Progetto d'esame
Modelli Statistici

Cazzolato Kevin SM3201245

Anno: 2024

1 Descrizione del problema e descrizione del dataset

Si vuole studiare il dataset "Tips" ed in particolare è di interesse la variabile "tip". Si vuole dunque determinare un modello di regressione lineare, se possibile, adeguato al fine di stimare la variabile tip. Una volta importato il Dataset si usa la funzione head() per sapere quali variabili sono all'interno di esso. Dunque queste risultano essere:

```
  total_bill  tip    sex smoker day   time size
1      16.99 1.01 Female     No Sun Dinner    2
2      10.34 1.66   Male     No Sun Dinner    3
3      21.01 3.50   Male     No Sun Dinner    3
4      23.68 3.31   Male     No Sun Dinner    2
5      24.59 3.61 Female     No Sun Dinner    4
6      25.29 4.71   Male     No Sun Dinner    4
```

Si nota quindi che le variabili all'interno del Dataset sono:

- toal_bill: il totale del conto dell'osservazione i-esima
- tip: la mancia lasciata (obiettivo della nostra analisi)
- sex: sesso dei clienti
- smoker: se fumatore
- day: giorno indicato (lu, ma,..)
- time: se a pranzo o a cena
- size: qt di persone coinvolte nel tavolo

possiamo inoltre verificare la tipologia delle variabili attraverso la funzione str():

```
'data.frame':   244 obs. of  7 variables:
 $ total_bill: num  17 10.3 21 23.7 24.6 ...
 $ tip       : num  1.01 1.66 3.5 3.31 3.61 4.71 2 3.12 1.96 3.23 ...
 $ sex       : chr  "Female" "Male" "Male" "Male" ...
 $ smoker    : chr  "No" "No" "No" "No" ...
 $ day       : chr  "Sun" "Sun" "Sun" "Sun" ...
 $ time      : chr  "Dinner" "Dinner" "Dinner" "Dinner" ...
 $ size      : int  2 3 3 2 4 4 2 4 2 2 ...
```

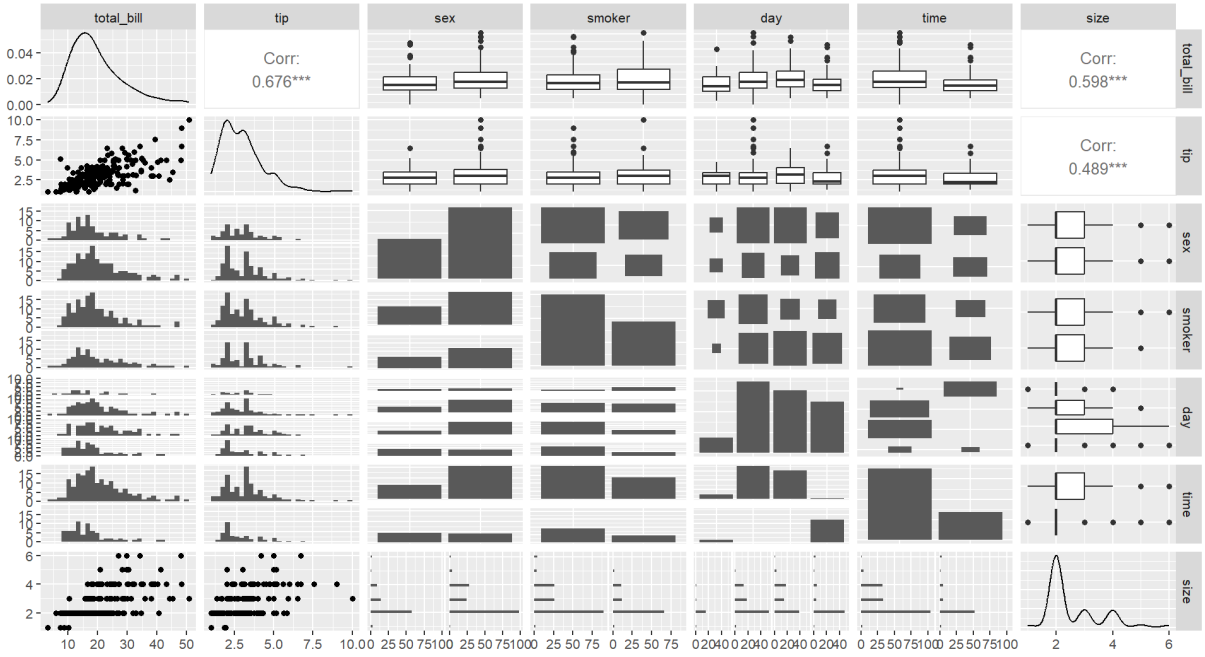
notiamo che il dataset è composto da 244 osservazioni totali, e che vi sono 3 variabili numeriche: total_bill, tip e size, mentre le restanti possono essere trasformate in fattori per semplicità e comodità, dunque si trasformano le variabili categoriali come segue:

```
data$sex = factor(data$sex, levels = c("Male", "Female"))
data$smoker = factor(data$smoker, levels = c("Yes", "No"))
data$day = factor(data$day, levels = c("Thur", "Fri", "Sat", "Sun"))
data$time = factor(data$time, levels = c("Lunch", "Dinner"))
```

