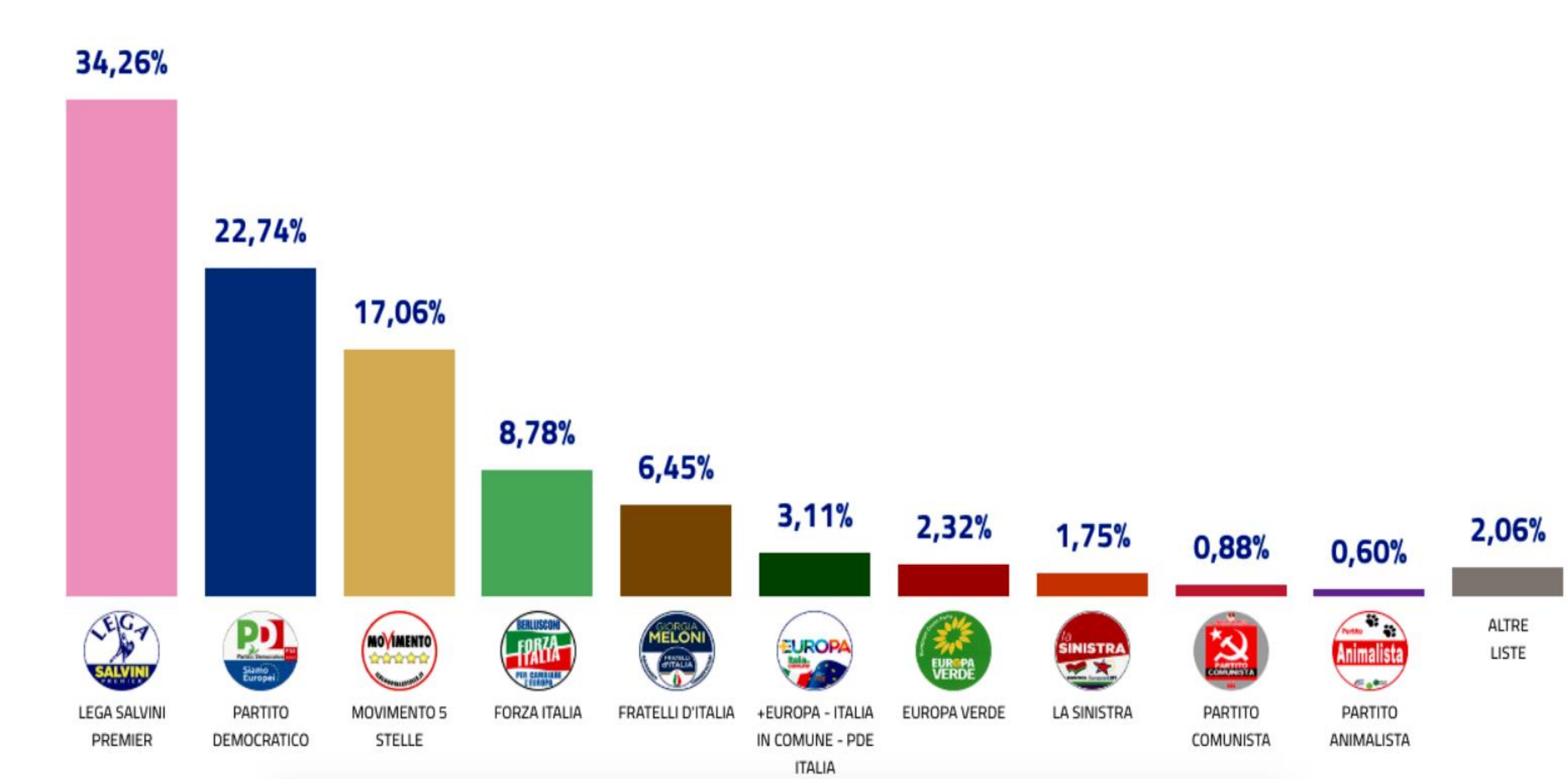
Elezioni Europee 2019



Struttura

- Raccolta e organizzazione dati
- Regressione multipla
- Analisi di Clustering
- Anova Two-Ways

Raccolta e organizzazione dati

Fonti

- ISTAT: Fattori socio-economici e demografici
- Ministero dell'Interno: Risultati elettorali dei comuni
- ACI: Infrastrutture stradali
- Il Sole 24 Ore: Indagini criminalità

Raccolta e organizzazione dati

Fonti

Estrapolazione dati diretta dalla sorgente HTML della pagina web con Python.

estrai_dati_criminalità.py - C:\Users\carlo\Desktop\prove_python\estrai_dati_criminalità.py (3.8.4rc1)

File Edit Format Run Options Window Help

```
from bs4 import BeautifulSoup
import requests
import pandas as pd

main_url = "https://lab24.ilsole24ore.com/indice-della-criminalita/indexT.php"
req = requests.get(main_url)
soup = BeautifulSoup(req.text, "lxml")

table = soup.find("table", id = "123")      #importo la tabella
rows = table.findAll('tr')
pos = []
citta = []
val1 = []
val2 = []
righe = table.findAll('tr')

for row in righe:                           #salvo i dati della tabella nelle liste
    cells = row.findAll('td')
    if cells != []:
        pos.append(cells[0].text)
        citta.append(cells[1].text)
        val1.append(cells[2].text)
        val2.append(cells[3].text)

d = {'citta':citta, 'n_denunce':val1,'denunce_su_abitante':val2}    #creo il dataframe
frame = pd.DataFrame(d)
```

Raccolta e organizzazione dati

Come organizzare i dati?

Raggruppamento per :

- Comuni: Eccessiva variabilità
- Regioni: Bassa numerosità ed eccessiva uniformazione
- Province: Numerosità e specificità adeguate

Regressione Multipla
Scelta delle covariate

- Popolazione
- Persone laureate
- Astenuti al voto
- Stranieri
- Persone nelle fasce di reddito più basse
- Tasso di disoccupazione
- Posti letto negli ospedali
- Fascia di età 1 (0-18)
- Fascia di età 2 (18-35)
- Fascia di età 3 (36-50)
- Fascia di età 4 (51-70)
- Km strade/popolazione
- Percentuale autostrade
- Km strade/estensione provincia

Risposta: percentuale dei voti del partito all'interno di ogni provincia.

E' stata eseguita una normalizzazione dei dati per avere una migliore leggibilità.

Regressione Multipla
Modello completo

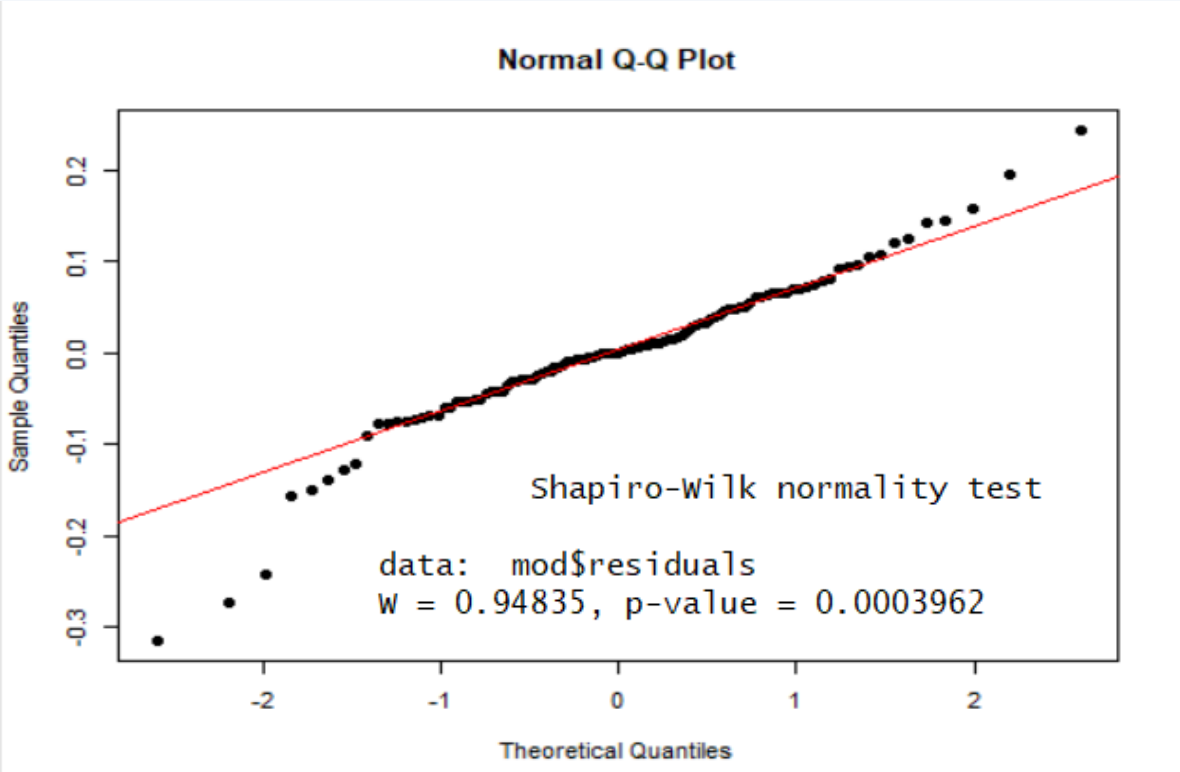
```
Call:
lm(formula = partito ~ ., data = tab[, 10:24])

Residuals:
     Min       1Q   Median       3Q      Max
-0.31428 -0.04194 -0.00026  0.04887  0.24281

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.4983026  0.1791962   2.781  0.00659 **
popol           0.0587000  0.0874410   0.671  0.50372
uni             0.1872115  0.0929317   2.015  0.04691 *
aste            0.1965900  0.0790365   2.487  0.01469 *
stranieri      -0.3566078  0.0868067  -4.108 8.71e-05 ***
redditobasso    0.2716736  0.0933096   2.912  0.00452 **
disoccup        0.2016118  0.0890441   2.264  0.02594 *
FasciaEta1     -0.3293698  0.1485569  -2.217  0.02911 *
FasciaEta2     -0.0600832  0.1686375  -0.356  0.72245
FasciaEta3      0.0447135  0.0632426   0.707  0.48136
FasciaEta4     -0.4367421  0.1433308  -3.047  0.00302 **
perc_posti_osp  0.0007575  0.0479466   0.016  0.98743
km_popol        0.0207875  0.0756278   0.275  0.78404
km_sup          0.0242227  0.0550147   0.440  0.66077
perc_autos      0.0617689  0.0955304   0.647  0.51953
denu_su_ab      0.1143245  0.0748053   1.528  0.12991
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09025 on 91 degrees of freedom
Multiple R-squared:  0.8845,    Adjusted R-squared:  0.8654
F-statistic: 46.44 on 15 and 91 DF,  p-value: < 2.2e-16
```

Primo modello: percentuale dei voti del M5S con tutti i predittori a disposizione.

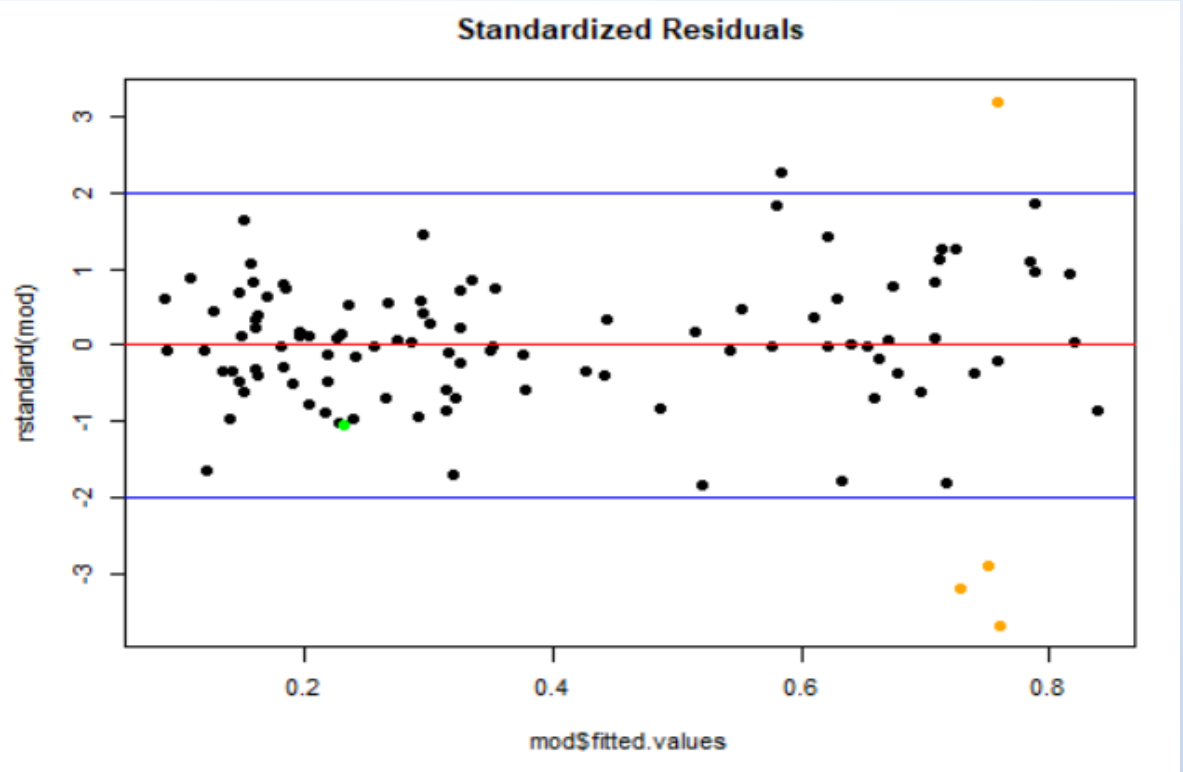
Il coefficiente di determinazione risulta elevato, però il numero di covariate utilizzate è ancora troppo alto.

