Package 'psfmi'

May 16, 2019

Type Package

Depends R (>= 3.4.0),

Imports survival (> 2.41-3), car (> 3.0-0), norm (>= 1.0-9.5), miceadds (> 2.10-14), mitools (> 2.3), foreign (> 0.8-69), pROC (> 1.11.0), rms (> 5.1-2), ResourceSelection (> 0.3-2), ggplot2 (> 2.2.1)

Title Prediction Model Selection and Performance Evaluation in Multiple Imputed Datasets

Version 0.1.0

Description Provides functions to apply pooling or backward selection for logistic or Cox regression prediction models in multiple imputed datasets. Backward selection can be done from the pooled model using Rubin's Rules (RR), the total covariance matrix (D1 method), pooling chi-square values (D2 method), pooling likelihood ratio statistics (D3) or pooling the median p-values. The model can contain continuous, dichotomous, categorical predictors and interaction terms between all type of these predictors. Continuous predictors can also be introduced as restricted cubic spline coefficients. It is also possible to force (spline) predictors or interaction terms in the model during predictor selection. The package also contains functions to generate apparent model performance measures over imputed datasets as ROC/AUC, R-squares, fit test values and calibration plots. A wrapper function over Frank Harrell's validate function is used for that. Bootstrap internal validation is performed in each imputed dataset and results are pooled. Backward selection as part of internal validation is optional and recommended. Also a function to externally validate logistic prediction models in multiple imputed datasets is available.
Eekhout (2017) <doi:10.1186/s12874-017-0404-7>.
Wiel (2009) <doi:10.1093/biostatistics/kxp011>.
Marshall (2009) <doi:10.1186/1471-2288-9-57>.

Encoding UTF-8

LazyData true

RoxygenNote 6.1.1

License GPL (>= 2)

URL https://github.com/mwheymans/psfmi

BugReports https://github.com/mwheymans/psfmi/issues

Suggests knitr, rmarkdown, testthat

VignetteBuilder knitr

NeedsCompilation no

Author Martijn Heymans [cre, aut],
Iris Eekhout [ctb]

Maintainer Martijn Heymans <mw.heymans@amsterdamumc.nl>

Repository CRAN

Date/Publication 2019-05-16 11:50:02 UTC

R topics documented:

D1_cox 2
D1_logistic 3
lbpmicox 3
lbpmilr 4
lbpmilr_dev 5
miperform_lr 6
mivalext_lr 8
psfmi_coxr 10
psfmi_D3 12
psfmi_lr 13

Index **16**

`D1_cox` *D1 method for Predictor selection called by psfmi_cox*

Description

`D1_cox` D1 pooling method

Usage

```
D1_cox(data, impvar, nimp, fm, names.var)
```

Arguments

`data`	Data frame or data matrix with stacked multiple imputed datasets. The original dataset that contains missing values must be excluded from the dataset.
`impvar`	A character vector. Name of the variable that distinguishes the imputed datasets.
`nimp`	A numerical scalar. Number of imputed datasets. Default is 5.
`fm`	regression formula from coxph object
`names.var`	list of predictors included in pooled regression model

Examples

```
D1_cox(data=lbpmicox, nimp=5, impvar="Impnr",
fm=survival::Surv(Time, Status) ~ Duration + Radiation + Onset,
names.var=list("Duration", "Radiation", "Onset"))
```

`D1_logistic` *D1 method for Predictor selection called by psfmi_lr*

Description

`D1_logistic` D1 pooling method

Usage

```
D1_logistic(data, impvar, nimp, fm, names.var)
```

Arguments

`data`	Data frame or data matrix with stacked multiple imputed datasets. The original dataset that contains missing values must be excluded from the dataset.
`impvar`	A character vector. Name of the variable that distinguishes the imputed datasets.
`nimp`	A numerical scalar. Number of imputed datasets. Default is 5.
`fm`	regression formula from glm object
`names.var`	list of predictors included in pooled regression model

Examples

```
D1_logistic(data=lbpmilr, nimp=5, impvar="Impnr",
fm=Chronic ~ Gender + Smoking + Function + JobControl,
names.var=list("Gender", "Smoking", "Function", "JobControl"))
```

`lbpmicox` *Example dataset for psfmi_coxr function*

Description

10 imputed datasets

Usage

```
data(lbpmicox)
```

Format

A data frame with 2650 observations on the following 18 variables.

