154880 (94.6%)
EUC2MNYN	
No	465 (0.3%)
Yes	327 (0.2%)
Missing	162935 (99.5%)
EUC3MNYN	
No	23 (0.0%)
Yes	25 (0.0%)
Missing	163679 (100.0%)

As shown above, variables `ECHLD_MNYN`, `ECVDMNYN`, `EUC1MNYN`, `EUC2MNYN`, `EUC3MNYN` have a lot of missing values, so we exclude them from the calculation of social service utilization scores. Thus the monthly social service utilization score for a participant can be 0 to 7.

Covariates fitted in the models

Demographic vairables:

- `ESEX`: sex (all "female" based on our inclusion criteria, so will not be in the models)
- `TRACE`: detailed race
- `EORIGIN`: is Spanish, Hispanic, or Latino, or not
- `EEDUC`: highest level of school completed or the highest degree received by December of (reference year)
- `EHLTSTAT`: self-reported health status (Confounder)

Individual's total cost of health care `HI` (sum of the following 2 variables)

- `TMDPAY`: non-premium medical out-of-pocket expenditures on medical care
- `TOTCMDPAY`: non-premium medical out-of-pocket expenditures for over-the-counter health-related products

Descriptive Statistics

We also studied the relationship of social service utilization for each individual and month. We first used a histogram and found that there's no significant difference for the distribution of social service utilization scores over different months. (**Figure 3**)

We then made a plot of all participants' social service utilization scores over months, and found a lot of overlapping. (**Figure 4**)

Table 1: Demographic charactersitics for older adult women in the SIPP study

	Overall
	(N=13322)
Detailed race	
American Indian or Alaska Native alone (AIAN)	106 (0.8%)
Asian alone	626 (4.7%)
Black alone	1548 (11.6%)
Black-AIAN	21 (0.2%)
Native Hawaiian or Other Pacific Islander alone (HP)	22 (0.2%)
Other 2 or more races	30 (0.2%)
White alone	10809 (81.1%)
White-AIAN	108 (0.8%)
White-Asian	21 (0.2%)
White-Black	31 (0.2%)
Is Spanish, Hispanic, or Latino	
Is Spanish, Hispanic, or Latino	1385 (10.4%)
No	11937 (89.6%)
Highest degree	
1 or more years of college, no degree	1538 (11.5%)
10th grade	240 (1.8%)
11th grade	238 (1.8%)
12th grade, no diploma	161 (1.2%)
1st, 2nd, 3rd or 4th grade	127 (1.0%)
5th or 6th grade	215 (1.6%)
7th or 8th grade	266 (2.0%)
9th grade	223 (1.7%)
Associate's degree	1335 (10.0%)
Bachelor's degree	2480 (18.6%)
Doctorate degree	209 (1.6%)
High School Graduate	3861 (29.0%)
Less than 1st grade	48 (0.4%)
Master's degree	1395 (10.5%)
Professional School degree	209 (1.6%)
Some college credit, but less than 1 year	777 (5.8%)
Health status	
Excellent	1584 (11.9%)
Fair	2414 (18.1%)
Good	4456 (33.4%)
Poor	968 (7.3%)
Very good	3674 (27.6%)
Missing	226 (1.7%)

Table 2: Social service utilization for older adult women in the SIPP study

	Overall
	(N=163727)
Monthly social service utilization (score)	
0	127733 (78.0%)
1	26394 (16.1%)
2	7021 (4.3%)
3	2116 (1.3%)
4	410 (0.3%)
5	51 (0.0%)
6	2 (0.0%)

Table 3: Health care spending for older adult women in the SIPP study

	Overall
	(N=163727)
Health care utilization (dollars)	
Mean (SD)	1230 (2530)
Median [Min, Max]	400 [0, 24500]