Regressione Multipla Diagnostica

Ipotesi di normalità dei residui

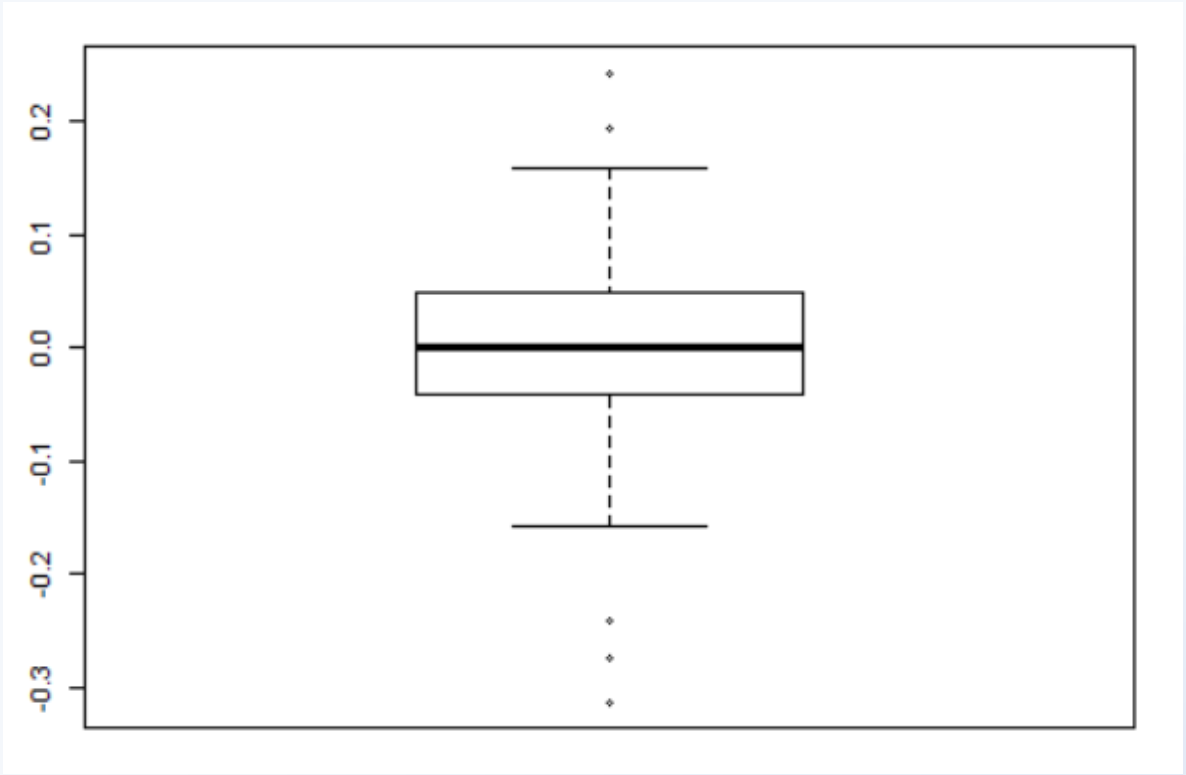


Ipotesi di omoschedasticità dei residui

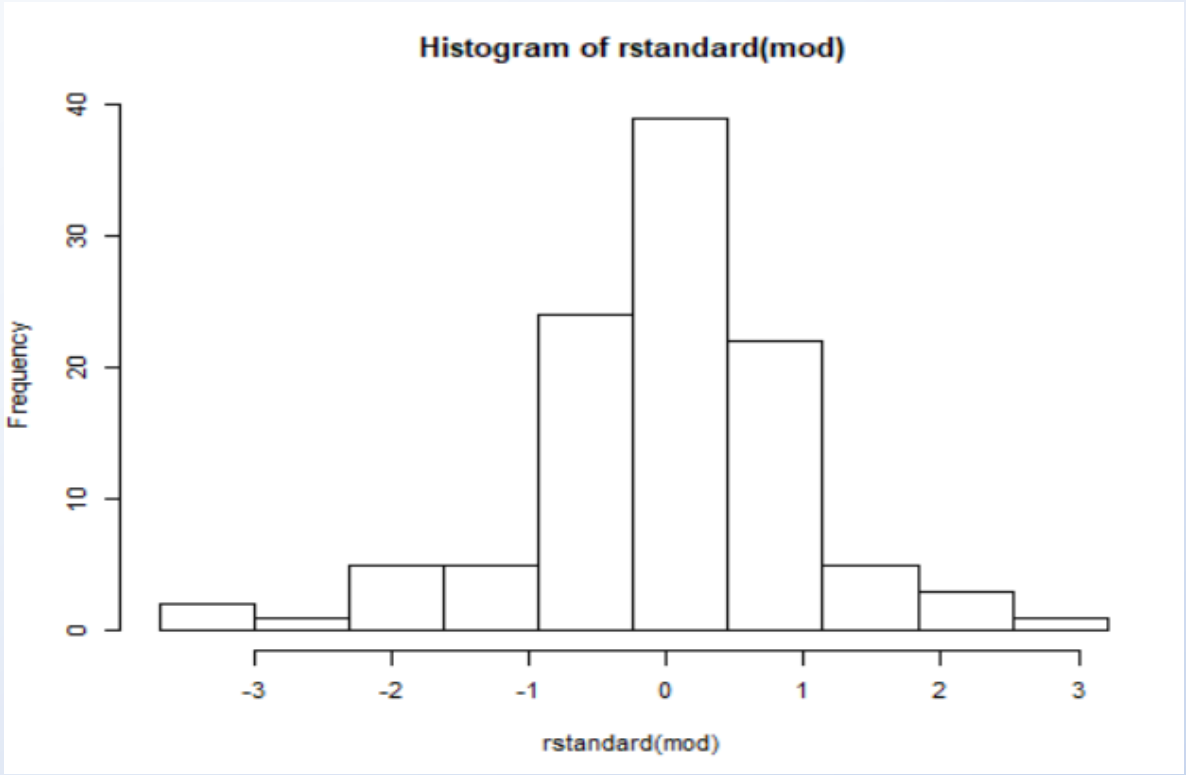


Regressione Multipla Diagnostica

Boxplot dei residui

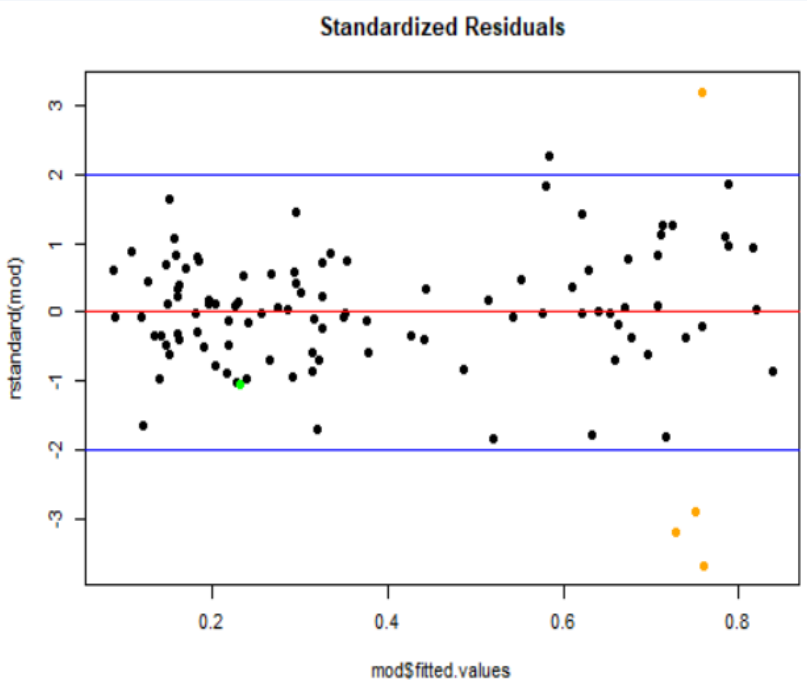


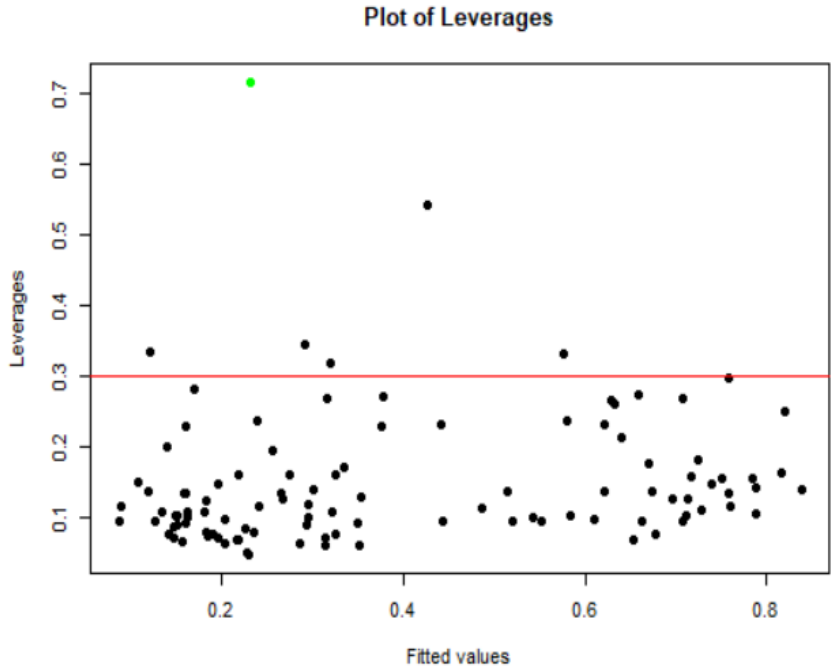
Istogramma dei residui

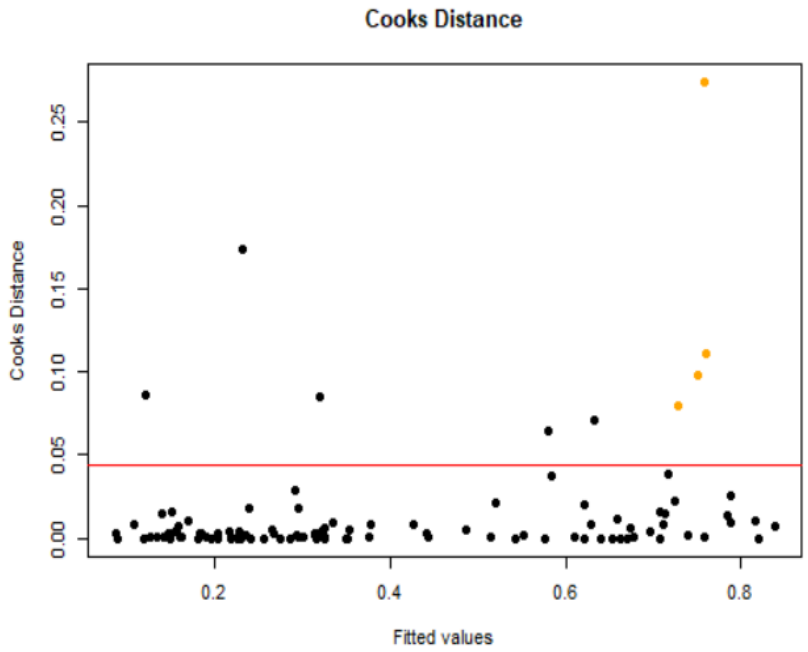


Regressione Multipla

Diagnostica - Outliers, Levareges, Punti influenti





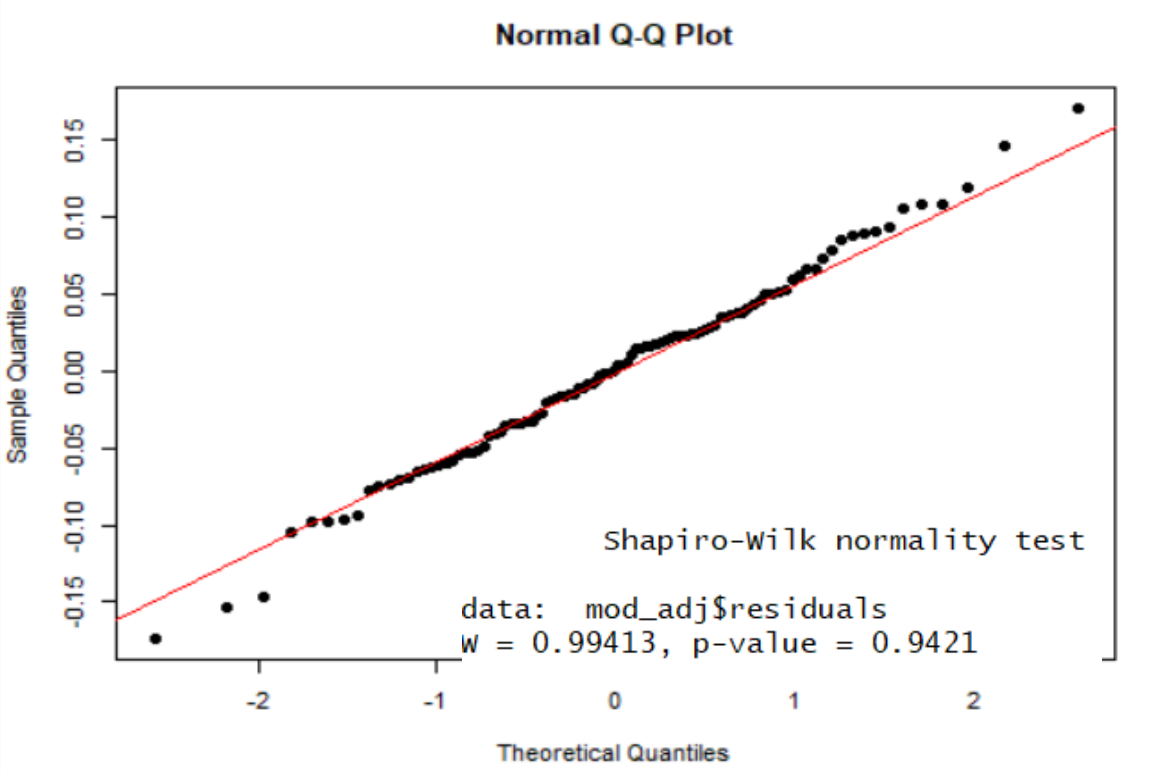


Si eliminano tre categorie di punti:

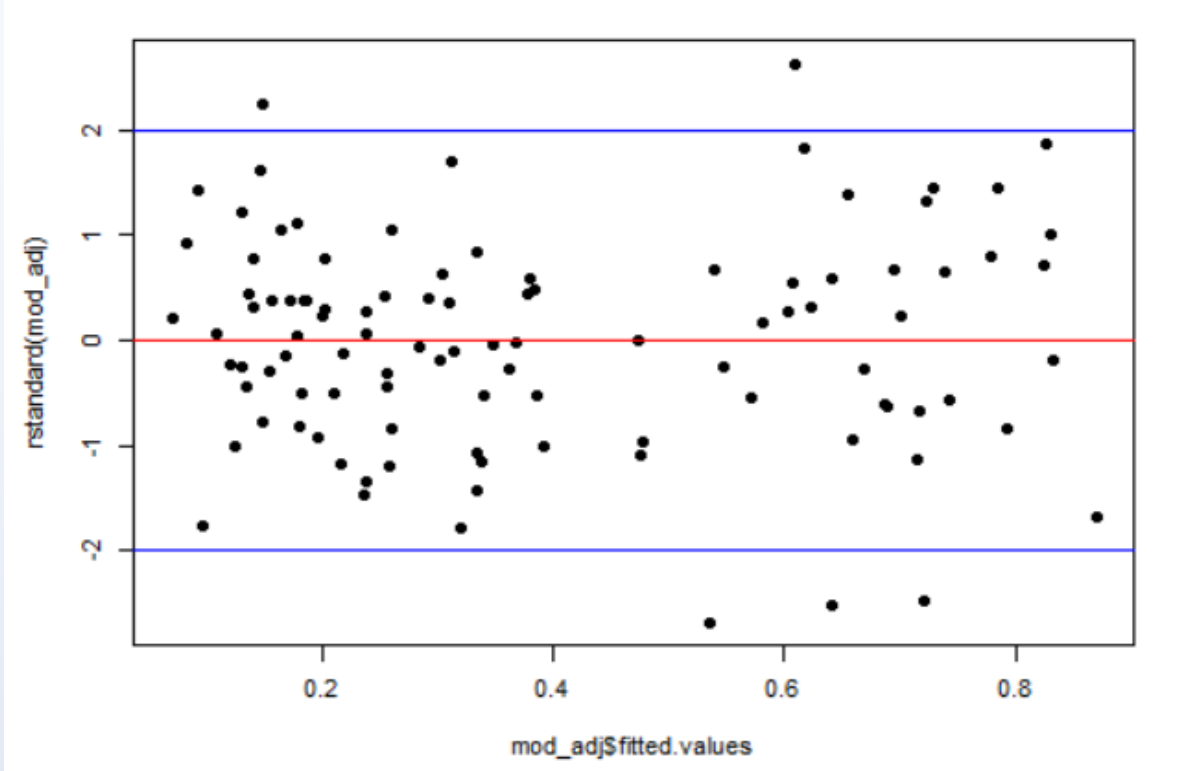
- Contemporaneamente outliers e leverages
- Contemporaneamente outlier e punti influenti
- Leverage massimo

