Determination of the Optimal Stratum Boundaries in the Monthly Retail Trade Survey in the Croatian Bureau of Statistics

Ivana JURINA (jurinal@dzs.hr)
Croatian Bureau of Statistics

Lidija GLIGOROVA (gligorovaL@dzs.hr)
Croatian Bureau of Statistics

ABSTRACT

This paper aims to compare the current sample design of the Monthly Retail Trade Survey in the SAS environment and the sample design with generalized Lavallee-Hidiroglou method of strata construction in the R environment. The sample for the Monthly Retail Trade Survey consists of all enterprises with 10 or more employees and the stratified random sample of enterprises with less than 10 employees. Stratification is made by the principal activity and by the number of employees. Allocation is done by applying the Neyman allocation method and the allocation variable is retail trade turnover. Using the strata.LH function from the stratification-package in R, optimal stratum boundaries with restriction to a certain economic activity group are determined, using the number of employees as a stratification variable and the retail trade turnover as a survey variable. In one case, the same level of precision for the estimated total turnover is taken into account and in another one, optimal stratum boundaries with the given sample size are determined. Furthermore, the results with different models, which take into account the discrepancy between stratification and survey variables, are compared.

Keywords: *Monthly Retail Trade Survey, strata construction, optimal boundaries, discrepancy*

JEL Classification: *C18, C87*

INTRODUCTION

The first business sample survey fully conducted by the Croatian Bureau of Statistics (CBS) was the Monthly Retail Trade Survey, which started in 1998. The sample frame is the Statistical Business Register (enterprises with certain types of economic activity) and is stratified according to economic activity codes and enterprise size class. The measure of size is the number of employees. In each stratum, simple random sample of enterprises is selected independently.

The sample is selected every two years and includes all enterprises with 10 or more persons employed, grouped in 5 size classes (take-all strata) and selected enterprises with less than 10 persons employed, grouped in 2 size-classes (take-some strata).

Allocation is done by applying the Neyman allocation method and the allocation variable is retail trade turnover. Appendix I shows size class boundaries, number of units in the frame and in the sample in Division 47, response rates and estimated population means of retail trade turnover according to size classes used in 2015 as one of the two stratification variables. The necessity of efficient stratification of business population is imposed by the attributes of business population: this population is generally unevenly distributed and tends to be very heterogeneous in size and characteristics. Business population is positively skewed with a relatively small number of large entities that have a large influence on the survey estimates, resulting in taking all the largest entities in the sample. Beside economic activities recorded in the Statistical Business Register, the complexity and heterogeneity of business population is closely related to the enterprise size.

The number of employees is continuous variable and for the purpose of stratification, size classes should be defined. Using the strata.LH function from the stratification-package in R, optimal stratum boundaries are determined with restriction to a certain economic activity group (Division 47 - Retail trade, except of motor vehicles and motorcycles), using the number of employees as a stratification variable and the retail trade turnover as a survey variable (univariate stratification).

Taking into account the discrepancy between the stratification variable and the survey variable, the Kozak's algorithm is applied considering a log linear model with mortality and heteroscedastic linear model.

ANALYSIS OF VARIABLES

The Statistical Business Register 2015 is used as the sampling frame. Firstly, the units for which information about the number of employees is missing and units that belong to the economic activity group other than Division 47 (Retail trade, except of motor vehicles and motorcycles) are excluded. Now there are 17 177 units in the frame. As the survey variable is not available in the frame for every unit, the relationship between the number of employees and the retail trade turnover from the sample survey data from January 2015 will be studied. The EMPLOYEES variable represents the number of employees and the P1_1501 variable represents the retail trade turnover.

From the Pearson's coefficient of correlation based on the sample data, it is noticeable that the number of employees and retail trade turnover are positively, linearly correlated. This is also visible from the plot with log-transformed values of these two variables (see Appendix II).

The lm function is applied to the formula that describes the variable turnover with the variable number of employees from the survey data, and the linear regression model is saved in a new variable reg. It is evident from the regression summary that variable number of employees is significant in the model. The regression diagnostic plots were also checked to assess the validity of the model. From Residuals vs Fitted plot, it can be assumed that residuals have a non-linear pattern. Normal Q-Q plot shows that residuals are normally distributed. The third plot is a scale-location plot (square rooted standardized residual vs. predicted value) and it is useful for checking the assumption of homoscedasticity. If the red line is flat and horizontal with equally and randomly spread data points, it can be assumed that there is no heteroscedasticity. If the red line has a positive slope to it, or if data points are not randomly spread out, this assumption is violated. From this plot, it will be assumed that there is no heteroscedasticity, and therefore, the log-linear model will be applied. Also, the studentized Breusch-Pagan test has been performed and it can be seen that the p-value is smaller than 0.05, which implies that there is a problem of heteroscedasticity. For that reason, the heteroscedastic linear model will also be applied to determine optimal strata boundaries.

From Residuals vs Leverage plot, it is evident that there are no influential cases (i.e. subjects) that determine a regression line.

