Package 'DOS'

October 12, 2022

Type Package

Title Design of Observational Studies

Version 1.0.0

Author Paul R. Rosenbaum

Maintainer Paul Rosenbaum <rosenbaum@wharton.upenn.edu>

Description Contains data sets, examples and software from the book Design of Observational Studies by Paul R. Rosenbaum, New York: Springer, <doi:10.1007/978-1-4419-1213-8>, ISBN 978-1-4419-1212-1.

Imports stats, MASS

Suggests optmatch

License GPL-2

Encoding UTF-8

LazyData true

NeedsCompilation no

Repository CRAN

Date/Publication 2018-08-31 19:20:03 UTC

R topics documented:

---

addalmostexact *Use a Penalty to Obtain a Near Exact Match*

Description

A near-exact match (also known as an almost exact match) for a nominal covariate maximizes the number of pairs that are exactly matched for the nominal covariate. This is done by adding a large penalty to the covariate distance between individuals with different values of the nominal covariate. The topic is discussed in Section 9.2 of Design of Observational Studies (2010).

Usage

```
addalmostexact(dmat, z, f, mult = 10)
```

Arguments

dmat	A distance matrix with one row for each treated individual and one column for each control. Often, this is either the Mahalanobis distance based on covariates, mahal(), or else a robust variant produced by smahal().
z	z is a vector that is 1 for a treated individual and 0 for a control. length(z) must equal sum(dim(dmat)).
f	A vector or factor giving the levels of the nominal covariate. length(f) must equal length(z).
mult	Let mx be the largest distance in dmat. Then mult*mx is added to each distance in dmat for which the treated and control individuals have different values of f. For large enough values of mult, this will ensure that a minimum distance match maximizes the number of individuals who are exactly matched for f.

Details

Very large values of mx may result in numerical instability or slow computations.

A small value of mx, say 1 or 1/10, may not maximize the number of exactly matched pairs, but it may usefully increase the number of exactly matched pairs.

Value

Returns the penalized distance matrix.

Author(s)

Paul R. Rosenbaum

References

Rosenbaum, P. R. (2010). Design of Observational Studies. New York: Springer. The method and example are discussed in Section 9.2.

Zubizarreta, J. R., Reinke, C. E., Kelz, R. R., Silber, J. H. and Rosenbaum, P. R. (2011). Matching for several sparse nominal variables in a case control study of readmission following surgery. The American Statistician, 65(4), 229-238. Combines near exact matching with fine balance for the same covariate.

Examples

```
data(costa)
z<-1*(costa$welder=="Y")
aa<-1*(costa$race=="A")
smoker=1*(costa$smoker=="Y")
age<-costa$age
x<-cbind(age,aa,smoker)
dmat<-mahal(z,x)
# Mahalanobis distances
round(dmat[,1:6],2)
# Mahalanobis distanced penalized for mismatching on smoking.
dmat<-addalmostexact(dmat, z, smoker, mult = 10)
# The first treated subject (row labeled 27) is a nonsmoker, but the
# third control (column 3) is a smoker, so there is a big penalty.
round(dmat[,1:6],2)
```

---

addcaliper   *Implement a Caliper Using a Penalty Function in Optimal Matching*

Description

A caliper forces a close match on a continuous covariate, often the propensity score (Rosenbaum and Rubin 1985). Alternatively or additionally, a caliper may be imposed on a prognostic score estimated from some other independent data set (Hansen 2008). Rosenbaum and Rubin (1985) suggested minimizing the Mahalanobis distance within calipers defined by the propensity score. If the caliper is implemented using a penalty function, the caliper cannot create infeasibility. Implementation of a caliper using a penalty function is discussed in Chapter 8 of Design of Observational Studies (2010).

Usage

```
addcaliper(dmat, z, p, caliper = 0.1, penalty = 1000)
```

Arguments

dmat   A distance matrix with one row for each treated individual and one column for each control. Often, this is either the Mahalanobis distance based on covariates, mahal(), or else a robust variant produced by smahal().

z   z is a vector that is 1 for a treated individual and 0 for a control. The number
                                        of treated subjects, sum(z), must equal the number of rows of dmat, and the
                                        number of potential controls, sum(1-z), must equal the number of columns of
                                        dmat.

p   A vector covariate to which the caliper will be applied. Often, p is the propensity
                                        score. length(p) must equal length(z).

caliper   A positive number. A penalty will be added to row i and column j of the distance
                                        matrix dmat if treated unit i and control unit j have values of p that differ in ab-
                                        solute value by more than caliper*sd(p), the default being a tenth of the standard
                                        deviation of p. This is different from the code in Section 13.11 of Rosenbaum
                                        (2010), where caliper defaulted to 0.2, rather than 0.1.

penalty   A positive number. penalty determines the magnitude of the penalty that is
                                        added to the distance matrix when the caliper is violated. Let pt and pc be the
                                        values of p for a treated individual and a control. If |pt-pc| <= caliper*sd(p), then
                                        nothing is added to dmat for these individuals. Otherwise, define gap = |pt-pc|-
                                        caliper*sd(p); then, the distance between these individuals in dmat is increased
                                        by gap*penalty. So there is no penalty inside the caliper, but dmat is increased
                                        rapidly as violation of the caliper increases in magnitude. A reasonable value of
                                        penalty will depend upon the units in which p is measured, and the default values
                                        are reasonable starting points for the propensity score expressed as a probability.

Details

By calling addcaliper several times, calipers may be imposed on several variables, say a propensity
score and age. If you use several penalties, say from addcaliper, addalmostexact, and fine, then you
need to pay some attention to their relative magnitudes to ensure that they express your view of the
relative importance of the conditions they impose.

Value

Returns the penalized distance matrix.

Author(s)

Paul R. Rosenbaum

References

Hansen, B. B. and Klopfer, S. O. (2006). Optimal full matching and related designs via network
flows. Journal of computational and Graphical Statistics, 15(3), 609-627. (optmatch package)

Hansen, B. B. (2007). Flexible, optimal matching for observational studies. R News, 7, 18-24.
(optmatch package)

Hansen, B. B. (2008). The prognostic analogue of the propensity score. Biometrika, 95(2), 481-
488.

Rosenbaum, P. R. (2010). Design of Observational Studies. New York: Springer. The method and
example are discussed in Section 9.2.

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched
sampling methods that incorporate the propensity score. The American Statistician, 39, 33-38.

