METODI INFORMATICI PER DATA SCIENCE
OUTLIERS

Vedremo

- individuazione Outliers cross-section
- individuazione Outliers per serie storiche

Introduzione

Outlier: ‘lies outside’

- ○ un’osservazione che non si adatta bene ad un modello
- ○ un’osservazione che non è vicino al centro dei dati

- **Outlier univariati**, quando si tratta di una sola variabile
- **Outlier in un modello**, riferito ad un insieme di variabili

Outliers per serie storiche R Packages:

- library(**univOutl**)
- Hidiroglou-Berthelot (1986) metodo implementato in library(**univOutl**)
- **alcune funzioni per lm()**
- library(**tsoutliers**)

Cross section

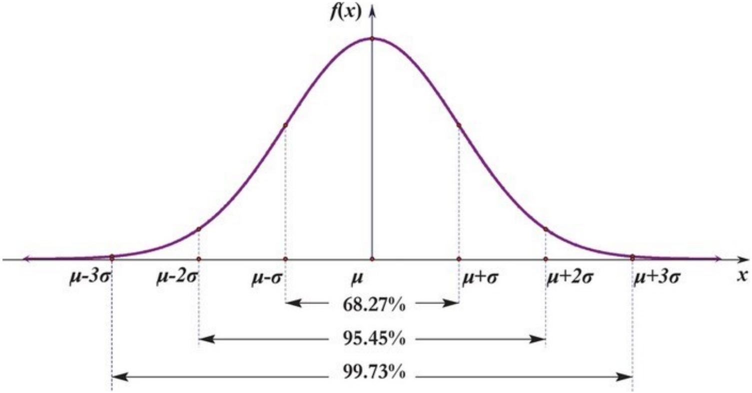
- Location and Scale-based intervals (principalmente riferiti alla distribuzione gaussiana)
- Metodo del Boxplot

Riferimenti:

- D'Orazio M. (2017). univOutl: Detection of Univariate Outliers.
  R package. version 0.2
  https://CRAN.R-project.org/package=univOutl

- Istat, CBS, SFSO and Eurostat (2007) Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys. Manual prepared by the EDIMBUS Project. http://ec.europa.eu/eurostat/documents/64157/4374310/30-Recommended+Practices-for-editing-and-imputation-in-cross- sectional-business-surveys-2008.pdf

Filtro di Hampel



- Outlier: osservazioni fuori intervallo $[\tilde{\mu} - k \times \tilde{\sigma}, \tilde{\mu} + k \times \tilde{\sigma}]$
- $\tilde{\mu}$ e $\tilde{\sigma}$ stime robuste di μ e σ, rispettivamente
- $k = 2, 2.5, 3$
- $\tilde{\mu} =$ mediana $= Q_{0.50}$ (max breakpoint del 50%)

Stima robusta σ

- $\tilde{\sigma} = IQR/a = (Q_{0.75} - Q_{0.25})/a$
- $\tilde{\sigma} = MAD = b \times med|x_i - med(x_i)|$
- $\tilde{\sigma} = S_n = c \times med_i\{med_j|x_i - x_j|\}$

Distribuzione gaussiana: a = 1.349, b = 1.4826, c = 1.1926

Con distribuzioni asimmetriche:

- trasformazione dei dati (log, Box-Cox)
- intervalli asimmetrici: $[\tilde{\mu} - k \times \tilde{\sigma}_L, \tilde{\mu} + k \times \tilde{\sigma}_R]$

$$\sigma_L = \frac{Q_2 - Q_1}{0.6745} \quad \sigma_R = \frac{Q_3 - Q_2}{0.6745} \tag{1}$$

Libreria univOutl

```
> library(univOutl)
> set.seed(123)
> x <- rnorm(30, 0, 1)
> x[5] <- -5
> x[15] <- 10
>out <- LocScaleB(x = x, k = 3, method='MAD')
No. of outliers in left tail: 1
No. of outliers in right tail: 1
> out$pars
      median        scale
 -0.07373326   1.06024220
>out$bounds
lower.low   upper.up
-3.254460   3.106993
>out$outliers
[1]  5 15
> x[out$outliers]
[1] -5  10
```

Metodo basato sul Boxplot

Outlier: osservazioni fuori intervallo $[f_l, f_u]$ (**f dette fence**)

- Tradizionale

$$f_l = Q_1 - k \times IQR \quad f_u = Q_3 + k \times IQR \tag{2}$$

- Fences asimmetrici (leggera asimmetria):

$$f_l = Q_1 - 2 * k \times (Q_2 - Q_1) \quad f_u = Q_3 + 2 * k \times (Q_3 - Q_2) \quad (3)$$

