Wprowadzenie do wnioskowania statystycznego z programem R część nr 1

Autor Daniel Kosiorowski

ZALECANY PODSTAWOWY PODRĘCZNIK DO KURSU: JOHN VERZANI (2014) USING R FOR INTRODUCTORY STATISTICS, CRC PRESS, BOCA RATON.

W trakcie naszych spotkań będziemy korzystać z następujących pakietów dodatkowych środowiska R, (które także należy sobie wcześniej zainstalować)

UsingR #pakiet stanowiący dodatek do klasycznego podręcznika Johna Verzaniego wymienionego powyżej.

robustbase #odporne procedury statystyczne

rrcov #odporne procedury statystyczne

MASS #instaluje się wraz z programem R, szereg technik statystycznych#

lattice#instaluje się wraz z programem R, wizualizacja danych#

quantreg #odporne regresje#

psych #wizualizacja i analiza danych, szereg użytecznych procedur statystycznych

car #dobre diagramy rozrzutu, klasyczna diagnostyka regresji

mvtnorm #wielowymiarowe rozkłady normalny i T Studenta

e1071 #szereg użytecznych procedur statystycznych

FNN #szereg procedur wykorzystujących zasadę k- najbliższych sąsiadów

Rcmdr # "user friendly" graficzna nakładka na R – zawierająca popularne techniki statystyczne, których z braku czasu nie udało się nam przedstawić

W trakcie spotkania będziemy wykorzystywać wymienione poniżej zbiory danych empirycznych. Zbiory te zapisano w postaci plików tekstowych. Będziemy je ładować do R Studio np. wykorzystując polecenie „Import Dataset" (należy zaznaczyć opcje First row as Names i Delimeter: Tab).

1. USA_DANE #badamy gospodarkę USA przez 654 miesiące ze względu na stopę bezrobocia rejestrowanego, średni czas pozostawania bez pracy, liczbę nowych inwestycji polegających na budowie domu, średniej wartości oszczędności gospodarstwa domowego.

2. FRANCJA # badamy gospodarkę Francji ze względu na poziom płacy minimalnej i stopę bezrobocia przez 17 lat.

3. COMPANIES dysponujemy próbą 100 przedsiębiorstw rozpatrywanych ze względu na 9 zmiennych ekonomicznych.

4. HOSPITAL #dysponujemy próbą 200 szpitali w USA, które rozważamy ze względu na 12 zmiennych organizacyjno-ekonomicznych.

5. HEART #Dane dotyczące czynników ryzyka zawału serca u mężczyzn.

#Błyskawiczne wprowadzenie do środowiska R raz jeszcze…#

#Przykład A#

```
dane3<-c(46.34, 50.34, 48.35, 53.74, 52.06, 49.45, 49.90, 51.25, 49.38, 49.31, 50.62,48.82, 46.90, 49.46, 51.17, 50.36, 52.18, 50.11, 52.49, 4.867) #definiujemy #wektor obserwacji#

mean(dane3) #liczymy średnią arytmetyczną#
dane3[3] #trzecia współrzędna wektora dane3 #
dane3[dane3>48] #elemety wektora dane3 większe od 48
mean(dane3[dane3>48]) #średnia z tych elementów wektora dane3, które są większe od 48#
length(dane3) #liczba elementów wektora#

dane4<-c(46.34, 50.34, 48.35, 53.74, 52.06, 49.45, 49.90, 51.25, 49.38, 49.31, 50.62,48.82, 46.90, 49.46, 51.17, 79.45, 76.80, 80.73, 76.10, 87.01)

dane5<-c(46.34, 50.34, 48.35, 53.74, 52.06, 49.45, 49.90, 51.25, 49.38, 49.31, 50.62,48.82, 46.90, 49.46, 51.17, 1.92, 0.71, 1.26, 0.32, -1.71)

mean(dane3)
dane3[3]
dane3[dane3>48]
mean(dane3[dane3>48])

length(dane3)

nowe_dane_1=dane3+dane4
nowe_dane_2=dane3*dane5
library(MASS) #wgrywamy pakiet MASS#
hist.FD(nowe_dane_2) #dobrej jakości histogram#
sum(nowe_dane_1)
W=rbind(dane3,dane5,dane4) #tworzymy macierz#
```

```
Z=cbind(dane3,dane5,dane4) #tworzymy macierz#
dim(W)
dim(Z)

par(mfrow=c(3,1)) #dzielimy okno graficzne: trzy wiersze i jedną kolumnę#
art<-rep(1,20) #wektor złożony z 20 jedynek#
plot(dane3,art,cex=3,main="dane 3")
abline(v=median(dane3),lwd=3,col="red")
abline(v=mean(dane3),lwd=3,col="blue")
art<-rep(1,20)
plot(dane4,art,cex=3,main="dane 4")
abline(v=median(dane4),lwd=3,col="red")
abline(v=mean(dane4),lwd=3,col="blue")
art<-rep(1,20)
plot(dane5,art,cex=3,main="dane 5")
abline(v=median(dane5),lwd=3,col="red")
abline(v=mean(dane5),lwd=3,col="blue")
par(mfrow=c(1,1)) #wracamy do oryginalnego okna graficznego#
library(lattice)
dotplot(dane3, cex=2)
dotplot(dane4, cex=2)
dotplot(dane5, cex=2)

mean(dane3) #średnia#
median(dane3) #mediana#
var(dane3) #wariancja#
sd(dane3) #odchylenie standardowe
mad(dane3) #mediana odchyleń absolutnych od mediany#
IQR(dane3) #rozstęp międzykwartylowy#
```

```
z3<-(dane3-mean(dane3))/sd(dane3) #klasyczna reguła trzech sigm#
round(z3,digit=2)
zz3<-(dane3-median(dane3))/mad(dane3) #poprawiona reguła trzech sigm#

round(zz3,digit=2)
par(mfrow=c(1,1))
par(cex=2,pch=19)
qqnorm(dane3)
qqline(dane3,col="red", lwd=3,cex=2,pch=19) #wykres kwantyl kwantyl#
par(cex=1,pch=19)
x1<-rnorm(20,10,2)
x2<-rnorm(3,15,2)
x3<-rnorm(2,25,3)
x<-c(x1,x2,x3)
x<-round(x)

par(mfrow=c(2,1)) #dzielimy okno graficzne na dwa wiersze i jedną kolumnę
plot(function(x) dnorm(x, log=FALSE), -5, 5,main = "gęstość rozkładu N(0,1)",
ylab="gęstość")
plot(function(x) pnorm(x, log.p=FALSE), -5, 5, main = "dystrybuanta rozkładu N(0,1)",
type="h",ylab="dystrybuanta")

par(mfrow=c(1,1)) #wracamy do oryginalnego okna graficznego#
x<-seq(-5,5,by=0.05)
g1<-dnorm(x,0,1)
g2<-dt(x,1)
plot(x,g1,type="l",lwd=2)
lines(x,g2,col="red",type="l",lwd=2)
legend("topleft",c("N(0,1)","t(1)"),fill=c("black","red"),cex=2)
```

library(UsingR) #wgrywamy pakiet Johna Verzani'ego

x = rnorm(100) #próba z rozkładu N(0,1)#

y = rt(100, df=3) #próba z rozkładu Student t(3)

