STA304/1003 H1 F - Summer 2014: Surveys, Sampling, and Observational Data

Lecture 6: Ratio and Regression Estimation

Ramya Thinniyam

June 3, 2014

Ratio and Regression Estimation

- Use supplemental information in the estimators
- Use variables that are correlated with the variable of interest to improve precision of estimators

Ratio Estimation in a SRS

- Two quantities x_i and y_i measured on each unit:
- x_i called auxiliary/subsidiary variable
- y_i is the variable of interest
- Population totals: $\tau_y = \sum_{i=1}^{N} y_i$ and $\tau_x = \sum_{i=1}^{N} x_i$
- Ratio: $R = \frac{\tau_y}{\tau_x} = \frac{\bar{y}_U}{\bar{x}_U} \left(= \frac{\mu_y}{\mu_x} \right)$
- A SRS of size n is taken and the information in both x and y is used to estimate R, τ_y, or $\bar{y}_U$

Note: textbook and lecture notes may use different notation - $\mu_x = \bar{x}_U$ and $\mu_y = \bar{y}_U$ all represent population means.

Make use of the population correlation coefficient

$$\rho = \frac{\sum_{i=1}^{N}(x_i - \bar{x}_U)(y_i - \bar{y}_U)}{(N-1)S_x S_y}$$

The higher the correlation, the better the estimation. i.e. stronger coefficient

Estimators using SRS:

- Estimator of Population Ratio, R: $r = \hat{R} = \frac{\bar{y}}{\bar{x}} = \frac{\hat{\tau}_y}{\hat{\tau}_x}$
- Estimator of Population Total, τ_y: $\hat{\tau}_{yr} = r\,\tau_x$
- Estimator of Population Mean, $\bar{y}_U$: $\hat{\bar{y}}_r = r\,\bar{x}_U$
  where τ_x and $\bar{x}_U$ are assumed to be known.

Why Use Ratio Estimation?

2) if N is unknown, estimate N by $\frac{T_x}{\bar{x}_U}$
ex. want to estimate # of fish that are larger than 12cm in a tank but total number N of fish is unknown
y_i = length of fish i
x_i = weight of fish i
weight of entire tank = T_x
Take a sample of n fish: Don't know $\bar{x}_U$
obtain $\bar{y}$ = sample mean length
$\bar{x}$ = sample mean weight
$\hat{t}_y = \frac{T_x}{\bar{x}} = \bar{y}$
↳ est. of N
ex: population of 4000 students. SRS of 400 students → total of 240 women and 160 men → 84 women plan to teach, 40 men plan to teach
using SRS: $\hat{t}$ = total # students planning to teach = $4000 \times (\frac{124}{400}) = 1240$
Now suppose you know pop. has 2700 W & 1300 M
$\hat{t} = (\frac{84}{240} \times 2700) + (\frac{40}{160} \times 1300)$
$\hat{t}$ women $\quad$ $\hat{t}$ men

proportion: denominator is fixed
$p = \frac{\sum_{i=1}^{N} y_i}{N}$ → pop size is fixed
$\hat{p} = \frac{\sum_{i=1}^{n} y_i}{n}$ → same for all samples (if sample size fixed)

1) To Estimate a Ratio: denominator will be different for each sample chosen

   Ex. average yield per acre, ratio of number fish caught to number of hours spent fishing, income per household, percentage of pages in 'Z' magazine with typos, etc.

   Both numerator and denominator are random quantities. Be careful - some ratio estimates may have denominators that look like a sample size.

2) To Estimate a Population Total when N is Unknown:

   Cannot use $\hat{\tau}_y = N\bar{y}$, but know $N = \frac{T_x}{\bar{x}_U}$ so estimate N as $\frac{T_x}{\bar{x}}$

3) To Increase Precision of estimates:

   Sampling distribution of ratio of two RVs with high correlation will have smaller variability if $\bar{x}$ & $\bar{y}$ are highly correlated. the sampling dist'n of $\frac{\bar{y}}{\bar{x}}$ has this property: $Var(\frac{\bar{y}}{\bar{x}}) \leq Var(\bar{y})$
   ratio est. $\quad$ SRS

4) To Adjust Estimates to Reflect Demographic Totals:

   Poststratification: ratio estimation within each strata/group

   4). to est. $\hat{t}_{women}$ $x_i = I(i \text{ is women})$ $y_i = I(i \text{ is a woman \& plans to teach})$
   $\hat{t}_{women} = (\bar{y}_w / \bar{x}_w) T_{x\,women} = (\frac{84}{240}) 2700 \Rightarrow$ ratio estimate using gender as auxilary variable.