`Impnr` a numeric vector

`patnr` a numeric vector

`Status` dichotomous event

`Time` continuous follow up time variable

`Duration` continuous

`Previous` dichotomous

`Radiation` dichotomous

`Onset` dichotomous

`Age` continuous

`Tampascale` continuous

`Pain` continuous

`Function` continuous

`Satisfaction` categorical

`JobControl` continuous

`JobDemand` continuous

`Social` continuous

`Expectation` a numeric vector

`Expect_cat` categorical

Examples

```
data(lbpmicox)
## maybe str(lbpmicox)
```

---

lbpmilr *Example dataset for psfmi_lr function*

Description

10 imputed datasets

Usage

```
data(lbpmilr)
```

Format

A data frame with 1590 observations on the following 17 variables.

`Impnr` a numeric vector

`ID` a numeric vector

`Chronic` dichotomous

`Gender` dichotomous

`Carrying` categorical

`Pain` continuous

`Tampascale` continuous

`Function` continuous

`Radiation` dichotomous

`Age` continuous

`Smoking` dichotomous

`Satisfaction` categorical

`JobControl` continuous

`JobDemands` continuous

`SocialSupport` continuous

`Duration` continuous

`BMI` continuous

Examples

```
data(lbpmilr)
## maybe str(lbpmilr)
```

---

lbpmilr_dev *Example dataset for mivalext_lr function*

Description

1 development dataset

Usage

```
data(lbpmilr_dev)
```

Format

A data frame with 108 observations on the following 16 variables.

`ID` a numeric vector

`Chronic` dichotomous

`Gender` dichotomous

`Carrying` categorical

`Pain` continuous

`Tampascale` continuous

`Function` continuous

`Radiation` dichotomous

`Age` continuous

`Smoking` dichotomous

`Satisfaction` categorical

`JobControl` continuous

`JobDemands` continuous

`SocialSupport` continuous

`Duration` continuous

`BMI` continuous

Examples

```
data(lbpmilr_dev)
## maybe str(lbpmilr_dev)
```

`miperform_lr` *Evaluate performance of logistic regression models over MI datasets*

Description

`miperform_lr` Evaluate Performance of logistic regression models

Usage

```
miperform_lr(data, nimp = 5, impvar = NULL, Outcome,
  predictors = NULL, cat.predictors = NULL, int.predictors = NULL,
  cal.plot = FALSE, plot.indiv = FALSE, int.val = FALSE,
  method = "boot", B = 250, bw = FALSE, rule = "p",
  type = "individual", p.val = 0.05, force = NULL)
```

Arguments

data	Data frame or data matrix with stacked multiple imputed datasets. The original dataset that contains missing values must be excluded from the dataset. The imputed datasets must be distinguished by an imputation variable, specified under impvar.
nimp	A numerical scalar. Number of imputed datasets. Default is 5.
impvar	A character vector. Name of the variable that distinguishes the imputed datasets.
Outcome	Character vector containing the name of the outcome variable.
predictors	Character vector with the names of the predictor variables. At least one predictor variable must be defined.
cat.predictors	A single string or a vector of strings to define the categorical variables. Default is NULL categorical predictors.
int.predictors	A single string or a vector of strings with the names of the variables that form an interaction pair, separated by a ":" symbol.
cal.plot	If TRUE a calibration plot is generated. Default is FALSE.
plot.indiv	If TRUE calibration plots of each imputed dataset are generated.
int.val	If TRUE performance measures are reported as a result of internal validation in each imputed datasets. This is a wrapper function of Frank Harrell´s validate function as part of the rms package.
method	"boot" is the default setting to generate bootstrap corrected performance measures.
B	The number of bootstrap resamples, default is 250.
bw	If TRUE backward selection is applied during bootstrap internal validation. Default is FALSE. Backward selection is done using the fastbw function of the rms package.
rule	Set at "p" for backward selection using the p-value as criterium when bw=TRUE.
type	Set at "individual" for backward selection of individual predictors when bw=TRUE.
p.val	P-value criterium for backward selection when bw=TRUE.
force	A vector of integers to define the variables that are forced in the model during backward selection. The integer value matches the order of the variable in the model (starting with the intercept).

