Tipping Dataset

Analisi Esplorativa dei dati

Il dataset in esame comprende 244 osservazioni effettuate da un cameriere riguardanti la mancia ricevuta durante un determinato periodo di tempo. È composto da due variabili quantitative, "*tip*" e "*total_bill*", e da quattro variabili categoriali, "*sex*", "*smoker*", "*day*" e "*time*", oltre alla variabile "*size*", che può essere interpretata sia come quantitativa che come categoriale.

```
##   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
```

```
## '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       : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 2 2 2 2 2 ...
##  $ smoker    : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ day       : Factor w/ 4 levels "Fri","Sat","Sun",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ time      : Factor w/ 2 levels "Dinner","Lunch": 1 1 1 1 1 1 1 1 1 1 ...
##  $ size      : int  2 3 3 2 4 4 2 4 2 2 ...
```

```
## The mean of tip is:  2.998279
```

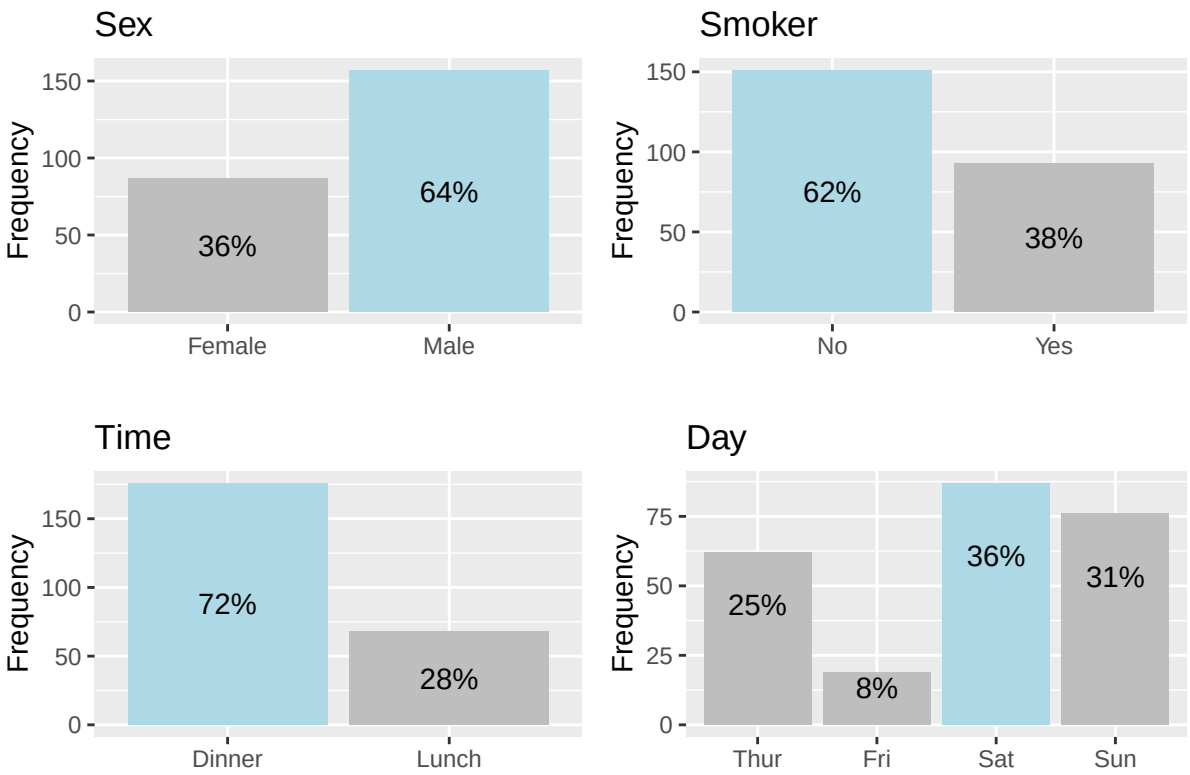
```
## The mean of total_bill is:  19.78594
```

```
##     total_bill   tip    sex smoker  day   time size
## 171      50.81 10.00   Male    Yes  Sat Dinner    3
## 213      48.33  9.00   Male     No  Sat Dinner    4
## 24       39.42  7.58   Male     No  Sat Dinner    4
## 60       48.27  6.73   Male     No  Sat Dinner    4
## 142      34.30  6.70   Male     No Thur  Lunch    6
## 215      28.17  6.50 Female    Yes  Sat Dinner    3
```

