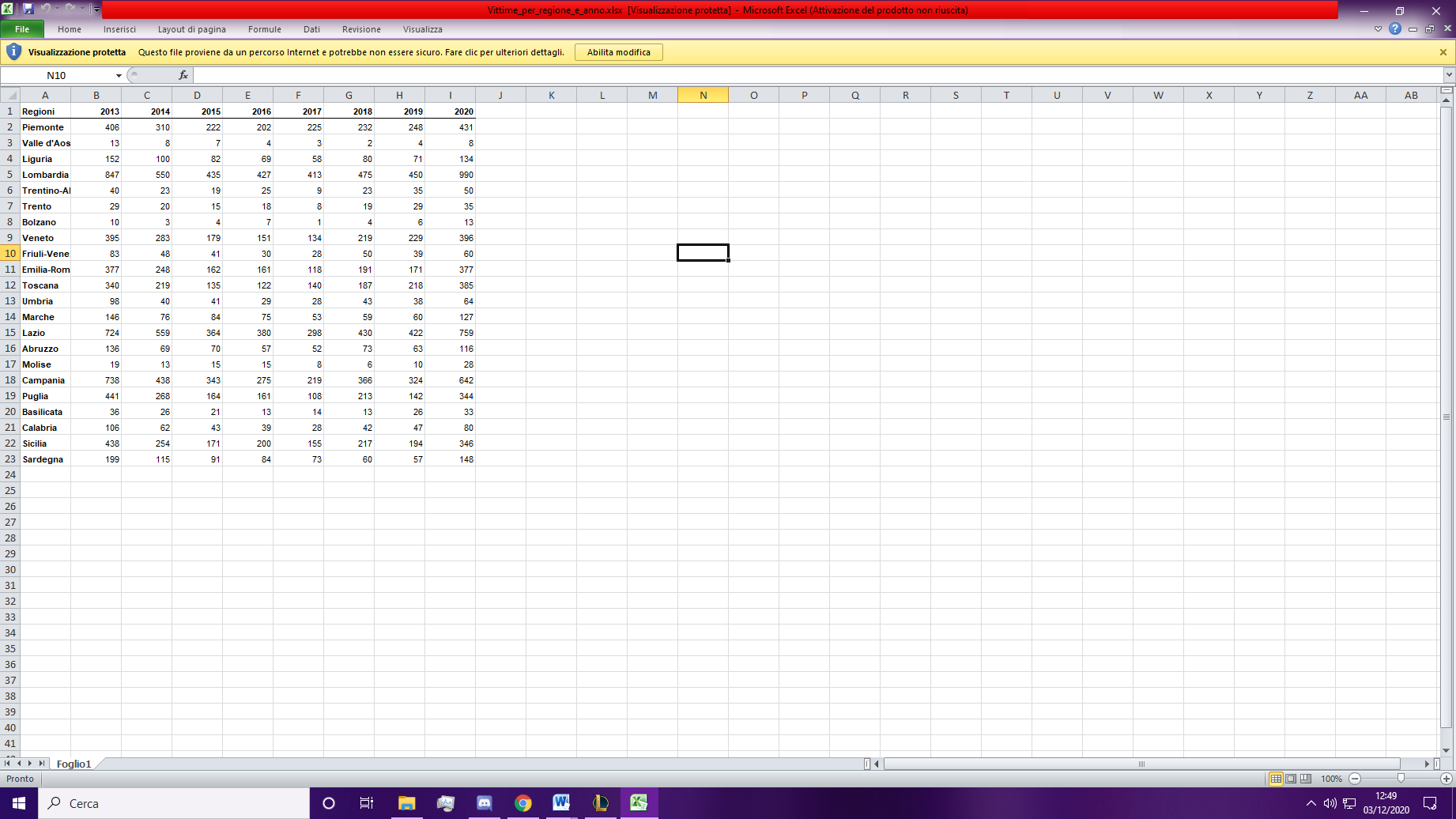
Statistica descrittiva univariata

Per questo progetto verranno presentati i dati delle vittime che hanno chiamato il 1522 (numero verde per vittime di violenza e stalking). Questo progetto si differenza dalla tabella degli utenti perché le vittime sono tutte quelle persone che hanno chiamato e sporto poi denuncia verso gli aggressori, mentre la tabella degli utenti comprende anche persone che hanno chiamato a scopo informativo e/o chiamato per denunciare eventi che non riguardano se stessi.

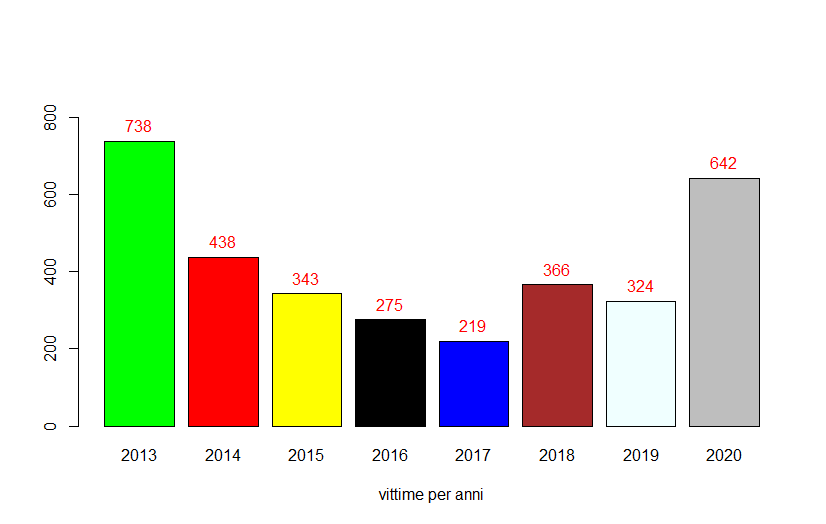
La tabella delle vittime fa riferimento al periodo marzo-giugno dal 2013 al 2020.

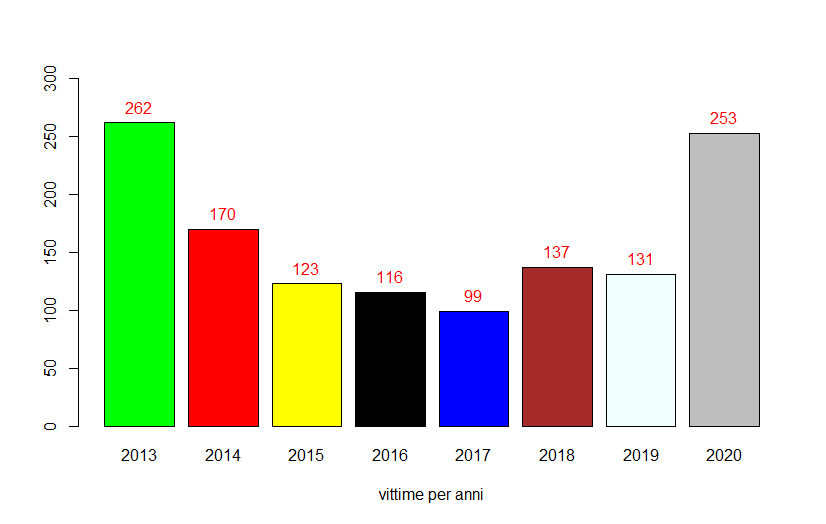
Tutti i confronti avverranno tra la media delle regioni delle vittime in Italia e le vittime in campania.

Di seguito verrà mostrata la tabella utilizzata:



Per mostrare meglio i numeri per anno, riguardante la media nazionale e i dati in campania, si è costruito questo grafici: Il primo è per le vittime in campania, il secondo rappresenta la media in Italia.





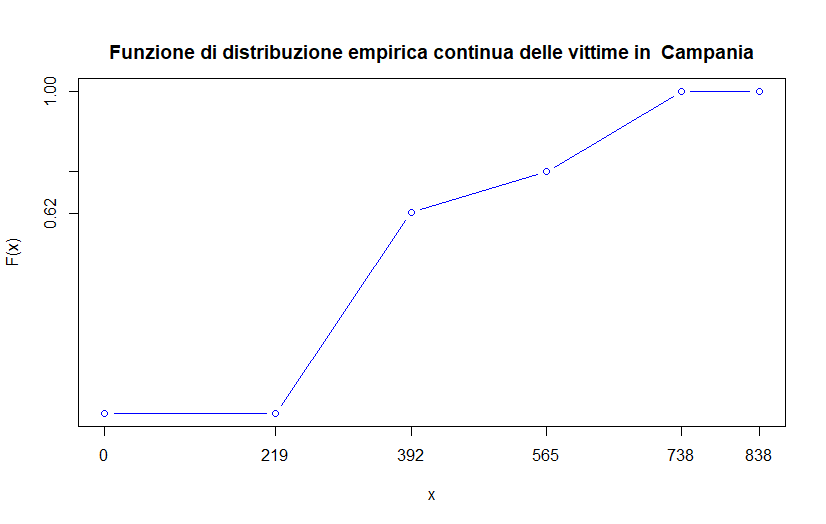
Funzione di distribuzione continua

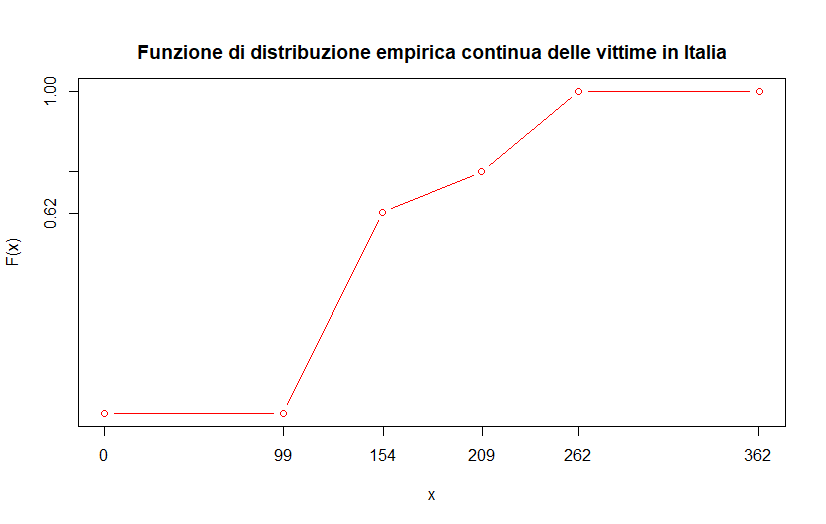
Per calcolare la funzione di distribuzione continua le osservazioni sono state divise in 3 classi:

Per la Campania: c1[219,392) , c2[392,565), c3[565,738]

Per la media nazionale: c1[99,154),c2[154,209),c3[209,262]

Di seguito verranno mostrati grafici che mostrano le frequenze di distribuzione continua in Campania e in Italia.