QQplot(x,y) #porównanie za pomocą „ładnego” wykresu kwantyl kwantyl#

simple.densityplot(x,y) #oszacowania jądrowe funkcji gęstościprawdopodobieństwa#

QQplot(x,y) porównanie za pomocą wykresu kwantyl kwantyl

QQplot(dane3,dane4) #co nieco bardziej złośliwe dane#

w1=rnorm(1000,2,3) #generujemy obserwacje z rozkładu N(2,3)

w2=rchisq(1000,2,5) #generujemy obserwacje z niecentralnego rozkładu chi kwadrat # $\chi^2(2,5)$ (2 stopnie swobody, parametr niecentralności wynosi 5)

w3=rexp(1000)*(-1) #próba z rozkładu skośnego#

w4=rnorm(1000,0,1)

w5=rt(1000,2,1.5) # generujemy obserwacje z niecentralnego rozkładu Studenta t(2, 1.5)

library(MASS)

hist.FD(w1)

hist.FD(w2)

hist.FD(w3)

#**wgrywamy zbiór danych HEART** #czynniki ryzyka zawału serca u mężczyzn#

attach(HEART) #dzięki tej komendzie korzystamy z nazw kolumn macierzy danych

library(e1071)
library(psych)

POPULARNA MIARA SKOŚNOŚCI

$$\frac{\sqrt{n}\sum_i (x_i-\overline{x})^3}{\left(\sum_i (x_i-\overline{x})^2\right)^{3/2}}$$

skewness(AGE)

POPULARNA MIARA KURTOZY

$$\frac{n\sum_i (x_i-\overline{x})^4}{\left((\sum_i x_i-\overline{x})^2\right)^2}-3.$$

kurtosis(AGE)

IQR(AGE) #rozstęp międzykwartylowy#
mad(AGE)

#Nasze funkcje licząca współczynniki asymetrii#
as_1<-function(z){mean((z)-median(z))/sd(z)}#prosty współczynnik asymetrii#
as_1(w1)
as_1(w2)
as_1(w3)
as_2<-function(z){sum((z-mean(z))^3)/(sd(z))^3}

Pytanie* Czy współczynnik as_2 można potraktować jako sensowną miarę skośności?

Zadanie: policz wartości współczynników skośności i kurtozy dla wektorów w1, w2, w3, w4, w5.

FAM=as.factor(FAMILY_DIS) #czynnik – występowanie choroby serca w rodzinie
boxplot(SMOKING~FAM)

ind_1=which(INFARCTION=='YES')

ind_2=which(INFARCTION =='NO') #wyodrębniamy dwie podpróby#

DATA_1=HEART[ind_1,]

DATA_2=HEART[ind_2,]

par(mfrow=c(1,2))

hist(DATA_1$AGE)

hist(DATA_2$AGE)

par(mfrow=c(1,1))

describe(AGE) #statystyki opisowe liczone w pakiecie psych#

SZACOWANIE PARAMETRÓW METODĄ NAJWIĘKSZEJ WIARYGODNOŚCI#

x <- rlnorm(100)

library(MASS)

wyn1=fitdistr(x, "lognormal")

wyn2=fitdistr(x, "exponential")

wyn1

wyn2

wyn1$loglik

wyn2$loglik

```
library(lattice) #wgrywamy pakiet zawierający szereg funkcji graficznych#
densityplot(DATA_1$AGE)
densityplot(DATA_2$AGE)
densityplot(w1)
densityplot(w2)
densityplot(w3)

AGE_den=density(AGE) # „podstawowe dla R” oszacowanie jądrowe gęstości #
plot(AGE_den,lwd=2,main= "oszacowanie gestosci ")

wyn_AGE_1=fitdistr(AGE, "lognormal")
wyn_AGE_2=fitdistr(AGE, "normal")
grid= seq(min(AGE),max(AGE),length=100)
lines(grid,dnorm(grid,wyn_AGE_2$estimate[1],wyn_AGE_2$estimate[2]),lwd=2,col="red")
lines(grid,dlnorm(grid,wyn_AGE_1$estimate[1],wyn_AGE_1$estimate[2]),lwd=2,col="green
")
```

Niech $f(x)$ będzie nieznaną funkcją gęstości, jej estymator jądrowy $\hat{f}(x)$ wyznaczony na podstawie próby $X^n = \{x_1,\ldots,x_n\}$ ma postać

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right),$$

gdzie K oznacza funkcję nazywaną jądrem h jest tzw. parametrem wygładzania (szerokością pasma, za pomocą którego kontrolujemy stosunek gładkości $\hat{f}$ do dokładności dopasowania do danych. Wybór jądra K ma mniejszy wpływ na zachowanie się $\hat{f}$, kluczowym jest jednak wybór parametru wygładzania h .

```
?density #opcje procedury (tzn. wybór szerokości pasma, jądra itd.)#
```

Zadanie: oszacować gęstość i dystrybuantę na podstawie poniżej zdefiniowanych wektorów obserwacji z1 i z2 wykorzystując metodę największej wiarygodności oraz estymator jądrowy.

```
#MIESZANINA 1#

x1<-rnorm(9000,0,1)
x2<-rnorm(1000,0,3)
```

```
URNA<-c(x1,x2)
z1<-sample(URNA,1000,replace=TRUE)

den_1<-density(z1)
plot(den_1,lwd=3)
```