STRATIFICATION TAKING INTO ACCOUNT A LOG-LINEAR MODEL WITH MORTALITY

This model considers the regression relationship between Y and X expressed by

$$Y = e^{\alpha+\beta\log(X)+\varepsilon} \tag{1}$$

where ε is assumed to be a 0-mean random variable, normally distributed with variance $\sigma^2_{\log}$ and independent from X, whereas α and $\beta_{\log}$ are the parameters to be estimated.

Estimated parameter values from the log-linear regression model are used between the stratification and survey variable to determine optimal stratum boundaries.

The values of the stratification variable, desired number of sampled strata, allocation rule, and target sample size n or target level of precision CV for the survey estimator need to be given in the strata.LH function.

As there are 7 size classes according to current stratification, the *Ls* argument will remain equal to 7. In that way, it is possible to compare how current stratum boundaries change. The *alloc* argument is a list that contains numeric objects q1, q2 and q3, which specify the allocation rule according to the general allocation scheme presented in Hidiroglou and Srinath (1993). We use Neyman allocation, which is obtained when q1=q3=0.5 and q2=0. The minimum number of units required in each sampled stratum (minNh) is set to 30. In the first two models, target sample size is given (2.393 units), while in the third and fourth model, the target level of precision of the estimated population mean for the survey variable is given (CV equal to 0.89%, which is a value from the survey in January 2015, using, beside the size class, an additional stratification variable, i.e. the economic activity). Response rates are not included in the first and the third model and non-response can be corrected *a posteriori,* by dividing the stratum sample sizes by the response rates. Response rates with values from January 2015 are added in the second and fourth model and they are taken into consideration when allocating the sample to the strata.

Model L1

```
loglinear1 <- strata.LH(x=FRAME$EMPLOYEES, n=2393, Ls=7,
alloc=c(0.5,0,0.5), model="loglinear", model.control=list(beta=reg$coef[2],
sig2=summary(reg)$sigma^2), algo.control = list(minNh=30))

Strata information:
           |type      ph rh|      bh      E(Y)        Var(Y)     Nh   nh    fh
stratum 1 |take-some  1  1 |     2.5     1.21          5.61  12575  516  0.04
stratum 2 |take-some  1  1 |     9.5     6.22        156.62   3478  753  0.22
stratum 3 |take-all   1  1 |    78.5    39.30       9636.44    983  983  1.00
stratum 4 |take-all   1  1 |   122.0   222.14     162491.82     38   38  1.00
stratum 5 |take-all   1  1 |   217.5   413.97     574581.99     30   30  1.00
stratum 6 |take-all   1  1 |   396.0   835.89    2370922.10     35   35  1.00
stratum 7 |take-all   1  1 | 12128.0  5879.52  597249554.18     38   38  1.00
Total                                                         17177 2393  0.14

Total sample size: 2393
Anticipated population mean: 20.31758
Anticipated CV: 0.005451871
```

In this first model, the non-response should be corrected *a posteriori*. Then total sample size is about 3450 units. Anticipated population mean and values E(Y) and Var(Y) are much lower than shown by the survey results, which indicates that the model used to describe the discrepancy between the stratification variable and the survey variable is not appropriate.

Model L2

```
loglinear2 <- strata.LH(x=FRAME$EMPLOYEES, n=2393, Ls=7,
alloc=c(0.5,0,0.5), rh=c(0.46, 0.73, 0.84, 0.91, 0.88, 0.95, 0.98),
model="loglinear", model.control=list(beta=reg$coef[2],
sig2=summary(reg)$sigma^2), algo.control = list(minNh=30))

Strata information:
          |type      ph   rh |      bh     E(Y)        Var(Y)    Nh   nh   fh
stratum 1 |take-some  1  0.46 |     1.5     1.00          3.21 10504  371  0.04
stratum 2 |take-some  1  0.73 |     4.5     3.23         38.43  4073  498  0.12
stratum 3 |take-some  1  0.84 |    11.5     9.91        357.15  1716  640  0.37
stratum 4 |take-all   1  0.91 |   157.0    59.74      27716.58   793  793  1.00
stratum 5 |take-all   1  0.88 |   256.5   535.26     964079.07    30   30  1.00
stratum 6 |take-all   1  0.95 |   442.0  1012.41    3419027.11    31   31  1.00
stratum 7 |take-all   1  0.98 | 12128.0  7117.76  747924250.02    30   30  1.00
Total                                                          17177 2393  0.14

Total sample size: 2393
Anticipated population mean: 20.31758
Anticipated CV: 0.06242225
```

When response rates are added in the model for the target sample size, anticipated CV increases from 0.55% to 6.24%. In the Model L1, units with more than 9 employees are included in the sample, while in the Model L2, this boundary is set to 11 employees.

In the third model, target CV is equal to the value from survey in January 2015, i.e. it is 0.89%.

