Functions_shapley

Pedro Salas and Paolo Brunori

July 2023

Documentation for the R functions used to estimate Shapley. They are defined to be used for the GEOM database.[1]

1 Libraries

Libraries are updated on the 1^{st} of March, 2023.

Library	Resource
resample	https://cran.r-project.org/web/packages/resample/index.html
gtools	https://cran.r-project.org/web/packages/gtools/index.html
dineq	https://cran.r-project.org/web/packages/dineq/index.html
utils	https://cran.r-project.org/web/packages/R.utils/index.html
data.table	https://cran.r-project.org/web/packages/data.table/index.html
string	https://cran.r-project.org/web/packages/stringr/index.html
stats	https://cran.r-project.org/web/packages/STAT/index.html
stringr	https://cran.r-project.org/web/packages/stringr/index.html

2 Code

2.1 Shapley

shapley(data, model, depname, wts, vars, type, ntree = 1, mincri = 0, minbu = 100, resample = 0.632, centiles = 99, order = 4, rel.ineq = TRUE)

- data: name of dataframe.
- model: formula, e.g., dependent independents.
- depname: name of the dependent variable (with quotation marks, e.g., "income").
- wts: name of the vector of weights (with quotation marks, e.g., "weights").

[1] Note: Some functions call a weights argument. If you do not want to use weights, use a vector of value 1.

- vars: vector of characters with names of circumstances.
- type: type of Shapley value decomposition. Available: "ctree", "ols" or "trafotree".
- ntree = number of trees or iterations. Default = 1.
- mincri: mincriterion (1-alpha) of each tree. Default value = 0.
- minbu: minbucket of each tree. Default value = 100.
- resample: share of observations extracted in each resampling. Default: 0.632.
- centiles: number of centiles to set tranches. Default = 99.
- order: order of the Bernstein Polynomial. Default: 4.
- rel.ineq: If TRUE, y_tilde is delivered to be estimated with relative inequality. If FALSE, y_tilde is delivered to be estimated with the variance.

The function returns three objects:

- shapval: marginal contribution of each variable.
- rel_shapval: relative contribution of each variable (should sum $\sim$ 100).
- rel_shap_max: indexed contribution of each variable. Maximum value is set to 100, and the rest is indexed accordingly.

2.2 Shapley EOp

shapley_eop(data, model, depname, wts, vars, ntree = 1, mincri = 0, minbu = 100, resample = 0.632, centiles = 99, order = 4, rel.ineq = TRUE, lenv = TRUE, share_lenv = 0.1)

- data: name of dataframe.
- model: formula, e.g., dependent independents.
- depname: name of the dependent variable (with quotation marks, e.g., "income").
- wts: name of the vector of weights (with quotation marks, e.g., "weights").
- vars: vector of characters with names of circumstances.
- ntree = number of trees or iterations. Default = 1.
- mincri: mincriterion (1-alpha) of each tree. Default value = 0.
- minbu: minbucket of each tree. Default value = 100.

- resample: share of observations extracted in each resampling. Default: 0.632.
- centiles: number of centiles to set tranches. Default = 99.
- order: order of the Bernstein Polynomial. Default: 4.
- rel.ineq: If TRUE, y_tilde is delivered to be estimated with relative inequality. If FALSE, y_tilde is delivered to be estimated with the variance.
- lenv: Estimate Shapley value decomposition of Lower Envelope.
- share_lenv: Estimate Shapley value decomposition of Robust Lower Envelope. Default bottom share of observations included: 0.º (10%).

The function returns seven objects. For the decomposition of ex-post IOp:

- shapval: marginal contribution of each variable.
- rel_shapval: relative contribution of each variable (should sum $\sim$ 100).
- rel_shap_max: indexed contribution of each variable. Maximum value is set to 100, and the rest is indexed accordingly.

For the decomposition of EOp and EOpX

- shapval_eop: marginal contribution of each variable.
- rel_shap_max_eop: indexed contribution of each variable. Maximum value is set to 100, and the rest is indexed accordingly.
- shapval_eopx: marginal contribution of each variable.
- rel_shap_max_eopx: indexed contribution of each variable. Maximum value is set to 100, and the rest is indexed accordingly.