#MIESZANINA 2#

```
X1<-rnorm(7000,0,10)
X2<-runif(500,-3.1,-3.0)
X3<-runif(500,3.1,3.2)
X5<-runif(500,1.1,1.2)
X6<-runif(500,-6.1,-6)
X7<-runif(500,5.1,5.2)
X8<-runif(500,-7.1,-7)

URNA<-c(X1,X2,X3,X5,X5,X7,X8)

z2<-sample(URNA,1000,replace=TRUE)

den_2<-density(z2)
plot(den_2,lwd=3)
```

#Porównanie własności estymatorów za pomocą symulacji#

#PRZYKŁAD B#

```
A<-matrix(c(1,2,3,4,5,6,7,8,9,10,11,12),nrow=3,ncol=4) #definiujemy macierz#
B<-matrix(c(1,2,3,4,5,6,7,8,9,10,11,12),nrow=4,ncol=3) #definiujemy macierz#
B

A[,2]
A[2,]
A[2,2]
apply(A,2,mean)
apply(A,1,mean) #czym różnią się te dwie komendy?#
A*A
A%*%A #mnożenie po współrzędnych#
A%*%B #klasyczny iloczyn macierzy#
C=A%*%B

t(C) #transpozycja macierzy#
suma<-c()

for(i in 1: nrow(B)) suma[i]<-sum(B[i,]) #prosta pętla#
suma
```

#PRZYKŁAD C# **#porównanie średniej i mediany z próby#**

```
library(MASS)
library(lattice)
x<-rnorm(1000,10,10)
hist(x,lwd=2,cex.lab=1.6)

Z<-matrix(nrow=1000,ncol=10)
for (i in 1:1000) Z[i,]<-rnorm(10,10,10)
means<-apply(Z,1,mean)
medians<-apply(Z,1,median)

hist(means,lwd=2,cex.lab=1.6,col=2)
hist(medians,lwd=2,cex.lab=1.6,col=2)
boxplot(means,medians,cex.axis=2)
```

porównanie mediany i przyciętej średniej ?

porównanie średniej przyciętej i średniej winsorskiej?

α – przycięta średnia: przytnij frakcję $\alpha / 2$ największych i $\alpha / 2$ najmniejszych obserwacji a następnie policz zwykłą średnią z pozostałych obserwacji.

α – winsorska średnia: zastąp $\alpha / 2$ największych i $\alpha / 2$ najmniejszych obserwacji kolejnymi najbliższymi obserwacjami a następnie policz średnią.

Przykład: 1,2,3,4,100

40% przycięta średnia = (2+3+4)/3

40% winsorska średnia = (2+2+3+4+4)/5

```
x<-rt(1000,2,10) #1000 obserwacji wygenerowanych z rozkładu Studenta o dwóch stopniach #swobody parametrze położenia 10#
densityplot(x,lwd=2,cex.lab=1.6) #oszacowanie gęstości prawdopodobieństwa#
Z<-matrix(nrow=1000,ncol=10) #definiujemy macierz, 1000 wierszy, 10 kolumn#
for (i in 1:1000) Z[i,]<-rt(10,2,10) #każdy wiersz macierzy zawiera próbę z rozkładu t(2,10)#
tr.mean=function(x) mean(x,tr=0.2) #funkcja licząca przyciętą średnią#
trmeans<-apply(Z,1,tr.mean) #liczymy średnią z każdego wiersza#
medians<-apply(Z,1,median) #liczymy median z każdego wiersza#

boxplot(trmeans,medians,cex.axis=2) #porównujemy średnią i medianę z próby#
par(mfrow=c(1,2)) #dzielimy okno graficzne na jeden wiersz i dwie kolumny#
hist.FD(trmeans) #histogram średnich przyciętych  z prób#
hist.FD(medians) #histogram median z prób#
```

Zadanie: Porównaj w analogiczny sposób tzn. za pomocą symulacji jednowymiarowe miary rozrzutu: odchylenie standardowe, medianę odchyleń absolutnych od mediany, rozstęp międzykwartylowy dla symetrycznych populacji, dla populacji o tłustych ogonach i dla populacji o lekkich ogonach, ale wykazujących skośność. Czy Twoim zdaniem istnieje estymator, który jest najlepszy w każdej z rozpatrywanych sytuacji?

#ESTYMACJA PRZEDZIAŁOWA#

proba_1=rnorm(20,5,2)

proba_2=rnorm(17,4,5)

proba_3=rlnorm(20,log(5.2),log(1.5))

t.test(proba_1,conf.level=0.9)

shapiro.test(proba_1) #popularny test normalności#

shapiro.test(proba_2)

shapiro.test(proba_3)

t.test(proba_3,conf.level=0.9)

 # WYBRANE TESTY ISTOTNOŚCI#

binom.test(70, 100,p=0.6) #test prawdopodobieństwa sukcesu, wskaźnika struktury#

t.test(proba_1, mu=5.5, alternative='greater',conf.level=0.9) #test Studenta#

t.test(proba_1, mu=6, alternative='two.sided',conf.level=0.9)

t.test(proba_1,conf.level = 0.90, alternative="less") #jednostronny przedział ufności#

dotplot(dane5,cex=1.4)

shapiro.test(dane5)

hist(dane5, prob=TRUE)

lines(density(dane5))

t.test(dane5, alternative='two.sided', mu=52, conf.level=.95) #test Studenta#

t.test(proba_1,proba_2, alternative='two.sided', conf.level=.95)

boxplot(proba_1,proba_2)

var.test(proba_1,proba_2, alternative='two.sided', conf.level=.95) #test równości wariancji

wilcox.test(proba_1,proba_2)

wilcox.test(dane3,dane5)

ks.test(proba_1,proba_2) #test Kołmogorowa-Smirnoffa#

Zadanie: wygeneruj dwie próby z rozkładów lognormalnych różniących się parametrami, przeprowadź testy Studenta oraz testy Wilcoxona i Kołmogorowa – Smirnoffa. Stosowanie którego z tych testów istotności wydaje Ci się najwłaściwsze w kontekście porównania populacji, które wygenerowały te próby.

