M2 Statistics & Econometrics

Survey sampling

Second computer lab session with R

This is the subject for the second computer lab session with R on survey sampling.

1 Data presentation

The population we consider is made of the 554 "communes" of the French Haute-Garonne department with less than 10000 inhabitants in 1999. The data source is the French census of 1999 (Insee). For these communes we have :

— the code of the communes (CODE_N)

— its name (COMMUNE)

— the code of the BVQ ("Bassin de vie quotidienne") the commune belongs to (BVQ_N)

— its population (POPSDC99)

— the number of dwellings or housing units (auxiliary variable $x =$ LOG)

— a stratification variable (stratlog) with 4 categories (1 if LOG <100, 2 if 100 $\leq$ LOG $<$ 300, 3 if 300 $\leq$ LOG $<$ 1000 and 4 if LOG $\geq$ 1000)

— the number of empty dwellings or housing units (variable of interest $y =$ LOGVAC).

The file rec99htegne is the survey frame. We will condider the variable of interest $y =$ LOGVAC, which is the number of empty housing units, and the variable stratlog that will be used as the post-stratification variable and the auxiliary variable $x =$LOG, which is the number of housing units.

2 The poststratified estimator

2.1 Objectives

A the end of this section you should know how to draw a sample with a SRSWOR design and how to estimate the total number of empty housing units with a Horvitz-Thompson and a poststratified estimator. You will also be able to compare the two estimators using simulations.

2.2 Preliminary results on the population

(1) Understand and adapt the following code in R and obtain the value of N, the total Y and the variance S_y^2 of y over the whole population.

```
#Data importation
rec99<-read.csv("K:/TP_RuizGazen/rec99htegne.csv")
#transformation of the columns in vectors
attach(rec99)

# Data details
attributes(rec99)
#Preliminary steps
dim(rec99)
ty=sum(LOGVAC)
ty
var(LOGVAC)
```

(2) Understand and use the following code in R to obtain the values of N_q, $q = 1, \ldots, 4$, the totals Y_q, $q = 1, \ldots, 4$, and the variances S_{yq}^2, $q = 1, \ldots, 4$, of y for the different poststrata.

```
table(stratlog)
by(LOGVAC,stratlog,sum)
by(LOGVAC,stratlog,var)
```

2.3 The one sample case

2.3.1 Drawing a sample with the SRSWOR sampling design and calculating the HT estimator

(3) We remind you the R code to draw a sample of communes of size 70 using a SRSWOR design and estimate the number of empty dwellings in the population using the HT estimator. Explain the different steps of the code. Give the estimated variance, standard deviation and coefficient of variation of your estimator.

```
set.seed(66542)  # needed if you want to obtain always the same results
si.rec99<-srswor(70,554)

ech.si <- svydesign(id=~CODE_N, weights=rep(554/70,70),fpc = rep(554, 70),
data=rec99[which(si.rec99==1),])

est<- svytotal(~LOGVAC, ech.si)

attributes(est)
SE(est)^2
SE(est)/est[1]
est[1]-1.96 * SE(est)
est[1]+1.96 * SE(est)
```

2.3.2 Calculating the poststratification estimator

(4) For the sample drawn in the previous subsection, calculate the poststratified estimator of the total number of empty housing units by using the “postStratify” function from the package survey. In order to understand how to do so, you can use the example from the help file `http://r-survey.r-forge.r-project.org/survey/html/postStratify.html`.

(5) Recall the definition of the poststratified estimator. Use a calculator and detail the calculus in order to check that you find the same values than R for the poststratified estimate.

2.4 Simulations

In order to compare the precision of the HT and the poststratified estimator, you have to draw 1000 samples.

2.4.1 Reminder for the SRSWOR design and the HT estimator

(6) We remind you the code in order to draw 1000 samples according to a SRSWOR sampling design with size $n = 70$ and to calculate the 1000 HT estimates for the total number of empty housings together with the empirical mean [1], the empirical (or Monte Carlo) standard deviation together with the Monte Carlo coefficient of variation in percentage (standard deviation divided by mean) for the 1000 HT estimates.

