Package 'formulaic'

October 8, 2019

Title Dynamic Generation and Quality Checks of Formula Objects

Version 0.0.4

Description

Many statistical models and analyses in R are implemented through formula objects. The formulaic package creates a unified approach for programmatically and dynamically generating formula objects. Users may specify the outcome and inputs of a model directly, search for variables to include based upon naming patterns, incorporate interactions, and identify variables to exclude. A wide range of quality checks are implemented to identify issues such as misspecified variables, duplication, a lack of contrast in the inputs, and a large number of levels in categorical data. Variables that do not meet these quality checks can be automatically excluded from the model. These issues are documented and reported in a manner that provides greater accountability and useful information to guide an investigation of the data.

Depends R (>= 3.1.0)

URL https://dachosen1.github.io/formulaic/index.html

BugReports https://github.com/dachosen1/formulaic/issues

License GPL-3

Encoding UTF-8

LazyData TRUE

RoxygenNote 6.1.1

Imports data.table, stats, DT

Suggests knitr, rmarkdown, testthat (>= 2.1.0)

VignetteBuilder knitr

NeedsCompilation no

Author David Shilane [aut],
Caffrey Lee [ctb],
Zichen Huang [ctb],
Anderson Nelson [ctb, cre]

Maintainer Anderson Nelson <an2908@columbia.edu>

Repository CRAN

Date/Publication 2019-10-08 21:00:06 UTC

R topics documented:

`add.backtick` *Add backtick*

Description

Function that add backticks to the input variables.

Usage

```
add.backtick(x, include.backtick = "as.needed")
```

Arguments

`x`	Character value specifying the name of input parameters.
`include.backtick`	specifies whether a backtick should be added. Parameter values should be either 'all' or 'as.needed'

`create.formula` *Create Formula*

Description

Create formula is a tool to automatically create a formula object from a provided variable and output names. Reduces the time required to manually input variables for modeling. Output can be used in linear regression, random forest, neural network etc. Create formula becomes useful when modeling data with multiple features. Reduces the time required for modeling and implementation :

Usage

```
create.formula(outcome.name, input.names = NULL, input.patterns = NULL,
  dat = NULL, interactions = NULL, force.main.effects = TRUE,
  reduce = FALSE, max.input.categories = 20,
  max.outcome.categories.to.search = 4, order.as = "as.specified",
  include.backtick = "as.needed", format.as = "formula",
  variables.to.exclude = NULL, include.intercept = TRUE)
```

Arguments

outcome.name	A character value specifying the name of the formula's outcome variable. In this version, only a single outcome may be included. The first entry of outcome.name will be used to build the formula.
input.names	The names of the variables with the full names delineated.
input.patterns	Includes additional input variables. The user may enter patterns – e.g. to include every variable with a name that includes the pattern. Multiple patterns may be included as a character vector. However, each pattern may not contain spaces and is otherwise subject to the same limits on patterns as used in the grep function.
dat	User can specify a data.frame object that will be used to remove any variables that are not listed in names(dat. As default it is set as NULL. In this case, the formula is created simply from the outcome.name and input.names.
interactions	A list of character vectors. Each character vector includes the names of the variables that form a single interaction. Specifying interactions = list(c("x", "y"), c("x", "z"), c("y", "z"), c("x", "y", "z")) would lead to the interactions x*y + x*z + y*z + x*y*z.
force.main.effects	This is a logical value. When TRUE, the intent is that any term included as an interaction (of multiple variables) must also be listed individually as a main effect.
reduce	A logical value. When dat is not NULL and reduce is TRUE, additional quality checks are performed to examine the input variables. Any input variables that exhibit a lack of contrast will be excluded from the model. This search is global by default but may be conducted separately in subsets of the outcome variables by specifying max.outcome.categories.to.search. Additionally, any input variables that exhibit too many contrasts, as defined by max.input.categories, will also be excluded.
max.input.categories	Limits the maximum number of variables that will be employed in the formula. As default it is set at 20, but users can still change at his/her convenience.
max.outcome.categories.to.search	A numeric value. The create.formula function includes a feature that identifies input variables exhibiting a lack of contrast. When reduce = TRUE, these variables are automatically excluded from the resulting formula. This search may be expanded to subsets of the outcome when the number of unique measured values of the outcome is no greater than max.outcome.categories.to.search. In this case, each subset of the outcome will be separately examined, and any inputs that exhibit a lack of contrast within at least one subset will be excluded.
order.as	User can specify the order the input variables in the formula in a variety of ways for patterns: increasing for increasing alphabet order, decreasing for decreasing alphabet order, column.order for as they appear in data, and as.specified for maintaining the user's specified order.
include.backtick	Add backticks if needed. As default it is set as 'as.needed', which add backticks when only it is needed. The other option is 'all'. The use of include.backtick =

"all" is limited to cases in which the output is generated as a character variable. When the output is generated as a formula object, then R automatically removes all unnecessary backticks. That is, it is only compatible when format.as != formula.

`format.as` The data type of the output. If not set as "formula", then a character vector will be returned.

