Live Session 10

Final Exam

- Week 14 : During the live session (In class portion)

- Part II (Take Home)
- Thursday August 10, 10.00a.m CT

- Submit on Monday August 14, midnight CT

Ratio estimator

- Need an auxiliary variable x that is correlated with y where

$$y_i = \hat{R}x_i$$

- Then can estimate

$$\hat{R} = \frac{\bar{y}}{\bar{x}}$$

- And form estimates of mean & total

$$\bar{y}_r = \bar{X}\frac{\bar{y}}{\bar{x}}$$

$$\hat{t}_{yr} = t_x \frac{\bar{y}}{\bar{x}}$$

Ratio estimator

- x_i is often called an auxiliary variable or subsidiary variable.
- Population size is N

$$t_x = \sum_{i=1}^{N} x_i$$

$$t_y = \sum_{i=1}^{N} y_i$$

Ratio estimator

- $True\ ratio\ (R)$

$$R = \frac{t_y}{t_x} = \frac{\bar{y}_U}{\bar{x}_U} \text{ where } \bar{y}_U = \frac{\sum_{i=1}^{N} y_i}{N} = \bar{Y}$$

$$\hat{R} = \frac{\bar{y}}{\bar{x}} = \frac{\hat{t}_y}{\hat{t}_x} \text{ where } \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

$$\hat{t}_y = N\bar{y}$$

Ratio estimators (con't)

- Then the ratio estimator of total is

$$\hat{t}_{yr} = \sum_{i=1}^{n} y_i + \sum_{i=n+1}^{N} \hat{y}_i = \sum_{i=1}^{n} y_i + \sum_{i=n+1}^{N} \hat{R} x_i$$

$$= n\bar{y} + \frac{\bar{y}}{\bar{x}}(t_x - n\bar{x}) = n\bar{y} + t_x \frac{\bar{y}}{\bar{x}} - n\bar{y}$$

$$\hat{t}_{yr} = t_x \frac{\bar{y}}{\bar{x}} = t_x \hat{R}$$

Ratio estimators (con't)

- Ratio estimator of total is

$$\hat{t}_{yr} = n\bar{y} + \frac{\bar{y}}{\bar{x}}(t_x - n\bar{x}) = t_x \frac{\bar{y}}{\bar{x}} = t_x \hat{R}$$

- The ratio estimator of mean is

$$\bar{y}_r = \bar{y} + \frac{\bar{y}}{\bar{x}}(\bar{X} - \bar{x}) = \bar{X}\frac{\bar{y}}{\bar{x}} = \bar{X}\hat{R} = \bar{x}_U \hat{R}$$

When is ratio estimator useful?

- Previous measurement of variable

- Other variable correlated with variable of interest
  - when one equals 0, the other does too

- Both numerator & denominator are items collected in the survey

Ratio estimation is most appropriate if a straight line through the origin summarizes the relationship between x_i and y_i

Examples of use of ratio estimation

- Economic/business surveys
  - Estimates of Inventory/sales
  - Estimates of total production based on sample
    - Estimate amount of juice from shipment of apples when know weight of shipment & juice/lb from sample

- Household surveys
  - Percentage of income spent on food

    (amount spent on food)/income

Estimation of ratio

- There are times when the objective is to estimate a ratio, rather than a mean or a total.
- In that case, you may not know X for the entire population, since your point is not to use X as a “helper” for Y.
- Example: Suppose you are interested in the profit per employee and the sale per employee among the 800 top-performing companies. You select a SRS of companies and estimate the ratio as

$$\hat{R} = \frac{\bar{y}}{\bar{x}}$$

SAS PROC SURVEYMEANS provides and optional statement for estimating the ratio.

Direct use of ratio estimation

Example: Suppose you are interested in percentages of pages in Good Housekeeping magazine that contain at least one advertisement. You might take an SRS of 10 issues from the most recent 60 issues of the magazines and for each issue measure the following.

x_i= total number of pages in issue *i*

y_i= total number of pages in issue *i* that contains at least one advertisement

The proportion of interest can be estimated as

$$\hat{R} = \frac{\hat{t}_y}{\hat{t}_x}$$

Variance estimate of ratio estimator in SAS Proc Surveymeans

- The default method for calculating standard error is the Taylor series method (TS)
- When sample size is small, the standard error of the ratio estimator is often underestimated, just as it is for a ratio itself. Then a replication method of variance estimation would give a more realistic measure of standard error.
  - Jackknife (JK)
  - Balanced Repeated Replication (BRR)

Ratio estimate in Proc Surveymeans

Ratio statement:

Ratio *varname1/varname2;*

Where

varname1 is the numerator

Varname2 is the denominator

Precision of ratio estimator

- Note that $Var(\hat{t}_{yr}) = Var\left(t_x \frac{\bar{y}}{\bar{x}}\right) = t_x^2 Var(\frac{\bar{y}}{\bar{x}})$
- Since $\bar{y}/\bar{x}$ is a ratio of sums, we use an approximation method to get its variance.
  - Taylor Series
  - BRR
  - Jackknife
- For large *n*, $Var(\hat{t}_{yr}) < Var(\hat{t}_y)$ when

$$\rho > \frac{1}{2}\frac{CV_x}{CV_y}$$

where $CV_x = S_x/\bar{X}$, and ρ is the correlation between *X* and *Y*.