Regressione Multipla
Diagnostica – Modello modificato

Ipotesi di normalità

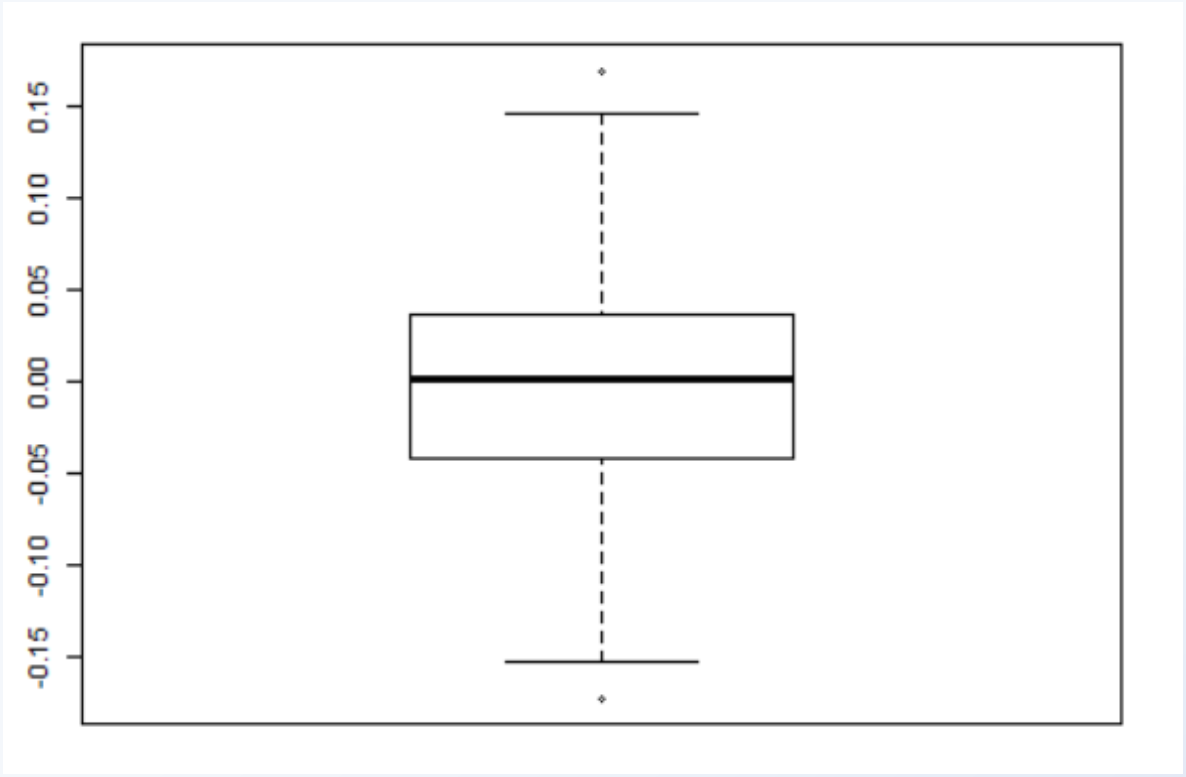


Ipotesi di omoschedasticità

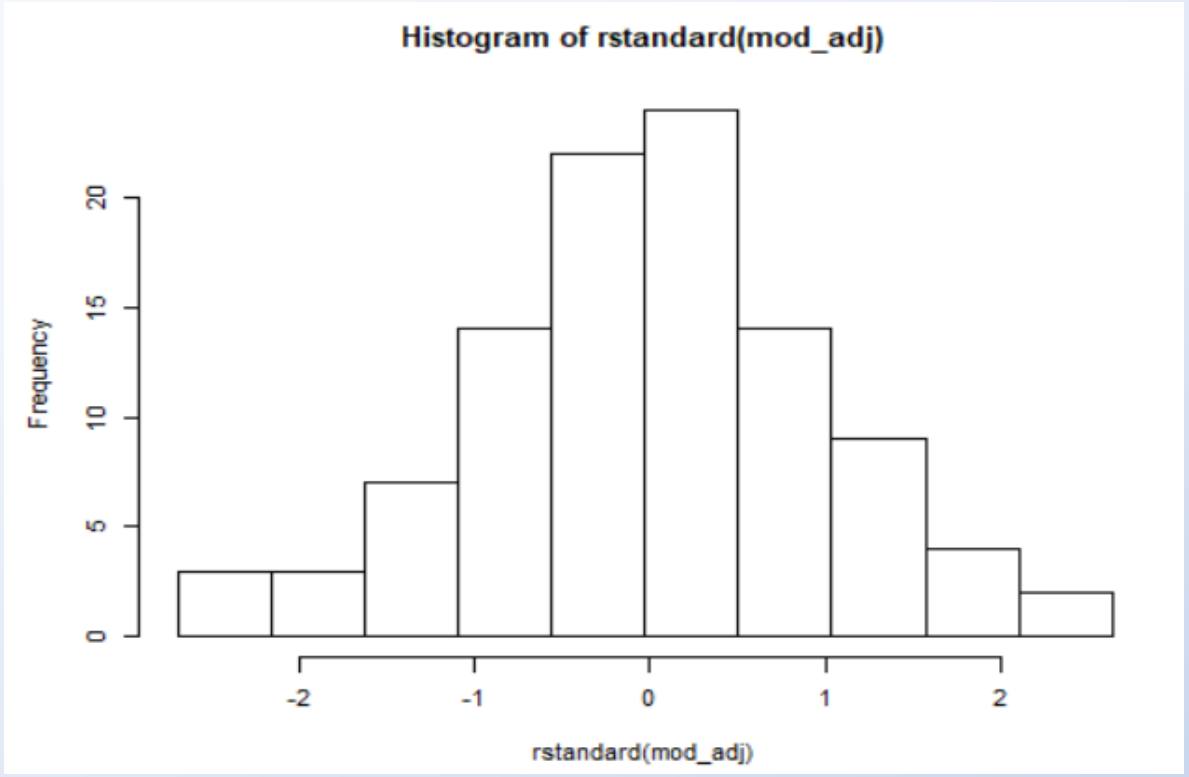


Regressione Multipla
Diagnostica – Modello modificato

Boxplot dei residui



Istogramma dei residui



Regressione Multipla
Modello modificato

```
Call:
lm(formula = partito_adj ~ ., data = tab[, 10:24])

Residuals:
      Min        1Q    Median        3Q       Max
-0.173340 -0.040782  0.001771  0.036552  0.170143

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.54314    0.13864   3.918 0.000179 ***
popol         -0.01915    0.07071  -0.271 0.787128
uni            0.26934    0.07318   3.681 0.000405 ***
aste           0.23819    0.06187   3.850 0.000227 ***
stranieri     -0.32983    0.06920  -4.766 7.56e-06 ***
redditobasso   0.36559    0.07286   5.018 2.79e-06 ***
disoccup       0.13483    0.06902   1.954 0.053997 .
FasciaEta1    -0.31936    0.11381  -2.806 0.006202 **
FasciaEta2    -0.20257    0.13074  -1.549 0.124969
FasciaEta3     0.02370    0.05048   0.469 0.639942
FasciaEta4    -0.51444    0.11167  -4.607 1.41e-05 ***
perc_posti_osp 0.01206    0.03698   0.326 0.745143
km_popol      -0.01953    0.05964  -0.327 0.744093
km_sup         0.02892    0.04321   0.669 0.505076
perc_autos     0.08080    0.12577   0.642 0.522278
denu_su_ab     0.07579    0.05851   1.295 0.198661
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06825 on 86 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.933,     Adjusted R-squared:  0.9214
F-statistic: 79.89 on 15 and 86 DF,  p-value: < 2.2e-16
```

Il coefficiente di determinazione è in aumento rispetto a prima ,tuttavia osserviamo che il modello presenta delle covariate non significative.

Nel modello c’è troppa confusione: bisogna ricercare le covariate che fittano al meglio il modello.

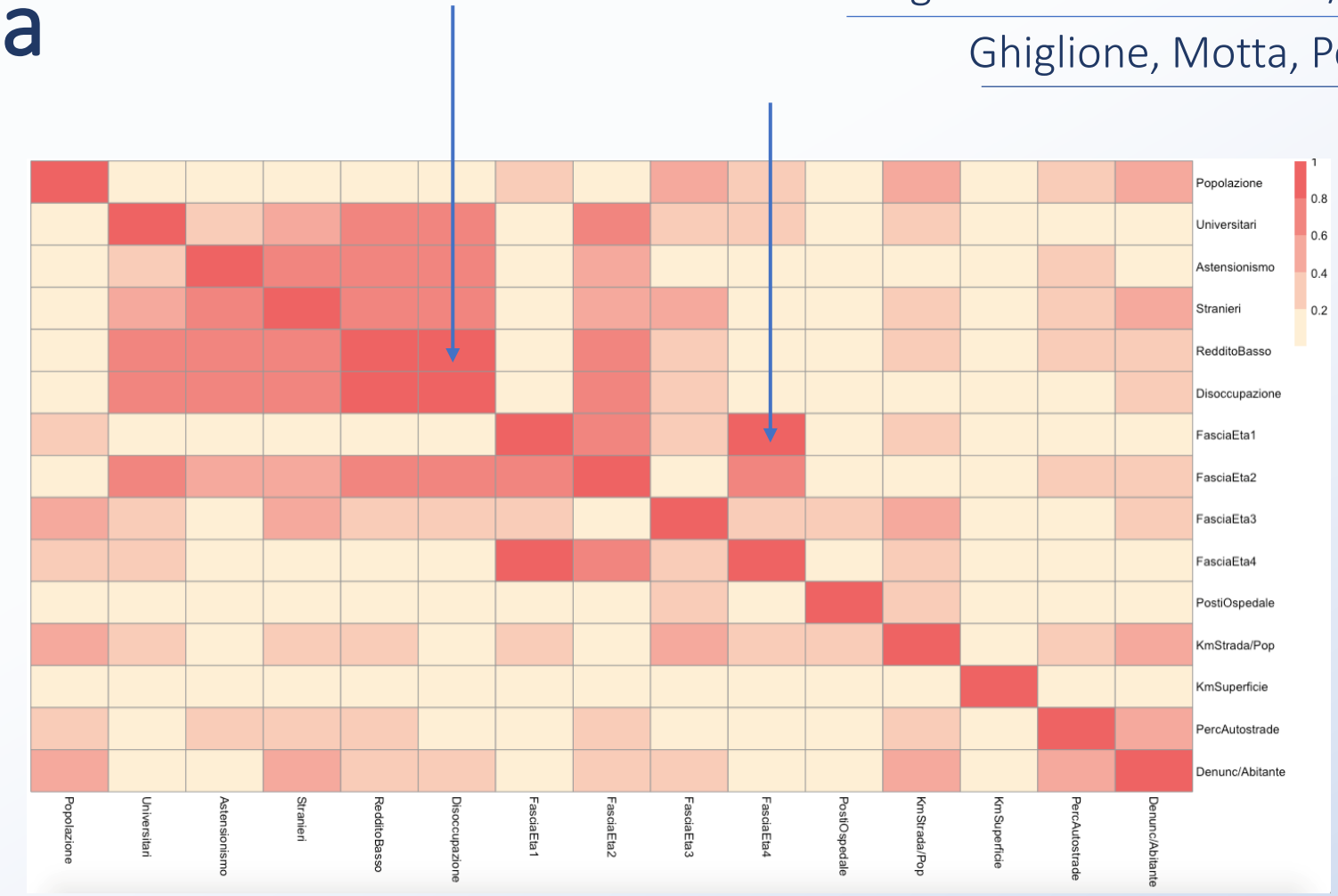
Si passa, quindi, all’analisi di collinearità e successivamente al metodo stepwise.

Regressione Multipla

Analisi di Collinearità

In questa fase verifichiamo che:

- Correlazione tra covariate, che non deve essere eccessivamente alta
- Indice VIF, che se superiore a 10 indica covariate molto legate



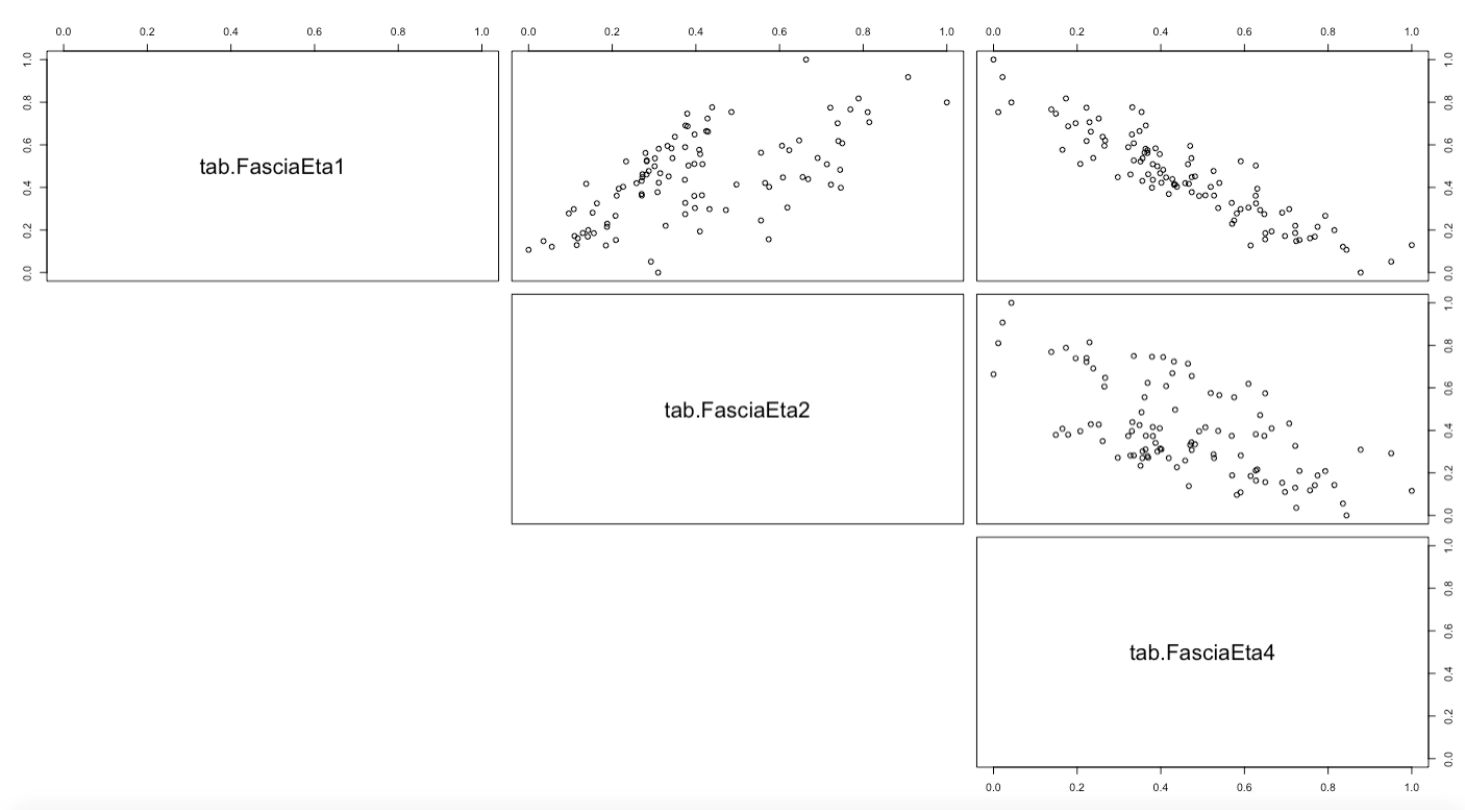

```
> print(vif(mod_adj))
      popol         uni        aste   stranieri redditobasso    disoccup  FasciaEta1
   1.978766    5.254748    4.188543    4.446439    9.273622    4.827492   11.793076
 FasciaEta2  FasciaEta3  FasciaEta4 perc_posti_osp   km_popol      km_sup  perc_autos
  17.383815    2.337879   12.031261    1.398016    2.378894    1.306102    1.874945
  denu_su_ab
    2.342644
```