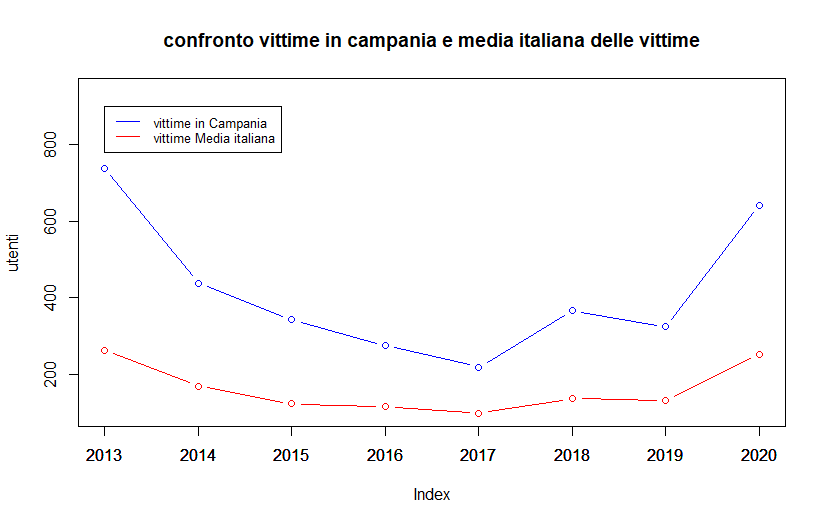




Indici di sintesi

Esistono particolari strumenti nella statistica che sono gli indici di sintesi. Questi strumenti servono a descrivere dati numerici. Esistono due tipi di questi strumenti; esistono quelli che misurano la centralità dei dati quali: media, mediana e moda; e quelli che misurano la dispersione dei dati: la varianza e la devianza standard.

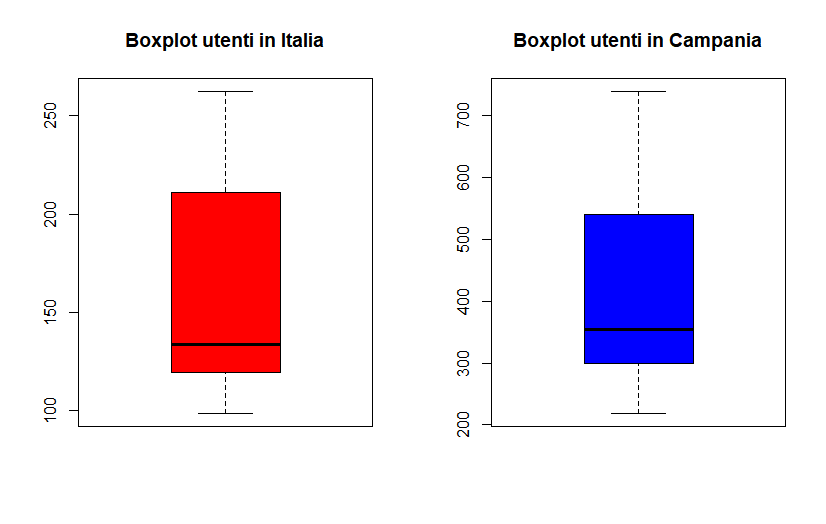
Di seguito è riportato il grafico che rappresenta le curve dei dati che si stanno analizzando, è stata preferenza del programmatore rappresentare le due curve in un solo grafo per mostrare meglio la differenza di numeri, ma andamento simile tra le vittime in media nazionale e le vittime in Campania



Il picco è presente in entrambi i casi nel 2013, per poi avere un andamento discendente fino al 2017 (anno del me too), per poi risalire dal 2018 e arrivare ad un incremento vertiginoso nel 2020.

E’ da ricordare che i numeri che si stanno analizzando fanno parte di denunce da donne, quindi, è lecito pensare che un grosso movimento quale il me too abbia dato coraggio alle donne che ricevevano abusi di farsi avanti e denunciare i propri aguzzini, nel 2020 l’obbligo della convivenza forzata ha solo incrementato quello che era già presente negli anni precedenti.

Di seguito verranno mostrati alcuni dettagli tramite boxplot riguardo le distribuzioni di frequenza e i vari quartili, la mediana e la media.



> summary(vittimecampania)

"Min. 1st Qu. Median Mean 3rd Qu. Max.

219.0 311.8 354.5 418.1 489.0 738.0

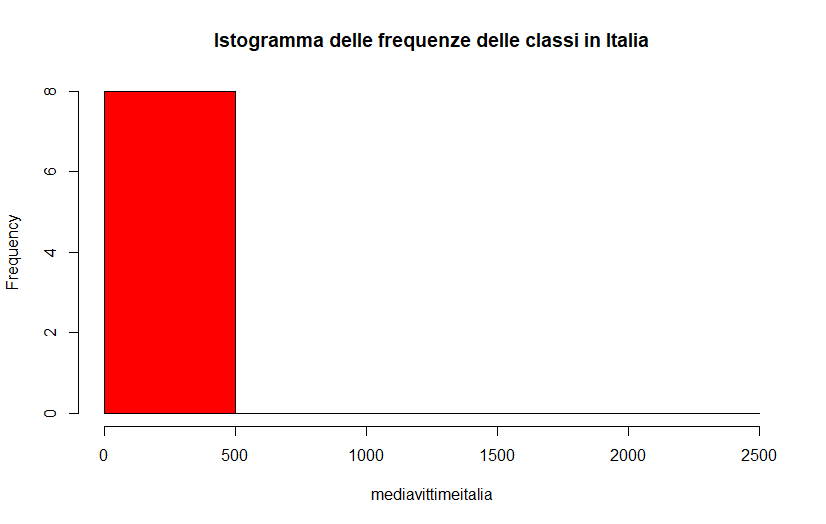
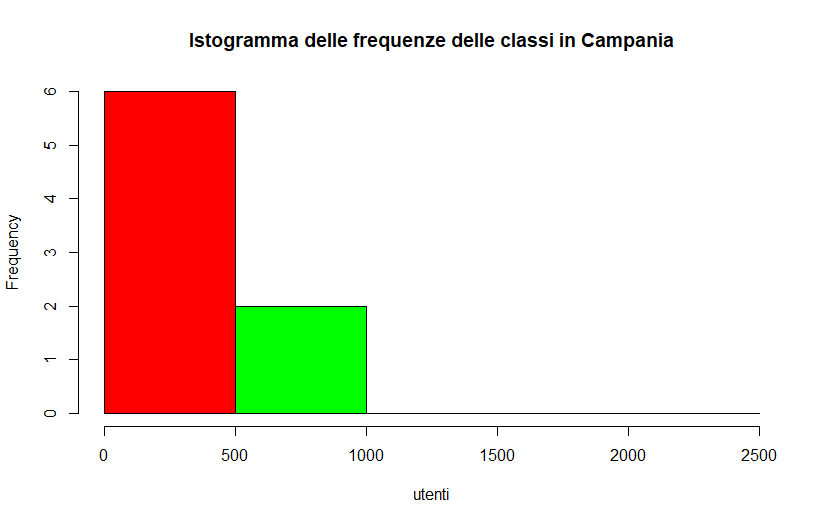
> summary(mediavittimeitalia)

Min. 1st Qu. Median Mean 3rd Qu. Max.

98.77 121.23 133.80 161.27 190.48 262.41

Per individuare la moda bisogna considerare le frequenze dei dati considerando la suddivisione nelle seguenti classi: C1 = [0, 500), C2 = [500, 1000), C3 = [1000, 1500) C4 = [1500, 2000), C5=[2000, 2500].

Sia per il primo grafico, che per il secondo, la classe modale è la prima.



