BIBLIOTECA DEI SAPERI

Metodi informatici per Data Science Outliers

Introduzione (Slide 1)

Benvenuti!
In questa lezione affronteremo i diversi approcci per rilevare i valori anomali in R, da tecniche semplici come le statistiche descrittive a tecniche più formali. Sebbene non esista una regola rigorosa o univoca in merito alla rimozione o meno dei valori anomali dal dataset prima di eseguire analisi statistiche, è abbastanza comune almeno rimuovere o imputare i valori anomali dovuti a un errore sperimentale o di misurazione.
Non risolveremo il problema se sia opportuno rimuovere i valori anomali o meno (né se attribuirli con la mediana, la media, la modalità o qualsiasi altro valore), ma capiremo come rilevarli nel caso i nostri dati siano di tipo cross-section o in serie storica. In alcuni casi, i valori anomali vengono mantenuti perché contengono informazioni preziose. Succede anche che le analisi vengano eseguite due volte, una con e una senza valori anomali per valutarne l'impatto sulle conclusioni.

Slide 2 Introduzione

Iniziamo da una semplice definizione di valore anomalo (outlier, ovvero in inglese 'lies outside'). In particolare, può essere definita outlier un'osservazione che non si adatta bene ad un modello o anche un'osservazione che non è vicino al centro dei dati.

Un'altra distinzione può essere considerata tra Outlier univariati (quando si tratta di una sola variabile) o Outlier in un modello (se riferito ad un insieme di variabili). Per il primo caso vedremo la libreria di R univOutl, per il secondo alcuni funzioni per lm(). Faremo degli esempi per serie storiche, utilizzando invece la libreria tsoutliers.

Slide 3: Cross section

Iniziamo dai modi in cui possono essere individuati i valori anomali in dati cross-section, descrivendo due gruppi di tecniche, e cioè:

1) quelli basati su **Location and Scale-based intervals** (principalmente riferiti alla distribuzione gaussiana)
2) e il **Metodo del Boxplot**

Due importanti riferimenti sono il **Manuale della libreria univOutl**, scritta da D'Orazio e delle **Linee guida** preparate nell'ambito delle attività di un progetto europeo **EDIMBUS** per il problema generale dell'editing e imputazione di dati mancanti.

Slide 4 Filtro di Hampel

Un primo metodo, noto come **Filtro di Hampel**, consiste nel considerare come valori anomali i valori al di fuori dell'intervallo centrato su una stima robusta della media della distribuzione (ad esempio la mediana) più o meno k volte una stima robusta della variabilità della distribuzione. K può essere 2 2,5 o 3.

Slide 5 Stima robusta di sigma

Esistono diverse stime robuste della variabilità. Tali stime possono differire in base alla distribuzione dei dati ed inoltre possono tener conto se questa è per esempio asimmetrica. In tal caso è possibile altrimenti operare una trasformazione dei dati.

Slide 6 Libreria univOUtl

Il metodo presentato nelle 2 precedenti slide è implementato nella funzione LocScaleB nella libreria univOutl. Per mostrarne l'output di tale funzione abbiamo creato come input un vettore simulato di 30 osservazioni dalla distribuzione normale standardizzata perturbato con l'introduzione di 2 valori anomali in corrispondenza della quinta e quindicesima osservazione con valori uguali a -5 e 10 rispettivamente. La funzione restituisce come output il vettore pars con la stima dei parametri di scala e mediana. Bounds che contiene il limite inferiore e superiore dell'intervallo al di fuori del quale le osservazioni sono considerati valori anomali. Inoltre, restituisce in outliers i valori anomali.

Slide 7 Metodo del boxplot

Il Boxplot è un grafico statistico che si utilizza per variabili quantitative. È molto utile per capire se la distribuzione è simmetrica, oppure asimmetrica e per confrontare la forma di più distribuzioni. Ma soprattutto permette di identificare in modo rapido e preciso valori anomali.

Il Boxplot permette di rappresentare sullo stesso grafico cinque tra le misure di posizione più utilizzate in statistica:

- il valore minimo
- il primo quartile (Q1)
- la mediana (Q2)
- il terzo quartile (Q3)
- ed il valore massimo di una variabile

Definendo come misura di dispersione lo scarto interquartilico (IQR) possiamo individuare come anomale le osservazioni fuori dall'intervallo definito dagli estremi definiti nell'equazione 2 (con k in genere uguale a 1,5). Le formule 3 e 4 mostrano alcune varianti della regola per distribuzioni moderatamente e fortemente asimmetriche.

