UNIVERSITÀ DEGLI STUDI DI SALERNO

Dipartimento di Informatica

CORSO DI LAUREA MAGISTRALE IN INFORMATICA

DATA SCIENCE E MACHINE LEARNING



PROGETTO DI STATISTICA E ANALISI DEI DATI

**Violenza di genere durante la pandemia: aggravamento o attenuazione dei numeri?**

|  |  |
| --- | --- |
| DOCENTE  Prof. Amelia Giuseppina Nobile | STUDENTI  Maria Natale, matricola: 0522500967  Gaetano Casillo, matricola:????????? |

ANNO ACCADEMICO 2020-2021

Sommario

[1 Introduzione 4](#_Toc58596987)

[1.1 Caso di studio 4](#_Toc58596988)

[1.2 Rappresentazione grafica 6](#_Toc58596989)

[2 Statistica descrittiva univariata 8](#_Toc58596990)

[2.1 Funzione di distribuzione empirica continua 8](#_Toc58596991)

[2.1.1 Funzione di distribuzione empirica continua Utenti 8](#_Toc58596992)

[2.1.2 Funzione di distribuzione empirica continua Vittime 9](#_Toc58596993)

[2.2 Indici di sintesi 10](#_Toc58596994)

[2.2.1 Indici di sintesi Utenti 11](#_Toc58596995)

[2.2.2 Indici di sintesi Vittime 16](#_Toc58596996)

[2.3 Forma della distribuzione di frequenze 19](#_Toc58596997)

[2.3.1 Forma della distribuzione di frequenze Utenti 21](#_Toc58596998)

[2.3.2 Forma della distribuzione di frequenze Vittime 21](#_Toc58596999)

[3 Statistica descrittiva bivariata 22](#_Toc58597000)

[3.1 Regressione lineare semplice 23](#_Toc58597001)

[3.1.1 Regressione lineare semplice Utenti 24](#_Toc58597002)

[3.1.2 Regressione lineare semplice Vittime 28](#_Toc58597003)

[3.2 Regressione lineare multipla 32](#_Toc58597004)

[3.2.1 Regressione lineare multipla Utenti 32](#_Toc58597005)

[3.2.2 Regressione lineare multipla Vittime 36](#_Toc58597006)

[4 Analisi dei cluster 40](#_Toc58597007)

[4.1 Suddivisione in cluster Utenti 42](#_Toc58597008)

[4.2 Suddivisione in cluster Vittime 48](#_Toc58597009)

[5 Variabile aleatoria esponenziale 54](#_Toc58597010)

[5.1 Stima del parametro non noto λ 57](#_Toc58597011)

[5.1.1 Stima puntuale 57](#_Toc58597012)

[5.1.2 Stima intervallare 58](#_Toc58597013)

[5.1.3 Confronto tra due popolazioni esponenziali 60](#_Toc58597014)

[5.2 Verifica delle ipotesi 62](#_Toc58597015)

[6 Bibliografia 63](#_Toc58597016)

# Introduzione

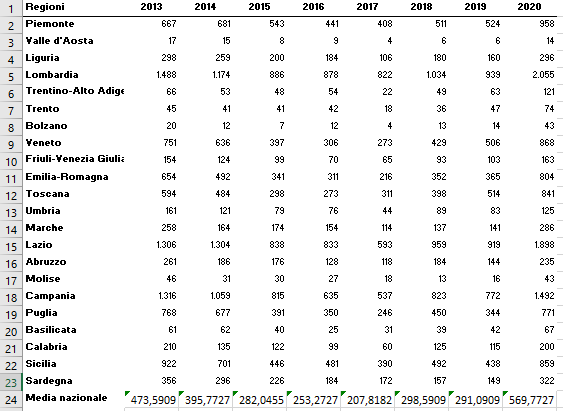
Si sente spesso parlare di violenza di genere, ma che cosa vuol dire? Le Nazioni Unite hanno definito la violenza di genere come “*ogni atto legato alla differenza di sesso che provochi o possa provocare un danno fisico, sessuale, psicologico o una sofferenza della donna, compresa la minaccia di tali atti, la coercizione o l’arbitraria privazione della libertà sia nella vita pubblica che nella vita privata*” [1]. Per supportare le vittime delle violenze di genere è stato attivato il numero verde 1522 attivo 24 ore su 24 offrendo accoglienza in italiano, inglese, francese, spagnolo e arabo.

## Caso di studio

Nel 2020 è stato vissuto il lockdown per 3 mesi, in questo periodo molto si è parlato del lato economico, della scuola, ma poco si è discusso del lato sociale di questo evento. Si è pensato pertanto di analizzare le chiamate e i messaggi effettuate al 1522, numero verde contro lo stalking e la violenza, confrontandole con lo stesso periodo (marzo/giugno) degli anni precedenti. In particolare, nell’analisi statistica effettuata si fa riferimento agli utenti e alle vittime per regione di provenienza e anno. Secondo quanto rilevato dall’Istat, che ha analizzato i dati messi a disposizione dal numero antiviolenza 1522, tra marzo e giugno 2020 le telefonate e le comunicazioni via chat con il centralino sono [più che raddoppiate](https://www.istat.it/it/archivio/246557) rispetto allo stesso periodo dell’anno precedente, con un +119,6%. Il numero complessivo di contatti validi nel periodo preso in considerazione è stato di 15.280, di cui circa un terzo poi trasferiti ad altri servizi come quello dei centri antiviolenza. Numeri assolutamente non paragonabili col passato [2].

Per l’analisi del fenomeno in esame si considerano i dati relativi agli utenti del numero antiviolenza 1522 effettuate nei mesi di marzo-giugno suddivisi per regione ed anno (2013-2020). In particolare, nell’analisi statistica univariata, verranno esaminate nei dettagli le curve relativi ai dati della regione Campania e la media delle chiamate degli utenti e delle vittime effettuate sull’intero territorio nazionale.

Nella seguente tabella vengono mostrati i dati relativi agli utenti del numero 1522 suddivisi per regione ed anno.



## Rappresentazione grafica

Di seguito vengono mostrati i due barplot relativi ai dati della Campania e della media sull’intero territorio nazionale per quanto riguarda la tabella Utenti. In entrambi i casi si può notare che la modalità a cui è associata la frequenza più alta è il 2020.

Il codice per realizzare il barplot della Campania:

png("grafici/chiamateEffettuateUtentiCampaniaFrequenza.png")

x<-barplot(utenti\_campania, xlab="Anni", ylab="Numero di chiamate effettuate", ylim=c(0,1800), col=1:9,

names.arg = colnames, main ="Numero di utenti al numero verde in Campania")

text(x, y=utenti\_campania, pos = 3, labels = utenti\_campania, col="red")

dev.off()

# Statistica descrittiva univariata

In questo capitolo verranno mostrati i risultati relativi all’analisi statistica univariata. In particolare, verrà mostrata la funzione di distribuzione empirica continua, i valori degli indici di sintesi, i quartili calcolati con i differenti algoritmi di R e gli indici di dispersione. Infine, verrà analizzata la forma della distribuzione di frequenze attraverso il calcolo della skewness campionaria e della curtosi campionaria. Le varie analisi verranno effettuate prendendo in esame i dati della Campania e della media nazionale negli anni 2013-2020, analizzando le tabelle Utenti.

