Usage of R in Official Statistics – Survey Data Analysis at the Statistical Office of the Republic of Slovenia

Jerneja PIKELJ (jerneja.pikelj@gov.si)
Statistical Office of the Republic of Slovenia

ABSTRACT

The paper has two practical purposes. The first one is to analyze how successfully R can be used for data analysis on surveys carried out by the Statistical Office of the Republic of Slovenia. In order to achieve this goal, we analyzed the data of the Monthly Statistical Survey on Earnings Paid by Legal Persons. The second purpose is to analyze how the assumption on the nonresponse mechanism, which occurs in the sample, impacts the estimated values of the unknown statistics in the survey. Depending on these assumptions, different approaches to adjust the problem caused by unit nonresponse are presented. We conclude the paper with the results of the analysis of the data and the main issues connected with the usage of R in official statistics.

Keywords: *Statistical Software R, Official Statistics, Social Statistics, Non-Response*

INTRODUCTION

SAS is the most commonly applied statistical software for data analysis at the Statistical Office of the Republic of Slovenia (SURS). Therefore, all the production processes, such like survey sampling, editing and imputation of micro data, weighting, aggregation and tabulation of the data, are carried out by SAS. We find R to be a very good and user friendly alternative to SAS, because it includes a wide set of packages implementing various functions, specifically formed for solving problems, which occur in official statistics. Because of our SAS oriented IT infrastructure, one of the main advantages of R statistical software is its ability to connect with other commercial software, like SAS. Other advantages worth taking into account are:

- Cost-free installation for the Windows and Linux operating systems
- ability to import (and export) the data from (to) Excel, txt, Access and Oracle, which are the most commonly used formats of the data at SURS
- easy data manipulation with large data sets
- continuous development and packages upgrade
- user-friendly graphical interface RStudio, which is available also for running from the Linux server

- large community of users providing technical support
- useful and systematically written documentation.

The main goal of this paper is to explore whether R can be used for data analysis on surveys carried out by SURS and to demonstrate some R statistical software tool functionalities on a practical example. We analyzed the data of the Monthly Statistical Survey on Earnings Paid by Legal Persons. The survey is based on full coverage, meaning that we monthly get the data on earnings paid out by around 50.000 legal persons of the public and private sector, or their units, registered for performing activity in the Republic of Slovenia. In comparison with other surveys carried out by SURS, this is one of largest surveys if the number of observed unit is considered. This is one of the reasons, why we have chosen this survey as a study case. It gives us an opportunity to analyze how R is able to deal with large data sets. The second reason why we have chosen this survey is the fact that the survey is based on a full coverage, offering us a very simple idea for simulation, which includes survey sampling, weight calculation and estimation of the unknown population total for one (or more) of the observed variables. This means, that on one hand we calculate the estimated value of the unknown parameter, while on the other hand we already know its real value. Consequently it is quite easy to see the difference between the estimated value and the real population total.

The focus of this paper is placed on different stages of the statistical process, ranging from data importing and exporting into R, survey sampling, data manipulation with large datasets, weight calculation and aggregation to some visualization of the results at the end.

IMPORT AND EXPORT OF THE DATA WITH R

The first step towards analyzing the data in R is getting the data into it. The R Project has a lot of packages, which allow us to read and write files in many different data formats. The most commonly used data formats at SURS are txt, tab, csv, Excel and Access. Simple datasets of like txt, csv and tab are easy to import and export, without any additional R package.

The most convenient package which includes functions for importing and exporting data into Excel and Access is RODBC. It includes simple functions for importing and exporting. For importing and exporting datasets in Excel format also the openxlsx package can be used. This package is dependent on the Rccp package, so we need to install both of them, in order to use the functions for importing and exporting data sets. The Openxlsx package allows us to use many other functionalities that are available in Excel, e.g. creating a new workbook **createWorkbook()**, adding new worksheets in the existing workbook **addWorksheet()**, writing a list of data frames to individual worksheets of a workbook **writeData()** and even changing or creating style of the datasets and format of its variables.

Our main concern was how to connect R to the Oracle database, because most of our micro data (and in some cases even macro and meta data) are placed in the Oracle database. The ROracle package offers many ways to connect to the database,

depending whether we are making the connection to the local or server database. For practical reasons we decided to connect to the Oracle database from RStudio Server edition. Our Oracle database also works on server and we need the following parameters to connect to it from RStudio:

- username
- password
- host
- SID
- Port.