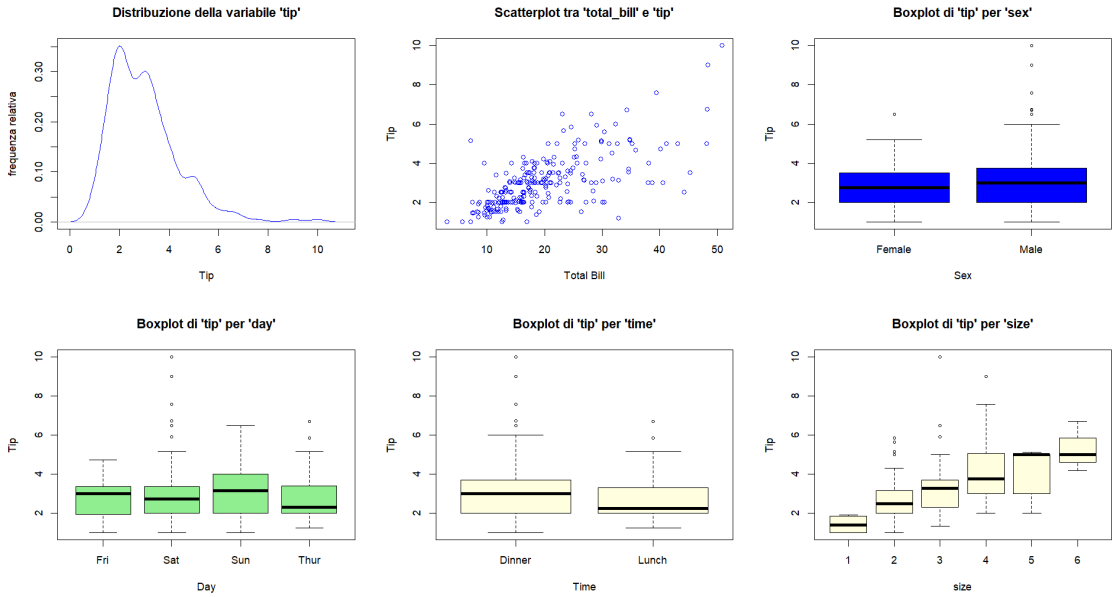
Attraverso la funzione any(is.na()) possiamo verificare se all'interno troviamo dei valori nulli da dover gestire, ma questi non sono presenti all'interno del dataset. Inoltre tutti i valori sono omogenei, dunque non abbiamo discrepanze di sintassi come "Si" e "si" o valori inattesi all'interno delle variabili numeriche (e.g. dei valori negativi o nulli). Si passa dunque ad un analisi esplorativa