Example of ratio estimation (Source: Sharon Lohr)

- The U.S. government conducts a Census of Agriculture every five years, collecting data on all farms in the 50 states.
- The file agri.csv contains data from SRS of 300 of the 3078 counties in 1987 and 1992. For this example, suppose we know the population totals for 1987, but only have 1992 information on the SRS 300 counties. When the sample quantity is measured at different times, the response of interest at an earlier time often makes an excellent auxiliary variable.
- Let

  y_i= total acreage of farms in county *i* in 1992

  x_i= total acreage of farms in county *i* in 1987

- In 1987 a total of t_x=964,470,625 acres were devoted to farms in the United States. The average acreage per county for the population is then

- $\bar{x}_U$=$\bar{X} = \frac{964{,}470{,}625}{3078} = 313{,}343.3$ acres of farms per county.

```
proc corr data = agsrs;
   var acres92 acres87;
   run;
```

The CORR Procedure

2 Variables:	acres92 acres87

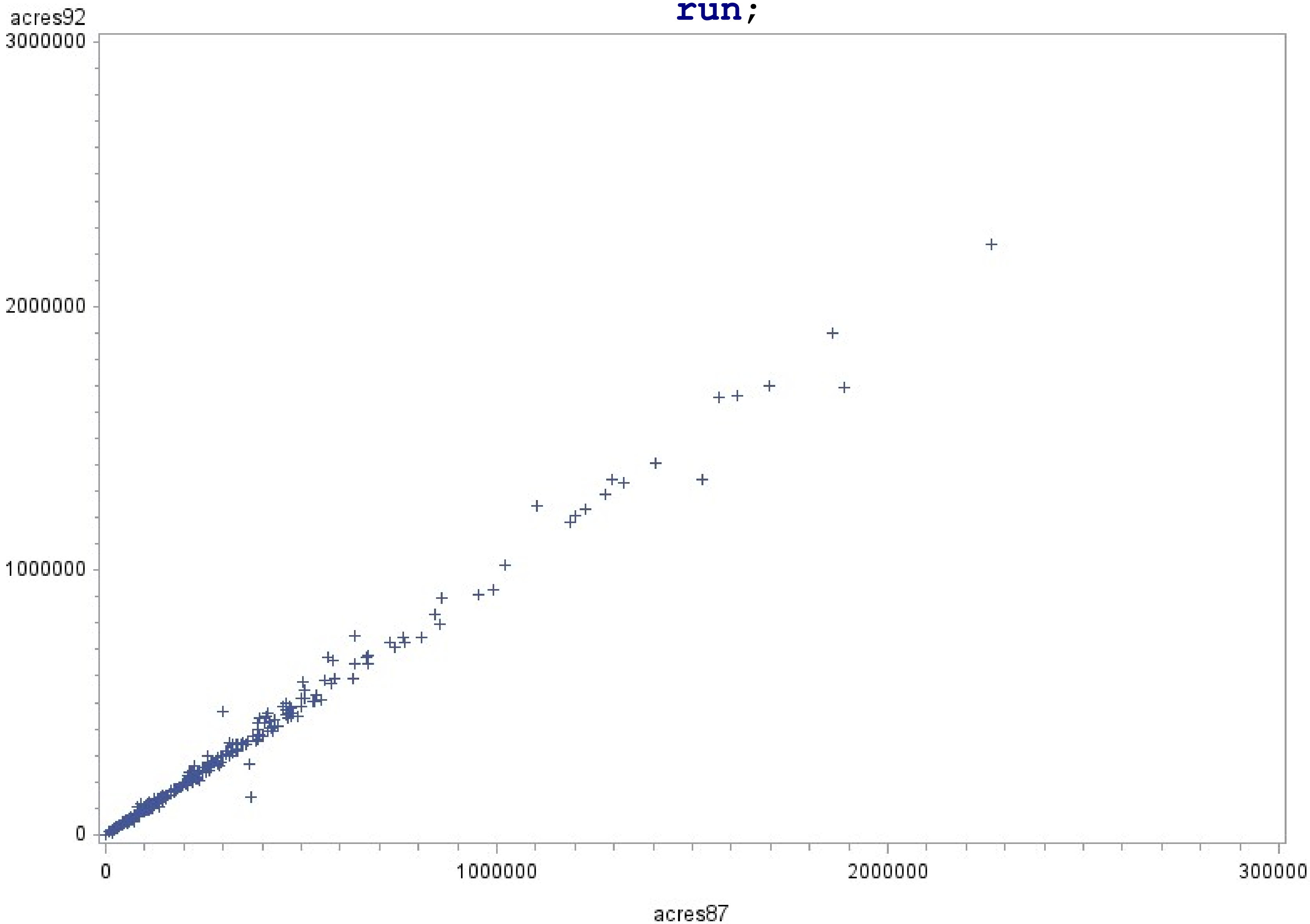
Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
acres92	300	297897	344552	89369114	0	2234262
acres87	300	301954	344830	90586117	0	2266986

