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

Ivana JURINA (jurinaI@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 Laval-lee-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 root of 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

\[ Y = e^{\alpha + \beta \log(X)} + \varepsilon \]

where \( \varepsilon \) is assumed to be a 0-mean random variable, normally distributed with variance \( \sigma^2_{\log} \) and independent from X, whereas \( \alpha \) 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 \textit{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

\begin{verbatim}
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))
\end{verbatim}

Strata information:

| 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 98 3 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 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 \textit{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.
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%.

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:
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.

**HETEROSEDASTIC 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:

\[ Y = \beta X + \varepsilon, \text{ where } \varepsilon \sim N(0; \sigma^2 X^\gamma) \]  

(2)

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

\[ \log(\hat{\sigma}^2) = \alpha + \gamma \log(x_i) \]  

(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 \( \sigma^2_{lin} \) and \( \beta_{lin} \) parameter:

\[
\text{beta.lin} \leftarrow \text{mean(SAMPLE1501$P1_1501$/SAMPLE1501$EMPLOYEES$)} \\
\text{sig2.lin} \leftarrow \text{var(SAMPLE1501$P1_1501$/SAMPLE1501$EMPLOYEE$)}
\]

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

\begin{verbatim}
Model H1 gives the anticipated population mean and values \(E(Y)\) and \(\text{Var}(Y)\) 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 \textit{a posteriori}, total sample size is 4025.

Response rates are added in the following model.

Model H2.
\end{verbatim}

<table>
<thead>
<tr>
<th>stratum</th>
<th>type</th>
<th>rh</th>
<th>E(Y)</th>
<th>Var(Y)</th>
<th>Nh</th>
<th>nh</th>
<th>fh</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>take-some</td>
<td>0.46</td>
<td>1.5</td>
<td>7.290950e+03</td>
<td>10504</td>
<td>674</td>
<td>0.06</td>
</tr>
<tr>
<td>2</td>
<td>take-some</td>
<td>0.73</td>
<td>2.5</td>
<td>2.043692e+04</td>
<td>2071</td>
<td>222</td>
<td>0.11</td>
</tr>
<tr>
<td>3</td>
<td>take-some</td>
<td>0.84</td>
<td>6.5</td>
<td>6.142158e+04</td>
<td>1264</td>
<td>517</td>
<td>0.41</td>
</tr>
<tr>
<td>4</td>
<td>take-all</td>
<td>0.88</td>
<td>11.25</td>
<td>3.288023e+06</td>
<td>332</td>
<td>332</td>
<td>1.00</td>
</tr>
<tr>
<td>5</td>
<td>take-all</td>
<td>0.95</td>
<td>442.0</td>
<td>4.707774e+07</td>
<td>78</td>
<td>78</td>
<td>1.00</td>
</tr>
<tr>
<td>6</td>
<td>take-all</td>
<td>0.98</td>
<td>12128.0</td>
<td>8.684400e+09</td>
<td>38</td>
<td>38</td>
<td>1.00</td>
</tr>
</tbody>
</table>

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:

| 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 than 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

<table>
<thead>
<tr>
<th>Stratum</th>
<th>CURRENT</th>
<th>MODEL H1</th>
<th>MODEL H2</th>
<th>MODEL H3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>bh</td>
<td>nh</td>
<td>bh</td>
<td>nh</td>
</tr>
<tr>
<td>stratum 1</td>
<td>5</td>
<td>922</td>
<td>1.5</td>
<td>678</td>
</tr>
<tr>
<td>stratum 2</td>
<td>10</td>
<td>347</td>
<td>3.5</td>
<td>414</td>
</tr>
<tr>
<td>stratum 3</td>
<td>20</td>
<td>641</td>
<td>8.5</td>
<td>488</td>
</tr>
<tr>
<td>stratum 4</td>
<td>50</td>
<td>267</td>
<td>20.5</td>
<td>354</td>
</tr>
<tr>
<td>stratum 5</td>
<td>100</td>
<td>96</td>
<td>195.0</td>
<td>381</td>
</tr>
<tr>
<td>stratum 6</td>
<td>250</td>
<td>58</td>
<td>396.0</td>
<td>40</td>
</tr>
<tr>
<td>stratum 7</td>
<td>12128</td>
<td>62</td>
<td>12128.0</td>
<td>38</td>
</tr>
</tbody>
</table>

Target value

| n | 2393 |
| n after correction of non-response | 2393 |
| Anticipated population mean | 427.16 |
| Anticipated CV | 2.35% |