Value

A miperform_lr object from which the following objects can be extracted: ROC results as ROC, R squared results as R2, Hosmer and Lemeshow test as HL_test, linear predictor pooled as LP_pooled, performance after internal validation as Int_val_pooled, and Outcome, nimp, impvar, predictors, cat.predictors, int.predictors, int.val.

References

Marshall A, Altman DG, Holder RL, Royston P. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. BMC Med Res Methodol. 2009;9:57.

F. Harrell. Regression Modeling Strategies. With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. Springer, New York, NY, 2015.

Van Buuren S. (2018). Flexible Imputation of Missing Data. 2nd Edition. Chapman & Hall/CRC Interdisciplinary Statistics. Boca Raton.

Harel, O. (2009). The estimation of R2 and adjusted R2 in incomplete data sets using multiple imputation. Journal of Applied Statistics, 36(10), 1109-1118

http://missingdatasolutions.rbind.io/

Examples

```
miperform_lr(data=lbpmilr, nimp=5, impvar="Impnr",
Outcome=c("Chronic"), predictors=c("Gender", "Pain",
"Tampascale","Smoking","Function", "Radiation", "Age"),
cat.predictors=c("Carrying", "Satisfaction"),
int.predictors=c("Carrying:Smoking", "Gender:Smoking"),
cal.plot=TRUE, plot.indiv = FALSE)
```

`mivalext_lr` *External Validation of logistic prediction models in MI datasets*

Description

`mivalext_lr` External validation of logistic prediction models

Usage

```
mivalext_lr(data.val = NULL, data.orig = NULL, nimp = 5,
  impvar = NULL, Outcome, predictors = NULL, lp.orig = NULL,
  cal.plot = FALSE, plot.indiv = FALSE, val.check = FALSE, g = 10)
```

Arguments

`data.val`	Data frame or data matrix with stacked multiple imputed validation datasets. The original dataset that contains missing values must be excluded from the dataset. The imputed datasets must be distinguished by an imputation variable, specified under impvar, and starting by 1.
`data.orig`	A single data frame or data matrix containing the original dataset that was used to develop the model. Used to estimate the original regression coefficients in case lp.orig is not provided.
`nimp`	A numerical scalar. Number of imputed datasets. Default is 5.
`impvar`	A character vector. Name of the variable that distinguishes the imputed datasets.
`Outcome`	Character vector containing the name of the outcome variable.
`predictors`	Character vector with the names of the predictor variables of the model that is validated.

lp.orig	Numeric vector of the original coefficient values that are externally validated.
cal.plot	If TRUE a calibration plot is generated. Default is FALSE.
plot.indiv	If TRUE calibration plots of each imputed dataset are generated. Default is FALSE.
val.check	logical vector. If TRUE the names of the predictors of the LP are provided and can be used as information for the order of the coefficient values as input for lp.orig. If FALSE (default) validation procedure is executed with coefficient values fitted in the order as used under lp.orig.
g	A numerical scalar. Number of groups for the Hosmer and Lemeshow test. Default is 10.

Details

The following information of the externally validated model is provided: pooled ROC curve (median and backtransformed after pooling log transformed ROC curves), pooled Nagelkerke R-Square value (median and backtransformed after pooling Fisher transformed values), pooled Hosmer and Lemeshow Test (using miceadds package), pooled coefficients when model is freely estimated in imputed datasets and the pooled linear predictor (LP), after the externally validated LP is estimated in each imputed dataset (provides information about miscalibration in intercept and slope). When the external validation is very poor, the R2 fixed can become negative due to the poor fit of the model in the external dataset (in that case you may report a R2 of zero).

