Stat 854 Project: Mirror-match Bootstrap

...

April 14 2024

Introduction

Method

Experiments

Related Works

Conclusion

Appendix

1. First, we pre-process the syc.txt in SAS, where we delete the missing value in numarr and keep only two column variables ‘stratum’ and ‘numarr’.

In addition, PROC FREQ is used to generate frequency counts for the stratum variable. After that, the processed dataset is exported to ‘syc_post.txt’.

SAS code


```
options pagesize=50 linesize=80;
options formchar='|----|+|---+=|-/\<>*';
title1 "SYC Data Preprocessing";
footnote "SYC Data Preprocessing";

/* creating dataset */
data SYC;
    infile "/home/u63744989/dataset/syc.txt" missover firstobs = 2 delimiter = ',';
    /* before read in the data, we manually delete the instructions on the top of the txt file */
    input stratum psu psusize initwt finalwt randgrp age race ethnicty educ
    sex livewith famtime crimtype everviol numarr probtn corrinst evertime
    prviol prprop prdrug prpub prjuv agefirst usewepn alcuse everdrug;

    /* creating labels */
    label stratum = "stratum";
    label numarr = "number of prior arrest";

    /* Filter out rows with missing numarr value*/
    if (numarr = 99) then delete;

    /* Keep only the necessary columns, i.e. stratum & numarr*/
    keep stratum numarr;
run;

/* sort the dataset by stratum in an ascending order*/
proc sort data=SYC;
    by stratum;
run;

/* output dataset to a csv file*/
proc export data=SYC
    outfile='/home/u63744989/dataset/syc_post.csv'
    dbms=csv
    replace;
run;

/* View overview of the stratum variable */
proc freq data=SYC;
  tables stratum / nocol;
  title "Overview of Stratum Variable";
run;

ods pdf file='/home/u63744989/figures/output_1.pdf';

ods _all_ close;
```

SAS output

Overview of Stratum Variable

The FREQ Procedure

stratum				
stratum	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	178	6.93	178	6.93
2	155	6.04	333	12.97
3	327	12.74	660	25.71
4	533	20.76	1193	46.47
5	568	22.13	1761	68.60
6	47	1.83	1808	70.43
7	7	0.27	1815	70.71
8	65	2.53	1880	73.24
9	71	2.77	1951	76.00
10	65	2.53	2016	78.54
11	83	3.23	2099	81.77
12	47	1.83	2146	83.60
13	93	3.62	2239	87.22
14	77	3.00	2316	90.22
15	101	3.93	2417	94.16
16	150	5.84	2567	100.00

SYC Data Preprocessing

2. Second, we read in the data in R and compute sampling weights.

```
# Read in the Survey of Youth in Custody
syc <- readr::read_csv(
  file = "data/syc_post.csv", # Tell it where the file is
  col_types = "nn", # Tell it that there are two columns, and they are "numeric" (n)
)

# glimpse the read data set
dplyr::glimpse(syc)
```

References