We can find the host, SID and port parameters in file named tnsnames.ora. In order to avoid technical issues searching for these parameters in the tnsnames.ora file and copying it into the R script every time we want to connect to the database, we found the following solution:

1. The tnsnames.ora dataset is copied in the folder which contains the Oracle instant client:

```
sudo cp tnsnames.ora /usr/lib/oracle/11.2/client64/lib
```

In case we need to add another TNS for some server, we add it in this dataset:

```
/usr/lib/oracle/11.2/client64/lib/tnsnames.ora
```

2. We also need to add variable TNS_ADMIN in the Config dataset for R Studio Server environment, which points to folder which contains Tnsnames.ora dataset

```
sudo nano /etc/R/Renviron
```

At the end of this dataset it is added:

```
location tnsnames.ora
TNS_ADMIN=/usr/lib/oracle/11.2/client64/lib
```

3. Restart the RStudio

```
sudo rstudio-server restart
```

4. Example of connection to the database

4.1 Install packages

```
library(DBI)
library(ROracle)
```

4.2 Create an Oracle Database instance and create one connection to a remote database

```
driver <- dbDriver("Oracle")
```

```
db <- "MYORADB"    #TNS alias
user <- "scott"
pwd <- "tiger"
con <- dbConnect(driver, username=user, password = pwd, dbname = db)
```

4.3 Look up the list of all the tables on schema

```
dbListTables(con, schema = "SCOTT", all = FALSE, full = TRUE)
```

4.4 Disconnect from the database

```
dbDisconnect(con)
```

This solution works only in cases when we have administrator rights on R-server.

The ROracle package includes many other convenience methods, like **`dbReadTable()`** for reading the table from the database, **`dbWriteTable()`** for writing new tables, **`dbExistTable()`** for checking whether the table exists in the database and **`dbRemoveTable()`** for removing the tables from the database.

DATA ANALYSIS

About the data

The statistical survey Monthly Report on Earnings Paid out by Legal Persons provides an insight into the amount of average monthly earnings and their changes in the Republic of Slovenia. Observation units are legal persons of the public and private sector, or their local kind of activity units, registered for performing activity in the Republic of Slovenia.

Key variables which are also the subject of our interest in the simulation study are:

- gross_earn – gross earnings paid out for the reference month
- net_earn – net earnings paid out for the reference month
- nr_emp – the number of persons in paid employment who received earnings for the reference month

Other important variables are:

- size_class - determined on the basis of the number of persons in paid employment
- NACE1 – the first digit od the Statistical Classification of Economic Activities in the European Community

For the purpose of testing we selected the data for November 2014.

Simulation

A good way to see and test how the mechanism of nonresponse, which occurs in the sample, impacts on the estimated value for the unknown parameter is to carry out a simulation. The unknown parameter, which we are planning to estimate, is the total number of persons in paid employment.

One step of our simulation includes: selecting the sample, selecting the respondents, depending on a given response mechanism, calculating the weights and estimating the unknown population total.

❖ **Sampling**

The sample was selected with one-stage stratified sample design. Strata are defined by two stratum variables: NACE1 and size_class. Units from the biggest size class are selected with certainty; units from other stratum were selected with random selection of units within the stratum.

❖ **Nonresponse**

Missing data are always present in surveys, no matter how hard we try to prevent this. So, when it happens, it is very good to know how to deal with it. At this point we use the Rubin's (Rubin, 1976) definitions to define three types of response mechanism in the sample.

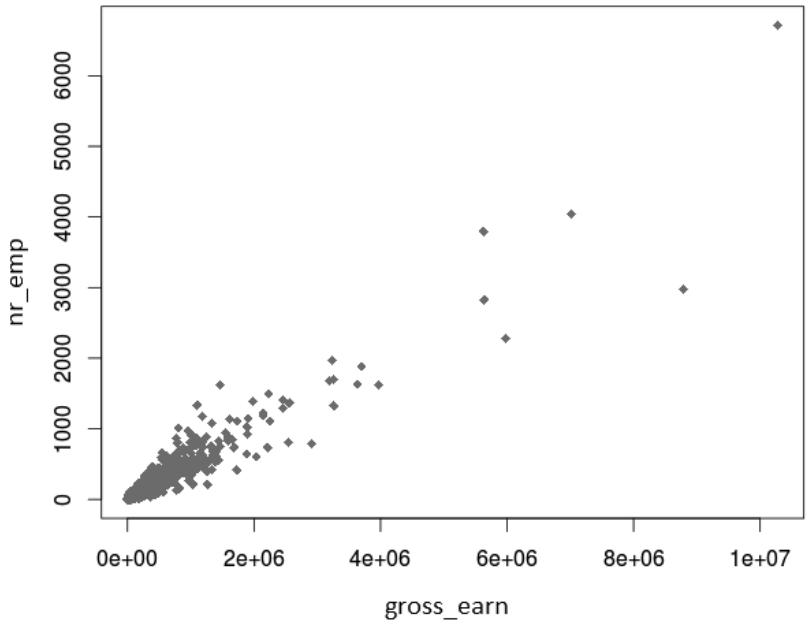
Our data do not contain the missing data, so we need to simulate this step. This can be done in three different ways. Firstly, we simulate the missing data completely at random (MCAR). Secondly, we assume that missing data occur at random (MAR) and simulate them within each stratum, and finally we simulate the missing data depending on the value of the ratio between the gross earnings paid out and the number of persons in paid employment who received earnings, considering on the assumption that the units with lower ratio have bigger chances to be a nonrespondent.

❖ **Estimation**

The simplest way to estimate the unknown total for the number of persons in paid employment is to use the Hrowitz-Thompson estimator. This estimator is unbiased in case of full response. Since we do not have full response, we can use an estimator which includes the correction for nonresponse.