## Funzione di distribuzione empirica continua

La funzione di distribuzione empirica continua viene utilizzata nel caso di dati continui che vengono strutturati in classi. Ad esempio, se si vuole considerare k classi distinte, le classi saranno così caratterizzate: C1 = [z0, z1), C2 = [z1, z2), … Ck=[zk-1, zk] con z0 < z1 < … < zk-1 < zk, dove z0 corrisponde al minimo delle osservazioni e zk corrisponde al massimo delle osservazioni. La funzione di distribuzione empirica continua viene calcolata a partire dalle frequenze relative cumulative associate alle varie classi.

Per calcolare la funzione di distribuzione continua relativa alla tabella Utenti le osservazioni sono state suddivise in tre classi. Per quanto riguarda la media nazionale le classi individuate sono le seguenti: C1 = [208, 329), C2 = [329, 450), C3 = [450, 570]. Per quanto riguarda la Campania le classi individuate sono le seguenti: C1 = [537, 855), C2 = [855, 1173), C3 = [1173, 1492].

Il seguente codice mostra come sono state calcolate le frequenze relative cumulative per le classi individuate per quanto riguarda la Campania.

minOsservazione = min(utenti\_campania)

maxOsservazione = max(utenti\_campania)

frequenza<-table(utenti\_campania)/length(utenti\_campania)

lung<-length(frequenza)

classe<-round((maxOsservazione-minOsservazione)/3, digits=0)

classi<-c(minOsservazione, minOsservazione+classe, minOsservazione+2\*classe, maxOsservazione)

frelclassi <-table (cut (utenti\_campania, breaks = classi,right = FALSE ))/ length (utenti\_campania)

Fcum <-cumsum (frelclassi)

Fcum[3]<-Fcum[3]+frequenza[lung]

Dopo aver calcolato le frequenze relative cumulative. sono stati quindi creati i grafici che mostrano le frequenze di distribuzione continue della Campania e dell’intera nazione.



## Indici di sintesi

Alcuni indici di sintesi utili a descrivere i dati sono media, mediana, moda, varianza, deviazione standard e coefficiente di variazione. Le prime tre sono misure di centralità dei dati mentre le altre misurano la loro dispersione.

Supponiamo di avere un insieme, x1 , x2 ,…, xn di n valori numerici. Si definisce **media campionaria** la quantità:

Dato un campione di dati ordinato in maniera crescente, si definisce la **mediana** (o **valore mediano**) come il valore/modalità assunto dalle unità statistiche che si trovano nel mezzo della distribuzione. Se n è dispari, la mediana sarà il valore in posizione (n+1)/2; se n è pari la mediana sarà la media aritmetica dei valori in posizione n/2 e n/2+1. La mediana, quindi è quel valore che divide a metà l’insieme dei dati ordinati. Oltre a questo indice si possono considerare altri indici di posizione detti quantili che consentono di suddividere l’insieme dei dati ordinati in un fissato numeri di parti uguali. In particolare, verranno considerati i quartili che consentono di dividere l’insieme dei dati ordinati in quattro parti uguali.

La **moda campionaria** di un insieme di dati è il valore a cui è associata la frequenza più elevata, non è obbligatorio che la moda esista in ogni insieme di dati e se esiste, è possibile che ne esista più di una; in questo caso, ogni valore è detto “valore modale”.

Nel grafico seguente vengono mostrate le due curve relative ai dati che si stanno analizzando.



Entrambe le curve mostrano una distribuzione di frequenze non simmetrica, in particolare inizialmente sono decrescenti ed hanno un picco massimo nell’ultimo anno 2020. Le due curve sono tra loro abbastanza simili.

Il grafico seguente mostra, invece, i boxplot di entrambi i campioni di dati per illustrare alcune caratteristiche della distribuzione di frequenza come centralità, dispersione, forma e la presenza di eventuali valori anomali. Il boxplot, detto anche “scatola con i baffi”, rappresenta una scatola i cui estremi sono Q1 (primo quartile)e Q3 (terzo quartile) tagliata da una linea orizzontale in corrispondenza di Q2 (secondo quartile). Sono inoltre presenti due ulteriori linee che rappresentano i baffi in alto e in basso. Il baffo inferiore corrisponde al valore più piccolo tra le osservazioni che risulta maggiore o uguale a , mentre il baffo superiore corrisponde al valore più grande delle osservazioni che risulta minore o uguale a . Se tutti i dati rientrano nell’intervallo , i baffi risultano essere posti in corrispondenza del minimo e del massimo dei dati del campione. I valori anomali al di fuori di tale intervallo vengono visualizzati sotto forma di punti nel grafico.



Entrambi i boxplot rivelano la presenza di asimmetria nei dati in quanto le distanze tra primo e terzo quartile dalla linea della mediana sono molto diverse tra loro. Si può intuire che le curve hanno una coda più allungata a destra e ciò verrà confermato attraverso il calcolo della skewness campionaria.

Utilizzando la funzione summary in R è possibile calcolare minimo, massimo, media, mediana, primo e terzo quartile.





Avendo ottenuto il valore dei quartili, è possibile calcolare il valore dei baffi del boxplot della Campania.

quindi il baffo inferiore è posto in corrispondenza del valore 537.

quindi il baffo superiore è posto in corrispondenza del valore 1492.

I valori sono compresi tra pertanto i baffi risultano essere posti in corrispondenza del minimo e del massimo del campione.

Valore dei baffi nel boxplot della media nazionale:

quindi il baffo inferiore è posto in corrispondenza del valore 208.

quindi il baffo superiore è posto in corrispondenza del valore 570.

Anche per quanto riguarda il campione della media nazionale i valori sono compresi tra pertanto i baffi risultano essere posti in corrispondenza del minimo e del massimo del campione.

Nei due boxplot non risultano esserci valori anomali.

Le medie campionarie dei due campioni di dati negli anni risultano essere:



Pertanto, è possibile vedere quali sono gli anni in cui ci sono state più chiamate rispetto alla media e gli anni in cui ci sono state meno chiamate.

|  |  |
| --- | --- |
| Media nazionale | |
| 2013 | 474 |
| 2014 | 396 |
| 2015 | 282 |
| 2016 | 253 |
| 2017 | 208 |
| 2018 | 299 |
| 2019 | 291 |
| 2020 | 570 |

|  |  |
| --- | --- |
| Campania | |
| 2013 | 1316 |
| 2014 | 1059 |
| 2015 | 815 |
| 2016 | 635 |
| 2017 | 537 |
| 2018 | 823 |
| 2019 | 772 |
| 2020 | 1492 |

Sia per la Campania che per la media nazionale gli anni in cui ci sono state più chiamate rispetto alla loro media sono 2013, 2014 e 2020.