Model L3

```
loglinear3 <- strata.LH(x=FRAME$EMPLOYEES, CV=0.0089, Ls=7,
alloc=c(0.5,0,0.5), model="loglinear", model.control=list(beta=reg$coef[2],
sig2=summary(reg)$sigma^2), algo.control = list(minNh=30))

Strata information:
          | type      ph rh |     bh     E(Y)        Var(Y)    Nh   nh   fh
stratum 1 | take-some  1  1|     2.5     1.21          5.61 12575  261  0.02
stratum 2 | take-some  1  1|     7.5     5.56        116.20  3137  296  0.09
stratum 3 | take-some  1  1|    20.5    18.37       1257.62  1006  313  0.31
stratum 4 | take-all   1  1|   157.0   100.28      54289.78   368  368  1.00
stratum 5 | take-all   1  1|   256.5   535.26     964079.07    30   30  1.00
stratum 6 | take-all   1  1|   442.0  1012.41    3419027.11    31   31  1.00
stratum 7 | take-all   1  1| 12128.0  7117.76  747924250.02    30   30  1.00
Total                                                      17177 1329  0.08

Total sample size: 1329
Anticipated population mean: 20.31758
Anticipated CV: 0.008891345
Note: CV=RRMSE (Relative Root Mean Squared Error) because takenone=0.
```

In this model, response rates are not included. After correcting the non-response *a posteriori,* total sample size is 1890 units. It is smaller than current sample size, but this model assumes much smaller variability of the survey variable than it is indicated in the survey.

Model L3 is simulated and corrected for non-response using the survey data from January 2015. The summary of results is given in the following table:

```
                         The SURVEYMEANS Procedure

                               Data Summary

          Number of Strata                        7
          Number of Observations               1890
          Sum of Weights                 36364.3237


                                Statistics

                                                 Std Error       Coeff of
Variable            N             Mean           of Mean         Variation
--------------------------------------------------------------------------
P1               1298       427.842470          6.867828         0.016052
--------------------------------------------------------------------------
```

As it is expected, CV=1.6% is larger than target CV=0.89% in the Model L3.

When response rates were added in each of the Ls sampled strata in Model L3, the algorithm did not converge because every initial boundary give sampled strata with less than 'minNh' units and/or with non-positive nh.

HETEROSCEDASTIC LINEAR MODEL

When the argument *model* is equal to "linear" it implies that the relationship between variables Y and X is described with heteroscedastic linear model:

$$Y = \beta X + \varepsilon \text{, where } \varepsilon \sim N(0; sig2X^{\gamma}) \tag{2}$$

To estimate γ, following Roshwalb (1987), we fit a given model and regress the log of the squared residuals on log(x) as follows:

$$\log(r_i^2) = \alpha + \gamma \log(x_i) \tag{3}$$

Based on sample data from January 2015, we get that the estimated value of gamma is equal to 1.487. According to Rivest (2002) we get the estimation of σ^2_{lin} and β_{lin} parameter:

```
beta.lin <- mean(SAMPLE1501$P1_1501/SAMPLE1501$EMPLOYEES)
sig2.lin <- var(SAMPLE1501$P1_1501/SAMPLE1501$EMPLOYEES)
```

The number of sampled strata remains equal to 7. The Neyman allocation and a minimum of 30 units in each sampled stratum (minNh) were used, the argument *model* was changed to "linear" and the required parameters were added.

Model H1

```
heteroscedastic1 <- strata.LH(x=FRAME$EMPLOYEES, n=2393, Ls=7,
alloc=c(0.5,0,0.5), model="linear", model.control=list(beta=beta.lin,
sig2=sig2.lin, gamma=1.487), algo.control = list(minNh=30))


Strata information:
          |      type rh |      bh       E(Y)         Var(Y)    Nh   nh    fh
stratum 1| take-some  1 |     1.5      45.22   7.290950e+03 10504  678  0.06
stratum 2| take-some  1 |     3.5     107.39   2.725574e+04  3313  414  0.12
stratum 3| take-some  1 |     8.5     242.13   9.397884e+04  2107  488  0.23
stratum 4| take-some  1 |    20.5     573.46   3.472334e+05   794  354  0.45
stratum 5|  take-all  1 |   195.0    2371.52   5.682434e+06   381  381  1.00
stratum 6|  take-all  1 |   396.0   12979.41   4.037654e+07    40   40  1.00
stratum 7|  take-all  1 |12128.0   64073.86   8.684400e+09    38   38  1.00
Total                                                       17177 2393  0.14

Total sample size: 2393
Anticipated population mean: 329.151
Anticipated CV: 0.009261417
```

Model H1 gives the anticipated population mean and values E(Y) and Var(Y) close to the results from the survey in January 2015, which indicates that this model much better describes the discrepancy between the stratification variable and the survey variable than the log-linear model with mortality.

After correcting the non-response *a posteriori*, total sample size is 4025.

Response rates are added in the following model.