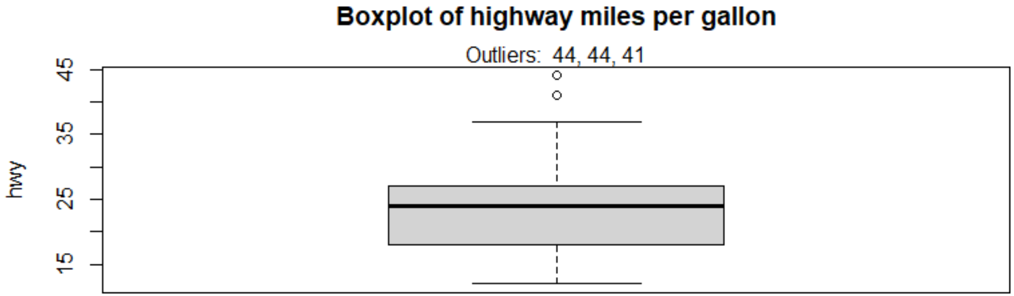
- Skewness-adjusted (moderata asimmetria, $-0.6 \leq M \leq 0.6$):

$$f_l = Q_1 - 1.5 \times e^{aM} \times IQR \quad f_u = Q_3 + 1.5 \times e^{bM} \times IQR \quad (4)$$

M è la misura della coppia media dell'asimmetria, quando $M > 0$ allora $a = -4$ e $b = 3$ ($a = -3$ e $b = 4$ con $M < 0$) (Vanderviere and Huber, 2008)

Metodo basato sul Boxplot in R

```
out <- boxplot.stats(mpg$hwy)$out
boxplot(mpg$hwy,ylab = "hwy",
  main = "Boxplot of highway miles per gallon")
mtext(paste("Outliers: ", paste(out, collapse = ", ")))
```



```
> library(univOutl)
> #1) method="resistant" 'standard' boxplot fences
>b <- boxB(mpg$hwy,k=1.5,method="resistant")
No. of outliers in left tail: 0
No. of outliers in right tail: 3
>mpg$hwy[b$outliers]
[1] 44 44 41
> #2) method="asymmetric" modifica standard
> # method per tener conto di (moderatemente) dati asimmetrici;
>b <- boxB(mpg$hwy,k=1.5,method="asymmetric")
No. of outliers in left tail: 0
No. of outliers in right tail: 4
>mpg$hwy[b$outliers]
[1] 37 44 44 41
> #3) method="adjbox" usa Hubert and Vandervieren (2008)
> #adjusted boxplot per distribuzioni asimmetriche
>b <- boxB(mpg$hwy,k=1.5,method="adjbox")
The MedCouple skewness measure is: -0.25
No. of outliers in left tail: 0
No. of outliers in right tail: 15
> mpg$hwy[b$outliers]
```

Tasso di crescita: $t_1 = 1$ e $t_2 = 2$

$t_1 = 1$	$t_2 = 1$
y_{11}	y_{12}
$\ldots$	$\ldots$
y_{i1}	y_{12}
$\ldots$	$\ldots$
y_{n1}	y_{n2}

Rilevazione di valori anomali sui rapporti $r_i = y_{i2}/y_{i1}$

Metodo di Hidiroglou-Berthelot (1986) per identificare gli outlier in variabili osservate in istanti diversi consiste nel derivare una variabile di punteggio basata sui rapporti r_i

Hidiroglou-Berthelot: Algoritmo

- All'inizio i rapporti sono centrati sulla loro mediana (r_M):

$$s_i = \begin{cases} 1 - r_M/r_i & \text{se } 0 < r_i < r_M \\ r_M/r_i - 1 & \text{se } r_i \geq r_M \end{cases}$$

- Quindi, per tenere conto della grandezza dei dati, si ricava il seguente punteggio:

$$E_i = s_i max(y_{i1}, y_{i2})^U \quad 0 \leq U \leq 1 \quad \text{usually} \quad (U = 0.5)$$

- Infine, l'intervallo è calcolato come:

$$(E_M - C \times d_{Q_1}, E_M + C \times d_{Q_3}) \qquad (5)$$

dove $d_{Q_1} = max(E_M - E_{Q_1}, |A \times E_M|)$ e $d_{Q_3} = max(E_{Q_3} - E_M, |A \times E_M|)$ con E_M, E_{Q_1}, E_{Q_3} i quartili degli E scores quando pct = 0.25 (default), A = 0.05 e $C \geq 4$ (di solito)

