Package 'NU.Learning'

July 21, 2025

Version 1.5

Date 2023-09-15

Title Nonparametric and Unsupervised Learning from Cross-Sectional Observational Data

Author Bob Obenchain [aut, cre], Stan Young [ctb]

Maintainer Bob Obenchain <wizbob@att.net>

Depends R (>= 3.5.0), cluster, lattice

Description Especially when cross-sectional data are observational, effects of treatment selection bias and confounding are best revealed by using Nonparametric and Unsupervised methods to ``Design" the analysis of the given data ...rather than the collection of ``designed data". Specifically, the ``effect-size distribution" that best quantifies a potentially causal relationship between a numeric y-Outcome variable and either a binary t-Treatment or continuous e-Exposure variable needs to consist of BLOCKS of relatively well-matched experimental units (e.g. patients) that have the most similar X-confounder characteristics. Since our NU Learning approach will form BLOCKS by ``clustering" experimental units in confounder X-space, the implicit statistical model for learning is One-Way ANOVA. Within Block measures of effect-size are then either [a] LOCAL Treatment Differences (LTDs) between Within-Cluster y-Outcome Means (``new" minus ``control") when treatment choice is Binary or else [b] LOCAL Rank Correlations (LRCs) when the e-Exposure variable is numeric with (hopefully many) more than two levels. An Instrumental Variable (IV) method is also provided so that Local Average y-Outcomes (LAOs) within BLOCKS may also contribute information for effect-size inferences when X-Covariates are assumed to influence Treatment choice or Exposure level but otherwise have no direct effects on y-Outcomes. Finally, a ``Most-Like-Me" function provides histograms of effect-size distributions to aid Doctor-Patient (or Researcher-Society) communications about Heterogeneous Outcomes. Obenchain and Young (2013) <doi:10.1080/15598608.2013.772821>; Obenchain, Young and Krstic (2019) <doi:10.1016/j.yrtph.2019.104418>.

License GPL-2

URL https://www.r-project.org, http://localcontrolstatistics.org

NeedsCompilation no

Repository CRAN

Date/Publication 2023-09-30 22:52:43 UTC

Contents

`NU.Learning-package` *NU.Learning: Nonparametric and Unsupervised Adjustment for Bias and Confounding*

Description

NU.Learning forms Local Treatment Differences (LTDs) or Local Rank Correlations (LRCs) within Clusters of experimental units (patients, etc.) who have been relatively well-matched on their baseline X-confounder characteristics. The resulting distribution of LTD/LRC effect-size estimates can be interpreted much like a Bayesian posterior. Yet these distributions have been formed, via Nonparametric and Unsupervised Preprocessing, in purely Objective Ways.

Details

Package:	NU.Learning
Type:	Package
Version:	1.5
Date:	2023-09-15
License:	GPL-2

UNSUPERVISED LOCAL TREATMENT DIFFERENCES or LOCAL RANK CORRELATIONS:

Multiple calls to ltdagg(K) or lrcagg(K) for varying numbers of clusters, K, are typically made after first invoking NUcluster() to hierarchically cluster patients in X-space and invoking NUsetup() to specify a numeric y-Outcome variable and a numeric treatment choice or exposure level measure, trex.

UNSUPERVISED INSTRUMENTAL VARIABLES = LOCAL AVERAGE y-OUTCOME EFFECTS:

An OBSERVED Propensity Score (PS) is defined here to be either (i) the local (within-cluster) fraction of experimental units (patients) receiving trex==1 (new) rather than trex==0 (control) or else (ii) a measure of "relative exposure" when the numeric trex measure has (many) more than 2 observed levels. Multiple calls to ivadj(K) for varying numbers of clusters, K, then yield alternative scatters of Local Average Outcomes (LAOs) for Clusters when plotted against their PS estimates and, thus, different possible linear fits or smooth.splines() yielding potentially different inferences about across-cluster Treatment or Exposure Effects.

CONFIRMATION and SENSITIVITY ANALYSES of LOCAL EFFECT-SIZE DISTRIBUTIONS:

For a given value of K = Number of Clusters requested, the output object from ltdagg(K) or lrcagg(K) can be input to confirm() to use (nonparametric) permutation theory to display visual evidence (empirical CDF comparisons) concerning the Question: "Does x-matching Truly Matter?" The NULL hypothesis here is that the x-Covariates used in Clustering / Matching of Experimental Units are actually IGNORABLE. Evidence against this hypothesis is provided when the observed LOCAL Effect-Size Distribution clearly deviates from the purely RANDOM, NULL distribution computed (to any desired precision) by randomly PERMUTING cluster ID labels across experimental units. Furthermore, the statistical significance of differences between the observed and random NULL distributions can be estimated using KSperm(), which simulates the random permutation distribution of the Kolmogorov-Smirnov D-statistic when many tied values occur in both distributions being compared. Finally, the NUcompare() function helps users of NU.Learning decide which Number of Clusters, K, optimizes Variance-Bias trade-offs. Larger values of K tend to yield smaller clusters with better matches and, thus, potentially reduced BIAS. On the other hand, smaller values of K usually yield local effect-size estimates with much lower Variability (higher Precision).

"Most-Like-Me" HISTOGRAMS for DOCTOR-PATIENT discussions of PERSONALIZED MEDICINE:

For a specified vector, xvec, of numerical values of the X-confounder variables used in the current CLUSTERING of eUnits, display histograms of observed LTD or LRC effect-sizes for (i) all available patients and (ii) for the specified number, NN, of "Nearest-Neighbors" in X-confounder space of the TARGET eUnit ...i.e. xvec defines "Me".

Author(s)

Bob Obenchain <wizbob@att.net>

References

McClellan M, McNeil BJ, Newhouse JP. (1994) Does More Intensive Treatment of Myocardial Infarction in the Elderly Reduce Mortality?: Analysis Using Instrumental Variables. *JAMA* **272**: 859-866.

Obenchain RL. (2010) The Local Control Approach using JMP. Chapter 7 of **Analysis of Observational Health Care Data using SAS**, *Cary, NC:SAS Press*, pages 151-192.

Obenchain RL, Young SS. (2013) Advancing Statistical Thinking in Observational Health Care Research. *J. Stat. Theory and Practice*, **7**: 456-469, doi:10.1080/15598608.2013.772821.

Lopiano KK, Obenchain RL, Young SS. (2014) Fair treatment comparisons in observational research. *Statistical Analysis and Data Mining*, **7**: 376-384, doi:10.1002/sam.11235.

Obenchain RL. NU.Learning-vignette. (2023) **NU.Learning_in_R.pdf** http://localcontrolstatistics.org

Rosenbaum PR, Rubin RB. (1983) The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* **70**: 41-55.

Rosenbaum PR, Rubin RB. (1984) Reducing Bias in Observational Studies Using Subclassification on a Propensity Score. *JASA* **79**: 516-524.

Rubin DB. (1980) Bias reduction using Mahalanobis metric matching. *Biometrics* **36**: 293-298.

Stuart EA. (2010) Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science* **25**: 1-21.