Model H2.

```
heteroscedastic2 <- strata.LH(x=FRAME$EMPLOYEES, n=2393, Ls=7,
alloc=c(0.5,0,0.5), rh=c(0.46, 0.73, 0.84, 0.91, 0.88, 0.95, 0.98),
model="linear",    model.control=list(beta=beta.lin,    sig2=sig2.lin,
gamma=1.487), algo.control = list(minNh=30))

Strata information:
         |      type    rh|      bh       E(Y)         Var(Y)    Nh   nh    fh
stratum 1|take-some  0.46|     1.5      45.22   7.290950e+03 10504  674  0.06
stratum 2|take-some  0.73|     2.5      90.44   2.043692e+04  2071  222  0.11
stratum 3|take-some  0.84|     6.5     181.62   6.142158e+04  2898  540  0.19
stratum 4|take-some  0.91|    21.5     495.03   2.960558e+05  1264  517  0.41
stratum 5| take-all  0.88|   112.5    2055.90   3.288023e+06   332  332  1.00
stratum 6| take-all  0.95|   442.0   11191.49   4.707774e+07    78   78  1.00
stratum 7| take-all  0.98| 12128.0   76133.22   1.029406e+10    30   30  1.00
Total                                                       17177 2393  0.14

Total sample size: 2393
Anticipated population mean: 329.151
Anticipated CV: 0.01876138
```

When response rates are added in the model, anticipated CV increases from 0.93% to 1.88% for a given sample size. The number of take-all strata is equal in both models and stratum boundaries do not differ significantly.

In the following model, CV is set equal to 0.89% as a target value.

Model H3

```
heteroscedastic3 <- strata.LH(x=FRAME$EMPLOYEES, CV=0.0089, Ls=7,
alloc=c(0.5,0,0.5), model="linear", model.control=list(beta=beta.lin,
sig2=sig2.lin, gamma=1.487), algo.control = list(minNh=30))
```

```
Strata information:
          |      type rh |      bh      E(Y)       Var(Y)    Nh   nh   fh
stratum 1 | take-some  1 |     2.5     52.67 9.737290e+03 12575 1008 0.08
stratum 2 | take-some  1 |     6.5    181.62 6.142158e+04  2898  584 0.20
stratum 3 | take-some  1 |    17.5    457.09 2.501240e+05  1148  467 0.41
stratum 4 |  take-all  1 |   103.5   1707.02 2.664696e+06   442  442 1.00
stratum 5 |  take-all  1 |   256.5   7725.88 2.029760e+07    53   53 1.00
stratum 6 |  take-all  1 |   442.0  15886.98 5.045336e+07    31   31 1.00
stratum 7 |  take-all  1 | 12128.0  76133.22 1.029406e+10    30   30 1.00
Total                                                     17177 2615 0.15

Total sample size: 2615
Anticipated population mean: 329.151
Anticipated CV: 0.008894913
Note: CV=RRMSE (Relative Root Mean Squared Error) because takenone=0.
```

After correcting the non-response *a posteriori*, total sample size is 4418, which is larger than the current sample size, but applying additional stratification variable economic activity in the model as in the current survey, the coefficient of variation of estimated survey variable population mean should be significantly decreased.

When response rates in each of the Ls sampled strata in Model H3 are added, the algorithm does not converge because every initial boundary give sampled strata with less than 'minNh' units and/or with non-positive nh.

COMPARISON OF THE RESULTS

The heteroscedastic model much better describes the survey variable then the log-linear model and only heteroscedastic model will be taken into further consideration.

The following table gives a review of the current model and three heteroscedastic models with different target values.

Table 1

		CURRENT		MODEL H1		MODEL H2		MODEL H3	
		bh	nh	bh	nh	bh	nh	bh	nh
	stratum 1	5	922	1.5	678	1.5	674	2.5	1008
	stratum 2	10	347	3.5	414	2.5	222	6.5	584
	stratum 3	20	641	8.5	488	6.5	540	17.5	467
	stratum 4	50	267	20.5	354	21.5	517	103.5	442
	stratum 5	100	96	195.0	381	112.5	332	256.5	53
	stratum 6	250	58	396.0	40	442.0	78	442.0	31
	stratum 7	12128	62	12128.0	38	12128.0	30	12128.0	30
Target value				n=2393		n=2393		CV=0.89%	
n		2393		2393		2393		2615	
n after correction of non-response		2393		4025		2393		4418	
Anticipated population mean		427.16		329.15		329.15		329.15	
Anticipated CV		**2.35%**		**0.93%**		**1.88%**		**0.89%**	

The data from Table 1 suggest that Model H2 applied in the Monthly Retail Trade Survey could give more precise estimate of the population mean than the current design with the same sample size. Models H1 and H2 suggest how to increase the current sample size and get the coefficient of variation of estimated population mean less than 1%. All heteroscedastic models suggest quite large lower boundary of stratum 7, which should be respected in the future survey sample design.

CONCLUSION

This paper gives a comparison of the current sample design of the Monthly Retail Trade Survey in the CBS and sample designs obtained by strata.LH function with generalized Lavallee-Hidiroglou method of strata construction in the R environment. The results will be additionally tested and incorporated in the survey.

The R environment offers a large variety of statistical methods, which thus become available to a wide range of statisticians.

