Wprowadzenie do regresji odpornej

Zajmiemy się teraz pewnymi statystykami, które są użyteczne, gdy dane zawierają wartości odstające. Oznaczmy medianę próby przez M_n, a średnią próby przez: $\overline{X}_n$. Wiemy, że: gdy $X_1,...,X_n$ są niezależnymi zmiennymi losowymi pobranymi z rozkładu symetrycznego o ograniczonej wariancji, to (z MPWL) $\overline{X}_n \to \mu$ i ponadto poprzez centralne twierdzenie graniczne

$$\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \xrightarrow{d} N(0,1).$$

Okazuje się, że mediana z próby również zbiega do mediany populacji, która w przypadku rozkładów symetrycznych jest średnią. Mówiąc konkretnie, dla rozkładu normalnego mamy:

$$\sqrt{n}\frac{M_n - \mu}{\sigma} \xrightarrow{d} N(0, \tfrac{\pi}{2}).$$

Przypomnijmy sobie *ARE*, czyli asymptotyczną względną efektywność. Jest to wariancja drugiego argumentu podzielona przez wariancję pierwszego, tj. $ARE(\overline{X}_n, M_n) = \frac{\pi}{2} > 1$. Im wyższe *ARE*, tym wiemy, że $\overline{X}_n$ jest lepszym estymatorem niż M_n dla rozkładu normalnego, ponieważ asymptotycznie ma mniejszą wariancję, chociaż oba zbiegają się do μ.

Kilka przykładów ARE

Weźmy rozkład Cauchy'ego z parametrami μ i σ. Przypomnijmy, że rozkład Cauchy'ego ma bardzo ciężkie ogony, a jego średnia i wariancja nie istnieją, wiec μ i σ są teraz odpowiednio parametrami położenia i skali. Załóżmy, że mamy: $X_1,...,X_n \sim \text{Cauchy}(\mu, \sigma^2)$. Mamy:

$$\overline{X}_n \sim \text{Cauchy}(\mu, \sigma^2) \text{ (przez funkcję charakterystyczną)}$$

$$M_n \sim N(\mu, (\tfrac{\pi}{4})^2 \tfrac{\sigma^2}{n})$$

$$ARE(\overline{X}_n, M_n) = (\tfrac{\pi}{4})^2 \tfrac{1}{n} \to 0$$

Rozważmy teraz rozkład *t*-Studenta z ν stopniami swobody

ν	$ARE(\overline{X}_n, M_n)$
≤ 2	0
3	0.62
4	0.89
5	1.041

W przypadku małych wartości ν rozkład *t* ma ciężkie ogony, natomiast wraz ze wzrostem liczby stopni swobody *t* zbiega się do rozkładu normalnego. Widzimy więc, że w przypadku rozkładu symetrycznego o grubych ogonach mediana jest lepszym estymatorem parametru położenia niż średnia.

Podejście heurystyczne

Zaczniemy od kilku dobrze znanych heurystyk do szacowania parametru położenia, gdy dane mają wartości odstające.

Def . α-średnia przycięta. Średnia α-przycięta najpierw porządkuje punkty danych, a następnie przycina część z obu ogonów i oblicza średnią z pozostałych punktów danych.

Def. α-średnia winsorowana. Usuwa α-część punktów danych z górnego i dolnego ogona i zastępuje je najbliższymi punktami danych, tj. najbardziej ekstremalnymi z pozostałych punktów danych.

Weźmy na przykład zbiór danych {2; 4; 5; 10; 100}.

20% średnia przycięta: (4 + 5 + 10)/3 = 6.33

20% średnia winsorowana: (4 + 4 + 5 + 10 + 10)/5 = 6.6

Lokalne i globalne miary odporności.