Value

A `mivalext_lr` object from which the following objects can be extracted: ROC results as `ROC`, R squared results (fixed and calibrated) as `R2 (fixed)` and `R2 (calibr)`, Hosmer and Lemeshow test as `HL_test`, coefficients pooled as `coef_pooled`, linear predictor pooled as `LP_pooled ext`, and `Outcome`, `nimp`, `impvar`, `val.check`, `g` and `coef.check`.

References

F. Harrell. Regression Modeling Strategies. With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. Springer, New York, NY, 2015.

Van Buuren S. (2018). Flexible Imputation of Missing Data. 2nd Edition. Chapman & Hall/CRC Interdisciplinary Statistics. Boca Raton.

http://missingdatasolutions.rbind.io/

Examples

```
mivalext_lr(data.val=lbpmilr, nimp=10, impvar="Impnr", Outcome="Chronic",
predictors=c("Gender", "factor(Carrying)", "Function", "Tampascale",  "Age"),
lp.orig=c(-9.2, -0.34, 0.92, 1.5, 0.5, 0.26, -0.02),
cal.plot=TRUE, plot.indiv=TRUE, val.check = TRUE)

mivalext_lr(data.val=lbpmilr, nimp=5, impvar="Impnr", Outcome="Chronic",
predictors=c("Gender", "factor(Carrying)", "Function", "Tampascale", "Age"),
lp.orig=c(-9.2, -0.34, 0.92, 1.1, -0.05, 0.26, -0.02),
cal.plot=TRUE, plot.indiv=TRUE, val.check = FALSE)
```

psfmi_coxr *Pooling and predictor selection function for Cox regression models in multiply imputed datasets*

Description

psfmi_coxr Pooling and backward selection for Cox regression models in multiply imputed datasets using different selection methods.

Usage

```
psfmi_coxr(data, nimp = 5, impvar = NULL, time, status,
  predictors = NULL, p.crit = 1, cat.predictors = NULL,
  spline.predictors = NULL, int.predictors = NULL,
  keep.predictors = NULL, knots = NULL, method = NULL,
  print.method = FALSE)
```

Arguments

data	Data frame or data matrix with stacked multiple imputed datasets. The original dataset that contains missing values must be excluded from the dataset. The imputed datasets must be distinguished by an imputation variable, specified under impvar, and starting by 1.
nimp	A numerical scalar. Number of imputed datasets. Default is 5.
impvar	A character vector. Name of the variable that distinguishes the imputed datasets.
time	Follow up time.
status	The status variable, normally 0=censoring, 1=event.
predictors	Character vector with the names of the predictor variables. At least one predictor variable has to be defined.
p.crit	A numerical scalar. P-value selection criterium.
cat.predictors	A single string or a vector of strings to define the categorical variables. Default is NULL categorical predictors.
spline.predictors	A single string or a vector of strings to define the (restricted cubic) spline variables. Default is NULL spline predictors.
int.predictors	A single string or a vector of strings with the names of the variables that form an interaction pair, separated by a ":" symbol.
keep.predictors	A single string or a vector of strings including the variables that are forced in the model during predictor selection. Categorical and interaction variables are allowed. See details.
knots	A numerical vector that defines the number of knots for each spline predictor separately.

`method` A character vector to indicate the pooling method for p-values to pool the total model or used during predictor selection. This can be "D1", "D2", or "MPR". See details for more information.

`print.method` logical vector. If TRUE full matrix with p-values of all variables according to chosen method (under method) is shown. If FALSE (default) p-value for categorical variables according to method are shown and for continuous and dichotomous predictors Rubin's Rules are used.

Details

The basic pooling procedure to derive pooled coefficients, standard errors, 95 confidence intervals and p-values is Rubin's Rules (RR). Specific procedures are available to derive pooled p-values for categorical (> 2 categories) and spline variables. print.method allows to choose between these pooling methods that are: "D1" is pooling of the total covariance matrix, "D2" is pooling of Chi-square values, and "MPR" is pooling of median p-values (MPR rule). Spline regression coefficients are defined by using the rcs function for restricted cubic splines of the rms package of Frank Harrell. A minimum number of 3 knots as defined under knots is needed.