Regressione Multipla

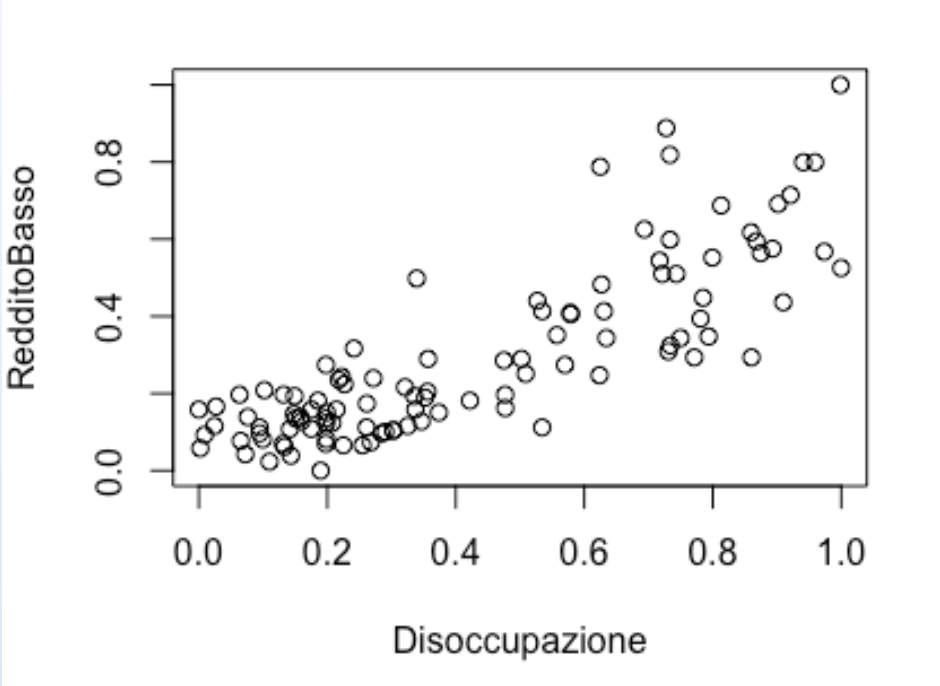
Ghiglione, Motta, Perelli

Analisi di Collinearità

- Si puo' notare come la correlazione sia evidente anche da uno scatterplot.



- Alla luce dell'analisi , rimuoviamo FasciaEta1, FasciaEta2 e RedditoBasso.



Regressione Multipla

Analisi di Collinearità

- Lo Shapiro Test mantiene p-value alto.

```
> print(shapiro.test(mod_new$residuals))

        Shapiro-Wilk normality test

data:  mod_new$residuals
W = 0.98879, p-value = 0.5538
```

- I coefficienti VIF sono ora soddisfacenti.

```
> print(vif(mod_new))
         popol            uni          aste     stranieri      disoccup    FasciaEta3    FasciaEta4
      1.724376       2.849888       3.213687       3.898973       3.857466       2.177688       1.829474
perc_posti_osp      km_popol        km_sup    perc_autos    denu_su_ab
      1.201064       2.262229       1.204803       1.743673       1.865522
```

Regressione Multipla

Metodo stepwise

Il modello ha ancora troppe covariate, molte delle quali poco significative. Applichiamo una procedura stepwise backward per selezionare le più significative.

```
14  mod_new <- step(mod_no_col, direction = "backward", k=qchisq(0.95,1))
```

- La differenza di R^2 adjusted tra inizio e fine degli step é pressoché nulla.

R2_adj_end	0.888428914488796
R2_adj_start	0.886913579751366

- VIF rimangono bassi

```
> print(vif(mod_new))
    popol       uni      aste  stranieri   disoccup FasciaEta4 denu_su_ab
 1.438662  2.273267  2.955705   3.229161   3.540417   1.430959   1.661019
```

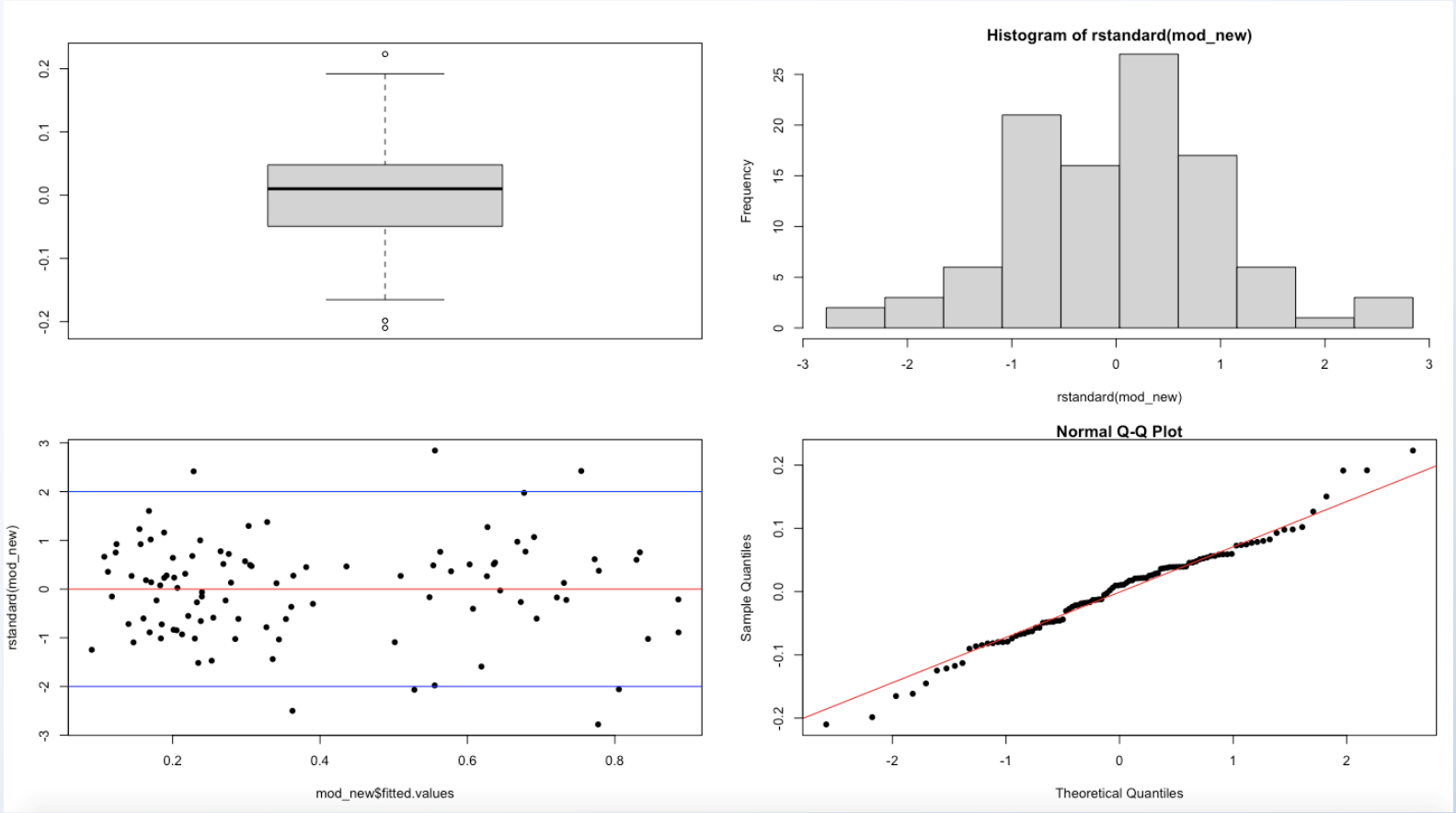
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.10814    0.06658   1.624 0.107676
popol       -0.20994    0.07182  -2.923 0.004340 **
uni          0.46269    0.05733   8.070 2.27e-12 ***
aste         0.37264    0.06191   6.019 3.36e-08 ***
stranieri   -0.28617    0.07024  -4.074 9.65e-05 ***
disoccup     0.25024    0.07040   3.554 0.000595 ***
FasciaEta4  -0.11967    0.04587  -2.609 0.010574 *
denu_su_ab   0.21160    0.05868   3.606 0.000500 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08129 on 94 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.8962,    Adjusted R-squared:  0.8884
F-statistic: 115.9 on 7 and 94 DF,  p-value: < 2.2e-16
```

Regressione Multipla

Diagnostica

- Shapiro test valido
- Residui non presentano pattern preoccupanti
- QQplot e istogramma soddisfacenti
- Boxplot e istogramma simmetrici

