Documentazione della qualità e qualità della documentazione

Premessa

L’esistenza e l’accessibilità di un’appropriata documentazione rappresenta un aspetto fondamentale per la gestione della qualità in ambito statistico e per la diffusione dell’informazione prodotta.

In particolare:

- consente agli utilizzatori interni ed esterni una corretta interpretazione dei dati e la valutazione della loro affidabilità;
- permette l’implementazione dei sistemi di gestione della qualità dei processi e dei prodotti.

Documentazione della qualità e qualità della documentazione

L’attività di documentazione

La predisposizione della documentazione richiede un notevole impegno da parte di un’organizzazione che preliminarmente deve sviluppare una strategia generale per la compilazione della documentazione, in cui gli obiettivi siano ben definiti, le priorità assegnate e che permetta di soddisfare tutte le molteplici necessità degli utilizzatori.

In particolare dovranno essere definite caratteristiche della documentazione quali:

- contenuto;
- livello di dettaglio;
- formato e modalità di diffusione.

È evidente che tale attività di specificazione dovrà tenere conto delle diverse potenziali tipologie di utilizzatori.

Documentazione della qualità e qualità della documentazione

I destinatari della documentazione

- Utente generico esterno
- Utente specializzato esterno
- Utente interno all'organizzazione (supervisore, progettista di indagine, responsabile di un settore di produzione dei dati, l'esperto di ingegnerizzazione di processi)

La documentazione sarà molto dettagliata e riguarderà tutti gli aspetti del processo di produzione (definizioni e classificazioni, strumenti di indagine, metodi, risultati, fonti di errore e misure degli errori campionari e non campionari) nel caso in cui sia rivolta agli utenti interni all'organizzazione.

L'utente esperto esterno sarà più interessato ai contenuti dell'indagine, alla comparabilità con altre fonti e ad una valutazione generale sulla qualità.

Per l'utilizzatore generico la documentazione dovrà essere sintetica, generale e chiara, con una terminologia comprensibile a tutti.

Documentazione della qualità e qualità della documentazione

La distinzione degli utilizzatori è anche utile per definire le forme di diffusione della stessa e dei dati.

Altri aspetti che dovrebbero guidare nella costruzione della documentazione riguardano:

- i costi (includendo lavoro e tempi di realizzazione);
- la raccolta e l'aggiornamento delle informazioni riportate;
- il grado di comparabilità e di standardizzazione;
- l'accessibilità da parte degli utilizzatori.

Lo sviluppo delle tecnologie informatiche nell'ambito della strutturazione dei sistemi informativi ha dato un contributo sostanziale nell'organizzazione coerente e standardizzata delle informazioni garantendo, tra l'altro, una maggiore facilità all'accesso dei dati di interesse.

Documentazione della qualità e qualità della documentazione

La lista di verifica per la documentazione dell'attività statistica

È uno strumento di supporto per documentare, seguendo uno schema strutturato e standard, le modalità di progettazione ed esecuzione dell'indagine e gli indicatori di qualità calcolati.

Consiste in un insieme di domande a risposte aperte per ciascuna fase di un processo statistico.

Ha una struttura modulare per consentire di documentare le diverse tipologie di processi produttivi e di produrre note metodologiche più o meno dettagliate, attraverso una pre-selezione di quesiti, in funzione delle esigenze conoscitive dei diversi profili dei destinatari della documentazione.

Documentazione della qualità e qualità della documentazione

Esiste anche una versione informatizzata della lista che agevola l'attività di documentazione consentendo di precostituire il percorso da seguire dopo aver scelto il profilo informativo e quello di utenza.

Nella versione informatizzata è possibile creare delle liste di verifica personalizzate selezionando i quesiti di interesse.Viene, inoltre, consentito il salvataggio in formato testo dei documenti compilati.

La lista di verifica può essere proficuamente utilizzata anche per la redazione del documento di progettazione e per il monitoraggio dell'esecuzione del processo costituendo, in questo caso, un valido ausilio per l'autovalutazione.

Documentazione della qualità e qualità della documentazione

Le integrazioni più recenti:

•domande relative alle rilevazioni di fonte amministrativa che consentono di descrivere sia il processo amministrativo alla base della raccolta dei dati (archivi amministrativi utilizzati, informazioni relative agli atti amministrativi all'origine dell'archivio, controlli di qualità effettuati dall'ente titolare del processo amministrativo) sia le analisi svolte per l'uso a fini statistici dei dati amministrativi (valenza statistica dell'archivio amministrativo, modalità di trasferimento dei dati raccolti con l'atto amministrativo);

•domande inerenti le elaborazioni su dati statistici che consentono di descrivere le fonti utilizzate, le metodologie di elaborazione dei dati e di integrazione fra le fonti, gli studi e gli eventuali interventi correttivi per migliorare la qualità dei dati di base;

•glossario contenente le definizioni dei principali termini utilizzati avente la duplice finalità di chiarire il significato di alcuni termini e di standardizzare il linguaggio fra gli addetti.

Documentazione della qualità e qualità della documentazione

Le tipologie di processo

Per l'identificazione delle varie tipologie di processo si prendono in esame:

- le modalità di acquisizione dei dati;
- la tipologia e la natura delle fonti dei dati.

Nel Programma Statistico Nazionale i processi vengono distinti in rilevazioni ed elaborazioni.

L'esplicitazione delle tipologie di processo deve essere chiara e definita a monte della documentazione del processo.

Documentazione della qualità e qualità della documentazione

I contenuti informativi per profilo di utente

Nella lista (composta da 122 quesiti) è possibile individuare dei sottoinsiemi di quesiti in base a due distinzioni:

1. tipologie di utenti destinatari della documentazione (esterni generici, esterni specializzati, interni);
2. criteri di valutazione dei quesiti (scala a tre livelli: molto rilevante, abbastanza rilevante, poco rilevante).

Esiste una gerarchia tra le categorie di utenti ed il sottoinsieme dei quesiti irrinunciabili per l'utenza generica potrebbe essere utilizzato come il set minimale con cui accompagnare le ricerche per le quali sono disponibili risorse esigue.