Library univOutl: Metodo di Hidiroglou-Berthelot

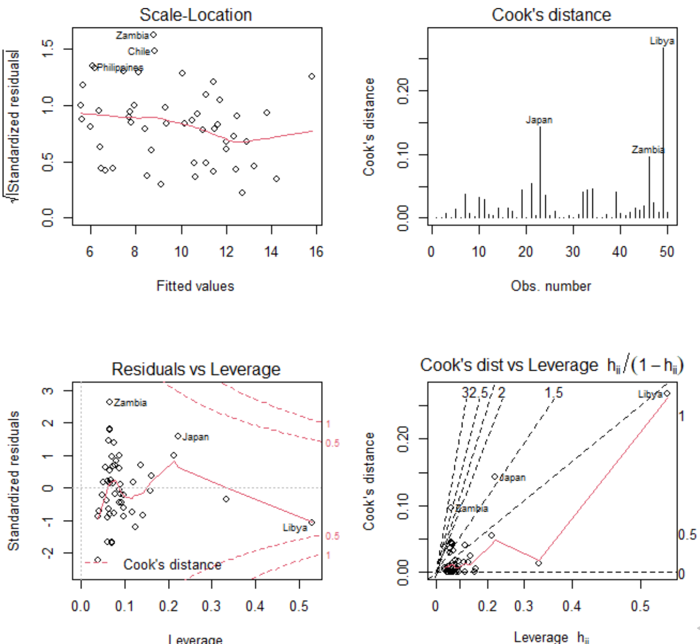
HBmethod(yt1, yt2, U=0.5, A=0.05, C=4, pct=0.25, id=NULL, std.score=FALSE, return.dataframe=FALSE, adjboxE=FALSE)

```
# genera dei dati
> set.seed(222)
> x0 <- rnorm(30, 50, 5)
> set.seed(333)
> rr <- runif(30, 0.9, 1.2)
> rr[10] <- 2
> x1 <- x0 * rr
> # run HBmethod with argument return.dataframe = TRUE
> out <- HBmethod(yt1 = x0, yt2 = x1,
+    return.dataframe = TRUE)
MedCouple skewness measure of E scores: 0.0637
Outliers found in the left tail: 0
Outliers found in the right tail: 1
```

Outlier in modelli

```
lm.SR <- lm(sr ~ pop15 + pop75 + dpi + ddpi,
data = LifeCycleSavings)
inflm.SR <- influence.measures(lm.SR)
plot(lm.SR,3)
plot(lm.SR,4)
plot(lm.SR,5)
plot(lm.SR,6)
```

Outlier in modelli grafico



Outlier e Serie storiche

- ESS guidelines su seasonal adjustment (Eurostat, 2015) https://ec.europa.eu/eurostat/documents/3859598/6830795/KS-GQ-15-001-EN-N.pdf
- Chen, C. and Liu, L. (2013) ‘Joint Estimation of Model Parameters and Outlier Effects in Time Series Joint Estimation of Model Parameters and Outlier Effects in Time Series’, 88(421), pp. 284–297

ESS guidelines sulla destagionalizzazione

Gli outlier sono valori anormali nella serie. Possono essere modellati in diversi modi:

- **Valori anomali additivi** (valori anomali in punti isolati della serie)
- **Temporary changes** (serie di valori anomali con effetti temporaneamente decrescenti sul livello della serie)
- **Level shifts** (serie di valori anomali con un effetto costante a lungo termine sul livello delle serie)
- **Ramps** (che descrivono una transizione graduale, lineare o quadratica tra due punti temporali a differenza del brusco cambiamento associato agli spostamenti di livello)
- **Spostamenti di livello temporanei** (dove lo spostamento di livello ha un effetto a breve termine piuttosto che a lungo termine)