`confirm` *Confirm that Clustering in Covariate X-space yields an "adjusted" LTD/LRC effect-size Distribution*

Description

For a given Number of Clusters, K, confirm() compares the observed distribution of LTDs or LRCs from relatively well-matched experimental units with the corresponding distribution from Purely Random Clusterings of experimental units. The larger are differences between the (blue) observed empirical CDF of effect-sizes and the (red) Purely Random CDF, the more potentially IMPORTANT are the "adjustments" resulting from focussing upon clustering (matching) of experimental units in X-space.

Usage

```
confirm(x, reps=100, seed=12345)
```

Arguments

`x`	An output object from ltdagg() or lrcagg() for a specified number of clusters, K.
`reps`	Number of simulation Replications, each with the same number, K, and sizes, N1, N2, ..., NK of Purely Random clusters.
`seed`	This (arbitrary) integer argument will be passed to the R set.seed() function. Knowing the value of this seed makes the output from confirm() reproducible.

Details

Making calls to confirm() for ltdagg() or lrcagg() objects resulting from different choices of K = Numbers of Clusters help the analyst decide which observed LTD or LRC effect-size distributions are (or are not) meaningfully different from Purely Random. When the X-covariates used in NUcluster() are truly "ignorable," then [i] all X-based clusters will be Purely Random, and [ii] both the number (K) and the sizes (N1, N2, ...,NK) of clusters formed will be meaningless and arbitrary. Thus the NU Strategy confirm() function simulates the empirical CDF for LTDs or LRCs resulting from purely random permutations of the Cluster ID numbers (1, 2, ...,K) assigned by ltdagg()

or lrcagg(). Each permutation yields K artificial "clusters" of sizes N1, N2, ..., NK. Simulation results are accumulated for the total number of random permutations specified in the "reps=" argument of confirm(). Calls to print.confirm() and plot.confirm() provide information on comparisons of empirical CDFs for the Observed and Purely Random LTD/LRC distributions, including calculation of an observed two-sample Kolmogorov-Smirnov D-statistic using stats::ks.test. This is a non-standard use of ks.test() because the distributions being compared are DISCRETE; both contain many within-cluster TIED effect-size estimates. The p-value computed by ks.test() is not reported or saved because it is badly biased downwards due to TIED estimates. Researchers wishing to simulate a p-value for the observed KS D-statistic that is adjusted for TIES can invoke KSperm(confirm()).

Value

An output list object of class confirm:

`hiclus`	Hierarchical clustering object created by the designated method.
`dframe`	Name of data.frame containing X, trex & Y variables.
`trtm`	Name of numerical trex variable.
`yvar`	Name of numerical Y-outcome variable.
`reps`	Number of overall Replications, each with the same numbers of requested clusters.
`seed`	Integer argument passed to set.seed(). Knowing which seed value was used in the call to confirm() makes not only the NULL distribution of observed LTDs or LRCs reproducible but also makes the NULL distribution of D-statistics (adjusted for ties) from a subsequent call to KSperm() reproducible.
`nclus`	Number of clusters requested.
`units`	Number of experimental units or patients.
`Type`	1 ==> LTDs, otherwise LRCs.
`NUmean`	Weighted Local Mean across Clusters.
`NUstde`	Weighted Std. Error across Clusters.
`RPmean`	Weighted Random Permutation Mean across Clusters.
`RPstde`	Weighted Random Permutation Std. Error across Clusters.
`KSobsD`	Output from print(ks.test()).
`NUdist`	data.frame of 5 key variables for all experimental units.
`dfconf`	data.frame of lstat = LTD or LRC values of max(length) = reps*units.

Author(s)

Bob Obenchain <wizbob@att.net>

References

Obenchain RL. (2010) The Local Control Approach using JMP. Chapter 7 of **Analysis of Observational Health Care Data using SAS**, *Cary, NC:SAS Press*, pages 151-192.

Obenchain RL. (2023) **NU.Learning_in_R.pdf** http://localcontrolstatistics.org

See Also

ltdagg and lrcagg.

---

ivadj *Instrumental Variable LAO Fitting and Smoothing*

Description

For a given number of patient clusters in baseline X-covariate space and a specified Y-outcome variable, smooth the distribution of Local Average Outcomes (LAOs) plotted versus Within-Cluster Propensity-like Scores: the Treatment Selection Fraction or the Relative Exposure Level.

Usage

```
ivadj(x)
```

Arguments

x	An output object from ltdagg() or lrcagg() using K Clusters in X-covariate space.

Details

Multiple invocations of ivadj(ltdagg()) or ivadj(lrgagg()) using varying numbers of clusters, K, can be made. Each invocation of ivadj() displays a linear lm() fit and a smooth.spline() fit to the scatter of LAO estimates plotted versus their within-cluster propensity-like score estimates.

Value

An output list object of class ivadj:

hclobj	Name of clustering object output by NUcluster().
dframe	Name of data.frame containing X, trtm & Y variables.
trtm	Name of the numeric treatment variable.
yvar	Name of the numeric outcome Y variable.
K	Number of Clusters Requested.
actclust	Number of Clusters actually produced.

Author(s)

Bob Obenchain <wizbob@att.net>

References

McClellan M, McNeil BJ, Newhouse JP. (1994) Does More Intensive Treatment of Myocardial Infarction in the Elderly Reduce Mortality?: Analysis Using Instrumental Variables. *JAMA* **272**: 859-866.

Obenchain RL. (2010) Local Control Approach using JMP. Chapter 7 of **Analysis of Observational Health Care Data using SAS**, *Cary, NC:SAS Press*, pages 151-192.

Obenchain RL. (2023) **NU.Learning_in_R.pdf** http://localcontrolstatistics.org

Rosenbaum PR, Rubin RB. (1983) The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* **70**: 41-55.

See Also

`ltdagg`, `lrcagg` and `NUcompare`.

Examples

```
# Running takes about 7 seconds...
data(pci15k)
xvars = c("stent", "height", "female", "diabetic", "acutemi", "ejfract", "ves1proc")
hclobj = NUcluster(pci15k, xvars)
NU.env = NUsetup(hclobj, pci15k, thin, surv6mo)
surv050 = ltdagg(50, NU.env)
iv050 = ivadj(surv050)
iv050
plot(iv050)
```

`KSperm` *Simulate a p-value for the significance of the Kolmogorov-Smirnov D-statistic from confirm().*

Description

For a given confirm() output object, KSperm() simulates the NULL distribution of LTDs or LRCs resulting from Purely Random Clusterings of experimental units within the parent data.frame. This NULL distribution is discrete because Local Effect-Size estimates are TIED within-clusters. The observed D-Statistic from confirm() is compared with new NULL order statistics computed by KSperm(), again using stats::ks.test. When KSperm() is called immediately after confirm() and the seed value used in confirm() is known, then both the simulated p-value and the additional NULL KS-D order statistics generated by KSperm() will all be reproducible.

Usage

```
KSperm(x, reps=100)
```

Arguments

x	An output object from confirm().
reps	This is the number of new NULL KS-D statistics to generated. Each experimental unit is used at most once within each full replication. No clusters will be empty, but some may be "uninformative".

