Package: SurrogateRsq (via r-universe)

September 15, 2024

Type Package

Title Goodness-of-Fit Analysis for Categorical Data using the Surrogate R-Squared

Version 0.2.1.9000

Maintainer Xiaorui (Jeremy) Zhu <zhuxiaorui1989@gmail.com>

Description To assess and compare the models' goodness of fit, R-squared is one of the most popular measures. For categorical data analysis, however, no universally adopted R-squared measure can resemble the ordinary least square (OLS) R-squared for linear models with continuous data. This package implement the surrogate R-squared measure for categorical data analysis, which is proposed in the study of Dungang Liu, Xiaorui Zhu, Brandon Greenwell, and Zewei Lin (2022) <doi:10.1111/bmsp.12289>. It can generate a point or interval measure of the surrogate R-squared. It can also provide a ranking measure of the percentage contribution of each variable to the overall surrogate R-squared. This ranking assessment allows one to check the importance of each variable in terms of their explained variance. This package can be jointly used with other existing R packages for variable selection and model diagnostics in the model-building process.

License GPL (>=2)

Encoding UTF-8

LazyData true

RoxygenNote 7.2.3

Depends R (>= 3.5.0), MASS (>= 7.3-54), PAsso (>= 0.1.10), progress (>= 1.2.0), scales (>= 1.1.1)

Suggests R.rsp, knitr, rmarkdown, testthat (>= 3.0.0), dplyr (>= 1.1.1)

Config/testthat/edition 3

URL https://xiaorui.site/SurrogateRsq/

BugReports https://github.com/XiaoruiZhu/SurrogateRsq/issues

VignetteBuilder R.rsp

Roxygen list(markdown = TRUE)

Repository https://xiaoruizhu.r-universe.dev

RemoteUrl https://github.com/xiaoruizhu/surrogatersq

RemoteRef HEAD

RemoteSha 1e78baf5aa4d4ac37d893510843b3f2df074d01d

Contents

RedWine   *Red wine quality dataset of the Portuguese "Vinho Verde" wine*

Description

A red wine tasting preference data used in the study of Cortez, Cerdeira, Almeida, Matos, and Reis 2009. This red wine contains 1599 samples and 12 variables including the tasting preference score of red wine and its physicochemical characteristics.

Usage

```
data(RedWine)
```

Format

A data frame with 1599 rows, quality score, and 11 variables of physicochemical properties of wines.

- quality Tasting preference is a rating score provided by a minimum of three sensory with ordinal values from 0 (very bad) to 10 (excellent). The final sensory score is the median of these evaluations.
- fixed.acidity The fixed acidity is the physicochemical property in unit (g(tartaric acid)/dm^3).
- volatile.acidity The volatile acidity is in unit g(acetic acid)/dm^3.
- citric.acid The citric acidity is in unit g/dm^3.
- residual.sugar The residual sugar is in unit g/dm^3.
- chlorides The chlorides is in unit g(sodium chloride)/dm^3.
- free.sulfur.dioxide The free sulfur dioxide is in unit mg/dm^3.

- `total.sulfur.dioxide` The total sulfur dioxide is in unit mg/dm^3.

- `density` The density is in unit g/cm^3.

- `pH` The wine's pH value.

- `sulphates` The sulphates is in unit g(potassium sulphates)/dm^3.

- `alcohol` The alcohol is in unit \

References

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. (2009), "Modeling wine preferences by data mining from physicochemical properties," Decision Support Systems, 47, 547–553. doi:10.1016/j.dss.2009.05.016

Examples

```
head(RedWine)
```

---

surr_rsq	*A function to calculate the surrogate R-squared measure.*

Description

It can provide the surrogate R-squared for a user specified model. This function will generate an S3 object of surrogate R-squared measure that will be called from other functions of this package. The generic S3 function `print` is also developed to present the surrogate R-squared measure.

Usage

```
surr_rsq(model, full_model, avg.num = 30, asym = FALSE, newdata = NULL, ...)
```

Arguments

`model`	A reduced model that needs to be investigated. The reported surrogate R-squared is for this reduced model.
`full_model`	A full model that contains all of the predictors in the data set. This model object should also contain the dataset for fitting the full model and the reduced model in the first argument.
`avg.num`	The number of replication for the averaging of surrogate R-squared.
`asym`	A logical argument whether use the asymptotic version of our surrogate R-squared. More details are in the paper Liu et al. (2023).
`...`	Additional optional arguments.

Value

An object of class `"surr_rsq"` is a list containing the following components:

surr_rsq   the surrogate R-squared value;

reduced_model   the reduced model under investigation. It should be a subset of the full model;

full_model   the full model used for generating the surrogate response. It should have passed
                initial variable screening and model diagnostics (see Paper for reference);

data   the dataset contains the response variable and all the predictors.

References

Zhu, X., Liu, D., Lin, Z., Greenwell, B. (2022). SurrogateRsq: an R package for categorical data
goodness-of-fit analysis using the surrogate R-squared

Liu, D., Zhu, X., Greenwell, B., & Lin, Z. (2023). A new goodness-of-fit measure for probit models:
Surrogate R2. British Journal of Mathematical and Statistical Psychology, 76(1), 192-210.