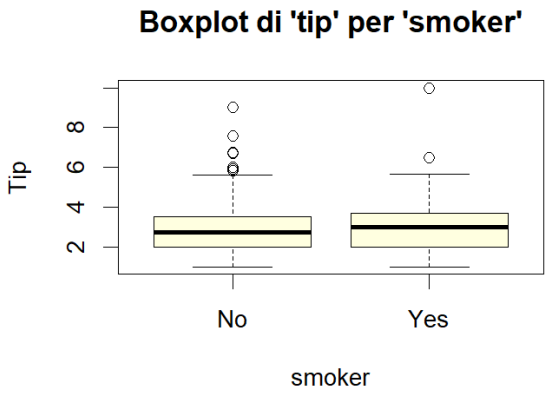
2 Analisi Esplorativa

Attraverso la libreria GGally e la funzione ggpairs si effettua una prima analisi esplorativa dei dati:



Si può notare che total_bill ha una forte correlazione con la variabile tip e size, quindi di probabile interesse per la nostra analisi, inoltre dato che l'obiettivo dell'analisi verte sulla variabile tip, possiamo concentrarci sin da subito sulle relazioni che intercorrono tra questa e le restanti variabili. Dunque viene analizzato il seguente grafico:





Possiamo dunque notare dal primo grafico che la variabile tip sembrerebbe seguire una distribuzione chi quadro ma questo non influenza la nostra analisi, in quanto l'ipotesi di gaussianità è richiesta sul modello $Y = \alpha + \beta x_i$ e sugli errori e non sulla variabile risposta. Inoltre dal secondo grafico, sembrerebbe che la variabile tip sia relazionata linearmente con la variabile Bill, anche se è presente una assimetria nella quantità dei dati (quindi in un futuro modello, se presa in considerazione la variabile total bill, sarebbe buona norma normalizzare i dati) possiamo anche notare una varianza eteroschedastica, in particolare all'aumentare della x la varianza sembra aumentare (un motivo in più per normalizzare). La variabile Sex non sembra essere influente per la nostra variabile riposta, in quanto le distribuzioni di Female e Male sembrano equivalersi, con una quantità di dati maggiore per Male. Nemmeno la variabile Day sembra essere significativa. Sembra esserci una leggera differenza tra la distribuzione Dinner e Lunch nella varibile Time, è necessario quindi eseguire una analisi più approfondita per verificare una possibile correlazione con la variabile Tip. Infine anche la variabile Size sembrerebbe inficiare sulla variabile Tip, in quanto all'aumentare del numero dei clienti al tavolo la mancia aumenta, anche se la crescita per le persone maggiori di 5 sembra decrescere (quindi potrebbe seguire una relazione logaritmica).

2.1 Analisi dipendenze con la variabile risposta

In questa sezione si vuole analizzare le possibili dipendenze e indipendenze intuite graficamente precedentemente con la variabile risposta.

2.1.1 variabili quantitative

Per effettuare tale analisi con le variabili quantitative possiamo usare la funzione cor() che indica la relazione lineare tra le variabili, quindi risulta essere:

```
            total_bill       tip      size
total_bill  1.0000000 0.6757341 0.5983151
tip         0.6757341 1.0000000 0.4892988
size        0.5983151 0.4892988 1.0000000
```

Tali risultati sono stati ottenuti con la formula:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \cdot \sum(Y_i - \bar{Y})^2}}$$

dove X_i e Y_i sono i valori delle osservazioni i-esime e $\bar{X}$ e $\bar{Y}$ sono le medie campionarie delle variabili X e Y, rispettivamente.
Dunque possiamo osservare che la variabile risposta è fortemente correlata (positivamente) con la variabile total bill e anche con la variabile size, anche se in maniera più contenuta. Inoltre la variabile

size è anch'essa fortemente correlata (positivamente) con la variabile total bill. Dunque potrebbe essere un buon modello quello in cui si considerano queste due variabili, data la forte relazione lineare.