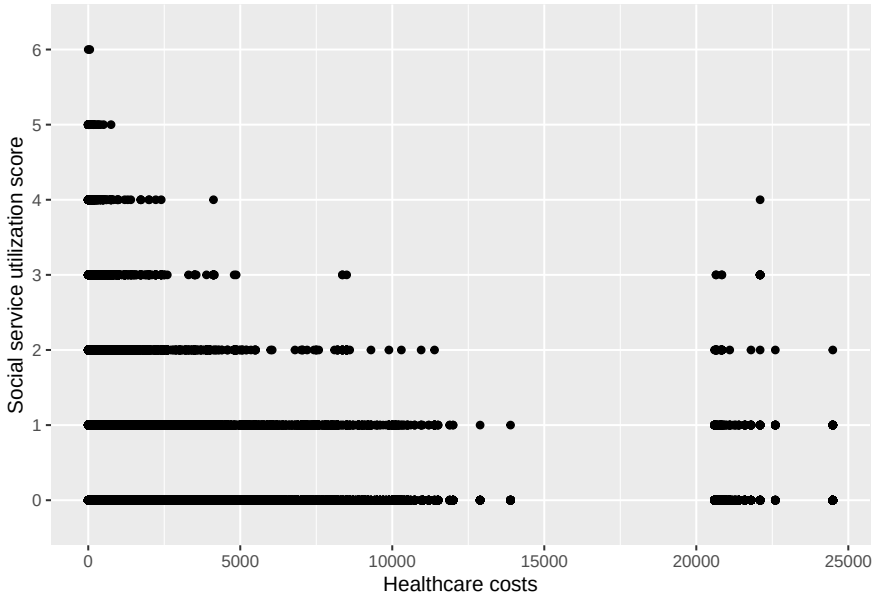

Figure 1: Social service utilization for different healthcare cost

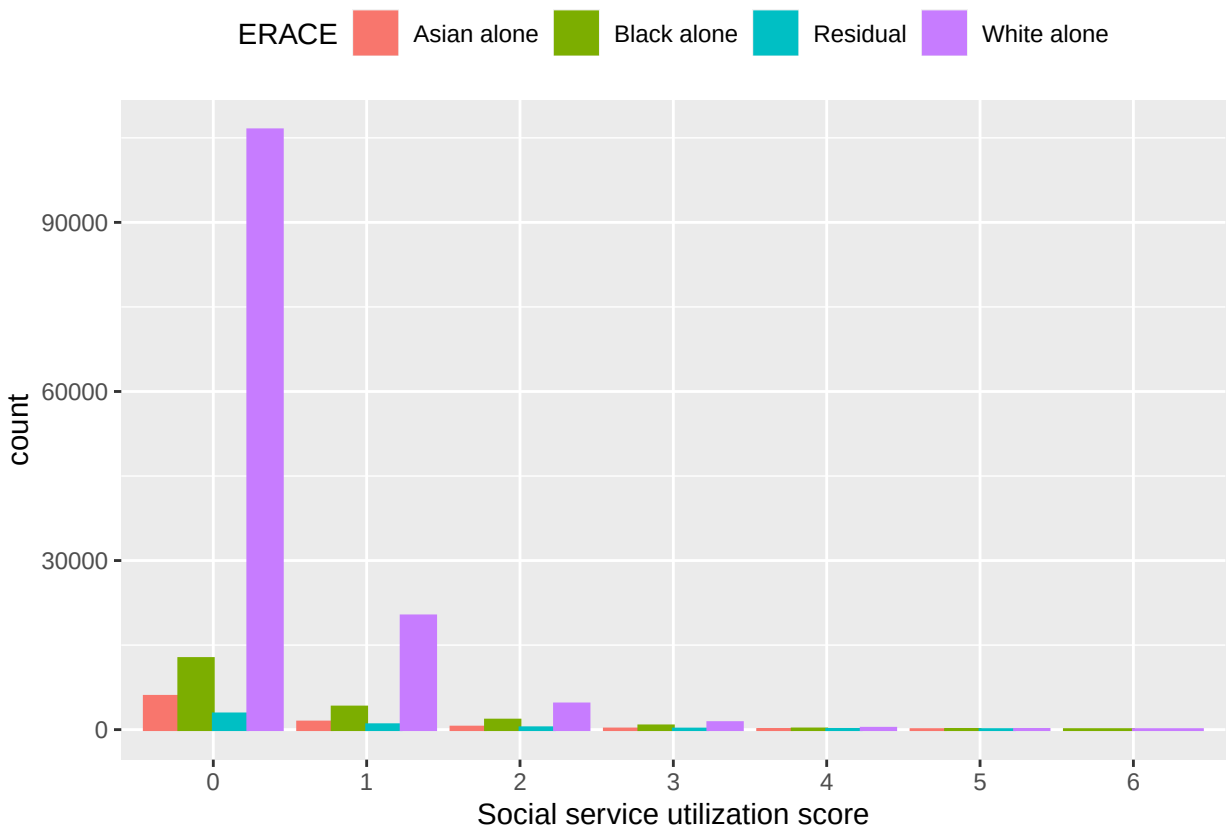

Figure 2: Histogram of social service utilization for different races

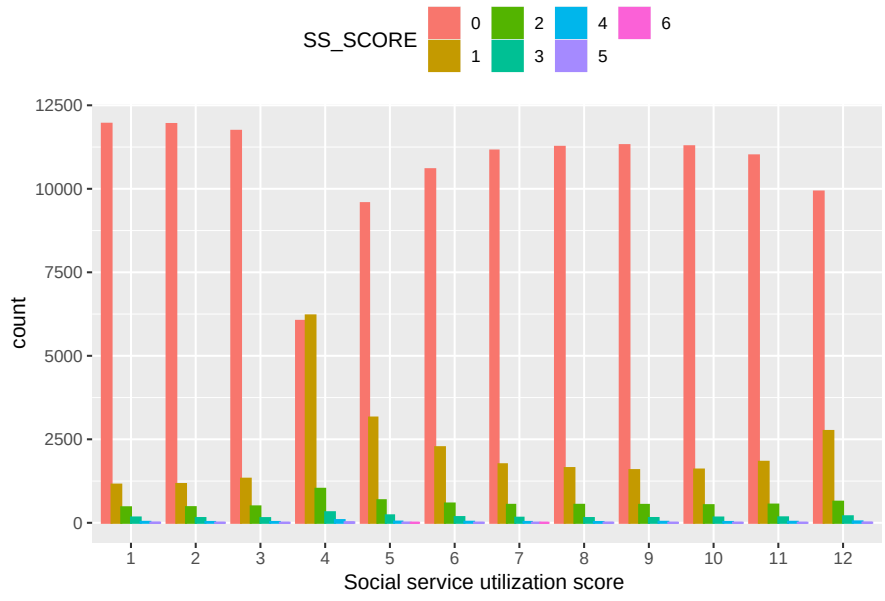

Figure 3: Social service utilization distribution over 12 months