Relying on a strong correlation between the number of persons in paid employment who received earnings (nr_emp) and gross earnings paid out for the reference month (gross_earn), which can also be seen from the scatter plot, a ratio estimator is used in order to estimate the total number of persons in paid employment who received earnings.

Correlation between gross earnings and number od persons in paid employment

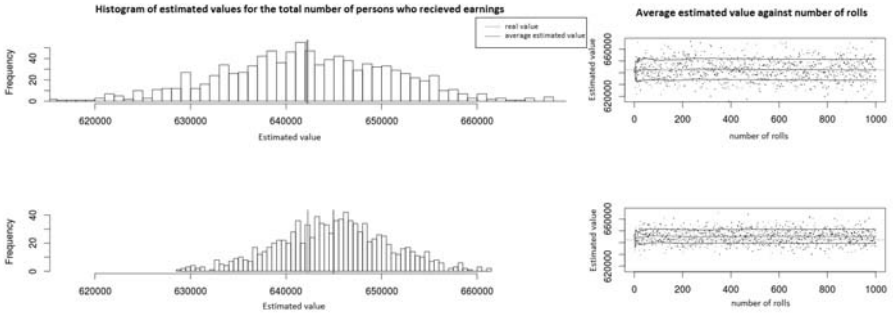


So we have two estimators that we are going to compare, depending on a given response mechanism. In order to find the best estimator, the weights for both of them are calculated at the stratum level.
For the sample selection and for determination of the respondents R sampling package was used while for weight calculation and estimation we used survey package.

Results

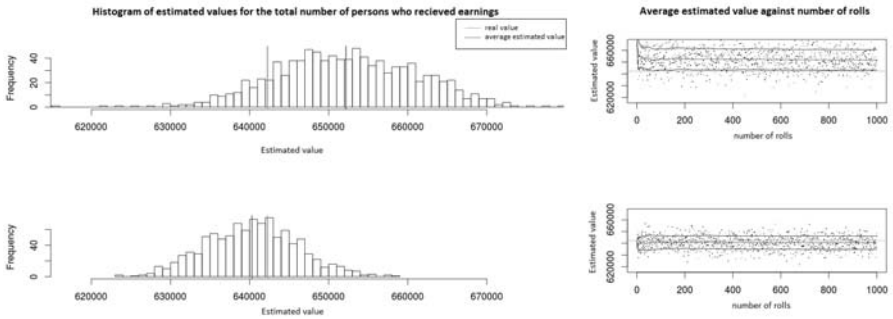
We repeated our simulation for 1000 times and visualize the results with histogram and line graph, depending on the assumed nonresponse mechanism. The first two pictures present results we got with generalized Horwitz-Thompson estimator and the second two pictures the results obtained by ratio estimator. Yellow line represents the true value for total number of persons in paid employment who received earnings, red line represents the mean of the estimated values and the blue line represents the standard error around the mean.
In the first picture, the results for simulation under MAR are presented. It can be clearly seen that the ratio estimator is slightly biased, whereas the generalized Horwitz-Thompson estimator is not. But, if we compare their variability, we can see that the ratio estimator is more precise.

Response mechanism: MAR



The second picture, where the results for simulation under NMAR are presented, indicates a bit different result. In this case the ratio estimator is more precise and less biased in comparison with generalized Horwitz-Thompson estimator. We can see that Horwitz-Thompson estimator does not give us good results if the missing data does not appear at random.

Response mechanism: NMAR



CONCLUSION

Beside the presented simulation, we also tested other complex algorithms for manipulation with larger datasets, creating complex survey design, estimating totals, means and rations and their variance.

R project is a powerful and user friendly tool for data analysis in official statistics. R contains packages with solutions that other standard tools do not have eg. small area estimation, disclosure control.

REFERENCES

1. C. E. Sarndal, B. Swenssoni, J. Wretman. Model Assisted Survey Sampling, Springer Series in Statistics. Springer-Verlag. New York, 1992.
2. SURS. ANNUAL QUALITY REPORT FOR THE SURVEY Monthly Report on Earnings Paid out by Legal Persons for 2013. Retreived on 25.02.2015 from: http://www.stat.si/doc/metodologija/kakovost/LPK_ZAPM_2013_eng.pdf
3. D. Mukhin, D. A. James and J. Luciani. Package ROracle, OCI based Oracle database interface for R. Retreived on 25.02.2015 from:

http://cran.r-project.org/web/packages/ROracle/ROracle.pdf

4. A.Walker, L. Braglia. Package openxlsx, Read, Write and Edit XLSX Files. http://galton.uchicago.edu/~eichler/stat24600/Admin/MissingDataReview.pdf available at: http://cran.r-project.org/web/packages/openxlsx/openxlsx.pdf
5. S. M. Lynch. Missing data. Retreived on 25.02.2015 from: http://www.princeton.edu/~slynch/soc504/missingdata.pdf.
6. T. D. Pigott. A Review of Methods for Missing Data. Retreived on 25.02.2015 from:http://galton.uchicago.edu/~eichler/stat24600/Admin/MissingDataReview.pdf