Analisi delle variabili categoriali

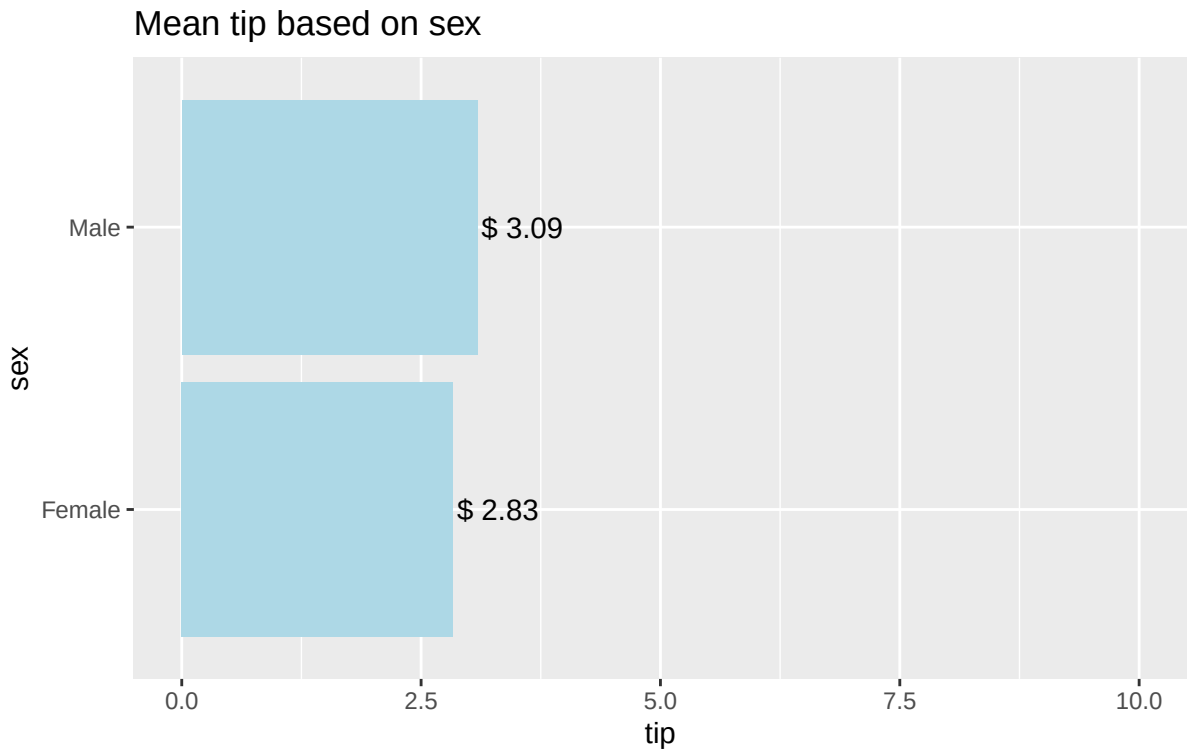
Cominciamo analizzando le variabili categoriali e la loro relazione con la variabile risposta "*tip*".

```
##      sex      smoker      day         time
##  Female: 87   No :151   Fri :19   Dinner:176
##  Male  :157   Yes: 93   Sat :87   Lunch : 68
##                         Sun :76
##                         Thur:62
```