```
        Shapiro-Wilk normality test

data:  mod_new$residuals
W = 0.98355, p-value = 0.2367
```



Regressione Multipla

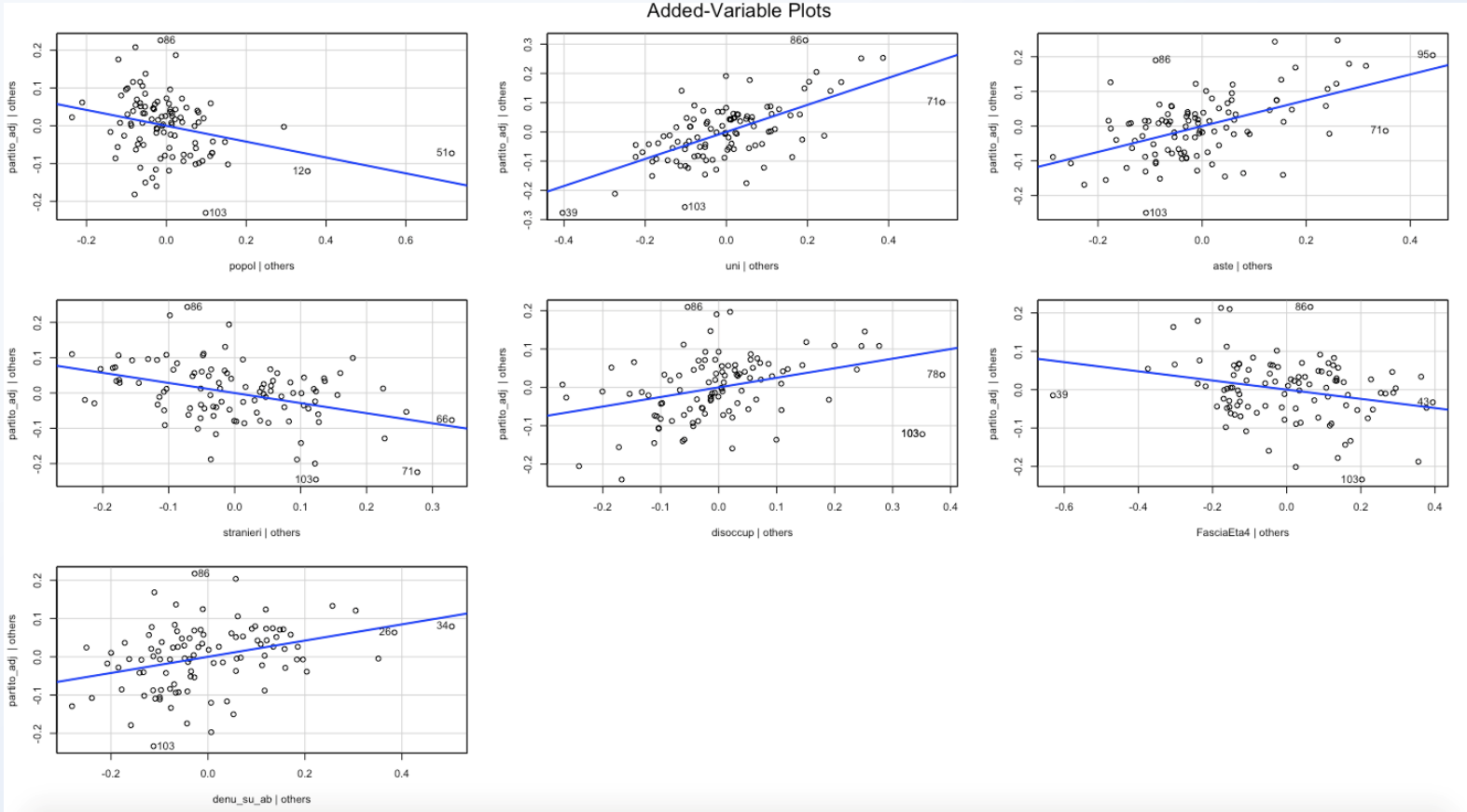
Interpretazione

L'astensionismo ha davvero danneggiato il Movimento5S?

Regressione Multipla

Interpretazione

I grafici di regressione parziale, permettono di valutare l'impatto netto di una singola covariata sulla risposta, mantenendo costanti le altre covariate.



Proprietà:

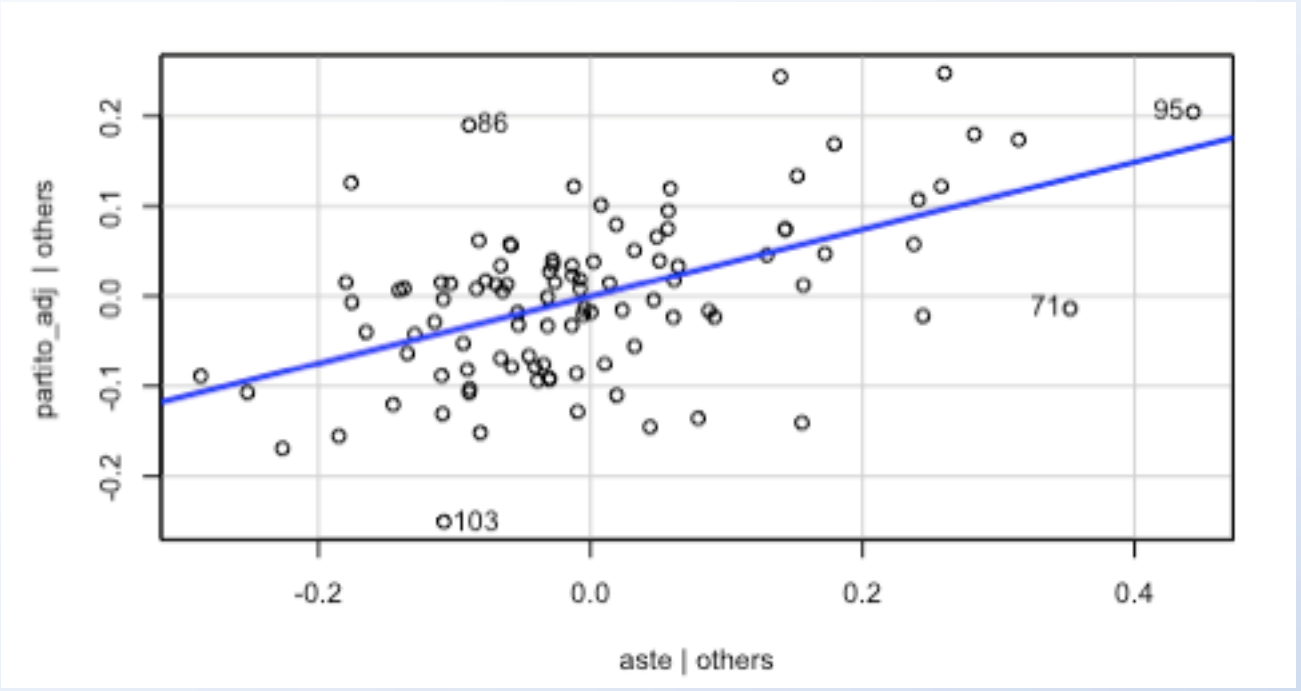
1. The least squares linear fit to this plot has the slope β_i and intercept zero.

Regressione Multipla

Interpretazione

	Estimate	Std. Error	t value	Pr(>|t|)	
aste	0.37264	0.06191	6.019	3.36e-08	***

La percentuale di voti ottenuti dal Movimento5S in una provincia aumenta all'aumentare dell'astensionismo.



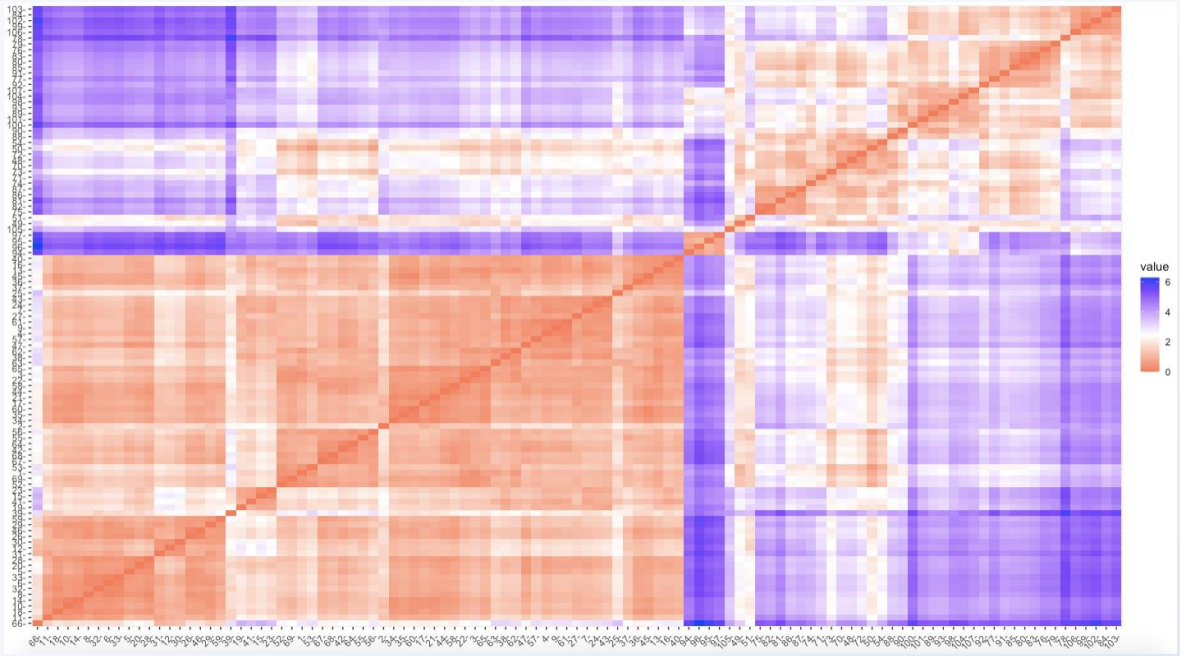
Se nelle province con alta astensione ci fosse stata piu' partecipazione alle urne, effettivamente il M5S avrebbe beneficiato di un numero maggiore di voti in termini assoluti.

Clustering

Metodo

Il clustering permette di aggregare il dataset in gruppi di dati ”simili”.

- Ogni provincia e’ rappresentabile come un punto in uno spazio N dimensionale.
- Calcoliamo la distanza euclidea tra ogni coppia di punti, raggruppiamo i punti vicini e rappresentiamo la distanza per ogni coppia di punti.



Coppie di punti lontani tra di loro finiranno in cluster distinti, punti vicini, cioè simili, finiranno nello stesso cluster.

In rosso coppie di punti vicine,
in blu coppie di punti lontane.

Clustering

Analisi fattori demografici

Il modello lineare mostra che é possibile spiegare i voti ottenuti dai partiti in base a fattori demografici specifici delle province.

Questi fattori si distribuiscono secondo un pattern ben definito?
Es: Grandi città vs piccoli paesi, zona costiera vs entroterra…

Sfruttando l'algoritmo *kmeans* di R, si ottiene una suddivisione ottimale in 2 cluster:

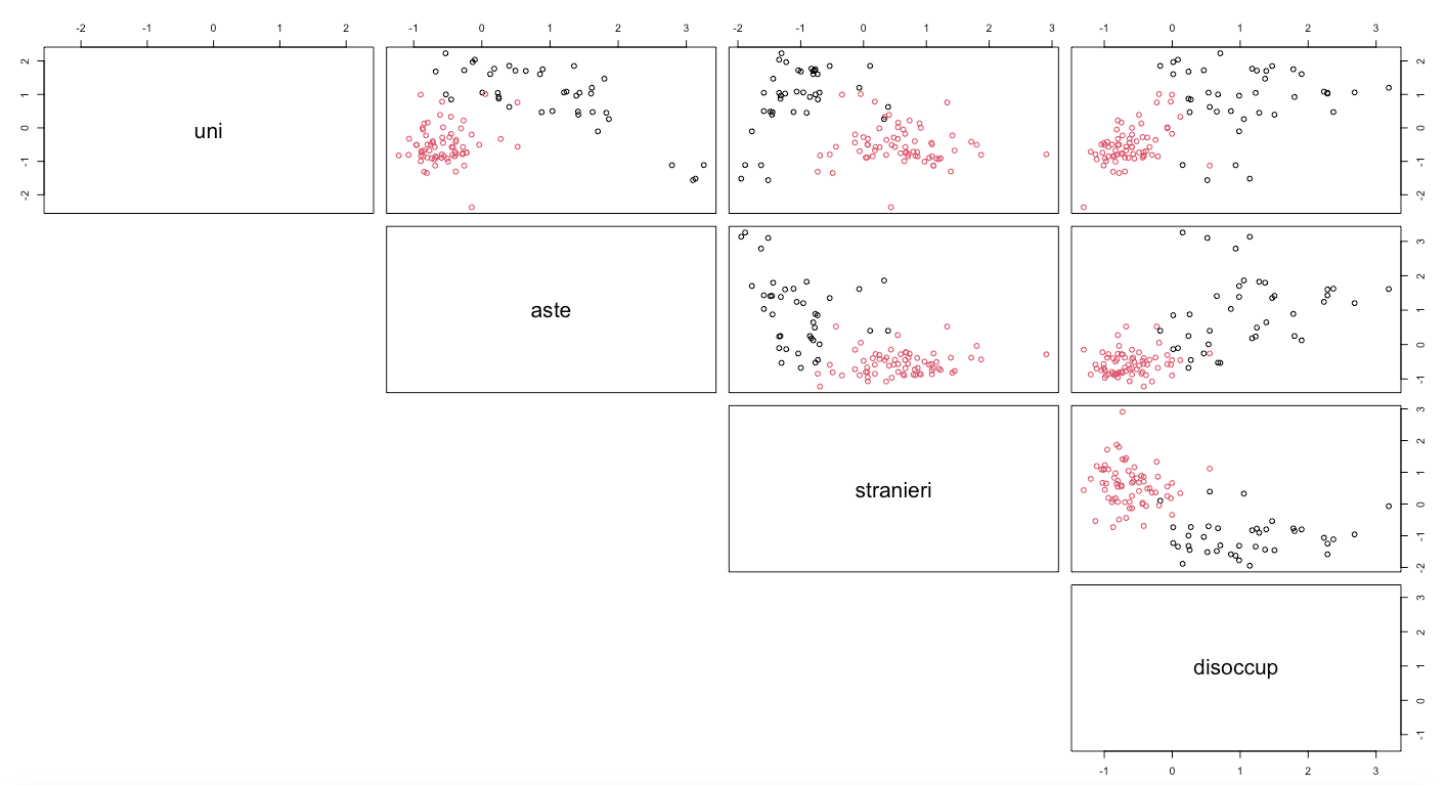
```
10  clust = kmeans(data,2);
```

Clustering

Analisi fattori demografici

Rispetto alle covariate significative del modello, il clustering sembra divider in maniera adeguata I dati in due gruppi ben distinti.

Ma si può visualizzare meglio…

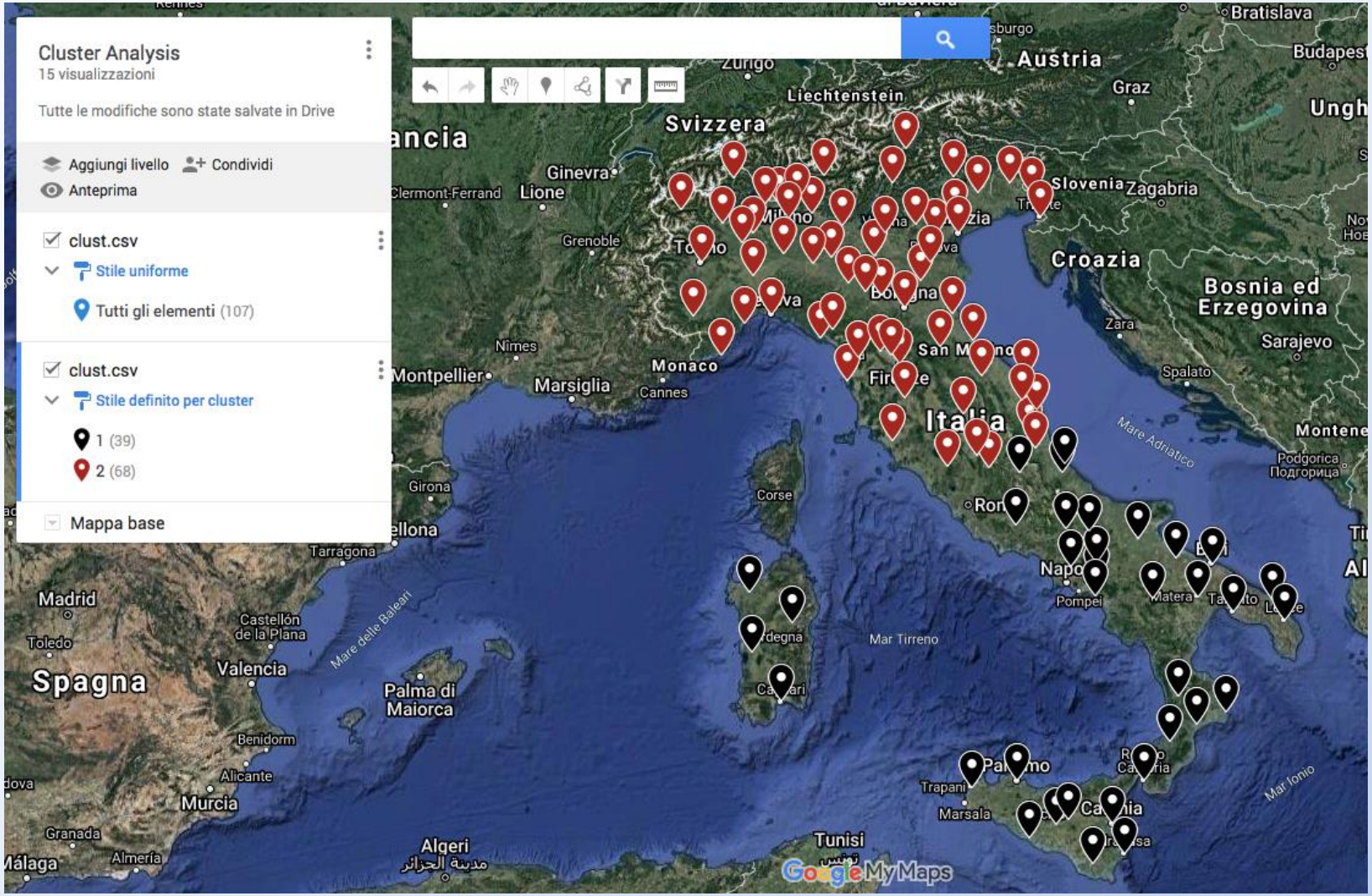


Clustering

Diagnostica

Combinando i risultati ottenuti con il servizio My Maps di Google, abbiamo ottenuto una rappresentazione geografica dei cluster.

L'algoritmo di clustering separa nettamente le province tra Nord e Sud, in base ai dati demografici ed economici considerati.



Two – ways ANOVA

Ghiglione, Motta, Perelli

Obbiettivo

Indagare se il risultato elettorale medio è stato influenzato dai due fattori categorici:

- Macro-regione geografica (Nord, Centro, Sud)
- Partito (Lega, PD, M5S)

e se tra i due fattori categorici c'è interazione.

Unità di analisi a livello provinciale.
Numerosità del campione e numerosità relative dei gruppi.

