Lecture 6 - Key

Ratio and Regression Estimation

Ratio Estimation

Suppose the research believes an auxiliary variable (or covariate) x is associated with the variable of interest y.

- variable of interest: *Number of bicycles rented*
- auxiliary variable: *amount of precipitation*

- variable of interest: *Income level*
- covariate: *number of years of education*

Situation: we have bivariate (X, Y) data and assume there is a positive proportional relationship between X and Y. *That is, on every sampling unit we take a pair of measurements and assume that* $Y \approx BX$ *for some value* $B > 0$.

Visual summary of ratio estimation.

The population correlation coefficient of x and y is:

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

where S_x and S_y are the population standard deviations of x and y.

There are two cases that may be be of interest to the researcher:

1. To estimate the ratio of two population characteristics. *The most common case is the population ratio of means or totals:*

$$B = \frac{\bar{y}_U}{\bar{x}_U} = \frac{t_y}{t_x}$$

2. To use the relationship *between X and Y to improve the estimation of t or $\bar{y}_U$.*

Note that ratio estimation *is not a sampling scheme, but rather a way to do estimation.*

The sampling plan will be to take a SRS of n pairs $(x_1, y_1), \ldots, (x_n, y_n)$ from the population of N pairs. We will use the following notation.

$$\bar{x}_U = \left(\sum_{i=1}^{N} x_i\right)/N \qquad t_x = \sum_{i=1}^{N} x_i \qquad \bar{y}_U = \left(\sum_{i=1}^{N} y_i\right)/N \qquad t_y = \sum_{i=1}^{N} y_i$$

$$\bar{x} = \left(\sum_{i=1}^{n} x_i\right)/n = \text{sample mean of x's} \qquad \bar{y} = \left(\sum_{i=1}^{n} y_i\right)/n = \text{sample mean of y's}$$

Estimation of B, $\bar{y}_U$, t_y

Case 1: t_x and $\bar{x_u}$ are known

First consider estimating B assuming t_x and $\bar{x}_U$ are known. The ratio estimator $\hat{B}$ is the ratio of the sample means and its estimated variance $\hat{V}(\hat{B})$ are

$$\hat{B} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{\bar{y}}{\bar{x}} \qquad \hat{V}(\hat{B}) = \left(\frac{N-n}{N\bar{x}^2}\right)\frac{s_e^2}{n},$$

where

$$s_e^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \hat{B}x_i)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 + \hat{B}^2\sum_{i=1}^{n} x_i^2 - 2\hat{B}\sum_{i=1}^{n} x_i y_i\right)$$

If $Y \approx BX$, *then* $y_i \approx \hat{B}x_i$. *Thus,* $\hat{B}x_i$ *can be considered the predicted value of* y_i *from a line through the origin (with intercept =0 and slope =* $\hat{B}$*.)*

The distribution of $\hat{B}$ is very complicated. For small samples $\hat{B}$ is likely to be skewed and is biased for B. For large samples, the bias is negligible (very small) and the distribution of $\hat{B}$ tends to be approximately normal.

Multiplication of $\hat{B}$ and $\hat{V}(\hat{B})$ by $\bar{x}_U$ and $\bar{x}_U^2$, respectively, yield the estimator $\hat{\bar{y}}_r$ for $\bar{y}_U$ and its estimated variance:

$$\begin{aligned} *\hat{\bar{y}}_r &= \left(\frac{\bar{y}}{\bar{x}}\right)\bar{x}_U * \\ *\hat{V}(\hat{\bar{y}}_r) &= \hat{V}(\hat{B}\bar{x}_U) = \bar{x}_U^2\hat{V}(\hat{B}) = \left(\frac{N-n}{N}\right)\left(\frac{\bar{x}_U}{\bar{x}}\right)^2\frac{s_e^2}{n}* \end{aligned}$$

$\hat{\bar{y}}_r$ is called the *ratio estimator of the population mean.*

By multiplying the above formulas by N and N^2, respectively, we get the estimator $\hat{t}_{yr}$ of t_y and its associated estimated variance:

$$\begin{aligned} \hat{t}_{yr} &= N\left(\frac{\bar{y}}{\bar{x}}\right)\bar{x}_U \\ \hat{V}(\hat{t_{yr}}) &= N(N-n)\left(\frac{\bar{x}_U}{\bar{x}}\right)^2\frac{s_e^2}{n} = \left(\frac{N-n}{N}\right)\left(\frac{t_x}{\bar{x}}\right)^2\frac{s_e^2}{n} \end{aligned}$$

t_{yr} is called the ratio estimator of the population total.