Tip vs Sex

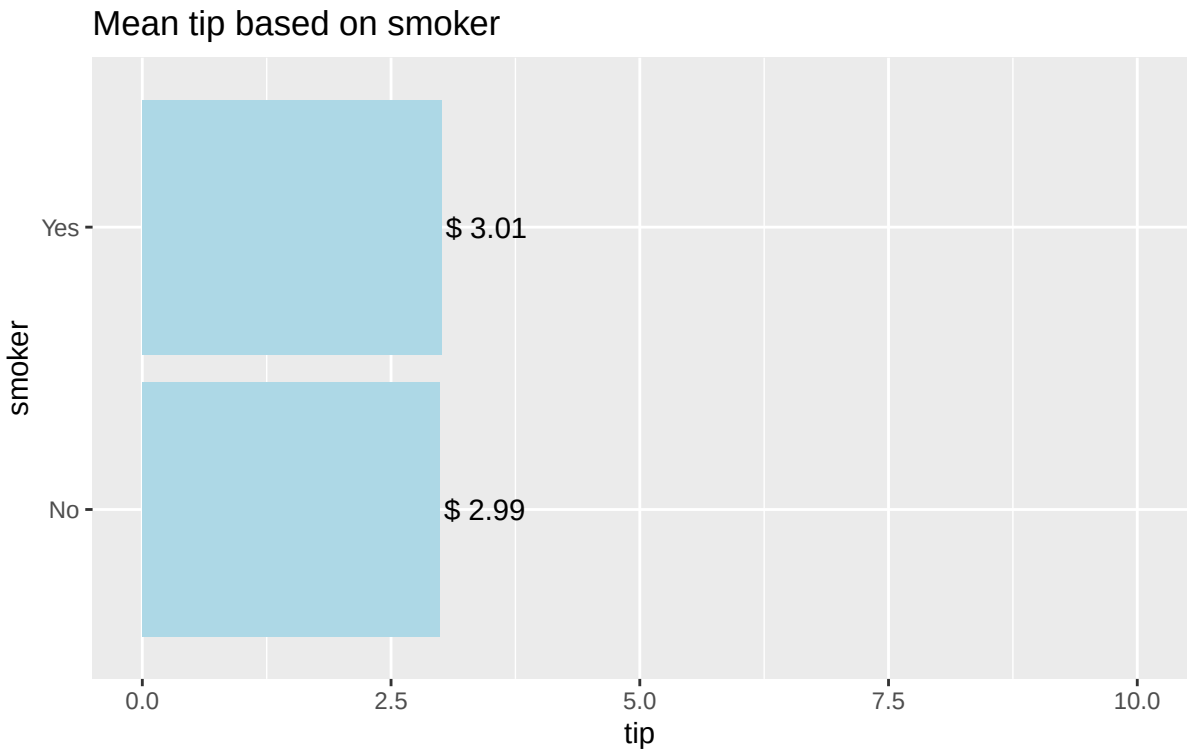
Ci si chiede se il sesso di chi lascia la mancia influisce sull'importo della mancia.




```
##
##  Welch Two Sample t-test
##
## data:  tip by sex
## t = -1.4895, df = 215.71, p-value = 0.1378
## alternative hypothesis: true difference in means between group Female and group Male is not equal to
## 95 percent confidence interval:
##  -0.5951448  0.0828057
## sample estimates:
## mean in group Female   mean in group Male
##             2.833448            3.089618
```

Il valore p è pari a 0.1378, quindi non c'è alcuna evidenza statistica per respingere l'ipotesi nulla che la differenza tra le medie dei due gruppi sia zero. In altre parole, non c'è alcuna differenza significativa tra la mancia media data dagli uomini e quella data dalle donne.

