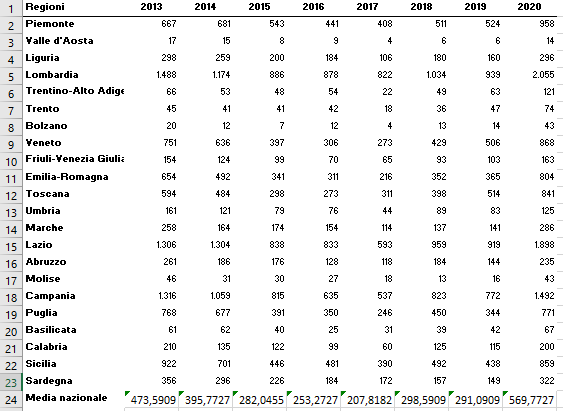
# Statistica descrittiva univariata

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, verranno esaminate nei dettagli le curve relativi ai dati della regione Campania e la media delle chiamate effettuate sull’intero territorio nazionale.



Sono stati creati due barplot, uno relativo ai dati della media nazionale e l’altro relativo ai dati in Campania.





## Funzione di distribuzione continua

Per calcolare la funzione di distribuzione continua sono stati suddivisi le osservazioni sono state suddivise in 3 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]. 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. 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 dati per illustrare alcune caratteristiche della distribuzione di frequenza come centralità, dispersione, forma e la presenza di eventuali valori anomali.



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. Dai grafici a barre in alto è possibile inoltre determinare la moda, che in entrambi i casi è associata all’anno 2020. In generale, la curva dei due dati è sostanzialmente simile anche se i dati relativi all’intera nazione sono più bassi in quanto sono calcolati dalla media di tutte le nazioni, che viene fortemente influenzata dai valori bassi presenti in molte regioni con meno abitanti rispetto alla Campania.





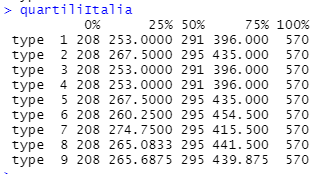
Per individuare la moda si considerano gli istogrammi delle frequenze dei dati considerando la loro suddivisione nelle seguenti classi: C1 = [0, 500), C2 = [500, 1000), C3 = [1000, 1500) C1 = [1500, 2000), C2=[2000, 2500].

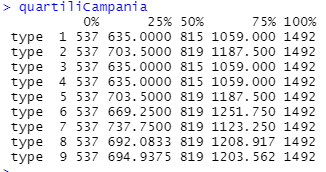
La classe modale per l’Italia è la prima, in particolare tutti i valori sono concentrati in quella classe. Per la Campania invece la classe modale è la seconda.





**Quartili con i differenti algoritmi di R**





**Indici di dispersione**



## Forma della distribuzione di frequenze

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



**Curtosi campionaria.** 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.

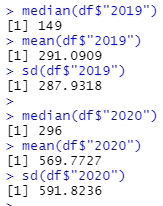


# Statistica descrittiva bivariata

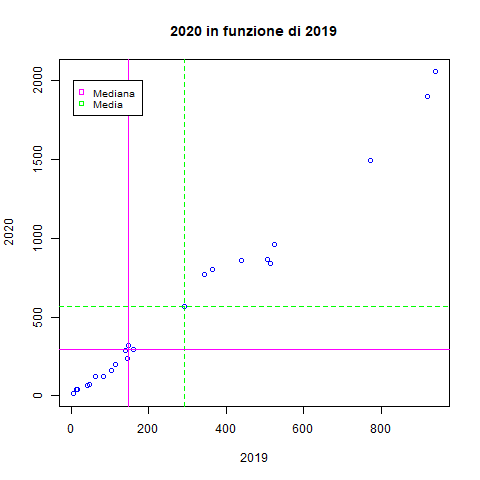
## Regressione lineare semplice

Le variabili che vengono considerate in quest’analisi sono le colonne della tabella relative ai dati del 2019 e del 2020. In particolare, la variabile indipendente è 2019, quella dipendente è 2020. Si calcolano i valori degli indici statistici mediana, media e deviazione standard dei dati relativi alle variabili considerate.

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



Successivamente, si realizza lo scatterplot ponendo sulle 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 sono posizionati lungo una retta ascendente quindi si può dedurre che esiste una correlazione positiva tra le variabili considerate.

