SEPTEMBER 19, 2023

LAB 7- REGRESSION AND RATIO ESTIMATOR

R-Markdown

A.MAYURI(2348133)

1MSTAT

LAB 7- REGRESSION AND RATIO ESTIMATOR

A.Mayuri(2348133)

2023-09-19

Question:

Mr. John selected a random sample using SRSWOR procedure of 21 states from a population of 50 states of a country. He collected information about the real estate farm loans and nonreal estate farm loans from the selected states given below. Now, Mr. John wants to estimate the average real estate farm loans assuming that the average nonreal estate farm loans in the country is known and is equal to $878.16. Use the ratio and regression estimators and give the estimates for this data set and discuss the 95% confidence interval. Also compare the results obtained using Ratio estimator and suggest the more efficient method to give estimate for this data set.

Variable of Interest

x-Nonreal estate farm loans

y-Real estate farm loans

Definition:

Ratio estimator: The ratio estimator is a statistical estimator for the ratio of means of two random variables. Ratio estimates are biased and corrections must be made when they are used in experimental or survey work. The ratio estimates are asymmetrical and symmetrical tests such as the t test should not be used to generate confidence intervals

Regression estimator: The ratio method of estimation uses the auxiliary information which is correlated with the study variable to improve the precision which results in the improved estimators when the regression of Y on X is linear and passes through the origin. When the regression of Y on X is linear, it is not necessary that the line should always pass through the origin. Under such conditions, it is more appropriate to use the regression type estimator to estimate the population means

Analysis

Step1: import dataset

```
library(readxl)
rare <- read_excel("C:/Users/mayur/Desktop/Mstat/tri sem 1/R/dataset/rare.xlsx")
View(rare)
attach(rare)
```

Step2: Calling the Packages

```
library(SDaA)
library(survey)
```

```
## Loading required package: grid
```

```
## Loading required package: Matrix
```

```
## Loading required package: survival
```

```
##
## Attaching package: 'survey'
```

```
## The following object is masked from 'package:graphics':
##
##     dotchart
```

Ratio estimator:

Step3: Understanding the dataset

```
head(rare)
```

```
## # A tibble: 6 × 3
##    s.no       x       y
##   <dbl>   <dbl>   <dbl>
## 1     1  348.    409.
## 2     2  431.     54.6
## 3     3  848.    908.
## 4     4 3929.   1343.
## 5     5  906.    316.
## 6     6    4.37    7.13
```

```
colnames(rare)
```

```
## [1] "s.no" "x"    "y"
```

```
nrow(rare)
```

```
## [1] 21
```

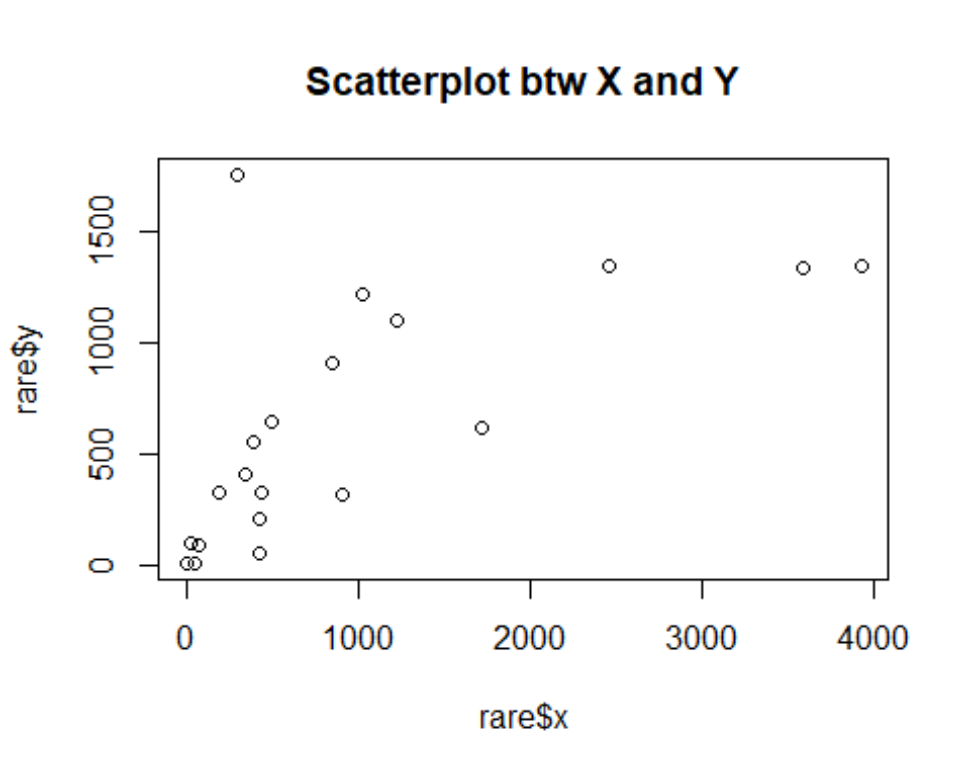
Inference:

thus we have 2 columns X,Y and 21 observations under each.