Examples

```
data(costa)
z<-1*(costa$welder=="Y")
aa<-1*(costa$race=="A")
smoker=1*(costa$smoker=="Y")
age<-costa$age
x<-cbind(age,aa,smoker)
dmat<-mahal(z,x)
# Mahalanobis distances
round(dmat[,1:6],2) # Compare with Table 8.5 in Design of Observational Studies (2010)
# Impose propensity score calipers
prop<-glm(z~age+aa+smoker,family=binomial)$fitted.values # propensity score
# Mahalanobis distanced penalized for violations of a propensity score caliper.
# This version is used for numerical work.
dmat<-addcaliper(dmat,z,prop,caliper=.5)
round(dmat[,1:6],2) # Compare with Table 8.5 in Design of Observational Studies (2010)
## Not run:
# Find the minimum distance match within propensity score calipers.
optmatch::pairmatch(dmat,data=costa)

## End(Not run)
# Conceptual versions with infinite distances for violations of propensity caliper.
dmat[dmat>20]<-Inf
round(dmat[,1:6],2) # Compare with Table 8.5 in Design of Observational Studies (2010)
```

---

angristlavy	*Class Size and Academic Performance – Maimonidies Rule*

Description

This data set is from Angrist and Lavy (1999). There are 86 pairs of two Israeli schools, one with slightly more than 40 students in the fifth grade, the other with 40 or fewer in the fifth grade, together with test scores in reading and math. This example is discussed in Chapters 1 and 5 of Design of Observational Studies (2010).

Usage

```
data("angristlavy")
```

Format

A data frame with 172 observations on the following 9 variables.

scode School ID

numclass Number of classes in the fifth grade, 1 or 2.

cohsize Total number of students in the fifth grade, near 40 for these schools.

avgmath Average grade in math in the fifth grade.

avgverb Average verbal grade in the fifth grade.

`tipuach` Percent of disadvantaged students. Used to form matched pairs.

`clasz` Average class size in the fifth grade, equal to cohsize/numclass

`z` 1 if cohsize<=40, 0 if cohsize>40.

`pair` pair ID, 1, 2, ..., 86

Details

As discussed by Angrist and Lavy (1999), Maimonidies rule requires that a class of more than 40 be divided to form two or more classes of at most 40, so there is a large discontinuity in class size at 40: at 40 students in the 5th grade, there is one class of 40, but at 41 students, there are two classes of average size 20.5. So the enrolement of one student should cut the class size roughly in half. Adherence to Maimonidies rule is good but not perfect. Pairs of schools were matched for the percent of disadvantaged students (tipuach).

References

Angrist, J. D. and Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. The Quarterly Journal of Economics, 114, 533-575.

Angrist, J. D. and Krueger, A. B. (1999). Empirical strategies in labor economics. In Handbook of Labor Economics (Vol. 3, pp. 1277-1366). Elsevier.

Rosenbaum, P. R. (2010). Design of Observational Studies. New York: Springer. This example is discussed in Chapters 1 and 5.

Examples

```
# Figure 1.1 in Chapter 1 of Design of Observational Studies (2010)
data(angristlavy)
attach(angristlavy)
grp<-factor(z,levels=c(1,0),labels=c("31-40","41-50"),ordered=TRUE)
par(mfrow=c(2,2))
boxplot(tipuach~grp,main="Disadvantaged",ylab="Percent")
boxplot(clasz~grp,main="Class Size",ylab="Students")
boxplot(avgmath~grp,main="Math",ylab="Average Score")
boxplot(avgverb~grp,main="Verbal",ylab="Average Score")
detach(angristlavy)
```

---

costa   *Welding and DNA-Protein Crosslinks*

Description

This data set is from Costa et al. (1993) and it describes 21 welders and 26 potential controls. All are men. The outcome is a measure of genetic damage; specifically, dpc is a measure of DNA-protein cross-links. There are 3 covariates, age, race and smoking. This tiny example is used to illustrate the concepts of multivariate matching in Chapter 8 of Design of Observational Studies. The example is useful because its tiny size permits close inspection of the details of multivariate matching, but its small sample size and limited number of covariates make it highly atypical of matching in observational studies.

Usage

```
data("costa")
```

Format

A data frame with 47 observations on the following 6 variables.

subject Within group ID number.

age Age in years.

race AA=African-American, C=Caucasian

smoker Y=yes, N=no

welder Y=yes/treated, N=no/control

dpc DNA-Protein Cross-links (percent)

Source

The data are from Costa et al. (1993). The data are used as a tiny example in Chapter 8 of Design of Observational Studies.

References

Costa, M., Zhitkovich, A. and Toniolo, P. (1993). DNA-protein cross-links in welders: molecular implications. Cancer research, 53(3), 460-463.

Rosenbaum, P. R. (2010). Design of Observational Studies. New York: Springer. This example is discussed in Chapter 8.

Examples

```
data(costa)
boxplot(costa$dpc~costa$welder,
  xlab="Control (N) or Welder (Y)",
  ylab="DNA-Protein Cross-links Percent")
```

---

dynarski *A Natural Experiment Concerning Subsidized College Education*

Description

The data are from Susan Dynarski (2003)'s study of the effects of a subsidy for college education provided Social Security to children whose fathers had died. The subsidy was eliminated in 1982. As presented here, the data are only a portion of the data used in Dynarski (2003), specifically the period from 1979-1981 when the subsidy was available. The data set here is Table 13.1 of Design of Observational Studies (1st edition), where it was used with the limited goal of illustrating matching techniques.

Usage

```
data("dynarski")
```

Format

A data frame with 2820 observations on the following 10 variables.

id identification number

zb treatment indicator, zb=1 if 1979-1981 with father deceased, or zb=0 if 1979-1981 with father alive

faminc family income, in units of tens of thousands of dollars

incmiss indicator for missing family income

black 1 if black, 0 otherwise

hisp 1 if hispanic, 0 otherwise

afqtpct test score percentile, Armed Forces Qualifications Test

edmissm indicator for missing education of mother

edm education of mother, 1 is <high school, 2 is high school, 3 is some college, 4 is a BA degree or higher

female 1 if female, 0 if male

Details

The examples reproduce steps in Chapter 13 of Rosenbaum (2010). Portions of the examples require you to load Ben Hansen's optmatch package and accept its academic license; these portions of the examples do not run automatically. Hansen's optmatch package uses the Fortran code of Bertsekas and Tseng (1988).

Source

Dynarski (2003).

References

Bertsekas, D. P. and Tseng, P. (1988). The relax codes for linear minimum cost network flow problems. Annals of Operations Research, 13(1), 125-190.

Dynarski, S. M. (2003). Does aid matter? Measuring the effect of student aid on college attendance and completion. American Economic Review, 93(1), 279-288.

Hansen, B. B. (2007). Flexible, optimal matching for observational studies. R News, 7, 18-24.

Hansen, B. B. and Klopfer, S. O. (2006). Optimal full matching and related designs via network flows. Journal of Computational and Graphical Statistics, 15(3), 609-627.

Rosenbaum, P. R. (2010). Design of Observational Studies. New York: Springer.