Per individuare la moda si considerano gli istogrammi delle frequenze dei dati considerando la loro suddivisione nelle seguenti cinque classi: C1 = [0, 500), C2 = [500, 1000), C3 = [1000, 1500) C4= [1500, 2000), C5= [2000, 2500].

classi<-c(0, 500, 1000, 1500, 2000, 2500)

fclassiCampania <-table (cut (utenti\_campania, breaks = classi,right = FALSE, dig.lab = 10))

for (i in 1:length(utenti\_campania)){

if(utenti\_campania[i]==2500)

fclassiCampania[3]<-fclassiCampania[3]+1

}

fclassiItalia <-table (cut (utenti\_nazione, breaks = classi,right = FALSE, dig.lab=10))

for (i in 1:length(utenti\_nazione)){

if(utenti\_nazione[i]==2500)

fclassiItalia[3]<-fclassiItalia[3]+1

}

png("grafici/istogrammaClassiCampania.png")

hist(utenti\_campania, breaks=classi, col=rainbow(3), main="Istogramma delle frequenze delle classi in Campania")

dev.off()

png("grafici/istogrammaClassiItalia.png")

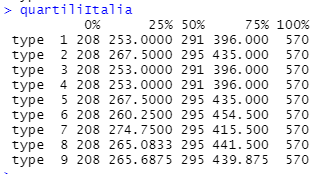
hist(utenti\_nazione, breaks=classi, col=rainbow(3), main="Istogramma delle frequenze delle classi in Italia")

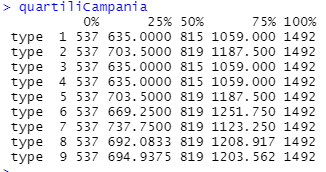
dev.off()

La classe modale per l’Italia è C1 = [0, 500), in particolare tutti i valori sono concentrati in quella classe. Per la Campania invece la classe modale risulta essere C2 = [500, 1000).



In R ci sono 9 diversi tipi di algoritmi che consentono di calcolare i quartili. Di seguito vengono mostrati i risultati ottenuti dal calcolo dei quartili per quanto riguarda gli utenti della Campania e della media nazionale.





Dopo aver considerato gli indici di posizione sono stati considerati gli indici di dispersione.

Avendo un insieme di dati numerici (), si definisce **varianza campionaria** e si indica con , la quantità:

Si definisce **deviazione standard campionaria** la radice quadrata della varianza ossia:

Assegnato un campione di dati numerici x1 , x2 ,…, xn , si definisce **coefficiente di variazione** il rapporto tra la deviazione standard campionaria e il modulo della media campionaria: .

Il seguente codice permette di mostrare i valori della varianza varianza, deviazione standard e coefficiente di variazione dei due campioni di dati.



La varianza e la deviazione standard di entrambi i campioni risultano essere dei valori grandi e da tali valori non si riesce ad avere una effettiva misura della dispersione, pertanto si considera il coefficiente di variazione. Inoltre, non è possibile da queste due misure effettuare un confronto delle dispersioni dei due campioni in quanto la media nazionale risulta avere valori numerici molto più bassi rispetto alla sola regione Campania.

Il coefficiente di variazione del campione di dati della Campania è circa 0.3567, mentre quello della media nazionale è circa 0.3551. I due coefficienti sono tra loro molto vicini, indicano quindi una dispersione dei dati attorno alla media molto simile. Il coefficiente di variazione di entrambi è più vicino allo 0 che ad 1 quindi i valori assunti dai due campioni non risultano essere molto sbilanciati tra i vari anni.

## Forma della distribuzione di frequenze

In questo paragrafo verranno descritti gli indici statistici che permettono di analizzare la forma della distribuzione di frequenze misurando se essa presenta asimmetrie (positive o negative) o se essa è più o meno piccata rispetto ad una distribuzione di frequenze normale standard. Prima di definire tali indici è utile introdurre il concetto di momento campionario e di momento centrato.

Assegnato un insieme di dati numerici x1 , x2 ,…, xn , si definisce **momento campionario** di ordine j la quantità:

Assegnato un insieme di dati numerici x1 , x2 ,…, xn , si definisce **momento campionario centrato** attorno alla media di ordine j la quantità:

La skewness campionaria permette di misurare la simmetria di una distribuzione di frequenze. Assegnato un insieme di dati numerici x1 , x2 ,…, xn , si definisce **skewness campionaria** il valore:

Se la distribuzione è simmetrica il valore γ1 è nullo, γ1 > 0 se la distribuzione ha un’asimmetria positiva (ovvero una coda a destra più allungata), γ1 < 0 se la distribuzione ha un’asimmetria negativa (ovvero una coda a sinistra più allungata).

Il codice per calcolare la skewness campionaria in R è:

skw <-function (x){

n<-length (x)

m2 <-(n -1) \*var (x)/n

m3 <- (sum ( (x- mean(x))^3) )/n

m3/(m2 ^1.5)

}

Applicando tale funzione ai due campioni di dati si ottengono i seguenti risultati.



Entrambe le distribuzioni di frequenze hanno un’asimmetria positiva, la distribuzione di frequenza ha quindi una coda più allungata a destra.

La **curtosi campionaria** è un indice che permette di misurare la densità dei dati intorno alla media.

Essa si calcola con la seguente equazione:

Dove è l’indice di Pearson e sono rispettivamente il momento centrato campionario di ordine 2 ed ordine 4.

Da notare anche che è indipendente dall’unità di misura dei dati.

Gli indici permettono di confrontare la dei dati con una densità di probabilità normale standard

* Se abbiamo una distribuzione di frequenze platicurtica, è quindi più piatta di una normale
* Se abbiamo una distribuzione di frequenze leptocurtica, è quindi più piccata di una normale
* Se abbiamo una distribuzione di frequenze normocurtica, è quindi piatta come una normale

Il codice per calcolare la curtosi in R è:

curt <-function (x){

n <-length (x)

m2 <-(n -1) \*var (x)/n

m4 <- (sum ((x-mean(x))^4) )/n

m4/(m2 ^2) -3

}

Applicando tale funzione ai due campioni di dati considerati si ottiene:



Il valore di entrambe le curtosi campionarie è negativo quindi entrambe le distribuzioni di frequenza sono meno piccate di una distribuzione di frequenze normale standard.

Confrontando i valori ottenuti da questi due indici si ha un’ulteriore conferma del fatto che l’andamento negli anni delle due curve considerate risulte essere molto simile anche se i dati relativi all’intera nazione sono più bassi in quanto sono ottenuti dalla media di tutte le nazioni, che viene fortemente influenzata dai valori bassi presenti in molte regioni con meno abitanti rispetto alla Campania.

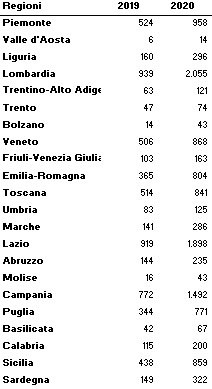
# Statistica descrittiva bivariata

In questo capitolo verranno mostrate le analisi di regressione lineare semplice e di regressione lineare multipla calcolando il modello lineare, i residui e il coefficiente di determinazione.

La statistica descrittiva bivariata si occupa dei metodi grafici e statistici atti a descrivere le relazioni che intercorrono tra due variabili X e YPer ottenere una misura quantitativa della correlazione tra le variabili si considera la covarianza campionaria.

