Introduction to Ratio and Regression Estimation

Introduction to Ratio Estimation

- Ratio estimation is a technique that uses available *auxiliary information* which is correlated with the variable of interest.

- Suppose that a variable $X$ is correlated with a variable of interest $Y$, and we have a paired random sample of $n$ observations $(x_i, y_i)$ for $i = 1, \ldots, n$.

Then, we define the ratio

$$R \equiv \frac{\tau_y}{\tau_x} = \frac{\mu_y}{\mu_x}$$

and the corresponding estimator is:

$$r \equiv \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{\bar{y}}{\bar{x}}.$$
• Both, the numerator and the denominator are random quantities.

• The estimated sampling variance of $r$ is

$$
\hat{\text{Var}}(r) = \left(1 - \frac{n}{N}\right) \frac{1}{\mu_x^2} \sum_{i=1}^{n} (y_i - r x_i)^2 \frac{1}{n(n - 1)}
$$

• The estimated variance can be written also in terms of the coefficient of correlation $\rho$:

$$
\hat{\text{Var}}(r) = \left(1 - \frac{n}{N}\right) \frac{1}{\mu_x^2} \frac{1}{n} \left(s_y^2 + r^2 s_x^2 - 2 r \hat{\rho} s_x s_y\right)
$$

• Note that we can substitute $\mu_x^2$ by its estimator $\overline{x}^2$ in both cases.
Ratio Estimate Examples

$X \equiv$ Family Size

$Y \equiv$ Food Consumption $\implies R \equiv$ Food Consumption per Capita

$X \equiv$ Labor Force Size

$Y \equiv$ Number Unemployed $\implies R \equiv$ Unemployment Rate

$X \equiv$ Cell Phones: 2000

$Y \equiv$ Cell Phones: 2005 $\implies R \equiv$ Increase Rate

$X \equiv$ Person – hours

$Y \equiv$ Number of Items Processed $\implies R \equiv$ Productivity Rate
**Example:**

It is interesting to know the relative change over a two-year period in the assessed value of homes in a given community. We take a simple survey sampling of $n = 20$ homes from the $N = 1000$ total homes in the community. We obtain the values for this year ($y$) and the corresponding values from two years ago ($x$) for each of the $n = 20$ homes included in the sample.

We want to estimate the relative change ($R$) in the assessed values for the $N = 1000$ homes (see the original example in Scheaffer et al. (1990))

Data are collected in two vectors:

\[
x = c(6.7, 8.2, 7.9, 6.4, 8.3, 7.2, 6.0, 7.4, 8.1, 9.3, 8.2, 6.8, 7.4, 7.5, 8.3, 9.1, 8.6, 7.9, 6.3, 8.9)
\]

\[
y = c(7.1, 8.4, 8.2, 6.9, 8.4, 7.9, 6.5, 7.6, 8.9, 9.9, 9.1, 7.3, 7.8, 8.3, 8.9, 9.6, 8.7, 8.8, 7.0, 9.4)
\]
N = 1000

n = length(x)

plot(x,y)

n = length(x)

r = sum(y)/sum(x)

r

var.r = (1-(n/N))*(1/mean(x)^2)*sum((y-r*x)^2)/(n*(n-1))

var.r

down = r - qnorm(0.975)*sqrt(var.r)

up = r + qnorm(0.975)*sqrt(var.r)

cat("Confidence Interval: ", "[", down, ";", up, "]", "\n")
• The ratio technique can be used to estimate a population total $\tau_y$ when we do not know $N$. In this case we must know the total of the *auxiliary variable* $x$, namely, $\tau_x$.

\[
\hat{\tau}_y = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} \cdot \tau_x = r \tau_x
\]

• The estimated variance of $\hat{\tau}_y$ is:

\[
\hat{\text{Var}}(\hat{\tau}_y) = \hat{\tau}_x^2 \left(1 - \frac{n}{N}\right) \frac{1}{\mu_x^2} \frac{\sum_{i=1}^{n} (y_i - rx_i)^2}{n(n - 1)}
\]

• In the same way, it can be estimated a population mean $\mu_y$, when we do not know $N$:

\[
\hat{\mu}_y = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} \cdot \mu_x = r \mu_x
\]

• The estimated variance of $\hat{\mu}_y$ is:

\[
\hat{\text{Var}}(\hat{\mu}_y) = \mu_x^2 \hat{\text{Var}}(r) = \left(1 - \frac{n}{N}\right) \frac{\sum_{i=1}^{n} (y_i - rx_i)^2}{n(n - 1)}
\]
We can substitute $\mu_x^2$ by its estimator $\overline{x}^2$.

See a **function** to calculate **ratio estimators** programmed in \texttt{R}.

```r
ratio.srs <- function(x, y, opt="Ratio", tauX=NA, N=NA) {
  # opt = "Tau" for the total of Y
  # opt = "Mu" for the mean of Y
  n <- length(x)
  if(is.na(N)) {fpc <- 1} else {fpc <- 1-(n/N)}
  ratio <- sum(y)/sum(x)
  if(is.na(tauX) & is.na(N)) {meanX <- mean(x)} else {meanX <- tauX/N}
  var.r <- fpc*(1/meanX^2)*(sum((y-ratio*x)^2)/n*(n-1))
}
```
switch(opt,

   "Ratio" = {theta <- ratio

           var.theta <- var.r},

   "Tau" = {theta <- ratio*tauX

            var.theta <- var.r*tauX^2},

   "Mu" = {theta <- ratio*meanX

            var.theta <- var.r*meanX^2}

)