Examples

```
#
data(dynarski)
# Table 13.1 of Rosenbaum (2010)
head(dynarski)
# Table 13.2 of Rosenbaum (2010)
zb<-dynarski$zb
zbf<-factor(zb,levels=c(1,0),labels=c("Father deceased","Father not deceased"))
table(zbf)
Xb<-dynarski[,3:10]

# Estimate the propensity score, Rosenbaum (2010, Section 13.3)
p<-glm(zb~Xb$faminc+Xb$incmiss+Xb$black+Xb$hisp
    +Xb$afqtpct+Xb$edmissm+Xb$edm+Xb$female,
    family=binomial)$fitted.values
# Figure 13.1 in Rosenbaum (2010)
boxplot(p~zbf,ylab="Propensity score",main="1979-1981 Cohort")

# Read about missing covariate values in section 13.4 of Rosenbaum (2010)

# Robust Mahalanobis distance matrix, treated x control
dmat<-smahal(zb,Xb)
dim(dmat)
# Table 13.3 in Rosenbaum (2010)
round(dmat[1:5,1:5],2)

# Add a caliper on the propensity score using a penalty function
dmat<-addcaliper(dmat,zb,p,caliper=.2)
dim(dmat)
# Table 13.4 in Rosenbaum (2010)
round(dmat[1:5,1:5],2)
## Not run:
# YOU MUST LOAD the optmatch package and accept its license to continue.
# Note that the optmatch package has changed since 2010.  It now suggests
# that you indicate the data explicitly as data=dynarski.

# Creating a 1-to-10 match, as in section 13.6 of Rosenbaum (2010)
# This may take a few minutes.
m<-fullmatch(dmat,data=dynarski,min.controls = 10,max.controls = 10,omit.fraction = 1379/2689)
length(m)
sum(matched(m))
1441/11 # There are 131 matched sets, 1 treated, 10 controls

# Alternative, simpler code to do the same thing
m2<-pairmatch(dmat,controls=10,data=dynarski)
# Results are the same:
sum(m[matched(m)]!=m2[matched(m2)])

# Housekeeping
im<-as.integer(m)
dynarski<-cbind(dynarski,im)
dm<-dynarski[matched(m),]
```

```
dm<-dm[order(dm$im,1-dm$zb),]

# Table 13.5 in Rosenbaum (2010)
which(dm$id==10)
dm[188:198,]
which(dm$id==396)
dm[23:33,]
which(dm$id==3051)
dm[1068:1078,]
# In principle, there can be a tie, in which several different
# matched samples all minimize the total distance.  On my
# computer, this calculation reproduces Table 13.5, but were
# there a tie, optmatch should return one of the tied optimal
# matches, but not any particular one.

## End(Not run)
```

---

fine   *Expand a Distance Matrix for Matching with Fine Balance.*

Description

In optimal pair matching with fine balance, expand a distance matrix to become a square matrix to enforce fine balance. The method is discussed in Chapter 10 of Design of Observational Studies (2010), and it is conceptually the simplest way to implement fine balance; therefore, it remains very useful for teaching and for self-study. See details. For practical work, consider also the rcbalance package, the designmatch package and the bigmatch package; see the references.

Usage

```
fine(dmat, z, f, mult = 100)
```

Arguments

dmat   A distance matrix with one row for each treated individual and one column for
            each control. Often, this is either the Mahalanobis distance based on covariates,
            mahal(), or else a robust variant produced by smahal(). The distance matrix
            dmat may have been penalized by addalmostexact() or addcaliper(). An error
            will result unless dmat has more columns than rows.

z   z is a vector that is 1 for a treated individual and 0 for a control. The number
            of treated subjects, sum(z), must equal the number of rows of dmat, and the
            number of potential controls, sum(1-z), must equal the number of columns of
            dmat.

f   A factor or vector to be finely balanced. Must have length(f)=length(z).

mult   A positive number, mult>0. Determines the penalty used to enforce fine balance
            as max(dmat)*mult. The distance matrix dmat may have been penalized by
            addalmostexact() or addcaliper(), and in this case it makes sense to set mult=1

or mult=2, rather than the default, mult=100. If dmat is already penalized, taking mult>1 creates a larger penalty for deviations from fine balance than the exisiting penalties.

Details

The method is discussed in Chapter 10 of Design of Observational Studies (2010), and it is conceptually the simplest way to implement fine balance. However, the expanded distance matrix can become quite large, so this simplest method is not the most efficient method in its use of computer storage. A more compact implementation uses minimum cost flow in a network (Rosenbaum 1989). Additionally, there are several extensions of fine balance, including near-fine balance (Yang et al. 2012), fine balance for several covariates via integer programming (Zubizarreta 2012, designmatch R-package), and refined balance (Pimentel et al. 2015, rcbalance R-package). Ruoqi Yu's bigmatch R-package implements fine balance and near-fine balance in very large matching problems.

Value

An expanded, square distance matrix with "extra" treated units for use in optimal pair matching. Any control paired with an "extra" treated unit is discarded, as are the "extra" treated units.

Author(s)

Paul R. Rosenbaum

References

Hansen, B. B. (2007). Flexible, optimal matching for observational studies. R News, 7, 18-24. (optmatch package)

Pimentel, S. D., Kelz, R. R., Silber, J. H. and Rosenbaum, P. R. (2015). Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons. Journal of the American Statistical Association, 110, 515-527. Introduces an extension of fine balance called refined balance.

Pimentel, S. D. (2016) Large, Sparse Optimal Matching with R Package rcbalance. Observational Studies, 2, 4-23. An introduction to the rcbalance package.

Pimentel, S. D. (2016). R Package rcbalance. A recommended R package implementing both fine balance and an extension, refined balance.

Rosenbaum, P. R. (1989). Optimal matching for observational studies. Journal of the American Statistical Association, 84(408), 1024-1032. Discusses and illustrates fine balance using minimum cost flow in a network.

Rosenbaum, P. R., Ross, R. N. and Silber, J. H. (2007). Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. Journal of the American Statistical Association, 102, 75-83. Discusses and illustrates fine balance using optimal assignment.

Rosenbaum, P. R. (2010). Design of Observational Studies. New York: Springer. The method and example are discussed in Chapter 10.

Yang, D., Small, D. S., Silber, J. H. and Rosenbaum, P. R. (2012). Optimal matching with minimal deviation from fine balance in a study of obesity and surgical outcomes. Biometrics, 68, 628-636.

Extension of fine balance useful when fine balance is infeasible. Comes as close as possible to fine balance. Implemented in Pimentel's rcbalance package.

Yu, Ruoqi (2018). R package bigmatch.

Zubizarreta, J. R., Reinke, C. E., Kelz, R. R., Silber, J. H. and Rosenbaum, P. R. (2011). Matching for several sparse nominal variables in a case-control study of readmission following surgery. The American Statistician, 65(4), 229-238. Combines near-exact matching with fine balance for the same covariate.

Zubizarreta, J. R. (2012). Using mixed integer programming for matching in an observational study of kidney failure after surgery. Journal of the American Statistical Association, 107, 1360-1371. Extends the concept of fine balance using integer programming. Implemented in R in the designmatch package.

