Progetto Modelli Statistici

Abdula Kalus

a.a. 2023/2024

Analisi dei dati sulle mancie

Il dataset in **tips.csv** contiene infromazioni registrate da un cameriere sulle mancie ricevute in un periodo di qualche mese. Obiettivo dell'analisi è investigare la relazione tra il valore della mancia e le caratteristiche presenti nel dataset. Carichiamo i dati e guardiamo le prime righe della matrice dei dati

Dataset head

total_bill	tip	sex	smoker	day	time	size
16.99	1.01	Female	No	Sun	Dinner	2
10.34	1.66	Male	No	Sun	Dinner	3
21.01	3.50	Male	No	Sun	Dinner	3
23.68	3.31	Male	No	Sun	Dinner	2
24.59	3.61	Female	No	Sun	Dinner	4
25.29	4.71	Male	No	Sun	Dinner	4

Il dataset è composto da 7 variabili e 244 osservazioni.

La variabile **y** è **tip**, e avrà il ruolo di variabile risposta.

- **tip**: mancia

Riportiamo sotto il significato delle variabili esplicative

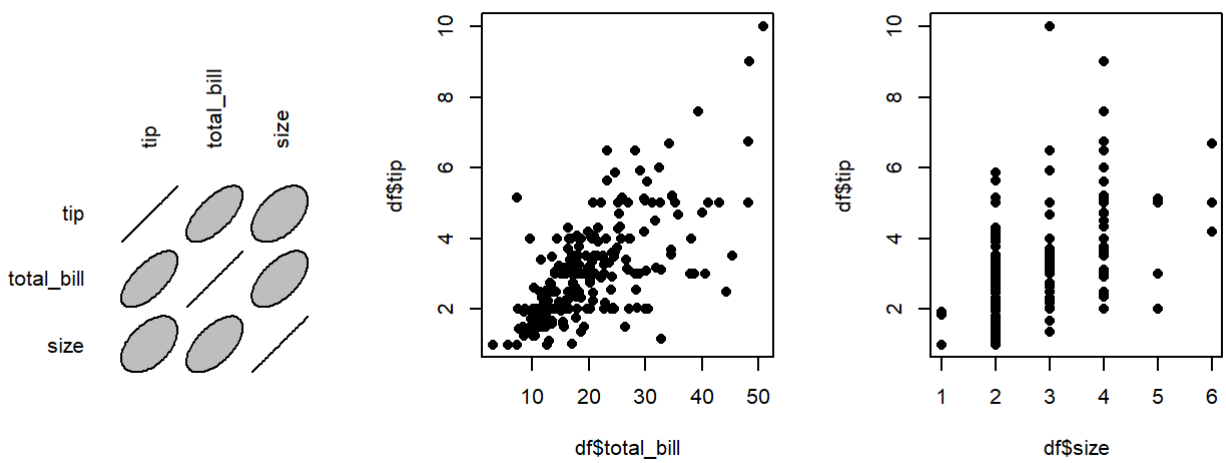
1. **total_bill**: conto totale
2. **sex**: sesso del cliente (*Male* o *Female*)
3. **smoker**: cliente fumaotre (*si* o *no*)
4. **day**: giorno della settimana
5. **time**: orario del servizio (*Lunch* o *Dinner*)
6. **size**: numero di persone

Analisi esplorativa

Summary

total_bill	tip	sex	smoker	day	time	size
Min. : 3.07	Min. : 1.000	Female: 87	No :151	Fri :19	Dinner:176	Min. :1.00
1st Qu.:13.35	1st Qu.: 2.000	Male :157	Yes: 93	Sat :87	Lunch : 68	1st Qu.:2.00
Median :17.80	Median : 2.900			Sun :76		Median :2.00
Mean :19.79	Mean : 2.998			Thur:62		Mean :2.57
3rd Qu.:24.13	3rd Qu.: 3.562					3rd Qu.:3.00
Max. :50.81	Max. :10.000					Max. :6.00