Załadujmy zbiór HOSPITAL. Zbiór ten zawiera dane dotyczące szpitali w USA z podziałem na grupy szpitali. Mamy w nim m. in. następujące zmienne grupujące:

Region: 1=South, 2=Northwest, 3=Midwest, 4=Southwest, 5=Rocky Mountain, 6= California, 7=Northeast;

Control: (typ własności) 1= government-nonfederal, 2= nongovernment-nonprofit, 3=for profit, 4=federal-government;

Service (typ szpitala) 1= ogólny, 2=wysoce specjalistyczny

attach(HOSPITAL)

ind_1=which(Service==1) #dzielimy szpitale ze względu na typ własności#

ind_2=which(Service==2)

HOSPITAL_1=HOSPITAL[ind_1,]

HOSPITAL_2=HOSPITAL[ind_2,]

dim(HOSPITAL_1)

dim(HOSPITAL_2)

dane_1=cbind(HOSPITAL_1$Beds,HOSPITAL_1$Personnel)

dane_2=cbind(HOSPITAL_2$Beds,HOSPITAL_2$Personnel)

library(MASS)

attach(HOSPITAL)

plot(Admissions~Personnel,cex=2)

lsreg=lm(Admissions~Personnel) #estymator NK#

abline(lsreg,lwd=2,cex=3,col='red')

summary(lsreg)

plot(lsreg) #klasyczna diagnostyka regresji#

names(HOSPITAL) <- make.names(names(HOSPITAL))

RegModel.1 <- lm(Admissions~Beds, data=HOSPITAL)

summary(RegModel.1)

influencePlot(RegModel.1, id.method="noteworthy", id.n=3)

#PROSTA JEDNOCZYNNIKOWA ANALIZA WARIANCJI (TEST ANOVA)

Control_1=as.factor(Control) #czynnik rodzaj nadzoru /własności szpitala#

boxplot(Beds~Control_1) #sprawdzamy spełnienie założenie stosowalności testu#

?bartlett.test #można użyć bardziej profesjonalnego sprawdzianu#

library(robustbase)

adjbox(Beds~Control_1) #sokrygowany ze względu na skośność boxplot#

library(rrcov)

adjbox(Beds~Control_1)

#Zadanie: Zaproponuj metodę wykrywania outlierów z wykorzystaniem obiektu adjbox

anal_war_1=aov(Beds~Control_1)

summary.aov(anal_war_1)

Zadanie: Przeprowadź test anova dla różnych regionów geograficznych USA i zmiennej beds/total expenditures. W jaki sposób można Twoim zdaniem „uodpornić" klasyczny test anova?

#PROSTA DWUCZYNNIKOWA ANALIZA WARIANCJI Z INTERAKCJĄ

Service_1=as.factor(Service) #drugi czynnik typ szpitala#

boxplot(Beds~Control_1+Service_1) #sprawdzamy założenie stosowalności testu

adjbox(Beds~Control_1+Service_1)

anal_war_2=aov(Beds~Control_1+Service_1+Control_1*Service_1)

summary.aov(anal_war_2)

ind_1=which(Service==1) #dzielimy szpitale na podgrupy#

ind_2=which(Service==2)

HOSPITAL_1=HOSPITAL[ind_1,]

HOSPITAL_2=HOSPITAL[ind_2,]

#TEST STUDENTA DLA DWÓCH NIEZALEŻNYCH PRÓB#

t.test(HOSPITAL_1$Admissions,HOSPITAL_2$Admissions)

#alternatywa dla testu Studenta?

#test Wilcoxona sumy rang#

#KILKA ZAGADNIEŃ Z WIELOWYMIAROWEJ ANALIZY STATYSTYCZNEJ#

PRZYKŁAD D

library(rrcov)

data(maryo) #klasyczny przykład pokazujący wady zwykłego współczynnika korelacji

plot(maryo,cex=2,pch=19)

imin <- which(maryo[,1]==min(maryo[,1])) # imin = 9

imax <- which(maryo[,1]==max(maryo[,1])) # imax = 19

maryo1 <- maryo

maryo1[imin,1] <- maryo[imax,1] #zamieniamy punkty w oryginalnym zbiorze maryo#

maryo1[imax,1] <- maryo[imin,1] #zamieniamy punkty w oryginalnym zbiorze maryo#

plot(maryo1,cex=2,pch=19)

cov(maryo) #**klasyczny estymator**#

	V1	**V2**
V1	0.560	0.531
V2	0.531	0.761

cov(maryo1) #**klasyczny estymator**#

	V1	**V2**
V1	0.5609	0.0363
V2	0.0363	0.7615

covMcd(x = maryo, alpha = 0.5) #**odporny estymator MCD**#

Robust Estimate of Location:

V1	**V2**
-0.2886	-0.2397

Robust Estimate of Covariance:

	V1	**V2**
V1	0.7147	0.743
V2	0.7430	1.015

covMcd(x = maryo1, alpha = 0.5) #**odporny estymator MCD**#

Robust Estimate of Location:

V1	**V2**
-0.2964	-0.2388

Robust Estimate of Covariance:

	V1	**V2**
V1	0.5945	0.5817
V2	0.5817	0.9082

PRZYKŁAD E

```
library(MASS)
library(rrcov)
data(maryo)
plot(maryo,cex=2)
cov(maryo)
```

cov.mve(maryo) #**estymator minimalnej elipsoidy objętości**#

cov.mcd(maryo) #estymator minimalnego wyznacznika macierzy #kowariancji

#PRZYKŁAD F#

```
library(rrcov)
data(maryo)
CovOgk(maryo)
cov(maryo)
```

Zadanie: Zastosuj estymatory MVE, MCD i OGK do zbioru USA_DANE

#PRZYKŁAD G# #TESTY HOTELLINGA T^2#

library(rrcov) #ważny pakiet zawierający m. in. testy Hotellinga T^2

```
?T2.test

#T2.test(x, y = NULL, mu = 0, conf.level = 0.95, method=c("c", "mcd"), ...)#
```

#Załadujmy zbiór danych HOSPITAL

```
library(aplpack)

#dwuwymiarowy wykres ramka wąsy#

attach(HOSPITAL)

bagplot(c(Personnel,Admissions),factor=5.5,create.plot=TRUE,approx.limit=300,
show.outlier=TRUE,show.looppoints=TRUE, show.bagpoints=TRUE,dkmethod=2,
show.whiskers=TRUE,show.loophull=TRUE,

show.baghull=TRUE,verbose=FALSE,pch=19,cex=1.2,
xlab="Personnel",ylab="Admission",cex.axis=1.8,main="Wykres pudełkowy 2D" )


#przygotowanie do przeprowadzenia testu Hotellinga T2

ind_1=which(Service==1)

ind_2=which(Service==2)

HOSPITAL_1=HOSPITAL[ind_1,]

HOSPITAL_2=HOSPITAL[ind_2,]

dim(HOSPITAL_1)

dim(HOSPITAL_1)

dane_1=cbind(HOSPITAL_1$Beds,HOSPITAL_1$Personnel)

dane_2=cbind(HOSPITAL_2$Beds,HOSPITAL_2$Personnel)
```

library(aplpack)