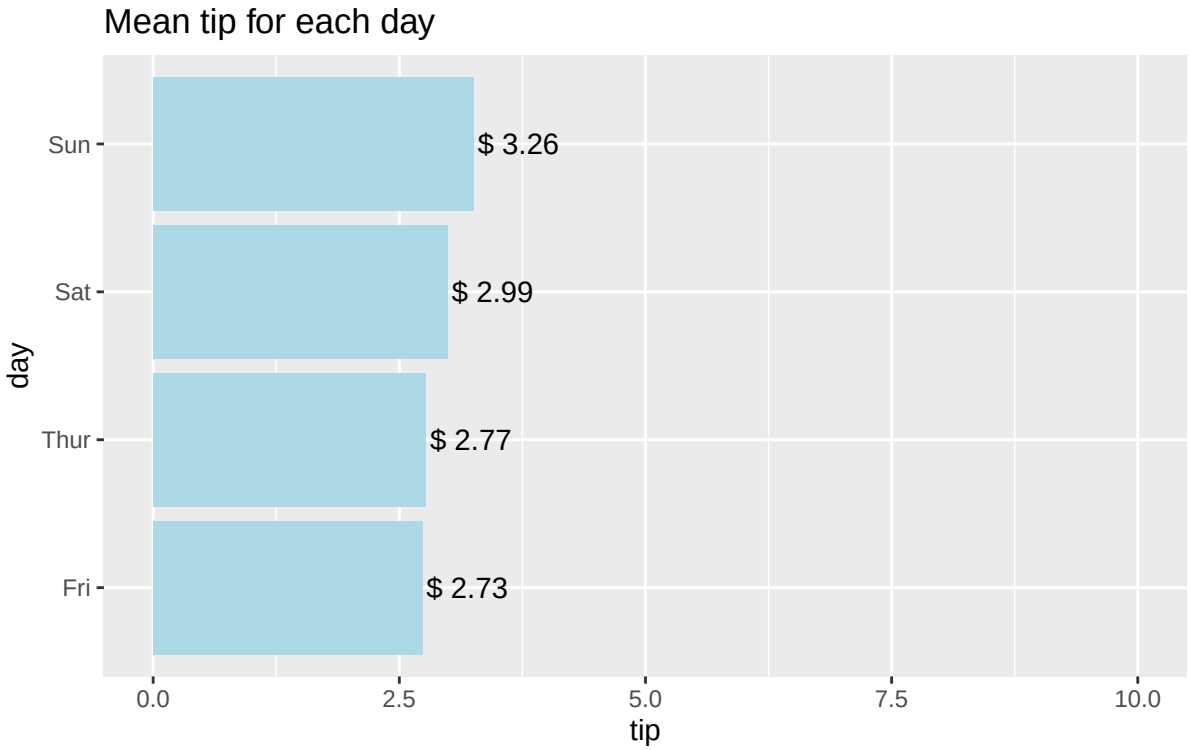
Tip vs Smoker




```
##
##  Welch Two Sample t-test
##
## data:  tip by smoker
## t = -0.091844, df = 192.26, p-value = 0.9269
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
## 95 percent confidence interval:
##  -0.3788291  0.3451184
## sample estimates:
##  mean in group No mean in group Yes
##          2.991854          3.008710
```

In questo caso le cose sono evidenti. Infatti anche intuitivamente si può supporre che l'essere fumatore o meno non debba avere alcuna influenza sull'importo della mancia.

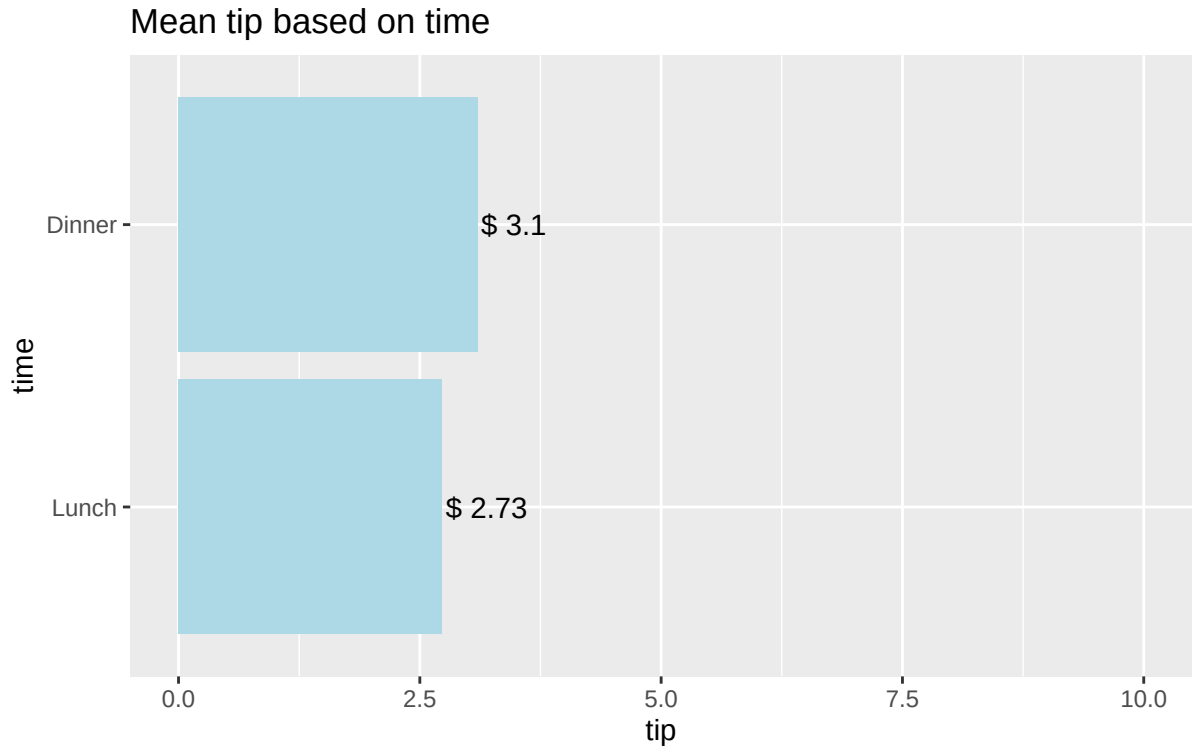
Tip vs Day




```
##              Df Sum Sq Mean Sq F value Pr(>F)
## day           3    9.5   3.175   1.672  0.174
## Residuals   240  455.7   1.899
```