Examples

```
data("RedWine")

full_formula <- as.formula(quality ~ fixed.acidity + volatile.acidity +
citric.acid+ residual.sugar + chlorides + free.sulfur.dioxide +
total.sulfur.dioxide + density + pH + sulphates + alcohol)

full_mod <- polr(formula = full_formula,
data=RedWine, method  = "probit")

select_model <- update(full_mod, formula. = ". ~ . - fixed.acidity -
citric.acid - residual.sugar - density")
surr_obj_sele_mod <- surr_rsq(model = select_model, full_model = full_mod,
                              data = RedWine, avg.num = 30)
print(surr_obj_sele_mod$surr_rsq, digits = 3)
```

---

surr_rsq_ci   *A function to calculate the interval estimate of the surrogate R-squared*
                            measure

Description

This function generates the interval measure of surrogate R-squared by bootstrap.

Usage

```
surr_rsq_ci(
  object,
  alpha = 0.05,
  B = 2000,
  asym = FALSE,
  parallel = FALSE,
  ...
)
```

Arguments

object	A object of class "surr_rsq" that is generated by the function "surr_rsq". It contains the following components: surr_rsq, reduced_model, full_model, and data.
alpha	The significance level alpha. The confidence level is 1-alpha.
B	The number of bootstrap replications.
asym	A logical argument whether use the asymptotic version of our surrogate R-squared. More details are in the paper of Liu et al. (2023).
parallel	logical argument whether conduct parallel for bootstrapping surrogate R-squared to construct the interval estimate. The clusters need to be registered through registerDoParallel(cl) beforehand.
...	Additional optional arguments.

Value

An list that contains the CI_lower, CI_upper.

Examples

```
data("RedWine")

full_formula <- as.formula(quality ~ fixed.acidity + volatile.acidity + citric.acid
+ residual.sugar + chlorides + free.sulfur.dioxide +
total.sulfur.dioxide + density + pH + sulphates + alcohol)

fullmodel <- polr(formula = full_formula,data=RedWine, method  = "probit")

select_model <- update(fullmodel, formula. = ". ~ . - fixed.acidity -
citric.acid - residual.sugar - density")

surr_rsq_select <- surr_rsq(select_model, fullmodel, data = RedWine, avg.num = 30)

# surr_rsq_ci(surr_rsq_select, alpha = 0.05, B = 2000, parallel = FALSE) # Not run, it takes time.

# surr_rsq_ci(surr_rsq_select, alpha = 0.05, B = 2000, parallel = TRUE) # Not run, it takes time.
```

---

surr_rsq_rank   *The contribution of each variable in the final model*

Description

This function calculates reduction of the surrogate R-squared goodness-of-fit of each variable to measure their relative explanatory power. This function creates a table containing the reductions of surrogate R-squared by removing each one of variables in the model.

Usage

```
surr_rsq_rank(object, avg.num = 30, var.set = NA, ...)
```

Arguments

object	A object of class "surr_rsq" that is generated by the function "surr_rsq". It contains the following components: surr_rsq, reduced_model, full_model, and data.
avg.num	The number of replication for the averaging of surrogate R-square.
var.set	A list that contains a few sets. Each component of these sets represents the variables that you want to examine for the contribution of goodness of fit. Then, for one component of this list, a model will fit by removing the specified variables.
...	Additional optional arguments.

Value

The default return is a list that contains the contribution of Surrogate R-squared for each variable in the full_model. If the var.set is specified, the return is a list of the contribution of the groups of variables in the var.set.

Examples

```
data("WhiteWine")

sele_formula <- as.formula(quality ~ fixed.acidity + volatile.acidity +
                           residual.sugar +  + free.sulfur.dioxide +
                           pH + sulphates + alcohol)

sele_mod <- polr(formula = sele_formula,
              data = WhiteWine,
              method = "probit")

sur1 <- surr_rsq(model = sele_mod,
              full_model = sele_mod,
              avg.num = 100)

rank_tab_sur1 <- surr_rsq_rank(object  = sur1,
                               avg.num = 30)
```

```
print(rank_tab_sur1)
```

WhiteWine   *White wine quality dataset of the Portuguese "Vinho Verde" wine*

Description

A white wine tasting preference data used in the study of Cortez, Cerdeira, Almeida, Matos, and Reis 2009. This white wine contains 4898 white vinho verde wine samples and 12 variables including the tasting preference score of white wine and its physicochemical characteristics.

Usage

```
data(WhiteWine)
```

Format

A data frame with 4898 rows, quality score, and 11 variables of physicochemical properties of wines.

- quality Tasting preference is a rating score provided by a minimum of three sensory with ordinal values from 0 (very bad) to 10 (excellent). The final sensory score is the median of these evaluations.
- fixed.acidity The fixed acidity is the physicochemical property in unit (g(tartaric acid)/dm^3).
- volatile.acidity The volatile acidity is in unit g(acetic acid)/dm^3.
- citric.acid The citric acidity is in unit g/dm^3.
- residual.sugar The residual sugar is in unit g/dm^3.
- chlorides The chlorides is in unit g(sodium chloride)/dm^3.
- free.sulfur.dioxide The free sulfur dioxide is in unit mg/dm^3.
- total.sulfur.dioxide The total sulfur dioxide is in unit mg/dm^3.
- density The density is in unit g/cm^3.
- pH The wine's pH value.
- sulphates The sulphates is in unit g(potassium sulphates)/dm^3.
- alcohol The alcohol is in unit \

References

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. (2009), "Modeling wine preferences by data mining from physicochemical properties," Decision Support Systems, 47, 547–553. doi:10.1016/j.dss.2009.05.016

Examples

```
head(WhiteWine)
```

Index