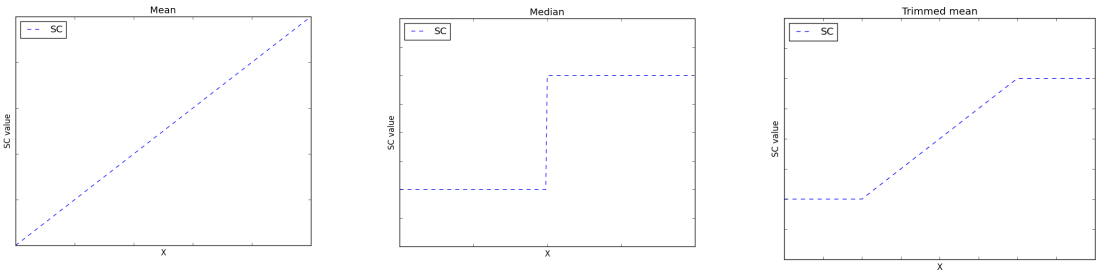
Def. Krzywa wrażliwości (*Sensitivity Curve*) mierzy wpływ zmiany jednego punktu danych na rozważany zbiór danych

$$SC_n(x,T) = \frac{T(X_1,\ldots X_{n-1},x) - T(X_1,\ldots X_{n-1})}{1/n}$$

Intuicyjnie *SC* jest analogonem populacyjnej pochodnej Gateaux względem rozkładu funkcjonału od dystrybuanty.

Jak wygląda krzywa wrażliwości SC dla średniej?

$$SC_n(x,\overline{X}_n) = \frac{((n-1)\overline{X}_{n-1} + x)/n - \overline{X}_{n-1})}{1/n} = x - \overline{X}_{n-1}$$



Na każdym z tych rysunków środek osi *x* jest prawdziwym parametrem lokalizacji.

(Mean). Zmiana dowolnego elementu wpływa na średnią. Im większa wartość *x* usuniętego elementu, tym większy wpływ na średnią.

(Median). Teraz rozważmy przypadek mediany. Dodanie lub usunięcie pojedynczego elementu przesuwa medianę do sąsiedniego elementu.

(Trimmed mean). W przypadku średniej przyciętej krzywa wrażliwości zachowuje się tak samo jak krzywa wrażliwości dla średniej, dopóki nie zaczniemy usuwać elementów, które znajdują się w αn elementach na odrzuconych końcach. W tym momencie krzywa staje się płaska.

Def. Funkcja wpływu to: $IF(x,T,F) = \lim_{n\to\infty} SC_n(x,T_n)$. Na przykład, gdy T_n jest średnią próby, mamy $IF(x,T,F) = x - T(F)$, gdzie $T(F)$ jest średnią populacji (jeśli średnia rzeczywiście istnieje). Jeśli funkcja wpływu statystyki jest ograniczona, to statystykę nazywamy odporną.

Do tej pory zajmowaliśmy się lokalnym pojęciom odporności.

Teraz rozważymy globalne pojęcia odporności.

Najpierw wprowadzimy obciążenie statystyki z próby jako:

$$bias(m,T_n,Z) = \sup_{Z'} |T_n(Z') - T_n(Z)|,$$

gdzie Z' oznacza zaburzony zbiór danych uzyskany przez zastąpienie m punktów danych w Z dowolnymi punktami danych. Tutaj $\sup_{Z'}$ można postrzegać jako grę dla dwóch graczy, w której jeden podaje drugiemu m danych z rozważanego zbioru danych , a drugi daje zwraca inne m danych tworząc zaburzony zbiór danych, który może sprawić, że statystyka będzie się różnić najbardziej jak to tylko jest możliwe od statystyki obliczonej z nieuszkodzonego zbioru danych.

Def. Punkt załamania (Breakdown point) jest zdefiniowany jako

$$\varepsilon^*(T_n,Z) = \min(\tfrac{m}{n} : bias(m,T_n,Z) = \infty \ .$$

Intuicyjnie, punkt załamania to ułamek punktów, które należy zastąpić w zbiorze danych , aby obciążenie statystyki stało się nieskończone. Dla średniej jest to asymptotycznie 0, dla mediany jest to 1/2, a dla średniej α-obciętej jest to asymptotycznie α .