Zubizarreta, J. R. designmatch. A recommended R package for fine balance using integer programming.

Examples

```
data(costa)
z<-1*(costa$welder=="Y")
aa<-1*(costa$race=="A")
smoker=1*(costa$smoker=="Y")
age<-costa$age
x<-cbind(age,aa,smoker)
dmat<-mahal(z,x)
# Mahalanobis distances
round(dmat[,1:6],2) # Compare with Table 8.5 in Design of Observational Studies (2010)
# Impose propensity score calipers
prop<-glm(z~age+aa+smoker,family=binomial)$fitted.values # propensity score
# Mahalanobis distanced penalized for violations of a propensity score caliper.
# This version is used for numerical work.
dmat<-addcaliper(dmat,z,prop,caliper=.5)
round(dmat[,1:6],2) # Compare with Table 8.5 in Design of Observational Studies (2010)
# Because dmat already contains large penalties, we set mult=1.
dmat<-fine(dmat,z,aa,mult=1)
dmat[,1:6] # Compare with Table 10.1 in Design of Observational Studies (2010)
dim(dmat) # dmat has been expanded to be square by adding 5 extras, here numbered 48:52
# Any control matched to an extra is discarded.
## Not run:
# Find the minimum distance match within propensity score calipers.
optmatch::pairmatch(dmat)
# Any control matched to an extra is discarded.  For instance, the optimal match paired
# extra row 48 with the real control in column 7 to form matched set 1.22, so that control
# is not part of the matched sample.  The harmless warning message from pairmatch
# reflects the divergence between the costa data.frame and expanded distance matrix.

## End(Not run)
# Conceptual versions with infinite distances for violations of propensity caliper.
dmat[dmat>20]<-Inf
round(dmat[,1:6],2) # Compare with Table 10.1 in Design of Observational Studies (2010)
```

`lead`	*Lead in Children*

Description

Data from Morton et al. (1982) concerning exposed children whose fathers worked in a battery plant where lead was used in the manufacture of batteries. Exposed children were matched to controls for age and neighborhood. For exposed children, also given are the father's level of exposure to lead at work (level) and the father's hygiene upon leaving the battery plant at the end of the day.

Usage

```
data("lead")
```

Format

A data frame with 33 observations on the following 6 variables.

`control` Blood lead level for the control, micrograms of lead per decaliter of whole blood.

`exposed` Blood lead level for the exposed/treated child, micrograms of lead per decaliter of whole blood.

`level` Father's level of exposure to lead: a factor with levels `high` `low` `medium`

`hyg` Father's hygiene before going home: a factor with levels `good` `mod` `poor`

`both` A factor built from level and hyg: a factor with levels `high.ok` `high.poor` `low` `medium`

`dif` Exposed-minus-control pair difference in blood lead levels.

Details

The data were assembled from two published tables in Morton et al. (1982). One matched pair with no control is omitted here. Small ambiguities in assembling a complete data set from two tables were resolved by a throw of the dice; however, it is a reasonable example to illustrate statistical methods. One table described the exposed-versus-control matched pairs, and these are as in the paper. The second table described the exposed individuals, their level of exposure and their hygiene, and again these are as in the paper. The two tables were linked using the blood lead level of the exposed children, and a couple of ties in these lead levels made for small ambiguities about which control responses belong with which hygienes and exposure levels for the exposed children. See, for instance, rows 18 and 22, where both exposed children have blood lead level 34.

Source

Data are from Morton et al. (1982). They were used as an example in Rosenbaum (1991, 2002, 2011, 2017).

References

Morton, D. E., Saah, A. J., Silberg, S. L., Owens, W. L., Roberts, M. A., & Saah, M. D. (1982). Lead absorption in children of employees in a lead-related industry. American Journal of Epidemiology, 115(4), 549-555.

Rosenbaum, P. R. (1991). Some poset statistics. The Annals of Statistics, 19(2), 1091-1097. <doi:10.1214/aos/1176348141>

Rosenbaum, P. R. (2002). Observational Studies (2nd edition). New York: Springer. Section 4.3.

Rosenbaum, P. R. (2011). A new U statistic with superior design sensitivity in matched observational studies. Biometrics, 67(3), 1017-1027. <doi:10.1111/j.1541-0420.2010.01535.x>

Rosenbaum, P. R. (2017). Observation and Experiment: An Introduction to Causal Inference. Cambridge, MA: Harvard University Press. Chapter 7.

Examples

```
data(lead)
# Reproduces parts of Table 2 in Rosenbaum (2011)
senU(lead$dif,gamma=5.8,m=8,m1=5,m2=8)
senU(lead$dif,gamma=5,m=5,m1=4,m2=5)

# m=2, m1=2, m2=2 is the U-statistic that closely
# resembles Wilcoxon's signed rank test.  Note
# that the results are almost the same.
senWilcox(lead$dif,gamma=5) # In Table 2
senU(lead$dif,gamma=5,m=2,m1=2,m2=2)
```

---

mahal   *Mahalanobis Distance Matrix for Optimal Matching*

Description

Computes a Mahalanobis distance matrix between treated individuals and potential controls. The method is discussed in Chapter 8 of Design of Observational Studies (2010).

Usage

```
mahal(z, X)
```

Arguments

z	z is a vector that is 1 for a treated individual and 0 for a control.
X	A matrix of continuous or binary covariates. The number of rows of X must equal the length of z.

Value

The distance matrix has one row for each treated individual (z=1) and one column for each potential control (z=0). The row and column names of the distance matrix refer to the position in z, 1, 2, ..., length(z).

Author(s)

Paul R. Rosenbaum

References

Hansen, B. B. and Klopfer, S. O. (2006). Optimal full matching and related designs via network flows. Journal of computational and Graphical Statistics, 15(3), 609-627. (optmatch package)

Hansen, B. B. (2007). Flexible, optimal matching for observational studies. R News, 7, 18-24. (optmatch package)

Rosenbaum, P. R. (2010). Design of Observational Studies. New York: Springer. The method and example are discussed in Chapter 8.

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. The American Statistician, 39, 33-38.

Rubin, D. B. (1980). Bias reduction using Mahalanobis-metric matching. Biometrics, 36, 293-298.