[1] The empirical mean in a context of simulations is also called the Monte Carlo mean.

```
# Simulations for SRSWOR
# Initialization
nsimu<-1000
nb.simul<-matrix(1:nsimu,nsimu,1);
tu.esti<-matrix(1,nsimu,1);
var.esti.tu<-matrix(1,nsimu,1);

# Simulations
for(i in 1:nsimu)
{
si.rec99<-srswor(70,554)
ech.si <- svydesign(id=~CODE_N, weights=rep(554/70,70),
      fpc = rep(554, 70),data=rec99[which(si.rec99==1),])
a<-svytotal(~LOGVAC, ech.si)
tu.esti[i]<-a[1]
var.esti.tu[i]<-SE(a)^2
}

#Monte Carlo mean and standard deviation
mean(tu.esti)
sqrt(var(tu.esti))
sqrt(var(tu.esti))/mean(tu.esti)

# Histogram
hist(tu.esti)
```

Run this code and check that you find a coefficient of variation around 19%.

2.4.2 Calculating the poststratified estimator for the 1000 simulations

(7) Adapt the code of the previous question in order to obtain 1000 poststratified estimates and calculate the empirical mean, the empirical (or Monte Carlo) standard deviation together with the Monte Carlo coefficient of variation in percentage (standard deviation divided by mean) for the 1000 poststratified estimates.

2.5 Comparison of the HT estimator with the poststratified estimator

(8) Use the previous results and draw histograms for the 1000 simulations in order to compare the HT and the poststratified estimator. Would you say that the results were expected ? Explain your answer.

3 The ratio estimator

3.1 Objectives

A the end of this section you should know how to draw a sample with a SRSWOR design and how to estimate the total number of empty housing units with a Horvitz-Thompson and a ratio estimator. You will also be able to compare the two estimators using simulations.

3.2 Preliminary results on the population

(1) Draw a scatterplot of the x and y variable. Make some comment about this plot.

3.3 The one sample case

3.3.1 Calculating the ratio estimator

(2) For the sample drawn in the previous subsection, calculate the ratio estimate of the total number of empty housing units by using the following instructions.

```
est.ratio<-svyratio(~LOGVAC,~LOG,ech.si)
predict(est.ratio, total=197314)
```

Understand and explain the previous two lines of code.

(3) Recall the definition of the ratio estimator. Use a calculator to detail the calculus in order to check that you find the same values than R for the ratio estimate.

3.4 Simulations

In order to compare the precision of the HT and the ratio estimator, you have to draw 1000 samples.

3.4.1 Calculating the ratio estimator for the 1000 simulations

(4) Adapt the code of the previous questions in order to obtain the 1000 ratio estimates and calculate the empirical mean, the empirical (or Monte Carlo) standard deviation together with the Monte Carlo coefficient of variation in percentage (standard deviation divided by mean) for the 1000 ratio estimates.

3.4.2 Comparison of the HT estimator with the ratio estimator

(5) Use the previous results for the 1000 simulations in order to compare the HT and the ratio estimator. Would you say that the results were expected? Explain your answer by explaining why or why not the ratio estimator is adapted to the data set.

4 The regression estimator

4.1 Objectives

A the end of this section you should know how to draw a sample with a SRSWOR design and how to estimate the total number of empty housing units with a Horvitz-Thompson and a regression estimator. You will also be able to compare the two estimators using simulations.

4.2 The one sample case

4.2.1 Calculating the regression estimator

(1) For the sample drawn in the previous subsection, calculate the simple regression estimate of the total number of empty housing units by using the following code.

```
ech.si.cal<-calibrate(ech.si,~LOG,c(554,197314))
svytotal(~LOGVAC, ech.si.cal)
```

Understand and explain the previous two lines of code.

(2) Recall the definition of the simple regression estimator as implemented by the code of the previous question. Use a calculator to detail the calculus in order to check that you find the same values than R for the regression estimate.

4.3 Simulations

In order to compare the precision of the HT and the regression estimator, you have to draw 1000 samples.

4.3.1 Calculating the regression estimator for the 1000 simulations

(3) Adapt the code of the previous questions in order to obtain the 1000 regression estimates and calculate the empirical mean, the empirical (or Monte Carlo) standard deviation together with the Monte Carlo coefficient of variation in percentage (standard deviation divided by mean) for the 1000 regression estimates.

4.3.2 Comparison of the HT estimator with the regression estimator

(4) Use the previous results for the 1000 simulations in order to compare the HT and the regression estimator. Would you say that the results were expected ? Explain your answer by explaining why or why not the regression estimator is adapted to the data set.