M-estymatory lub ML (Maximum Likelihood) type estimators

Przypomnijmy estymację ML (największej wiarygodności) parametru położenia

$$X_1,...,X_n \sim f(X,\theta)$$

$$L(X_1,...,X_n;\theta) = \prod_{i=1}^{n} f(X_i,\theta)$$

$$\hat{\theta} \leftarrow \arg\min \sum_{i=1}^{n} -\log(f(X_i,\theta))$$

$$\frac{\partial L}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log(f(X_i,\theta)) = 0$$

W ostatnim wierszu po prostu przyrównujemy funkcję wynikową (Fisher score function) do zera i rozwiązujemy równanie względem $\hat{\theta}$. Jak się okazuje, M- estymatory uogólniają MLE. Zamiast minimalizować score function $-\log(f(X_i,\theta))$ zminimalizujemy teraz funkcję straty $\rho(X_i,\theta)$.

Niech $\psi(z) = \dfrac{d\rho(z)}{dz}$. Ta funkcja zastąpi funkcję wynikową. Wobec tego M-estymacja polega na

$$\hat{\theta} \leftarrow \arg\min \sum_{i=1}^{n} \rho(X_i, \theta)$$

$$\sum_{i=1}^{n} \psi(X_i, \theta) = 0 .$$

Rozsądna funkcja ρ powinna mieć następujące własności:

- $\rho(e) \geq 0$
- $\rho(0) = 0$
- $\rho(e) = \rho(-e)$
- $\rho(e_i) \geq \rho(e_{i'})$ dla $|e_i| > |e_{i'}|$

Przykłady. Jeśli $\hat{\theta}$ jest średnią próbkową, to otrzymamy ją przyjmując kwadratową funkcję straty $\rho(x) = x^2$. Medianę natomiast otrzymujemy przyjmując $\rho(x) = |x|$. Rzeczywiście

$$\rho(X, \theta) = (X - \theta)^2, \psi(X, \theta) = 2(X - \theta) \rightarrow \sum_i (X_i - \theta) = 0 \rightarrow \hat{\theta} = \overline{X}_n$$

$$\rho(X, \theta) = |X - \theta|, \psi(X, \theta) = \text{sign}(X - \theta) \rightarrow \sum_i \text{sign}(X_i - \theta) = 0 \rightarrow \hat{\theta} = M_n$$

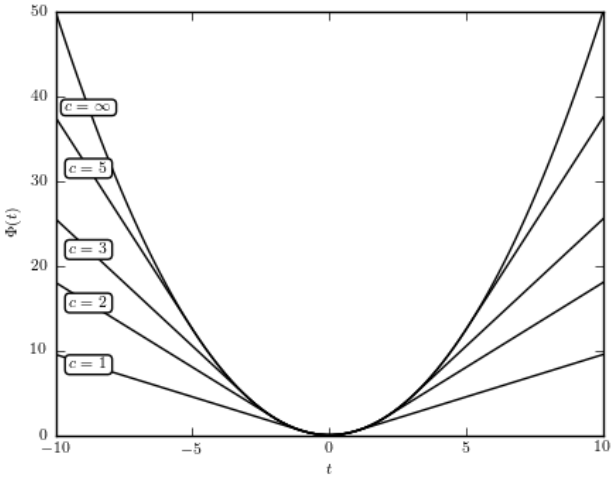
Okazuje się, że funkcja wpływu M-estymatora jest proporcjonalna do jego funkcji ψ w punkcie x. Tak więc, dopóki funkcja $\psi(x)$ jest ograniczona dla wszystkich x, mamy estymator odporny.