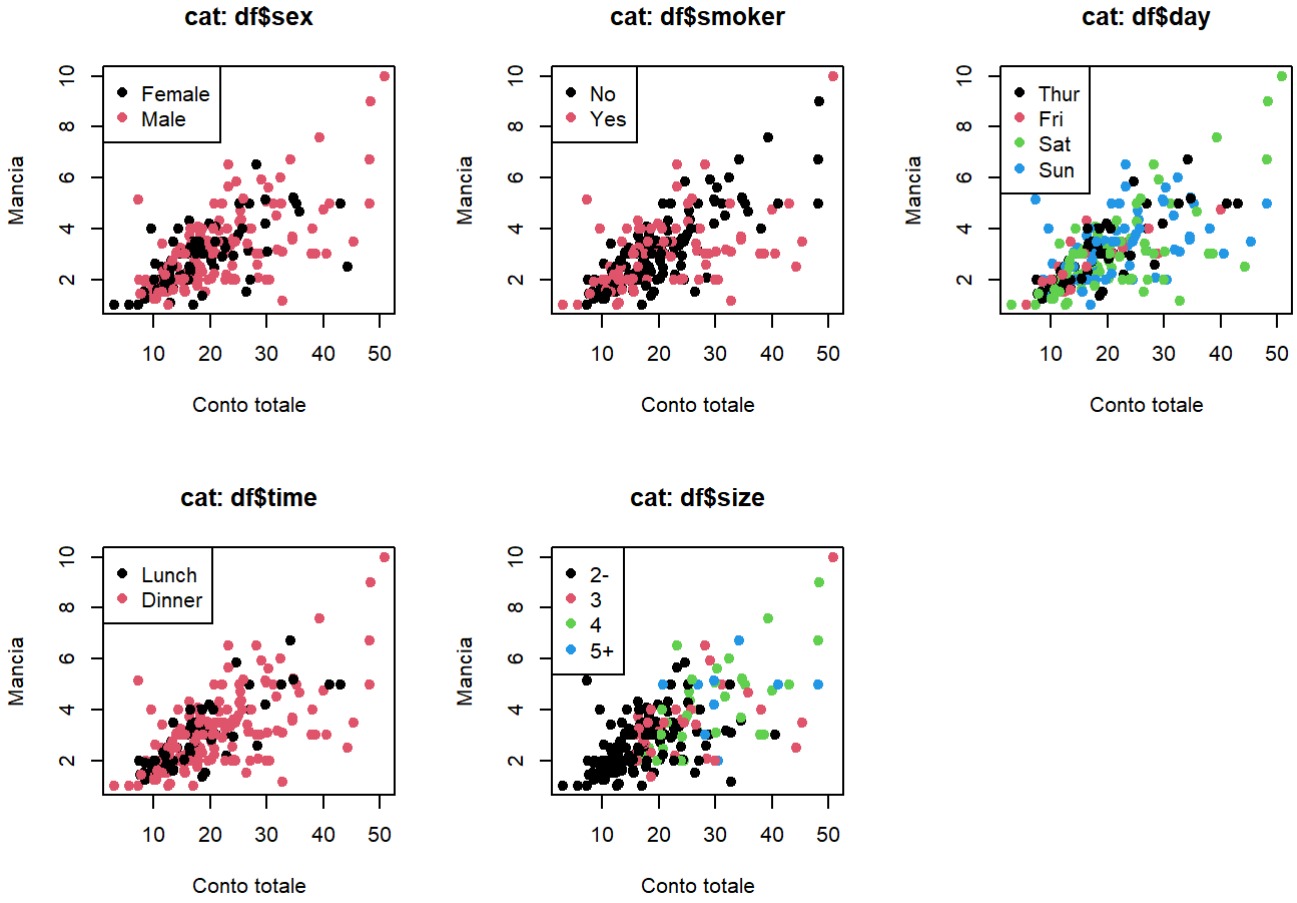
Possiamo vedere dalla tabella riassuntiva ottenuta con il comando *summary* alcune sintesi statistiche sul dataset, sono presenti tre variabili continue (*total_bill*, *tip*, *size*) e quattro varibile categoriali (*sex*, *smoker*, *day*, *time*). Inoltre possiamo osservare che non è segnalata la presenza di valori NA.



Dal grafico rappresentante la correlazione tra le variabili quantitaitve possiamo vedere che ci sono delle buone correlazioni tra la variabile risposta e le variabili esplicative, tuttavia anche le due variaibili esplicative sono correlate tra di loro. Osseravando il grafico di dispersione della variabile size è stato deciso di convertirla da numerica a fattore.

Diagramma di dispersione in base ai gruppi delle variabili qualitative

Per osservare se nel grafico di dipsersione tra le mancie ed il conto toale sono presenti dei pattern tra i gruppi delle variabili qualitative, si osservino i grafici seguenti



Possiamo osservare che non sono presenti dei pattern nei diagrammi di dispersione rispetto alle rispettive variabili qualitative, i dati risultano distribuiti in modo omogeneo.