Details

The observed value of the Kolmogorov-Smirnov D-statistic from confirm() is used here, but its "p.value" from ks.test() is not because it is badly biased downwards. This bias results because the distribution of LTDs or LRCs across clusters is always discrete, due to TIED values within clusters that typically also vary in size. Thus, KSperm() generates "reps" additional, independent, NULL values of KS-D and saves their order statistics. Finally, KSperm() compares the Observed KS-D from confirm() with its simulated NULL order statistics to estimate an appropriately "adjusted" p-value, pv.adj. Note that the simulated pv.adj value estimate cannot be less than 1/(reps).

Value

An output list object of class KSperm:

hiclus	Hierarchical clustering object created by the designated method.
dframe	Name of data.frame containing X, t & Y variables.
trtm	Name of numerical treatment/exposure variable.
yvar	Name of numerical y-Outcome variable.
Type	1 ==> LTDs, otherwise LRCs.
reps	Number of overall Replications, each with the same number, K, of requested clusters.
nclus	Number of clusters requested.
units	Number of experimental units or patients.
obsD	Observed numerical value of KS D-statistic from confirm()
Dvec	Vector of order statistics for simulated NULL KS D-statistics.
pv.adj	Simulated p-value adjusted for TIES within discrete LTD/LRC distributions.

Author(s)

Bob Obenchain <wizbob@att.net>

References

Obenchain RL. (2010) Local Control Approach using JMP. Chapter 7 of **Analysis of Observational Health Care Data using SAS**, *Cary, NC:SAS Press*, pages 151-192.

Obenchain RL. (2019) **NU.Learning_in_R.pdf** http://localcontrolstatistics.org

See Also

confirm, ltdagg and lrcagg.

`lrcagg` *Calculate the observed Distribution of LRCs in NU.Learning*

Description

For a given number, K, of Clusters of Experimental Units in baseline X-covariate space, lrcagg() calculates the observed distribution of "Local Rank Correlations" (LRCs) across Clusters ...where each LRC = cor(trex, Y, method = "spearman") within a Cluster, trex is a numeric measure of Exposure, and Y is a numeric measure of Outcome.

Usage

```
lrcagg(K, envir)
```

Arguments

`K`	Number of Clusters in baseline X-covariate space.
`envir`	R environment output by a previous call to NUsetup().

Details

Multiple calls to lrcagg(K) for varying numbers of clusters, K, are typically made after first invoking NUcluster() to hierarchically cluster patients in X-space and then invoking NUsetup() to specify a Y Outcome variable and a continuous, numerical treatment Exposure: trex. lrcagg() computes an observed LRC Distribution, updates information stored in its envir object, and outputs an object that is typically saved in the user's .GlobalEnv to allow subsequent use by print.lrcagg(), plot.lrcagg(), confirm() or KSperm(). Uninformative Clusters (those containing only 1 or 2 experimental units) contribute **NA** values to the LRCtabl$LRC and LRCdist$LRC objects within the lrcagg() output list.

Value

An output list of 12 objects, of class lrcagg:

`hclobj`	Name of clustering dendrogram object created by NUcluster().
`dframe`	Name of data.frame containing X, trex & Y variables.
`trex`	Name of numerical treatment/exposure level variable.
`yvar`	Name of outcome Y variable.
`K`	Number of Clusters Requested.
`actclust`	Number of Clusters delivered.
`LRCtabl`	data.frame with 5 columns and K rows for Clusters.
`LRCtabl$c`	Cluster ID Factor, "1", "2", ..., "K".
`LRCtabl$LRC`	Numerical value of Local Treatment Difference for a Cluster.
`LRCtabl$w`	Integer value of "weight" = Cluster Size.

LRCtabl$LAO	Numerical value of within-cluster Local Average Outcome (Y-value).
LRCtabl$PS	Numerical value of Local Relative Propensity for Exposure, 0.0 to 1.0.
LRCdist	data.frame with 5 columns and same number of rows as the data: dframe.
LRCdist$c.K	Cluster ID Variable of the form: "c.K"
LRCdist$ID	Observation ID Variable for the rows of the input dframe.
LRCdist$y	Numerical values of Y-Outcomes for Experimental Units.
LRCdist$t	Numerical values of Treatment-Exposure Levels for Experimental Units.
LRCdist$LRC	Numerical values of the LRC for the Cluster containing each Unit.
infoclus	Integer value of Number of Informative Clusters.
infounits	Integer value of Number of Units within Informative Clusters.
LRCmean	Numerical value of mean(LRCdist$LRC) = Weighted Average of LRCtabl$LRC values.
LRCstde	Numerical value of sqrt(var(LRCdist$LRC)) = Weighted Standard Deviation of LRCtabl$LRC values.

Author(s)

Bob Obenchain <wizbob@att.net>

References

Obenchain RL. (2010) The Local Control Approach using JMP. Chapter 7 of **Analysis of Observational Health Care Data using SAS**, *Cary, NC:SAS Press*, pages 151-192.

Obenchain RL. (2019) **NU.Learning_in_R.pdf** http://localcontrolstatistics.org

See Also

ivadj, ltdagg and NUcompare.

Examples

```
data(radon)
xvars = c("obesity", "over65", "cursmoke")
hclobj = NUcluster(radon, xvars)
e = NUsetup(hclobj, radon, lnradon, lcanmort)
lrc050 = lrcagg(50, e)
lrc050
plot(lrc050, e)
```

`ltdagg` *Calculate the Observed Distribution of LTDs in NU.Learning*

Description

For a given number, K, of Clusters of Experimental Units in baseline X-covariate space, ltdagg() calculates the observed distribution of "Local Treatment Differences" (LTDs) of the form LTD = ((mean(Y) for units receiving trtm==1) - (mean(Y) for units receiving trtm==0)).

Usage

```
ltdagg(K, envir)
```

Arguments

`K`	Number of Clusters in baseline X-covariate space.
`envir`	R environment output by a previous call to NUsetup().

Details

Multiple calls to ltdagg(K) for varying numbers of clusters, K, are typically made after first invoking NUcluster() to hierarchically cluster patients in X-space and then invoking NUsetup() to specify a Y Outcome variable and a two-level, numerical treatment variable: trtm. ltdagg() computes an observed LTD Distribution, updates information stored in its envir object, and outputs an object that is typically saved in the user's .GlobalEnv to allow subsequent use by print.ltdagg(), plot.ltdagg(), confirm() or KSperm(). Uninformative Clusters (those containing either only trtm==1 units or else only trtm==0 units) contribute **NA** values to the LTDtabl$LTD and LTDdist$LTD objects within the ltdagg() output list object.

Value

An output list of 12 objects, of class ltdagg:

`hiclus`	Name of clustering object created by NUcluster().
`dframe`	Name of data.frame containing X, trtm & Y variables.
`trtm`	Name of treatment factor variable.
`yvar`	Name of outcome Y variable.
`K`	Number of Clusters Requested.
`actclust`	Number of Clusters delivered.
`LTDtabl`	data.frame with 5 columns and K rows for Clusters.
`LTDtabl$c`	Cluster ID Factor, "1", "2", ..., "K".
`LTDtabl$LTD`	Numerical value of Local Treatment Difference for a Cluster.
`LTDtabl$w`	Integer value of "weight" = Cluster Size.
`LTDtabl$LAO`	Numerical value of within-cluster Local Average Outcome (Y-value).