If N is unknown but we know N is large relative to n, then the FPC $(N-n)/N \approx 1$, typically researchers will replace the FPC with 1.

Example: Suppose we can predict MSU statistics students future income, using the average GPA in statistics courses. Let y be income and x be GPA. Suppose that the $\bar{y} = \$70,000$ and $\bar{x} = 3.50$. Sampled from students that have taken STAT 446.

Compute $\hat{B} = \frac{70,000}{3.50} = 20,000$.

It is known that the average GPA for all statistics students is 3.00. Compute $\hat{\bar{y}}_r = \hat{B}\bar{x}_U = 60,000$.

Case 2: t_x and $\bar{x}_U$ are unknown

If t_x and $\bar{x}_U$ are unknown, it will not affect the estimator $\hat{B} = \bar{y}/\bar{x}$. *It will, however, affect the estimators* $\hat{y}_r$ *and* $\hat{t}_{yr}$ *that depend on* t_x *and* $\bar{x}_U$.

In such cases, it is common to replace t_x with $N\bar{x}$ or replace $\bar{x}_U$ with $\bar{x}$. then $\hat{\bar{y}}_r = \bar{y}$ *and* $\hat{t_y r} = N\bar{y}$

This will yield:

$$\hat{V}(\hat{\bar{y}}_r) \approx \left(\frac{N-n}{n}\right)\frac{s_e^2}{n} \qquad\qquad \hat{V}(\hat{t}_{yr}) = N(N-n)\frac{s_e^2}{n}$$

When $\bar{x}$ is larger than $\bar{x}_U$, $\hat{V}(\hat{\bar{y}}_r)$ and $\hat{V}(\hat{t}_{yr})$ tend to be too large as variance estimates. Similarly, when $\bar{x}$ is smaller than $\bar{x}_U$, $\hat{V}(\hat{\bar{y}}_r)$ and $\hat{V}(\hat{t}_{yr})$ tend to be too small as variance estimates.

Example: Demonstration of Bias in Ratio Estimation

There are $N = 4$ sampling units in the population, where x is the number of aspen trees and y is the number of Ponderosa pine trees:

x_i value	67	63	66	69
y_i value	68	62	64	70

Let $n = 2$, then there are 6 possible samples. The total abundances are $t_x = 265$ and $t_y = 264$. Therefore, $B = 264/265 = 0.9962$. Also, $S_x^2 = 6.250$ and $S_y^2 \approx 13.3$.

Sample	Units	$\hat{t}_{yr}$	$\hat{t}_{SRS}$
1	1,2	*265*	260
2	1,3	*263.0075*	264
3	1,4	*268.8971*	276
4	2,3	258.8372	*252*
5	2,4	265	*264*
6	3,4	263.0370	*268*

Note that $E(\hat{t}_{yr}) = 263.963$*, so the ratio estimator is biased.*

Recall that the $MSE(\hat{\theta}) = E\left[(\hat{\theta}-\theta)^2\right]$, it turns out the ratio estimator has much smaller variance.

```r
t.y <- 265
t.yr <- c(265,263.0075,268.8971,258.8372,265,263.0370)
t.srs <- c(260,264,276,252,264,268)
MSE.ratio <- mean((t.yr - t.y)^2); MSE.ratio
```

```
## [1] 10.16515
```

```r
MSE.srs <- mean((t.srs - t.y)^2); MSE.srs
```

```
## [1] 54.33333
```

Bias and MSE of Ratio Estimators

As we have seen ratio estimators are biased. The bias occurs in the ratio estimation because $E\left[\frac{\bar{y}}{\bar{x}}\right] \neq \frac{E[\bar{y}]}{E[\bar{x}]}$. *That is the expected value of the ratio is not equal to the ratio of the expected values.*

However, when used appropriately the *reduction in variance from the ratio estimator will offset the presence of bias.*

Also for large samples, the estimators t_{yr} and $\bar{y}_r$ will be approximately normally distributed.

The bias of $\hat{\bar{y}}_r$ as well as ($\hat{t}_{yr}$ and $\hat{B}$) will be small if

1. *the sample size n is large*

2. *the sampling fraction n/N is large*

3. *S_x is small*

4. *the correlation, R, is close to 1.*

Note that if the $x's$ all have the same value the ratio estimator reduces to the SRS estimator $\bar{y}$ and is unbiased.