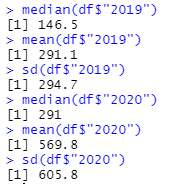
## Regressione lineare semplice

Il modello di regressione lineare semplice viene utilizzato per spiegare la relazione che esiste tra una variabile dipendente Y e una variabile indipendente X. In questa analisi verrà considerata come variabile indipendente l’anno 2019 e come variabile dipendente l’anno 2020.



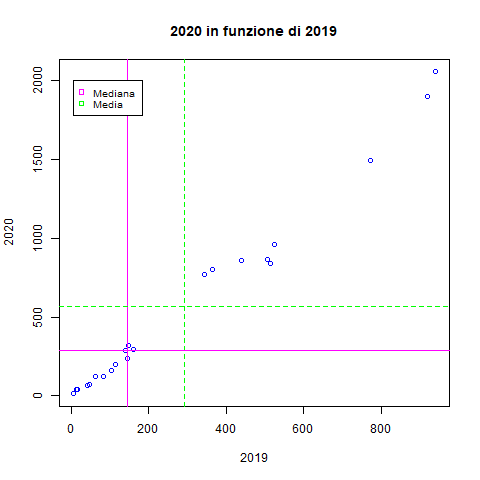
Si calcolano gli indici di posizione e di dispersione relativi alle due coppie di variabili.

Si nota che sia mediana, sia media che deviazione standard sono maggiori per la variabile Y.



Un primo passo per indagare l’eventuale dipendenza tra due variabili X e Y consiste nel disegnare il diagramma di dispersione o scatterplot. Il grafico che si ottiene mira ad evidenziare se le coppie di punti presentano qualche forma di regolarità.

Nello scatterplot si pone sull’asse delle ascisse la variabile indipendente 2019 e sulle ordinate la variabile dipendente 2020. Vengono poi tracciate delle linee orizzontali e verticali in corrispondenza delle mediane e delle medie delle due variabili.



Dallo scatterplot si nota che i dati (a parte qualche punto che si discosta un po’ di più) sono posizionati lungo una retta ascendente quindi si può dedurre che esiste una correlazione lineare positiva tra le due variabili considerate.

Per ottenere una misura quantitativa della correlazione tra le variabili si calcola la covarianza campionaria, che è così definita:

*Assegnato un campione bivariato (x1,y1),(x2,y2), ...,(xn,yn) di una variabile quantitativa bi-dimensionale (X,Y), siano 𝑥̅ e 𝑦̅ rispettivamente le medie campionarie di x1,x2, ...,xn e di y1,y2, ...,yn. La covarianza campionaria tra le due variabili X e Y è così definita:*

Se le variabili sono correlate positivamente, se le variabili sono correlate negativamente, se le variabili non sono correlate.



La covarianza tra le due variabili risulta essere 177155, pertanto esiste una correlazione lineare positiva tra le due variabili come si poteva già intuite dal grafico dello scatterplot.

Per ottenere una misura quantitativa della correlazione tra le variabili si può anche considerare il coefficiente di correlazione campionario, che è così definito:

*Assegnato un campione bivariato (x1,y1),(x2,y2), ...,(xn,yn) di una variabile quantitativa bidimensionale (X,Y), siano 𝑥̅ e sx la media campionaria e la deviazione standard di x1,x2, ...,xn ed inoltre siano 𝑦̅ e sy la media campionaria e la deviazione standard di y1,y2, ...,yn . Il* ***coefficiente di correlazione campionario*** *tra le due variabili X e Y è cosi definito:*

Il coefficiente di correlazione campionario misura la forza del legame di natura lineare esistente tra due variabili quantitative. In particolare, e il suo valore indica la direzione della retta interpolante.

* : (correlazione perfetta negativa), tutti i punti sono allineati lungo una retta discendente;
* (correlazione negativa), i punti sono posizionati in una nuvola attorno ad una retta interpolante discendente;
* : (nessuna correlazione), i punti sono completamente dispersi in una nuvola che non presenta alcuna evidente direzione di natura lineare;
* : (correlazione positiva), i punti sono posizionati in una nuvola attorno ad una retta interpolante ascendente;
* : (correlazione perfetta positiva), tutti i punti sono allineati lungo una retta ascendente.



Siccome il coefficiente di correlazione è uguale a 0.9924 che è prossimo ad 1 i dati presentano un’altissima correlazione positiva. Per calcolare la retta interpolante di questi punti si utilizza il modello di regressione lineare semplice.

Il modello di regressione lineare semplice è esprimibile tramite l’equazione di una retta che riesce ad interpolare la nuvola di punti dello scatterplot meglio di tutte le altre possibili rette, dove α è l’intercetta e β è il coefficiente angolare. Se β>0 la retta di regressione è crescente, se β<0 la retta di regressione è discendente, se β=0 la retta è orizzontale. L’intercetta α corrisponde invece al punto di intersezione della retta interpolante con l’asse delle ordinate.

Il seguente codice permette di realizzare lo scatterplot relativo ai dati del 2019 e del 2020 con la retta interpolante stimata.

png("grafici/bivariata/scatterPlotUtenti2020\_2019RettaRegressione.png")

plot(df$"2019", df$"2020", main="Retta di regressione 2020 in funzione di 2019", col="blue",

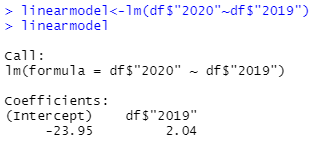
xlab="2019", ylab="2020")

abline(lm(df$"2020"~df$"2019"), col="magenta")

dev.off()



La funzione lm(y˜x) permette di eseguire le analisi di regressione lineare della variabile dipendente y in funzione della variabile indipendente x.

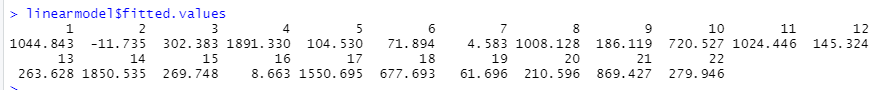


L’intercetta α vale -23.97, mentre il coefficiente angolare β vale 2.04. Siccome il coefficiente angolare ha segno positivo, la retta è ascendente. L’equazione della retta interpolante risulta pertanto:

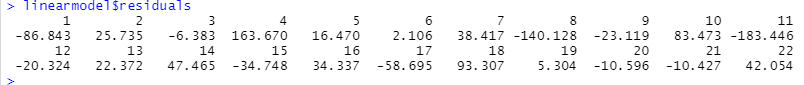
Dopo aver calcolato la retta interpolante, è possibile notare che esistono degli scostamenti tra i valori osservati del campione e i valori stimati attraverso la retta di regressione. Le differenze tra le ordinate dei punti dei valori osservati e le ordinate dei punti dei valori stimati prendono il nome di residui. Se si indica con il valore osservato e con il valore stimato, i **residui** sono così definiti:

La media campionaria dei residui è nulla.

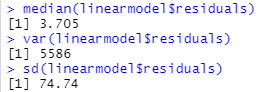
Il codice seguente permette di visualizzare i valori stimati.