LTDtabl$PS	Numerical value of Propensity Score = Local Fraction of Experimental Units receiving trtm==1; 0.0 <= PS <= 1.0.
LTDdist	data.frame with 5 columns and same number of rows as the data: dframe.
LTDdist$c.K	Factor values within c("1", "2", ..., "K").
LTDdist$ID	Observation ID Variable for the rows of the input dframe.
LTDdist$y	Numerical value of the Y-Outcome for an Experimental Unit.
LTDdist$t	Numerical value of trtm (0 or 1) for an Experimental Unit.
LTDdist$LTD	Numerical value of the LTD for the Cluster containing each Exp. Unit.
infoclus	Integer value of Number of Informative Clusters.
infounits	Integer value of Number of Units within Informative Clusters.
LTDmean	Numerical value of mean(LTDdist$LTD) = Weighted Average of LTDtabl$LTD values.
LTDstde	Numerical value of sqrt(var(LTDdist$LTD)) = Weighted Standard Deviation of LTDtabl$LTD values.

Author(s)

Bob Obenchain <wizbob@att.net>

References

Obenchain RL. (2010) Local Control Approach using JMP. Chapter 7 of **Analysis of Observational Health Care Data using SAS**, *Cary, NC:SAS Press*, pages 151-192.

Obenchain RL. (2019) **NU.Learning_in_R.pdf** http://localcontrolstatistics.org

See Also

ivadj, lrcagg and NUcompare.

Examples

```
# Running takes more than 7 seconds...
data(pci15k)
xvars = c("stent", "height", "female", "diabetic", "acutemi", "ejfract", "ves1proc")
hclobj = NUcluster(pci15k, xvars)
NUe = NUsetup(hclobj, pci15k, thin, surv6mo)
surv050 = ltdagg(50, NUe)
surv050
plot(surv050, NUe)
```

mlme *Create a «Most-Like-Me» data.frame for a specified X-Confounder vector: xvec*

Description

For a Given X-confounder Vector (xvec), sort all experimental units (eUnits) in an ltdagg() or lrcagg() output object into the strictly non-decreasing order of their distances from this X-Vector, which defines the TARGET eUnit: "Me". Plots of mlme() objects and displays of mlme.stats() are then used to Visualize and Summarize "Mini-" « LOCAL effect-size Distributions » for different Numbers of "Nearest Neighbor" eUnits.

Usage

```
mlme(envir, hcl, NUagg, xvec )
```

Arguments

envir	Environment output by a call to the NUsetup() function.
hcl	Name of a NUcluster() output object created using a cluster::diana or stats::hclust method.
NUagg	A data.frame object output by ltdagg() or lrcagg() containing LOCAL effect-size Estimates for eUnits within Clusters defined in X-covariate space.
xvec	A suitable vector of the Numerical values for the X-Confounder variables, used in the current CLUSTERING, that define the eUnit: "Me".

Details

For example, in demo(radon), the eUnits are 2881 US "Counties", and the NUagg object is of type lrcagg() because radon exposure is a continuous variable. But, in demo(pci15k), the eUnits are 15487 "Patients," and the NUagg object is of type ltdagg() because treatment choice (thin) is Binary (0 = "No", 1 = "Yes").

Value

An output list object of class mlme:

xvec	The xvec vector input to mlme().
Type	Either "LTD" or "LRC".
xvars	Names of the X-Confounder variables specified in NUsetup().
varx	The vector of Variances of the xvars variabes, used in rescaling distances.
outdf	The output data.frame of sorted "Nearest Neighbor" candidate eUnits.

Author(s)

Bob Obenchain <wizbob@att.net>

References

Obenchain RL. NU.Learning-vignette. (2023) **NU.Learning_in_R.pdf** http://localcontrolstatistics.org

See Also

plot.mlme,print.mlme,mlme.stats

Examples

```
# Running takes about 7 seconds...
data(pci15k)
xvars = c("stent", "height", "female", "diabetic", "acutemi", "ejfract", "ves1proc")
hclobj = NUcluster(pci15k, xvars)
NU.env = NUsetup(hclobj, pci15k, thin, surv6mo)
surv0500 = ltdagg(500, NU.env)
xvec11870 = c( 0, 162, 1, 1, 0, 57, 1)
mlmeC5H = mlme(envir = NU.env, hcl = hclobj, NUagg = surv0500, xvec = xvec11870 )
plot(mlmeC5H) # using default "NN" and "breaks" settings...
```

---

mlme.stats *Print Summary Statistics for One or More "Most-Like-Me" Histogram Pairs.*

Description

Print Summary Statistics for Local effect-size (LTD or LRC) Distributions associated with given Numbers of "Nearest-Neighbors" in X-confounder Space.

Usage

```
mlme.stats(x, NN = 50, ...)
```

Arguments

x	An object output by mlme.data().
NN	Number(s) of "Nearest Neighbors" displayed in Histogram(s). NN can be either a single integer like NN = 40 or a combination of integers like NN = c(50, 250, 2500).
...	Other arguments passed on to print().

Value

NULL

Author(s)

Bob Obenchain <wizbob@att.net>

See Also

plot.mlme,print.mlme,mlme

NUcluster — *Hierarchical Clustering of experimental units (such as patients) in X-covariate Space*

Description

Form the full, hierarchical clustering tree (dendrogram) for all units (regardless of Treatment/Exposure status) using Mahalonobis distances computed from specified baseline X-covariate characteristics.

Usage

```
NUcluster(dframe, xvars, method="ward.D")
```

Arguments

dframe	Name of data.frame containing baseline X covariates.
xvars	List of names of X variable(s).
method	Hierarchical Clustering Method of "diana", "ward.D", "ward.D2", "complete", "average", "mcquitty", "median" or "centroid".

Details

The first step in applying NU.Learning to data is to hierarchically cluster experimental units in baseline X-covariate space ...thereby creating "Blocks" of relatively well-matched units. NUcluster first calls stats::prcomp() to calculate Mahalanobis distances using standardized and orthogonal Principal Coordinates. NUcluster then uses either the divisive cluster::diana() method or one of seven agglomerative methods from stats::hclust() to compute a dendrogram tree. The hclust function is based on Fortran code contributed to STATLIB by F. Murtagh.

Value

An output list object of class NUcluster, derived from cluster::diana or stats::hclust.

dframe	Name of data.frame containing all baseline X-covariates.
xvars	List of 1 or more X-variable names.
method	Hierarchical Clustering Method: "diana", "ward.D", "ward.D2", "complete", "average", "mcquitty", "median" or "centroid".
hclobj	Hierarchical clustering object created by the designated method.

Author(s)

Bob Obenchain <wizbob@att.net>

References

Kaufman L, Rousseeuw PJ. (1990) **Finding Groups in Data. An Introduction to Cluster Analysis**. New York: John Wiley and Sons.

Kereiakes DJ, Obenchain RL, Barber BL, et al. (2000) Abciximab provides cost effective survival advantage in high volume interventional practice. *Am Heart J* **140**: 603-610.

Murtagh F. (1985) Multidimensional Clustering Algorithms. **COMPSTAT Lectures** 4.