Slide 8 Metodo del boxplot in R

Per produrre un Boxplot il cui nome completo è "**box and whiskers plot**", che in italiano è spesso tradotto come "diagramma a scatola e baffi" in R può essere utilizzata la funzione boxplot. La scatola (il box) è compresa tra il primo e terzo quartile e mostra l'ampiezza della metà centrale della distribuzione. L'altezza della scatola è infatti pari al range interquartile (IQR) e contiene il 50% centrale delle osservazioni, quelle comprese tra il primo ed il terzo quartile. La linea all'interno della scatola invece rappresenta la mediana. I due segmenti che partono dalla scatola e si prolungano verso l'alto e verso il basso sono detti "baffi". I baffi indicano la dispersione dei valori inferiori al primo quartile e superiori al terzo quartile non classificati come outliers. Estraendo dall'output della funzione boxplot l'oggetto stats si possono salvare esattamente quali sono le osservazioni anomale.

Slide 9 Metodo basato sul boxplot nella libreria univOutl

Utilizzando la funzione boxB nella libreria univOutl siamo in grado di utilizzare anche le regole che nelle precedenti formule abbiamo descritto per variabili con distribuzioni asimmetriche e personalizzare la regola con diverse scelte ad esempio della costante k.

Slide 10 Tasso di crescita: $t_1=1$ e $t_2=2$

Supponiamo ora di svolgere un'indagine campionaria finalizzata al calcolo della variazione dell'ammontare di una certa variabile quantitativa y osservata ad un tempo t_1 ed un tempo t_2. Il metodo, proposto originariamente da **Hidiroglou e Berthelot** (1986) per l'individuazione di dati anomali si basa sul ricorso a soglie di accettazione per i tassi di variazione r_i calcolati per ogni unità.

Slide 11 Hidiroglou-Berthelot: Algoritmo

Per trattare in modo equilibrato sia le variazioni positive che quelle negative, consideriamo la trasformazione degli r_i in nuovi valori s_i dove r_M è la mediana dei rapporti. In tal modo, metà dei valori di s_i risultano positivi e l'altra metà negativi. Per tener conto della diversa “grandezza” delle osservazioni, i valori s_i vengono trasformati in nuovi valori Ei dove U è un parametro compreso tra 0 e 1. E_i è detto effetto associato all'i-esima unità, e l'esponente U consente di controllare l'importanza da associare alla "grandezza" dell'unità stessa

In pratica, questa trasformazione consente di dare più importanza a relativamente piccole variazioni in grandi valori rispetto a grandi variazioni in piccoli valori. In tal modo è poi possibile definire i limiti inferiore e superiore della regione di accettazione della formula (5) in cui la costante C permette di ampliare o restringere l'ampiezza dell'intervallo di accettazione. Tutte le unità in cui i corrispondenti effetti Ei assumono valori esterni all’intervallo sono da considerare outlier.

Slide 12 Library univOutl: metodo di Hidiroglou-Berthelot

Il metodo di Hidiroglou-Berthelot è implementato nella funzione HBMethod della libreria univOutl. In particolare, in questo esempio il metodo è applicato a dei dati simulando in un vettore x0 le osservazioni al primo istante e calcolando quelle al secondo instante nel vettore x1 ottenute applicando al vettore x0 i tassi di crescita in rr per cui la decima osservazione è stata perturbata.

Slide 13 Outlier in modelli

Influence.measure() è un insieme di funzioni che può essere utilizzata per calcolare alcune diagnostiche di regressione per valutare la presenza di osservazioni anomale o influenti quando si è stimato un modello di regressione.

Slide 14 Outlier in modelli grafico

Consente di produrre molto facilmente dei grafici per alcune di queste misure di influenza, come ad esempio la **Distanza di Cook**, che misura quanto cambia l'intera funzione di regressione quando viene eliminato l'i-esima osservazione. O il **Grafico dei residui standardizzati** che permette di evidenziare le osservazioni che hanno un valore alto di tale residuo.