Documentazione della qualità e qualità della documentazione

Le informazioni fondamentali comuni alle diverse tipologie di studi

- Dati identificativi del processo di produzione dei dati
- Caratteristiche del fenomeno e della popolazione oggetto di studio
- Livello di disaggregazione e tempestività nella diffusione dei dati
- Tutela della riservatezza

Per l'utenza specialistica

- Approfondimenti sulle classificazioni
- Modalità di stratificazione e selezione del campione
- Scostamenti tra popolazione obiettivo e archivi e metodologie di integrazione fra questi
- Interventi correttivi effettuati sui dati di base e trattamento statistico degli outliers e delle mancate risposte
- Verifiche esogene sulla qualità degli archivi

Documentazione della qualità e qualità della documentazione

I quesiti rilevanti per l'organizzazione

- Indicazione dell'ufficio responsabile del processo
- Acquisizione dei dati (strutture esterne di rilevazione, team di rilevatori, modalità di formazione e supervisione)
- Revisione codifica e registrazione dei dati (interna/esterna, manuale/automatica, numero di persone utilizzate, modalità di formazione e supervisione)
- Aspetti di architettura hardware e software utilizzati

Tecniche di indagine

Le fasi operative

Con il termine generico “fasi operative” si indicano tutti gli aspetti connessi alle modalità di contatto ed alla raccolta dei dati presso le unità di rilevazione ed alla preparazione dei dati per l’analisi statistica.

Nel caso in cui la tecnica di indagine prescelta preveda l’uso di un questionario si possono distinguere le seguenti fasi:

- preparazione di supporti cartacei e/o di procedure informatizzate che siano strumentali al controllo dell’errore non campionario generabile nelle fasi operative;
- rilevazione;
- codifica dei quesiti aperti;
- registrazione dati su supporto magnetico;
- revisione automatica o interattiva.

Tecniche di indagine

Le fasi operative

La preparazione di supporti cartacei e/o di procedure informatizzate strumentali al controllo dell'errore non campionario è un'attività trasversale che potrebbe teoricamente impattare su ogni fase operativa.

Il fine di tale attività è quello di aiutare il supervisore dell'indagine a controllare gli errori non campionari verificando la coerenza delle informazioni generate dalle singole fasi operative o dai sistemi dei controlli previsti per ognuna di esse.

A seconda dei casi si potrebbe decidere di creare della modulistica e/o delle applicazioni informatiche per raccogliere informazioni aggiuntive sull'indagine.

Tecniche di indagine

Le fasi operative

La rilevazione ha tre obiettivi fondamentali:

1. individuare l'unità di rilevazione e convincerla a partecipare all'indagine;
2. raccogliere i dati in modo neutrale (non influenzando il rispondente);
3. rendere l'esperienza gradevole per il rispondente al fine di facilitare eventuali contatti futuri (indagini longitudinali; indagini ripetute; ritorni sul campo; indagini di controllo).

Il raggiungimento di tali obiettivi può essere agevolato preparando con cura tutte le attività connesse a tale fase e creando un ambiente favorevole in cui vi sia una convinta partecipazione all'indagine da parte dei soggetti interessati.

Tecniche di indagine

Le fasi operative

In sostanza, è necessario curare con estrema attenzione la redazione del questionario e la creazione di eventuali modelli ausiliari e/o procedure informatizzate finalizzati ad agevolare le operazioni di:

- contatto delle unità di rilevazione e di gestione della raccolta;
- tempistica ed interazione fra le strutture preposte alla rilevazione;
- campagne di sensibilizzazione dei rispondenti;
- formazione del personale;
- supervisione delle operazioni e recupero delle informazioni incomplete;
- controllo dell’errore non campionario.

Tecniche di indagine

Le fasi operative

Nella fase di codifica dei quesiti aperti si procede all'associazione alle informazioni pervenute sotto forma di linguaggio libero di un insieme finito di codici corrispondenti ad una classificazione precostituita (ad esempio quella ATECO o ICD-9).

E' un operazione molto complessa e richiede l'impiego di codificatori esperti. Il codificatore ha, in generale, la possibilità di consultare le liste di codici su un computer e sfruttare quindi tutte le possibilità di ricerca offerte da tale mezzo.

Gli errori che si commettono più frequentemente riguardano l'interpretazione della descrizione, la trascrizione del codice, inadeguatezza della classificazione stessa, difficoltà a ricercare il codice, condizioni lavorative.

Tecniche di indagine

Le fasi operative

La fase di registrazione su supporto informatico dei dati raccolti sul questionario cartaceo consiste, in generale, nell'immissione dei dati al computer da parte di un operatore che digita su una tastiera esattamente ciò che legge sul questionario.

Normalmente tale attività viene affidata a personale non specializzato e pertanto rappresenta una delle principali fonti di errore non campionario.

Frequentemente le operazioni di acquisizione dei dati su supporto informatico vengono appaltate all'esterno e questo contribuisce ad aumentare l'errore.

La scelta di modalità tecniche per l'inserimento dei dati che prevedono la creazione di maschere di acquisizione con avvisi di digitazioni non ammissibili è quella che garantisce una migliore qualità del dato registrato (ovviamente esiste sempre qualche forma di errore che non può essere evitata).

Tecniche di indagine

Le fasi operative

L’attività di registrazione può essere semplificata da una progettazione del questionario che curi la leggibilità delle risposte e dei relativi codici ad esse associati (incluse le mancate risposte ed il non pertinente) e da un tracciato record che tenga conto della variabilità del fenomeno (campo età con tre posizioni).

E’, inoltre, buona norma pre-codificare tutte le domande aperte presenti.

In indagini condotte su un numero limitato di soggetti per le quali non è accettabile nessun livello di errore è da prevedere la doppia imputazione dei dati da parte di due diversi operatori.

Tale fase è assente quando la rilevazione viene effettuata con i sistemi CA o con questionari a lettura ottica.

In quest’ultimo caso l’operatore svolge il ruolo di supervisore delle operazioni svolte dalla macchina e dal software di controllo della lettura.

Tecniche di indagine

Le fasi operative

Con la revisione automatica si procede ad identificare i valori mancanti o incongruenti delle variabili rilevate ed ad intervenire su tali valori per mezzo di procedure informatizzate.

La revisione interattiva prevede l'automazione della sola fase di individuazione dell'errore, lasciando all'operatore il compito di eseguire le correzioni al terminale.

L'obiettivo che ci si propone con tale attività è quello di *minimizzare l'effetto degli errori riscontrati* sulle successive fasi di elaborazione e sull'informazione prodotta (vedi stime).