Obenchain RL. (2010) Local Control Approach using JMP. Chapter 7 of **Analysis of Observational Health Care Data using SAS**, *Cary, NC:SAS Press*, pages 151-192.

Rubin DB. (1980) Bias reduction using Mahalanobis metric matching. *Biometrics* **36**: 293-298.

See Also

NUsetup, ltdagg and lrcagg.

Examples

```
data(radon)
xvars  = c("obesity", "over65", "cursmoke")
hclobj = NUcluster(radon, xvars)  # ...using default method = "ward.D"
plot(hclobj)
```

NUcompare — *Display NU Sensitivity Graphic for help in choice of K = Number of Clusters*

Description

This function displays Box-Whisker diagrams that compare Treatment Effect-Size distributions for different values of K = Number of Clusters requested in X-covariate space. After an initial call to NUsetup(), the analyst typically makes multiple calls to either ltdagg() or lrcagg() for different values of K. The analyst then invokes NUcompare() to see how choice of K changes the location, spread and/or skewness of the distribution of Treatment Effect-Size estimates across Clusters. Variance-Bias trade-offs occur as K increases; large values of K may reduce Bias, but they definitely inflate the Variance of LTD and LRC distributions.

Usage

```
NUcompare(envir)
```

Arguments

envir	R environment output by an earlier call to NUsetup().

Details

The third phase of NU.Learning is called EXPLORE and uses graphical Sensitivity Analyses to show how Treatment Effect-Size distributions change with choice of NU parameter settings. Choice of K = Number of Clusters requested is guided, primarily, by NUcompare() graphics. Equally important are the analyst's choices of (i) which [and how many] of the available baseline X-covariates to "adjust for" and (ii) which clustering algorithm and dissimilarity metric to use. Unfortunately, changing these latter choices requires the analyst to essentially "start over" ...i.e. invoking NUcluster() with changed arguments, followed by an invocation of NUsetup() with a different 1st argument. To change only one's choice of y-Outcome variable and/or the Treatment/Exposure variable, a new NUsetup() invocation is all that is needed.

Value

NULL

Author(s)

Bob Obenchain <wizbob@att.net>

References

Obenchain RL. (2010) Local Control Approach using JMP. Chapter 7 of **Analysis of Observational Health Care Data using SAS**, *Cary, NC:SAS Press*, pages 151-192.

Obenchain RL. (2015) **NU_Confirm_Guidelines.pdf** http://localcontrolstatistics.org

Obenchain RL. (2023) **NU.Learning_in_R.pdf** http://localcontrolstatistics.org

Rubin DB. (1980) Bias reduction using Mahalanobis metric matching. *Biometrics* **36**: 293-298.

Tukey JW. (1977) **Exploratory Data Analysis**, *New York: Addison-Wesley*, Section 2C.

See Also

ltdagg, ivadj and lrcagg.

Examples

```
# Running takes more than 7 seconds...
data(pci15k)
xvars   = c("stent", "height", "female", "diabetic", "acutemi", "ejfract", "ves1proc")
hclobj  = NUcluster(pci15k, xvars)
NU.env  = NUsetup(hclobj, pci15k, thin, surv6mo)
surv050 = ltdagg( 50, NU.env)
surv100 = ltdagg(100, NU.env)
surv200 = ltdagg(200, NU.env)
NUcompare(NU.env)
```

NUsetup *Specify KEY parameters used in NU.Learning to "design" analyses of Observational Data.*

Description

Invoke NUsetup() to specify the name of the Hierarchical Clustering object output by NUcluster() and the name of the data.frame containing all desired X-covariates, the Treatment/Exposure variable and the Y-Outcome variable. It is ESSENTIAL to save the Environment output by NUsetup() as a named object within the user's .GlobalEnv space.

Usage

```
NUsetup(hclobj, dframe, trex, yvar)
```

Arguments

hclobj	Name of a NUcluster() output object created using a cluster::diana or stats::hclust method.
dframe	Name of the data.frame containing all X-covariates, the Treatment/Exposure variable and the Y-Outcome variable.
trex	Name of the numerical Treatment/Exposure variable.
yvar	Name of the numerical Y-Outcome variable.

Value

The environment output by NUsetup() must be saved to the user's .GlobalEnv space. It's contents will be automatically updated by calls to other NU.Learning functions:

aggdf	data.frame with 4 columns and 1 row for each call to ltdagg() or lrcagg().
aggdf$Label	Factor value of "LTD" or "LRC".
aggdf$Blocks	K = integer Number of Clusters requested.
aggdf$LTDmean or aggdf$LRCmean	numerical value of cluster mean of LTD or LRC estimates.
aggdf$LTDstde or aggdf$LRCstde	numerical value of the within-cluster standard deviation.
boxdf	data.frame of 2 variables ...for input to boxplot() by NUcompare().
boxdf$NUstat	LTD or LRC estimate for a single experimental unit from ltdagg() or lrcagg().
boxdf$K	Number of Cluters used in forming the LTD or LRC estimate for each Experimental Unit.
Kmax	Maximum Number of Clusters so that Average Size will be >= 12 experimental units.
LTDmax or LRCmax	Maximum Treatment Effect-Size estimate across Clusters.

LTDmin or LRCmin	Minimum Treatment Effect-Size estimate across Clusters.
NumLevels	Integer number of distinct Levels of the Treatment/Exposure variable: trex.
pars	Character data.frame with 4 columns and 1 row.
pars[1, 1]	Name of the diana or hclust object created by NUcluster().
pars[1, 2]	Name of data.frame containing the X, Treatment/Exposure and Y variables.
pars[1, 3]	Name of Treatment/Exposure variable within data.frame pars[1,2].
pars[1, 4]	Name of Y-outcome variable within data.frame pars[1,2].

Author(s)

Bob Obenchain <wizbob@att.net>

References

Obenchain RL. (2010) Local Control Approach using JMP. Chapter 7 of **Analysis of Observational Health Care Data using SAS**, *Cary, NC:SAS Press*, pages 151-192.

Obenchain RL. (2023) **NU.Learning_in_R.pdf** http://localcontrolstatistics.org

See Also

ltdagg, ivadj and lrcagg.

Examples

```
# Running takes about 7 seconds...
data(pci15k)
xvars  = c("stent", "height", "female", "diabetic", "acutemi", "ejfract", "ves1proc")
hclobj = NUcluster(pci15k, xvars)
NUe    = NUsetup(hclobj, pci15k, thin, surv6mo)
ls.str(NUe)
```

---

pci15k — *Six-month Survival, Cardiac cost and Baseline Covariate data for 15,487 PCI patients.*

Description

Using observational data on 996 patients who received a Percutaneous Coronary Intervention (PCI) at Ohio Heart Health, Lindner Center, Christ Hospital, Cincinnati (Kereiakes et al, 2000), we generated this much larger dataset via "plasmode simulation."