2.1.2 variabili categoriali

Per le variabili categoriali possiamo usare il **test chi quadro** (lo eseguiamo solo sulle variabili categoriali perchè si assume che X, Y possono assumere solo un numero finito), il che consiste nel costruire una tabella di contingenza e verificare l'ipotesi nulla:

$$H_0 : p_{ij} = p_{i.} * p_{.j}$$

$\forall(i, j)$, dove p_{ij} sono le probabilità teoriche congiunte (in alternativa si possono usare gli stimatori di MV: frequenze empiriche relative congiunte), mentre $p_{i.}, p_{.j}$ sono le probabilità teoriche marginali. In R possiamo usare la funzione chisq.test() che ci fornisce la statistica χ^2, dove:

$$\chi^2 = \sum_{i=1}^{s} \sum_{j=1}^{t} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

i suoi gradi di libertà e il p-value. Rifiuteremo l'ipotesi nulla al di sotto di 0.05. Dunque otteniamo le seguenti:

```
        Pearson's Chi-squared test

data:  table(data$smoker, data$tip)
X-squared = 112.13, df = 122, p-value = 0.728
```

```
        Pearson's Chi-squared test

data:  table(data$day, data$tip)
X-squared = 329.53, df = 366, p-value = 0.9148
```

```
        Pearson's Chi-squared test

data:  table(data$sex, data$tip)
X-squared = 118.12, df = 122, p-value = 0.5826
```

```
        Pearson's Chi-squared test

data:  table(data$time, data$tip)
X-squared = 108.56, df = 122, p-value = 0.8026
```

Dunque non respingiamo l'ipotesi nulla di indipendenza tra queste variabili e la variabile risposta, inoltre avendo valori molto alti per il p-value possiamo avere un forte dubbio che siano indipendenti.

3 creazione dei modelli

Possiamo quindi passare alla creazione di un primo modello di regressione lineare: data la dimensione contenuta del dataset possiamo applicare un stepwise algorithm partendo da tutte le variabili all'interno del dataset. Dunque utilizzando il criterio AIC, definito come: $AIC = 2k - 2ln(L)$ (dove k=numero di parametri e L è il valore massimizzato della funzione di verosimiglianza del modello stimato), la funzione stepAIC() della libreria MASS ci restituirà il modello con l'AIC più basso, nel nostro caso il miglior modello risulta essere quello con le variabili total bill e size. Potremmo, però, pensare che

un altro modello valido possa essere quello in cui consideriamo la relazione tra total bill e size (che abbiamo visto avere una forte relazione lineare), dunque utilizzare il seguente modello

```
modello2=lm(tip~total_bill*size,data=data)
```

possiamo notare che il primo modello ha un AIC, seppur di poco, più basso rispetto al secondo (703 vs 705), dunque si cercano ulteriori conferme attraverso il criterio BIC, definito come $BIC = -2ln(L) + kln(n)$. Si ottengono, rispettivamente, i seguenti risultati: (717,723), dunque un ulteriore conferma del fatto che il primo modello potrebbe essere più adeguato. Per avere un ulteriore conferma possiamo usare il metodo delle k fold incrociate per stimare gli errori. Attraverso il seguente codice:

```
k_fold <- function(set, k, modelli, norm) {
  errori_complessivi = vector("list", length(modelli)) #lista di liste
  fold_size= nrow(set) %/% k #parte inf del modulo

  for (i in 1:k) {
    # Seleziona i dati di test e di addestramento
    test_indices = ((i - 1) * fold_size + 1):(i * fold_size)
    test_data = set[test_indices, ]
    training_data = set[-test_indices, ]

    if(norm==T){
    training_data = normalizzazione_min_max(training_data)
    test_data = normalizza_test_data(test_data, training_data)
    }

    for (j in seq_along(modelli)) {
      mod = training(modelli[[j]], training_data)
      prediction = predict(mod, newdata = test_data)
      errore_parziale = abs(test_data$tip - prediction)
      errori_complessivi[[j]] = c(errori_complessivi[[j]], errore_parziale)
    }
  }

  # Calcola le medie degli errori per ogni modello
  medie = sapply(errori_complessivi, mean)

  return(medie)
}

training <- function(f,dataset){
  return(lm(f,data=dataset))
}
```