5) To Adjust for Non-Response:

   Instead of using $\hat{\tau}_y = N\bar{y}$, if smaller units are less likely to respond to survey we can use ratio estimator $\hat{\tau}_y = \frac{T_x}{\bar{x}}\bar{y}$.

Example: Number of Animals in National Park

A national park has total acreage of 2.1km^2, divided into 2873 units by m^2. We wish to estimate the total number of animals in the park by sampling 20 units.

$N=2873$ block units sampling unit = block/unit
$n=20$ x_i = size of unit i in m^2
y_i = # animals in unit i.

Define this as a ratio estimation problem. Give a point estimate for each of the following:

$\tau_x = 2.1 \times 1000^2 = 2100{,}000\ m^2$
x & y are positively correlated by scatterplot
$\bar{y}=25.4$, $\bar{x}=731$ (from output).

a) population ratio of interest

$R = \frac{\tau_y}{\tau_x}$ $\hat{R} = r = \frac{\bar{y}}{\bar{x}} = \frac{25.4}{731} = 0.0347$ animals per m^2

b) population total number of animals

total # of animals in park: τ_y, $\hat{\tau}_y = r\tau_x = 0.0347 \times 2100000 = 72968.54$

c) population mean number of animals

Mean # animals: $\bar{y}_u$ $\hat{\bar{y}}_u = \hat{\bar{y}}_{reg} = r\bar{x}_u = 0.0347\frac{\tau_x}{N} = 0.0347\frac{2100000}{2873}$

$\hat{\bar{y}}_{reg} = 25.3980$

or $\hat{\bar{y}}_{reg} = \frac{\hat{\tau}_y}{N} = \frac{72968.54}{2873} = 25.3980$

R Code and Output

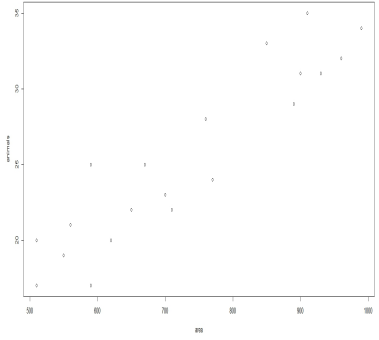
```
> parkdata <- read.csv("ratio_parkanimals.csv")
```

```
> parkdata
   area animals
1   890      29
2   770      24
3   900      31
4   560      21
5   930      31
6   670      25
7   590      17
8   510      17
9   550      19
10  960      32
11  910      35
12  700      23
13  710      22
14  760      28
15  990      34
16  510      20
17  590      25
18  650      22
19  850      33
20  620      20
```

```
> attach(parkdata)
```

```
> mean(animals)
[1] 25.4
> mean(area)
[1] 731
```

```
> plot(area,animals)
```



Properties of Ratio Estimators in SRS

- Bias: Usually biased for $\bar{y}_U$ and τ_y (unlike the estimators $\bar{y}$ and $N\bar{y}$ in SRS).

  Bias occurs because of the factor $\frac{\bar{x}_U}{\bar{x}}$ in $\hat{\bar{y}}_r$.

- Variance: More precise (lower variance) estimators due to the high correlation between x and y.

  Reduced variance usually compensates for the bias.

- Mean Squared Error (MSE): Use MSE to compare estimators (tradeoff between bias and variance).

  • Low variance means the estimator is precise - different samples lead to similar estimates ⇒ low bias

  • Unbiasedness or low bias means the average of estimates from different samples is equal (approximately equal) the parameter of interest

  • Low MSE means the estimator is accurate - different samples yield estimates that are close to the true value and close to each other