Tramite l'analisi della varianza si nota che il valore p è 0.174, che è più grande di 0.05, quindi non si può rifiutare l'ipotesi nulla e non c'è evidenza sufficiente per dimostrare che ci sia una differenza significativa tra i valori medi di "tip" per i diversi giorni della settimana.

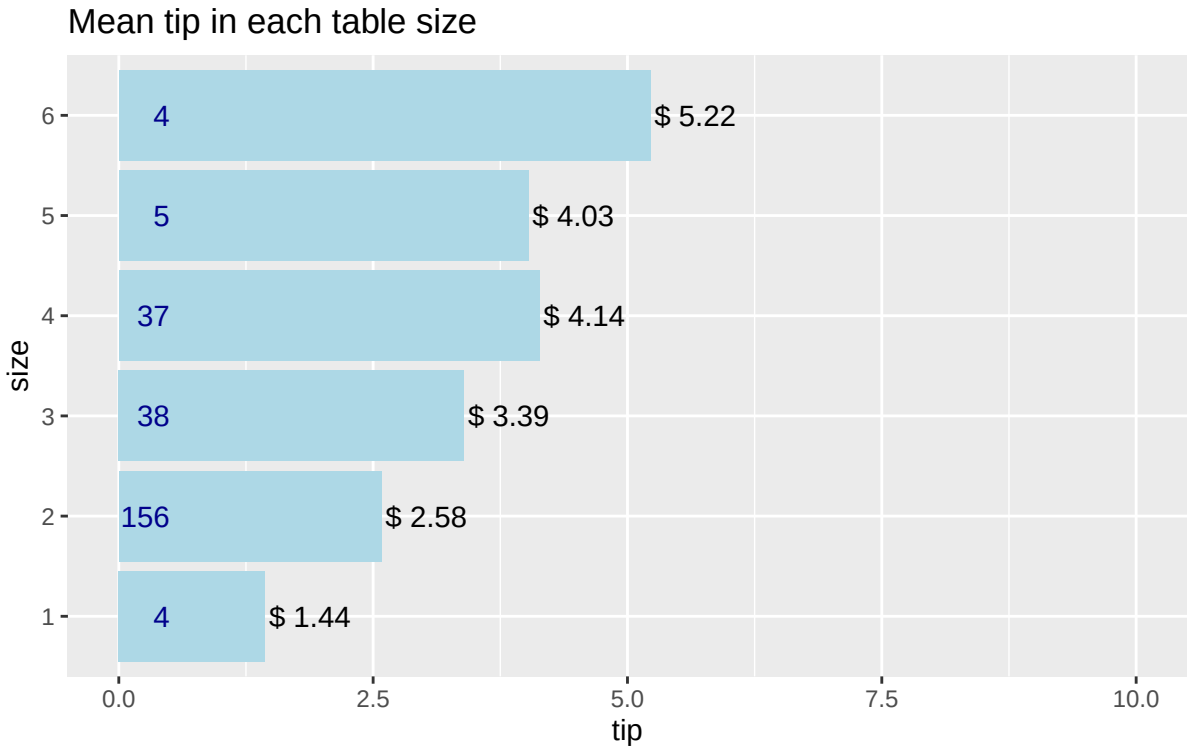
Tip vs Time




```
##
##  Welch Two Sample t-test
##
## data:  tip by time
## t = 2.0593, df = 144.07, p-value = 0.04126
## alternative hypothesis: true difference in means between group Dinner and group Lunch is not equal to
## 95 percent confidence interval:
##  0.01505364 0.73411080
## sample estimates:
## mean in group Dinner  mean in group Lunch
##             3.102670             2.728088
```

Sembra che la mancia media sia significativamente diversa a cena e a pranzo.