Il seguente codice permette di visualizzare i residui, ossia di quanto le ordinate dei valori osservati si discostano dai valori stimati.



La mediana, la varianza e la deviazione standard dei residui assumono i seguenti valori. Non è possibile calcolare il coefficiente di variazione in quanto la media campionaria dei residui è 0.



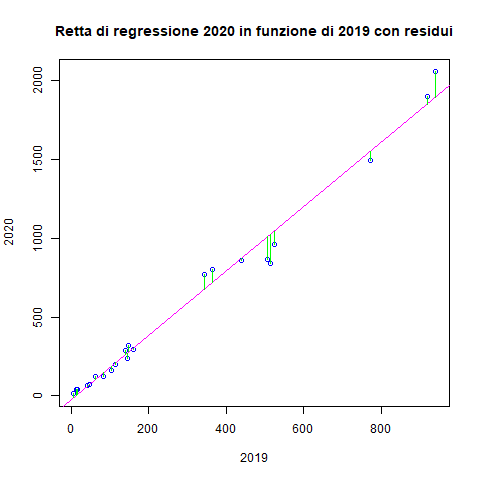
Di seguito viene mostrato il grafico che rappresenta lo scatterplot dei punti, la retta di regressione e dei segmenti verticali che rappresentano i residui.

plot(df$"2019", df$"2020", main="Retta di regressione 2020 in funzione di 2019 con residui", col="blue",

xlab="2019", ylab="2020")

abline(linearmodel, col="magenta")

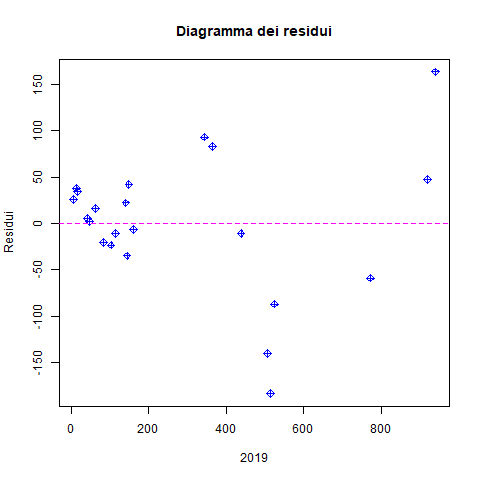
segments (df$"2019", linearmodel$fitted.values, df$"2019", df$"2020" ,col="green")



Un esame più accurato del modo con cui la retta di regressione interpola i dati e di come i residui si dispongano intorno alla retta interpolante influenzandone la posizione, può essere ottenuto attraverso il diagramma dei residui che è un grafico in cui i valori dei residui sono posti sull’asse delle ordinate e quelli della variabile indipendente sull’asse delle ascisse.

plot(df$"2019", residui, main="Diagramma dei residui", xlab="2019", ylab="Residui", col="blue", pch =9)

abline (h=0, col ="magenta",lty=2)



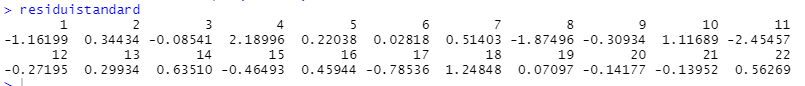
La linea tratteggiata è posizionata su 0 che indica la media campionaria dei residui. Si nota che i punti sono disposti casualmente attorno alla retta orizzontale e non si evidenzia nessun comportamento particolare nella distribuzione dei punti. La posizione della retta di regressione è fortemente influenzata dalla presenza di eventuali valori anomali che si discostano in modo significativo dagli altri. L’analisi dei residui aiuta ad individuare eventuali punti isolati (valori anomali) dovuti ad errori nella stima. Tali valori possono perturbare significativamente la stima dei parametri di regressione e influenzare l’interpretazione dei residui. Eliminando i valori anomali la varianza campionaria dei residui diminuisce.

Spesso è utile calcolare i residui standardizzati, così definiti:

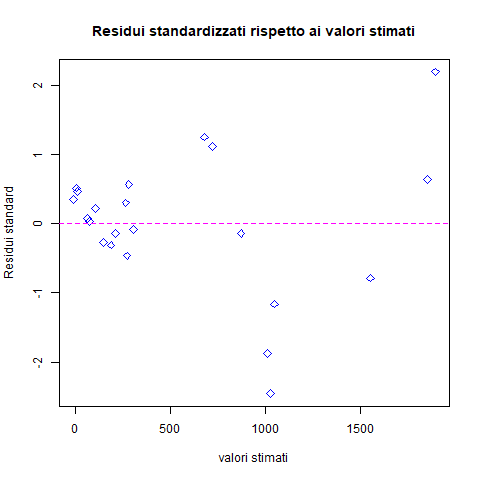
I residui standardizzati sono caratterizzati da media nulla e varianza unitaria.

residui<-linearmodel$residuals

residuistandard<-residui/sd(residui)



Successivamente è stato realizzato un grafico che mostra sulle ordinate i residui standardizzati e sulle ascisse i valori stimati.



La maggior parte dei residui standardizzati si concentra nell’intervallo [-1,1] e sono quei residui in corrispondenza di quei valori osservati che hanno lo stesso andamento dei valori attesi. Ci sono comunque valori che si discostano maggiormente dai propri valori attesi come Piemonte, Lombardia, Veneto, Toscana, Emilia-Romagna e Campania.

Per conoscere quanto la retta di regressione si adatta ai dati considerati si calcola il **coefficiente di determinazione** che viene definito come il rapporto tra la varianza dei valori osservati e la varianza dei valori stimati. Un coefficiente di determinazione prossimo ad 1 indica che tutti i punti tendono ad allinearsi lungo la retta di regressione, mentre un coefficiente di determinazione prossimo a 0 indica una completa incapacità della retta di rappresentare la distribuzione dei dati considerati. Per il modello di regressione lineare semplice il coefficiente di determinazione corrisponde al quadrato del coefficiente di correlazione. È possibile ottenere il valore del coefficiente di determinazione in questo modo:

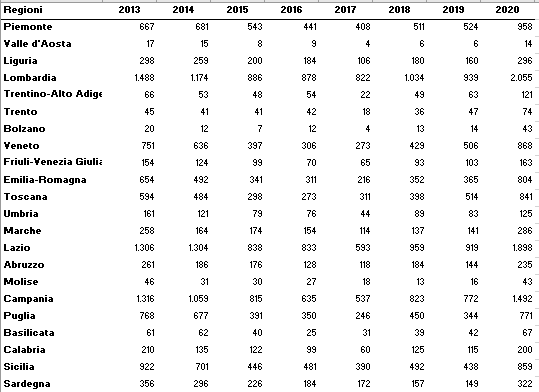


In questo caso il coefficiente di correlazione vale 0.9848. Siccome è prossimo ad 1, significa che la retta descrive bene i dati considerati, infatti anche dai grafici visti precedentemente si nota che gli scostamenti dalla retta sono molto piccoli.