#dwuwymiarowy wykres ramka wąsy jako sprawdzian założeń stosowalności testu#

```
par(mfrow=c(1,2))

bagplot(dane_1,xlab="beds",ylab="personnel",main="grupa szpitali nr 1")

bagplot(dane_2,xlab="beds",ylab="personnel",main="grupa szpitali nr 2")

par(mfrow=c(1,1))
```

T2.test(dane_1,dane_2)

Polecenie: przeprowadź testy Hotellinga dla jednej grupy oraz dla dwóch grup niezależnych dla zbioru HOSPITAL. W przypadku testu dla jednej próby interesują nas zmienne: przyjęcia/wydatki całkowite oraz liczba łóżek/wydatki całkowite. W przypadku testu dla dwóch prób interesuje nas porównanie regionów „South" i „California"

wskazówka: pakiet rrcov

PRZYKŁAD* #skośny rozkład Studenta i normalny 2D + boxplot2D #

#Jak zachowuje się test Hotellinga T^2 w przypadku tego typu odstępstwa od jego założeń?#

library(sn)

```
?rmst
xi <- c(1, 1)
Omega <- diag(2)
Omega[2,1] <- Omega[1,2] <- 0.5
alpha <- c(2,2)
#alpha<-c(2,-2)#
rnd <- rmst(1000, xi, Omega, alpha, 3)
plot(rnd)
cov(rnd)
eigen(cov(rnd))
library(aplpack)
bagplot(rnd,factor=5.5,create.plot=TRUE,approx.limit=300,
show.outlier=TRUE,show.looppoints=TRUE,
show.bagpoints=TRUE,dkmethod=2,
show.whiskers=TRUE,show.loophull=TRUE,
show.baghull=TRUE,verbose=FALSE,pch=19,cex=1.2,xlab="",ylab="",cex.axis=1.8)

library(MASS)
Service_1=as.factor(Service)
dane=cbind(Beds,Admissions)
analiza_1=qda(dane,Service_1) #kwadratowa funkcja dyskryminacyjna#
analiza_2=lda(dane,Service_1) #liniowa funkcja dyskryminacyjna#
wyniki_1=as.double(predict(analiza_1,dane)$class)
etykiety_1=as.double(Service_1)
wyniki_2=as.double(predict(analiza_2,dane)$class)
etykiety_2=as.double(Service_1)
```

#WYBRANE REGRESJE NIEPARAMETRYCZNE I ODPORNE#

#PRZYKŁAD H

```
x<-seq(-10,10,by=0.2)
n<-length(x)
eps<-rt(n,2)
y<-2*x+1+eps

REG<-lm(y~x)
summary(REG)
plot(x,y,lwd=2,pch=3,cex.axis=1.8)

abline(REG,lwd=2,col="red")
round(REG$residuals,digit=2)
round(REG$fitted,digit=2)
RES<-REG$residuals
FIT<-REG$fitted
plot(RES,cex=3)
plot(FIT,RES,cex=3)
abline(h=0,lwd=2,col="red")
```

Zadanie: Załaduj zbiór danych FRANCJA. Oszacuj równanie regresji stopa bezrobocia(zmienna objaśniana) vs. płaca minimalna (zmienna objaśniająca) za pomocą metody najmniejszych kwadratów. Skomentuj „stopień złośliwości" danych. Czy Twoim zdaniem pośród danych występują obserwacje odstające?
Czy popełniamy błąd specyfikacji decydując się wykorzystać prostą zależność liniową pomiędzy zmiennymi?

```
attach(FRANCJA)
plot(FRANCJA$MW,FRANCJA$UR,cex=2)
lsreg=lm(FRANCJA$UR~FRANCJA$MW)

abline(lsreg,lwd=2,cex=3,col='red')

summary(lsreg)

RES2<-rlm(FRANCJA$UR~FRANCJA$MW) #popularny M-estymator Hubera#

summary(RES2)

abline(RES2,lwd=5,col="blue")

deviance(RES2)

###Regresja najmniejszej mediany kwadratów###

RES4<-lqs(FRANCJA$UR~FRANCJA$MW,method="lms") #najmniejsza mediana
kwadratów#

RES4

RES5<-lqs(FRANCJA$UR~FRANCJA$MW,method="lts") # najmniejsze #przycięte
kwadraty
```

RES5

abline(RES4,lwd=2,col="green")

abline(RES5,lwd=2,col="pink")

#NIEPARAMETRYCZNA REGRESJA LOESS#

library(car)

plot(FRANCJA$UR~FRANCJA$MW, xlab=" FRANCJA$MW ", ylab=" FRANCJA$UR", data=FRANCJA)

with(FRANCJA, lines(lowess(FRANCJA$MW, FRANCJA$UR, f=0.5, iter=0), lwd=2))

#Załadujmy zbiór danych HOSPITAL#

attach(HOSPITAL)

plot(Admissions~Personnel,cex=1.4)

lines(lowess(Personnel, Admissions, f=0.5, iter=0), lwd=2) #regresja loess#

#KOLEJNA ALTERNATYWA DLA REGRESJI NK – regresja lokalnie wielomianowa) #

library(KernSmooth)

n <- 50

x <- runif(n)

f <- function (x) { cos(3*x) + cos(5*x) }

y <- f(x) + 0.5*rnorm(n)

plot(y~x,cex=1.8)

curve(f(x), add=T, lty=2,lwd=2)

bw <- dpill(x,y) #automatyczny wybór parametru wygładzania#

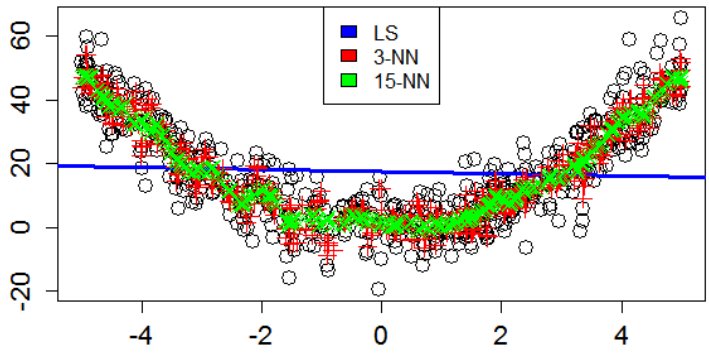
lines(locpoly(x,y,degree=0, bandwidth=bw), col='red',lwd=2)

lines(locpoly(x,y,degree=1, bandwidth=bw), col='green',lwd=2)

lines(locpoly(x,y,degree=2, bandwidth=bw), col='blue',lwd=2)

legend(par("usr")[1], par("usr")[3], yjust=0, c("stopien=0", "stopien=1", "stopien=2"), lwd=2, lty=1, col=c('red', 'green', 'blue'))