Otteniamo che la media degli errori del primo modello è di 0.7515757 mentre per il secondo modello è di 0.7595539, per finire quindi analizziamo il Summary dei due modelli:

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.9279 -0.5547 -0.0852  0.5095  4.0425

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.668945   0.193609   3.455  0.00065 ***
total_bill  0.092713   0.009115  10.172  < 2e-16 ***
size        0.192598   0.085315   2.258  0.02487 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.014 on 241 degrees of freedom
Multiple R-squared:  0.4679,    Adjusted R-squared:  0.4635
F-statistic: 105.9 on 2 and 241 DF,  p-value: < 2.2e-16
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.8817 -0.5877 -0.0762  0.5070  4.0409

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.918271   0.509878   1.801 0.072964 .
total_bill      0.081881   0.022429   3.651 0.000321 ***
size            0.093269   0.206384   0.452 0.651734
total_bill:size 0.003941   0.007453   0.529 0.597491
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.015 on 240 degrees of freedom
Multiple R-squared:  0.4685,    Adjusted R-squared:  0.4618
F-statistic: 70.51 on 3 and 240 DF,  p-value: < 2.2e-16
```

Possiamo dunque notare che tutti i coefficienti, nel primo modello, sono significati, a differenza del secondo dove solo total bill è significativo. Possiamo notare, oltre ad una F statistic maggiore nel primo modello rispetto al secondo, che anche l'R quadro aggiustato, ovvero

$$R^2_{adj} = 1 - \left(\frac{(1 - R^2) \cdot (n - 1)}{n - k - 1} \right)$$

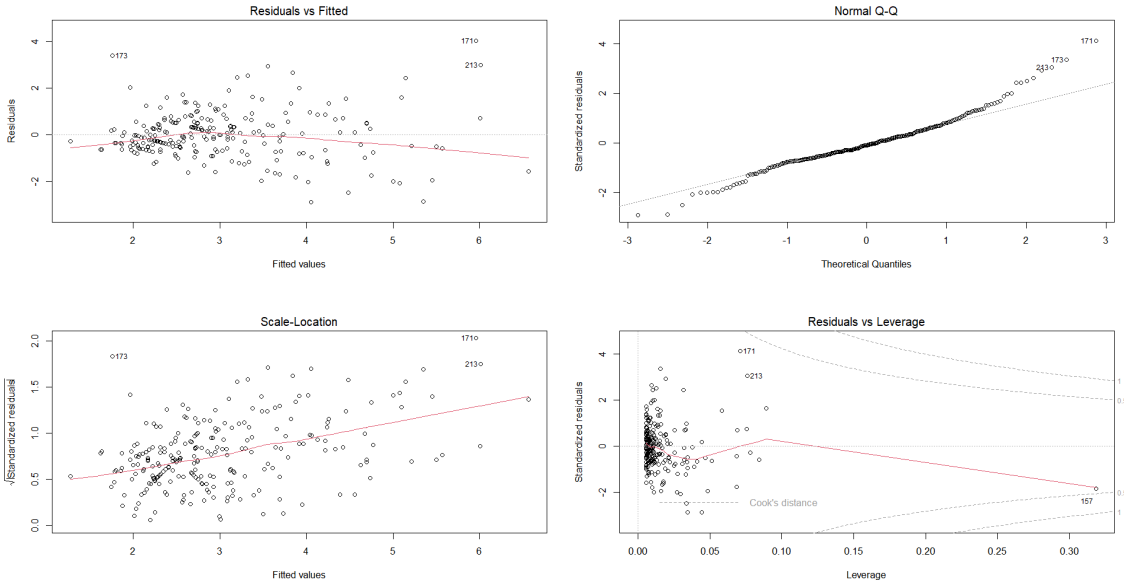
con

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

risulta essere maggiore nel primo modello. Dunque si decide di adottare il primo modello come modello di riferimento.

4 Analisi dei Residui