**Quartili con i differenti algoritmi di R**

> print(mediavittime)

0% 25% 50% 75% 100%

type 1 98.77273 115.6364 131.0455 169.6364 262.4091

type 2 98.77273 119.3636 133.7955 211.3182 262.4091

type 3 98.77273 115.6364 131.0455 169.6364 262.4091

type 4 98.77273 115.6364 131.0455 169.6364 262.4091

type 5 98.77273 119.3636 133.7955 211.3182 262.4091

type 6 98.77273 117.5000 133.7955 232.1591 262.4091

type 7 98.77273 121.2273 133.7955 190.4773 262.4091

type 8 98.77273 118.7424 133.7955 218.2652 262.4091

type 9 98.77273 118.8977 133.7955 216.5284 262.4091

> print(campania)

0% 25% 50% 75% 100%

type 1 219 275.0000 343.0 438.00 738

type 2 219 299.5000 354.5 540.00 738

type 3 219 275.0000 343.0 438.00 738

type 4 219 275.0000 343.0 438.00 738

type 5 219 299.5000 354.5 540.00 738

type 6 219 287.2500 354.5 591.00 738

type 7 219 311.7500 354.5 489.00 738

type 8 219 295.4167 354.5 557.00 738

type 9 219 296.4375 354.5 552.75 738

**Indici di dispersione**

> var(vittimecampania)

[1] 32884.41

> sd(vittimecampania)

[1] 181.3406

> coefficienteVariazioneCampania=sd(vittimecampania)/abs(mean(vittimecampania))

> coefficienteVariazioneCampania

[1] 0.4336995

> var(mediavittimeitalia)

[1] 3954.902

> sd(mediavittimeitalia)

[1] 62.88801

> coefficienteVariazioneItalia=sd(mediavittimeitalia)/abs(mean(mediavittimeitalia))

> coefficienteVariazioneItalia

[1] 0.3899619

Forma della distribuzione di frequenze

skw(vittimecampania)

[1] 0.8126429

> skw(mediavittimeitalia)

[1] 0.8410663

>

> curt(vittimecampania)

[1] -0.7215766

> curt(mediavittimeitalia)

[1] -0.9292483

**Skewness campionaria**. Per quanto riguarda la skewness campionaria, l’indice è positivo per entrambi i dati, ciò quindi significa che nella distribuzione di frequenze, la coda di destra è più allungata.

**Curtosi campionaria.** Il valore di entrambe le curtosi campionarie è negativo e quindi la distribuzione di frequenze è più piatta di una normale.

Statistica descrittiva bivariata

Regressione lineare semplice

In questa sezione verranno utilizzate le colonne del dataset delle vittime del 2019 e del 2020, con la colonna del 2020 come variabile dipendente e quella del 2019 come variabile indipendente.

Si calcolano i valori degli indici statistici mediana, media e deviazione standard per le due colonne.

Sia mediana, che media e deviazione standard sono maggiori per la colonna relativa al 2020.

> median(dataframe$"2019")

[1] 61.5

> mean(dataframe$"2019")

[1] 131.0455

> sd(dataframe$"2019")

[1] 134.5527

> print("cambio anno")

[1] "cambio anno"

> median(dataframe$"2020")

[1] 130.5

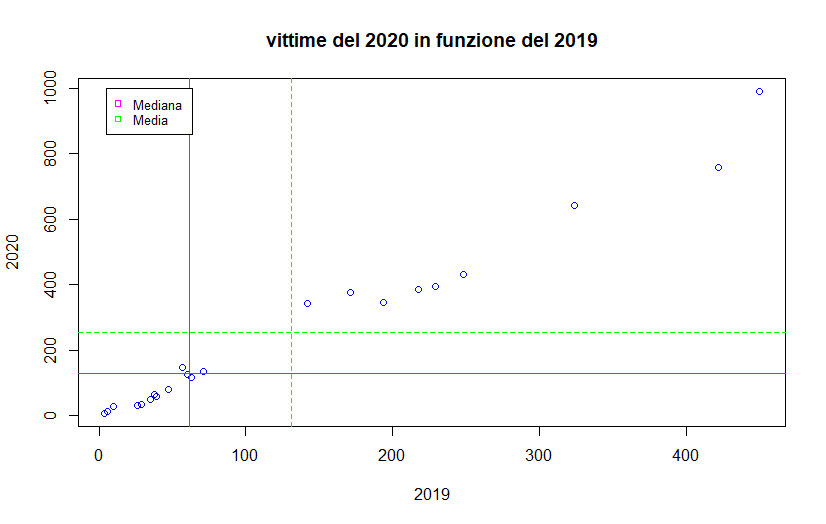
> mean(dataframe$"2020")

[1] 253

> sd(dataframe$"2020")

[1] 269.0815

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 può notare come tutti i dati siano posizionati lungo una rette ascendente quindi si può dedurre che esiste una correlazione positiva tra le variabili considerate.

Per vedere se ciò è matematicamente provato, bisogna controllare la covarianza e la correlazione campionaria: la covarianza deve essere positiva e la correlazione deve essere prossima ad 1 .

> cov(dataframe$"2019", dataframe$"2020")

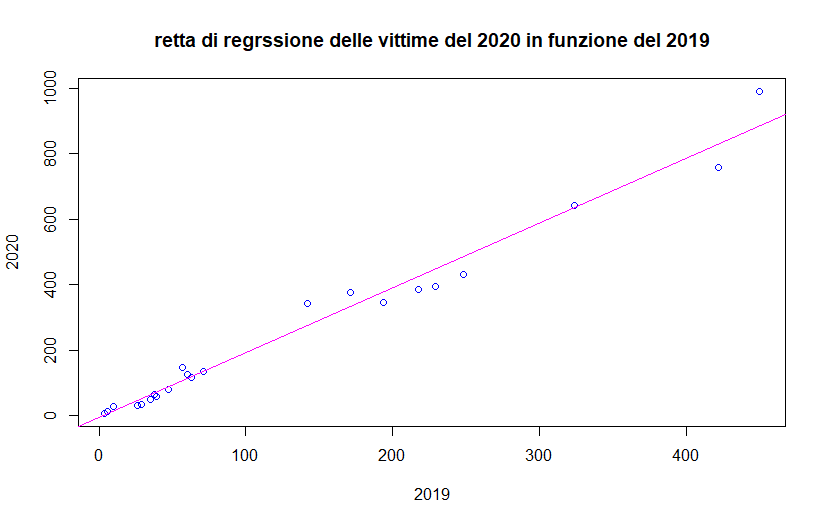
[1] 35798.62

> cor(dataframe$"2019",dataframe$"2020")

[1] 0.9887581

Entrambe le condizioni sono soddisfatte. Le due variabili sono linearmente correlati poiché la correlazione è prossima ad 1, ed è positiva poiché la covarianza è positiva.

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 -6.122, mentre il coefficiente angolare vale 1.977. Siccome il coefficiente angolare è positivo, la retta è ascendente:

Call:

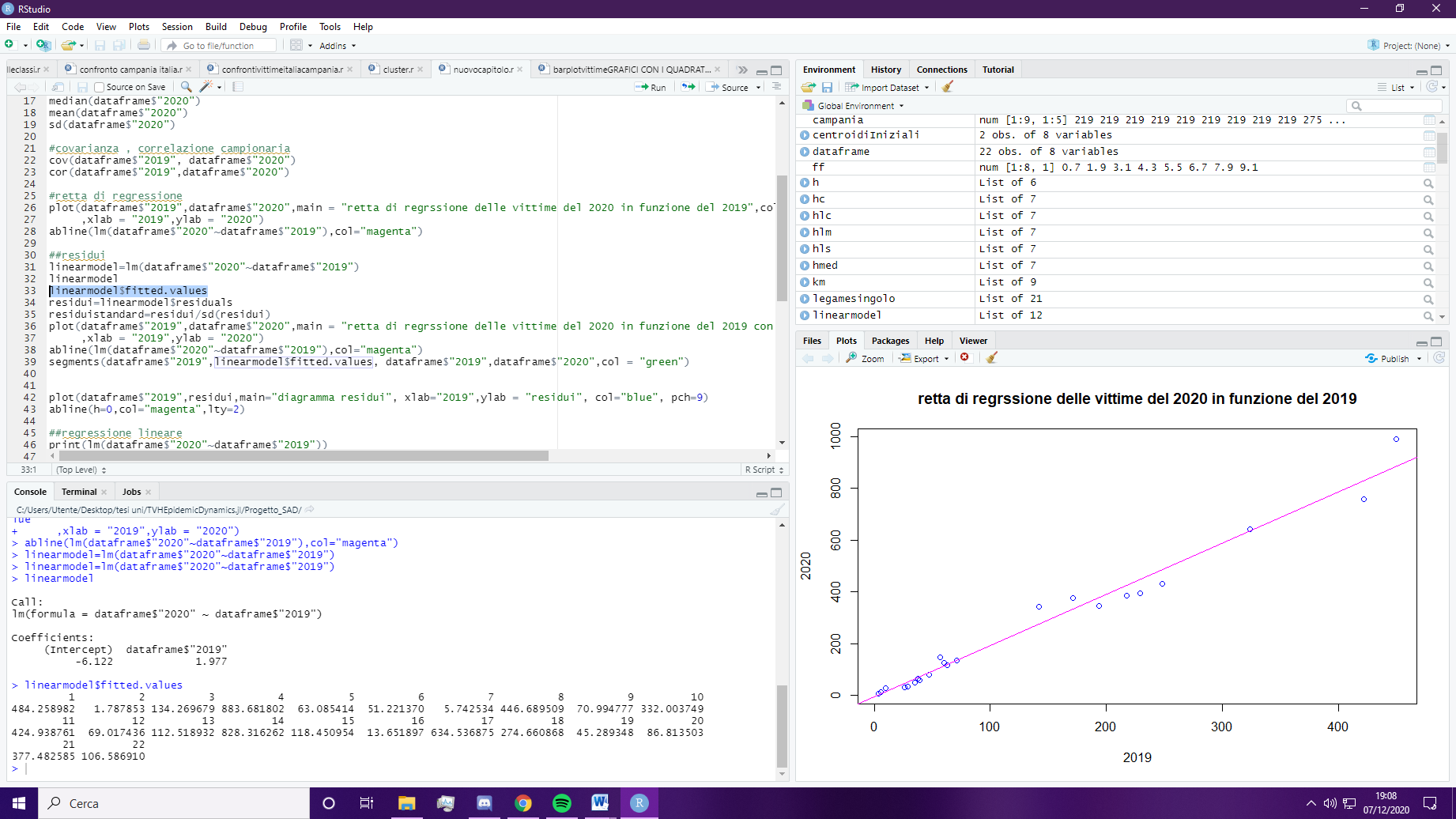
lm(formula = dataframe$"2020" ~ dataframe$"2019")

Coefficients:

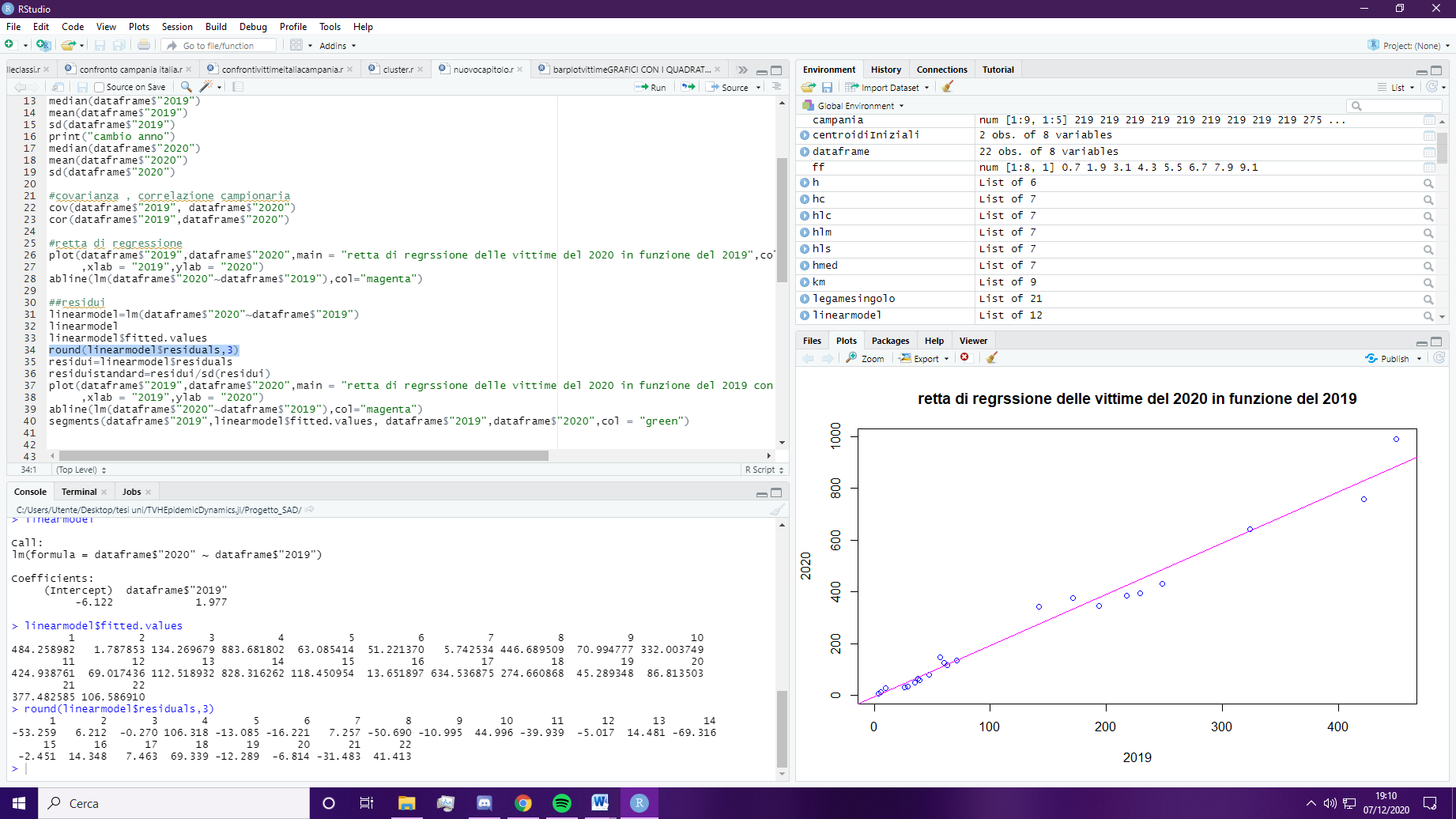
(Intercept) dataframe$"2019"

-6.122 1.977

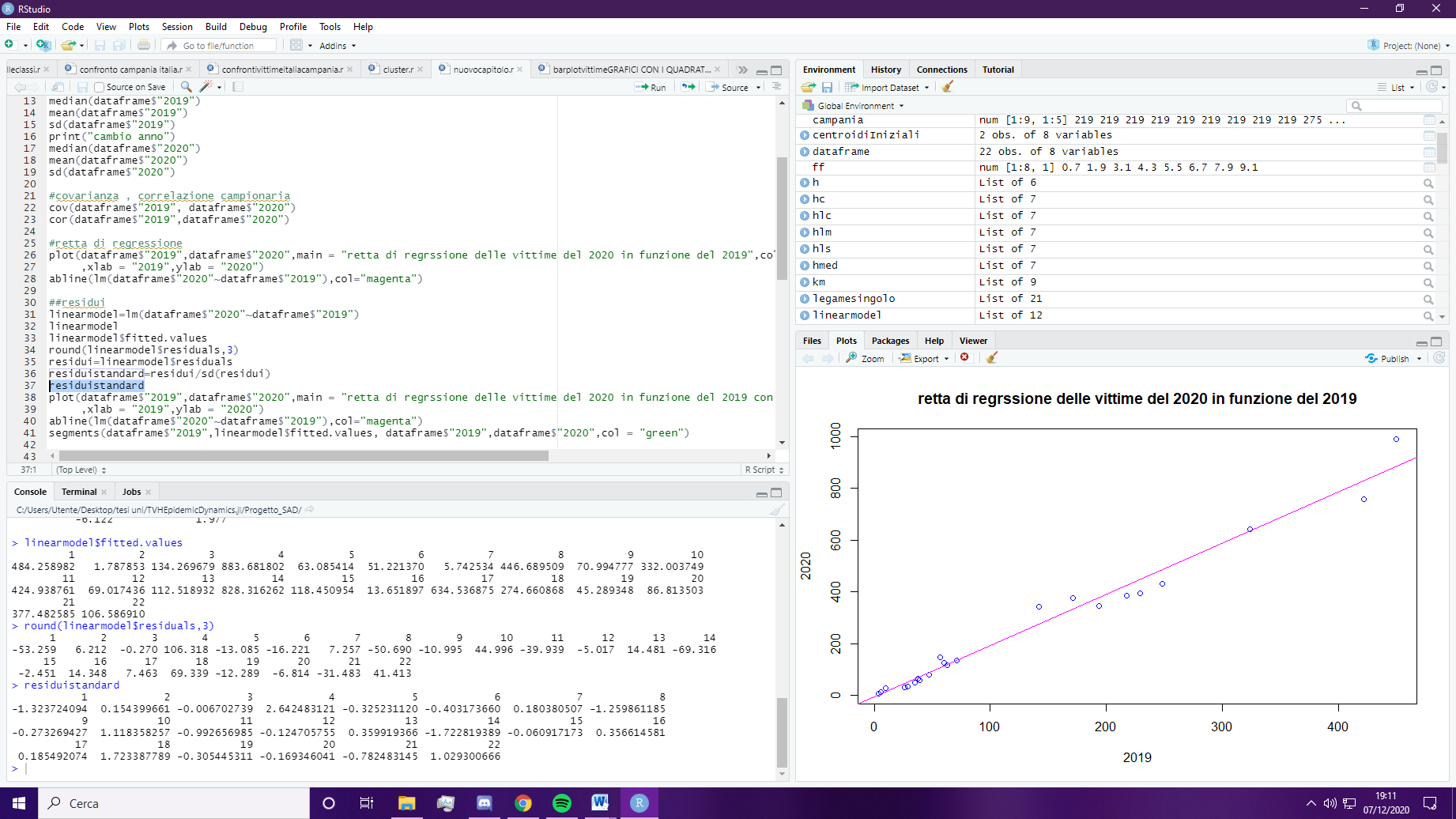
Il codice seguente permette di visualizzare i valori stimati.



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

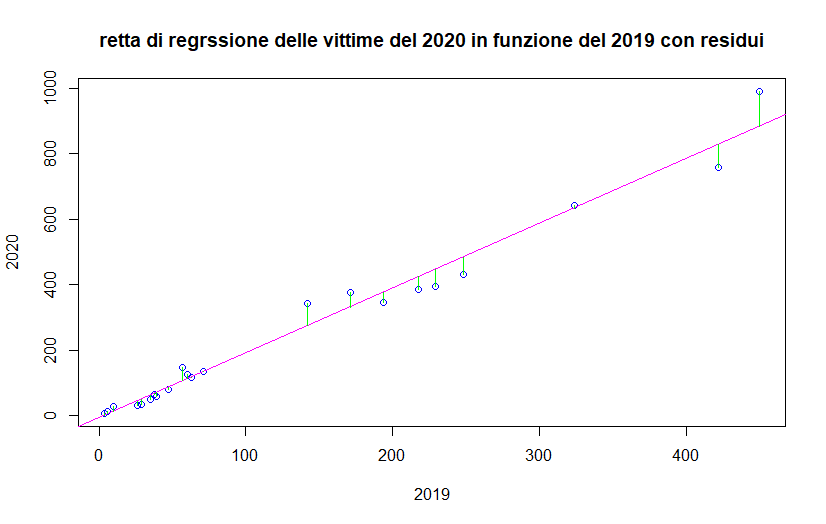


Valore dei residui standardizzati rispetto alla deviazione standard. Si può osservare che i valori sono molto piccoli.