Value

A `psfmi_coxr` object from which the following objects can be extracted: pooled model as `RR_model`, pooled p-values according to pooling method as `multiparm_p`, predictors excluded at each step as `coef.excl_step`, and `impvar`, `nimp`, `method`, `p.crit`, `predictors`, `cat.predictors`, `keep.predictors`, `int.predictors`, `spline.predictors`, `knots`, `print.method`.

References

Eekhout I, van de Wiel MA, Heymans MW. Methods for significance testing of categorical covariates in logistic regression models after multiple imputation: power and applicability analysis. BMC Med Res Methodol. 2017;17(1):129.

Enders CK (2010). Applied missing data analysis. New York: The Guilford Press.

van de Wiel MA, Berkhof J, van Wieringen WN. Testing the prediction error difference between 2 predictors. Biostatistics. 2009;10:550-60.

Marshall A, Altman DG, Holder RL, Royston P. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. BMC Med Res Methodol. 2009;9:57.

Van Buuren S. (2018). Flexible Imputation of Missing Data. 2nd Edition. Chapman & Hall/CRC Interdisciplinary Statistics. Boca Raton.

http://missingdatasolutions.rbind.io/

Examples

```
pool_coxr <- psfmi_coxr(data=lbpmicox, nimp=5, impvar="Impnr", time="Time",
status="Status", predictors=c("Duration", "Radiation", "Onset"), p.crit=1,
method="D1", cat.predictors=c("Expect_cat"))
pool_coxr$RR_Model
pool_coxr$multiparm_p
```

```
pool_coxr <- psfmi_coxr(data=lbpmicox, nimp=5, impvar="Impnr", time="Time",
status="Status", predictors=c("Previous",  "Radiation", "Onset",
"Function", "Tampascale" ), p.crit=0.05, cat.predictors=c("Expect_cat"),
int.predictors=c("Tampascale:Radiation",
"Expect_cat:Tampascale"), keep.predictors = "Tampascale", method="D2")
pool_coxr$RR_Model
pool_coxr$multiparm_p
```

---

psfmi_D3 *Meng & Rubin pooling method called by psfmi_lr*

Description

psfmi_D3 Function to pool using Meng & Rubin pooling method

Usage

```
psfmi_D3(data, nimp, impvar, Outcome, P, p.crit, print.method)
```

Arguments

data	Data frame or data matrix with stacked multiple imputed datasets. The original dataset that contains missing values must be excluded from the dataset.
nimp	A numerical scalar. Number of imputed datasets. Default is 5.
impvar	A character vector. Name of the variable that distinguishes the imputed datasets.
Outcome	Character vector containing the name of the outcome variable.
P	Character vector with the names of the predictor variables. At least one predictor variable has to be defined.
p.crit	A numerical scalar. P-value selection criterium.
print.method	logical vector. If TRUE full matrix with p-values of all variables according to chosen method (under method) is shown. If FALSE (default) p-value for categorical variables according to method are shown and for continuous and dichotomous predictors Rubin's Rules are used

Examples

```
psfmi_D3(data=lbpmilr, nimp=5, impvar="Impnr",
P=c("Gender", "Smoking", "Function", "JobControl"),
Outcome="Chronic", print.method = FALSE)
```

psfmi_lr *Pooling and Predictor selection function for Logistic regression models in multiply imputed datasets*

Description

psfmi_lr Pooling and backward selection for Logistic regression prediction models in multiply imputed datasets using different selection methods.

Usage

```
psfmi_lr(data, nimp = 5, impvar = NULL, Outcome, predictors = NULL,
  p.crit = 1, cat.predictors = NULL, spline.predictors = NULL,
  int.predictors = NULL, keep.predictors = NULL, knots = NULL,
  method = NULL, print.method = FALSE)
```