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

APPENDIX I

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

<table>
<thead>
<tr>
<th>stratum</th>
<th>Type</th>
<th>Number of employees</th>
<th>Nh</th>
<th>nh (number of selected units)</th>
<th>Response rate</th>
<th>( \bar{y} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>take-some</td>
<td>&lt; 5</td>
<td>14,577</td>
<td>922</td>
<td>0.43</td>
<td>74.8</td>
</tr>
<tr>
<td>2</td>
<td>take-some</td>
<td>&lt; 10</td>
<td>1,476</td>
<td>347</td>
<td>0.73</td>
<td>280.4</td>
</tr>
<tr>
<td>3</td>
<td>take-all</td>
<td>&lt; 20</td>
<td>641</td>
<td>641</td>
<td>0.84</td>
<td>588.8</td>
</tr>
<tr>
<td>4</td>
<td>take-all</td>
<td>&lt; 50</td>
<td>267</td>
<td>267</td>
<td>0.91</td>
<td>1238.9</td>
</tr>
<tr>
<td>5</td>
<td>take-all</td>
<td>&lt; 100</td>
<td>96</td>
<td>96</td>
<td>0.88</td>
<td>5215.3</td>
</tr>
<tr>
<td>6</td>
<td>take-all</td>
<td>&gt;= 250</td>
<td>58</td>
<td>58</td>
<td>0.95</td>
<td>9112.6</td>
</tr>
<tr>
<td>7</td>
<td>take-all</td>
<td>&gt;= 250</td>
<td>62</td>
<td>62</td>
<td>0.98</td>
<td>66137.2</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>17,177</td>
<td>2,393</td>
<td>0.67</td>
<td></td>
</tr>
</tbody>
</table>

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

Romanian Statistical Review nr. 4 / 2017
R code and output

```r
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)) # Pearon'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

```

Romanian Statistical Review nr. 4 / 2017
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
(1521 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: mnsol = 1000, idopti = nh, minNh = 30, maxiter = 10000,
maxstep = 20, maxstill = 200, rep = 5, trumany = TRUE

Strata information:

<table>
<thead>
<tr>
<th>type</th>
<th>ph</th>
<th>rh</th>
<th>bh</th>
<th>E(Y)</th>
<th>Var(Y)</th>
<th>Nh</th>
<th>nh</th>
<th>fh</th>
</tr>
</thead>
<tbody>
<tr>
<td>stratum 1</td>
<td>take-some</td>
<td>1 1</td>
<td>2.5</td>
<td>1.21</td>
<td>5.61</td>
<td>12575</td>
<td>516</td>
<td>0.04</td>
</tr>
<tr>
<td>stratum 2</td>
<td>take-some</td>
<td>1 1</td>
<td>9.5</td>
<td>6.22</td>
<td>156.62</td>
<td>3478</td>
<td>753</td>
<td>0.22</td>
</tr>
<tr>
<td>stratum 3</td>
<td>take-all</td>
<td>1 1</td>
<td>78.5</td>
<td>39.30</td>
<td>9636.44</td>
<td>983</td>
<td>983</td>
<td>1.00</td>
</tr>
<tr>
<td>stratum 4</td>
<td>take-all</td>
<td>1 1</td>
<td>122.0</td>
<td>222.14</td>
<td>162491.82</td>
<td>38</td>
<td>38</td>
<td>1.00</td>
</tr>
<tr>
<td>stratum 5</td>
<td>take-all</td>
<td>1 1</td>
<td>217.5</td>
<td>413.97</td>
<td>57481.99</td>
<td>30</td>
<td>30</td>
<td>1.00</td>
</tr>
<tr>
<td>stratum 6</td>
<td>take-all</td>
<td>1 1</td>
<td>396.0</td>
<td>835.89</td>
<td>2370922.10</td>
<td>35</td>
<td>35</td>
<td>1.00</td>
</tr>
<tr>
<td>stratum 7</td>
<td>take-all</td>
<td>1 1</td>
<td>12128.0</td>
<td>5879.52</td>
<td>597249554.18</td>
<td>38</td>
<td>38</td>
<td>1.00</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>17177 2393 0.14</td>
</tr>
</tbody>
</table>
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:

<table>
<thead>
<tr>
<th>type</th>
<th>ph</th>
<th>rh</th>
<th>bh</th>
<th>E(Y)</th>
<th>Var(Y)</th>
<th>Nh</th>
<th>nh</th>
<th>fh</th>
</tr>
</thead>
<tbody>
<tr>
<td>stratum 1</td>
<td>take-some</td>
<td>1</td>
<td>0.46</td>
<td>1.5</td>
<td>1.00</td>
<td>3.21</td>
<td>10504</td>
<td>371</td>
</tr>
<tr>
<td>stratum 2</td>
<td>take-some</td>
<td>1</td>
<td>0.73</td>
<td>4.5</td>
<td>3.23</td>
<td>38.43</td>
<td>4073</td>
<td>498</td>
</tr>
<tr>
<td>stratum 3</td>
<td>take-some</td>
<td>1</td>
<td>0.84</td>
<td>11.5</td>
<td>9.91</td>
<td>357.15</td>
<td>1716</td>
<td>1640</td>
</tr>
<tr>
<td>stratum 4</td>
<td>take-all</td>
<td>1</td>
<td>0.81</td>
<td>157.0</td>
<td>59.74</td>
<td>27710.58</td>
<td>793</td>
<td>793</td>
</tr>
<tr>
<td>stratum 5</td>
<td>take-all</td>
<td>1</td>
<td>0.88</td>
<td>256.5</td>
<td>535.26</td>
<td>964079.07</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>stratum 6</td>
<td>take-all</td>
<td>1</td>
<td>0.95</td>
<td>442.0</td>
<td>1012.41</td>
<td>3419027.11</td>
<td>31</td>
<td>31</td>
</tr>
<tr>
<td>stratum 7</td>
<td>take-all</td>
<td>1</td>
<td>0.98</td>
<td>12128.0</td>
<td>7117.76</td>
<td>747924250.02</td>
<td>30</td>
<td>30</td>
</tr>
</tbody>
</table>
Total | | | | | 17177 | | 2393 | | 0.14 |

Total sample size: 2393
Anticipated population mean: 20.31758
Anticipated CV: 0.008890745
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:

<table>
<thead>
<tr>
<th>type</th>
<th>ph</th>
<th>rh</th>
<th>bh</th>
<th>E(Y)</th>
<th>Var(Y)</th>
<th>Nh</th>
<th>nh</th>
<th>fh</th>
</tr>
</thead>
<tbody>
<tr>
<td>stratum 1</td>
<td>take-some</td>
<td>1</td>
<td>0.46</td>
<td>1.5</td>
<td>1.00</td>
<td>3.21</td>
<td>10504</td>
<td>371</td>
</tr>
<tr>
<td>stratum 2</td>
<td>take-some</td>
<td>1</td>
<td>0.73</td>
<td>4.5</td>
<td>3.23</td>
<td>38.43</td>
<td>4073</td>
<td>498</td>
</tr>
<tr>
<td>stratum 3</td>
<td>take-some</td>
<td>1</td>
<td>0.84</td>
<td>11.5</td>
<td>9.91</td>
<td>357.15</td>
<td>1716</td>
<td>1640</td>
</tr>
<tr>
<td>stratum 4</td>
<td>take-all</td>
<td>1</td>
<td>1.01</td>
<td>157.0</td>
<td>59.74</td>
<td>27710.58</td>
<td>793</td>
<td>793</td>
</tr>
<tr>
<td>stratum 5</td>
<td>take-all</td>
<td>1</td>
<td>0.88</td>
<td>256.5</td>
<td>535.26</td>
<td>964079.07</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>stratum 6</td>
<td>take-all</td>
<td>1</td>
<td>0.95</td>
<td>442.0</td>
<td>1012.41</td>
<td>3419027.11</td>
<td>31</td>
<td>31</td>
</tr>
<tr>
<td>stratum 7</td>
<td>take-all</td>
<td>1</td>
<td>0.98</td>
<td>12128.0</td>
<td>7117.76</td>
<td>747924250.02</td>
<td>30</td>
<td>30</td>
</tr>
</tbody>
</table>
Total | | | | | 17177 | | 2393 | | 0.14 |