NIEPARAMETRYCZNA REGRESJA K- NAJBLIŻSZYCH SĄSIADÓW



Regresja k – najbliższych sąsiadów

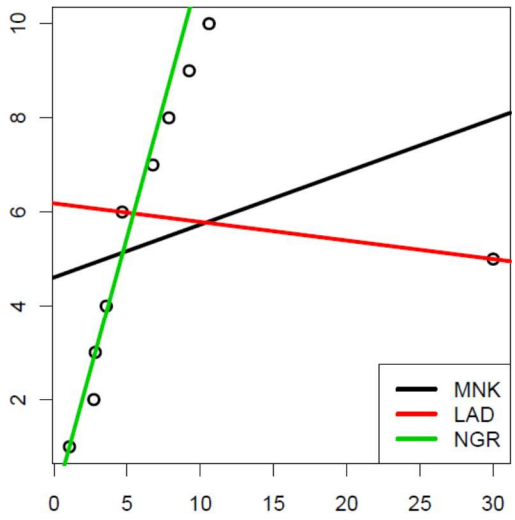
Dysponujemy próbą $\{y_i, \mathbf{x}_i\} \subset \mathbb{R}^{p+1}$, definiujemy estymator regresji k – najbliższych sąsiadów jako

$$\hat{f}(\mathbf{x}) = ave(y_i : x_i \in N_k(\mathbf{x})) ,$$

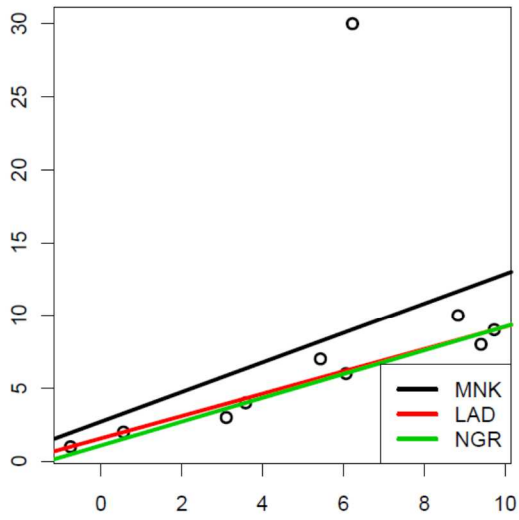
gdzie ave oznacza średnią, $N_k(\mathbf{x})$ oznacza sąsiedztwo punktu $\mathbf{x}$ zawierające k najbliższych (w sensie pewnej odległości) punktów do $\mathbf{x}$.

PRZYKŁAD I

```
library(FNN)
X<-runif(500,-5,5)
Y<-2*X^2+rnorm(500,0,7)
plot(Y~X,cex=1.6,xlab="",ylab="",cex.axis=1.3)
reg<-lm(Y~X)
abline(reg,lwd=3,col="blue")
library(robustbase)
regmm<- lmrob(Y ~X, method = "MM") # MM regresja odporna  Yohai'a #
abline(regmm,col= "brown",lwd=3)
pac.knn3<- knn.reg(X, y=Y, k=3) #regresja 3-najbliższych sąsiadów#
pac.knn15<-knn.reg(X, y=Y, k=15) #regresja 15-najbliższych sąsiadów#
points(X, pac.knn3$pred, col=2, pch=3,cex=1.4)
points(X, pac.knn15$pred, col="green", pch=4,cex=1.4)
legend("top",c("LS","3-NN","15-NN"),fill=c("blue","red","green"),cex=1.3)
```

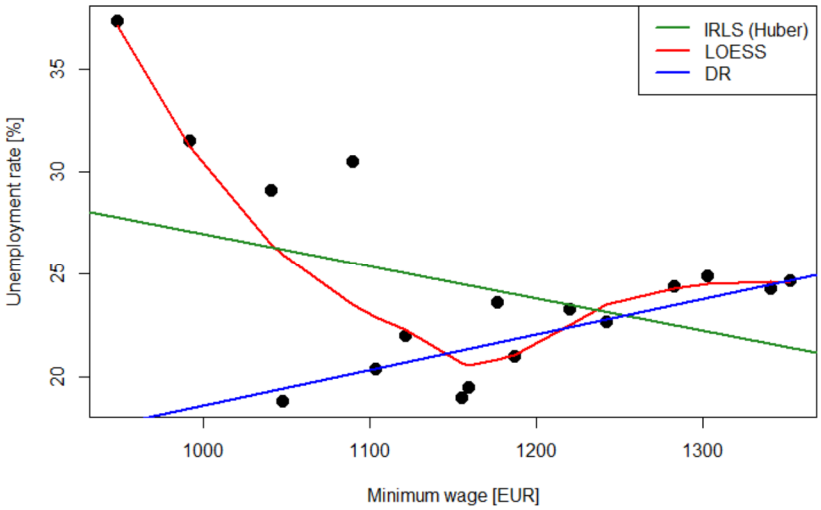


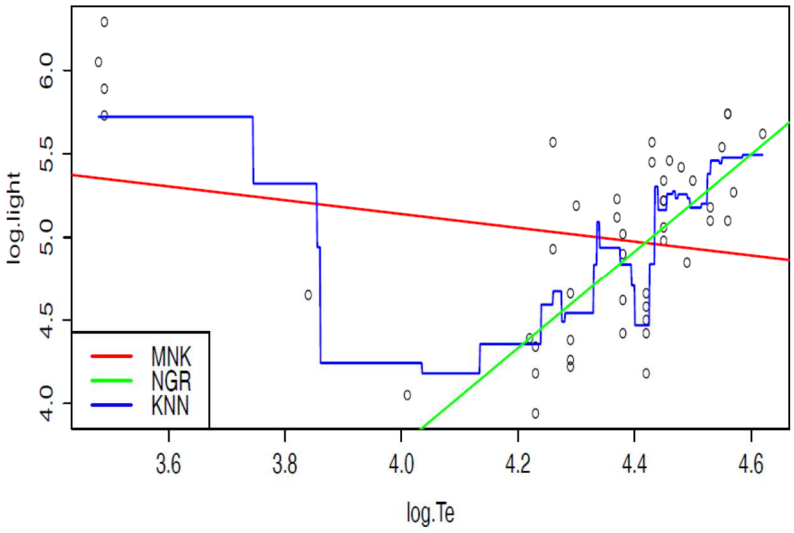
(a) Obserwacja odstająca ze względu na zmienną objaśniającą.