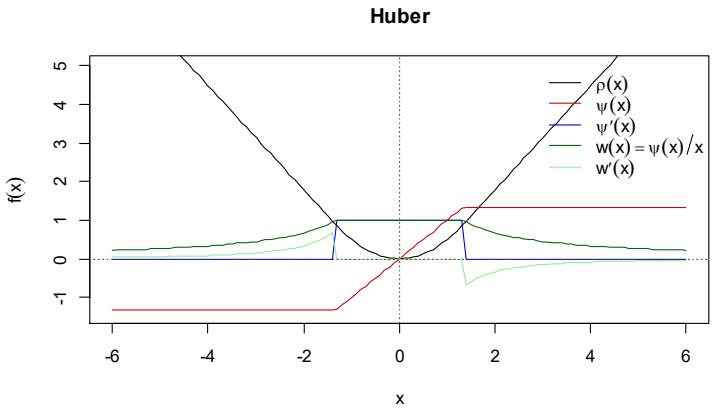
Funkcja straty Hubera o łączy stratę kwadratową i stratę modułową w sposób adaptacyjny.

$$\rho(x) = \begin{cases} \frac{1}{2} x^2; & |x| \leq c \\ c(|x| - \frac{c}{2}); & |x| > c \end{cases}$$

$$\psi(x) = \begin{cases} x; & |x| \leq c \\ c\,\text{sign}(x); & |x| > c \end{cases}$$

Intuicyjnie, gdy punkty danych nie są zbyt duże, funkcja straty Hubera karze jak strata kwadratowa, w przeciwnym razie karze jak strata modułowa. Parametr c to parametr dostrajający wybierany na podstawie danych. Sam Huber lubił ustawiać c na 1,345, co prowadzi do 95% efektywności, jeśli $X_1, \ldots, X_n \sim N(0,1)$.

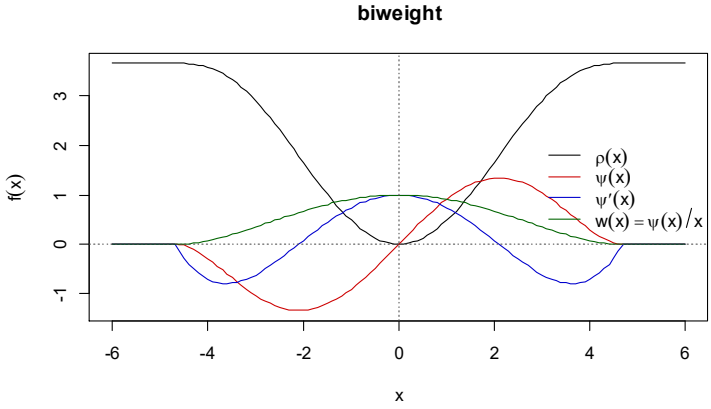




Tukey bisquare (biweight)

$$\rho(x) = \frac{c^2}{6}\begin{cases} 1-(1-(\frac{x}{c})^2)^3; \mid x \mid \leq c \\ 1 \quad ; \mid x \mid > c \end{cases}$$

$$\psi(x) = \begin{cases} x(1-(\frac{x}{c})^2)^2; \mid x \mid \leq c \\ 0 \quad ; \mid x \mid > c \end{cases}$$



Inne funkcje straty zob. psi_function.pdf

Do tej pory zakładaliśmy niejawnie, że znamy parametr skali lub wariancję. Ale co, jeśli trzeba to ten parametr oszacować? Wartości odstające będą miały znacznie większy wpływ na wariancję niż średnią, więc nie możemy użyć odchylenia standardowego próby. Zamiast tego możemy użyć estymatora MAD: mediana odchylenia bezwzględnego. Właściwie jest to bardziej MADAM (Median Absolute Deviation Around the Median), mediana odchylenia bezwzględnego wokół mediany i użyć M- estymatorów do przeskalowanych wyników, tzn.

$$\hat{\mu} \leftarrow \arg\min_{\mu} \sum_{i=1}^{n} \rho(\tfrac{X_i-\mu}{s})$$

MLE dla gęstości postaci $s^{-1} f(\tfrac{x-\mu}{s})$ prowadzi do równania dla parametru skali

$$\sum_{i=1}^{n} \psi(\tfrac{x_i-\mu}{s})(\tfrac{x_i-\mu}{s}) = n$$

które nie jest "odporne" (i prowadzi do estymatora obciążonego dla rozkładu normalnego). Modyfikujemy więc to równanie zastępując je równaniem

$$\sum_{i=1}^{n} \chi(\tfrac{x_i-\mu}{s}) = (n-1)\gamma\,,$$

z ograniczoną funkcją χ i parametrem γ zapewniającym zgodność dla rozkładu normalnego (więc $\gamma = E_{\chi}(N)$. Przykładem jest "Huber proposal 2" $\chi(x) = \psi(x)^2 = \min(|x|, c)^2$