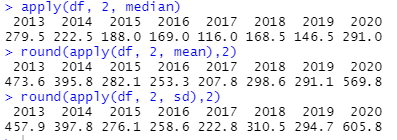
## Regressione lineare multipla

Il modello di regressione lineare multipla viene utilizzato per spiegare la relazione tra una variabile quantitativa Y detta variabile dipendente e le variabili quantitative indipendenti X1, X2, …, Xp.

Il data frame considerato è il seguente e si utilizza il modello di regressione lineare multipla per spiegare la relazione le variabili indipendenti: 2013, 2014, 2015, 2016, 2017, 2018, 2019 e la variabile dipendente: 2020.

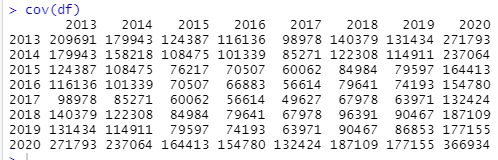


Di seguito vengono mostrati i valori degli indici di posizione e di dispersione (mediana, media e deviazione standard) relativi alle variabili considerate.



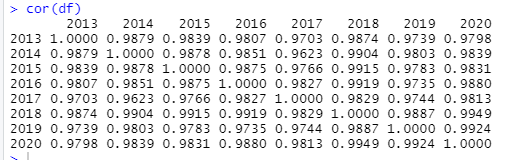
Media, mediana e deviazione standard sono maggiori per la variabile 2020.

Di seguito viene mostrata la matrice delle covarianze che contiene sulla diagonale principale la varianza delle singole colonne del dataframe, mentre gli altri elementi rappresentano le covarianze tra le coppie di variabili.



Da questa matrice si può notare come tutte le coppie di variabili siano tra di loro positivamente correlate. Tali numeri sono però elevati e non suggeriscono quanto sia forte il legame tra le variabili pertanto viene considerato il coefficiente di correlazione.

Di seguito viene quindi mostrata la matrice delle correlazioni che contiene tutte le correlazioni lineari tra le coppie di variabili, ossia misura la forza del legame di natura lineare esistente tra tutte le coppie di variabili quantitative. La matrice delle correlazioni contiene 1 sulla diagonale principale.



Si nota che esiste una forte correlazione lineare tra tutte le variabili considerate.

Il seguente grafico visualizza in un’unica finestra tutti gli scatterplot ottenuti mettendo in relazione le varie coppie di variabili.

pairs(df, main="Scatterplot per le coppie di variabili", col="blue")



Anche dal grafico si nota che le coppie di variabili sono positivamente correlate.

Il grafico precedente è ottenuto mediante il codice:

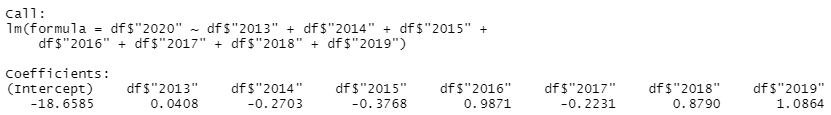
Il modello di regressione lineare multipla con p variabili indipendenti è esprimibile attraverso l’equazione:

Dove:

* è l’intercetta, ossia il valore di Y quando X1=X2= …= Xp=0;
* sono i regressori. In particolare, rappresenta l’inclinazione di Y rispetto alla variabile X1 tenendo costanti le variabili X2, X3, …,Xp, …, rappresenta l’inclinazione di Y rispetto alla variabile Xp tenendo costanti le variabili X1, X2, …,Xp-1.

Utilizzando il modello di regressione lineare multipla si ottiene:

multiplelinearmodel<-lm(df$"2020"~df$"2013" +df$"2014" + df$"2015" + df$"2016" + df$"2017" + df$"2018" + df$"2019")

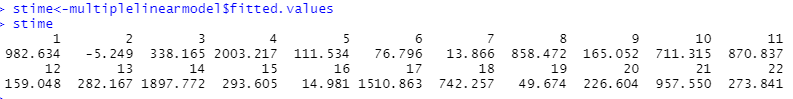


Da cui si ricava che l’intercetta è -18.6585 e i regressori sono: 0.0408, -0.2703, -0.3768, 0.9871, -0.2231, 0.8790, 1.0864.

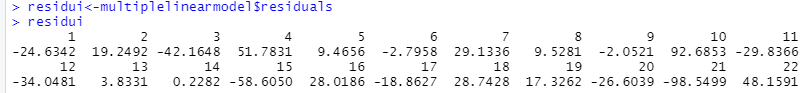
I segni dei regressori β1, β4, β6, β7 sono positivi: questo indica che all’aumentare del numero di utenti nel 2013, 2016, 2018 e 2019 aumenta il numero di utenti nel 2020. Mentre i regressori β2, β3 , β5 sono negativi quindi all’aumentare del numero di utenti nel 2014, 2015, 2017 diminuisce il numero di utenti nel 2020.

Il regressore β1 è prossimo allo zero, questo indica che il numero di utenti nel 2013 non incide in maniera significativa il numero di utenti nel 2020.

Il codice seguente permette di visualizzare i valori stimati rispetto al modello di regressione lineare multipla.



Il seguente codice permette di visualizzare i residui.



Successivamente sono stati calcolati i residui standardizzati.

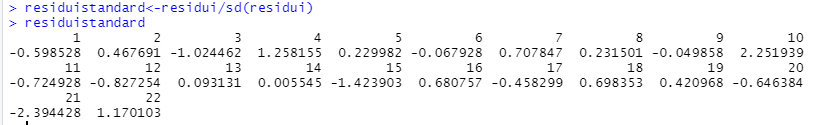
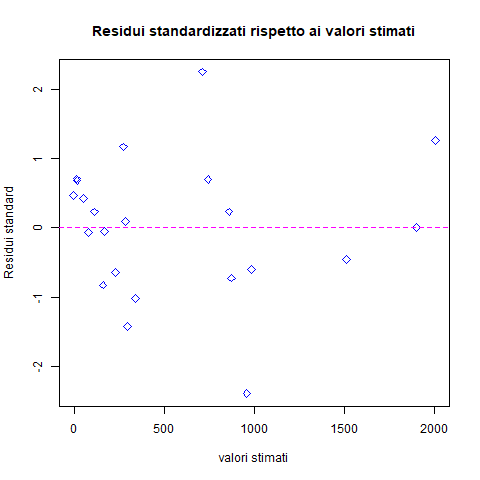


Grafico che mostra i residui standardizzati in funzione dei valori stimati.

plot(stime, residuistandard, main="Residui standardizzati rispetto ai valori stimati", xlab="valori stimati"

, ylab="Residui standard", pch=5, col="blue")

abline (h=0, col ="magenta",lty =2)



La linea tratteggiata è posizionata su 0 che indica la media campionaria dei residui. Si nota che i punti sono disposti casualmente attorno alla retta orizzontale e non si evidenzia nessun comportamento particolare nella distribuzione dei punti. La maggior parte dei punti sono concentrati nell’intervallo [-1,1] pertanto gli scostamenti dei valori osservati rispetto ai valori stimati risultano essere molto bassi. Solo per qualche regione tali scostamenti sono più elevati come Liguria, Emilia-Romagna, Abruzzo, Sicilia e Sardegna.

Anche in questo caso il coefficiente di determinazione è prossimo ad 1, infatti vale 0.9954. Il modello di regressione lineare multipla descrive bene i dati considerati.