REFERENCES

1. **Baillargeon, S.** and **Rivest, L.-P.**, 2011, The Construction of Stratified designs in R with the package stratification. Survey Methodology, 37, pp. 53-65.
2. **Baillargeon, S.** and **Rivest, L.-P.**, 2017, stratification: Univariate Stratification of Survey Populations. https://CRAN.R-project.org/package=stratification.
3. **Chambers, R.** and **Clark, R.,** 2012, An Introduction to Model-Based Survey Sampling with Applications. Oxford University Press, Inc., New York.
4. **Godfrey, J., Roshwalb, A., & Wright, R.,** 1984. Model-based stratification in inventory cost estimation. Journal of Business & Economic Statistics, 2, 1-9.
5. **Godfrey, J., Roshwalb, A., & Wright, R.,** 1987, A New Approach for Stratified Sampling in Inventory Cost Estimation. Journal of Practice and Theory, 7, 54-70.
6. **Henry, K.** and **Valliant, R.,** 2009, Comparing Sampling and Estimation Strategies in Establishment Populations. Survey Research Methods, Vol.3, No.1, pp. 27-44.
7. **Henry, K. and Valliant, R.,** 2006,, Comparing Strategies to Estimate a Measure of Heteroscedasticity. Proceedings of the Section on Survey Research Methods, Washington DC: American Statistical Association, 3118-3125.
8. **Khan, M. G. M., Rao, D., Ansari, A. H.** and **Ahsan, M. J.,** 2014, Determining Optimum Strata Boundaries and Sample Sizes for Skewed Population with Log-normal Distribution. Communications in Statistics - Simulation and Computation, DOI: 10.1080/03610918.2013.819917.
9. **Khan, M. G. M., Prasad, V. D.** and **Rao, D. K.,** 2014, Optimum Stratification Using Auxiliary Variables. Aligarh Journal of Statistics, Vol. 34 (2014), 85-104.
10. **Rivest, L. P.,** 2002. A generalization of the Lavallee and Hidiroglou algorithm for stratification in business surveys. Survey Methodology, 28(2), 191-198.
11. **Särndal, Swensson,** and **Wretman,** 1992, Model Assisted Survey Sampling. Springer-Verlag, New York.

APPENDIX I

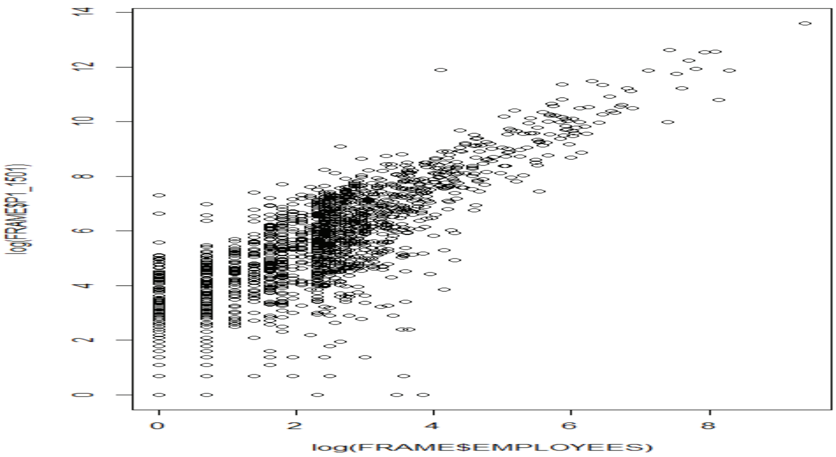
Information about the sample design of the Monthly Retail Trade Survey used in 2015

stratum	Type	Number of employees	Nh	nh (number of selected units)	Response rate	$\bar{y}$
1	take-some	< 5	14.577	922	0.43	74.8
2	take-some	< 10	1.476	347	0.73	280.4
3	take-all	< 20	641	641	0.84	588.8
4	take-all	< 50	267	267	0.91	1238.9
5	take-all	< 100	96	96	0.88	5215.3
6	take-all	< 250	58	58	0.95	9112.6
7	take-all	>= 250	62	62	0.98	66137.2
		Total	17.177	2.393	0.67	
Estimated population mean using simple expansion estimator and stratification according to economic activity and size classes: 375.718						
CV: 0.0089						
Estimated population mean using simple expansion estimator and stratification according to size classes: 427.163						
CV:0.0235						

