R package **stratification** summary table

Sophie Baillargeon and Louis-Paul Rivest

November 1, 2024

A summary table of the package **stratification** can be found in the appendix of Baillargeon and Rivest (2011). Since the publication of this paper, the package has been updated (see the NEWS file for more details). At the end of this short note you will find an update of this summary table that reflects the changes made to the package. This table aims at providing a quick reference for the R package **stratification**. It lists the five public functions in **stratification** and their arguments. The following notes complete the table

(1) According to the general allocation scheme (Hidiroglou and Srinath, 1993). The stratum sample sizes are proportional to $N_h^{2q_1}\bar{Y}_h^{2q_2}S_{yh}^{2q_3}$ (see `help(stratification)` for more details).

(2) The elements of the `model.control` argument depend on the model :
- `loglinear` model with mortality :

$$Y = \begin{cases} \exp(\alpha + \texttt{beta}\ \log(X) + \texttt{epsilon}) & \text{with probability } p_h \\ 0 & \text{with probability } 1 - p_h \end{cases}$$

where $\texttt{epsilon} \sim N(0, \texttt{sig2})$ is independent of X. The parameter p_h is specified through `ph`, `ptakenone` and `pcertain`.
- heteroscedastic `linear` model :

$$Y = \texttt{beta}X + \texttt{epsilon} \qquad \text{where} \quad \texttt{epsilon} \sim N(0, \texttt{sig2}\ X^{\texttt{gamma}})$$

- `random` replacement model:

$$Y = \begin{cases} X & \text{with probability } 1 - \texttt{epsilon} \\ Xnew & \text{with probability } \texttt{epsilon} \end{cases}$$

where $Xnew$ is a random variable independent of X with the same distribution as X.

The following table presents `model.control` default values according to the model.

model	`beta`	`sig2`	`ph`	`ptakenone`	`pcertain`	`gamma`	`epsilon`
`"loglinear"`	1	0	`rep(1,Ls)`	1	1	–	–
`"linear"`	1	0	–	–	–	0	–
`"random"`	–	–	–	–	–	–	0

(3) The default value of `initbh` is the boundaries obtained with the cumulative root frequency method of Dalenius and Hodges (1959) for Kozak's algorithm, and the set of arithmetic starting points of Gunning and Horgan (2007) for Sethi's algorithm. If `takenone`=1 and `initbh` is of size `Ls`-1, the initial boundary of the take-none stratum is set to the first percentile of `X`.

(4) The following table summarize information about elements of `algo.control`. For a complete description of every element see `help(strata.LH)`. Sethi's algorithm has only one customizable parameter, the maximal number of iterations `maxiter`. However, for Kozak's algorithm, every parameter in the table below apply.

parameter	description	format	default
`maxiter`	maximal number of iterations	positive integer	500 (Sethi) or 10 000 (Kozak)
`minsol`	if the number of solutions is below `minsol` $\Rightarrow$ complete enumeration	integer ≥ 2 and $\leq$ 2 000 000	10 000
`idopti`	identification of stratum sample sizes used in optimization criteria calculation	`"nh"` or `"nhnonint"`	`"nh"`
`minNh`	minimum size for sampled strata	integer ≥ 2	2
`maxstep`	maximal step for boundary modification	integer ≥ 2	$^\star Nu/10$, rounded up and truncated to 100
`maxstill`	maximal number of iterations without boundary modification	positive integer	`maxstep`*10, bounded between 50 and 500
`rep`	number of repetition of the algorithm	integer ≥ 1	5
`trymany`	indicator for trying many initial stratum boundaries	`TRUE` or `FALSE`	`TRUE`

$^\star Nu$ = number of unique values of the stratification variable X (without considering the units in the certainty stratum)

References

Baillargeon, S. and Rivest L.-P. (2011). The construction of stratified designs in R with the package stratification. *Survey Methodology*, **37**(1), 53-65. `http://www.statcan.gc.ca/pub/12-001-x/2011001/article/11447-eng.pdf`

Dalenius, T. and Hodges, J.L., Jr. (1959). Minimum variance stratification, *Journal of the American Statistical Association*, **54**, 88-101.

Gunning, P. and Horgan, J. M. (2007). Improving the Lavallée and Hidiroglou algorithm for stratification of skewed populations, *Journal of Statistical Computation and Simulation*, **77**, 277-291

Hidiroglou, M. A. , and Srinath, K. P. (1993). Problems associated with designing subannual business surveys, *Journal of Business and Economic Statistics*, **11**, 397-405

argument	strata.cumrootf	strata.geo	strata.LH	strata.bh	var.strata	description	format	default
x	✓	✓	✓	✓		stratification variable	numeric vector	none (x is mandatory)
bh				✓		stratum boundaries	numeric vector, length $L-1$	none (bh is mandatory)
n	✓	✓	✓	✓		target total sample size	positive integer	none (n or CV is mandatory)
CV	✓	✓	✓	✓		target CV or RRMSE	numeric	none (n or CV is mandatory)
Ls	✓	✓	✓	✓		number of sampled strata	integer ≥ 2	3
certain	✓	✓	✓	✓		x-indices for units a priori chosen to be in the sample	numeric vector	NULL (no certainty stratum)
alloc	✓	✓	✓	✓		allocation specification (1)	list (q1,q2,q3) where qi≥ 0	Neyman (q1=q3=0.5, q2=0)
takenone			✓	✓		number of take-none strata	0 or 1	0
bias.penalty			✓	✓		penalty for the bias	numeric $\in [0,1]$	1
takeall			✓	✓		number of take-all strata	one of $\{0,1,\ldots,$Ls$-1\}$	0
takeall.adjust				✓		indicator of adjustment for take-all strata	TRUE or FALSE	TRUE (as in the rest of the package)
rh	✓	✓	✓	✓	✓	anticipated response rates	numeric (vector or not)	rep(1,Ls) or rh from strata
rh.postcorr					✓	indicator of posterior correction for non-response	TRUE or FALSE	FALSE (no correction)
model	✓	✓	✓	✓	✓	model identification	"none", "loglinear", "linear"* or "random"* $\rightarrow$	"none" (*unavailable with Sethi's algo)
model.control	✓	✓	✓	✓	✓	model's parameter specification (2)	list (beta, sig2, ph ptakenone, pcertain, gamma, epsilon)	depends on model, but equivalent to model="none"
nclass	✓					number of classes	integer $\geq$ Ls	$min(15$Ls$, Nu)$
initbh			✓			initial stratum boundaries	numeric vector	depends on algo (3)
algo			✓			algorithm identification	"Kozak" or "Sethi"	"Kozak"
algo.control			✓			algorithm's parameters specification (4)	list (maxiter, minsol, idopti, minNh, maxstep, maxstill, rep, trymany)	depends on algo
strata					✓	stratified design	strata object	none (strata is mandatory)
y					✓	study variable	numeric vector	NULL (model given instead)

R package **stratification** summary table