adottato il modello precedente, si vuole ora condurre un analisi dei residui $(y_i - \hat{y}_i)$ per verificare le ipotesi di linearità, gaussianità degli errori e l'omoschedasticità. Dunque possiamo plottare i seguenti grafici:



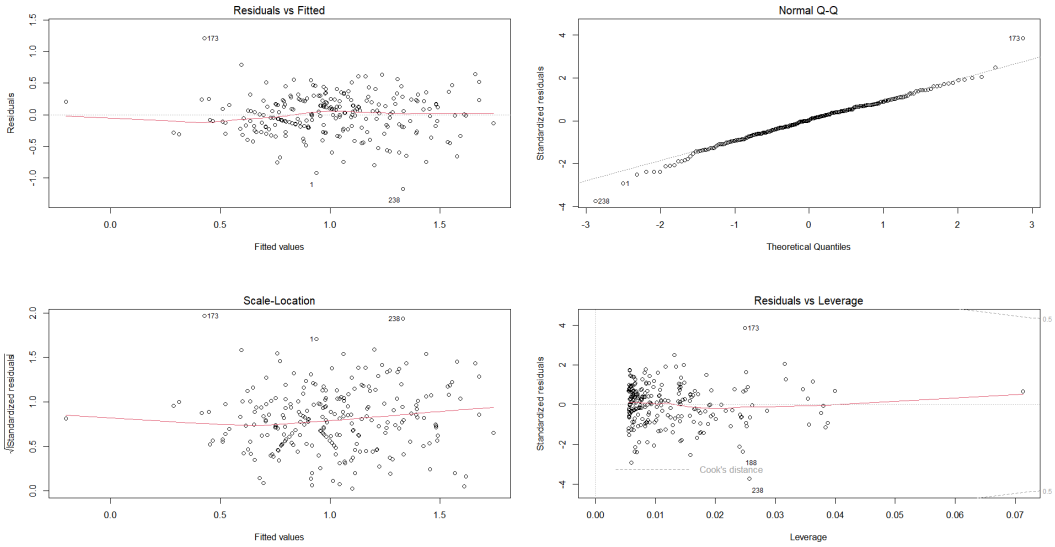
Possiamo notare l'assenza di pattern nel primo grafico, il che ci fa pensare che l'ipotesi di linearità sia rispettata. Possiamo inoltre notare che i residui, sempre dal primo grafico, sembrano essere eteroschedastici, infatti per x maggiori la varianza sembra aumentare, inoltre sembra esserci una forte assimetria delle osservazioni nell'asse delle y. Mentre sembra essere rispettata la simmetra con l'asse delle X. Dal secondo grafico notiamo che l'andamento è abbastanza conforme con quello gaussiano se non per le code che si discostano leggermente dalla retta. Infine, nel terzo grafico sembra esserci un pattern di linearità. Dunque tale modello non sembra essere pienamente efficiente così com'è, è quindi necessario attuare una normalizzazione dei dati per avere una varianza unica e aumentare l'ipotesi di gaussianità e linearità.

4.1 Normalizzazione

Si attuano varie normalizzazioni ai dati: Z_score, min_max ed in fine una trasformazione logaritmica dei dati. Otteniamo così 3 modelli allenati su 3 dataset differenti, possiamo ripetere i passi eseguiti precedentemente per la scelta del modello:

	media degli errori	AIC	BIC
Z score	4.28	545	559
Min Max	0.3847693	368	354
log	0.2484026	139	153

Dunque possiamo conludere, senza idagare oltre, che la normalizzazione logaritmica è quella più efficace per il nostro dataset. Infatti possiamo ora notare da questi residui



che la normalità sembra essere rispettata, anche la gaussianità e infine anche l'ipotesi di eteroschedasticità,

5 Rappresentazione del modello