Examples

```
data(costa)
z<-1*(costa$welder=="Y")
aa<-1*(costa$race=="A")
smoker=1*(costa$smoker=="Y")
age<-costa$age
x<-cbind(age,aa,smoker)
dmat<-mahal(z,x)
# Mahalanobis distances
round(dmat[,1:6],2) # Compare with Table 8.5 in Design of Observational Studies (2010)
# Impose propensity score calipers
prop<-glm(z~age+aa+smoker,family=binomial)$fitted.values # propensity score
# Mahalanobis distanced penalized for violations of a propensity score caliper.
# This version is used for numerical work.
dmat<-addcaliper(dmat,z,prop,caliper=.5)
round(dmat[,1:6],2) # Compare with Table 8.5 in Design of Observational Studies (2010)
## Not run:
# Find the minimum distance match within propensity score calipers.
optmatch::pairmatch(dmat,data=costa)

## End(Not run)
# Conceptual versions with infinite distances for violations of propensity caliper.
dmat[dmat>20]<-Inf
round(dmat[,1:6],2) # Compare with Table 8.5 in Design of Observational Studies (2010)
```

---

pinto   *Welding and DNA-Protein Crosslinks*

Description

This data set is from Pinto et al. (2000) and it describes 22 professional painters and 22 controls matched for age. All are men. The outcome is a measure of genetic damage, namely the frequency of micronuclei in 3000 oral epithelial cells scraped from the cheek, recorded as micronuclei per 1000 cells. The data are used as an example in Chapter 5 of Design of Observational Studies, where the data illustrate the subtle relationship between doses of treatment and sensitivity to unmeasured biases. The dose of treatment is years of work as a professional painter, from 1.6 years to 40 years for painters, and it is highly correlated with age for painters, but it is zero for controls of all ages.

Usage

```
data("pinto")
```

Format

A data frame with 44 observations on the following 6 variables.

id ID number from Pinto et al. (2000).

pair Pair number, 1 to 22. The pair numbers are different from Design of Observational Studies, where pairs were sorted by doses.

group painter or control

age age in years

years years of work as a professional painter, 0 for controls

longEx Painter in this pair had long exposure, TRUE if years>=4 years.

mn Micronuclei per 1000 cells

Source

The data are from Pinto et al. (2000). The data are used as an example in Chapter 5 of Design of Observational Studies. The later chapter "Anticipated Patterns of Response" shows that the pattern seen in this example is expected in theory, namely focusing on high-dose pairs makes the study more insensitive to unmeasured biases, despite the loss of sample size.

References

Pinto, D., J. M. Ceballos, G. Garcia, P. Guzman, L. M. Del Razo, E. Vera, H. Gomez, A. Garcia, and M. E. Gonsebatt (2000). Increased cytogenetic damage in outdoor painters. Mutation Research/Genetic Toxicology and Environmental Mutagenesis 467, 105-111.

Rosenbaum, P. R. (2003). Does a dose response relationship reduce sensitivity to hidden bias?. Biostatistics, 4, 1-10.

Rosenbaum, P. R. (2004). Design sensitivity in observational studies. Biometrika, 91, 153-164. Does design sensitivity calculations with doses of treatment.

Rosenbaum, P. R. (2010). Design of Observational Studies. New York: Springer. This example is discussed in Chapter 5.

Examples

```
data(pinto)
par(mfrow=c(1,3))
attach(pinto)
boxplot(mn~group,ylim=c(0,6),main="All",ylab="Micronuclei")
boxplot(mn[!longEx]~group[!longEx],ylim=c(0,6),main="Short Ex",ylab="Micronuclei")
boxplot(mn[longEx]~group[longEx],ylim=c(0,6),main="Long Ex",ylab="Micronuclei")

# Calculations in Table 5.5 of Design of Observational Studies (2010)
d<-mn[group=="painter"]-mn[group=="control"] # 22 pair differences
senWilcox(d,gamma=1)
senWilcox(d,gamma=2) # sensitive to gamma=2
senWilcox(d,gamma=3.3)
dLong<-d[longEx[group=="painter"]] # 12 pairs with long exposure
senWilcox(dLong,gamma=3.3) # insensitive to gamma=3.3
```

---

schoket   *DNA Damage in Aluminum Production Workers*

Description

This data set is from Schoket et al. (1991) and is discussed in Chapter 16 of Design of Observational Studies (2010). The data describe 25 aluminum production workers (w) and 25 controls (c) matched for age and smoking. The outcome is a measure of genetic damage, namely DNA adducts per 10^8 nucleotides. The data are used as an example in Chapter 16 of Design of Observational Studies, where the data illustrate the possibility that some treated individuals are strongly affected by treatment, while others are unaffected, so the average treatment effect may be small, but the effect on affected individuals may be large and evident in data.

Usage

```
data("schoket")
```

Format

A data frame with 25 observations on the following 9 variables.

pair Pair number, 1 to 25.

idw Worker ID from Schoket et al. (1991).

agew Worker age in years

smokingw Worker cigarettes per day

adductsw Worker DNA adducts

idc Control ID from Schoket et al. (1991).

agec Control age in years

smokingc Control cigarettes per day

adductsc Control DNA adducts

Source

The data are from Schoket et al. (1991). The data are used as an example in Chapter 16 of Design of Observational Studies (2010).

References

Conover, W. J. and Salsburg, D. S. (1988). Locally most powerful tests for detecting treatment effects when only a subset of patients can be expected to" respond" to treatment. Biometrics, 189-196.

Rosenbaum, P. R. (2007). Confidence intervals for uncommon but dramatic responses to treatment. Biometrics, 63(4), 1164-1171.

Rosenbaum, P. R. (2010). Design of Observational Studies. New York: Springer. This example is discussed in Chapter 16.

Rosenbaum, P. R. (2011). A new U statistic with superior design sensitivity in matched observational studies. Biometrics, 67(3), 1017-1027.

Schoket, B., Phillips, D. H., Hewer, A. and Vincze, I. (1991). 32P-postlabelling detection of aromatic DNA adducts in peripheral blood lymphocytes from aluminium production plant workers. Mutation Research/Genetic Toxicology, 260(1), 89-98.

Stephenson, W. R. (1981). A general class of one-sample nonparametric test statistics based on subsamples. Journal of the American Statistical Association, 76(376), 960-966.

Examples

```
data(schoket)
attach(schoket)
plot(sort(adductsc),sort(adductsw),ylim=c(0,6.4),xlim=c(0,6.4),
   xlab="DNA adducts for controls",ylab="DNA adducts for workers",
   main="Quantile-Quantile Plot") # Compare with Chapter 16
abline(0,1) # line of equality
legend(4,1,lty=1,"x=y")
boxplot(adductsw,adductsc,ylim=c(0,6.4),ylab="DNA adducts",names=c("Worker","Control"))
d<-adductsw-adductsc
senWilcox(d,gamma=1)
senWilcox(d,gamma=1.5) # sensitive to gamma=1.5
```

---

senU   *Sensitivity Analysis for a New U Statistic*

Description

Sensitivity analysis for the new U statistic of Rosenbaum (2011). For m=m1=m2, this is the test of Stephenson (1981). The ranks proposed by Stephenson closely approximate the optimal ranks proposed by Conover and Salzburg (1988) for detecting a treatment that has a large effect on a small subpopulation and no effect on most of the population; see Rosenbaum (2007). The example reproduces some results from Chapter 16 of Design of Observational Studies (2010).