W regresji liniowej rozkład estymatora najmniejszych kwadratów (OLS) parametru położenia jest analogiczny do rozkładu średniej próbki: kilka skrajnych przypadków może w dużym stopniu określić wartość estymatora OLS. Oczywiście dostępne są metody diagnostyczne do wykrywania potencjalnie wpływowych punktów. Inne podejście polega na zastąpieniu zwykłej metody najmniejszych kwadratów metodą odporną, na którą w mniejszym stopniu wpływają punkty odstające i wpływowe, a zatem może dawać użyteczne wyniki poprzez uwzględnienie niezgodnych danych.

Rozważamy model liniowy, który zapisujemy jako

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i .$$

Zakładamy, że model liniowy jest prawidłowy, ale rozkład błędów może mieć „ciężkie ogony” (*heavy-tailed*) i produkować wiele odstających obserwacji (*outliers*). Mając estymator $\mathbf{b}$ parametru $\boldsymbol{\beta}$ mamy oceny

$$\hat{y}_i = b_0 + b_1 x_{i1} + \cdots + b_p x_{ip} = \mathbf{x}_i^T \mathbf{b}$$

i reszty $e_i = y_i - \hat{y}_i = y_i - \mathbf{x}_i^T \mathbf{b}$.

Jeżeli założymy, że funkcja gęstości rozkładu błędu jest skalowana tzn. $s^{-1} f(\frac{e}{s})$ i przyjmiemy $\rho = -\log f$ estymator największej wiarygodności minimalizuje

$$\sum_{i=1}^{n} \rho\left(\frac{y_i - \mathbf{x}_i^T \mathbf{b}}{s}\right) + n \log s .$$

Jeśli parametr skali s jest znany i przyjmując $\psi = \rho'$, to estymator $\mathbf{b}$ parametru $\boldsymbol{\beta}$ jest rozwiązaniem nieliniowego układu równań $\sum_{i=1}^{n} \psi\left(\frac{y_i - \mathbf{x}_i^T \mathbf{b}}{s}\right) \mathbf{x}_i^T = \mathbf{0}$.

Aby ułatwić obliczenia, uczynimy to ostatnie równanie podobnym do równań estymujących dla znanego problemu, takiego jak ważona metoda najmniejszych kwadratów. W tym celu zdefiniujemy funkcję wagi

$$w_i = w\left(\frac{y_i - \mathbf{x}_i^T \mathbf{b}}{s}\right) = \frac{\psi\left(\frac{y_i - \mathbf{x}_i^T \mathbf{b}}{s}\right)}{\frac{y_i - \mathbf{x}_i^T \mathbf{b}}{s}},$$

a równanie $\sum_{i=1}^{n} \psi\left(\frac{y_i - \mathbf{x}_i^T \mathbf{b}}{s}\right) \mathbf{x}_i^T = \mathbf{0}$ można zapisać w postaci

$$\sum_{i=1}^{n} w_i \left(\frac{y_i - \mathbf{x}_i^T \mathbf{b}}{s}\right) \mathbf{x}_i^T = \mathbf{0} .$$

Problem jest nieco skomplikowany, bo wagi zależą od reszt, reszty zależą od estymowanych współczynników, a estymowane współczynniki zależą od wag. Iteracyjne rozwiązanie tego problemu zwane IRLS (*Iteratively Rewighted Least Squares*) przebiega następująco:

1. Wybrać początkowe oszacowanie $\mathbf{b}^{(0)}$, np. takie jak z OLS
2. W każdej iteracji t wyznaczyć reszty $e_i^{(t-1)}$ i odpowiadające im wagi $w_i^{(t-1)} = w[\frac{e_i^{(t-1)}}{s}]$ z poprzedniej iteracji
3. Wyznaczyć nowe oszacowanie UMNK postaci $\mathbf{b}^{(t)} = (\mathbf{X}^T\mathbf{W}^{(t-1)}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{(t-1)}\mathbf{y}$, gdzie $\mathbf{X}$ jest macierzą planu eksperymentu , $\mathbf{W}^{(t-1)} = \mathrm{diag}\{w_i^{(t-1)}\}$ jest aktualną macierzą wag

Powtarzać punkty 2 i 3 aż do momentu zbieżności wektora współczynników. Asymptotyczna macierz kowariancji współczynników dana jest wzorem

$$V(\mathbf{b}) = \frac{E(\psi^2)}{[E(\psi')]^2}(\mathbf{X}^T\mathbf{X})^{-1}.$$

Alternatywnie możemy estymować parametr skali s tak jak w przypadku estymacji ML

$$\sum_{i=1}^{n}\psi(\tfrac{y_i-\mathbf{x}_i^T\mathbf{b}}{s})(\tfrac{y_i-\mathbf{x}_i^T\mathbf{b}}{s}) = n$$

które nie jest "odporne" (i prowadzi do estymatora obciążonego dla rozkładu normalnego).
Modyfikujemy więc to równanie zastępując je równaniem

$$\sum_{i=1}^{n}\chi(\tfrac{y_i-\mathbf{x}_i^T\mathbf{b}}{s}) = (n-p)\gamma,$$

z ograniczoną funkcją χ i parametrem γ zapewniającym zgodność dla rozkładu. Przykładem jest "Huber proposal 2" $\chi(x) = \psi(x)^2 = \min(|x|, c)^2$

Dlaczego warto rozważyć stosowanie odpornych metod estymacji

1. Użytkownicy, nawet eksperci w dziedzinie statystyki, nie zawsze sprawdzają dane.
2. Ostra decyzja o utrzymaniu lub odrzuceniu obserwacji jest marnotrawstwem. Możemy zrobić lepiej, zmniejszając wagę wątpliwych obserwacji, niż je odrzucając, chociaż możemy chcieć odrzucić całkowicie błędne obserwacje.
3. Wykrycie wartości odstających w danych wielowymiarowych lub o dużej strukturze może być trudne lub nawet niemożliwe.
4. Odrzucenie wartości odstających wpływa na rozkład, który należy skorygować. W szczególności wariancje będą niedoszacowane na podstawie „oczyszczonych” danych

Regresja odporna a regresja oporna (Robust vs resistant regression)

Koncepcja regresji resistant dotyczy dobrego zachowania się modelu regresji w przypadku obecności nieprawidłowych punktów danych. Regresja oporna ma wysoki punkt załamania, zbliżający się do 50%. Rozważamy zastąpienie najmniejszych kwadratów jedną z metod

- LMS- (Least Median of Squares) najmniejsza mediana kwadratów: zminimalizuj medianę kwadratów reszt $\min_b \text{median}_i |y_i - \mathbf{x}_i^T\mathbf{b}|^2$

  Bardziej ogólnie, dla LQS zminimalizuj pewien kwantyl (powiedzmy 80%) kwadratów reszt.

- LTS - Least Trimmed Squares) najmniejsze przycięte kwadraty: zminimalizuj sumę kwadratów dla q reszt. Początkowo q obejmowało nieco ponad 50%, ale S-PLUS przeszło na 90%. $\min_b \sum_{i=1}^{q} |y_i - \mathbf{x}_i^T\mathbf{b}|^2_{(i)}$

Jest wiele funkcji w R które realizują estymację odporną robust i resistant. W pakiecie MASS

rlm()

lqs()