Definiamo errore un valore che implica la violazione di regole logico-formali (anche indicate con il termine di compatibilità) inerenti:

- il campo di variazione di una variabile (range dei codici ammissibili);
- le relazioni intercorrenti tra le variabili;
- le relazioni formali stabilite dalle norme di compilazione dei modelli cartacei.

Tecniche di indagine

Le fasi operative

Le procedure informatizzate che effettuano l'imputazione dei valori errati si classificano in base alla tipologia di errori trattati. In particolare, gli errori possono essere suddivisi in:

- errori sistematici, per i quali è possibile supporre che, per sottopopolazioni identificabili, esista un unico valore corretto con il quale effettuare l'imputazione (individuo in età inferiore ai 14 catalogato automaticamente tra la popolazione non attiva);
- errori casuali, per i quali, comunque siano identificate le sottopopolazioni di unità, ci si deve aspettare un margine di variabilità rispetto alle possibili correzioni effettuabili (incompatibilità di alcuni valori della professione con i titolo di studio).

Le tecniche di imputazione che trattano la prima tipologia di errori operano secondo l'applicazione di regole deterministiche del tipo "SE-ALLORA", mentre le seconde sostituiscono i valori errati ricorrendo a valutazioni di tipo probabilistico (da utilizzare per correggere grandi moli di dati raccolti su unità statistiche abbastanza omogenee).

Tecniche di indagine

Le fasi operative

Nell’applicare le procedure di revisione automatica bisogna avere come obiettivo iniziale quello di correggere gli errori sistematici e successivamente quelli casuali.

Con l’editing selettivo si effettuano correzioni solo sulle variabili o sulle unità statistiche ritenute più influenti sulle stime di interesse. In generale, queste tecniche sono applicate sotto forma di revisione interattiva con l’obiettivo, una volta individuato l’errore, di ritornare sul campo contattando le unità che hanno fatto registrare il problema.

In conclusione, il processo di revisione automatica può essere visto come un modo per aumentare la qualità dei dati raccolti, incorporando in essi una serie di conoscenze, esprimibili sotto forma di proposizioni logiche relative al fenomeno indagato ed al processo di produzione dell’informazione.

DATA SCREENING

Valutare la qualità e la validità dei dati in ingresso prima di procedere all'analisi

Le procedure di *data screening*:

- aiutano ad utilizzare in modo appropriato un set di dati;
- puntano ad isolare le particolarità dei dati;
- permettono di individuare gli adattamenti da implementare in prospettiva dell'applicazione delle statistiche multivariate.

La sequenza di operazioni per lo screening dei dati è, di solito, la seguente:

1. analisi dell'accuratezza dei dati attraverso lo studio delle statistiche descrittive univariate e dei grafici delle distribuzioni (valori fuori i range; deviazioni standard e medie plausibili; attribuzione corretta dei codici per l'identificazione dei dati mancanti e non pertinenti);

2. analisi della matrice di correlazione;

3. valutazione dell'ammontare e della distribuzione dei dati mancanti *(missing data)*;

4. rispetto delle assunzioni (controllo dei valori di curtosi e simmetria, probability plots; controlli per la non-linearità e l'omoschedasticità; scelta della trasformazione più idonea; controllo dei risultati dopo la trasformazione).

5. identificazione e trattazione dei casi estremi *(outliers)*
   a. "*outliers*" univariati;
   b. "*outliers*" multivariati;

6. valutazione delle variabili per la multicollinearità e singolarità.

Oss.: l'ordine con cui sono eseguite le procedure di *screening* è molto importante: le decisioni iniziali influenzano tutte le procedure successive.

Ad esempio in presenza di variabili non-normali e di casi estremi spesso si decide di trasformare le variabili o eliminare i casi.

Se si procede prima con la trasformazione si hanno alte probabilità che gli outliers diminuiscano, tuttavia se si procede prima con l'eliminazione le probabilità di trovare variabili non-normali diminuiscono.

Accuratezza dei dati

Il modo migliore per verificare l'accuratezza con cui è stato costituito il set di dati si basa essenzialmente sull'esame delle statistiche descrittive e delle rappresentazioni grafiche delle variabili effettuabile tramite l'impiego delle procedure di esplorazione statistica dei dati previste pacchetti statistici.

N.B. Nel caso in cui il data set sia costituito da pochi dati il metodo che porta ai migliori risultati è basato sul confronto diretto del file di dati con i dati originali.

Accuratezza dei dati

Il controllo di ammissibilità dei dati registrati interesserà essenzialmente aspetti quali:

•il campo di variazione delle variabili e l'eventuale correzione delle modalità che fuoriescono dall'intervallo predefinito;

•la coerenza di un dato con altre informazioni contenute nello stesso record e l'eventuale segnalazione delle modalità palesemente errate e l'eventuale correzione delle modalità che è possibile correggere;

•la verosimiglianza di osservazioni delle quali sia nota la distribuzione e che si presentino come improponibili;

•la correttezza delle codifiche utilizzate.

Correlazioni “veritiere”

La maggior parte delle procedure multivariate analizzano le **strutture delle correlazioni** tra le variabili.

E’ quindi fondamentale che le correlazioni (sia fra due variabili continue che fra una dicotomica ed una continua) siano le più accurate possibile.

Correlazioni “veritiere”

Correlazioni più grandi di quelle che si riscontrerebbero nella popolazione:

più variabili forniscono le stesse informazioni. Tale situazione si verifica ad esempio quando gli stessi “*items*” sono utilizzati in più di una variabile (scale della personalità, indici di status economico, indici che misurano lo stato di salute, ecc.). Se tra due misure composte esiste molta sovrapposizione è consigliabile introdurne una soltanto nell’analisi.

Correlazioni più piccole di quelle che si riscontrerebbero nella popolazione:

il campione considerato non è rappresentativo per la popolazione (range di casi ristretto), o si è in presenza di una distribuzione non omogenea dei casi tra le modalità di variabili dicotomiche.