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





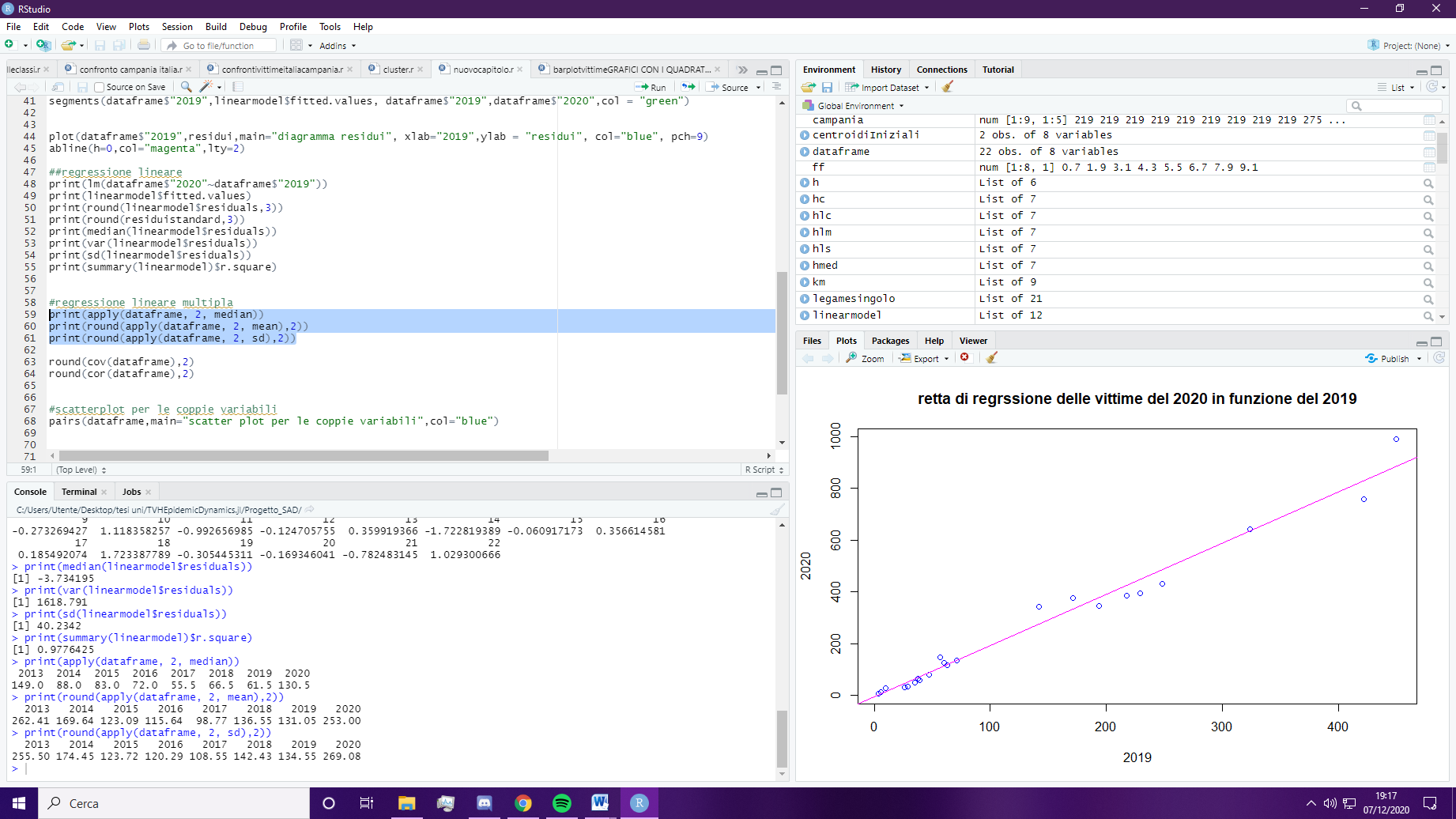
Per valutare quanto la retta di regressione si adatta ai dati si calcola il coefficiente di determinazione che si calcola effettuando il rapporto tra la varianza dei valori stimati tramite la retta di regressione e la varianza dei valori osservati. In questo caso il coefficiente di correlazione vale 0.9776425. 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.



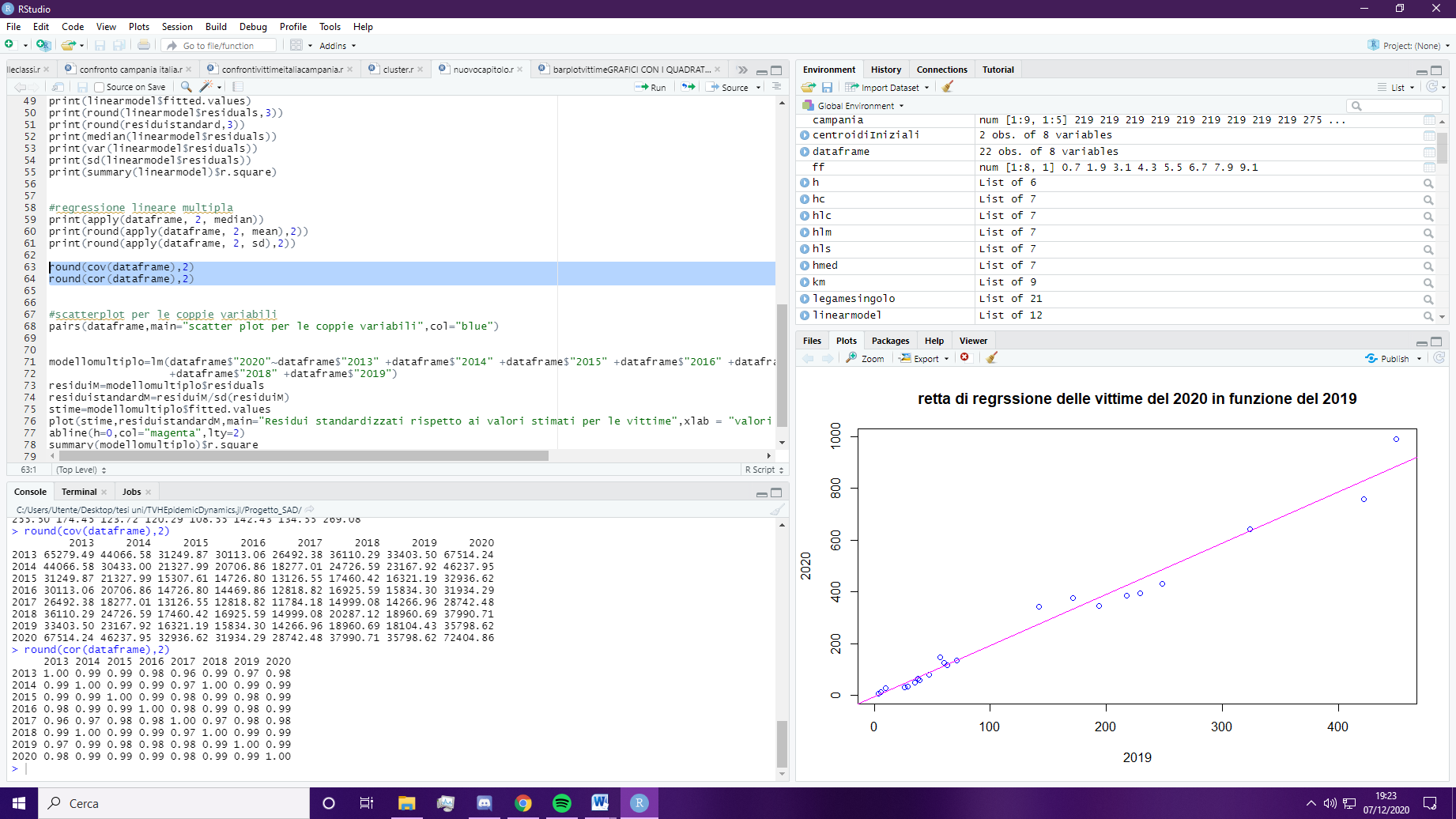
Regressione lineare multipla

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:

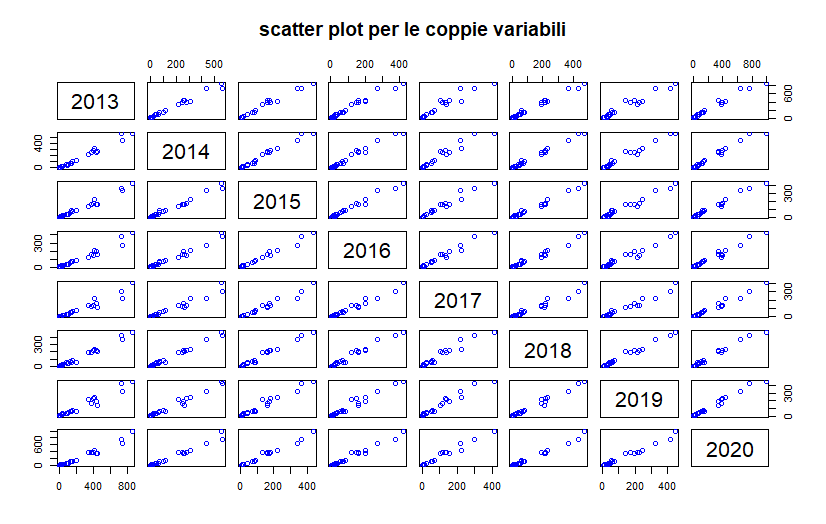


Matrice delle covarianze e sotto la matrice delle correlazioni che contiene tutte le correlazioni lineari tra le coppie di variabili:

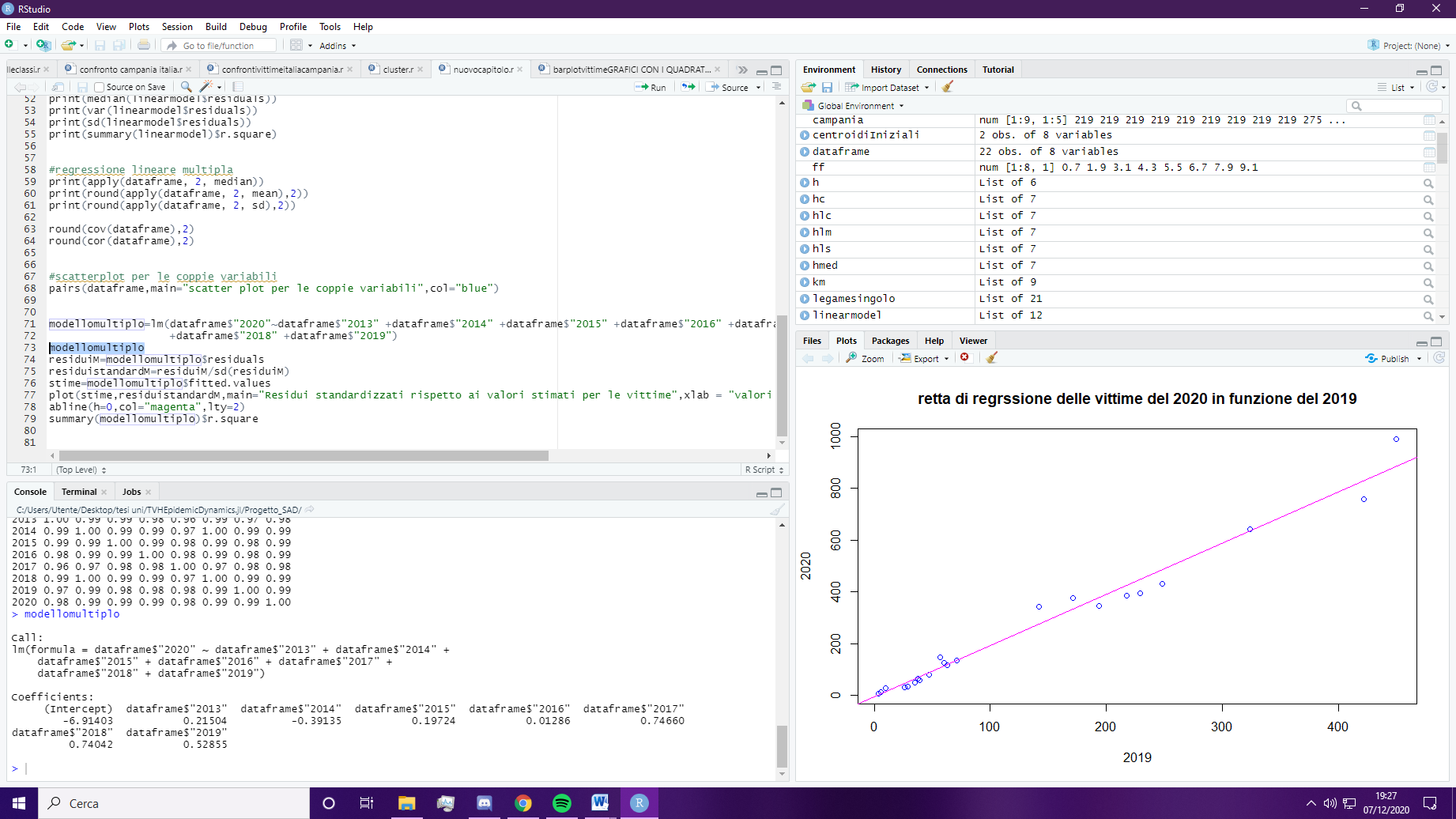


Si nota che esiste una forte correlazione lineare tra tutte le variabili considerate.

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:

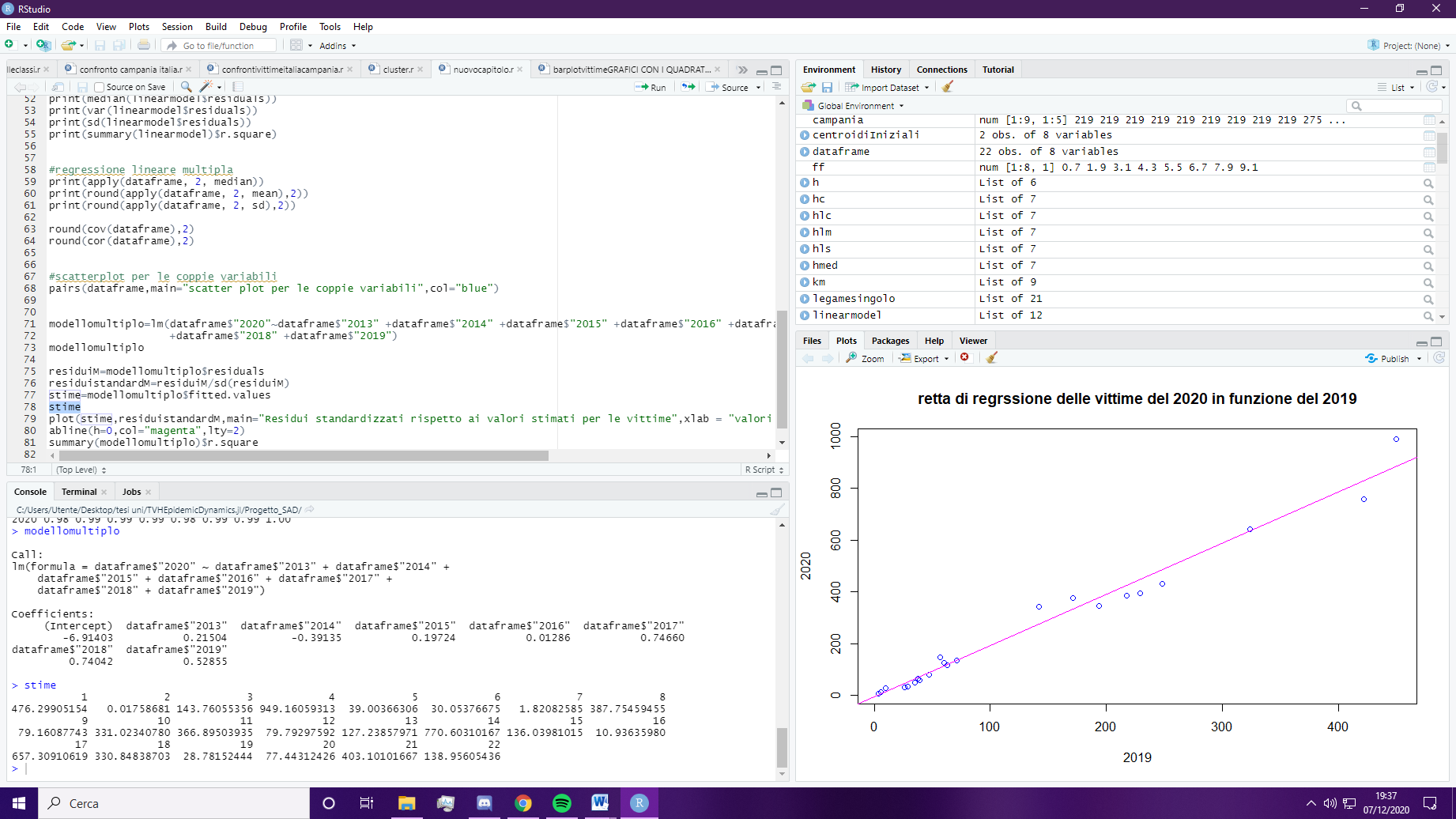


Da cui si ricava che l’intercetta è -6.91403 e i regressori sono: 0.21504, -0.39135, 0.19724, 0.01286, - 0.74660, 0.74042, 0.52855.

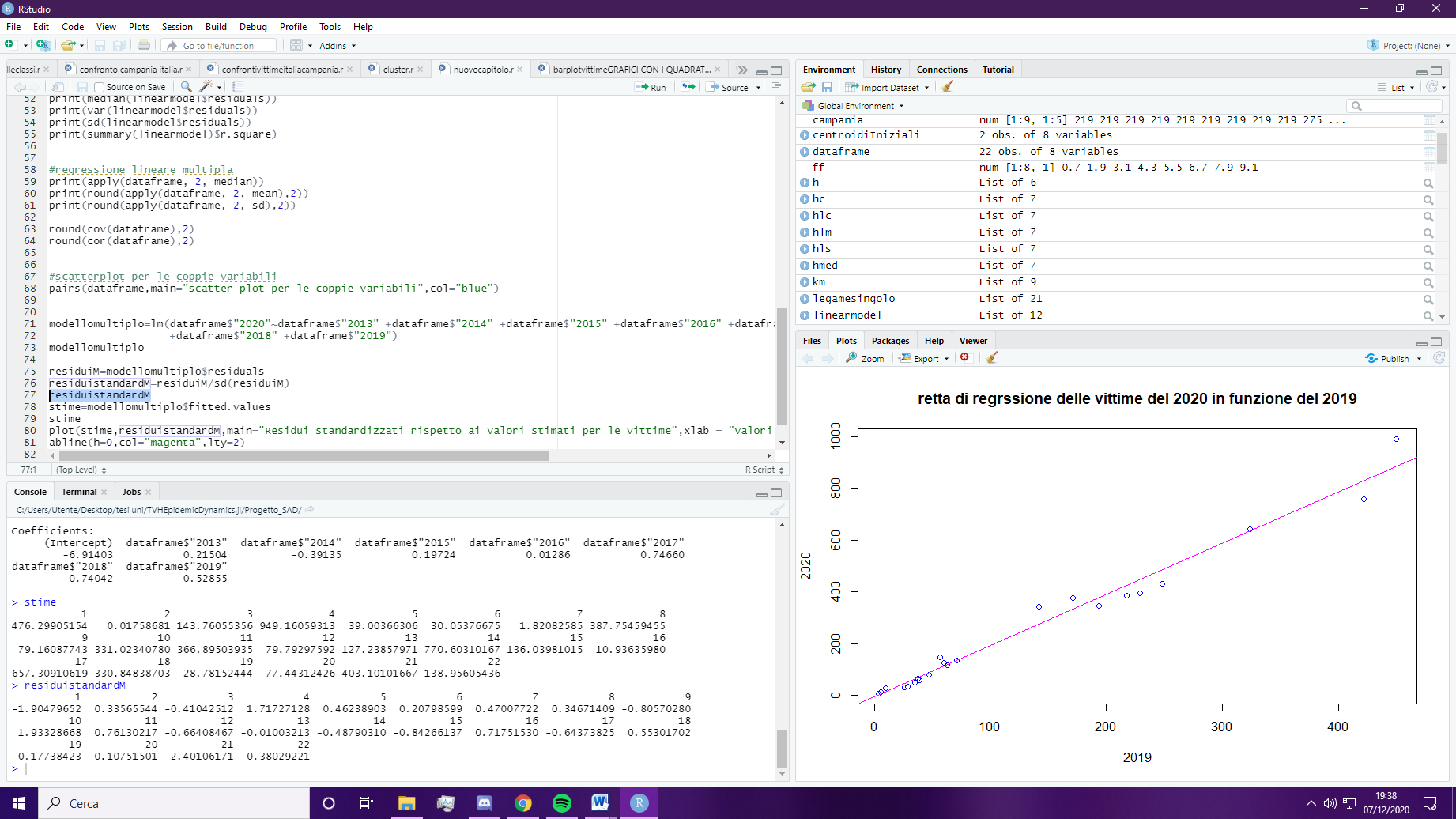
I segni dei regressori b1, b3, b4, b5, b6, b7, b8 sono positivi: questo indica che all’aumentare del numero di utenti nel 2013, 2015, 2016, 2017,2018,2019 aumenta il numero di utenti nel 2020. Mentre il regressore b2,è negativo quindi all’aumentare del numero di utenti nel 2014 diminuisce il numero di utenti nel 2020.