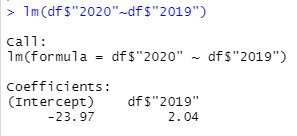
Per vedere se esiste tale correlazione si calcolano la covarianza e la correlazione campionaria. Da questo calcolo si evince che i dati dei due vettori 2019 e 2020 sono positivamente correlati essendo la covarianza positiva. Inoltre, il coefficiente di correlazione è uguale a 0.9923597 che è prossimo ad 1, quindi come indicato dallo scatterplot esiste una forte correlazione lineare tra i dati del 2019 e i dati del 2020.



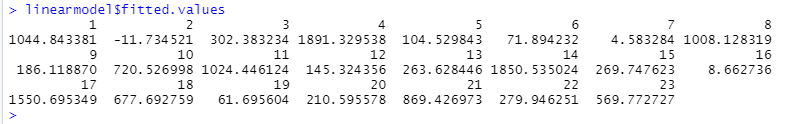
Il seguente grafico mostra lo scatterplot relativo ai dati del 2019 e del 2020 con la retta interpolante stimata.



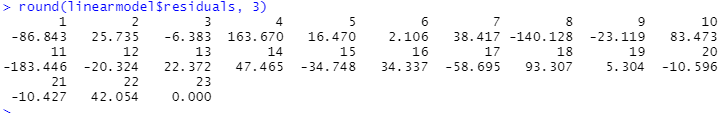
Il seguente codice permette di ottenere il modello di regressione lineare per le due variabili. In particolare, l’intercetta vale -23.97, mentre il coefficiente angolare vale 2.04. Siccome il coefficiente angolare è positivo, la retta è ascendente.



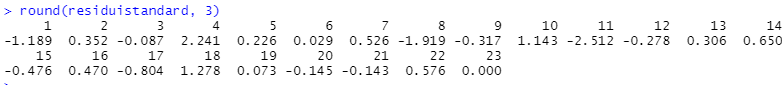
Il codice seguente permette di visualizzare i valori stimati.



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



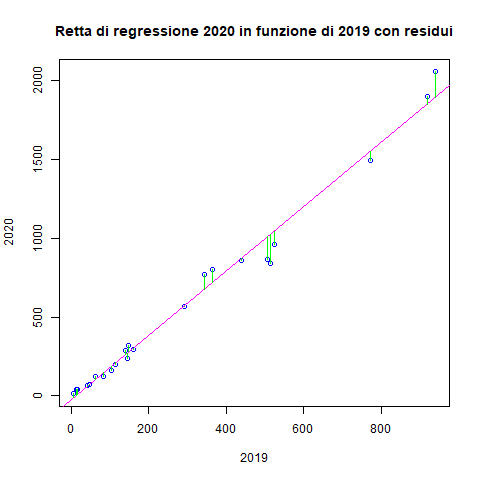
Valore dei residui standardizzati rispetto alla deviazione standard. Si può osservare che i valori sono molto piccoli.



Le seguenti linee di codice mostrano i valori della mediana, della varianza e della deviazione standard dei residui.







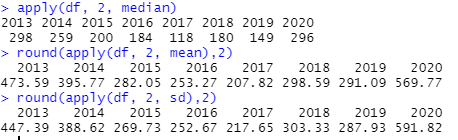
Per valutare quanto la retta di regressione si adatta ai dati si calcola il coefficiente di determinazione che si calcola effettuando il rapporto tra la varianza dei valori stimati tramite la retta di regressione e la varianza dei valori osservati. In questo caso il coefficiente di correlazione vale 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.



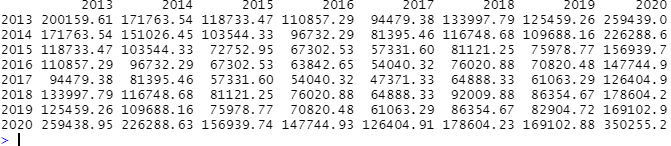
## Regressione lineare multipla

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

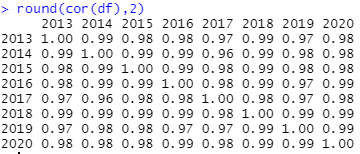
Valore degli indici di posizione e di dispersione (mediana, media e deviazione standard) relativi alle variabili:



Matrice delle covarianze:



La matrice delle correlazioni che contiene tutte le correlazioni lineari tra le coppie di variabili.