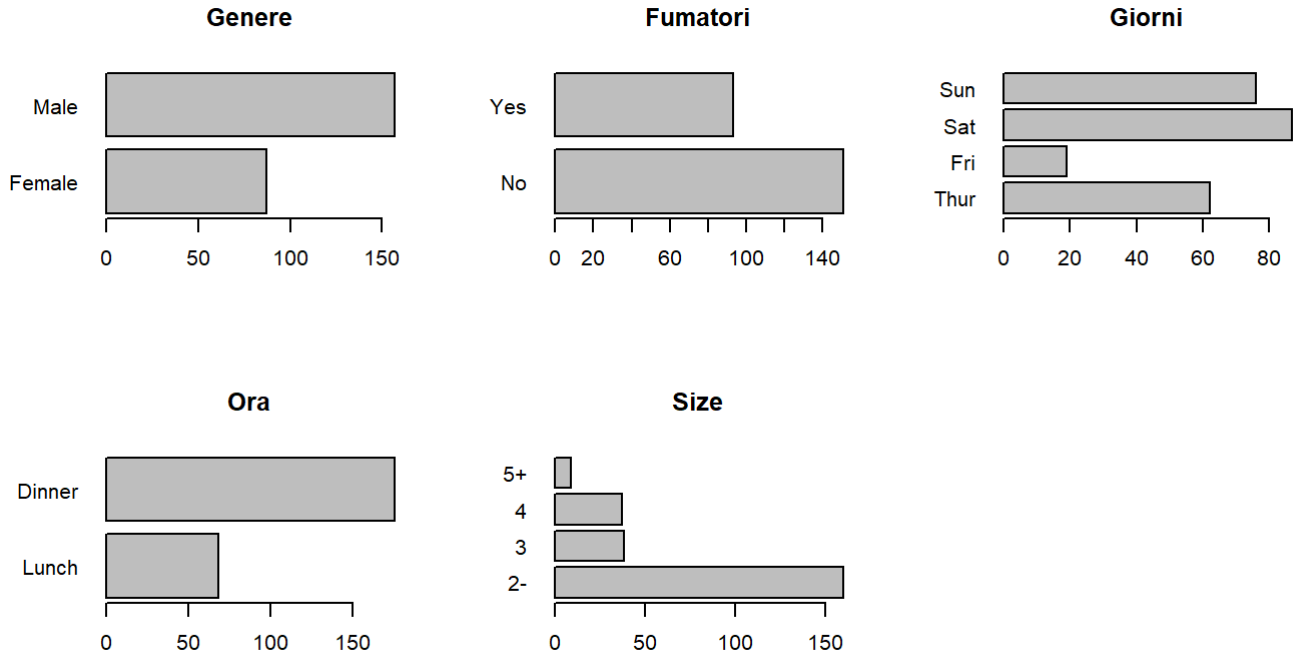
Panoramica delle variabili

I grafici seguenti mostrano delle carattersitche generali delle varaibili, in particolar modo:

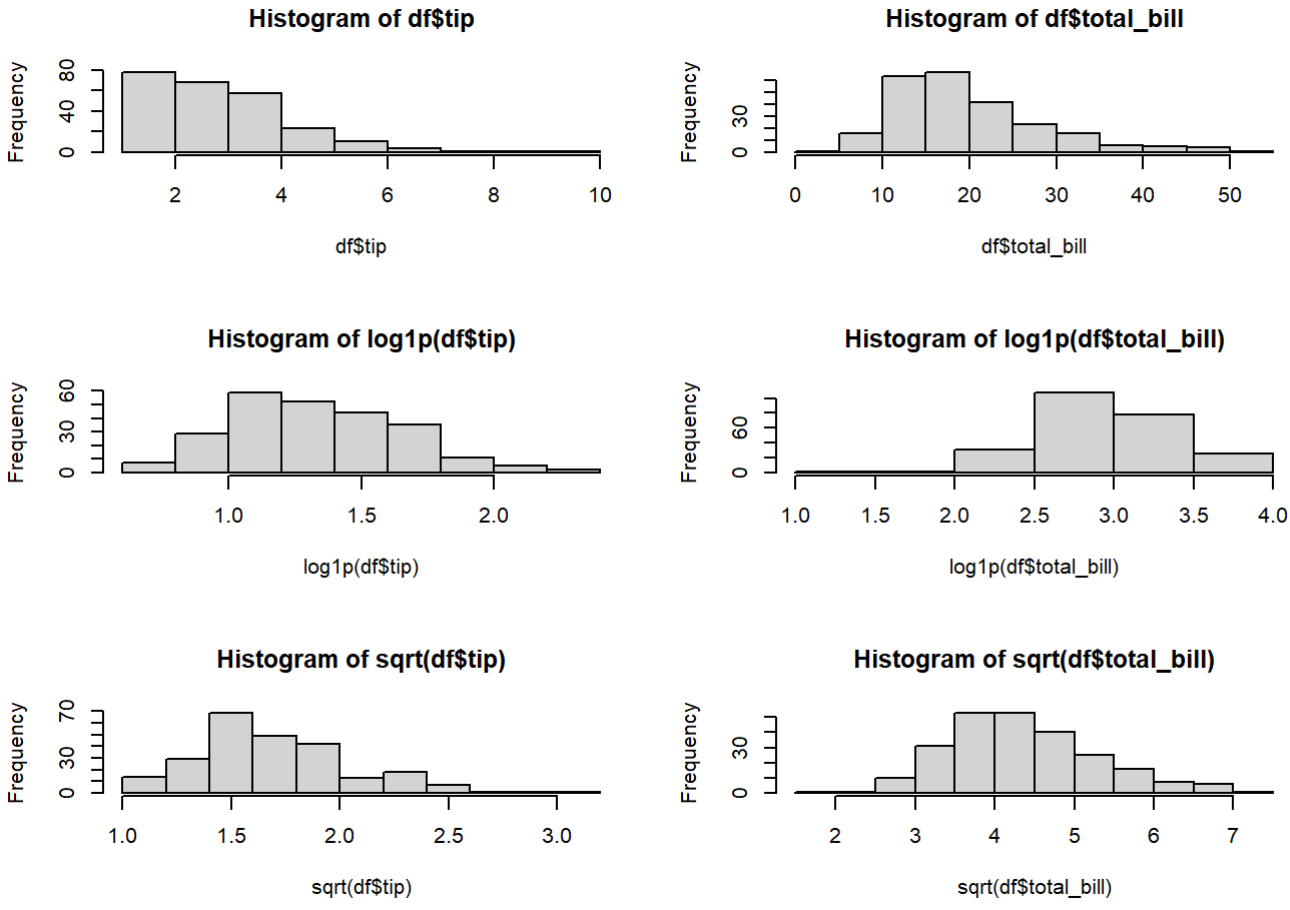
- la frequenza di ciascun gruppo nelle varie variabili esplicative
- la distribuzione della variabile **tip** e __total__bill__