Estimators using SRS

Let $res_i = y_i - rx_i$: ith residual from fitting the line $y = rx + \epsilon$

$s_r^2 = \frac{1}{n-1}\sum_{i\in S} r_i^2$: sample variance of residuals

- Estimator of Population Mean, $\bar{y}_U$: $\hat{\bar{y}}_r = r\,\bar{x}_U$; τ_x and $\bar{x}_U$ are known.
  - $Bias(\hat{\bar{y}}_r) = -Cov(r, \bar{x})$
  - $Bias(\hat{\bar{y}}_r) \approx \frac{1}{\bar{x}_U}[r\ V(\bar{x}) - Cov(\bar{x}, \bar{y})] = (1 - \frac{n}{N})\ \frac{1}{n\bar{x}_U}(rS_x^2 - \rho S_x S_y)$
  - $V(\hat{\bar{y}}_r) \approx MSE(\hat{\bar{y}}_r) \approx (1 - \frac{n}{N})\ \frac{S_y^2 - 2r\rho S_x S_y + r^2 S_x^2}{n}$
  - $\hat{V}(\hat{\bar{y}}_r) = (1 - \frac{n}{N})\left(\frac{\bar{x}_U}{\bar{x}}\right)^2 \frac{s_r^2}{n}$
- Estimator of Population Ratio, R: $r = \frac{\bar{y}}{\bar{x}} = \frac{\hat{\tau}_y}{\hat{\tau}_x}$
  - $\hat{V}(r) = (1 - \frac{n}{N})\ \frac{s_r^2}{n}\frac{1}{\bar{x}_U^2}$: if population mean $\bar{x}_U$ is unknown, approximate using $\bar{x}$ in formula.
- Estimator of Population Total, τ_y: $\hat{\tau}_{yr} = r\,\tau_x$
  - $\hat{V}(\hat{\tau}_{yr}) = (1 - \frac{n}{N})\ (\frac{\tau_x}{\bar{x}})^2 \frac{s_r^2}{n}$
  - $\hat{V}(\hat{\tau}_{yr}) = N^2\,(1 - \frac{n}{N})\ \frac{s_r^2}{n}$ if N and $\bar{x}_U$ are unknown.

With large sample sizes, approximate $100(1-\alpha)\%$ CIs are:
$\hat{\bar{y}}_r \pm z_{\alpha/2}SE(\hat{\bar{y}}_r)$, $r \pm z_{\alpha/2}SE(r)$, $\hat{\tau}_{yr} \pm z_{\alpha/2}SE(\hat{\tau}_{yr})$

Example: Number of Animals in National Park

Use the park data to find a 95% CI for the population ratio of the mean of the number of animals to mean of unit size.

linear model ← no intercept

```
> regmodel <- lm(animals ~ 0 + area)
> summary(regmodel)
Call:
lm(formula = animals ~ 0 + area)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4695 -1.3574 -0.4492  1.6631  4.5305

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
area 0.0346942  0.0006732   51.54   <2e-16 ***
---
Residual standard error: 2.25 on 19 degrees of freedom
Multiple R-squared: 0.9929,     Adjusted R-squared: 0.9925
F-statistic:  2656 on 1 and 19 DF,  p-value: < 2.2e-16

> res = residuals(regmodel)
> var(res)
[1] 5.061587
res2 = animals -  (0.0346942*area)
> var(res2)
[1] 5.06159
> mean(animals)
[1] 25.4
> mean(area)
[1] 731
> length(animals)
[1] 20
```

x ~ variable

regmodel_name ← lm(y ~ 0 + x)

response; no intercept

95% CI for R: $r \pm 1.96\sqrt{(1-\frac{n}{N})\frac{S_r^2}{n\bar{x}_u^2}}$

$\bar{x}_u = \frac{T_x}{N} = 730.94$ NOT known → use sample $\bar{x}$

$\bar{x} = 731$

$0.034 \pm 1.96\sqrt{(1-\frac{20}{2873})\frac{5.0616}{20(731)^2}} = (0.0334, 0.0360)$

$S_r^2 = 5.0616$ — from output

When is Ratio Estimation better than SRS?

If x and y are perfectly correlated, then the estimators from SRS and Ratio are the same and there is no estimation error.

$$MSE(\hat{\bar{y}}_r) \leq MSE(\bar{y}) \text{ iff } \rho \geq \frac{rS_x}{2S_y} = \frac{CV(x)}{2\ CV(y)}$$

$\rightarrow$ If CVs are approximately equal, then Ratio estimation will be better when the correlation between x and y is bigger than 0.5.

Most appropriate to use ratio estimation when we can model the relationship between x and y with a straight line through origin and variance of y is proportional to x

Regression Estimation in SRS

Ratio estimation works best when we can fit data by a straight line through the origin.

In some cases, we have a model with an intercept, i.e.
$\hat{y}_i = a + bx_i$ where a: intercept, b: slope

Ordinary Least Squares estimates are:

$$
\begin{aligned}
b &= \frac{\sum_{i \in \mathcal{S}} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i \in \mathcal{S}} (x_i - \bar{x})^2} \\
a &= \bar{y} - b\bar{x}
\end{aligned}
$$

$y_i = a + b x_i$

$\hat{\beta}_0$ $\hat{\beta}_1$

To estimate $\bar{y}_u$: plug in $\bar{x}_u$ in regression eqn

Use correlation to increase precision of estimators.