B <- 2*sqrt(var.theta)

   cat("Parameter",theta,"

   cat("Variance Parameter",var.theta,"

   cat("Confidence Interval: ","[",theta-B,";",theta+B,"]","\n
)}
We can consider an example about the ratio of prizes between a couple of years (1994 and 1996).

```r
price.94 <- c(48.2, 30.236, 0.919, 1.109, 1.043, 0.768, 1.892, 0.899, 0.917, 1.457, 0.789, 0.505, 0.440, 1.604, 1.674, 2.530, 0.506)
price.96 <- c(49.231, 31.438, 1.121, 1.318, 1.260, 0.875, 1.848, 1.002, 1.308, 1.652, 0.886, 0.593, 0.481, 1.210, 1.735, 3.307, 0.622)
ratio.srs(x=price.94, y=price.96, opt="Ratio")
```
Alternative for ratio estimators using **weighted regression**:

We use the variables of the data from example of *Synthetic Data* (p. 15).

```r
fit <- lm(formula = rent ~ -1 + income,  # the ’-1’ removes the intercept
data = srs, weights = 1/income)  # the weight is specified as 1/X

# Standard error formula of the ratio estimator which
# includes the finite population correction (fpc) factor n/N
ratio.se <- function(mux, s.diff, n, N) {
  sqrt((1/mux^2)*((s.diff^2)/n)*(1-(n/N)))
}
```
# Derive the correct fpc corrected Standard Error of the ratio

```r
define the correct fpc corrected Standard Error of the ratio

reg.ratio.se <- ratio.se(
  mux = mean(srs$income),
  s.diff = sd(fit$resid),  # We use the model residuals
  n = nrow(srs),
  N = N
)

cat( "Estimated ratio: ", round(fit$coeff, 3), "\n", ’ (SE = ’, round(reg.ratio.se, 5), ’)’,
    sep="", "\n" )
```
Ratio Estimation in Stratified Random Sampling

- There are two different methods to construct estimators of a ratio in stratified sampling.

- **Separate Ratio Estimator**: Estimate the ratio of $\mu_y$ to $\mu_x$ within each stratum and then form a weighted average of the separated estimates.

- **Combined Ratio Estimator**: Compute the usual $\bar{y}_{st}$ and $\bar{x}_{st}$, then use their quotient as an estimator of $\frac{\mu_y}{\mu_x}$.

- If the stratum sample sizes are large (more than 20) it is better to use separate ratio estimators. Otherwise, if the sample sizes are small or the within-stratum ratios are approximately equal, it is better to use combined ratio estimators.
Introduction to Regression Estimation

• When the auxiliary variable $X$ is a predetermined (non-random) variable, we can obtain an alternative estimator to the ratio estimator.

• It is based on the concept of least squared method and it is known as regression estimation.

• Assuming there is a linear relationship between $X$ and $Y$

$$
\hat{y}_i = a + bx_i = \bar{y} + b(x_i - \bar{x})
$$

with paired observations $(x_i, y_i)$ for $i = 1, \ldots n$. Then the estimator of a population mean $\mu_y$ is

$$
\hat{\mu}_y = \bar{y} + b(\mu_x - \bar{x})
$$

where

$$
b = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}
$$
• The estimated variance of \( \hat{\mu}_{yL} \) is

\[
\hat{\text{Var}}(\hat{\mu}_{yL}) = \left( \frac{N - n}{Nn} \right) \left( \frac{1}{n - 2} \right) \left[ \sum_{i=1}^{n} (y_i - \bar{y})^2 - b^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \right] = \\
= \left( \frac{N - n}{Nn} \right) \cdot MSE
\]

where \( MSE \) is the mean square error from the standard simple linear regression.

• In general, the ratio estimator is most appropriate when the relationship between \( x \) and \( y \) is linear through the origin. Otherwise, in general, it is better to use regression estimators.
Example of ratio and regression estimators with the library `survey` of R:

```r
# SYNTHETIC DATA
mydata <- rbind(matrix(rep("nc",165),165,1,byrow=TRUE),
                matrix(rep("sc",70),70,1,byrow=TRUE))
myx <- 100*runif(235)
myy <- myx*1.2+rnorm(235)
mydata <- cbind.data.frame(mydata,c(rep(1,100),rep(2,50),rep(3,15),
                                   rep(1,30),rep(2,40)),myx,myy)
names(mydata) <- c("state","region","income","rent")
N <- dim(mydata)[[1]]
n <- 50

# Selection of a sample
srs_rows <- sample(N,n)
srs <- mydata[srs_rows,
```
# Export data to Stata format

```r
library(foreign)
write.dta(srs,"C:/QM/mydataratio.dta")
```

```r
library(survey)
srs$popsize <- N
dsrs <- svydesign(id=~1, fpc=~popsize, data=srs)
summary(dsrs)
```

```r
svyratio(~rent, ~income, design=dsrs)
```

```r
eso <- svyglm(rent~income, design=dsrs)
svyplot(rent~income, design=dsrs,style="bubble" ,xlab="Income",ylab="Rent")
```

```r
summary(eso)
plot(eso)
```
Example of ratio and regression estimators with **Stata**:

```
use C:\QM\mydataratio.dta, clear

count

* Compute **weights** and the **factor** of population correction

gen pw = 235/50

gen fpc = 235

* Set the sampling design

svyset [pweight=pw], fpc(fpc)

* Ratio estimator

svy: ratio rent/income

* Regression estimator

svy: regress rent income
```