ESS guidelines sulla destagionalizzazione: Opzioni

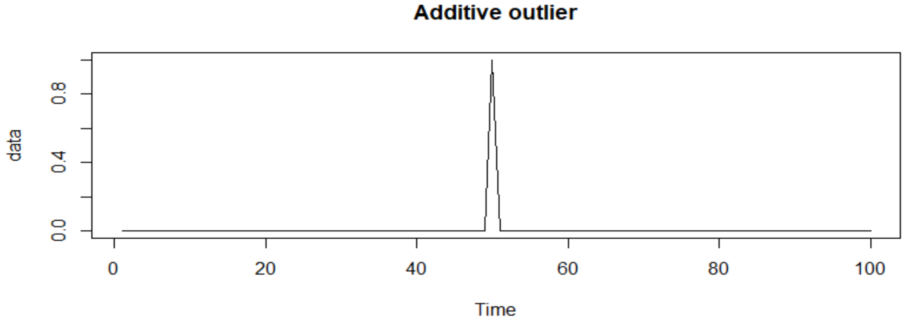
A) Le serie dovrebbero essere controllate per valori anomali di diverso tipo. Una volta identificati, i valori anomali causati da errori di dati dovrebbero essere corretti nei dati (grezzi) non rettificati prima del pretrattamento. I valori anomali rimanenti dovrebbero essere spiegati/modellati utilizzando tutte le informazioni disponibili. I valori anomali per i quali esiste una chiara interpretazione (es. scioperi, conseguenze di cambiamenti nella politica di governo, cambiamenti di territorio che interessano paesi o aree economiche, ecc.) sono inclusi come regressori nel modello, anche se i loro effetti sono leggermente inferiori alla soglia di significatività generale

B) Come A), ma con una procedura completamente automatica per rilevare e correggere i valori anomali

C) Nessun trattamento preliminare dei valori anomali

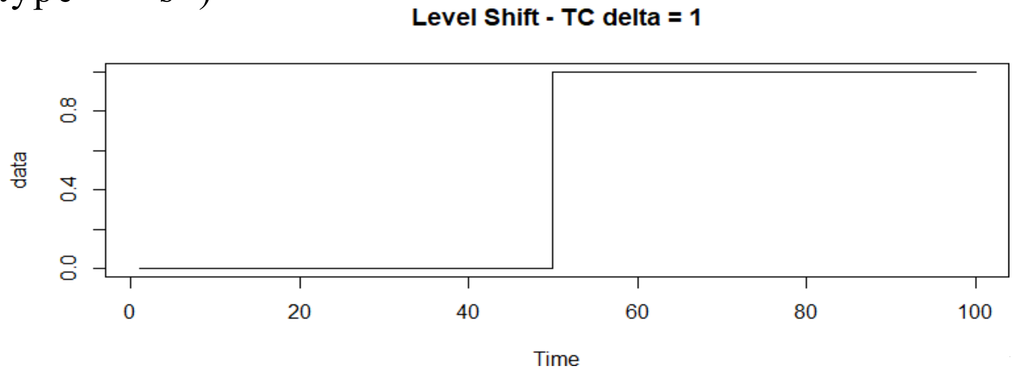
Outlier Additivo (AO)

```
a <- rep(0, 100)
a[50] <- 1
ao <- filter(a, filter = 0, method = "recursive")
par(mfrow=c(1,1))
plot(ao, ylab= "data",main = "Additive outlier", type = "l
```



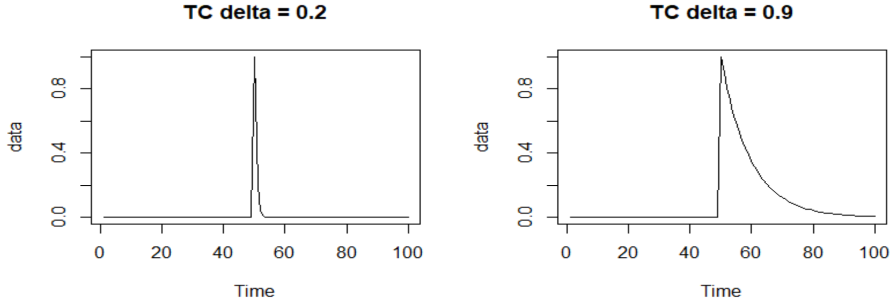
Level shift (LS)

```
par(mfrow=c(1,1))
ls <- filter(a, filter = 1, method = "recursive")
plot(ls, ylab= "data",main = "Level Shift - TC delta = 1",
 type = "s")
```



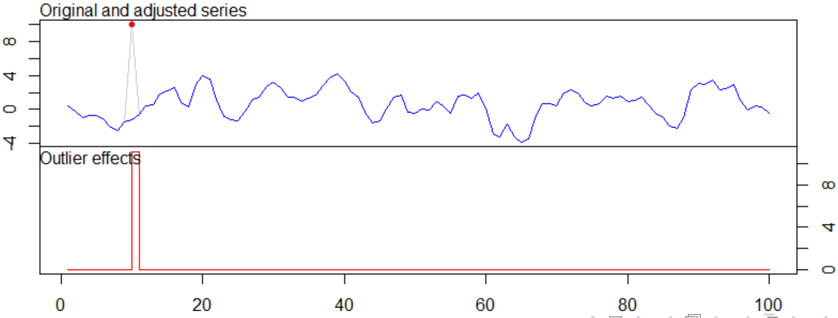
Temporary change (TC)

```
a_0_2 <- filter(a, filter = 0.2, method = "recursive")
a_0_9 <- filter(a, filter = 0.9, method = "recursive")
par(mfrow=c(1,2))
plot(a_0_2, ylab= "data",main = "TC delta = 0.2")
plot(a_0_9, ylab= "data",main = "TC delta = 0.9")
```