Regression Estimators

- Estimator of $\bar{y}_U$:
  - If $\bar{x}_U$ is known, $\hat{\bar{y}}_{reg} = a + b\bar{x}_U = \bar{y} + b(\bar{x}_U - \bar{x})$
  - $Bias(\hat{\bar{y}}_{reg}) = -Cov(b, \bar{x})$
  - $V(\hat{\bar{y}}_{reg}) \approx MSE(\hat{\bar{y}}_{reg}) = \left(1 - \frac{n}{N}\right) \frac{s_d^2}{n}$

    where $d_i = y_i - [\bar{y}_U + b(x_i - \bar{x}_U)]$ a type of residual.
  - $\hat{V}(\hat{\bar{y}}_{reg}) = \left(1 - \frac{n}{N}\right) \frac{1}{n} \left(\frac{\sum_{i=1}^{n} (y_i - (a+b)x_i)^2}{n-2} \right) = \left(1 - \frac{n}{N}\right) \frac{MSE}{n}$

    where MSE is the MSE from the standard simple linear regression of y on x.

Ex: regression estimation (motivate a non-zero intercept)
- interested in mean sugar content of Mr. Christie choc chip cookies.

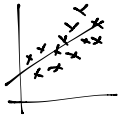
y_i = sugar content of ith cookie

x_i = # of choc chips on the ith cookie

Number of choc chips has high correlation with sugar content.
But a cookie can have a non-zero sugar content even if it has no chips.

$\hat{y}_i = a + bx_i$

sugar content → # choc chips



Example: Dead Trees

To estimate the number of dead trees in an area, we divide the plot into 100 square plots and count the number of dead trees on a photograph of each plot. Photos counts are quick to do but sometimes trees are misclassified or not detected. So we select a SRS of 25 plots and do field counts of dead trees. The population mean number of dead trees from the photo count is 11.3.

Set this up as a regression estimation problem. Find a 95% CI for the mean number of dead trees from the field count using the output.

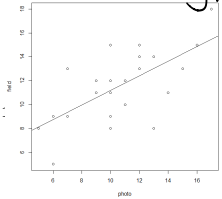
y_i = fields count of the ith plot
x_i = photo count of the ith plot
$N = 100$
$n = 25$
sampling unit : a plot
observation unit: a tree
$\hat{y}_i = a + b\hat{x}_i$
want CI for $\bar{y}_U$

(See next page)

R Code and Output

```
> deadtrees <- read.csv("reg_deadtrees.csv")
> deadtrees
   field photo
1     15    10
2     14    12
3      9     7
4     14    13
.
.
25     8    10
> attach(deadtrees)
> regmodel <- lm(field ~ photo)
> summary(regmodel)

Call:
lm(formula = field ~ photo)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0319 -1.8053  0.1947  1.4212  3.8080

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.0593     1.7635   2.869 0.008676 **
photo         0.6133     0.1601   3.832 0.000854 ***
---
Residual standard error: 2.406 on 23 degrees of freedom
Multiple R-squared: 0.3896,     Adjusted R-squared: 0.3631
F-statistic: 14.68 on 1 and 23 DF,  p-value: 0.0008538
```

```
> mean(field)
[1] 11.56
> mean(photo)
[1] 10.6
> var(field)
[1] 9.09
> var(photo)
[1] 9.416667
> var(residuals(regmodel))
[1] 5.548341
> anova(regmodel)

Analysis of Variance Table
Response: field
          Df Sum Sq Mean Sq F value Pr(>F)
photo      1  85.00   85.00  14.681 0.0008538
Residuals 23 133.16    5.79

> plot(photo,field)
> abline(regmodel)
```



$\bar{y}=11.56$ $a=5.0593$

$\bar{x}=10.6$ $b=0.6133$

$MSE=5.79$ ① ↳ from ANOVA

or ② $=(2.406)^2$ ↳ residual standard error from 'lm' summary

③ $MSE=\frac{n-1}{n-2}S_r^2$ (small adjustment)

$=\frac{24}{23}5.5483$

$S_r^2=\frac{1}{n-1}\sum_{i=1}^{n}[y_i-(a+bx_i)^2]=\frac{1}{n-1}SSE$

$\hat{\bar{y}}_{reg}=\bar{y}+b(\bar{x}_u-\bar{x})$ $\bar{x}_u=11.3$ given in question

$=11.56+0.6133(11.3-10.6)$

$=11.99$

95% CI for $\bar{y}_u$:

$\hat{\bar{y}}_{reg}\pm 1.96\sqrt{(1-\frac{n}{N})\frac{MSE}{n}}=[11.1731, 12.8069]$

a
b
$\sqrt{MSE}$
25-2