Dati Mancanti

Uno dei più gravi problemi che si possono presentare durante l'analisi è la mancanza di dati per alcune variabili *(missing values).*

La serietà del problema dipende essenzialmente dalla:

- struttura dei dati mancanti: la struttura dei dati mancanti deve essere casuale, altrimenti è difficile generalizzare i risultati che si ottengono; assumere l'andamento casuale è molto rischioso, è infatti sempre necessario testarlo (un metodo consiste nel costruire una variabile *"dummy"* che separi i casi in base all'assenza o alla presenza della variabile con i dati mancanti. Individuati i gruppi si può calcolare un test della differenza fra le medie delle variabili di interesse. Se non si rilevano differenze significative allora il modo in cui si trattano i dati mancanti non è così critico, ovviamente se si escludono le analisi rispetto alla variabile "incriminata");

continua

Dati Mancanti

• ammontare dei dati mancanti: se in un campione abbastanza grande i dati mancanti sono pochi il problema è quasi irrilevante ed esistono molte tecniche risolutive; al contrario se i dati mancanti sono molti e il campione è di piccole dimensioni il problema è piuttosto rilevante e difficilmente risolvibile (sfortunatamente non esistono ancora indicazioni su quanti dati mancanti possano essere accettati per un campione di una certa dimensione).

Oss.: la scelta del metodo per trattare i dati mancanti è fondamentale poiché influisce sui risultati finali dell'analisi.

Dati Mancanti

I metodi più comunemente utilizzati per il trattamento dei dati mancanti

1. Eliminazione di casi
2. Eliminazione di variabili
3. Stima (imputazione) dei dati mancanti
4. Utilizzo del valore mancante come dato
5. Replica dell’analisi con e senza missing

Dati Mancanti

1. Eliminazione di casi

   Si eliminano i casi con i dati mancanti, questa soluzione è ottimale *sse* i dati mancanti interessano poche unità rappresentanti un sottocampione casuale dell'intero campione.

2. Eliminazione di variabili

   Si eliminano le variabili in cui c'è concentrazione di dati mancanti, ma *sse* sono in numero ridotto e poco critiche per l'analisi oppure se sono fortemente correlate con altre variabili.

Dati Mancanti

Considerazioni generali sul metodo di eliminazione di casi o di variabili

Se i dati mancanti sono distribuiti a macchia tra le unità e le variabili questa procedura può portare ad una perdita sostanziale di casi, eventualità che determina conseguenze importanti in caso:

- di ricerche dove le unità sono state raggruppate in disegni sperimentali e la perdita anche di una sola unità richiede aggiustamenti per trattare con celle di dimensioni disuguali;
- di raccolta di dati estremamente costose sia in termini economici che di tempo;
- notevole diminuzione della potenza statistica dell'analisi.

N.B. In molti pacchetti statistici l'eliminazione dei casi è l'opzione di default

Dati Mancanti

3.Stima (imputazione) dei dati mancanti

- Utilizzo delle *conoscenze a priori.*

Consiste nella sostituzione da parte del ricercatore del dato mancante con un valore ipotizzato sulla base di conoscenze approfondite.

L'utilizzo di questa procedura è ragionevole quando lo studioso ha lavorato nell'ambito della specifica area di ricerca per lungo tempo, il campione è grande ed i dati mancanti sono pochi.

Nel caso in cui il dato mancante interessi una variabile continua un'alternativa è costituita dalla trasformazione della variabile da continua a dicotomica (ad esempio "alto" e "basso") per prevedere con buona dose di certezza a quale categoria il caso possa essere attribuito. Ovviamente tale soluzione comporta una perdita di informazioni.

Nel caso di indagini longitudinali si potrebbe sostituire l'ultimo valore osservato, ovviamente se si ha l'aspettativa che non siano intervenuti cambiamenti.

continua

Dati Mancanti

- Inserimento dei *valori medi*.

Si inseriscono le medie calcolabili a partire dai dati a disposizione.

Questa procedura è molto utile in assenza di ogni altro tipo di informazione; tuttavia, spesso, si ha la diretta conseguenza di una notevole diminuzione della varianza che implica una riduzione del coefficiente di correlazione.

Tale metodo è stato molto applicato in passato. Oggi c'è la tendenza ad usare le soluzioni più convincenti messe a disposizione dai software statistici.

continua

Dati Mancanti

Oss.: le procedure appena citate sono entrambe delle soluzioni estreme, nel primo caso si dà totale libertà all'analista, mentre nel secondo caso si applica un metodo fortemente conservatore.

Una strategia alternativa è rappresentata da un compromesso tra i due metodi e precisamente l'inserimento della media del gruppo del caso con il dato mancante (la riduzione della varianza entro i gruppi potrebbe aumentare le differenze tra i gruppi in modo spurio).

continua

Dati Mancanti

- Utilizzo della *regressione*.

La stima del dato mancante è ottenuta scrivendo una equazione di regressione in cui la variabile con i casi mancanti è usata come VD e le altre sono utilizzate come VI. L'equazione è stimata sulla base dei casi completi ed è impiegata per prevedere i casi mancanti. A volte si reitera la procedura fino a quando i valori stimati nel passo precedente sono simili a quelli ottenuti (essi cioè convergono).

Questo metodo è sicuramente il più sofisticato ed è sicuramente più oggettivo di quello che si basa sull'ipotesi del ricercatore e non così cieco come il semplice inserimento della media.

continua

Dati Mancanti

Svantaggi

- si migliora in modo fittizio l'adattamento dei dati (i valori mancanti sono stati stimati a partire dalle altre variabili, pertanto è probabile che essi siano più consistenti con i punteggi delle variabili utilizzate che non i valori veri);
- si riduce la varianza poiché il valore stimato è probabilmente molto vicino al valore della media;
- il metodo "funziona" se le variabili utilizzate per la stima sono dei buoni predittori per la variabile con i valori mancanti;
- le stime ottenute sono utilizzabili sse assumono valori nel range dei valori assunti dalla variabile nei casi completi.

Dati Mancanti

- Uso di *algoritmi*.

1. Metodo della massima verosimiglianza
2. Imputazioni multiple (multiple imputation)
3. Massimizzazione delle aspettative (expectation maximization)

continua

Dati Mancanti

- Uso della *matrice di correlazione con dati mancanti*.

Si usano tutte le coppie di valori disponibili per calcolare ogni correlazione della matrice.

Ogni correlazione della matrice può quindi essere calcolata sulla base di un numero diverso e di un diverso sottoinsieme di casi a seconda della struttura dei dati mancanti.

Poiché l'errore standard della distribuzione campionaria di *r* è basata sul numero di casi, alcune correlazioni saranno meno stabili di altre nella stessa matrice di correlazione.

Questo metodo può essere applicato nel caso in cui si dispone di campioni grandi e siano presenti pochi valori mancanti.