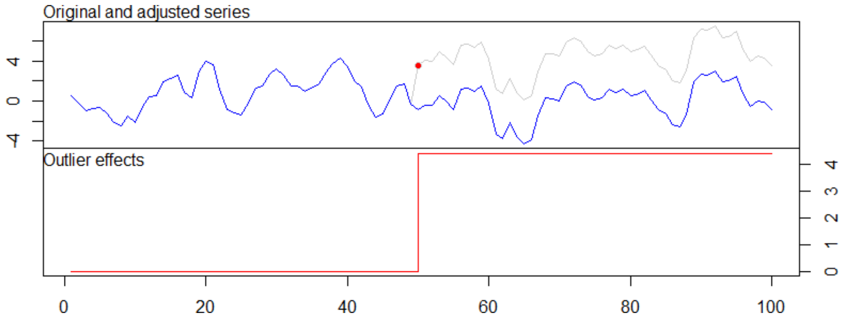
Come individuare AO

```
library(tsoutliers)
set.seed(12345)
y=arima.sim(model=list(ar=.8,ma=.5),n.start=158,n=100)
y[10]=10
y=ts(y,freq=1,start=1)
plot(y,type='o')
b <- tso(y)
plot(b)
```



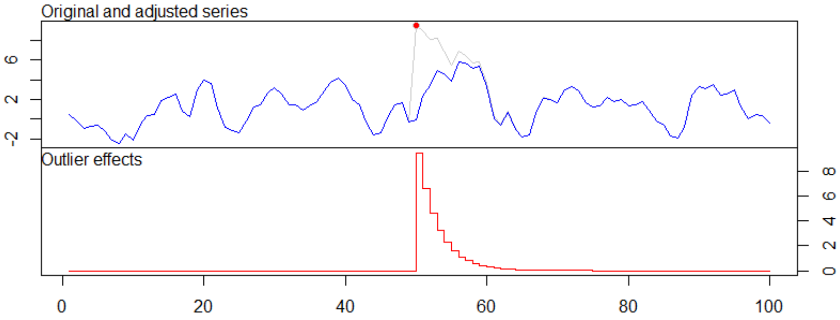
Come individuare LS

```
library(tsoutliers)
set.seed(12345)
y=arima.sim(model=list(ar=.8,ma=.5),n.start=158,n=100)
z <- y+4*ls
b<-tso(z)
plot(b)
```



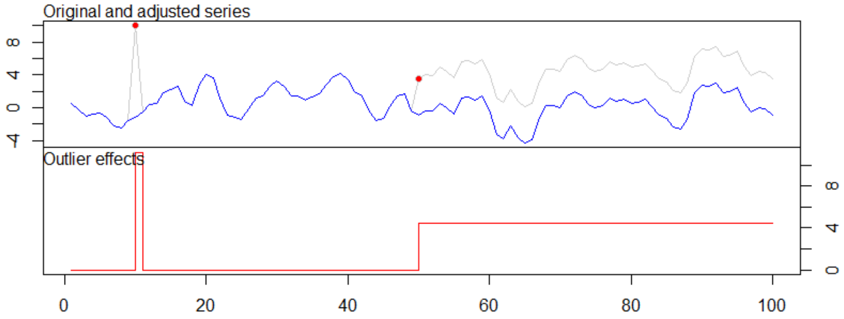
Come individuare TC

```
library(tsoutliers)
set.seed(12345)
y=arima.sim(model=list(ar=.8,ma=.5),n.start=158,n=100)
z <- y+10*a_0_9
b<-tso(z)
plot(b)
```



Come individuare AO e LS

```
library(tsoutliers)
set.seed(12345)
y=arima.sim(model=list(ar=.8,ma=.5),n.start=158,n=100)
y[10] <- 10
z <- y+4*ls
b<-tso(z)
plot(b)
```



Riepilogo e conclusioni finali

- individuazione Outliers cross-section
- individuazione Outliers per serie storiche