Slide 15 Outlier e Serie storiche

La presenza di Outlier, ossia di osservazioni anomale nella serie storica causate da eventi straordinari, può determinare significative distorsioni nella stima dei coefficienti dei modelli per serie storica. In questa parte faremo riferimento a due principali fonti, e cioè:;

- le **Linee guida di Eurostat** sulla destagionalizzazione che dedica ampio spazio al problema degli outlier
- e l’**Articolo di Chen e Liu** le cui tecniche sono implementate nel pacchetto R che presesenteremo.

Slide 16 ESS guidelines sulla destagionalizzazione

Le **Linee guida Eurostat** descrivono i metodi con cui possono essere modellati i valori anomali fornendo anche una tassonomia nota nella letteratura sulle Serie storiche. In particolare:

- abbiamo valori anomali additivi (valori anomali in punti isolati della serie)
- oppure temporary changes (serie di valori anomali con effetti temporaneamente decrescenti sul livello della serie)
- poi level shifts (serie di valori anomali con un effetto costante a lungo termine sul livello delle serie)
- o anche ramps (che descrivono una transizione graduale, lineare o quadratica tra due punti temporali a differenza del brusco cambiamento associato agli spostamenti di livello)
- ed infine spostamenti di livello temporanei (dove lo spostamento di livello ha un effetto a breve termine piuttosto che a lungo termine)

Slide 17 Le Linee guida, inoltre, forniscono tre opzioni per il trattamento dei dati anomali che vanno da quella da evitare che non richiede nessun trattamento (opzione C) ad una opzione A per la quale le Serie dovrebbero essere controllate per valori anomali di diverso tipo (che vedremo dopo). Una volta identificati, i valori anomali causati da errori di dati dovrebbero essere corretti nei dati (grezzi). 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 variabili esplicative nel modello. Un'opzione intermedia B richiede una procedura completamente automatica per rilevare e correggere i valori anomali.

Slide 18 Outlier Additivo

La **Funzione impulso** è utile per rappresentare valori anomali additivi. Può essere pensato come un caso speciale del modello del temporary changes con delta = 0 (vedremo più avanti il modello del temporary change). Possiamo rappresentarlo graficamente con l'aiuto della funzione filter().

Slide 19 Level shift

Una **Funzione step** è utile per rappresentare i valori anomali dello spostamento di livello. Può essere pensato come un caso speciale del modello del temporary changes con delta = 1 (di nuovo vedremo più avanti il modello del cambiamento transitorio). Anche qui possiamo rappresentarlo con l'aiuto della funzione filter().

Slide 20 Temporary change

Una **Funzione di modifica transitoria** è utile per rappresentare valori anomali temporary change. Possiamo rappresentarlo graficamente con l'aiuto della funzione filter(). Vengono considerati due valori di delta per mostrare come la variazione transitoria varia di conseguenza.

Slide 21 Come individuare AO

Con l'aiuto della **Funzione tso()** all'interno del pacchetto tsoutliers, identifichiamo se sono presenti valori anomali in alcune serie simulate con la funzione arima.sim(). In particolare generiamo una serie storica y in cui alla decima osservazione sostituiamo 10. La funzione tso identifica tale outlier come mostrato dal grafico dell'output salvato in b della funzione.

Slide 22 Come individuare LS

In questo esempio la serie storica simulata y è perturbata con un cambio di livello pari a 4. La **Funzione tso** identifica tale outlier come mostrato dal grafico dell'output salvato in b della funzione.

Slide 23 Come individuare TC

Ricordando l'oggetto a_09 creato precedentemente, in questo esempio la Serie storica simulata y è perturbata introducendo un valore anomalo temporary change con delta=0.9. La **funzione tso** identifica tale outlier come mostrato dal grafico dell'output salvato in b della funzione.

Slide 24 Come individuare AO e LS

In molti casi, le Serie storiche possono presentare più valori anomali e di diverso tipo, come per la Serie simulata in questo esempio. Se utilizziamo la procedura implementata in tso, in questo esempio vengono individuati un outlier di tipo additivo e uno level shift.

Slide 25 Conclusioni

Bene, siamo giunti alla fine di questa video lezione.

Ti ricordo che abbiamo approfondito diversi approcci utili per rilevare i valori anomali R da tecniche semplici come le statistiche descrittive a tecniche più formali.

Grazie per l'attenzione!