Dati Mancanti

4.Utilizzo del *valore mancante come dato*

La presenza di dati mancanti non è sempre una situazione negativa poiché il fatto che un valore sia mancante potrebbe essere un buon predittore della variabile di interesse nella ricerca.

In genere la procedura utilizzata si basa sulla costruzione di una variabile “dummy” e sulla sostituzione del valore medio al dato mancante in modo tale che tutti i casi possano essere esaminati.

La variabile dummy è quindi usata come un’altra variabile nell’analisi.

Dati Mancanti

5.*Replica dell'analisi con e senza missing*

Dopo la scelta del metodo per il trattamento dei dati mancanti è molto importante verificarne gli effetti.

In genere questa procedura si basa sul confronto delle analisi ottenute rispettivamente sui casi completi e sui casi incompleti.

Nell'eventualità vi siano forti differenze il metodo scelto potrebbe essere poco opportuno e condurre a risultati inaffidabili.

In tal caso è necessario analizzare i motivi che hanno portato a tali differenze e valutare quale risultato si approssima di più alla realtà oppure riportare entrambi i risultati.

Dati Mancanti

Riepilogo 1

- Osservare la struttura dei missing per determinare se la loro distribuzione è casuale
- Eliminare i casi è ragionevole solo se la struttura dei missing è casuale e solo pochissimi casi sono mancanti e quei casi sono mancanti su variabili diverse. Se c'è evidenza di assenza di casualità si preferiscono i metodi che preservano tutti i casi per le analisi successive
- L'eliminazione di una variabile con molti dati mancanti è accettabile fintanto che la variabile non è critica per l'analisi. Se la variabile è importante si può usare una variabile dicotomica che codifica il fatto che i punteggi sono mancanti in aggiunta alla sostituzione del valore medio. Tale soluzione consente di analizzare tutti i casi e le variabili

Dati Mancanti

Riepilogo 2

- E’ conveniente evitare la sostituzione della media a patto che la proporzione di valori mancanti sia molto piccola e non ci siano altre opzioni disponibili
- Usare la conoscenza a priori richiede un notevole livello di certezza da parte del ricercatore circa l’ambito disciplinare allo studio e dei risultati attesi
- E’ possibile applicare metodi di regressione che non richiedono software specializzati, ma generalmente tali metodi comportano molti svantaggi
- Il metodo EM rappresenta l’approccio più semplice e più ragionevole alla sostituzione dei dati mancanti. Richiede però la disponibilità di pacchetti specializzati e la circostanza che i punteggi siano mancanti in modo casuale

Dati Mancanti

Riepilogo 3

- Il metodo della sostituzione multipla, che incorpora le procedure EM, rappresenta uno strumento più idoneo perché può essere applicato indipendentemente dalla struttura dei valori mancanti
- L'uso della matrice di correlazione con dati mancanti è accattivante perché non richiede ulteriori passaggi. Ovviamente bisogna disporre di un software che consente di inserire tale opzione e comunque è da usare quando i dati mancanti sono distribuiti equamente sulle variabili e quindi non ci sono variabili con molti dati mancanti. I problemi collegati a tale metodo sono minori se il data set è grande e il numero di missing è ridotto
- Ripetere l'analisi con e senza i missing data è altamente raccomandato indipendentemente dal metodo scelto se la proporzione di dati mancanti è alta e il data set è piccolo

OUTLIERS

Osservazioni con valori estremamente devianti dalla media, *outliers*, si trovano non di rado durante le elaborazioni di dati e spesso provocano distorsioni nella stima dei parametri. La presenza di un *outlier* può essere determinata dalle seguenti ragioni:

- scorretta introduzione dei dati;
- errore nella specificazione dei codici che indicano i dati mancanti;
- l'*outlier* non appartiene alla popolazione dalla quale si intendeva campionare;
- distribuzione nella popolazione della variabile, rispetto alla quale abbiamo trovato l'*outlier*, con valori più estremi rispetto ad una distribuzione normale.

Outliers

Gli *outliers* si distinguono in:

- Univariati: casi con valori estremi su una sola variabile;
- Multivariati: casi con valori estremi su due o più variabili.

Un caso con valore sulla variabile età pari a 15 è assolutamente plausibile così come un caso che sulla variabile reddito annuo fa registrare un valore pari a 45.000 euro. Ma un caso che presenti un valore di 15 sulla variabile età e di 45.000 su quella del reddito è decisamente inusuale ed è altamente probabile che sia un *outlier*.

Outliers

Outliers univariati su variabili dicotomiche

Casi che cadono nella classe “sbagliata” di una variabile con frequenze molto difformi delle due modalità (ricorda: variabili dicotomiche con i casi concentrati al 90% in una delle due categorie dovrebbero essere eliminate sia perché i coefficienti di correlazione fra queste variabili e le altre sono più bassi e sia perché i casi appartenenti alla modalità con frequenza minore hanno maggiore influenza di quelli appartenenti alla categoria più numerosa).

Outliers univariati su variabili continue

Casi che presentano punteggi standardizzati molto elevati. In generale punteggi superiori a 3,29 sono considerati potenziali *outliers* (la scelta della soglia va fatta anche sulla base della dimensione campionaria; infatti con campioni molto numerosi possono presentarsi solo pochi punteggi standardizzati superiori a 3,29).

Outliers

Come individuare *outliers* univariati

⇒ Esaminare le distribuzioni di frequenze semplici, i valori minimi e massimi, le medie e le deviazioni standard di ogni variabile