Usage

```
data(pci15k)
```

Format

A data frame of 11 variables on 15,487 patients; no NAs.

patid Patient ID number: 1 to 15487.

surv6mo Binary PCI Survival variable: 1 => Survival for at least 6 months following PCI, 0 => Survival for less than 6 months.

cardcost Cardiac related costs incurred within 6 months of patient's initial PCI; numeric value in 1998 dollars; costs were truncated by death for the 404 patients with surv6mo == 0.

thin Numeric treatment selection indicator: thin = 0 implies usual PCI care alone; thin = 1 implies usual PCI care augmented by either planned or rescue treatment with a new blood thinning agent.

stent Coronary stent deployment; numeric, with 1 meaning YES and 0 meaning NO.

height Height in centimeters; numeric integer from 133 to 198.

female Female gender; numeric, with 1 meaning YES and 0 meaning NO.

diabetic Diabetes mellitus diagnosis; numeric, with 1 meaning YES and 0 meaning NO.

acutemi Acute myocardial infarction within the previous 7 days; numeric, with 1 meaning YES and 0 meaning NO.

ejfract Left ejection fraction; numeric value from 17 percent to 77 percent.

ves1proc Number of vessels involved in the patient's initial PCI procedure; numeric integer from 0 to 5.

References

Kereiakes DJ, Obenchain RL, Barber BL, et al. Abciximab provides cost effective survival advantage in high volume interventional practice. *Am Heart J* 2000; **140**: 603-610.

Gadbury GL, Xiang Q, Yang L, Barnes S, Page GP, Allison DB. Evaluating Statistical Methods Using Plasmode Data Sets in the Age of Massive Public Databases: An Illustration Using False Discovery Rates. *PLOS Genetics* 2008; **4**: 1-8, e1000098 (Open Access).

Obenchain RL. (2023) **NU.Learning_in_R.pdf** http://localcontrolstatistics.org

Examples

```
data(pci15k)
str(pci15k)
```

`plot.ivadj` *Display an Instrumental Variable (LAO) plot with Linear and smooth.spline Fits*

Description

For a given number of patient clusters, K, in baseline X-covariate space and a specified Y-outcome variable, display the distribution of Local Average Outcomes (LAOs) plotted versus Within-Cluster Propensity-like Scores: Treatment Selection Fractions or Relative Exposure Levels.

Usage

```
  ## S3 method for class 'ivadj'
plot(x, maxsiz = 0.15, ...)
```

Arguments

x	An object output by ivadj() for K given Clusters in baseline X-covariate space.
maxsiz	Radius of the Circle plotting symbol for the largest Cluster. Usually < 0.6
...	Other arguments passed on to plot().

Value

NULL

Author(s)

Bob Obenchain <wizbob@att.net>

See Also

plot.ltdagg, plot.lrcagg

---

plot.lrcagg *Display Visualizations of an Observed LRC Distribution in NU.Learning*

Description

Display a Histogram, Box-Whisker Diagram and/or empirical Cumulative Distribution Function depicting the Observed Local Rank Correlation (LRC) Distribution across K Clusters.

Usage

```
  ## S3 method for class 'lrcagg'
plot(x, envir, show="all", breaks="Sturges", ...)
```

Arguments

x	An object output by lrcagg() for K = Number of Clusters in baseline X-covariate space.
envir	R environment output by a previous call to NUsetup().
show	Choice of "all", "seq", "hist", "boxp", or "ecdf".
breaks	Parameter setting for hist(); May be an integer value ...like 25 or 50.
...	Other arguments passed on to plot().

Value

NULL

Author(s)

Bob Obenchain <wizbob@att.net>

See Also

plot.ltdagg

---

plot.ltdagg *Display Visualizations of an Observed LTD Distribution in NU.Learning*

Description

Display a Histogram, Box-Whisker Diagram and/or empirical Cumulative Distribution Function depicting the Observed Local Treatment Difference (LTD) Distribution across K Clusters.

Usage

```
  ## S3 method for class 'ltdagg'
plot(x, envir, show="all", breaks="Sturges", ...)
```

Arguments

x	An object output by ltdagg() for K = Number of Clusters in baseline X-covariate space.
envir	R environment output by a previous call to NUsetup().
show	Choice of "all", "seq", "hist", "boxp", or "ecdf".
breaks	Parameter setting for hist(); May be an integer value ...like 25 or 50.
...	Other arguments passed on to plot().

Value

NULL

Author(s)

Bob Obenchain <wizbob@att.net>

See Also

plot.lrcagg

`plot.mlme` *Display a Pair (or Pairs) of Histograms showing LOCAL effect-sizes for Patients "Most-Like-Me".*

Description

Display Pair(s) of Histograms of Local effect-size (LTD or LRC) Distributions for a specified Number (or combinations of Numbers) of "Nearest-Neighbors in X-confounder Space.

Usage

```
  ## S3 method for class 'mlme'
plot(x, NN=50, breaks=50, ...)
```

Arguments

`x`	An object output by mlme().
`NN`	Number(s) of Nearest Neighbors displayed in Bottom Histogram(s). NN can be a single integer like NN = 40 or a combination of integers like NN = c(50, 250, 2500).
`breaks`	Integer number of breaks in the Top Histogram for the full LTD or LRC distribution. Because the Bottom Histogram may include only a few Nearest Neighbors, it is always displayed using breaks = "Sturges".
`...`	Other arguments passed on to plot().

Value

NULL

Author(s)

Bob Obenchain <wizbob@att.net>

See Also

`mlme.stats`,`print.mlme`,`mlme`

pmdata *Particulate Matter, Mortality and Other data for 2980 US Counties*

Description

This data.frame combines 122 variables from the 5 sources referenced below. Several PM variables appear to be predictions from EPA "CMAQ" models rather than values from validated measuring instruments. NU.Learning concepts are illustrated in demo(pmdata) using Clustering of 2973 Counties and Parishes within the contiguous 48 US States and Washington, D.C.

Usage

```
data(pmdata)
```

Format

This data.frame contains 122 variables for 2,980 US counties. A total of 738 "NA"s imply that only about two tenths of one percent of these 363,560 values are missing.

fips Federal Information Processing Standard code; 4 or 5 digits; 2980 unique values

C50 Cluster ID Number between 1 and 50. Total of 50 unique values

LRC50 Local (Spearman) Rank Correlation between Bvoc and AACRmort within Cluster

County County or Parish name is a Factor variable (character code)

State State name is a 2-Character Factor code; 49 unique levels

Deaths CDC: Total number of Deaths in the County in 2016

Population CDC: Total population of County in 2016

CRmort CDC: Crude Rate of Circulatory-Respiratory Mortality for the County in 2016