Usage

```
senU(d, gamma = 1, m = 2, m1 = 2, m2 = 2, conf.int = FALSE,
     alpha = 0.05, alternative = "greater", exact = NULL)
```

Arguments

d	A vector of treated-minus-control matched pair differences in outcomes.
gamma	gamma >= 1 is the value of the sensitivity parameter.
m	See m2.
m1	See m2.
m2	If (m,m1,m2) are three integers such that $1 <= m1 <= m2 <= m$, then the triple (m,m1,m2) defines a U statistic. If (m,m1,m2) = (1,1,1), then the U statistic is the sign test statistic. If (m,m1,m2) = (2,2,2), then it is the U statistic that closely approximates Wilcoxons signed rank test. If m=m1=m2, then the U statistic is the test of Stephenson (1981). The general U statistic is discussed in Rosenbaum (2011).
conf.int	If conf.int=TRUE, the a 1-alpha confidence interval and an interval of point estimates is returned in addition to the P-value testing no treatment effect.
alpha	Coverage rate of the confidence interval. With probability at least 1-alpha, the confidence interval will cover the treatment effect providing the bias in treatment assignment is at most gamma.
alternative	If alternative = "greater" or alternative = "less", then one-sided tests and intervals are returned. If alternative = "twosided", then both one sided tests are done, with the smaller P-value doubled to yield a two-sided P-value. If alternative = "twosided", the confidence interval is the intersection of two one-sided 1-alpha/2 confidence intervals.
exact	If exact is NULL, then exact is set to TRUE if length(d) <= 50, and is set to FALSE if length(d) > 50. The ranks used by the U statistic involve combinatorial coefficiencts that grow rapidly with increasing sample size. If exact=TRUE, these ranks are computed exactly using expression (8) in Rosenbaum (2011). If exact=FALSE, the ranks are computed by an asymptotic approximation that does not involve large combinatorial coefficients, specifically expression (9) in Rosenbaum (2012).

Details

The senWilcox function uses a large sample Normal approximation to the distribution of Wilcoxon's signed rank statistic. When gamma=1, it should agree with the wilcox.test() function in the stats package with exact=FALSE and correct=FALSE. The example reproduces the example of the large-sample approximation in Section 3.5 of Design of Observational Studies. Note that the confidence intervals in Table 3.3 of that book are exact, not approximate, so they are slightly different.

Value

pval	The upper bound on the P-value testing no effect in the presence of a bias in treatment assignment of at most gamma. If the bias in treatment assignment is

at most gamma, and if there is no treatment effect, then there is at most an alpah chance that the P-value is less than alpha, this being true for all 0<alpha<1.

estimate   If conf.int=TRUE, the interval of point estimates of an additive treatment effect in the presence of a bias in treatment assigment of at most gamma. If gamma=1, then you are assuming ignorable treatment assignment or equivalently no unmeasured confounding, so the interval collapses to a point, and that point is the usual Hodges-Lehmann point estimate.

conf.int   If conf.int=TRUE, the a 1-alpha confidence interval for an additive treatment effect in the presence of a bias in treatment assignment of at most gamma. If gamma=1, then this is the usual confidence interval obtained by inverting the Wilcoxon test, and it would be appropriate in a paired randomized experiment.

Note

The test of Stephenson (1981) uses ranks similar to those of Conover and Salzburg (1988) which were designed to have high power when most people are unaffected by treatment, but a small subpopulation is strongly affected; see Rosenbaum (2007). This is the situation discussed in Chapter 16 of Design of Observational Studies (2010). Even for pair differences d that are Normal with expectation tau and constant variance, the Wilcoxon test tends to exaggerate the degree of sensitivity to unmeasured bias. Compare the Wilcoxon test and the U statistic with (m,m1,m2) = (5,4,5) in the Normal situation. In a randomized experiment (gamma=1), the two tests have the same Pitman efficiency. However, as the number of pairs increases with tau=0.5, the Wilcoxon test has limiting sensitivity to bias of gamma=3.2 while (m,m1,m2) = (5,4,5) has limiting sensitivity 3.9, and (m,m1,m2) = (8,7,8) has limiting sensitivity 5.1. See Rosenbaum (2011) for specifics.

Author(s)

Paul R. Rosenbaum

References

Conover, W. J. and Salsburg, D. S. (1988). Locally most powerful tests for detecting treatment effects when only a subset of patients can be expected to" respond" to treatment. Biometrics, 189-196.

Hodges Jr, J. L. and Lehmann, E. L. (1963). Estimates of location based on rank tests. The Annals of Mathematical Statistics, 598-611.

Rosenbaum, P. R. (1993). Hodges-Lehmann point estimates of treatment effect in observational studies. Journal of the American Statistical Association, 88(424), 1250-1253.

Rosenbaum, P. R. (2007). Confidence intervals for uncommon but dramatic responses to treatment. Biometrics, 63(4), 1164-1171. <doi:10.1111/j.1541-0420.2007.00783.x> >

Rosenbaum, P. R. (2010). Design of Observational Studies. New York: Springer. The method and example are discussed in Chapter 16.

Rosenbaum, P. R. (2011). A new U statistic with superior design sensitivity in matched observational studies. Biometrics, 67(3), 1017-1027. <doi:10.1111/j.1541-0420.2010.01535.x>

Schoket, B., Phillips, D. H., Hewer, A. and Vincze, I. (1991). 32P-postlabelling detection of aromatic DNA adducts in peripheral blood lymphocytes from aluminium production plant workers. Mutation Research/Genetic Toxicology, 260(1), 89-98.

Stephenson, W. R. (1981). A general class of one-sample nonparametric test statistics based on subsamples. Journal of the American Statistical Association, 76(376), 960-966.

Examples

```
data("schoket")
d<-schoket$adductsw-schoket$adductsc

# With the defaults, m=2, m1=2, m2=2, the U-statistic is very
# similar to Wilcoxon's signed rank statistic
senWilcox(d,gamma=1)
senU(d,gamma=1)

# With m=1, m1=1, m2=1, the U-statistic is the sign test
senU(d,gamma=1,m=1,m1=1,m2=1)
prop.test(sum(d>0),length(d),p=.5,
     alternative="greater",correct=FALSE)$p.value

# With m=m1=m2, this is the test of Stephenson (1981) whose ranks are similar to
# those of Conover and Salzburg (1988); see Rosenbaum (2007).

# The calculations that follow reproduce the sensitivity analysis for the
# data of Schoket et al. () in Chapter 16 of Desgin of Observational Studies (2010).
senU(d,gamma=1,m=2,m1=2,m2=2)
senU(d,gamma=1,m=5,m1=5,m2=5)

senU(d,gamma=1.5,m=2,m1=2,m2=2)
senU(d,gamma=1.5,m=5,m1=5,m2=5)

senU(d,gamma=1.8,m=2,m1=2,m2=2)
senU(d,gamma=1.8,m=5,m1=5,m2=5)

senU(d,gamma=2,m=2,m1=2,m2=2)
senU(d,gamma=2,m=5,m1=5,m2=5)

data(lead)
# Reproduces parts of Table 2 in Rosenbaum (2011)
senU(lead$dif,gamma=5.8,m=8,m1=5,m2=8)
senU(lead$dif,gamma=5,m=5,m1=4,m2=5)

# m=2, m1=2, m2=2 is the U-statistic that closely
# resembles Wilcoxon's signed rank test.  Note
# that the results are almost the same.
senWilcox(lead$dif,gamma=5) # In Table 2
senU(lead$dif,gamma=5,m=2,m1=2,m2=2)
```

senWilcox   *Sensitivity Analysis for Wilcoxon's Signed-rank Statistic*

Description

Sensitivity analysis for Wilcoxon's signed rank statistic in observational studies. Performs a sensitivity analysis for the P-value, the Hodges-Lehmann estimate of an additive treatment effect, and the confidence interval for the effect.