Si puo osservare dalle frequenze che:

- gli uomini tendono a dare più spesso la mancia
- analogo per i non fumatori
- la maggiro parte delle mancie è data durante gli weekend
- la maggiro parte delle mancie è data durante le cene
- la maggior parte delle mancie è data dai tavoli meno numerosi

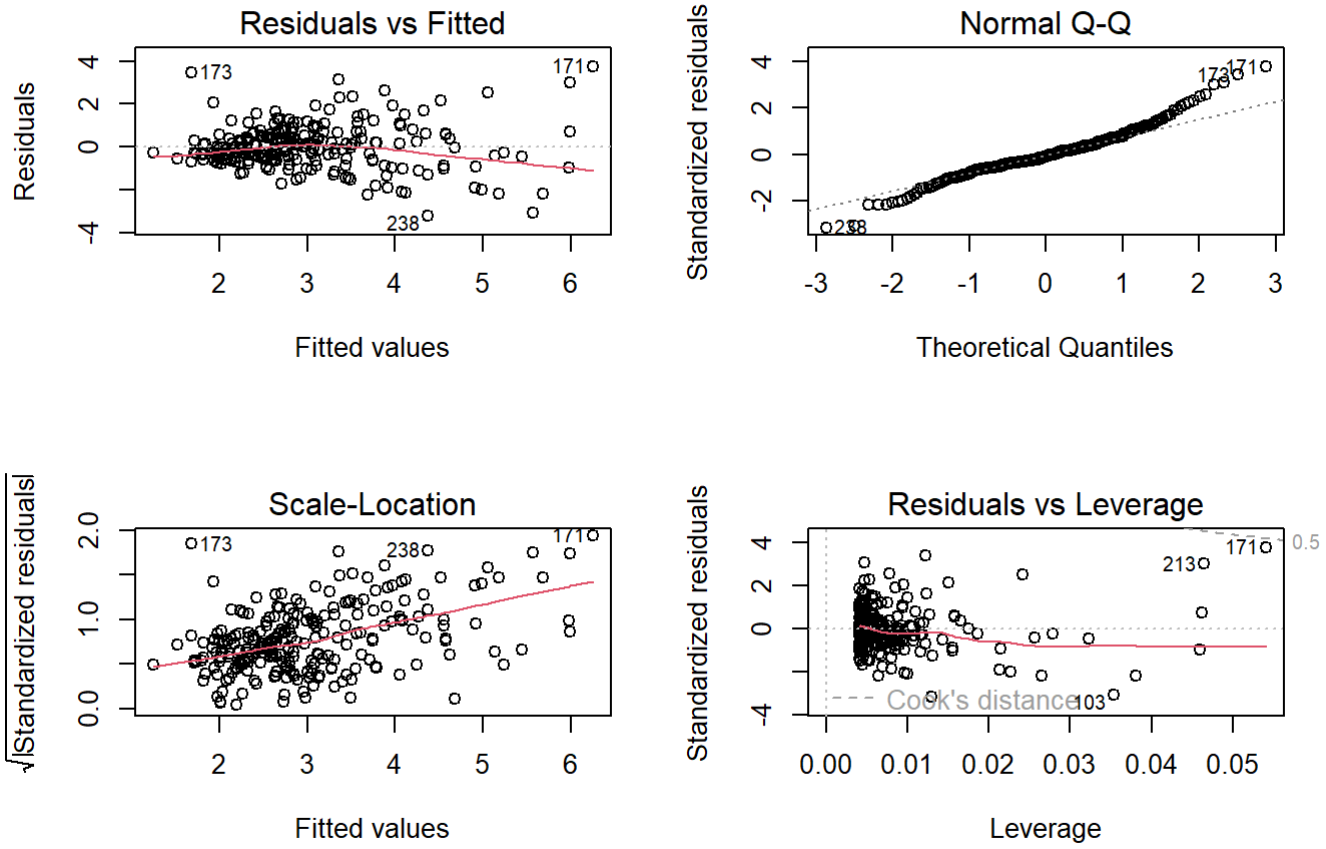


Per quanto riaguarda la distribuzione delle mancie si può notare che è decisamente asimmetrica. Questo si riflette sulla distribuzione dei residui del modello lineare per gli incassi. Per questo motivo viene proposta una trasformazione logaritimica e una sotto radice. La trasformazione logaritmica e radice appare, almeno marginalmente, normalizzante.