(b) Obserwacja odstająca ze względu na zmienną objaśnianą.

„REGRESJA ODPORNA VS. NIELINIOWA VS. NIEPARAMETRYCZNA”





Regresja odporna vs regresja nieparametryczna vs regresja semiparametryczna?

Statystyka odporna = statystyka nieparametryczna?

REGRESJA ODPORNA

Dysponujemy próbą $\mathbf{Z}^n = \{(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)\}$,

gdzie $y_i \in \mathbb{R}$ oznaczają odpowiedzi, $\mathbf{x}_i \in \mathbb{R}^p$ oznaczają predyktory – wektory wierszowe

Naszym celem jest predykcja Y za pomocą $\mathbf{X}^T\beta$.

Oznaczmy resztę dla danego β za pomocą $r_i(\beta) = y_i - \mathbf{x}_i^T\beta$.

Estymator najmniejszej mediany kwadratów) (regresja najmniejszej mediany kwadratów, kwadratów) (LMS) została gruntownie opisana przez Rousseeuw (1984), jednak była antycypowana kilka lat wcześniej wcześniej przez Franka Hampela.

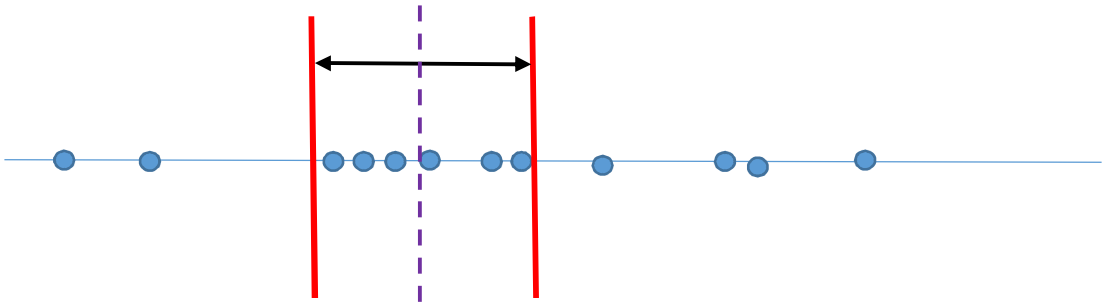
$$\hat{\beta} = \min_{\beta} Med\left\{ r_i(\beta)^2 : i = 1, \ldots, n \right\},$$

równoważnie

$$\min_{\beta} MAD\left\{ r_i(\beta) \right\}.$$

Rousseeuw (1984) wykazał wysoką odporność tej regresji oraz jej dobre własności statystyczne.

Wariant estymatora LMS będący 1D estymatorem położenia: **SHORTH** (brak polskiej nazwy) – środek najkrótszego przedziału zawierającego połowę obserwacji.



UWAGI: Estymator LMS niej jest jak to bywa zazwyczaj $\sqrt{n}$ ale jedynie $\sqrt[3]{n}$ zgodny tzn. $\sqrt{n}\left\|\hat{\beta} - \beta\right\| \to O_p(\infty)$, $\sqrt[3]{n}\left\|\hat{\beta} - \beta\right\| = O_p(1)$, estymator LMS nie jest lokalnie stabilny oraz wymaga bardzo wielkich prób.

W przypadku odpornej regresji mierzymy dobroć dopasowania za pomocą **alternatywnego współczynnika determinacji**

$$R^2 = 1 - \frac{med\left|r_i\right|}{MAD(y_i)}.$$

W roku 1984 Rousseeuw i Yohai (1984) zaproponowali klasę S estymatorów wprowadzając kryterium

$$\min_{\beta} S(r_i(\beta)),$$

gdzie S oznacza miarę rozrzutu (dla $S = MAD$ otrzymamy estymator LMS, dla $S = SD$ otrzymamy estymator NK). Rousseeuw i Yohai (1984) jako miarę rozrzutu zaproponowali monotoniczny M estymator rozrzutu. Ich estymator był $\sqrt{n}$ zgodny i asymptotycznie normalny. Estymatory te mają jednak pewne problemy z efektywnością.

Są wysoce odporne zarówno w przypadku obserwacji odstających, co do zmiennej objaśnianych jak i w przypadku zmiennych objaśniających.

Warto wspomnieć o **MM – estymatorze** zaproponowanym przez Yohai'a (1987) w celu usunięcia mankamentów S – estymatora związanych z jego efektywnością przy zachowaniu jego wyśmienitej odporności. Dla MM estymatora BP=0.5 i cechuje go bardzo wysoka efektywność przy normalności populacji (BP to tzw. punkt załamania próby skończonej – maksymalna frakcja „złych" obserwacji w próbie nie powodująca „załamania" jakości procedury statystycznej).

W tej metodzie estymacji w pierwszym kroku stosujemy wysoce odporny S- estymator regresji, który minimalizuje M – miarę rozrzutu reszt (pierwsze M) . Następnie oszacowaną miarę rozrzutu traktuje się jako ustaloną dla kolejnego kroku, w którym szacuje się parametry regresji za pomocą M – estymatora (drugie M). MM- estymator dostępny jest w pakiecie **robustbase.**

#PRZYKŁAD J METODA BOOTSTRAP – OBCIĄŻENIE I WARIANCJA #ESTYMATORA

Interesuje nas oszacowanie parametru θ nieznanego rozkładu F , powiedzmy $\theta(F)$.

Rozważmy próbę złożoną z n – obserwacji $X^n = \{x_1,...,x_n\}$ pobraną z F , niech F_n oznacza rozkład empiryczny wyznaczony na podstawie X^n (jest to rozkład dyskretny przyporządkowujący prawdopodobieństwo $1/n$ każdemu punktowi próby).

Obliczamy wartość statystyki $\hat{\theta} = \theta(F_n)$ jako wartość estymatora $\theta(F)$.

Podstawowe pytanie: Jak dobry jest nasz estymator?