APPENDIX II

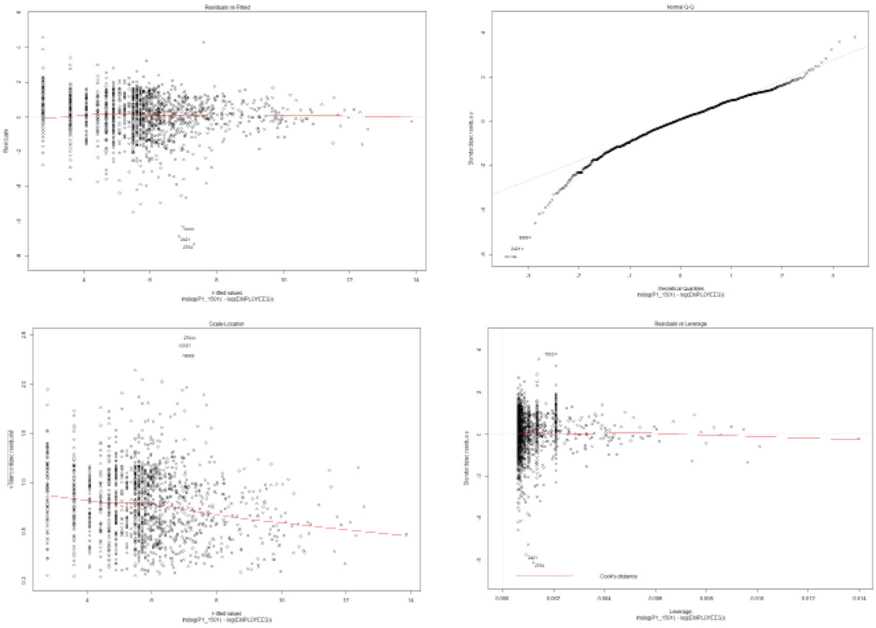
R code and output

```
FRAME<-trg1_okvir15[!is.na(trg1_okvir15$EMPLOYEES),]  # excluding units with missing
number of employees

FRAME  <- FRAME[which(substr(FRAME$NKD2007Glavna,1,2)=='47')    , ]
# keeping units from Division 47
# 17.177 units remaining in the frame

cor.test((FRAME$EMPLOYEES), (FRAME$P1_1501))

         Pearson's product-moment correlation

data:  (FRAME$EMPLOYEES) and (FRAME$P1_1501)
t = 100.0989, df = 1654, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9193091 0.9329863
sample estimates:
      cor
0.9264529

plot (log(FRAME$EMPLOYEES), log(FRAME$P1_1501))
```




```
reg <- lm(log(P1_1501)~log(EMPLOYEES), data = FRAME)
summary(reg)

Call:
lm(formula = log(P1_1501) ~ log(EMPLOYEES), data = FRAME)

Residuals:
    Min       1Q  Median      3Q     Max
-7.3120 -0.6822  0.1221  0.7834  4.5436
```

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.76828    0.05484   50.48   <2e-16 ***
log(EMPLOYEES)  1.18015    0.01966   60.03   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.199 on 1654 degrees of freedom
  (15521 observations deleted due to missingness)
Multiple R-squared:  0.6854, Adjusted R-squared:  0.6852
F-statistic:  3604 on 1 and 1654 DF,  p-value: < 2.2e-16
```



```
loglinear1 <- strata.LH(x=FRAME$EMPLOYEES, n=2393, Ls=7,
alloc=c(0.5,0,0.5), model="loglinear", model.control=list(beta=reg$coef[2],
 sig2=summary(reg)$sigma^2), algo.control = list(minNh=30))

loglinear1
Given arguments:
x = FRAME$EMPLOYEES
n = 2393, Ls = 7, takenone = 0, takeall = 0
allocation: q1 = 0.5, q2 = 0, q3 = 0.5
model = loglinear: beta = 1.180153, sig2 = 1.436693, ph = 1 1 1 1 1 1 1
algo = Kozak: minsol = 1000, idopti = nh, minNh = 30, maxiter = 10000,
              maxstep = 20, maxstill = 200, rep = 5, trymany = TRUE

Strata information:
         |      type ph rh|      bh    E(Y)         Var(Y)    Nh   nh   fh
stratum 1|take-some  1  1|     2.5    1.21           5.61 12575  516 0.04
stratum 2|take-some  1  1|     9.5    6.22         156.62  3478  753 0.22
stratum 3| take-all  1  1|    78.5   39.30        9636.44   983  983 1.00
stratum 4| take-all  1  1|   122.0  222.14      162491.82    38   38 1.00
stratum 5| take-all  1  1|   217.5  413.97      574581.99    30   30 1.00
stratum 6| take-all  1  1|   396.0  835.89     2370922.10    35   35 1.00
stratum 7| take-all  1  1| 12128.0 5879.52   597249554.18    38   38 1.00
Total                                                     17177 2393 0.14
```

```
Total sample size: 2393
Anticipated population mean: 20.31758
Anticipated CV: 0.005451871
Note: CV=RRMSE (Relative Root Mean Squared Error) because takenone=0.
```

```
loglinear2 <- strata.LH(x=FRAME$EMPLOYEES, n=2393, Ls=7,
alloc=c(0.5,0,0.5), rh=c(0.46, 0.73, 0.84, 0.91, 0.88, 0.95, 0.98),
model="loglinear", model.control=list(beta=reg$coef[2],
sig2=summary(reg)$sigma^2), algo.control = list(minNh=30))
```

```
loglinear2
Given arguments:
x = FRAME$EMPLOYEES
n = 2393, Ls = 7, takenone = 0, takeall = 0
allocation: q1 = 0.5, q2 = 0, q3 = 0.5
model = loglinear: beta = 1.180153, sig2 = 1.436693, ph = 1 1 1 1 1 1 1
algo = Kozak: minsol = 1000, idopti = nh, minNh = 30, maxiter = 10000,
              maxstep = 20, maxstill = 200, rep = 5, trymany = TRUE