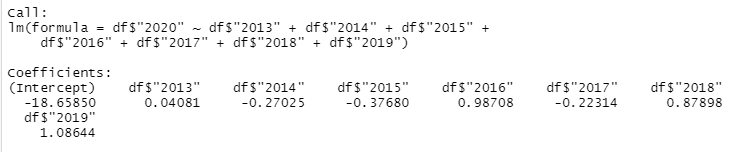


Si nota che esiste una forte correlazione lineare tra tutte le variabili considerate.

Il seguente grafico visualizza in un’unica finestra tutti gli scatterplot ottenuti mettendo in relazione le varie coppie di variabili. Da tale grafico si può dedurre che le variabili sono altamente correlate e si intuisce che avranno un coefficiente di correlazione quasi pari ad 1.



Utilizzando il modello di regressione lineare multipla si ottiene:

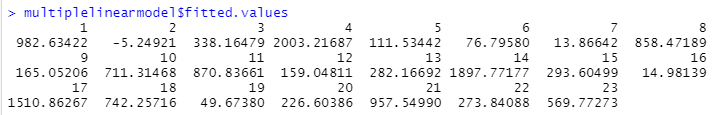


Da cui si ricava che l’intercetta è -18.65850 e i regressori sono: 0.04081, -0.27025, -0.37680, 0.98708, -0.22314, 0.87898, 1.08644.

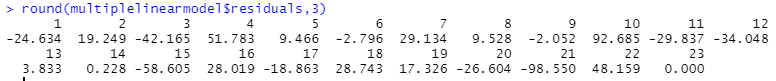
I segni dei regressori b1, b4, b6, b7 sono positivi: questo indica che all’aumentare del numero di utenti nel 2013, 2016, 2018 e 2019 aumenta il numero di utenti nel 2020. Mentre i regressori b2, b3, b5 sono negativi quindi all’aumentare del numero di utenti nel 2014, 2015, 2017 diminuisce il numero di utenti nel 2020.

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

Valori stimati rispetto al modello di regressione multipla.



Residui dei valori osservati rispetto ai valori stimati.



Valori dei residui standardizzati rispetto alla deviazione standard.

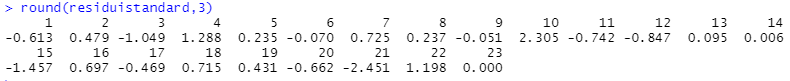
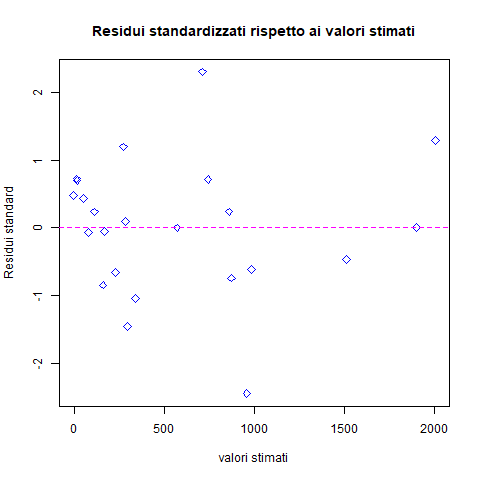


Grafico che mostra i residui standardizzati in funzione dei valori stimati.



I punti indicano dove si collocano i residui rispetto ai valori stimati. Non si evidenzia nessuna tendenza particolare rispetto alla retta orizzontale che rappresenta la media dei residui (0).

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.



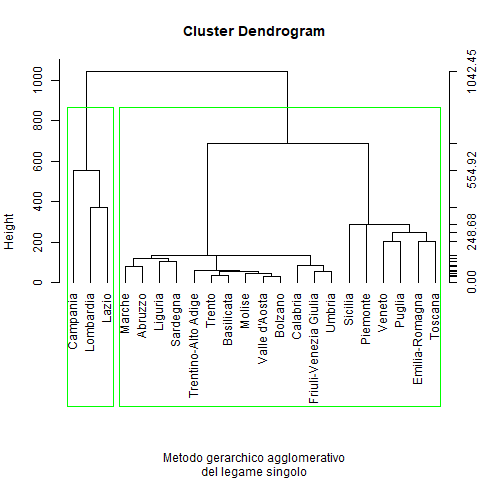
# Analisi dei cluster

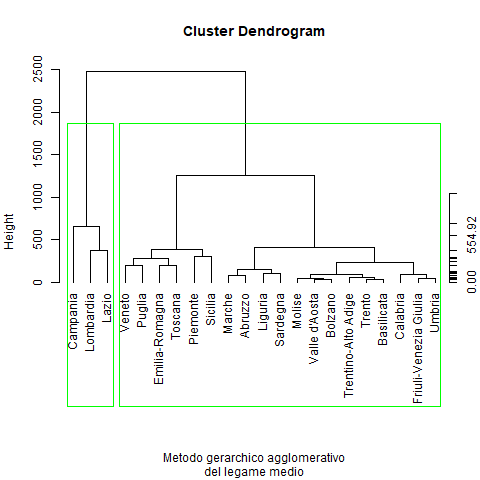
Tutti i metodi gerarchici: legame singolo, legame medio, legame completo, metodo del centroide e metodo della mediana hanno fornito il seguente partizionamento in due cluster.