Nel seguente modello possiamo constatare l'osservazioen precedente ovvero che il modello senza nessuna trasformazione non rispetta l'omoschedasticità e la normalità.

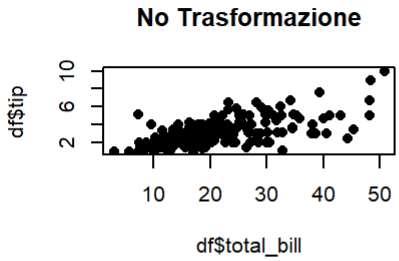
```
fit_nT <- lm(tip ~ total_bill, data = df)
```

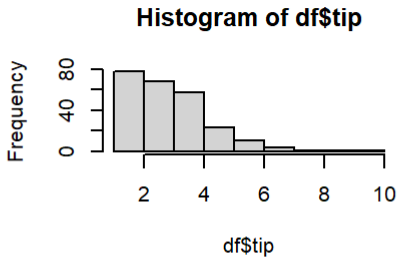


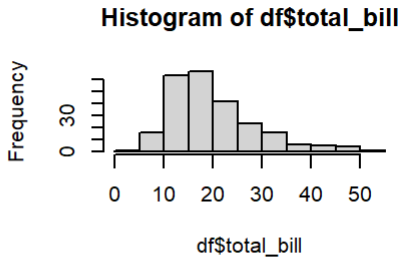
E' opportuna perciò adottare una delle trasformazione proposte, dopo alcune analisi preliminari risultano soddisfacenti il passaggio al logaritmo o alla radice.

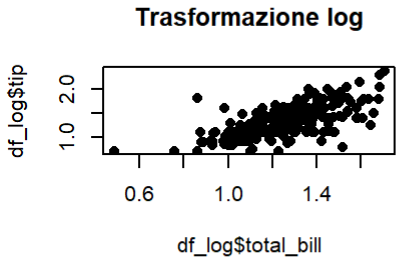
Creazione del modello

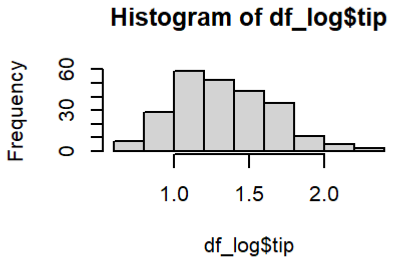
Dai seguenti grafici possiamo vedere che le due trasformazioni hanno un effetto positivo sul grafico di dispersione e sulle dsitribuzioni delle variabili.