Strata information:
          |      type  ph    rh |      bh      E(Y)        Var(Y)     Nh    nh    fh
stratum 1|  take-some   1  0.46 |     1.5      1.00          3.21  10504   371  0.04
stratum 2|  take-some   1  0.73 |     4.5      3.23         38.43   4073   498  0.12
stratum 3|  take-some   1  0.84 |    11.5      9.91        357.15   1716   640  0.37
stratum 4|   take-all   1  0.91 |   157.0     59.74      27716.58    793   793  1.00
stratum 5|   take-all   1  0.88 |   256.5    535.26     964079.07     30    30  1.00
stratum 6|   take-all   1  0.95 |   442.0   1012.41    3419027.11     31    31  1.00
stratum 7|   take-all   1  0.98 |12128.0   7117.76  747924250.02     30    30  1.00
Total                                                               17177  2393  0.14

Total sample size: 2393
Anticipated population mean: 20.31758
Anticipated CV: 0.06242225
Note: CV=RRMSE (Relative Root Mean Squared Error) because takenone=0.
```

```
loglinear3 <- strata.LH(x=FRAME$EMPLOYEES, CV=0.0089, Ls=7,
alloc=c(0.5,0,0.5), model="loglinear", model.control=list(beta=reg$coef[2],
sig2=summary(reg)$sigma^2), algo.control = list(minNh=30))
```

```
loglinear3
Given arguments:
x = FRAME$EMPLOYEES
CV = 0.0089, Ls = 7, takenone = 0, takeall = 0
allocation: q1 = 0.5, q2 = 0, q3 = 0.5
model = loglinear: beta = 1.180153, sig2 = 1.436693, ph = 1 1 1 1 1 1 1
algo = Kozak: minsol = 1000, idopti = nh, minNh = 30, maxiter = 10000,
              maxstep = 20, maxstill = 200, rep = 5, trymany = TRUE

Strata information:
          |      type ph rh|      bh     E(Y)       Var(Y)     Nh    nh    fh
stratum 1| take-some   1  1|     2.5     1.21          5.61  12575   259  0.02
stratum 2| take-some   1  1|     7.5     5.56        116.20   3137   293  0.09
stratum 3| take-some   1  1|    19.5    17.99       1189.74    982   294  0.30
stratum 4|  take-all   1  1|   157.0    96.24      51447.13    392   392  1.00
stratum 5|  take-all   1  1|   256.5   535.26     964079.07     30    30  1.00
stratum 6|  take-all   1  1|   442.0  1012.41    3419027.11     31    31  1.00
stratum 7|  take-all   1  1| 12128.0  7117.76  747924250.02     30    30  1.00
Total                                                       17177  1329  0.08

Total sample size: 1329
Anticipated population mean: 20.31758
Anticipated CV: 0.008890745
Note: CV=RRMSE (Relative Root Mean Squared Error) because takenone=0.
```

```
loglinear4 <- strata.LH(x=FRAME$EMPLOYEES, CV=0.0089, Ls=7,
alloc=c(0.5,0,0.5), rh=c(0.46, 0.73, 0.84, 0.91, 0.88, 0.95, 0.98),
model="loglinear", model.control=list(beta=reg$coef[2],
sig2=summary(reg)$sigma^2))

Warning messages:
1: the algorithm cannot be run because every initial boundaries give
sampled strata with less than 'minNh' units and/or with non-positive nh
2: divisions by zero occured in the computations, therefore some
statistics do not have finite values


# HETEROSCEDASTIC LINEAR MODELS


SAMPLE1501<-FRAME[!is.na(FRAME$P1_1501),]

reg2 <- lm(P1_1501 ~ EMPLOYEES, data = SAMPLE1501)
reg2

Call:
lm(formula = P1_1501 ~ EMPLOYEES, data = SAMPLE1501)

Coefficients:
(Intercept)    EMPLOYEES
    -191.57        64.68

res <- (reg2$residuals)^2

reg3 <- lm(log(res)~log(SAMPLE1501$EMPLOYEES))
reg3

Call:
lm(formula = log(res) ~ log(SAMPLE1501$EMPLOYEES))

Coefficients:
              (Intercept)  log(SAMPLE1501$EMPLOYEES)
                    8.058                      1.487

beta.lin <- mean(SAMPLE1501$P1_1501/SAMPLE1501$EMPLOYEES)
sig2.lin <- var(SAMPLE1501$P1_1501/SAMPLE1501$EMPLOYEES)

heteroscedastic1 <- strata.LH(x=FRAME$EMPLOYEES, n=2393, Ls=7,
alloc=c(0.5,0,0.5), model="linear", model.control=list(beta=beta.lin,
sig2=sig2.lin, gamma=1.487), algo.control = list(minNh=30))

heteroscedastic1

Given arguments:
x = FRAME$EMPLOYEES
n = 2393, Ls = 7, takenone = 0, takeall = 0
allocation: q1 = 0.5, q2 = 0, q3 = 0.5
model = linear: beta = 45.22049, sig2 = 7290.946, gamma = 1.487
algo = Kozak: minsol = 1000, idopti = nh, minNh = 30, maxiter = 10000,
              maxstep = 20, maxstill = 200, rep = 5, trymany = TRUE