```
Numerosità macro-regioni
  N   C   S
141  66 114
--------------------
Numerosità partiti
 sx  dx m5s
107 107 107
```

```
                tab$macroreg
tab$partito   C  N  S
         dx  22 47 38
         m5s 22 47 38
         sx  22 47 38
```

	voti	partito	circoscr	macroreg
1	27.576993	sx	A	N
2	30.290627	dx	A	N
3	17.534723	m5s	A	N
4	17.367787	sx	A	N
5	40.063221	dx	A	N
6	14.794782	m5s	A	N
7	26.349995	sx	A	N
8	34.848514	dx	A	N
9	16.103412	m5s	A	N
10	21.876332	sx	A	N
11	38.786108	dx	A	N
12	15.085583	m5s	A	N
13	19.855094	sx	A	N
14	51.085497	dx	A	N
15	6.890215	m5s	A	N
16	20.601116	sx	A	N

Two – ways ANOVA

Analisi interazione

Per stabilire se si debba considerare l'interazione o no nel modello, si disegnano i due *interaction plot*:

- risposta vs fattore 1 facendo variare il fattore 2
- risposta vs fattore 2 facendo variare il fattore 1

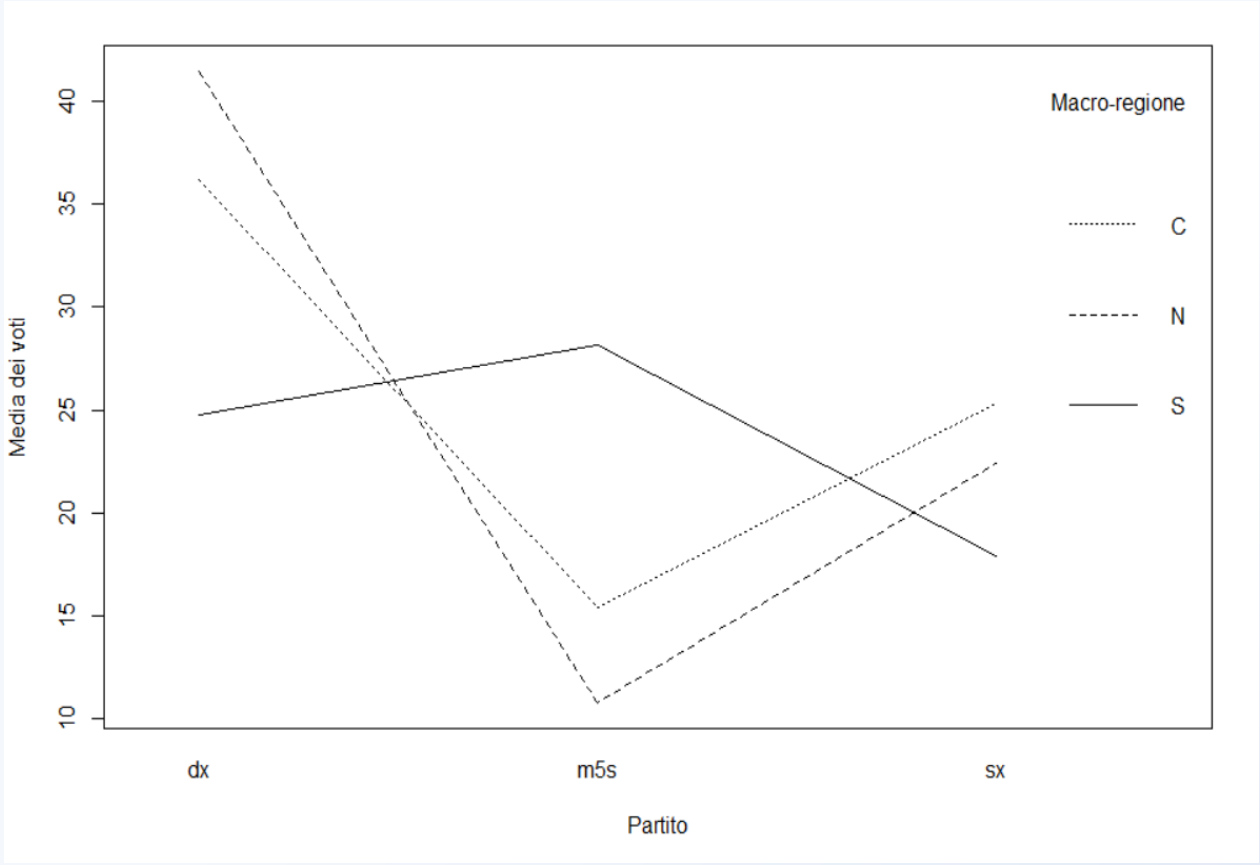
Se i grafici risultano paralleli, il modello corretto è senza interazione, altrimenti il modello corretto è con l'interazione.

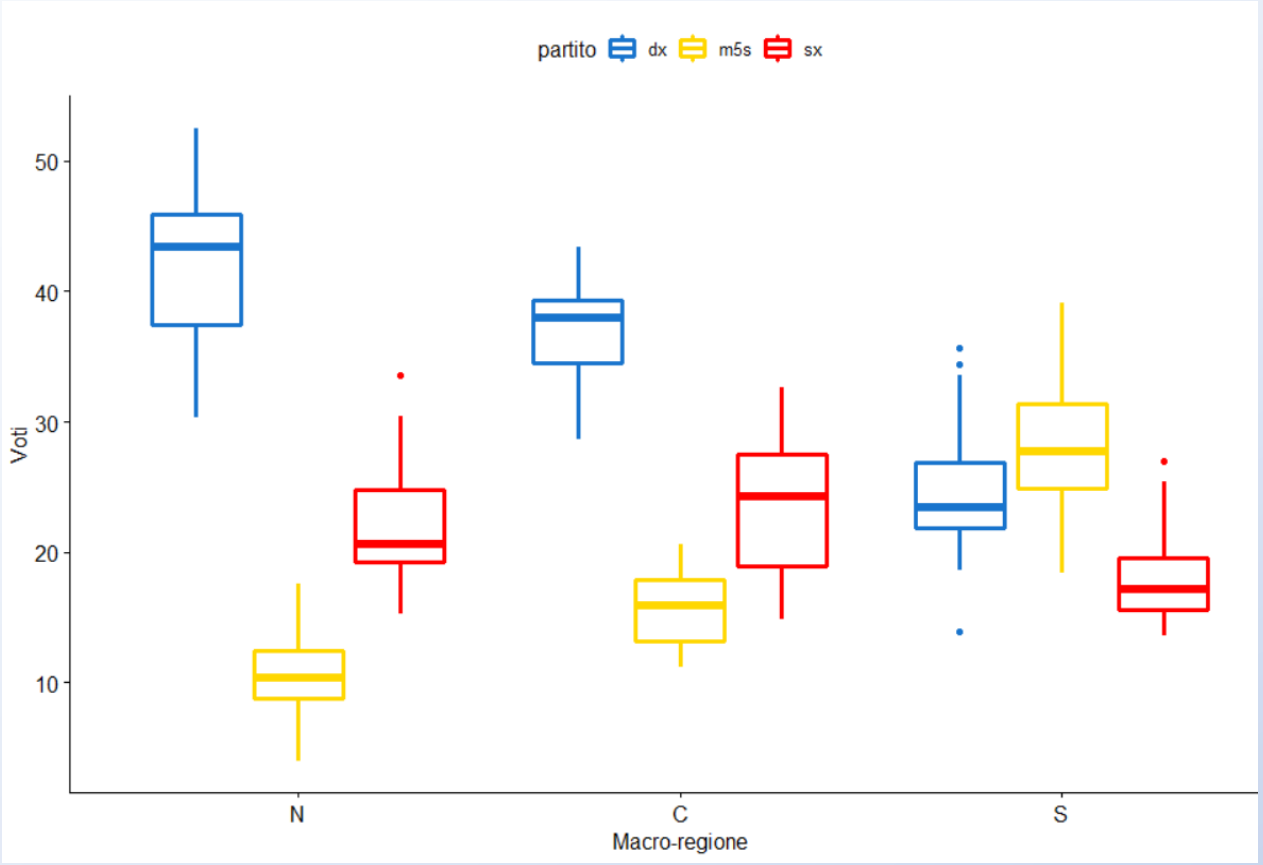
Two – ways ANOVA

Progetto MMIS A.A. 2019/2020

Ghiglione, Motta, Perelli

Analisi interazione, Voti vs Partito



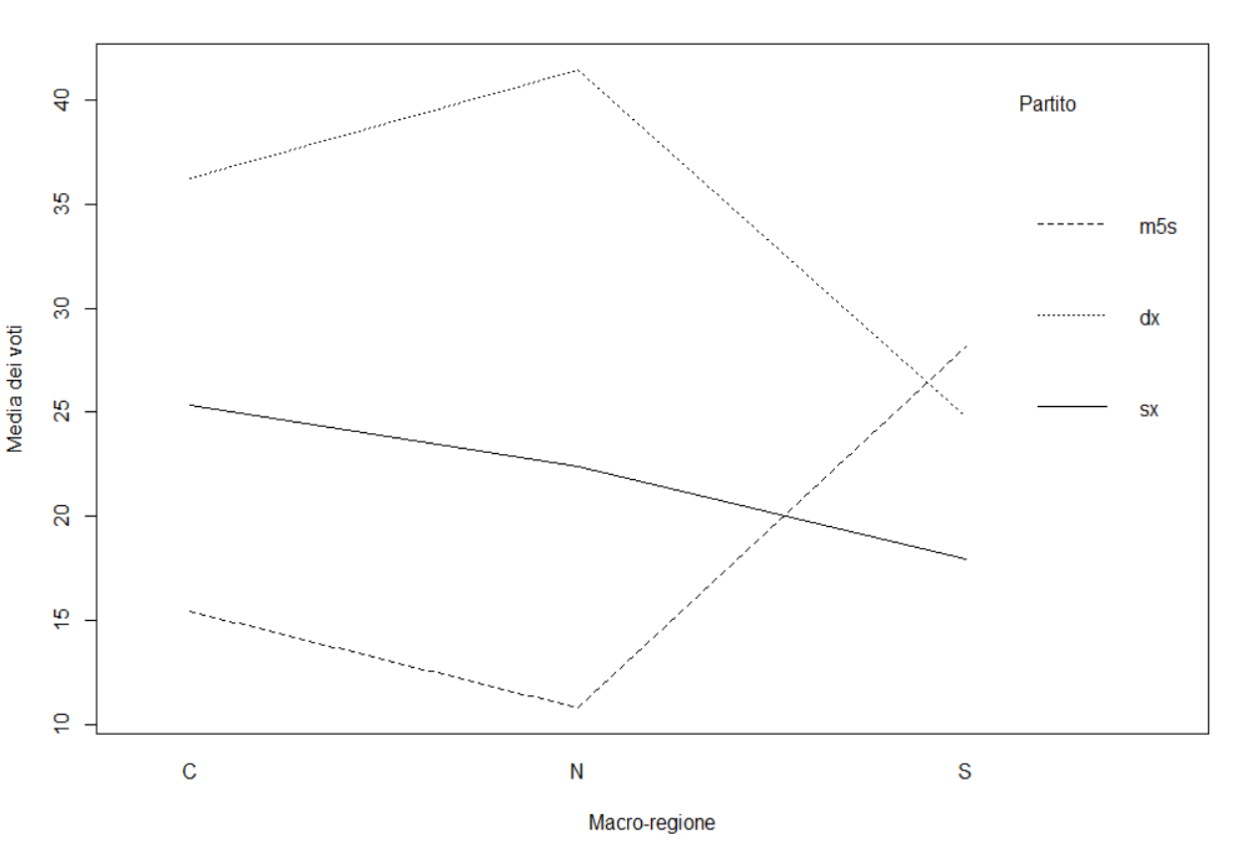


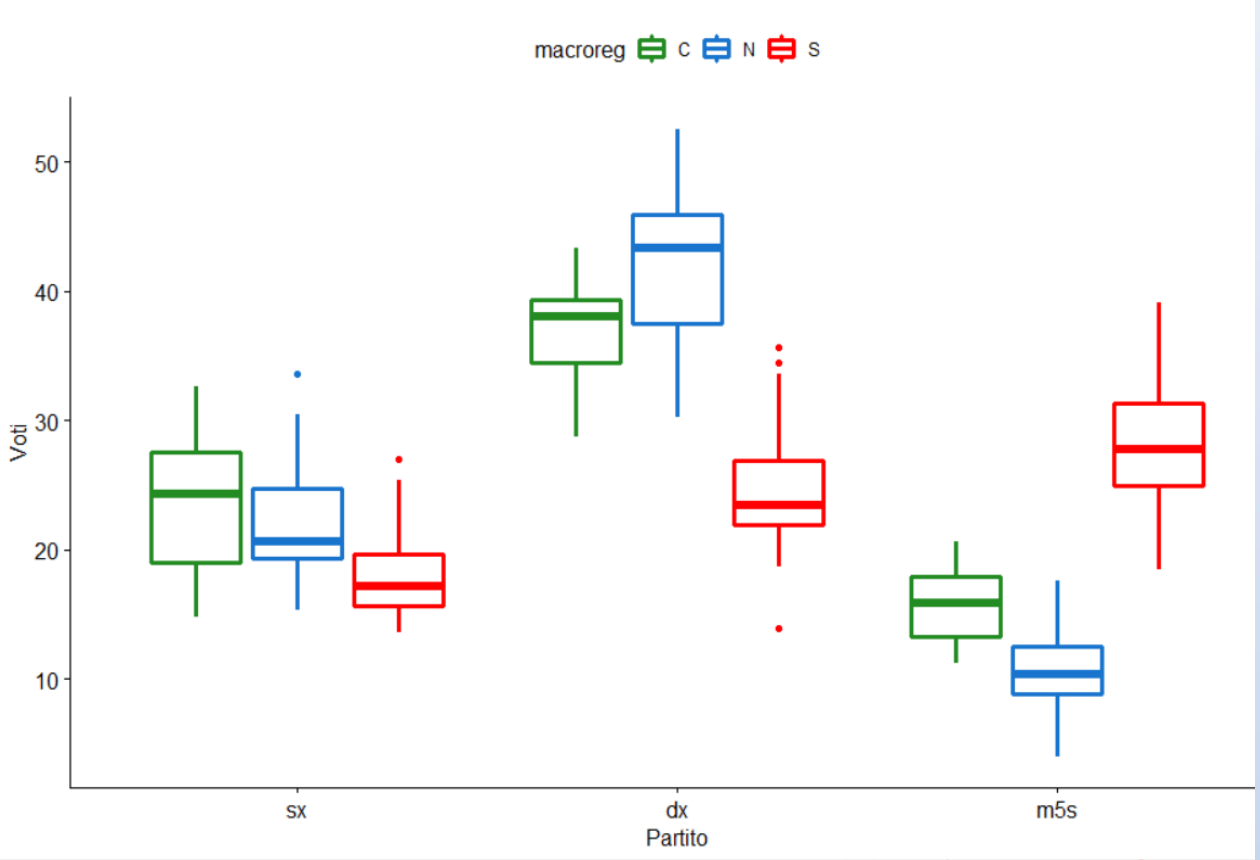
Two – ways ANOVA

Progetto MMIS A.A. 2019/2020

Ghiglione, Motta, Perelli

Analisi interazione, Voti vs Macro-regione





Two – ways ANOVA

Analisi interazione

In entrambi i casi, i grafici non sono paralleli, quindi il modello corretto è con interazione.

$$\boldsymbol{y} = \beta_o + \beta_1 \boldsymbol{z_1} + \beta_2 \boldsymbol{z_2} + \beta_3 \boldsymbol{z_1 z_2} + \boldsymbol{\varepsilon}$$

$$\varepsilon_k \sim N(0, \sigma^2) \qquad y_k \sim N(\mu_{ij}, \sigma^2) \quad i = 1, \dots, g;\ j = 1, \dots, b;\ k = 1, \dots, n_{ij}$$

```
mod <- aov(tab$voti ~ tab$partito*tab$macroreg)
```

Two – ways ANOVA

Analisi ipotesi modello

Per poter procedere con l'analisi del modello ANOVA, è necessario che due ipotesi siano verificate:

- omoschedasticità tra i gruppi dei residui (Barlett's Test, Levene's Test)
- normalità intra-gruppo dei dati (Shapiro-Wilk test)

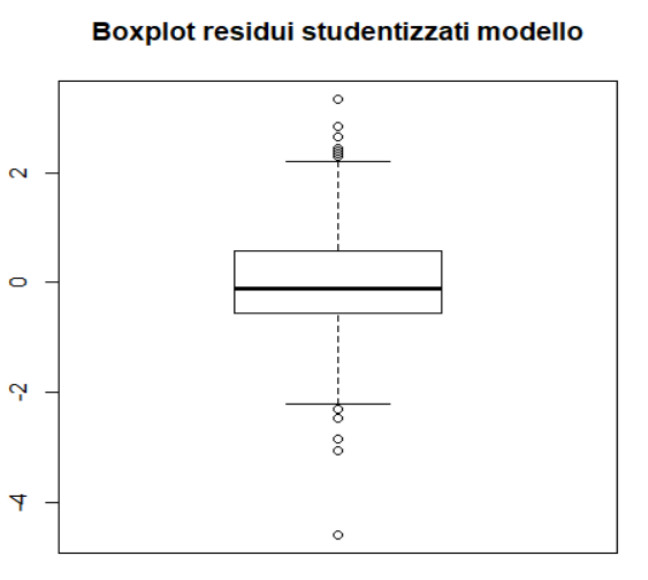
```
Analisi ipotesi modello
--------------------
        Bartlett test of homogeneity of variances

data:  tab$voti and as.factor(tab$partito):as.factor(tab$macroreg)
Bartlett's K-squared = 70.309, df = 8, p-value = 4.265e-12