Szczegóły plik- Robust.pdf

Plik Robust regresion.R

```
library(MASS)
#Vanebles&Ripley
help(phones)
ph<-as.data.frame(phones)

phones.lm<-lm(calls~year,data=phones)
attach(phones)
plot(year,calls)
detach(phones)
abline(phones.lm$coef) #model dopasowany klasyczną MNK (OLS)
help(rlm)

abline(rlm(calls~year,phones,maxit=50),lty=2,col=2)
abline(lqs(calls~year,phones),lty=3,col=3)
#wykonując poniższą instrukcję należy wskazać na rysunku gdzie umieścić legendę
#funkcja locator(1) odczytuje pozycję kursora gdy wciśniemy lewy przycisk myszki
legend(locator(1),lty=1:3,col=1:3,legend=c("least squares","M-estimate","LTS"))

summary(lm(calls~year,data=phones))

summary(rlm(calls~year,phones,maxit=50))
summary(rlm(calls~year,scale.est="proposal 2",phones,maxit=50))
summary(rlm(calls~year,psi=psi.bisquare,phones,maxit=50))

#zob. plik Robust.pdf
lqs(calls~year,data=phones)#domyślnie method="lts"
```

```
lqs(calls~year,data=phones,method="lms")#least median of squares
lqs(calls~year,data=phones,method="lts")#least trimmed squares
lqs(calls~year,data=phones,method="S")
summary(rlm(calls~year,data=phones,method="MM"))#kombinacja resistant lqs i efektywnośco M estymatorów
psi.huber(x, k = 1.345, deriv = 0)
psi.hampel(u, a = 2, b = 4, c = 8, deriv = 0)
psi.bisquare(u, c = 4.685, deriv = 0)



x<-seq(-6,6,by=0.1)
plot(x, psi.huber(x)*x, lwd=2, col='orange', type='l')
plot(x, psi.huber(x), lwd=2, col='orange', type='l')
plot(x, psi.hampel(x)*x, lwd=2, col='red', type='l')
plot(x, psi.bisquare(x)*x, lwd=2, col='blue', type='l')
library(robustbase)
plot(huberPsi, x, ylim=c(-1.4, 5), leg.loc="topright", main=FALSE)
title(main="Huber")
plot(hampelPsi, x, ylim=c(-1.4, 5), leg.loc="topright", main=FALSE)
title(main="Hampel")
source(system.file("xtraR/plot-psiFun.R", package = "robustbase", mustWork=TRUE))
p.psiFun(x, "biweight", par = 4.685)
title(main="biweight")

#Zbiór z wieloma odstającymi obserwacjami
library(faraway)
data(star)
plot (star$temp, star$light, xlab="log(Temperature)", ylab="log(Light Intensity)")
ga <- lm (light~ temp, star)
abline (ga)
#Czy są odstające obserwacje?
range (rstudent (ga))
ga <- lm (light ~ temp, data=star, subset=( temp > 3.6))
abline (ga, lty=2)
ga_rob <- rlm (light ~ temp, data=star, maxit=50)
abline(rlm(light ~ temp, data=star, maxit=50),col="blue")
abline(lqs(light ~ temp, data=star,method="S"),col="red")
```

Przedstawiamy dwa zestawy danych pochodzące z chemii analitycznej (Abbey, 1988; Analytical Methods Committee, 1989a,b). Zestaw danych abbey zawiera 31 oznaczeń zawartości niklu ($\mu g\ g{-}1$) w SY-3, kanadyjskiej skale syenitowej, a chem zawiera 24 oznaczenia miedzi ($\mu g\ g{-}1$) w mące pełnoziarnistej. Dane te są częścią większego badania, które sugeruje $\mu = 3{,}68$.

```
#przykład 1
sort(chem)
mean(chem)
plot(density(chem))
median(chem)
mad(chem) #estymator MAD parametru skali
unlist(huber(chem)) #odporna estymacja parametru położenia z użuciem estymatora skali MAD
unlist(hubers(chem))#odporna estymacja parametru położenia i skali (Huber Proposal 2)

#przykład 2
sort(abbey)
```

```
mean(abbey)
plot(density(abbey))
median(abbey)
mad(abbey) #estymator MAD parametru skali
unlist(huber(abbey)) #odporna estymacja parametru położenia z użuciem estymatora skali MAD
unlist(hubers(abbey))#odporna estymacja parametru położenia i skali (Huber Proposal 2)
```