# Analisi dei cluster

L’**analisi dei cluster** è una tecnica matematica usata in informatica e altre discipline, essa si basa sul considerare diversi tipi di dati (numerici, persone, misure) ed unirli in gruppi che contengono tutti elementi che hanno somiglianze tra di loro. La creazione dei cluster può essere effettuata in diversi metodi, ma tutte le tecniche hanno in comune lo scopo di rendere quanto più possibili omogenei gli elementi all’interno di un gruppo e rendere quanto il più eterogenei i gruppi così che il grado di associazione sia alto tra membri dello stesso gruppo e basso tra membri di gruppi diversi.

Le tecniche di raggruppamento tendono ad unire quei dati che sono tra di loro simili e svolgono questo lavoro basandosi sul concetto che ogni elemento di un certo insieme di dati ha delle caratteristiche osservabili, possono essere il colore degli occhi per le persone, o possono essere le denunce al numero verde 1522 fatte di anno in anno per una regione ed è per questo che si usano le funzioni distanza tra i vettori delle caratteristiche che servono a calcolare le misure metriche di somiglianza.

La misura di distanza utilizzata nel progetto è la **metrica Euclidea** così definita:

Dove è il valore della k-esima caratteristica dell’individuo I.

Ottenuta la distanza dei vari elementi è necessario raggrupparli in cluster. Esistono due tipologie per creare i cluster.

* **Metodi gerarchici**: mirano a costruire gerarchie di cluster; si dividono in due tipologie di approcci diversi: L’approccio agglomerativo è un approccio “bottom-up”, si parte dall’inserire ogni elemento in un singolo cluster e si procede ad accorparli a due a due; l’approccio divisivo è un approccio “top-down” che da un singolo cluster che comprende tutti gli elementi viene diviso in tanti sotto cluster. Tutti i metodi gerarchici producono una struttura ad albero chiamata “dendogramma”.
* **Metodi non gerarchici**: permettono di riposizionare elementi di un cluster qualora venga notato che un elemento piazzato in cluster conviene spostarlo in un altro, di questo metodo fa parte l’algoritmo k-means.

Tra i metodi gerarchici ci sono:

* **Metodo del legame singolo**: Presa la matrice delle distanze, si parte dalla distanza 0; si trovano gli elementi che hanno la distanza minore, si rimuovono quegli elementi e si cercano i primi elementi che hanno la distanza minore. Se ci sono più coppie di elementi che hanno la distanza minore nella matrice se ne sceglie uno arbitrariamente.
* **Metodo del legame completo**: La distanza tra due gruppi g1 e g2, con n1 e n2 individui, è definita come la massima tra tutte le distanze di n1 e n2, questo metodo privilegia la differenza tra i gruppi piuttosto che l’omogeneità del gruppo stesso.
* **Metodo del legame medio**: nel metodo del legame medio si considera, come distanza tra due gruppi, la media di tutte le distanze calcolate a due a due tra tutti gli elementi dei due gruppi
* **Metodo del centroide**: la distanza tra i gruppi g1 e g2 è calcolata sulle medie campionarie dei due gruppi. La particolarità di questo metodo è che tende ad avere un effetto gravitazionale: I gruppi più grandi tendono ad assorbire i gruppi più piccoli.
* **Metodo della mediana**: il metodo è simile a quello del centroide, ma non è dipendente dalla numerosità del gruppo. Quando due gruppi si uniscono, il nuovo centroide è calcolato come la semisomma dei due gruppi precedenti.

Tra i metodi non gerarchici, il metodo usato nel progetto è stato “**k-means**”, l’algoritmo funziona in diversi step:

1. Si fissano a priori il numero dei cluster scegliendo però elementi che hanno determinate caratteristiche
2. Si considerano tutti gli elementi e si attribuisce ad ognuno un cluster basandosi sulla distanza minore dal punto di riferimento scelto per ogni cluster
3. Si ricalcolano i centroidi dei k gruppi costituendo il nuovo punto di riferimento per i cluster così ottenuti
4. Si rivalutano le distanze per ogni unità rispetto ai centroidi dei vari cluster. Se un elemento x ha una distanza minore ad un altro centroide rispetto a quello del proprio cluster, si riposiziona l’elemento.
5. Si ricalcolano i centroidi.
6. Si ripete dallo step 4, se si arriva ad un punto in cui non ci sono stati spostamenti tra elementi dei cluster, l’algoritmo si conclude.

Per la suddivisione in cluster si è scelto di considerare la suddivisione in 2 cluster. Tuttavia, al posto di considerare il data frame con i dati originali, si è scelto di scalarli per ottenere dei dati più piccoli.

Il seguente codice permette di calcolare la matrice delle distanze euclidee a partire dal data frame Z scalato.

d<-dist(Z, method="euclidean", diag=TRUE, upper=TRUE)

**Metodo del legame singolo**

hls<-hclust(d, method="single")

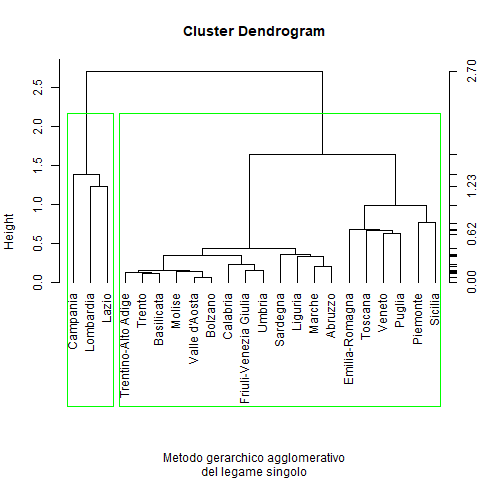
png("grafici/cluster/dendrogrammaUtenti\_LegameSingolo.png")

plot(hls, hang=-1, xlab="Metodo gerarchico agglomerativo", sub="del legame singolo")

rect.hclust(hls, k=2, border="green")

axis(side=4, at=round(c(0, hls$height),2))

dev.off()



**Metodo del legame medio**

hlm<-hclust(d, method="average")

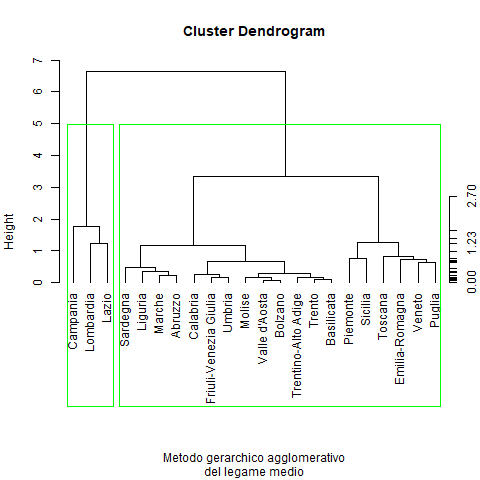
png("grafici/cluster/dendrogrammaUtenti\_LegameMedio.png")

plot(hlm, hang=-1, xlab="Metodo gerarchico agglomerativo", sub="del legame medio")

rect.hclust(hlm, k=2, border="green")

axis(side=4, at=round(c(0, hls$height),2))

dev.off()