Strata information:
         |      type  rh |      bh      E(Y)        Var(Y)     Nh    nh    fh
stratum 1|  take-some  1 |     1.5     45.22  7.290950e+03  10504   678  0.06
stratum 2|  take-some  1 |     3.5    107.39  2.725574e+04   3313   414  0.12
stratum 3|  take-some  1 |     8.5    242.13  9.397884e+04   2107   488  0.23
stratum 4|  take-some  1 |    20.5    573.46  3.472334e+05    794   354  0.45
stratum 5|   take-all  1 |   195.0   2371.52  5.682434e+06    381   381  1.00
stratum 6|   take-all  1 |   396.0  12979.41  4.037654e+07     40    40  1.00
stratum 7|   take-all  1 | 12128.0  64073.86  8.684400e+09     38    38  1.00
Total                                                       17177  2393  0.14
```

```
Total sample size: 2393
Anticipated population mean: 329.151
Anticipated CV: 0.009261417
Note: CV=RRMSE (Relative Root Mean Squared Error) because takenone=0.

heteroscedastic2 <- strata.LH(x=FRAME$EMPLOYEES, n=2393, Ls=7,
alloc=c(0.5,0,0.5), rh=c(0.46, 0.73, 0.84, 0.91, 0.88, 0.95, 0.98),
model="linear", model.control=list(beta=beta.lin, sig2=sig2.lin,
gamma=1.487), algo.control = list(minNh=30))

heteroscedastic2
Given arguments:
x = FRAME$EMPLOYEES
n = 2393, Ls = 7, takenone = 0, takeall = 0
allocation: q1 = 0.5, q2 = 0, q3 = 0.5
model = linear: beta = 45.22049, sig2 = 7290.946, gamma = 1.487
algo = Kozak: minsol = 1000, idopti = nh, minNh = 30, maxiter = 10000,
              maxstep = 20, maxstill = 200, rep = 5, trymany = TRUE

Strata information:
          |      type   rh|      bh      E(Y)        Var(Y)    Nh   nh   fh
stratum 1 | take-some 0.46|     1.5     45.22  7.290950e+03 10504  674 0.06
stratum 2 | take-some 0.73|     2.5     90.44  2.043692e+04  2071  222 0.11
stratum 3 | take-some 0.84|     6.5    181.62  6.142158e+04  2898  540 0.19
stratum 4 | take-some 0.91|    21.5    495.03  2.960558e+05  1264  517 0.41
stratum 5 |  take-all 0.88|   112.5   2055.90  3.288023e+06   332  332 1.00
stratum 6 |  take-all 0.95|   442.0  11191.49  4.707774e+07    78   78 1.00
stratum 7 |  take-all 0.98| 12128.0  76133.22  1.029406e+10    30   30 1.00
Total                                                        17177 2393 0.14

Total sample size: 2393
Anticipated population mean: 329.151
Anticipated CV: 0.01876138
Note: CV=RRMSE (Relative Root Mean Squared Error) because takenone=0.

heteroscedastic3 <- strata.LH(x=FRAME$EMPLOYEES, CV=0.0089, Ls=7,
alloc=c(0.5,0,0.5),   model="linear",   model.control=list(beta=beta.lin,
sig2=sig2.lin, gamma=1.487), algo.control = list(minNh=30))

heteroscedastic3
Given arguments:
x = FRAME$EMPLOYEES
CV = 0.0089, Ls = 7, takenone = 0, takeall = 0
allocation: q1 = 0.5, q2 = 0, q3 = 0.5
model = linear: beta = 45.22049, sig2 = 7290.946, gamma = 1.487
algo = Kozak: minsol = 1000, idopti = nh, minNh = 30, maxiter = 10000,
              maxstep = 20, maxstill = 200, rep = 5, trymany = TRUE

Strata information:
          |      type rh|     bh      E(Y)        Var(Y)    Nh   nh   fh
stratum 1 |take-some  1|    2.5     52.67  9.737290e+03 12575 1008 0.08
stratum 2 |take-some  1|    6.5    181.62  6.142158e+04  2898  584 0.20
stratum 3 |take-some  1|   17.5    457.09  2.501240e+05  1148  467 0.41
stratum 4 | take-all  1|  136.0   1830.74  3.299017e+06   458  458 1.00
stratum 5 | take-all  1|  256.5   8797.22  2.150262e+07    37   37 1.00
stratum 6 | take-all  1|  442.0  15886.98  5.045336e+07    31   31 1.00
stratum 7 | take-all  1|12128.0  76133.22  1.029406e+10    30   30 1.00
Total                                                   17177 2615 0.15

Total sample size: 2615
Anticipated population mean: 329.151
Anticipated CV: 0.008894913
Note: CV=RRMSE (Relative Root Mean Squared Error) because takenone=0.
```

```
heteroscedastic4 <- strata.LH(x=FRAME$EMPLOYEES, CV=0.0089, Ls=7,
alloc=c(0.5,0,0.5), rh=c(0.46, 0.73, 0.84, 0.91, 0.88, 0.95, 0.98),
model="linear",    model.control=list(beta=beta.lin,    sig2=sig2.lin,
gamma=1.487), algo.control = list(minNh=30))
Warning messages:
1: the algorithm cannot be run because every initial boundaries give
sampled strata with less than 'minNh' units and/or with non-positive nh
2: divisions by zero occured in the computations, therefore some
statistics do not have finite values
```