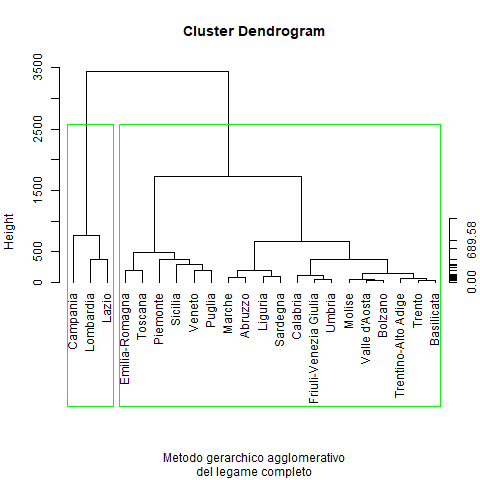
Primo cluster: 19 individui

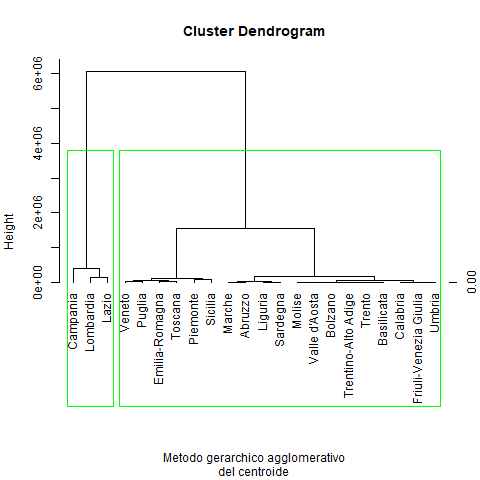
Secondo cluster: 3 individui

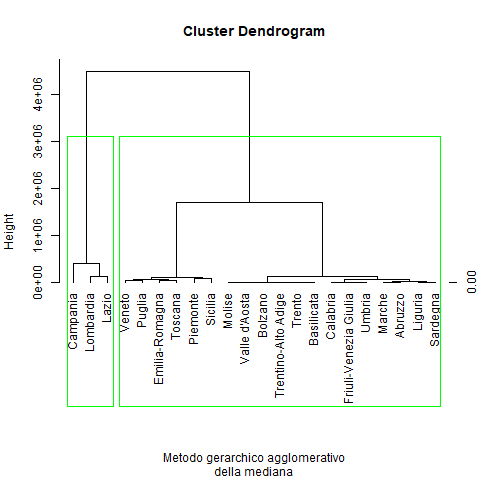
|  |  |
| --- | --- |
| Cluster 1 | Piemonte, Valle d’Aosta, Liguria, Trentino-Alto Adige, Trento, Bolzano, Veneto, Friuli-Venezia Giulia, Emilia-Romagna, Toscana, Umbria, Marche, Abruzzo, Molise, Puglia, Basilicata, Calabria, Sicilia, Sardegna |
| Cluster 2 | Lombardia, Lazio, Campania |











La misura di non omogeneità totale trT nel data frame considerato risulta essere uguale a 23327101.

La misura di non omogeneità all’interno del primo cluster trH1 risulta essere uguale a 7304986.

La misura di non omogeneità all’interno del secondo cluster trH2 risulta essere uguale a 340968.7.

Pertanto, la misura di non omogeneità tra i cluster risulta essere trB=trT-trH1-trH2= 15681146.

Il rapporto =  **0.6722287**.

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 |

La misura di non omogeneità totale trT nel data frame considerato risulta essere uguale a 23327101.

La misura di non omogeneità all’interno del primo cluster trH1 risulta essere uguale a 5812676.7.

La misura di non omogeneità all’interno del secondo cluster trH2 risulta essere uguale a 637988.3.

Pertanto, la misura di non omogeneità tra i cluster risulta essere trB=trT-trH1-trH2=16876436.

Il rapporto = **0.7234691**.

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