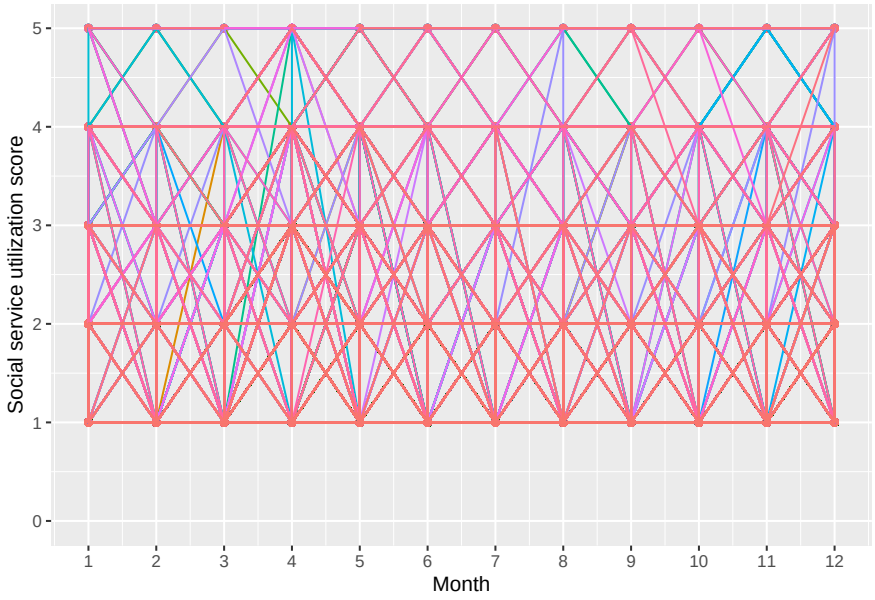

Figure 4: Social service utilization changes over 12 months for all participants

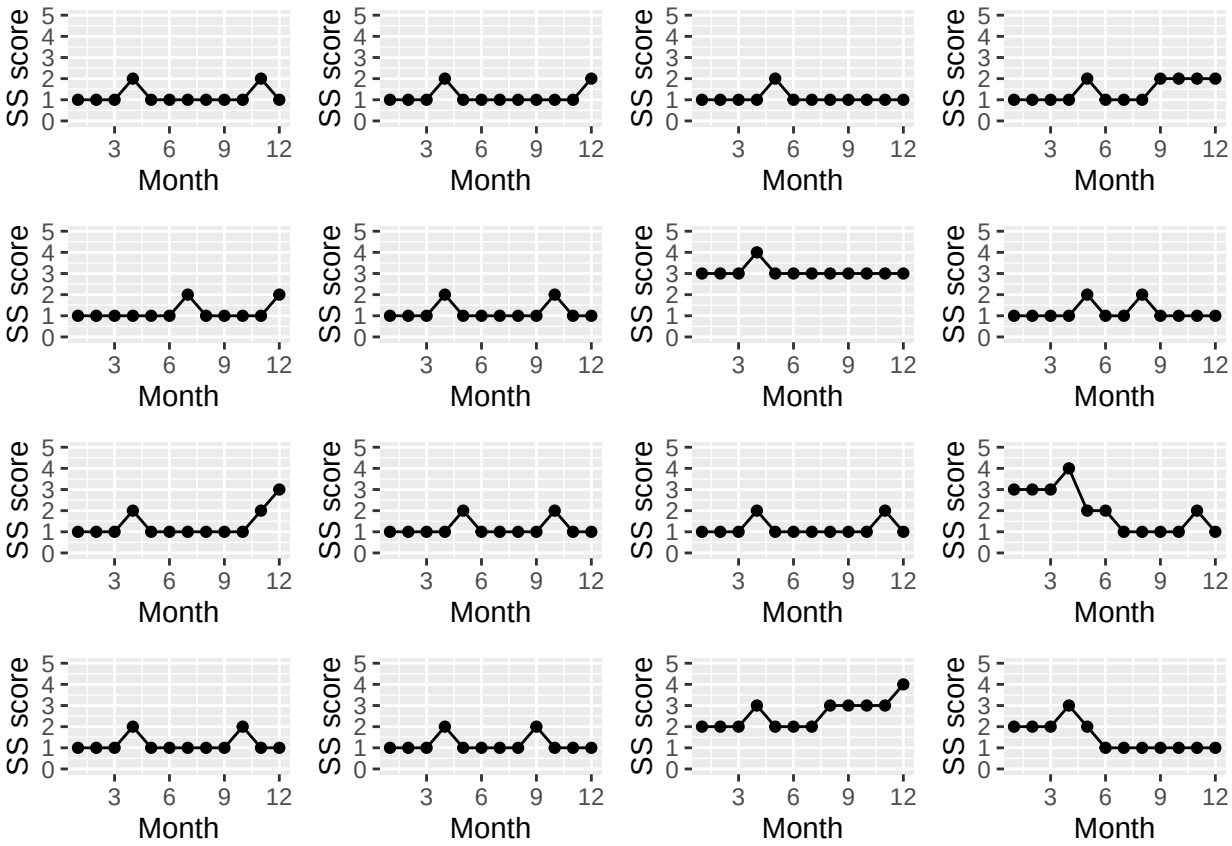

Figure 5: Social service utilization changes over 12 months for participants with a change

To study the difference in more details, we went ahead to make a single plot for the trend for each participant. After carefully screening, we found most participants have the trend of having a smaller score around summer, and a larger score over spring and fall. We randomly picked some of the participants shown here as an illustration of the trends. (**Figure 5**) This trend further instructed us to put month as a covariate in our models.