**Metodo del legame completo**

hlc<-hclust(d, method="complete")

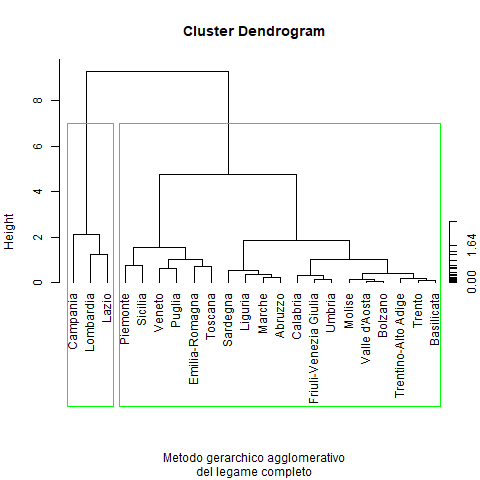
png("grafici/cluster/dendrogrammaUtenti\_LegameCompleto.png")

plot(hlc, hang=-1, xlab="Metodo gerarchico agglomerativo", sub="del legame completo")

rect.hclust(hlc, k=2, border="green")

axis(side=4, at=round(c(0, hls$height),2))

dev.off()



**Metodo del centroide**

d2<-d^2

hc<-hclust(d2, method="centroid")

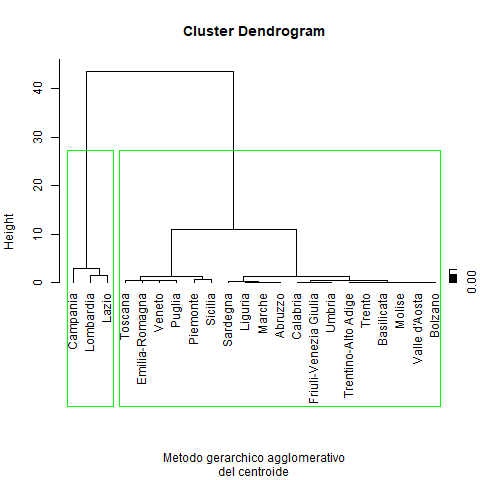
png("grafici/cluster/dendrogrammaUtenti\_MetodoCentroide.png")

plot(hc, hang=-1, xlab="Metodo gerarchico agglomerativo", sub="del centroide")

rect.hclust(hc, k=2, border="green")

axis(side=4, at=round(c(0, hls$height),2))

dev.off()



**Metodo della mediana**

hmed<-hclust(d2, method="median")

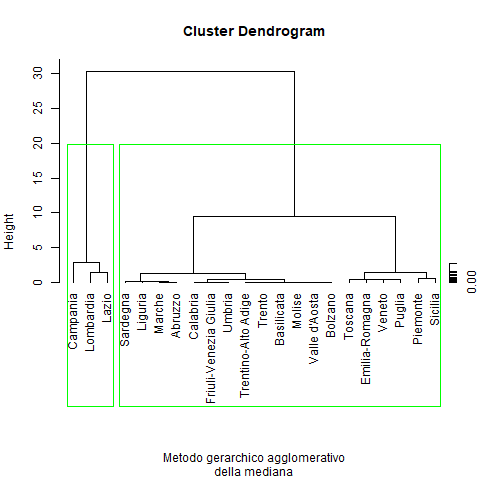
png("grafici/cluster/dendrogrammaUtenti\_MetodoMediana.png")

plot(hmed, hang=-1, xlab="Metodo gerarchico agglomerativo", sub="della mediana")

rect.hclust(hmed, k=2, border="green")

axis(side=4, at=round(c(0, hls$height),2))

dev.off()



Tutti i metodi gerarchici: legame singolo, legame medio, legame completo, metodo del centroide e metodo della mediana hanno fornito il seguente partizionamento in due cluster.

Primo cluster: 19 individui

Secondo cluster: 3 individui

|  |  |
| --- | --- |
| Cluster 1 | Piemonte, Valle d’Aosta, Liguria, Trentino-Alto Adige, Trento, Bolzano, Veneto, Friuli-Venezia Giulia, Emilia-Romagna, Toscana, Umbria, Marche, Abruzzo, Molise, Puglia, Basilicata, Calabria, Sicilia, Sardegna |
| Cluster 2 | Lombardia, Lazio, Campania |

Per valutare quanto questa suddivisione è “buona” si calcolano le misure di non omogeneità.

Si mostra in R il codice per il calcolo delle misure di non omogeneità per i cluster ottenuti con il metodo del legame singolo. Siccome il partizionamento ottenuto è uguale anche per gli altri metodi i risultati saranno uguali.

n<-nrow(Z)

trH<-(n -1)\*sum (apply(Z,2,var))

taglio<-cutree(hls, k=2)

num <-table (taglio)

tagliolist<-list(taglio)

agvar <- aggregate (Z, tagliolist , var)[, -1]

trH1 <-(num [[1]]-1) \* sum (agvar [1, ])

trH2 <-(num [[2]]-1) \* sum (agvar [2, ])

trB<-trH-trH1-trH2

rapportoLegameSingolo<-trB/trH

La misura di non omogeneità totale trH nel data frame considerato risulta essere uguale a 168.

La misura di non omogeneità all’interno del primo cluster trH1 risulta essere uguale a 52.6722.

La misura di non omogeneità all’interno del secondo cluster trH2 risulta essere uguale a 2.631797.

Pertanto, la misura di non omogeneità tra i cluster risulta essere trB=trT-trH1-trH2= 112.696.

Il rapporto =  **0.6708096**.

La suddivisione ottenuta con i metodi gerarchici risulta essere abbastanza buona in quanto si avvicina quasi al 70%.

Successivamente, è stato utilizzato anche il metodo non gerarchico K-means considerando sempre la suddivisione delle regioni in due cluster.

Il metodo non gerarchico K-means ha fornito il seguente partizionamento in due cluster.

Primo cluster: 9 individui

Secondo cluster: 13 individui

|  |  |
| --- | --- |
| Cluster 1 | Piemonte, Lombardia, Veneto, Emilia-Romagna, Toscana, Lazio, Campania, Puglia, Sicilia |
| Cluster 2 | Valle d’Aosta, Liguria, Trentino-Alto Adige, Trento, Bolzano, Friuli-Venezia Giulia, Umbria, Marche, Abruzzo, Molise, Basilicata, Calabria, Sardegna |

km <-kmeans (Z, centers=2, iter.max =10, nstart =1)

rapportoKMeans<-km$betweenss/km$totss

Il rapporto = **0.7129243**.

La suddivisione in cluster ottenuta con il metodo non gerarchico K-means risulta essere migliore in quanto supera il 70%.

# Bibliografia

|  |  |
| --- | --- |
| [1] | «Che cos'è la violenza di genere,» [Online]. Available: https://www.comune.venezia.it/it/content/cos%C3%A8-la-violenza-di-genere#:~:text=Le%20Nazioni%20Unite%20in%20occasione,minaccia%20di%20tali%20atti%2C%20la. |