Il regressore del 2016 è prossimo allo zero, ciò vuol dire che l’aumento di vittime nel 2016 non infice molto sulle vittime del 2020.

Valori stimati rispetto al modello di regressione multipla:



Valori dei residui standardizzati rispetto alla deviazione standard.



Residui dei valori osservati rispetto ai valori stimati.

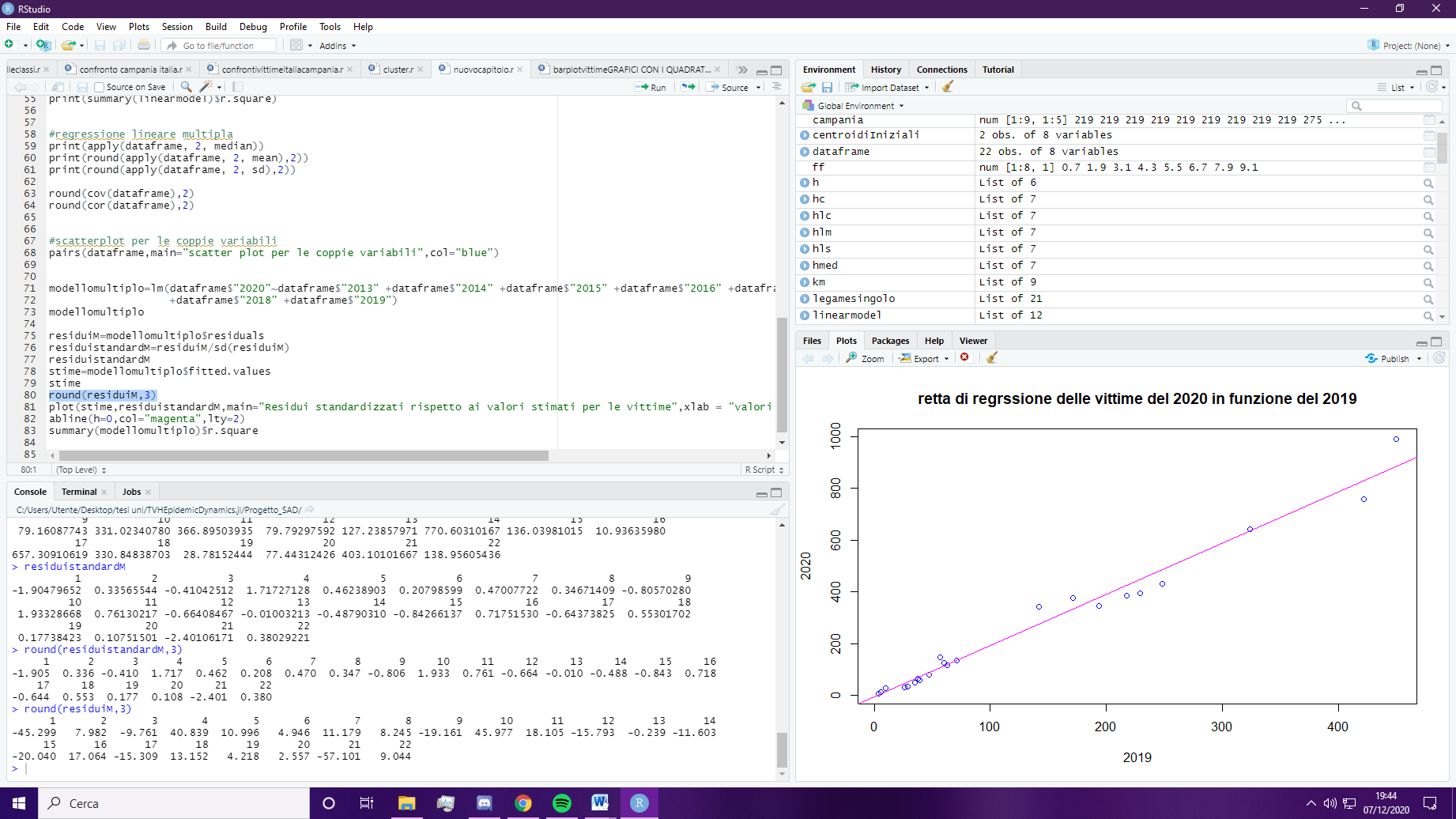
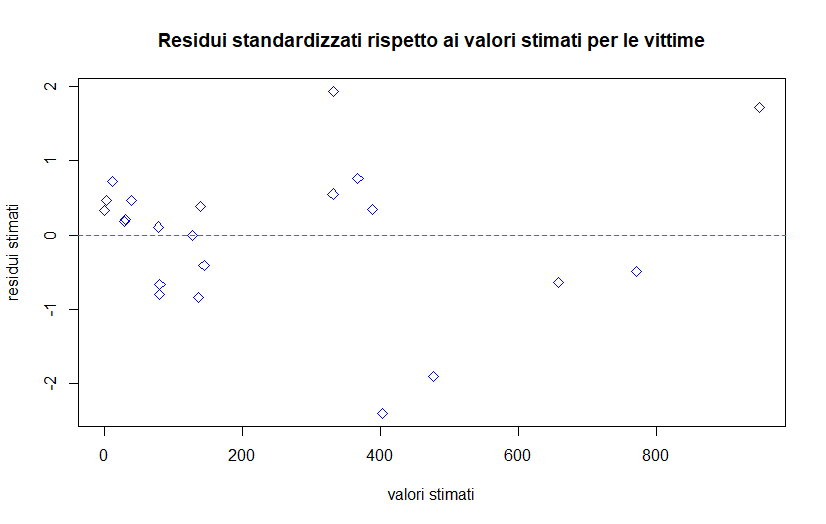
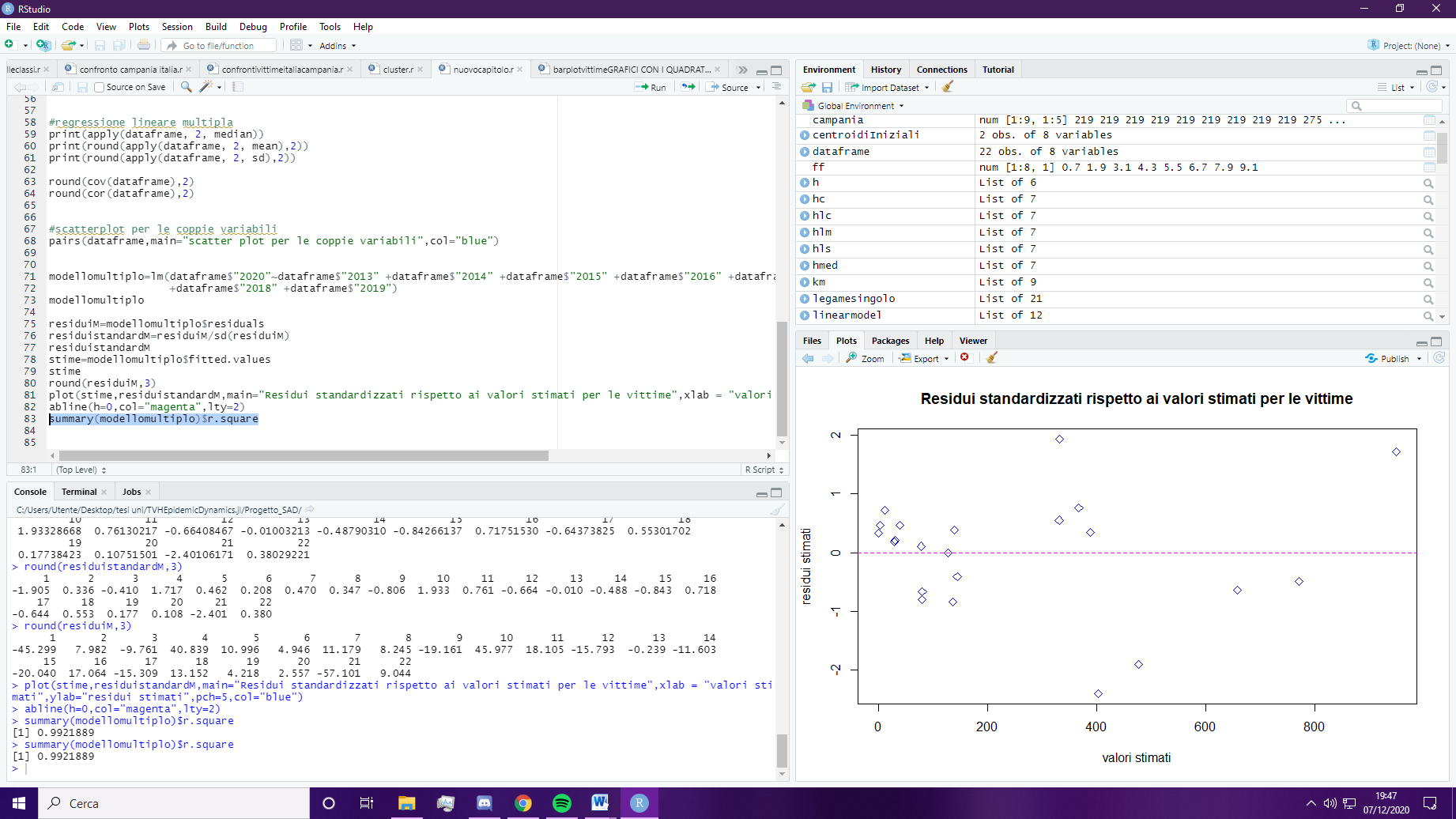