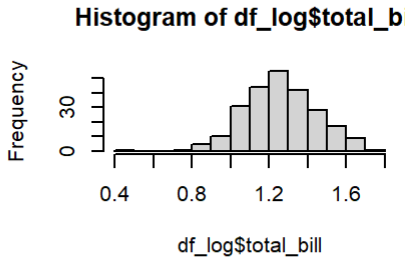


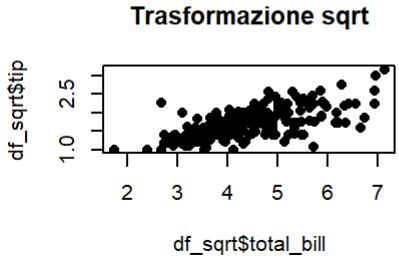


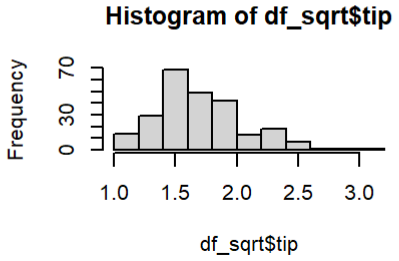


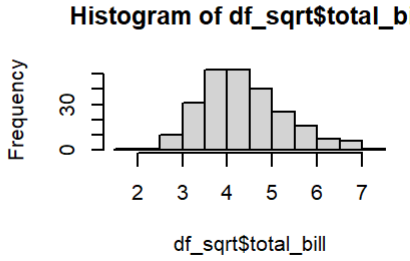










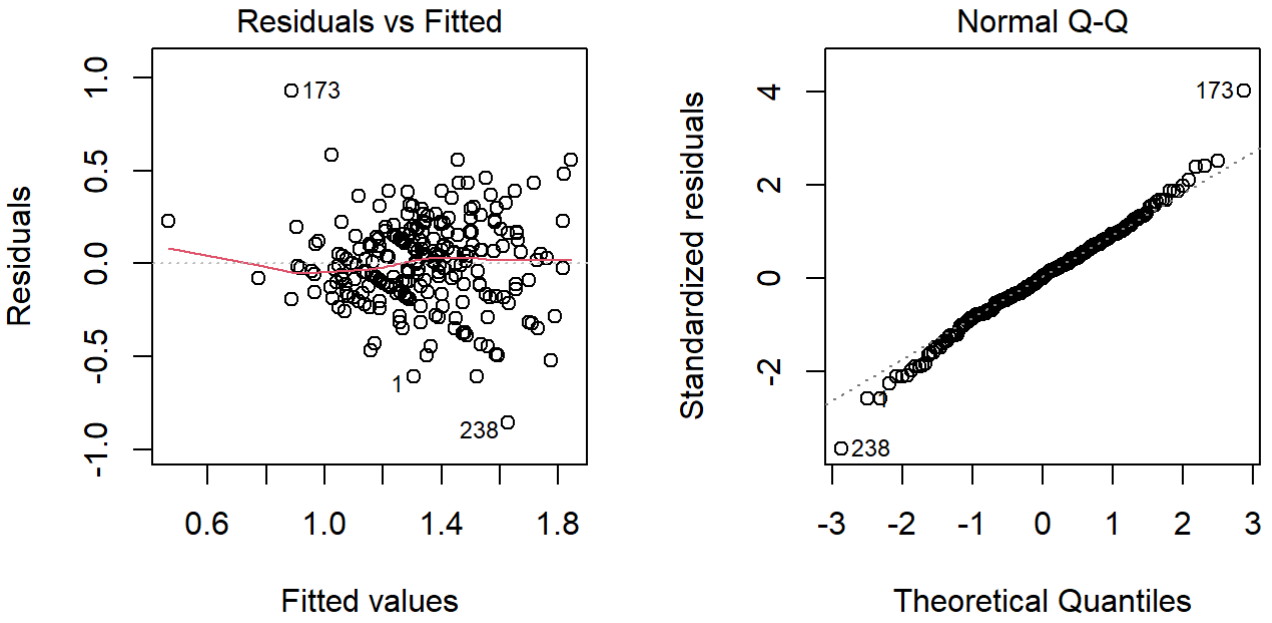


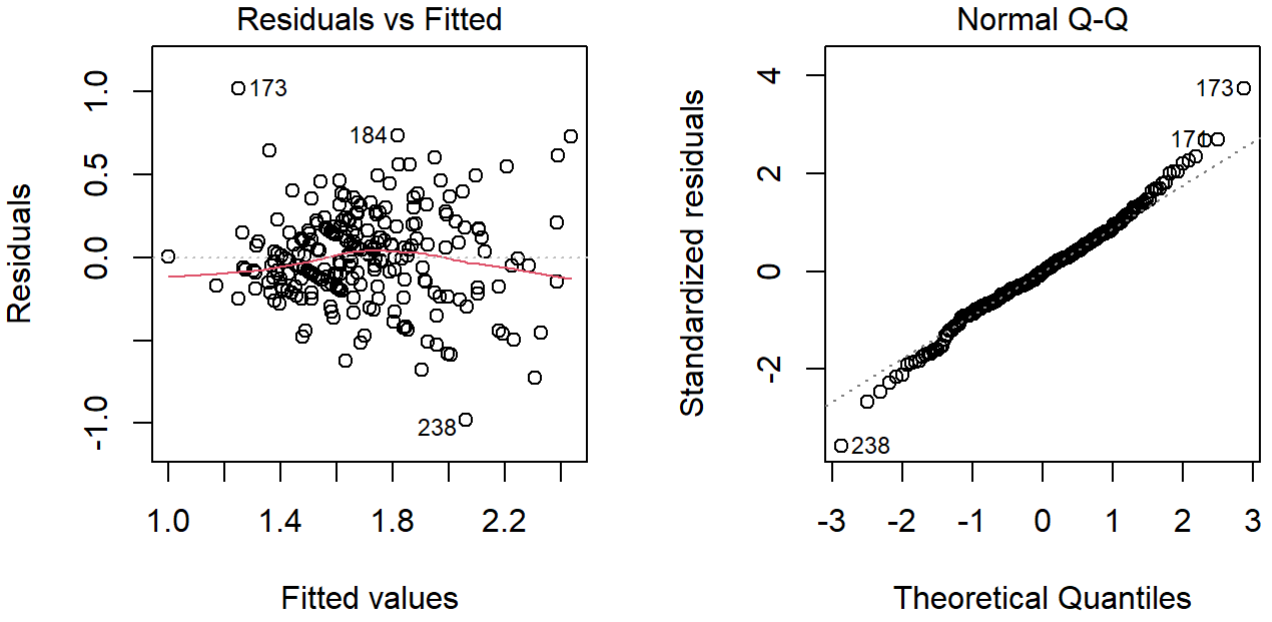
fit_lA & fit_sA

I primi due modelli sono dei modelli semplici tra la variabile esplicativa **total_bill** e la varaibile risposta **tip** con le rispettive trasformazioni:

- **fit_lA** modello log
- **fit_sA** modello sqrt

```
par(mfrow=c(1,2))
fit_lA <- lm(tip ~ total_bill, data = df_log)
fit_sA <- lm(tip ~ total_bill, data = df_sqrt)
```





Si ha che la variabile **total_bill** ha un effetto signficativo e positivo sulla variabile risposta **tip**, nel senso che quest'ultimo cresce al crescere del conto totale.

Inoltre dal grafico dei residui possiamo osservare che nessuno dei due modelli ha un grafico di dispersione ideale, non è rispettata appieno l'omoschedasticità, invece dal **qq-plot** osserviamo che sulle che la normalità non molto rispettata.

Summary modello fit_lA

```
##
## Call:
## lm(formula = tip ~ total_bill, data = df_log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.85404 -0.13198  0.00283  0.14743  0.92969
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.08648    0.10011  -0.864    0.389
## total_bill   1.13123    0.07885  14.347   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2343 on 242 degrees of freedom
## Multiple R-squared:  0.4596, Adjusted R-squared:  0.4574
## F-statistic: 205.8 on 1 and 242 DF,  p-value: < 2.2e-16
```

Dal summary del modelli possiamo vedere che:

- l'intercetta nel modello è negativo (un valore **non** realistico)
- solamente il secondo coefficenti risulta statisticamente significativo
- con un \(R^{2}\) pari a 0.4574 si ha che la percentuale di variabilità spiegata dal modello è buona

Summary modello fit_sA

```
##
## Call:
## lm(formula = tip ~ total_bill, data = df_sqrt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.97853 -0.16729 -0.00683  0.16039  1.01971
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.53106    0.08239   6.445 6.19e-10 ***
## total_bill   0.26688    0.01852  14.408  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2759 on 242 degrees of freedom
## Multiple R-squared:  0.4617, Adjusted R-squared:  0.4595
## F-statistic: 207.6 on 1 and 242 DF,  p-value: < 2.2e-16
```

per quanto rigiarda il secondo modello:

- l'intercetta è positivo
- entrambi i coefficenti risultano statisticamente significativi nel secondo modello
- con un \(R^{2}\) pari a 0.4595 si ha che la percentuale di variabilità spiegata dal modello è buona

fit_sB

```
fit_sB <- lm(tip ~ size, data = df_sqrt)
```

```
##
## Call:
## lm(formula = tip ~ size, data = df_sqrt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.69640 -0.23229  0.00295  0.19584  1.35812
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.56921    0.02618  59.931  < 2e-16 ***
## size3        0.23494    0.05977   3.931 0.000111 ***
## size4        0.42730    0.06042   7.072 1.66e-11 ***
## size5+       0.54140    0.11346   4.772 3.18e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3312 on 240 degrees of freedom
## Multiple R-squared:  0.2308, Adjusted R-squared:  0.2211
## F-statistic:    24 on 3 and 240 DF,  p-value: 1.282e-13
```

dal summary possiamo vedere che:

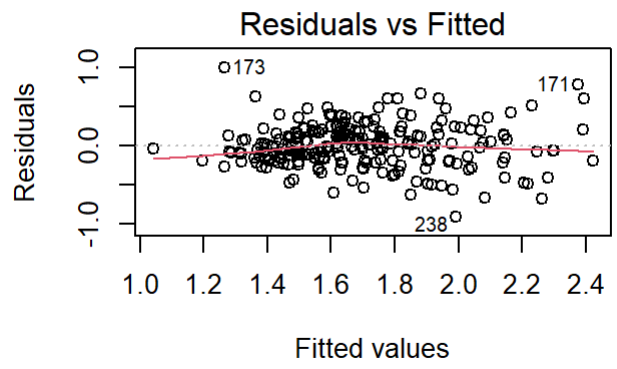
- l'intercetta è positivo
- tutti i coefficenti risultano statisticamente significativi
- si ha un \(R^{2}\) pari a 0.2211, minore rispetto al modello **fit_sA**