`variables.to.exclude`
A character vector. Any variable specified in variables.to.exclude will be dropped from the formula, both in the individual inputs and in any associated interactions. This step supersedes the inclusion of any variables specified for inclusion in the other parameters.

`include.intercept`
A logical value. When FALSE, the intercept will be removed from the formula.

Details

Return as the data type of the output. If not set as "formula", then a character vector will be returned. The input.names and names of variables matching the input.patterns will be concatenated to form the full list of input variables.

Examples

```
n <- 10
dd <- data.table::data.table(w = rnorm(n= n), x = rnorm(n = n), pixel_1 = rnorm(n = n))
dd[, pixel_2 := 0.3 * pixel_1 + rnorm(n)]
dd[, y := 5 * x + 3 * pixel_1 + 2 * pixel_2 + rnorm(n)]

create.formula(outcome.name = "y", input.names = "x", input.patterns = c("pi", "xel"), dat = dd)
```

---

reduce.existing.formula

Reduce Existing Formula

Description

The reduce.existing.formula function was designed to perform quality checks and automatic removal of impractical variables can also be accessed when an existing formula has been previously constructed. This method uses natural language processing techniques to deconstruct the components of a formula.

Usage

```
reduce.existing.formula(the.initial.formula, dat,
  max.input.categories = 20, max.outcome.categories.to.search = 4,
  force.main.effects = TRUE, order.as = "as.specified",
  include.backtick = "as.needed", format.as = "formula")
```

Arguments

the.initial.formula	is an object of class "formula" or "character" that states the inputs and output in the form y ~ x1 + x2.
dat	Data frame, list or environment (or object coercible by as.data.frame to a data frame) containing the variables in the model.
max.input.categories	Limits the maximum number of variables that will be employed in the formula.As default it is set at 20, but users can still change at his/her convenience.
max.outcome.categories.to.search	A numeric value. The create.formula function includes a feature that identifies input variables exhibiting a lack of contrast. When reduce = TRUE, these variables are automatically excluded from the resulting formula. This search may be expanded to subsets of the outcome when the number of unique measured values of the outcome is no greater than max.outcome.categories.to.search. In this case, each subset of the outcome will be separately examined, and any inputs builthat exhibit a lack of contrast within at least one subset will be excluded.
force.main.effects	This is a logical value. When TRUE, the intent is that any term included as an interaction (of multiple variables) must also be listed individually as a main effect.
order.as	rearranges its first argument into ascending or descending order.
include.backtick	Add backticks to make a appropriate variable
format.as	The data type of the output. If not set as "formula", then a character vector will be returned.

Examples

```
data('snack.dat')
the.initial.formula <- Awareness ~ .

reduce.existing.formula(the.initial.formula = the.initial.formula,dat = snack.dat,
  max.input.categories = 30)$formula
```

---

snack.dat *Simulated Marketing Data*

Description

contains information from the (fictionalized) marketing survey's data.

Usage

```
snack.dat
```

Format

A data frame of 23000 rows and 23 columns

User ID Character values assigning a unique customer value

Age Numeric values displaying the age of customer in years

Gender Character value describing gender of the customer

Income Numeric values displaying the income of the customer

Region Numeric values describing the region of the customer

Persona Character value describing the customer persona:"Millennial Muncher" "Righteous Reviewer" "Mainstream Maynard" "Savvy Samantha" "Easygoing Edith" "Old School Oliver"

Product Character value describing product consumed by the customer

Awareness Numeric values displaying the customer awareness level

BP_For_Me_0_10 Numeric value displaying brand perception survey result scale (0-10)

BP_Fits_Budget_0_10 Numeric value displaying brand perception survey results for budget scale (0-10)

BP_Tastes_Great_0_10 Numeric value displaying brand perception survey results for tastes scale (0-10)

BP_Good_To_Share_0_10 Numeric value displaying brand perception survey results for good to share scale (0-10)

BP_Like_Logo_0_10 Numeric value displaying brand perception survey results for like logo scale (0-10)

BP_Special_Occasions_0_10 Numeric value displaying brand perception survey results for special occasion scale (0-10)

BP_Everyday_Snack_0_10 Numeric value displaying brand perception survey results for everyday snack scale (0-10)

BP_Healthy_0_10 Numeric value displaying brand perception survey results for healthy scale (0-10)

BP_Delicious_0_10 Numeric value displaying brand perception survey results for delicious scale (0-10)

BP_Right_Amount_0_10 Numeric value displaying brand perception survey results for right amount scale (0-10)

BP_Relaxing_0_10 Numeric value displaying brand perception survey results for relaxing scale (0-10)

Consideration Numeric displaying if the customer would consider this product 1: Yes, 0: No

Consumption Numeric displaying if the customer would consume this product 1: Yes, 0: No

Satisfaction Numeric displaying if the customer was satisfied by this product 1: Yes, 0: No

Advocacy Numeric displaying if the customer would advocate for this product 1: Yes, 0: No

Age Group Categorical variable that breaks the Users into 4 different groups

Income Group Categorical variable that breaks the Users into 5 different levels

Source

"Randomly generated data"

Index