Grafico che mostra i residui standardizzati in funzione dei valori stimati.



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 0.9921889. Il modello di regressione lineare multipla descrive bene i dati considerati



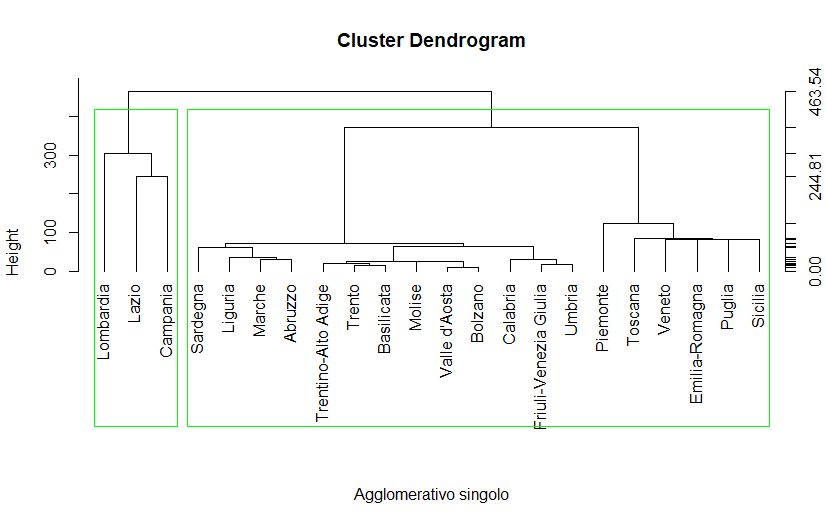
Analisi dei cluster

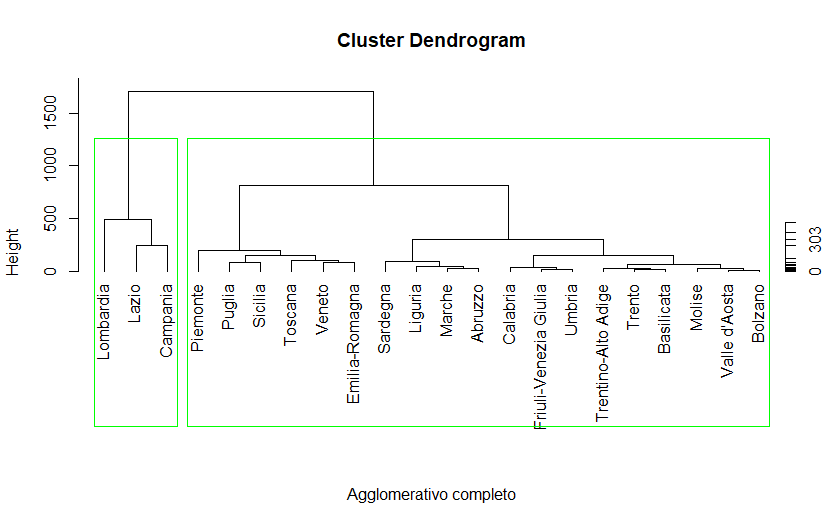
Tutti i metodi gerarchici: legame singolo, legame medio, legame completo, metodo del centroide e metodo della mediana hanno fornito il seguente partizionamento in due cluster.

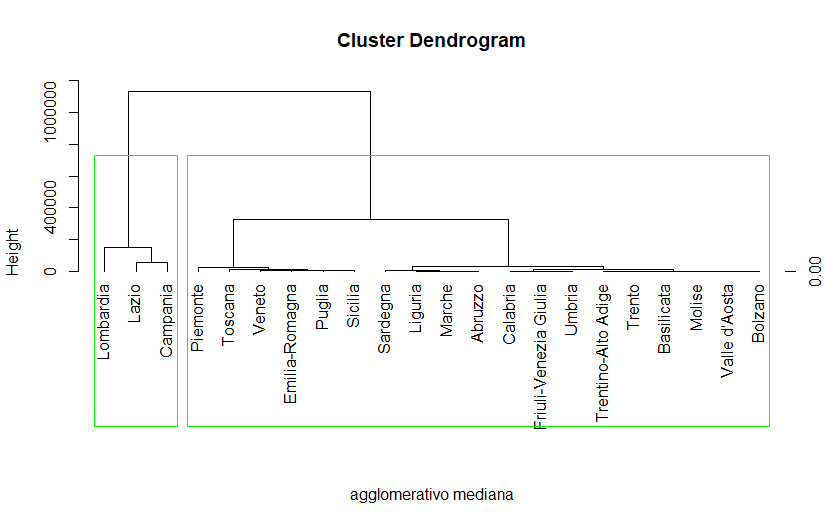
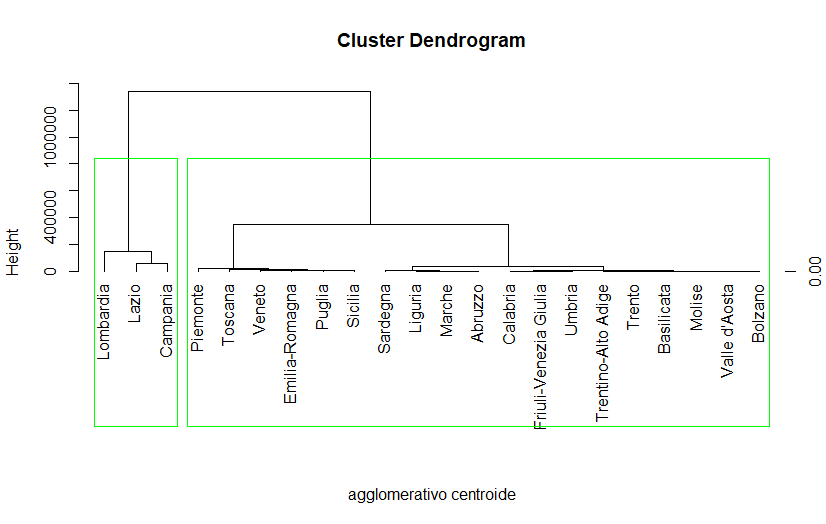
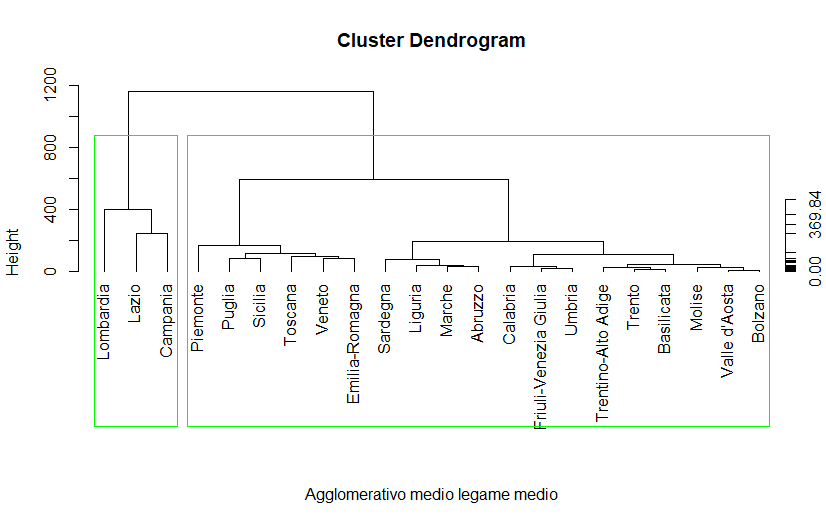
Primo cluster: 19 individui

Secondo cluster: 3 individui

|  |  |
| --- | --- |
| Cluster 2 | 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 1 | Lombardia, Lazio, Campania |







La misura di non omogeneità totale trT nel data frame considerato risulta essere uguale a: 5209482

La misura di non omogeneità all’interno del primo cluster trH1 risulta essere uguale a 1611162.

La misura di non omogeneità all’interno del secondo cluster trH2 risulta essere uguale a 131426.

Pertanto, la misura di non omogeneità tra i cluster risulta essere trB=trT-trH1-trH2= 3466893.

Il rapporto =  **0.6654968**.

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

.

Primo cluster: 13 individui

Secondo cluster: 9 individui

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

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 131830.2.

La misura di non omogeneità all’interno del secondo cluster trH2 risulta essere uguale a 1310843.8.

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

Il rapporto =  **0.7230677**.

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