Code Appendix

```
knitr::opts_chunk$set(echo = FALSE,message = FALSE,warning = FALSE)
library(tidyverse)
library(mlmRev)
library(lme4)
library(arm)
library(haven)
library(table1)
library(Hmisc)
library(ggplot2)
library(ggpubr)
library(expss)
library(dplyr)
library(gridExtra)
library(flextable)

## ------------- working directories for Hanyi ----------
wd_hanyi = '/Users/hanyiwang/Desktop/Social-Service-Participation-Among-Older-Adult-Women-in-Washington-
path_hanyi = c("../SIPPdata/sipp.csv")

## ------------- read data ----------
setwd(wd_hanyi)
sipp = read.csv(path_hanyi)

unique = sipp %>% distinct(UQID, .keep_all=TRUE) # unique participants

table1(~ECHLD_MNYN+ECVDMNYN+EUC1MNYN+EUC2MNYN+EUC3MNYN,data = sipp)
## ------------- modify data set ----------
sipp = sipp %>%
  mutate(SS_SCORE = factor(SS_SCORE)) %>%
  mutate(MONTHCODE = factor(MONTHCODE))

## ------------- modify labels and units ----------
unique = apply_labels(unique,
                      ESEX = "Sex",
                      ERACE = "Race",
                      TRACE = "Detailed race",
                      EORIGIN = "Is Spanish, Hispanic, or Latino",
                      EEDUC = "Highest degree",
                      EHLTSTAT = "Health status")

sipp = apply_labels(sipp,
                      SS_SCORE = "Monthly social service utilization",
                      HI = "Health care utilization")
```

```r
units(sipp$HI) = "dollars"
units(sipp$SS_SCORE) = "score"
## ------------- descriptive tables ----------
demographic = table1(~+TRACE+EORIGIN+EEDUC+EHLTSTAT,data=unique, caption = "Demographic charactersitics

ss_score = table1(~(SS_SCORE),data=sipp, caption = "Social service utilization for older adult women in

hi = table1(~HI,data=sipp, caption = "Health care spending for older adult women in the SIPP study")

# t1flex(demographic) %>% save_as_docx(path="demographic.docx")

demographic
ss_score
hi
## ------------- SS_SCORE VS HI ----------
ggplot(sipp,aes(x=HI,y=SS_SCORE)) +
  geom_point() +
  labs(y = "Social service utilization score", x = "Healthcare costs")


## ------------- SS_SCORE VS race ----------
ggplot(sipp, aes(x=(SS_SCORE), color=ERACE,fill=ERACE)) +
  geom_histogram(position = "dodge", binwidth=2,stat="count") +
  labs(x = "Social service utilization score") +
  theme(legend.position="top")

## ------------- SS_SCORE VS month ----------
#### boxplot
ggplot(sipp, aes(x=(MONTHCODE), color=SS_SCORE,fill=SS_SCORE)) +
  geom_histogram(position = "dodge", binwidth=2,stat="count") +
  labs(x = "Social service utilization score ") +
  theme(legend.position="top")
#### all individuals
sipp = sipp %>%
  mutate(SS_SCORE = as.numeric(SS_SCORE)) %>%
  mutate(MONTHCODE = as.numeric(MONTHCODE)) %>%
  mutate(UQID = factor(UQID))


ggplot(sipp,  aes(x = MONTHCODE,y = SS_SCORE,color = UQID)) +
  geom_point()+ geom_line() +
  ylim(0,5) + theme(legend.position = "none") + scale_x_continuous(breaks=seq(0,12,by=1)) +
  labs(y = "Social service utilization score",x = "Month")


#### randomly pick individuals
# rd = sample(nrow(unique), 30)
# sipp_rd = sipp[sipp$UQID %in% unique[rd,]$UQID,]
# sipp_rd = sipp_rd %>%
#   mutate(SS_SCORE = as.numeric(SS_SCORE)) %>%
#   mutate(MONTHCODE = as.numeric(MONTHCODE)) %>%
#   mutate(UQID = factor(UQID))
#
```

```r
# ggplot(sipp_rd,  aes(x = MONTHCODE,y = SS_SCORE,color = UQID)) +
#   geom_point()+
#   geom_line() +
#   ylim(0,5) +
#   theme(legend.position = "none")

#### all individuals in seperate plots

# p = list()
# for (i in 1:30){
#   sipp_rd = sipp[sipp$UQID == factor(unique$UQID)[i+1000],]
#   p[[i]] = ggplot(sipp_rd,  aes(x = MONTHCODE,y = SS_SCORE)) +
#     geom_point()+ geom_line() + ylim(0,5) + scale_x_continuous(breaks=seq(0,12,by=3))
# }
#
# do.call(grid.arrange,p)


# after carefully screening the individuals, we found most participants have this trend of having a sma

in_ch = c(9, 6,16,45, 57,319,71,67, 80,94,95,117, 315,316,1013,1015)

p = list()
for (i in 1:length(in_ch)){
  sipp_rd = sipp[sipp$UQID == factor(unique$UQID)[in_ch[i]],]
  p[[i]] = ggplot(sipp_rd,  aes(x = MONTHCODE,y = SS_SCORE)) +
    geom_point()+ geom_line() + ylim(0,5) + scale_x_continuous(breaks=seq(0,12,by=3)) +
    labs(y = "SS score",x = "Month")

}

do.call(grid.arrange,p)
```