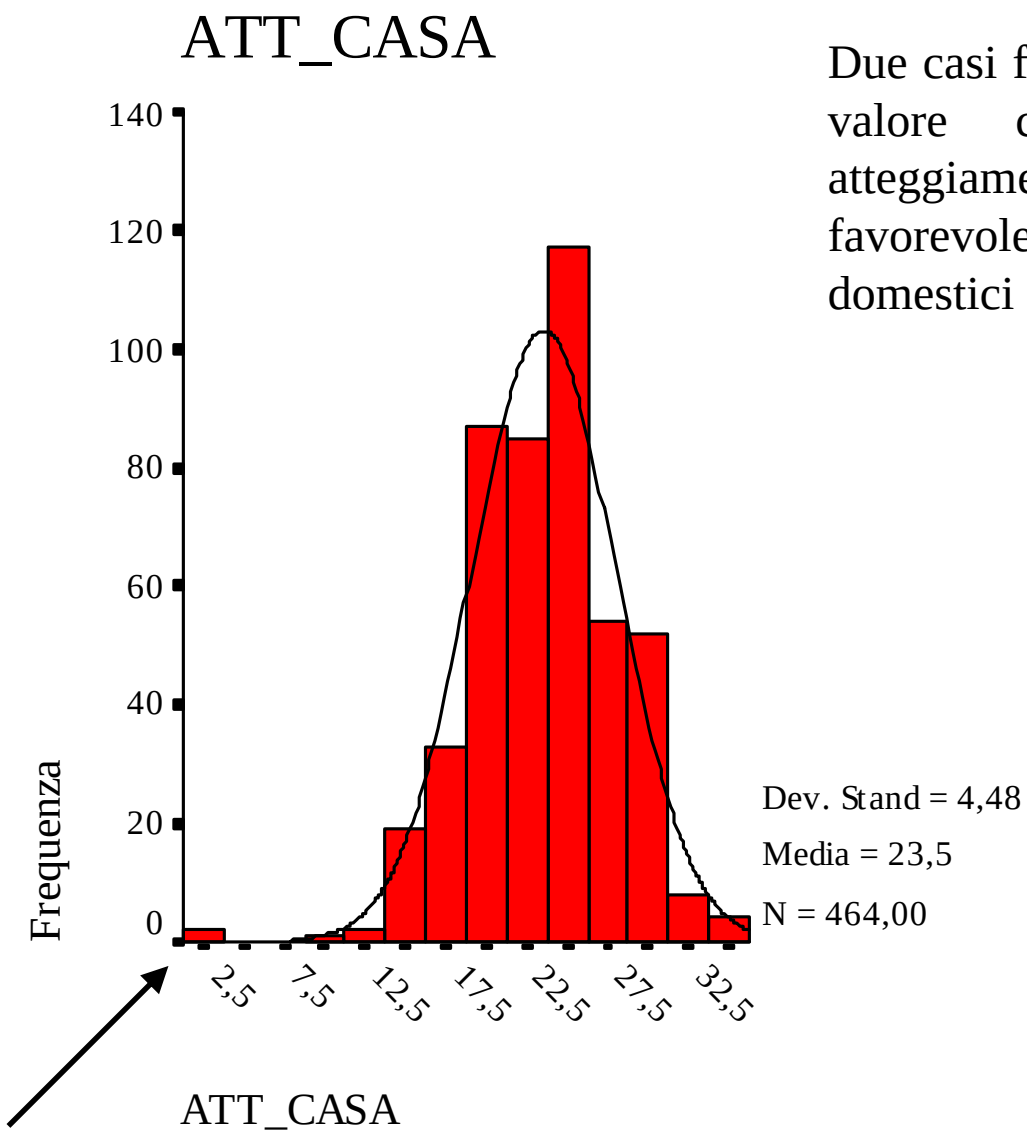
⇒ Utilizzare metodi grafici, generando per ogni variabile istogrammi, grafici a scatola (box plot), grafici di probabilità normale e detrendizzati

Outliers

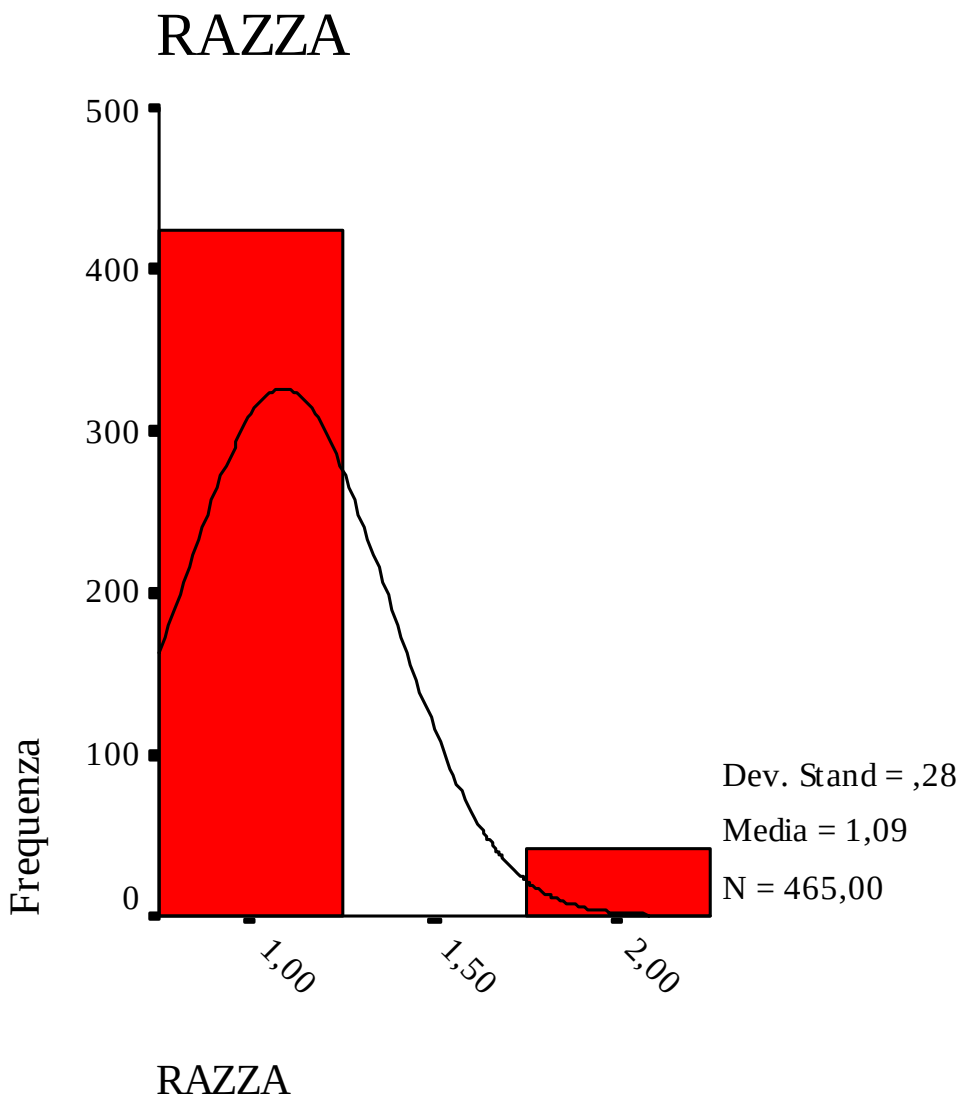
Gli istogrammi sono facilmente interpretabili. Generalmente si riscontra nell'esame del grafico una concentrazione di dati attorno alla media e la presenza di casi che se ne discostano in entrambe le direzioni. Un caso estremo si individua poiché sembra non appartenere alla distribuzione.