---------------------
Levene's Test for Homogeneity of Variance (center = median)
       Df F value    Pr(>F)
group   8  5.4731 1.784e-06 ***
      312
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
---------------------
Shapiro-Wilk test sui gruppi
       dx:C        dx:N        dx:S       m5s:C       m5s:N       m5s:S        sx:C        sx:N        sx:S
0.106786258 0.102871969 0.040488248 0.378089268 0.795024561 0.978254634 0.411011071 0.004918852 0.003890775
```

Two – ways ANOVA

Analisi ipotesi modello e correzione

- P-value Barlett e Levene Test molto bassi, l'ipotesi di omoschedasticità dei residui è da rifiutare.
- Tre gruppi su nove hanno P-value dello Shapiro-Wilk test sotto 0.05, per un terzo dei gruppi l'ipotesi di normalità dei dati non è verificata.

Le ipotesi del modello ANOVA non sono soddisfatte, bisogna fare delle correzioni.

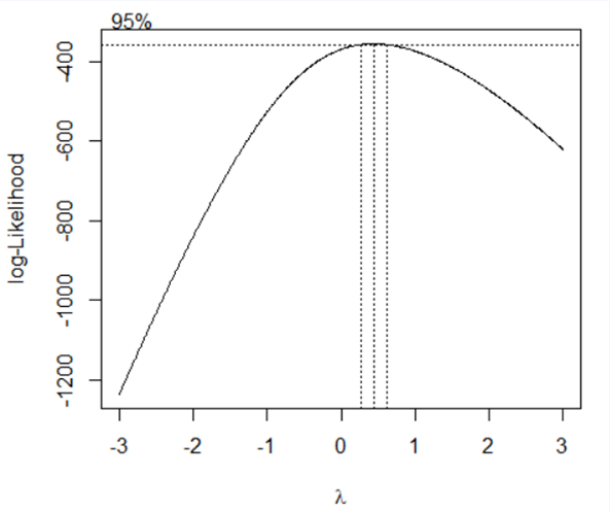


Si osserva dal boxplot un certo numero di dati con valori dei residui outlier.

Si è deciso pertanto trascurare questi dati con residui molto elevati (13 punti su 321 osservazioni).

Two – ways ANOVA

Correzione modello



Applicando una trasformazione Box-Cox, il risultato della massimizzazione della likelihood va interpretato approssimandolo con il valore notevole più vicino.

```
Lambda risultato dal box-cox
[1] 0.44
----------------------
Lambda utilizzato
[1] 0.5
```

Si ottiene così il modello corretto:

$$\sqrt{\boldsymbol{y}} = \beta_o + \beta_1 \boldsymbol{z_1} + \beta_2 \boldsymbol{z_2} + \beta_3 \boldsymbol{z_1 z_2} + \boldsymbol{\varepsilon}$$

```
mod_cox <- aov(tab$voti^lambda ~ tab$partito*tab$macroreg)
```

Two – ways ANOVA

Analisi ipotesi modello corretto

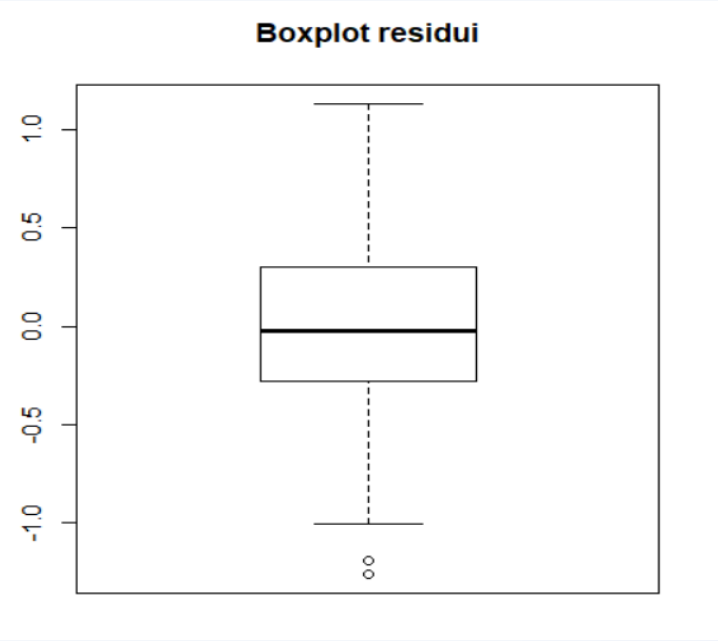
```
Analisi ipotesi modello trasformato
--------------------
        Bartlett test of homogeneity of variances

data:  tab$voti^lambda and as.factor(tab$partito):as.factor(tab$macroreg)
Bartlett's K-squared = 10.381, df = 8, p-value = 0.2393

--------------------
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   8  1.2715 0.2579
      299
--------------------
Shapiro-Wilk test sui gruppi
      dx:C       dx:N       dx:S      m5s:C      m5s:N      m5s:S       sx:C       sx:N       sx:S
0.12097323 0.09666936 0.29860132 0.38253061 0.83446918 0.96771435 0.39088661 0.10356863 0.02036186
```

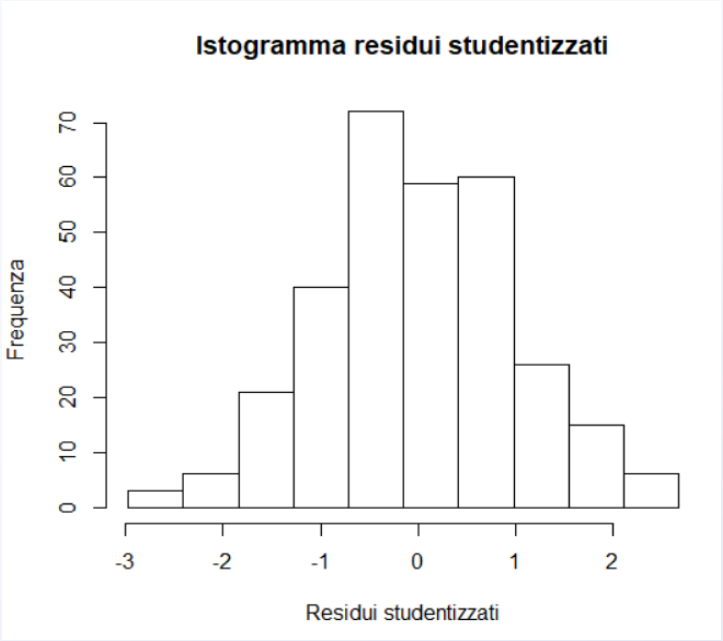
- P-value Barlett e Levene Test elevati, l'ipotesi di omoschedasticità dei residui non è da rifiutare.
- Solo un gruppo su nove ha P-value dello Shapiro-Wilk test sotto 0.05, per otto gruppi su nove l'ipotesi di normalità dei dati è accettabile.

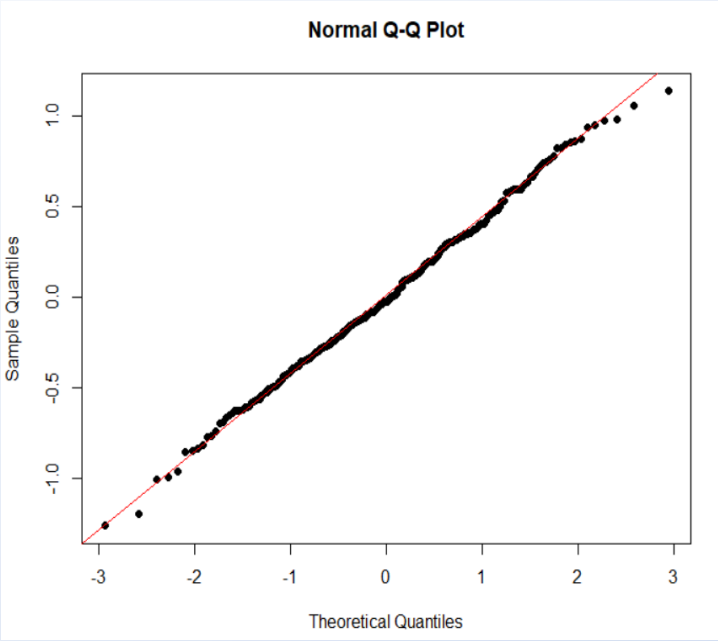
Le ipotesi del modello ANOVA risultano verificate in maniera soddisfacente, si può procedere con l'analisi del modello.

Two – ways ANOVA

Diagnostica modello corretto







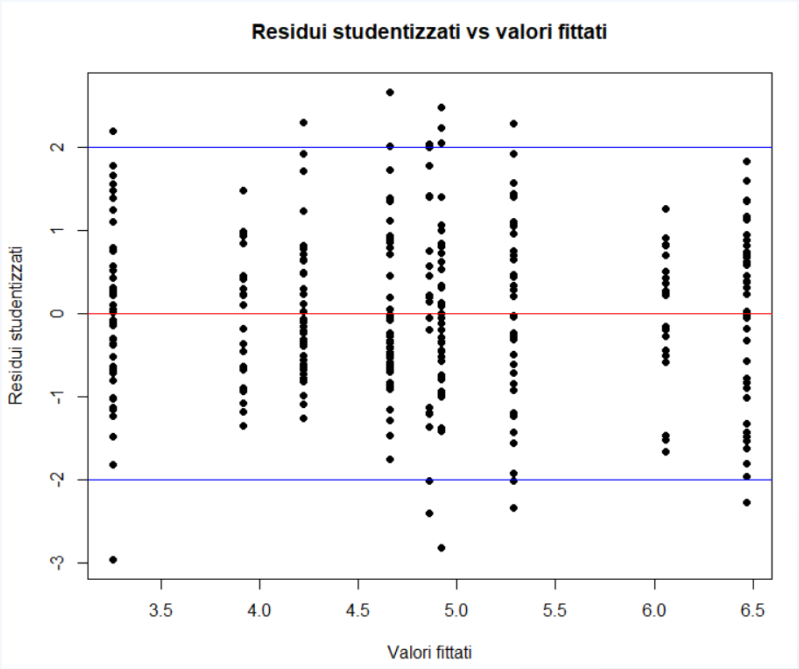

```
        Shapiro-Wilk normality test

data:  mod_cox$residuals
W = 0.9975, p-value = 0.9234
```

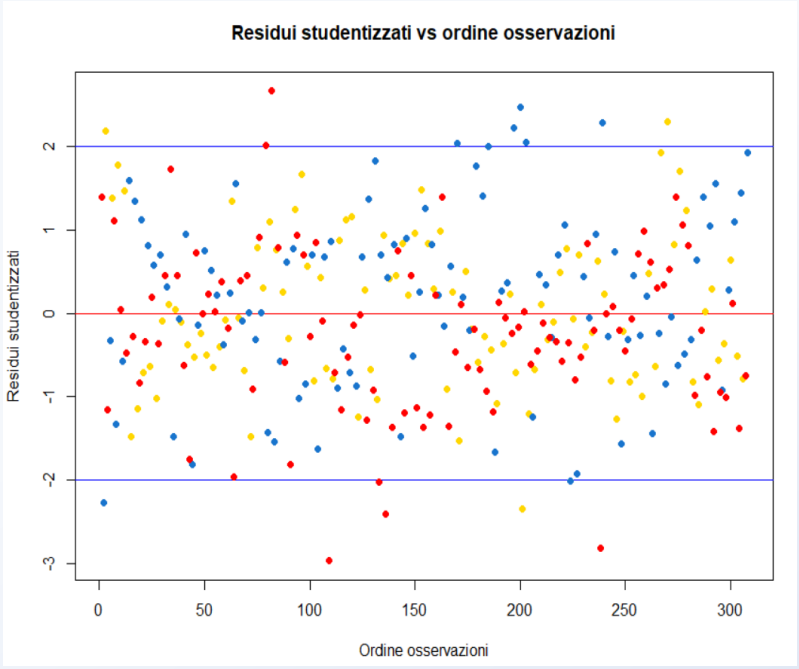
In seguito alla trasformazione di Box-Cox, osservando i vari grafici di diagnostica e il risultato dello Shapiro-Wilk test, le ipotesi di omoschedasticità e normalità dei residui risultano soddisfatte.

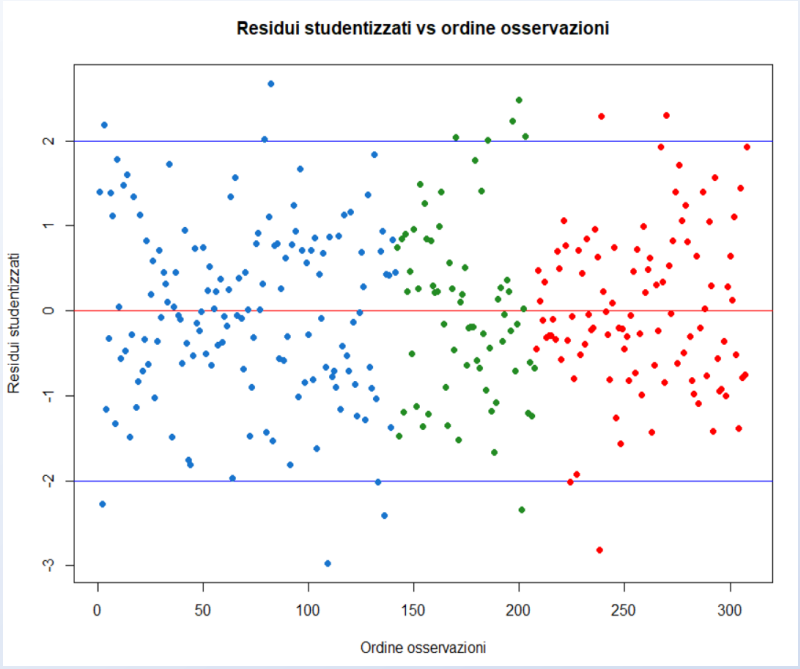
Two – ways ANOVA

Ghiglione, Motta, Perelli

Diagnostica modello corretto







I residui studentizzati risultano distribuiti in modo uniforme attorno al valore zero.

Distinguendo con colori diversi i partiti nel grafico a sinistra e le macro-regioni nel grafico a destra, si osserva come i residui non presentino pattern particolari in funzione dei due fattori.

Two – ways ANOVA

Test Anova

Nell'Anova two-ways si effettua una scomposizione della varianza rispetto alle medie dei due fattori e della loro interazione.

$$\mathbb{E}[X_{ijk}] = \beta_0 + \tau_i + \rho_j + \gamma_{ij}$$
$$i = 1, \dots, g;\ j = 1, \dots, b;\ k = 1, \dots, n$$

$$SS_{tot} = \sum_{j=1}^{b} \sum_{i=1}^{g} \sum_{k=1}^{n} (x_{ijk} - \bar{x})^2 = SS_1 + SS_2 + SS_{int} + SS_{res}$$

$$SS_1 = \sum_{i=1}^{g} bn\,(\bar{x}_{i\cdot} - \bar{x})^2; \qquad MS_1 = SS_1 \,/\, (\mathrm{g} - 1)$$
$$SS_2 = \sum_{j=1}^{b} gn\,(\bar{x}_{j\cdot} - \bar{x})^2; \qquad MS_2 = SS_2 \,/\, (\mathrm{b} - 1)$$
$$SS_{int} = \sum_{i=1}^{g} \sum_{j=1}^{b} n\,(\bar{x}_{ij} - \bar{x}_{i\cdot} - \bar{x}_{j\cdot} + \bar{x})^2; \qquad MS_{int} = SS_{in} \,/\, (\mathrm{g} - 1)(\mathrm{b} - 1)$$
$$SS_{res} = \sum_{j=1}^{b} \sum_{i=1}^{g} \sum_{k=1}^{n} (x_{ijk} - \bar{x}_{ij})^2; \qquad MS_W = SS_{res} \,/\, \mathrm{gb}(\mathrm{n} - 1)$$

Two – ways ANOVA

Test Anova

Vengono fatti tre test differenti:

- Test 1: effetto fattore 1 sulle medie
- Test 2: effetto fattore 2 sulle medie
- Test 3: effetto interazione fattori sulle medie

$$H_{01}: \tau_i = 0 \quad i = 1, \dots, g$$

$$H_{02}: \rho_j = 0 \quad j = 1, \dots, b$$

$$H_{03}: \gamma_{ij} = 0 \quad i = 1, \dots, g; j = 1, \dots, b$$

Come nell'ANOVA one way, i test si basano sul confronto tra *varianza between* e *varianza within* il cui rapporto è distribuito come una Fisher.

$$\frac{MS_1}{MS_W} \sim \mathrm{F}(\mathrm{g}-1, \mathrm{gb}(\mathrm{n}-1))$$

$$\frac{MS_2}{MS_W} \sim \mathrm{F}(\mathrm{b}-1, \mathrm{gb}(\mathrm{n}-1))$$

$$\frac{MS_{int}}{MS_W} \sim \mathrm{F}((\mathrm{g}-1)(\mathrm{b}-1), \mathrm{gb}(\mathrm{n}-1))$$

Two – ways ANOVA

Test Anova