CrudeL95 CDC: Lower 95% confidence limit for Crude Rate of CR Mortality

CrudeU95 CDC: Upper 95% confidence limit for Crude Rate of CR Mortality

CrudeSE CDC: Standard Error for Crude Rate of CR Mortality

AACRmort CDC: Age Adjusted Rate of Circulatory-Respiratory Mortality in 2016

AACRL95 CDC: Lower 95% confidence limit for Age Adjusted Rate of CR Mortality

AACRU95 CDC: Upper 95% confidence limit for Age Adjusted Rate of CR Mortality

AACR_SE CDC: Standard Error for Age Adjusted Rate of CR Mortality

TotDeathPct CDC: Total Death Percentage - Circulatory-Respiratory Mortality in 2016

lat EPA: County Latitude used in EPA "CMAQ" model calculations

long EPA: County Longitude used in EPA "CMAQ" model calculations

RHpct EPA: Relative Humidity Percentage for 2016

SFCtmpC EPA: Surface Temperature in Degrees Centigrade for 2016

NO2 EPA: Nitrogen Dioxide level (NO2.ppbV) for 2016

O3 EPA: Ozone level (O3.ppbV) for 2016

pmCL EPA: Chlorine level in Particulate Matter (PM25_CL.ugm3) for 2016

pmEC EPA: Ethylene Carbonate level in Particulate Matter (PM25_EC.ugm3) for 2016

pmNA EPA: Sodium level in Particulate Matter (PM25_NA.ugm3) for 2016

pmMG EPA: Magnesium level in Particulate Matter (PM25_MG.ugm3) for 2016

pmK EPA: Potassium level in Particulate Matter (PM25_K.ugm3) for 2016

pmCA EPA: Calcium level in Particulate Matter (PM25_CA.ugm3) for 2016

pmNH4 EPA: Ammonium level in Particulate Matter (PM25_NH4.ugm3) for 2016

pmNO3 EPA: Nitrate level in Particulate Matter (PM25_NO3.ugm3) for 2016

pmOC EPA: Organic Compounds in pmTOT [fine particulate matter] (PM25_OC.ugm3) for 2016

pmOM EPA: OM compounds in pmTOT (PM25_OM.ugm3) for 2016

pmOTHR EPA: Other Compounds in pmTOT (PM25_OTHR.ugm3) for 2016

pmSO4 EPA: Sulfate Compounds in pmTOT (PM25_SO4.ugm3) for 2016

pmFE EPA: Ferrous Compounds in pmTOT (PM25_FE.ugm3) for 2016

pmSI EPA: Silicon Compounds in pmTOT (PM25_SI.ugm3) for 2016

pmTI EPA: Titanium Compounds in pmTOT (PM25_TI.ugm3) for 2016

pmMN EPA: Manganese Compounds in pmTOT (PM25_MN.ugm3) for 2016

pmAL EPA: Aluminum Compounds in pmTOT (PM25_AL.ugm3) for 2016

pmUNSPCRS EPA: UNSPCRS Compounds in pmTOT (PM25_UNSPCRS.ugm3) for 2016

pmPOA EPA: Primary Organic Aerosols in pmTOT (PM25_POA.ugm3) for 2016

pmSOA EPA: Secondary Organic Aerosols in pmTOT (PM25_SOA.ugm3) for 2016; pmSOA = Avoc+Bvoc

pmGLY EPA: Glycemic Secondary Organic Aerosols in pmTOT (PM25_GLYSOA.ugm3) for 2016

pmOLGB EPA: OLGB compounds in pmTOT (PM25_OLGB.ugm3) for 2016

pmISOP EPA: ISOP compounds in pmTOT (PM25_ISOP.ugm3) for 2016

pmEPOX EPA: EPOX compounds in pmTOT (PM25_EPOX.ugm3) for 2016

pmSQT EPA: SQT compounds in pmTOT (PM25_SQT.ugm3) for 2016

pmMTN EPA: MTN compounds in pmTOT (PM25_MTN.ugm3) for 2016

pmMT EPA: MT compounds in pmTOT (PM25_MT.ugm3) for 2016

pmTOT EPA: Total (fine) Particulate Matter (PM25_TOT.ugm3) for 2016

SFCtmpK EPA: Surface Temperature in Degrees Kelvin for 2016

CardioRes CDC: Cardio Respiratory Rate (rate I00J98.per100000.cdc) for 2016

POPcdc CDC: County Population (population.cdc) for 2016

POP5yracs CDC: 5yracs Population (population.people.5yracs) for 2016

PREMdeath Premature Deaths per 100K Residents ...UWPHI for 2018

POFHealth Poor or Fair Health rate (Poor.or.fair.health) ...UWPHI for 2018

PPHdays Poor Physical Health days (Poor.physical.health.days) ...UWPHI for 2018

PMHdays Poor Mental Health days (Poor.mental.health.days) ...UWPHI for 2018

LBW Low Birth Weight rate (Low.birthweight) ...UWPHI for 2018

ASmoke Adult Smoking Percentage (Adult.smoking) ...UWPHI for 2018

AObes Adult Obesity Percentage (Adult.obesity) ...UWPHI for 2018

FEnv Food Environment Index (Food.environment.index) ...UWPHI for 2018

PhysInAct Physical Inactivity (Physical.inactivity) ...UWPHI for 2018

ExercOPS Access to Exercise Opportunities ...UWPHI for 2018

ExsDrink Excessive Drinking Rate (Excessive.drinking) ...UWPHI for 2018

AIDrivD Alcohol Impaired Driving Deaths ...UWPHI for 2018

STInfect Sexually Transmitted Infections ...UWPHI for 2018

TBirths Teenage Births (Teen.births) ...UWPHI for 2018

Uninsur Uninsured Residences (Uninsured) ...UWPHI for 2018

PCDocs Primary Care Physicians (Primary.care.physicians) ...UWPHI for 2018

Dentists Dentists (Dentists) ...UWPHI for 2018

PrevntHS Preventable Hospital Stays (Preventable.hospital.stays) ...UWPHI for 2018