Arguments

data	Data frame or data matrix with stacked multiple imputed datasets. The original dataset that contains missing values must be excluded from the dataset. The imputed datasets must be distinguished by an imputation variable, specified under impvar, and starting by 1.
nimp	A numerical scalar. Number of imputed datasets. Default is 5.
impvar	A character vector. Name of the variable that distinguishes the imputed datasets.
Outcome	Character vector containing the name of the outcome variable.
predictors	Character vector with the names of the predictor variables. At least one predictor variable has to be defined.
p.crit	A numerical scalar. P-value selection criterium.
cat.predictors	A single string or a vector of strings to define the categorical variables. Default is NULL categorical predictors.
spline.predictors	A single string or a vector of strings to define the (restricted cubic) spline variables. Default is NULL spline predictors. See details.
int.predictors	A single string or a vector of strings with the names of the variables that form an interaction pair, separated by a ":" symbol.
keep.predictors	A single string or a vector of strings including the variables that are forced in the model during predictor selection. Categorical and interaction variables are allowed.
knots	A numerical vector that defines the number of knots for each spline predictor separately.
method	A character vector to indicate the pooling method for p-values to pool the total model or used during predictor selection. This can be "D1", "D2", "D3" or "MPR". See details for more information.

print.method logical vector. If TRUE full matrix with p-values of all variables according to chosen method (under method) is shown. If FALSE (default) p-value for categorical variables according to method are shown and for continuous and dichotomous predictors Rubin's Rules are used.

Details

The basic pooling procedure to derive pooled coefficients, standard errors, 95 confidence intervals and p-values is Rubin's Rules (RR). Specific procedures are available to derive pooled p-values for categorical (> 2 categories) and spline variables. print.method allows to choose between these pooling methods that are: "D1" is pooling of the total covariance matrix, "D2" is pooling of Chi-square values, "D3" is pooling Likelihood ratio statistics (method of Meng and Rubin) and "MPR" is pooling of median p-values (MPR rule). Spline regression coefficients are defined by using the rcs function for restricted cubic splines of the rms package of Frank Harrell. A minimum number of 3 knots as defined under knots is needed.

Value

A `psfmi_lr` object from which the following objects can be extracted: pooled model as `RR_model`, pooled p-values according to pooling method as `multiparm_p`, predictors excluded at each step as `coef.excl_step`, and `impvar`, `nimp`, `Outcome`, `method`, `p.crit`, `predictors`, `cat.predictors`, `keep.predictors`, `int.predictors`, `spline.predictors`, `knots`, `print.method`.

References

Eekhout I, van de Wiel MA, Heymans MW. Methods for significance testing of categorical covariates in logistic regression models after multiple imputation: power and applicability analysis. BMC Med Res Methodol. 2017;17(1):129.

Enders CK (2010). Applied missing data analysis. New York: The Guilford Press.

Meng X-L, Rubin DB. Performing likelihood ratio tests with multiply-imputed data sets. Biometrika.1992;79:103-11.

van de Wiel MA, Berkhof J, van Wieringen WN. Testing the prediction error difference between 2 predictors. Biostatistics. 2009;10:550-60.

Marshall A, Altman DG, Holder RL, Royston P. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. BMC Med Res Methodol. 2009;9:57.

Van Buuren S. (2018). Flexible Imputation of Missing Data. 2nd Edition. Chapman & Hall/CRC Interdisciplinary Statistics. Boca Raton.

http://missingdatasolutions.rbind.io/

Examples

```
pool_lr <- psfmi_lr(data=lbpmilr, nimp=5, impvar="Impnr", Outcome="Chronic",
predictors=c("Gender", "Smoking", "Function", "JobControl",
"JobDemands", "SocialSupport"), method="D1")
pool_lr$RR_Model
pool_lr$multiparm_p
```

```
pool_lr <- psfmi_lr(data=lbpmilr, nimp=5, impvar="Impnr", Outcome="Chronic",
predictors=c("Gender", "Smoking", "Function", "JobControl",
"JobDemands", "SocialSupport"), p.crit = 0.05, method="D1")
pool_lr$RR_Model
pool_lr$multiparm_p
```

Index

∗Topic **datasets**
lbpmicox, 3
lbpmilr, 4
lbpmilr_dev, 5

D1_cox, 2
D1_logistic, 3

lbpmicox, 3
lbpmilr, 4
lbpmilr_dev, 5

miperform_lr, 6
mivalext_lr, 8

psfmi_coxr, 10
psfmi_D3, 12
psfmi_lr, 13