Total sample size: 1329
Anticipated population mean: 20.31758
Anticipated CV: 0.00890745
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:
<table>
<thead>
<tr>
<th>type</th>
<th>rh</th>
<th>bh</th>
<th>E(Y)</th>
<th>Var(Y)</th>
<th>Nh</th>
<th>nh</th>
<th>fh</th>
</tr>
</thead>
<tbody>
<tr>
<td>stratum 1</td>
<td>take-some</td>
<td>1</td>
<td>1.5</td>
<td>45.22</td>
<td>7.290950e+03</td>
<td>10504</td>
<td>678</td>
</tr>
<tr>
<td>stratum 2</td>
<td>take-some</td>
<td>1</td>
<td>3.5</td>
<td>107.39</td>
<td>2.725574e+04</td>
<td>3313</td>
<td>414</td>
</tr>
<tr>
<td>stratum 3</td>
<td>take-some</td>
<td>1</td>
<td>8.5</td>
<td>242.13</td>
<td>9.397884e+04</td>
<td>2107</td>
<td>488</td>
</tr>
<tr>
<td>stratum 4</td>
<td>take-some</td>
<td>1</td>
<td>20.5</td>
<td>573.46</td>
<td>3.472334e+05</td>
<td>794</td>
<td>354</td>
</tr>
<tr>
<td>stratum 5</td>
<td>take-all</td>
<td>1</td>
<td>195.0</td>
<td>2371.52</td>
<td>5.682434e+06</td>
<td>381</td>
<td>381</td>
</tr>
<tr>
<td>stratum 6</td>
<td>take-all</td>
<td>1</td>
<td>396.0</td>
<td>12979.41</td>
<td>4.037654e+07</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>stratum 7</td>
<td>take-all</td>
<td>1</td>
<td>12128.0</td>
<td>64073.86</td>
<td>8.684400e+09</td>
<td>38</td>
<td>38</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>17177</td>
<td>2393</td>
<td>0.14</td>
</tr>
</tbody>
</table>
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:
\(x = \text{FRAME$EMPLOYEES}\)
\(n = 2393, \text{Ls = 7, takenone = 0, takeall = 0}\)
allocation: \(q_1 = 0.5, q_2 = 0, q_3 = 0.5\)
model = linear: \(\beta = 45.22049, \sigma^2 = 7290.946, \gamma = 1.487\)
algo = Kozak: minsol = 1000, idopti = nh, minNh = 30, maxiter = 10000, maxstep = 20, maxstill = 200, rep = 5, trymany = TRUE

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

Total sample size: 2393
Anticipated population mean: 329.151
Anticipated CV: 0.008894913
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:
\(x = \text{FRAME$EMPLOYEES}\)
\(CV = 0.0089, \text{Ls = 7, takenone = 0, takeall = 0}\)
allocation: \(q_1 = 0.5, q_2 = 0, q_3 = 0.5\)
model = linear: \(\beta = 45.22049, \sigma^2 = 7290.946, \gamma = 1.487\)
algo = Kozak: minsol = 1000, idopti = nh, minNh = 30, maxiter = 10000, maxstep = 20, maxstill = 200, rep = 5, trymany = TRUE

Strata information:
<table>
<thead>
<tr>
<th>type</th>
<th>rh</th>
<th>bh</th>
<th>E(Y)</th>
<th>Var(Y)</th>
<th>Nh</th>
<th>nh</th>
<th>fh</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>take-some</td>
<td>1</td>
<td>2.5</td>
<td>52.67</td>
<td>9.737290e+00</td>
<td>1008</td>
<td>0.08</td>
</tr>
<tr>
<td>2</td>
<td>take-some</td>
<td>1</td>
<td>6.5</td>
<td>181.62</td>
<td>6.142158e+04</td>
<td>2898</td>
<td>584</td>
</tr>
<tr>
<td>3</td>
<td>take-some</td>
<td>1</td>
<td>17.5</td>
<td>457.09</td>
<td>2.501240e+05</td>
<td>1148</td>
<td>467</td>
</tr>
<tr>
<td>4</td>
<td>take-all</td>
<td>1</td>
<td>136.0</td>
<td>1830.74</td>
<td>3.299017e+06</td>
<td>458</td>
<td>458</td>
</tr>
<tr>
<td>5</td>
<td>take-all</td>
<td>1</td>
<td>256.5</td>
<td>8797.22</td>
<td>2.150262e+07</td>
<td>37</td>
<td>37</td>
</tr>
<tr>
<td>6</td>
<td>take-all</td>
<td>1</td>
<td>442.0</td>
<td>15886.98</td>
<td>5.045336e+07</td>
<td>31</td>
<td>31</td>
</tr>
<tr>
<td>7</td>
<td>take-all</td>
<td>1</td>
<td>12128.0</td>
<td>76133.22</td>
<td>1.029406e+10</td>
<td>30</td>
<td>30</td>
</tr>
</tbody>
</table>
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(0.5,0.5), 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 occurred in the computations, therefore some 
statistics do not have finite values