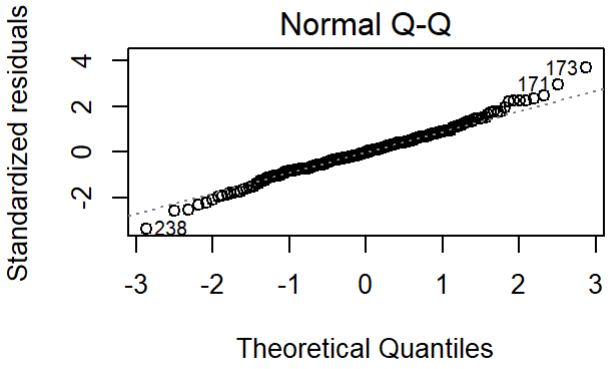
```
## Analysis of Variance Table
##
## Response: tip
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## size        3  7.8978 2.63261  23.999 1.282e-13 ***
## Residuals 240 26.3268 0.10969
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

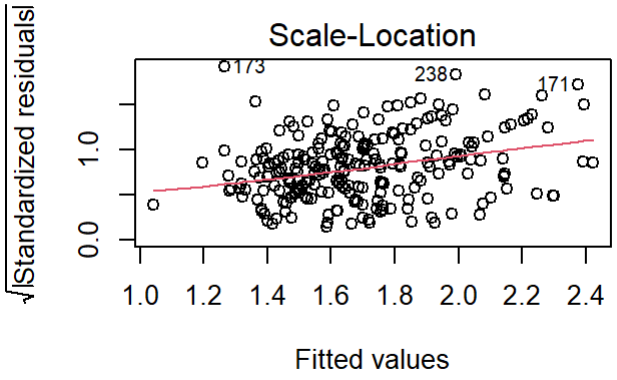
I risultati del test mostrano chiaramente che non vi è evidenza per ritenere che tutte le medie siano uguali (o equivalentemente che i coefficienti \(\beta_{j}\) , j = 2, . . . , 5 siano congiuntamente pari a zero)

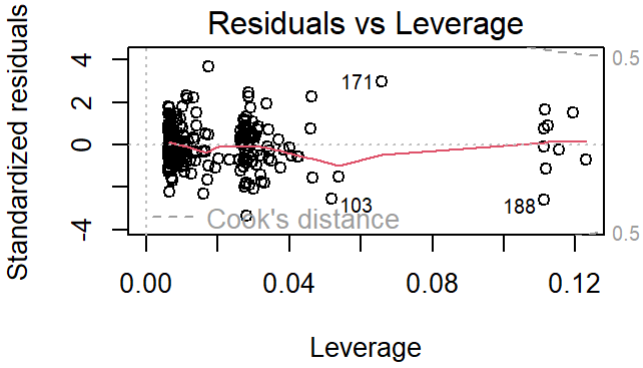
fit_sAB

```
fit_sAB <- lm(tip ~ total_bill + size, data = df_sqrt)
```









Summary fit_sAB

```
##
## Call:
## lm(formula = tip ~ total_bill + size, data = df_sqrt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.91091 -0.17078 -0.00947  0.16065  1.00297
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.62259    0.09338   6.667 1.79e-10 ***
## total_bill   0.23910    0.02294  10.424  < 2e-16 ***
## size3        0.04817    0.05279   0.912   0.3624
## size4        0.10849    0.05878   1.846   0.0662 .
## size5+       0.14077    0.10181   1.383   0.1680
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2752 on 239 degrees of freedom
## Multiple R-squared:  0.4712, Adjusted R-squared:  0.4623
## F-statistic: 53.24 on 4 and 239 DF,  p-value: < 2.2e-16
```

dal summary possiamo vedere che:

- l'intercetta è positivo
- solamente il coefficente **total_bill** risulta statisticamente significativo
- si ha un \(R^{2}\) pari a 0.4623, poco maggiore rispetto al modello **fit_sA**, la complessità è aumentata di molto tuttavia il tradeoff non è vantaggioso

Scelta del modello

```
##                          R2_adj       AIC        BIC
## log(total_bill)       0.4573789 -11.77823  -1.286721
## sqrt(total_bill)      0.4594946  68.04409  78.535593
## size                  0.2211495 159.15589 176.641730
## sqrt(total_bill)+size 0.4623227  69.72032  90.703328
```

Si può osservare che il modello con \(R^{2}_{adj}\) più elevato è l'ultimo, tuttavia non va considerato il modello migliroe in quanto la complessità di tale modello è più elevata rispetto agli altri, ed in confronto al modello **sqrt(total_bill)** la variabilità spiegata non è drasticamente più elevata, inoltre bisogna ricordare che solo il coefficente **total_bill** è siginficativa in quel modello invece gli altri no.

N.B. L' **AIC** ed il **BIC** non possono essere usati per confrontare due modelli che non modellano le stesse variabiabili{red}

Per la nota precedente escludiamo la trasforamzione logaritmica per il confronto dei modelli con i test **AIC** e **BIC**, osservando i valori questi due test per i modelli restanti vediamo che i valori minori per entrambi i test corrispondono al modello **sqrt(total_bill)**.

Il modello con il logaritmo non viene considerato perchè ha un intercetta negativo e non significativo per questo motivo giungiamo alla conlcusione che il modello migliore deve essere uno di quelli con la trasformazione **sqrt**.

Quindi il modello scelto dopo tutte le considerazione precedenti è **sqrt(total_bill)**.