The approximate MSE of $\hat{\bar{y}}_r$ will be small when:

1. *the sample size n is large*

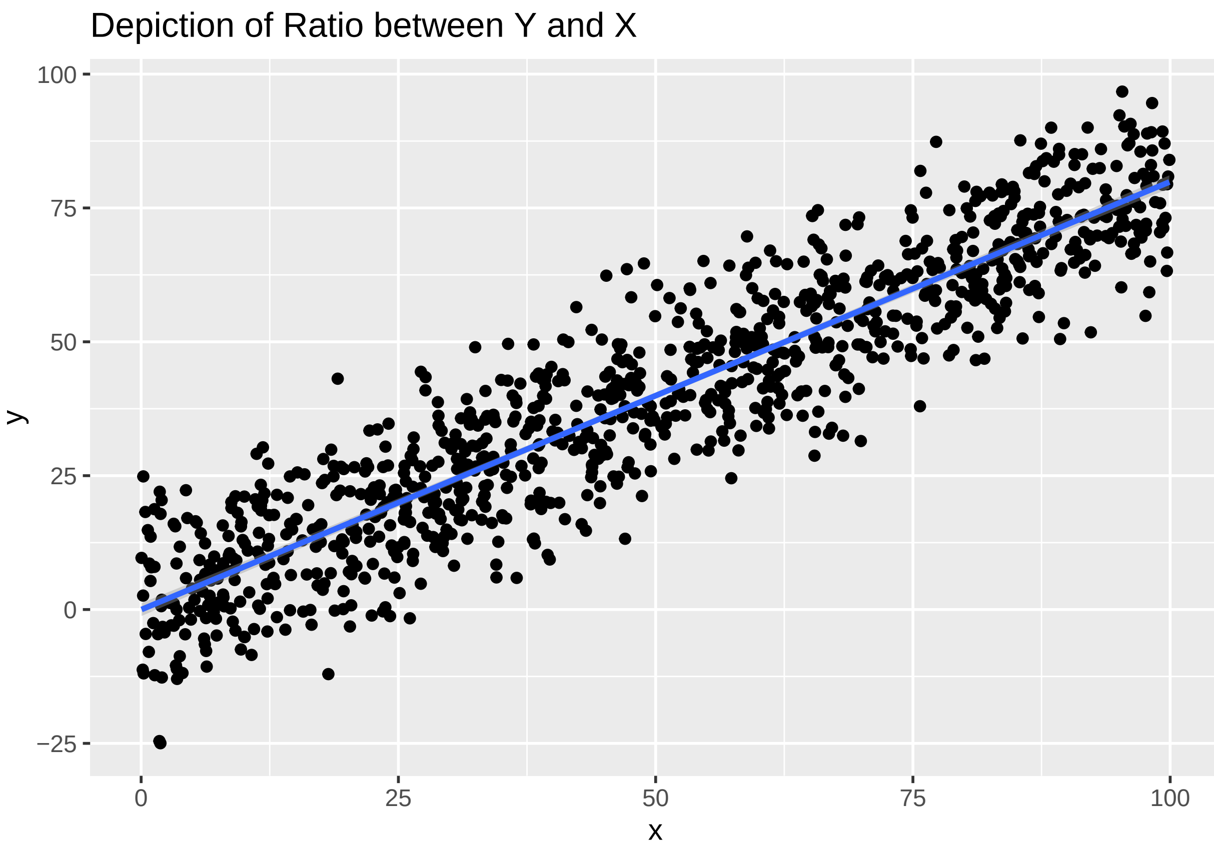
2. *the sampling fraction n/N is large*

3. *the deviations (residuals) $y_i - Bx_i$ are small*

4. *the correlation, R, is close to 1.*

Ratio Estimation in R

```r
# Simulate Ratio Estimation Data
set.seed(10162019)
N <- 1000
x <- runif(N,0,100)
t.x <- sum(x)
x.mean <- mean(x)
B <- .8
sigsq <- 100
y <- x*B + rnorm(N,0,sqrt(sigsq))
t.y <- sum(y)
ratio_df <- tibble(x=x, y=y)
```




```r
# Take SRS
n <- 50
sample_vals <- ratio_df %>% sample_n(n)

# Compute Estimates of t.y
y_bar <- sample_vals %>% summarize(mean(y)) %>% pull()
SRS_estimate <- N * y_bar

B <- y_bar / sample_vals %>% summarize(mean(x)) %>% pull()

ratio_estimate <- B * x.mean * N
```

In this case, $t_y = 3.9668 \times 10^4$. Based on a single sample, the ratio estimator is off by 812 and the SRS estimator is off by 1739