Usage

```
senWilcox(d, gamma = 1, conf.int = FALSE, alpha = 0.05, alternative = "greater")
```

Arguments

d	A vector of treated-minus-control matched pair differences in outcomes.
gamma	gamma >= 1 is the value of the sensitivity parameter.
conf.int	If conf.int=TRUE, the a 1-alpha confidence interval and an interval of point estimates is returned in addition to the P-value testing no treatment effect.
alpha	Coverage rate of the confidence interval. With probability at least 1-alpha, the confidence interval will cover the treatment effect providing the bias in treatment assignment is at most gamma.
alternative	If alternative = "greater" or alternative = "less", then one-sided tests and intervals are returned. If alternative = "twosided", then both one sided tests are done, with the smaller P-value doubled to yield a two-sided P-value. If alternative = "twosided", the confidence interval is the intersection of two one-sided 1-alpha/2 confidence intervals.

Details

The senWilcox function uses a large sample Normal approximation to the distribution of Wilcoxon's signed rank statistic. When gamma=1, it should agree with the wilcox.test() function in the stats package with exact=FALSE and correct=FALSE. The example reproduces the example of the large-sample approximation in Section 3.5 of Design of Observational Studies. Note that the confidence intervals in Table 3.3 of that book are exact, not approximate, so they are slightly different.

Value

pval	The upper bound on the P-value testing no effect in the presence of a bias in treatment assignment of at most gamma. If the bias in treatment assignment is at most gamma, and if there is no treatment effect, then there is at most an alpah chance that the P-value is less than alpha, this being true for all 0<alpha<1.
estimate	If conf.int=TRUE, the interval of point estimates of an additive treatment effect in the presence of a bias in treatment assigment of at most gamma. If gamma=1, then you are assuming ignorable treatment assignment or equivalently no unmeasured confounding, so the interval collapses to a point, and that point is the usual Hodges-Lehmann point estimate.
conf.int	If conf.int=TRUE, the a 1-alpha confidence interval for an additive treatment effect in the presence of a bias in treatment assignment of at most gamma. If gamma=1, then this is the usual confidence interval obtained by inverting the Wilcoxon test, and it would be appropriate in a paired randomized experiment.

Note

The Wilcoxon test is not the best test for use in sensitivity analyses – it tends to exaggerate how sensitive a study is to bias, saying it is more sensitive than it truly is. Other, less familiar, test statistics avoid this issue; see Rosenbaum (2010b, 2011, 2013, 2014, 2015). Learn about this using a shinyapp, at https://rosenbap.shinyapps.io/learnsenShiny/

Author(s)

Paul R. Rosenbaum

References

Hodges Jr, J. L. and Lehmann, E. L. (1963). Estimates of location based on rank tests. The Annals of Mathematical Statistics, 598-611.

Hollander, M., Wolfe, D. and Chicken, E. (2013) Nonparametric Statistical Methods. (3rd edition) New York: John Wiley.

Lehman, E. L. (1975). Nonparametrics. San Francisco: Holden-Day. Reprinted by Prentice-Hall and Springer.

Rosenbaum, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. Biometrika, 74(1), 13-26.

Rosenbaum, P. R. (1993). Hodges-Lehmann point estimates of treatment effect in observational studies. Journal of the American Statistical Association, 88(424), 1250-1253.

Rosenbaum, P. R. (2002). Observational Studies. New York: Springer. Wilcoxon's test is discussed in Section 4.3.3.

Rosenbaum, P. R. (2007). Sensitivity Analysis for M Estimates, Tests, and Confidence Intervals in Matched Observational Studies. Biometrics, 63(2), 456-464. R-packages sensitivitymult and sensitivitymv

Rosenbaum, P. R. (2010). Design of Observational Studies. New York: Springer. The method and example are discussed in Section 3.5.

Rosenbaum, P. R. (2010b). Design sensitivity and efficiency in observational studies. Journal of the American Statistical Association, 105(490), 692-702.

Rosenbaum, P. R. (2011). A new U statistic with superior design sensitivity in matched observational studies. Biometrics, 67(3), 1017-1027.

Rosenbaum, P. R. (2013). Impact of multiple matched controls on design sensitivity in observational studies. Biometrics, 69(1), 118-127. R-packages sensitivitymv, sensitivitymult and sensitivityfull

Rosenbaum, P. R. (2014). Weighted M statistics with superior design sensitivity in matched observational studies with multiple controls. Journal of the American Statistical Association, 109(507), 1145-1158. R-package sensitivitymw

Rosenbaum, P. R. (2015). Bahadur efficiency of sensitivity analyses in observational studies. Journal of the American Statistical Association, 110(509), 205-217.

Examples

```
data(werfel)
d<-werfel$serpc_p-werfel$cerpc_p

# Reproduces the approximate one-sided P-value computed in Section 3.5 of Rosenbaum (2010).
senWilcox(d,gamma=3)

# Reproduces parts of Tables 4.3 and 4.4 in Rosenbaum (2002)
data(lead)
senWilcox(lead$dif,gamma=1,conf.int=TRUE,alternative="twosided")
senWilcox(lead$dif,gamma=2,conf.int=TRUE,alternative="twosided")

# Agrees with the usual Wilcoxon procedures when gamma=1.
senWilcox(d,gamma=1,conf.int=TRUE,alternative="twosided")
stats::wilcox.test(d,conf.int=TRUE,exact=FALSE,correct=FALSE)
```

---

senWilcoxExact   *Exact Sensitivity Analysis for Wilcoxon's Signed-rank Statistic*

Description

Exact sensitivity analysis for Wilcoxon's signed rank statistic in observational studies. Performs a sensitivity analysis for the one-sided P-value. The method can be used in small samples without ties; however, it is primarily of theoretical interest, as the large sample approximation in senWilcox() is fine for most samples of practical size.