Pearson Correlation Coefficients, N = 300
Prob > |r| under H0: Rho=0

	acres92	acres87
acres92	1.00000	0.99581 <.0001
acres87	0.99581 <.0001	1.00000

```
proc gplot data = agsrs;
   plot acres92 * acres87;
run;
```



```
proc surveymeans data=agsrs total=3078 mean stderr clm
sum clsum  ;
   var acres92 acres87;  /* need both in var statement */
   ratio 'acres92/acres87' acres92/acres87;
   weight sampwt;
   ods output Statistics=statsout Ratio=ratioout;
   run;


proc print data=ratioout;
run;
```

The SAS System

Obs	RatioLabel	NumeratorName	DenominatorName	Ratio	StdErr	LowerCL	UpperCL
1	acres92/acres87	acres92	acres87	0.986565	0.005750	0.97524871	0.99788176

$$\hat{R} = \frac{\bar{y}}{\bar{x}} = \frac{297897}{301954} = 0.9865$$

$$\bar{y}_r = \bar{X}\hat{R} = \bar{x}_U\hat{R} = 313\,343.283 * 0.986565 = 309133.59$$

```
data ratioout1;
   set ratioout;
   Xbar = 313343.283;
   ratiomean = ratio*xbar;
   semean = stderr*xbar;
   lowercls = lowercl*xbar;
   uppercls = uppercl*xbar;

proc print data = ratioout1;
run;
```

The SAS System

Obs	RatioLabel	NumeratorName	DenominatorName	Ratio	StdErr	LowerCL	UpperCL	Xbar	ratiomean	semean	lowercls	uppercls
1	acres92/acres87	acres92	acres87	0.986565	0.005750	0.97524871	0.99788176	313343.28	309133.59	1801.87	305587.63	312679.55

Performing ratio estimation in SAS

- SURVEYMEANS does not have a built-in function for calculating the ratio estimator of mean.
- It is easy to make a ratio estimator from SURVEYMEANS because you can request an estimate of a ratio
- So to compute a ratio estimator, simply define a new variable

$$\text{m}xy = \bar{x}_U * y$$

and then proceed with computing the ratio

$$\frac{\overline{mxy}}{\bar{x}}$$

Another way to get the mean estimate of y from ratio estimation

```
data agsrs;
   infile agsrs delimiter= ',' firstobs = 2;
   input county $ state $ acres92 acres87 ;
   sampwt = 3078/300;
mxy=313343.283*acres92;* mean estimate of y from
ratio estimation;
run;


proc surveymeans data=agsrs total=3078 mean stderr clm sum
clsum  ;
var acres92 acres87;
ratio  mxy/acres87;
weight sampwt;
ods output Statistics=statsout Ratio=ratioout4;
run;
```

The SAS System

Obs	NumeratorName	DenominatorName	Ratio	StdErr	LowerCL	UpperCL
1	mxy	acres87	309134	1801.871993	305587.633	312679.548

$$\hat{t}_{yr} = t_x \frac{\bar{y}}{\bar{x}} = t_x \hat{R} = 964470625 * 0.9865652371 = 951513191$$

```
data ratioout2;
   set ratioout;
   xtotal = 964470625;
   ratiosum = ratio*xtotal;
   sesum = stderr*xtotal;
   lowercls = lowercl*xtotal;
   uppercls = uppercl*xtotal;

proc print data = ratioout2;
run;
```

Obs	RatioLabel	NumeratorName	DenominatorName	Ratio	StdErr	LowerCL	UpperCL	xtotal	ratiosum	sesum
1	acres92/acres87	acres92	acres87	0.986565	0.005750	0.97524871	0.99788176	964470625	951513190.87	5546161.99

lowercls	uppercls
940598734.13	962427647.61

Performing ratio estimation in SAS

- SURVEYMEANS does not have a built-in function for calculating the ratio estimator of total.
- It is easy to make a ratio estimator from SURVEYMEANS because you can request an estimate of a ratio
- So to compute a ratio estimator, simply define a new variable

$$txy = t_x * y$$

and then proceed with computing the ratio

$$\frac{\overline{txy}}{\bar{x}}$$

Another way to get the total estimate of y from ratio estimation

```
data agsrs;
   infile agsrs delimiter= ',' firstobs = 2;
   input county $ state $ acres92 acres87 ;
sampwt = 3078/300;
txy=964470625*acres92;* total estimate of y from ratio
estimation;
run;


proc surveymeans data=agsrs total=3078 mean stderr clm
sum clsum  ;
   var acres92 acres87;
   ratio  txy/acres87;
   weight sampwt;
   ods output Statistics=statsout Ratio=ratioout3;
   run;
```

The SAS System

Obs	NumeratorName	DenominatorName	Ratio	StdErr	LowerCL	UpperCL
1	txy	acres87	951513191	5546162	940598734	962427648

R links

- http://r-survey.r-forge.r-project.org/survey/html/svyratio.html
- https://stats.idre.ucla.edu/r/faq/how-can-i-do-ratio-estimation-with-survey-data/
- https://rpubs.com/trjohns/survey-ratioreg