```
                          Df Sum Sq Mean Sq F value Pr(>F)
tab$partito                2 159.76   79.88 431.913 <2e-16 ***
tab$macroreg               2   1.25    0.62   3.371 0.0357 *
tab$partito:tab$macroreg   4 141.75   35.44 191.607 <2e-16 ***
Residuals                299  55.30    0.18
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
13 observations deleted due to missingness
```

Il p-value del terzo test è molto basso, rifiuto l’ipotesi H_{03}, tra i due fattori c’è interazione.

Ciò è in linea con le osservazioni dei dati tramite gli *interaction plot.*

Two – ways ANOVA

Analisi post-hoc, test di Tukey

Tramite la tecnica post-hoc di Tukey, è possibile:

- fare test sulla differenza delle medie tra i gruppi
- trovare intervallo di confidenza della differenza delle medie tra i gruppi

Dato il modello con k gruppi di cardinalità n, il gruppo A e B di medie $\bar{x}_A$ e $\bar{x}_B$ con $\bar{x}_A > \bar{x}_B$

$$H_0: \mu_A = \mu_B$$
$$H_1: \mu_A \neq \mu_B$$

$$\frac{\bar{x}_A - \bar{x}_B}{\sqrt{S_p^2/n}} < q \sim Q_{k,N-k} \quad sotto\ H_0 \qquad con\ N = nk\ e\ S_p^2\ varianza\ pooled\ del\ modello$$

$Q_{k,N-k}$ è la *studentized range distibuition*.

Dati k campioni casuali di n osservazioni $N(\mu, \sigma^2)$ e di ciascuno si calcola la media $\bar{x}_i$, si ha che:

$$\frac{\bar{x}_{max} - \bar{x}_{min}}{\sqrt{S_p^2/n}} \sim Q_{k,N-k}$$

Two – ways ANOVA

Analisi post-hoc, test di Tukey

Posso costruire il test (unilatero) di livello α di regione critica.

$$RC_\alpha = \{ \bar{x}_A - \bar{x}_B > \sqrt{\frac{S_p^2}{n}} Q_{1-\alpha,k,N-k} \}$$

Posso costruire l'intervallo di confidenza di livello 1 - α .

$$IC_{1-\alpha} = \{ |\bar{x}_A - \bar{x}_B| < \sqrt{\frac{S_p^2}{n}} Q_{1-\alpha/2,k,N-k} \}$$

Se i gruppi A e B hanno numerosità diverse n_A e n_B, si utilizza la correzione di Tukey-Kramer.

$$RC_\alpha = \{ \bar{x}_A - \bar{x}_B > \sqrt{\frac{S_p^2}{2}\left(\frac{1}{n_A} + \frac{1}{n_B}\right)} Q_{1-\alpha,k,N-k} \}$$

$$IC_{1-\alpha} = \{ |\bar{x}_A - \bar{x}_B| < \sqrt{\frac{S_p^2}{2}\left(\frac{1}{n_A} + \frac{1}{n_B}\right)} Q_{1-\alpha/2,k,N-k} \}$$

Two – ways ANOVA

Analisi post-hoc, test di Tukey

Applicando il test di Tukey al modello, si osserva che tutti i p-value sono bassi, quindi rifiuto l'ipotesi che le medie siano uguali a coppie:

```
Test Tukey partiti al Nord
                diff       lwr       upr       p adj
m5s:N-dx:N -3.213766 -3.499098 -2.928434 8.91065e-13
sx:N-dx:N  -1.810765 -2.102292 -1.519237 8.91065e-13
sx:N-m5s:N  1.403001  1.119427  1.686576 8.91065e-13
--------------------
Test Tukey partiti al Centro
                diff        lwr        upr        p adj
m5s:C-dx:C -2.1415608 -2.5515215 -1.7316000 8.910650e-13
sx:C-dx:C  -1.1983900 -1.6182420 -0.7785379 9.778844e-13
sx:C-m5s:C  0.9431708  0.5279995  1.3583421 3.321086e-10
--------------------
Test Tukey partiti al Sud
                diff         lwr        upr        p adj
m5s:S-dx:S  0.3650527  0.05469181  0.6754135 8.496500e-03
sx:S-dx:S  -0.7008257 -1.01118653 -0.3904648 4.320244e-10
sx:S-m5s:S -1.0658783 -1.37416319 -0.7575935 9.583445e-13
```

```
Test Tukey macro-regioni Lega
                diff         lwr        upr        p adj
dx:N-dx:C  0.4127354  0.05359446  0.7718764 1.13957e-02
dx:S-dx:C -1.1366654 -1.50380610 -0.7695248 9.75886e-13
dx:S-dx:N -1.5494009 -1.85238303 -1.2464187 8.91065e-13
--------------------
Test Tukey macro-regioni PD
                diff        lwr        upr        p adj
sx:N-sx:C -0.1996392 -0.5633448  0.1640665 7.368654e-01
sx:S-sx:C -0.6391011 -1.0103253 -0.2678770 5.348194e-06
sx:S-sx:N -0.4394620 -0.7386509 -0.1402731 2.235886e-04
--------------------
Test Tukey macro-regioni M5S
                diff       lwr        upr        p adj
m5s:N-m5s:C -0.6594696 -1.006601 -0.3123387 2.888181e-07
m5s:S-m5s:C  1.3699480  1.009949  1.7299471 8.968382e-13
m5s:S-m5s:N  2.0294176  1.736262  2.3225729 8.910650e-13
```

Two – ways ANOVA

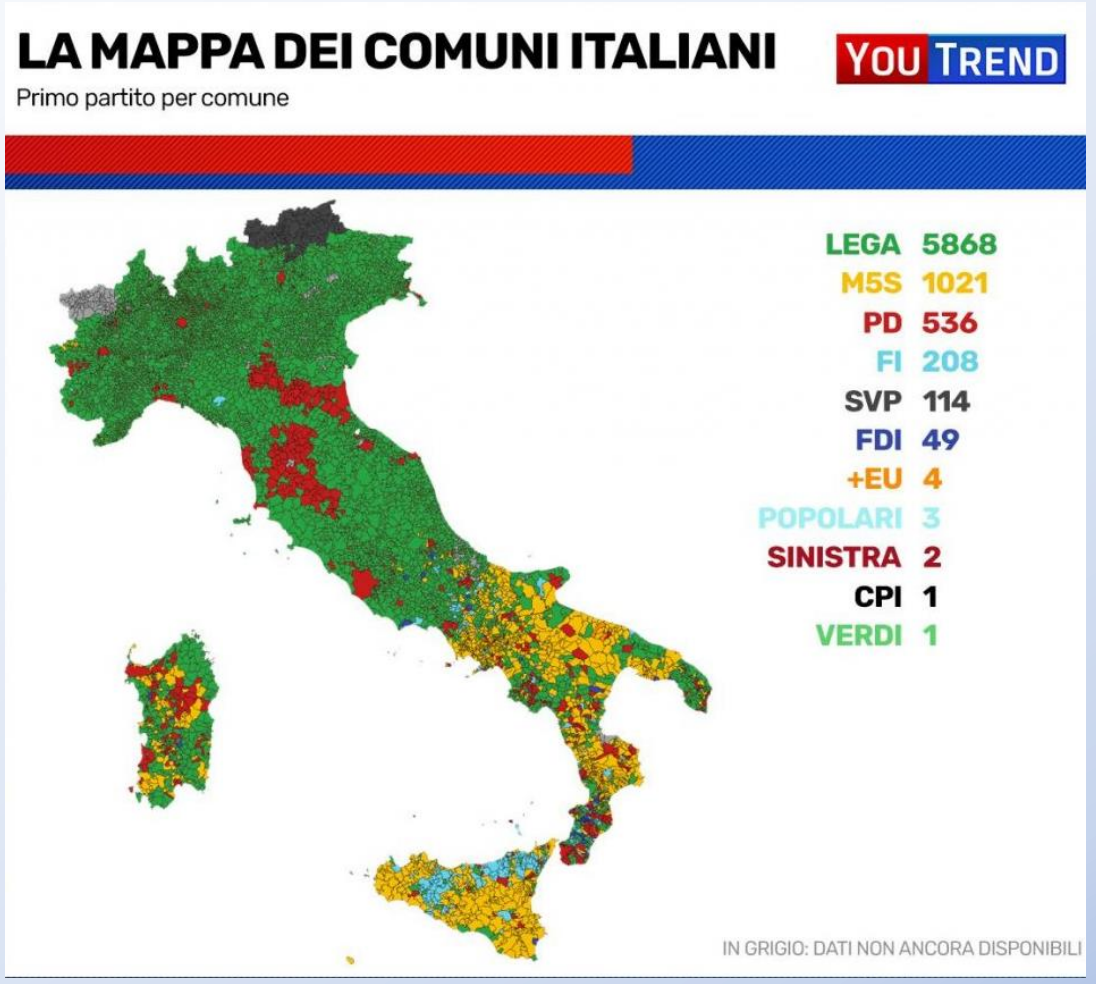
Conclusioni

Ghiglione, Motta, Perelli

Elezioni europee: al Nord la Lega supera il 40%, M5s primo partito a Sud e nelle isole

HUFFPOST

Ma quindi, è vero che, come è stato proclamato dai giornali e come sembra dalle mappe della distribuzione del voto, la geografia ha avuto un impatto determinante sui risultati dei partiti?



Two – ways ANOVA

Ghiglione, Motta, Perelli

La nostra risposta è…

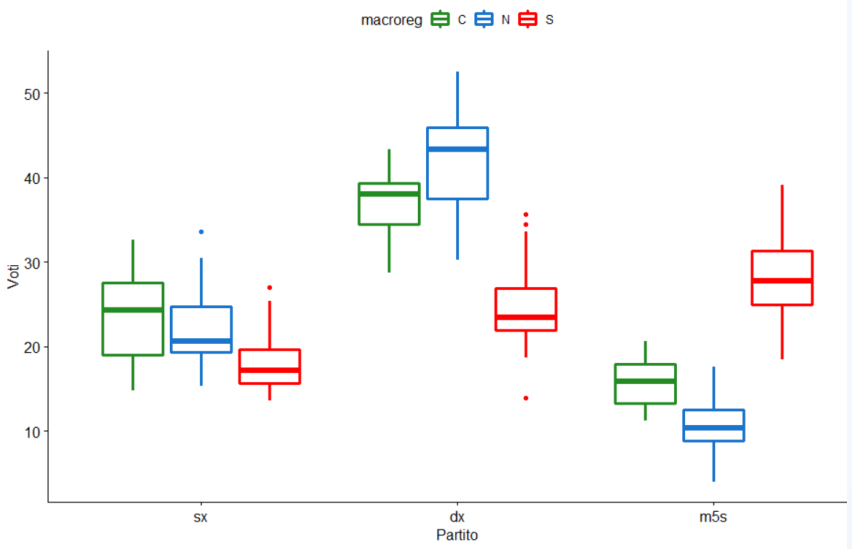
Two – ways ANOVA

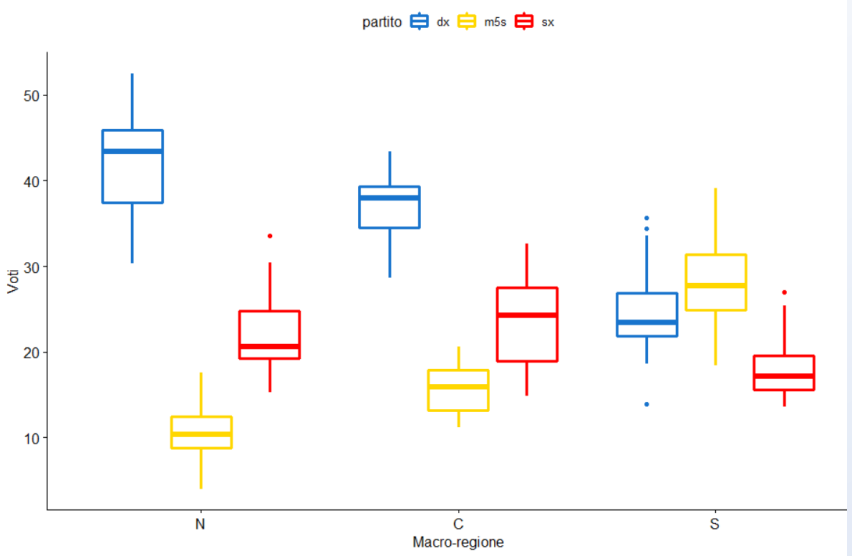
Ghiglione, Motta, Perelli

La nostra risposta è…

Two – ways ANOVA

Conclusioni





Anche se è ben noto come, storicamente, il comportamento elettorale del Nord, del Centro e del Sud Italia sia diverso, l'applicazione del modello ANOVA Two-ways ha permesso di dimostrare che, statisticamente, le performance elettorali dei partiti siano decisamente influenzate dalla suddivisione geografica dell'Italia in macro-aree.

Alla luce di ciò, suggeriamo a chi volesse in futuro fare indagini più approfondite sulle cause demografiche e sociali dei risultati elettorali, di tenere in forte considerazione questo fattore.

Grazie per l’attenzione da Carlo, Paolo e Leonardo.