Step4: Plotting

We can use scatterplot to understand the relationship between the X and Y variables

```
plot(rare$x,rare$y, main="Scatterplot btw X and Y")
```



Now, Mr. John wants to estimate the average real estate farm loans assuming that the average nonreal estate farm loans in the country is known and is equal to $878.16.

```
ratio=svydesign(ids=~1, weights = ~1, data = rare)
svyratio(numerator = ~y, denominator = ~x,
         design = ratio)
```

```
## Ratio estimator: svyratio.survey.design2(numerator = ~y, denominator = ~x,
 design = ratio)
## Ratios=
##           x
## y 0.6695023
## SEs=
```

```
##           x
## y 0.1379684
```

Step5:

estimated mean and total using ratio estimator using(Rcap*Xbar) for y(bar)

```
#estimated mean using ratio estimator
0.6695023*878.16
```

```
## [1] 587.9301
```

```
#estimated total using ratio estimator
0.6695023*(878.16*21)
```

```
## [1] 12346.53
```

```
# we had to assume the mean non real loans as 878.16 and number of observatio
n as 21
```

Thus we get a mean estimation for real estate farm loans as 583.93 dollars and a 12346.53 dollars as the estimated total for real estate farm loans.

Step6: Confidence Interval

We have to use t distribution since the sample size is less than 30 the value of t(alpha/2) with 20 DF {(N-1)df} is 2.086 using t table.

```
r=mean(rare$y)/mean(rare$x)
r
```

```
## [1] 0.6695023
```

```
v=var(rare$y)/var(rare$x)
SE=sqrt(v)
SE
```

```
## [1] 0.4840727
```

```
Upper_confidence_level=r+2.086*SE
Upper_confidence_level
```

```
## [1] 1.679278
```

```
Lower_confidence_level=r-2.086*SE
Lower_confidence_level
```

```
## [1] -0.3402733
```

Hence at 95% confidence level os 5% of level of significance for the r value will lieof population will lie between [0,1.67]

Regression Estimator

Step7: Estimate b value

```
rerare=rare$x
set.seed(123)
reg_model=lm(y ~0+x, weights = rerare, data = rare)
summary(reg_model)
```

```
##
## Call:
## lm(formula = y ~ 0 + x, data = rare, weights = rerare)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -13745.4    -83.6    662.1   9848.5  28284.1
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## x  0.39778    0.03358   11.84 1.71e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12180 on 20 degrees of freedom
## Multiple R-squared:  0.8752, Adjusted R-squared:  0.869
## F-statistic: 140.3 on 1 and 20 DF,  p-value: 1.712e-10
```

Thus estimated b value is 0.39778

Step8: Find population mean

```
# now regression estimator is ybar_reg=ybar+b(Xbar-xbar)
b=0.39778
ybar=mean(rare$y)
xbar=mean(rare$x)
Xbar=878.16
ybar
```

```
## [1] 602.121
```

```
xbar
```

```
## [1] 899.3561
```

```
#estimated population mean using regression estimator
ybar_reg=ybar+b*(Xbar-xbar)
ybar_reg
```

```
## [1] 593.6896
```

Estimated popultion mean is 593.6896.

Step9: To find standard error

standard error- First find out the unbiased estimate of variance ybar_reg by using the formula and then take the square root of it. For this we need r, N, n, sample mean square of y variable i.e. sy^2.

note: N=50 n=21

```
r=cor(rare$y,rare$x)
N=50
n=21
sy2=var(rare$y)
# Formula application
v=((N-n)/(N*n))*(sy2-(r^2)*sy2)
v
```

```
## [1] 4639.497
```

```
SE=sqrt(v)
SE
```

```
## [1] 68.11386
```

We observe that the population variance is 4639.497. and the population std deviation is 68.11386

Conclusion:

the ratio estimator is more efficient than that of regression estimator because.

1) lesser variability

2) There is less no absolute linear relation between X and Y variable

THE END