#OBCIĄŻENIE ESTYMATORA:

$$E_F(\hat{\theta}(F_n)) - \theta(F) \text{ , gdzie } E_F(X) = \int_{-\infty}^{\infty} xf(x)dx \text{ , } f(x) \text{ gęstość } F$$

#WARIANCJA: Z jaką zmiennością od próby do próby należy się liczyć

$$Var = E_F\left[(\hat{\theta} - E_F(\hat{\theta}))^2\right] .$$

#METODA BOOTSTRAP

Dysponujemy próbą $X^n = \{x_1,...,x_n\}$ z F .

1. Liczymy $\hat{\theta}' = \theta(F_n)$ jako nasze wstępne oszacowanie $\theta(F)$
2. Niech $X^{n^*} = \{x_1^*,...,x_n^*\}$ oznacza bootstrapową próbę pobraną **<u>ze zwracaniem</u>** z $X^n = \{x_1,...,x_n\}$
   - Szacujemy parametr θ za pomocą $\hat{\theta}^*(F_n^*)$
   - Generujemy K prób bootstrapowych i liczymy na ich podstawie K oszacowań
     $$\theta^{1*}(F_n^{1*}) \text{ , } \theta^{2*}(F_n^{2*}),...,\theta^{K*}(F_n^{K*})$$
   - Bootstrapowe przybliżenie obciążenia i wariancji estymatora $\hat{\theta}$ wynoszą

$$Bias(\hat{\theta}) = \tilde{\theta}^* - \hat{\theta}' \text{ , gdzie } \tilde{\theta}^* = \frac{1}{K}\sum_{i=1}^{K}\theta^{i*} ,$$

$$Var(\hat{\theta}) = \frac{1}{K-1}\sum_{i=1}^{K}\left(\theta^{i*} - \tilde{\theta}^*\right)^2 .$$

```
dane<-c(2,3,3,5,6,6,2,7,1,4)
xx<-mean(dane)
x1<-sample(dane,10,replace=TRUE)
x2<-sample(dane,10,replace=TRUE)
x3<-sample(dane,10,replace=TRUE)
```

```
x4<-sample(dane,10,replace=TRUE)
data
x1
x2
x3
x4
xb<-(mean(x1)+mean(x2)+mean(x3)+mean(x4))/4
xb-xx

bias.boot.mean<-function(dane,k){
resb<-c()
for (i in 1:k)
resb[i]<-mean(sample(dane,length(dane),replace=TRUE))
bias<-mean(dane)-mean(resb)
return(bias)
}
bias.boot.mean(rnorm(10),200)
bias.boot.mean(rnorm(100),200)
bias.boot.mean(rnorm(1000),200)

var.boot.mean<-function(dane,k){
mb<-(mean(dane)-bias.boot.mean(dane,k))
resvar<-c()
for (i in 1:k)
resvar[i]<-(mean(sample(dane,length(dane),replace=TRUE))-(mb+mean(dane)))^2
varb<-sum(resvar)/(k-1)
return(varb)
}
```

```
#zwiększamy wielkość próby#
sqrt(var.boot.mean(rt(10,3,3),200))
sqrt(var.boot.mean(rt(100,3,3),200))
sqrt(var.boot.mean(rt(1000,3,3),200))

#zwiększamy liczę powtórzeń bootstrapowych#
sqrt(var.boot.mean(rt(30,3,3),50))
sqrt(var.boot.mean(rt(30,3,3),100))
sqrt(var.boot.mean(rt(30,3,3),300))
sqrt(var.boot.mean(rt(30,3,3),500))
```

```
#NIECO BARDZIEJ PROFESJONALNIE#
#PRZYKŁAD K  #STOSUJEMY PAKIET boot  (musimy go sobie doinstalować)#
library(boot)
meaan<-function(data,indeksy){d<-data[indeksy]  #funkcja do liczenia średniej#
m<-sum(d)/length(indeksy)
return(m)}
data<-c(1,3,4,4,9,6,7,2,3,4,5)
#data<-c(1,3,4,4,9,6,7,2,3,4,55)# #zbiór do zastanowienia się#
#data<-rnorm(30,2,3)
#data<-rexp(100,10)
W<-boot(data,meaan,R=999)
W$t[1:20,]
W
plot(W)
```

```
meed<-function(data,indeksy){d<-data[indeksy] #funkcja do liczenia mediany #

m<-median(d)

return(m)}


W1<-boot(data,meed,R=999)

W1$t[1:20,]

W1

plot(W1)
```

Literatura uzupełniająca:

1. Kosiorowski, D., (2012a) Wstęp do statystyki odpornej Kurs z wykorzystaniem środowiska R, Wydawnictwo UEK w Krakowie, Kraków.

2. Kosiorowski, D. (2012b), Generalizations of a Boxplot in a Decision Support Process, in: Knowledge-Economy-Society, Transfer of Knowledge in the Contemporary Economy, red. P. Lula, B. Mikuła, A. Jaki, Fundation of the Cracow University of Economics, str. 271 – 283

3. Kosiorowski, D. Jaśko, P. (2016a) Prognozowanie warunkowej kowariancji z wykorzystaniem odpornych estymatorów rozrzutu MCD i PCS w analizie portfelowej, Przegląd Statystyczny R. LXIII – ZESZYT 2 – 2016,149 – 171.

4. Kosiorowski, D., Jaśko, P. (2016b) Prognozowanie warunkowej kowariancji z wykorzystaniem odpornych estymatorów rozrzutu MCD i PCS w analizie portfelowej, Raport Techniczny Katedry Statystyki Uniwersytetu Ekonomicznego w Krakowie 2/2016.

5. Kosiorowski, D., Rydlewski, J. P., Snarska M., (2017). Detecting a structural change in functional time series using local Wilcoxon statistic. *Statistical Papers*. doi: 10.1007/s00362-017-0891-y

6. Kosiorowski D, Zawadzki Z (2014, 2017) DepthProc: An R package for robust exploration of multidimensional economic phenomena. http://arxiv.org/pdf/1408.4542.pdf. (pakiet DepthProc 2.0 do ściągnięcia z CRAN)

7. Kosiorowska, E., Kosiorowski, D., Zawadzki, Z. (2015) Evaluation Of The Fourth Millennium Development Goal Realisation Using Robust And Nonparametric Tools Offered By The Data Depth Concept, Folia Oeconomica Stetinensia 15(23)/1, 34-52.