DiabMNT Diabetes Monitoring (Diabetes.monitoring) ...UWPHI for 2018

MammoSC Mammography Screening (Mammography.screening) ...UWPHI for 2018

SomCOL Some College Education (Some.college) ...UWPHI for 2018

UnEMP Unemployment Rate (Unemployment) ...UWPHI for 2018

ChildPOV Children Living in Poverty (Children.in.poverty) ...UWPHI for 2018

IncomIEQ Income Inequality (Income.inequality) ...UWPHI for 2018

ChildSPH Children In Single-Parent Households ...UWPHI for 2018

SocASOC Social Associations (Social.associations) ...UWPHI for 2018

VioCRM Violent Crime Rate (Violent.crime) ...UWPHI for 2018

InjyDths Injury Death Rate (Injury.deaths) ...UWPHI for 2018

AirPolpm Air Pollution Particulate Matter ...UWPHI for 2018

DrnkWtVi Drinking Water Violations (Drinking.water.violations) ...UWPHI for 2018

SevrHOUS Severe Housing Problems (Severe.housing.problems) ...UWPHI for 2018

DrivATW Driving Alone to Work (Driving.alone.to.work) ...UWPHI for 2018

LComutA Long Commute - Driving Alone to Work ...UWPHI for 2018

PAAM Premature Age Adjusted Mortality ...UWPHI for 2018

FrqPhysD Frequent Physical Distress ...UWPHI for 2018

FrqMentD Frequent Mental Distress ...UWPHI for 2018

DiabPrev Diabetes Prevalence (Diabetes.prevalence) ...UWPHI for 2018

FoodInSec Food Insecurity (Food.insecurity) ...UWPHI for 2018

LimAHFood Limited Access to Healthy Foods ...UWPHI for 2018

DrugOdDM Drug Overdose Deaths Model predictions ...UWPHI for 2018

InsufSlp Insufficient Sleep (Insufficient.sleep) ...UWPHI for 2018

UnInsAds Uninsured Adults (Uninsured.adults) ...UWPHI for 2018

UnInsCls Uninsured Children (Uninsured.children) ...UWPHI for 2018

HCareCost Health Care Costs (Health.care.costs) ...UWPHI for 2018

OthPrimCP Other Primary Care Providers ...UWPHI for 2018

MHHIncome Median Household Income ...UWPHI for 2018

ChildFRPL Children Eligible for Free or Reduced-Price Lunch ...UWPHI for 2018

Population County Population (Population) ...UWPHI for 2018

AGELess18 Residents below 18 Years of Age ...UWPHI for 2018

A65oOVR Residents 65 or Older (X..65.and.older) ...UWPHI for 2018

NHispAfA Non-Hispanic African-American Residents ...UWPHI for 2018

AmINalsN American Indian or Alaskan Natives ...UWPHI for 2018

Asian Asian Residents (X..Asian) ...UWPHI for 2018

NHawOPI Native Hawaiian and Other Pacific Islanders ...UWPHI for 2018

Hispanic Hispanic Residents (X..Hispanic) ...UWPHI for 2018

NHispWht Non-Hispanic White Residents ...UWPHI for 2018

LoProEngl Low Proficiency in English (not.proficient.in.English) ...UWPHI for 2018

Females Female Residents ...UWPHI for 2018

Rural Rural Residents ...UWPHI for 2018

pmOA EPA: Organic Aerosols in pmTOT (PM25_OA.ugm3) for 2016

Avoc EPA: Anthroprogenic [man-made] Volatile Organic Compounds in pmTOT for 2016

pmSEAspry EPA: Sea Spray components in pmTOT (PM25_SOAAVOC.ugm3) for 2016

pmDUST EPA: Dust components in pmTOT (PM25_DUST.ugm3) for 2016

pmNH4NO3 EPA: Ammonium Nitrate components in pmTOT (PM25_NH4NO3.ugm3) for 2016

pmSOOT EPA: Soot components in pmTOT (PM25_SOOT.ugm3) for 2016

isop EPA: SOA Isoprenes (PM25_SOAISOPRENE.ugm3) for 2016

terp EPA: SOA Terpenes (PM25_SOATERPENE.ugm3) for 2016

Bvoc EPA: Biogenic (natural) Volatile Organic Compounds for 2016; Bvoc = isop + terp

References

Obenchain RL. and Young SS. (2022), EPA Particulate Matter Data - Analyses using Local Control Strategy. (24 pages, 22 figures) https://doi.org/10.48550/arXiv.2209.05461

Pye, H., Ward-Caviness, C., Murphy, B., Appel, K., and Seltzer, K. (2021). Secondary organic aerosol association with cardiorespiratory disease mortality in the united states. Nature Communications, 12.7215 https://doi.org/10.1038/s41467-021-27484-1

Pye, H. [EPA] (2021), Data For Secondary Organic Aerosol and Cardiorespiratory Disease Mortality. https://doi.org/10.5281/zenodo.5713903

University of Wisconsin, Population Health Institute. https://uwphi.pophealth.wisc.edu [UWPHI] UWPHI@med.wisc.edu

Young SS. and Obenchain RL. (2022), "EPA particulate matter data...Analyses using Local Control Strategy" https://doi.org/10.5061/dryad.63xsj3v58

Examples

```
data(pmdata)
str(pmdata, list.len=122)
```

`print.mlme` *Print Summary Statistics on Local effect-size Estimates for Patients "Most-Like-Me".*

Description

Display "Most-Like-Me" Summary Statistics for LOCAL effect-size (LTD or LRC) Distributions of "Nearest-Neighbors" in X-confounder Space.

Usage

```
  ## S3 method for class 'mlme'
print(x, ...)
```

Arguments

`x`	An object output by mlme().
`...`	Other arguments passed on to print().

Value

NULL

Author(s)

Bob Obenchain <wizbob@att.net>

See Also

`mlme.stats`,`plot.mlme`,`mlme`

radon *Radon exposure and lung cancer mortality data for 2,881 US counties in 46 States.*

Description

Federal EPA and state government agencies have been reporting observational data at the US County level since about 1980. The data given here include 5 potential X-confounder variables of the relationship between lung cancer mortality and radon exposure; they were amassed and checked by Goran Krstic, Fraser Health Authority, Vancouver, BC, Canada.

Usage

```
data(radon)
```

Format

A data frame of 11 variables for 2881 US counties. One Missing Value; row 778 for Shannon County, SD, fips == 46113, has hhincome == NA.

fips County FIPS code. Codes are 4 or 5 digit integers; 2881 unique values.

state State Factor variable (2-character codes); 46 unique levels.

county County or Parish Factor variable (character codes); 1703 unique levels.

lcanmort Lung Cancer Mortality rate (deaths per 100,000 person-years), 1980-2004.

radon County Radon Exposure level in picocuries per liter (pCi/L) for some unspecified period within 1986-1992; rounded to nearest single decimal place.

lnradon Natural logarithm of County Radon Exposure level. Radon levels reported as 0.0 for 10 US counties are Windsorized here to ln(0.05), which is roughly -3.

obesity Percentage of County Residents considered Obese (age adjusted), 2008.

over65 Percentage of County Residents of Age 65 and over, 2000 Census.

cursmoke Percentage of County Residents who Currently Smoke, 1997-2003.

evrsmoke Percentage of County Residents who Ever Smoked, 1997-2003.

hhincome Average Median HouseHold Income in Thousands ($1,000), 1989-2004.

References

Krstic G, Obenchain RL. (2016) Radon dataset documentation and downloads. http://localcontrolstatistics.org

Obenchain RL. (2018) **RADON_short.pdf** http://localcontrolstatistics.org 40 PPT Slides and Commentary in Notes Pages format.

Examples

```
data(radon)
str(radon)
```

`reveal.data` *Create a data.frame for use in Prediction of a LTD/LRC effect-size Distribution*

Description

reveal.data() forms a data.frame by sorting and appending the LTD or LRC exposure effect-size measures from ltdagg() or lrcagg() – as well as a Cluster membership-number variable – to a copy of the data.frame specified in NUsetup(). In the fourth and final REVEAL Phase of NU.Learning, a stretch-goal is to predict variation in LTD/LRC effect-size distributions using the known (baseline) X-covariate characteristics of experimental units. For example, the data.frame output by reveal.data() is suitable for input to party::ctree() as well as to a number of other "less Visual" prediction methods available in **R**.

Usage

```
reveal.data(x, clus.var="Clus", effe.var="eSiz")
```

Arguments

`x`	An output object resulting from a call to ltdagg() or lrcagg().
`clus.var`	Quoted NAME for the Cluster-Number variable.
`effe.var`	Quoted NAME for the LTD/LRC effect-size variable.

Value

The desired data.frame:

`outdf`	A data.frame containing clus.var, effe.var plus (X, trex & Y) variables.

Author(s)

Bob Obenchain <wizbob@att.net>

References

Obenchain RL. (2023) **NU.Learning_in_R.pdf** http://localcontrolstatistics.org

See Also

`ltdagg`, `lrcagg`, and `NUsetup`.

Index