I box plot sono di interpretazione ancora più immediata. Racchiudono le osservazioni che si posizionano attorno alla mediana in scatole; i casi che si trovano lontano dai bordi della scatola rappresentano degli estremi.

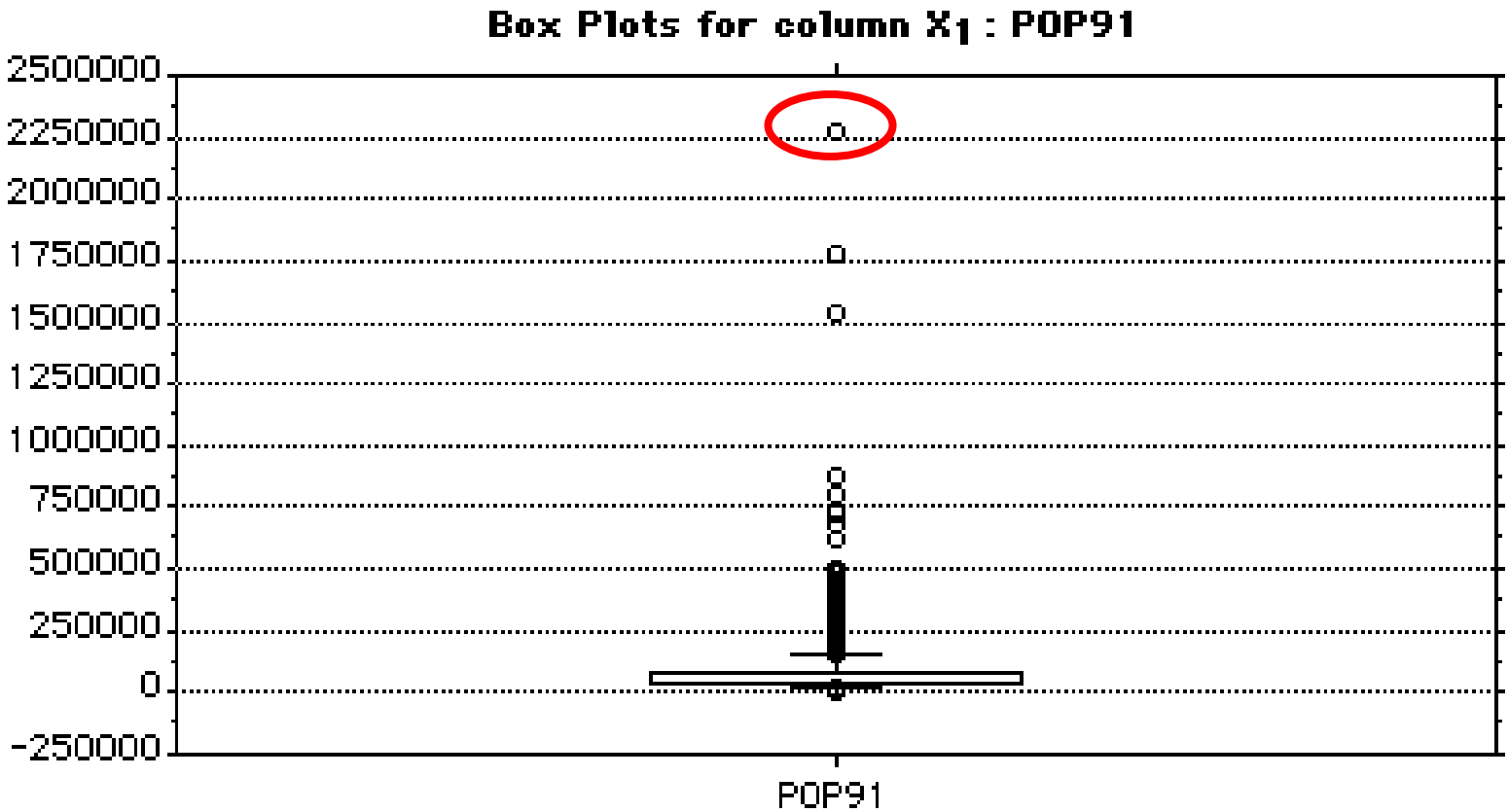
I grafici di probabilità normale e detrendizzati sono generati per stabilire la normalità della distribuzione di una variabile, comunque gli *outliers* univariati sono visibili in questi grafici come punti che giacciono ad una distanza considerevole dagli altri.



Due casi fanno registrare un valore che indica un atteggiamento estremamente favorevole verso i lavori domestici


RAZZA
500
400
300
200
100
0
Frequenza
Dev. Stand = ,28
Media = 1,09
N = 465,00
1,00
1,50
2,00
RAZZA