Usage

```
senWilcoxExact(d, gamma = 1)
```

Arguments

d   A vector of treated-minus-control matched pair differences in outcomes. There
                must be no ties in |d| when computing the exact distribution. If ties are present,
                use senWilcox() instead.

gamma   gamma >= 1 is the value of the sensitivity parameter.

Details

The exact method is discussed in Section 3.9 of Design of Observational Studies (2010) and is illustrated in Section 3.5. Tables 3.2 and 3.3 of Section 3.5 use these exact calculations.

Value

The upper bound on the one-sided, upper-tailed P-value testing no treatment effect in the presence of a bias in treatment assignment of at most gamma.

Note

The senWilcox() function uses a large-sample approximation, adding confidence intervals and point estimates.

Author(s)

Paul R. Rosenbaum

References

Pagano, M. and Tritchler, D. (1983). On obtaining permutation distributions in polynomial time. Journal of the American Statistical Association, 78, 435-440.

Rosenbaum, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. Biometrika, 74(1), 13-26.

Rosenbaum, P. R. (2010). Design of Observational Studies. New York: Springer. The method and example are discussed in Sections 3.5 and 3.9.

Examples

```
data(werfel)
d<-werfel$serpc_p-werfel$cerpc_p

# Reproduces the exact one-sided P-value computed in Section 3.9 of Rosenbaum (2010).
senWilcoxExact(d,gamma=2)

# Agrees with the usual Wilcoxon procedures when gamma=1.
senWilcoxExact(d,gamma=1)
stats::wilcox.test(d,alternative="greater")

# Reproduces the one-sided confidence interval for gamma=3 in Table 3.3 of Rosenbaum (2010)
senWilcoxExact(d-0.0935,gamma=3)
senWilcoxExact(d-0.0936,gamma=3)
```

---

smahal   *Robust Mahalanobis Distance Matrix for Optimal Matching*

Description

Computes a robust Mahalanobis distance matrix between treated individuals and potential controls. The usual Mahalnaobis distance may ignore a variable because it contains one extreme outlier, and it pays too much attention to rare binary variables, but smahal addresses both issues.

Usage

```
smahal(z, X)
```

Arguments

z   z is a vector that is 1 for a treated individual and 0 for a control.

X   A matrix of continuous or binary covariates. The number of rows of X must
                            equal the length of z.

Details

The usual Mahalnaobis distance may ignore a variable because it contains one extreme outlier, and
it pays too much attention to rare binary variables, but smahal addresses both issues.

To address outliers, each column of x is replaced by a column of ranks, with average ranks used
for ties. This prevents one outlier from inflating the variance for a column, thereby making large
differences count little in the Mahalanobis distance.

Rare binary variables have very small variances, p(1-p) for small p, so in the usual Mahalanobis
distance, a mismatch for a rare binary variable is counted as very important. If you were matching
for US states as 49 binary variables, mismatching for California would not be very important,
because p(1-p) is not so small, but mismatching for Wyoming is very important because p(1-p) is
very small. To combat this, the variances of the ranked columns are rescaled so they are all the
same, all equal to the variance of untied ranks. See Chapter 8 of Design of Observational Studies
(2010).

Value

The robust distance matrix has one row for each treated individual (z=1) and one column for each
potential control (z=0). The row and column names of the distance matrix refer to the position in z,
1, 2, ..., length(z).

Author(s)

Paul R. Rosenbaum

References

Rosenbaum, P. R. (2010). Design of Observational Studies. New York: Springer. The method and
example are discussed in Chapter 8.

Examples

```
data(costa)
z<-1*(costa$welder=="Y")
aa<-1*(costa$race=="A")
smoker=1*(costa$smoker=="Y")
age<-costa$age
x<-cbind(age,aa,smoker)
dmat<-smahal(z,x)
# Mahalanobis distances
round(dmat[,1:6],2) # Compare with Table 8.6 in Design of Observational Studies (2010)
# Impose propensity score calipers
prop<-glm(z~age+aa+smoker,family=binomial)$fitted.values # propensity score
# Mahalanobis distanced penalized for violations of a propensity score caliper.
```

```
# This version is used for numerical work.
dmat<-addcaliper(dmat,z,prop,caliper=.5)
round(dmat[,1:6],2) # Compare with Table 8.6 in Design of Observational Studies (2010)
## Not run:
# Find the minimum distance match within propensity score calipers.
optmatch::pairmatch(dmat,data=costa)

## End(Not run)
# Conceptual versions with infinite distances for violations of propensity caliper.
dmat[dmat>20]<-Inf
round(dmat[,1:6],2) # Compare with Table 8.6 in Design of Observational Studies (2010)
```

---

werfel *Welding Fumes and DNA Damage*

Description

This data set from Werfel et al. (1998) describes 39 electric arc welders and 39 controls matched for age and smoking. All are men. The outcome is a measure of genetic damage; specifically, erpcp_p is a measure of DNA single strand breakage and DNA-protein cross-links using elution rates through polycarbonate filters with proteinase K.

Usage

```
data("werfel")
```

Format

A data frame with 39 observations on the following 6 variables.

cage Age in years of the control in a matched pair.

csmoke NS=nonsmoker, S=smoker for the control in a pair

cerpc_p erpcp_p for the control in a pair

sage Age in years of the welder in a matched pair.

ssmoke NS=nonsmoker, S=smoker for the welder in a pair

serpc_p erpcp_p for the welder in a pair

Source

The data are from Werfel et al. (1998). It is used as an example in Section 3.5 of Design of Observational Studies. It is also discussed in Fogarty (2016) and Rosenbaum (2007, 2015).

References

Fogarty, C. B. (2016). Sensitivity analysis for the average treatment effect in paired observational studies. arXiv preprint arXiv:1609.02112.

Rosenbaum, P. R. (2007). Sensitivity analysis for M estimates, tests, and confidence intervals in matched observational studies. Biometrics, 63(2), 456-464.

Rosenbaum, P. R. (2010). Design of Observational Studies. New York: Springer. This example is discussed in Section 3.5.

Rosenbaum, P. R. (2015). Two R packages for sensitivity analysis in observational studies. Observational Studies, 1(1), 1-17.

Werfel, U., Langen, V., Eickhoff, I., Schoonbrood, J., Vahrenholz, C., Brauksiepe, A., Popp, W. and Norpoth, K. (1998). Elevated DNA single-strand breakage frequencies in lymphocytes of welders exposed to chromium and nickel. Carcinogenesis, 19(3), 413-418.

Examples

```
data(werfel)
d<-werfel$serpc_p-werfel$cerpc_p

# Reproduces the approximate one-sided P-value computed in Section 3.5 of Rosenbaum (2010).
senWilcox(d,gamma=3)

# Agrees with the usual large sample Wilcoxon procedures when gamma=1.
senWilcox(d,gamma=1,conf.int=TRUE,alternative="twosided")
stats::wilcox.test(d,conf.int=TRUE,exact=FALSE,correct=FALSE)
```

Index