Grafici

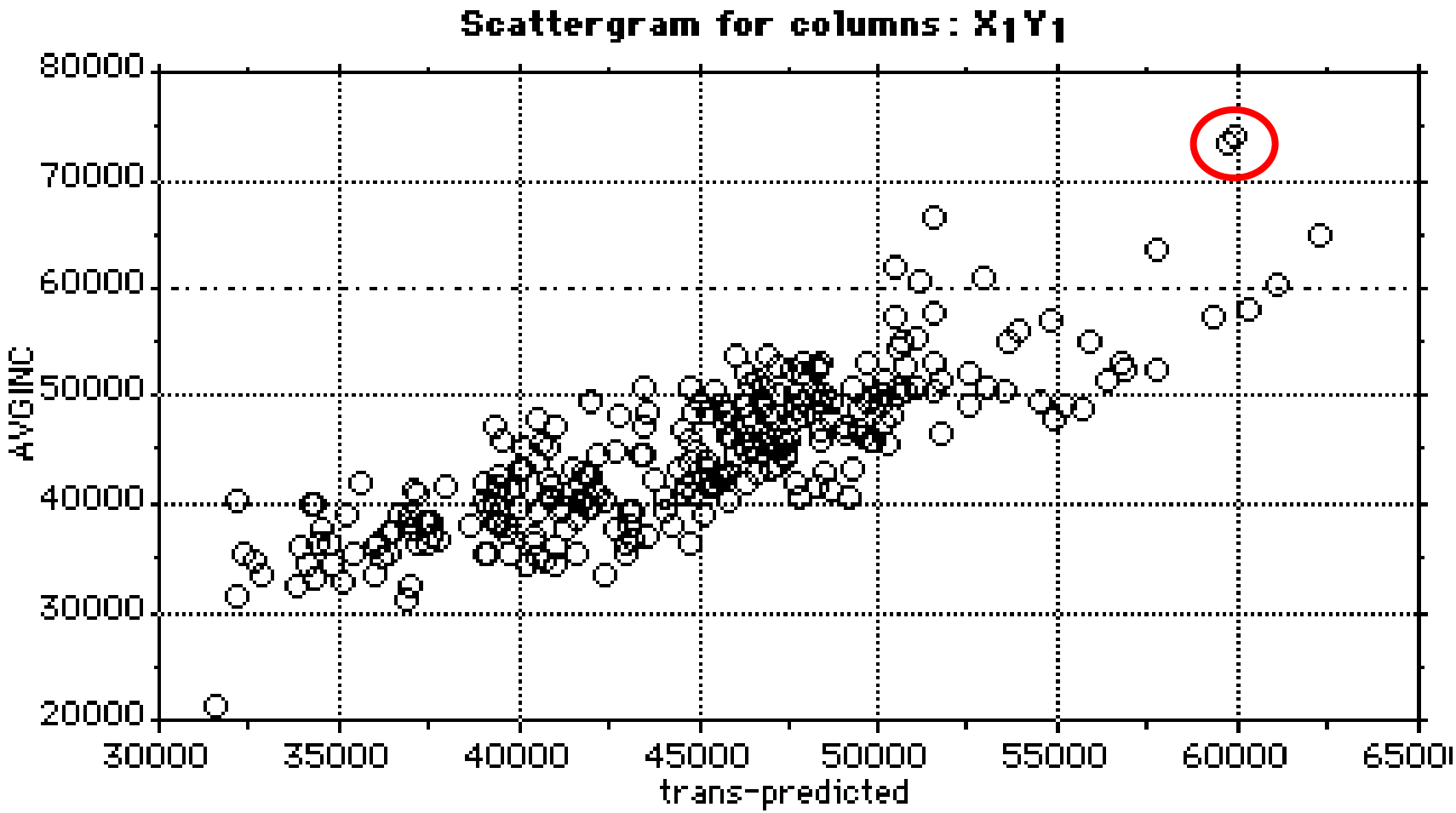

Box Plots for column X1 : POP91
2500000
2250000
2000000
1750000
1500000
1250000
1000000
750000
500000
250000
0
-250000
POP91

Outliers

Come individuare *outliers* multivariati

⇒ Calcolare la distanza di Mahalanobis, la leverage, la discrepanza e l'influenza (sebbene queste tre ultime statistiche siano state sviluppate nell'ambito dell'analisi di regressione adesso è possibile ottenerle anche applicando altri metodi multivariati)

⇒ Generare scatterplot per individuare outliers bivariati


Scattergram for columns: X1 Y1
80000
70000
60000
50000
40000
30000
20000
AVGINC
30000
35000
40000
45000
50000
55000
60000
65000
trans-predicted

Outliers

Suggerimenti

1. Individuare gli *outliers* univariati
2. Individuare gli *outliers* multivariati
3. Verificare se gli *outliers* univariati sono anche *outliers* multivariati
4. Decidere come trattare gli *outliers*
5. Ripetere la ricerca per accertarsi di aver eliminato tutti i casi estremi (potrebbe, infatti, accadere, che i casi estremi eliminati nascondano altri *outliers*). Se la ricerca sembra ripetersi all'infinito conviene provare ad effettuare l'analisi con e senza gli ultimi casi estremi evidenziati. Se i risultati non cambiano il processo si può interrompere

Outliers

Perché i casi sono estremi

E’ essenziale determinare le variabili rispetto alle quali i casi sono estremi per:

1. Capire se i casi individuati dovevano far parte del campione selezionato
2. Se scegliamo di modificare i valori piuttosto che procedere all’eliminazione dei casi estremi è necessario sapere quali punteggi andare a cambiare
3. Ottenere informazioni sul tipo di casi rispetto ai quali non è possibile generalizzare i risultati

Outliers

Descrivere i casi estremi

Se ci sono solo pochi *outliers* multivariati è ragionevole esaminarli individualmente.

Se siamo, invece, in presenza di numerosi *outliers* conviene esaminarli come gruppo per verificare se esiste qualche variabile che separa il gruppo di casi estremi dal resto delle osservazioni.

Operativamente si procede con la creazione di una variabile dicotomica che assume valore zero per i casi estremi ed uno per tutti gli altri casi. Tale variabile è quindi usata come VD in un'analisi discriminante, logistica oppure di regressione. La tecnica di ingresso delle variabili da utilizzare è la stepwise che fa entrare nell'equazione solo le variabili che discriminano gli *outliers*.

Outliers

Ridurre l'influenza dei casi estremi

1. Valutare la possibilità che una sola variabile sia responsabile per la presenza della maggior parte di *outliers*. Se questa è la circostanza prendere in considerazione la possibilità di eliminare la variabile (ovviamente questa soluzione costituisce una buona scelta solo se la variabile è fortemente correlata con altre variabili oppure non è critica per l'analisi da svolgere).
2. Verificare se gli *outliers* appartengono alla popolazione dalla quale si intendeva estrarre il campione. Se essi non appartengono alla popolazione possono essere eliminati senza pregiudicare la generalizzazione alla popolazione di interesse.
3. Se i casi estremi sono di natura univariata il loro impatto può essere ridotto trasformando la variabile per renderla più vicina ad una normale.

Outliers

Outliers nei risultati

Alcuni casi potrebbero non adattarsi in modo soddisfacente all'interno di una soluzione; i punteggi previsti per quei casi dal modello adottato sono molto diversi dai punteggi osservati.

Se si decidesse di eliminare o cambiare i punteggi per tali casi e poi di ripetere l'analisi si potrebbe "far apparire" i risultati migliori di quello che realmente sono.

Pertanto sarebbe opportuno limitarsi ad operazioni di "retrofitting" soltanto in ambito di analisi esplorativa.

NB: è necessario documentare in maniera adeguata ogni decisione volta all'eliminazione di casi o variabili, o al cambiamento dei valori di variabili