Tip vs Size



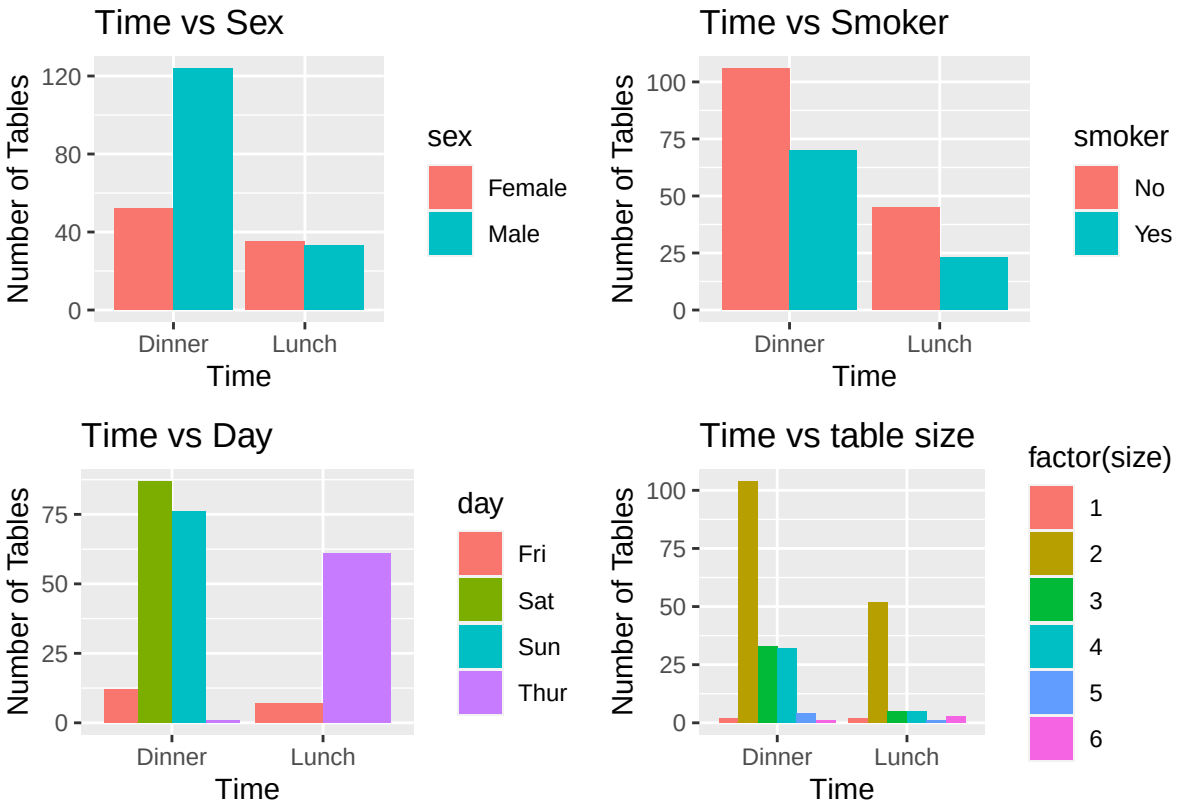

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## factor(size)   5  115.6  23.128   15.75 2.17e-13 ***
## Residuals    238  349.6   1.469
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Questi risultati mostrano che esiste una relazione statisticamente significativa tra la dimensione del gruppo (numero di persone al tavolo) e la mancia data al cameriere.

Sarà utile tenere conto di queste informazioni nella ricerca del modello di regressione lineare.

Relazioni fra le variabili categoriali

Analizziamo le relazioni che possono intercorrere tra la variabile "*time*" e le altre variabili categoriali:

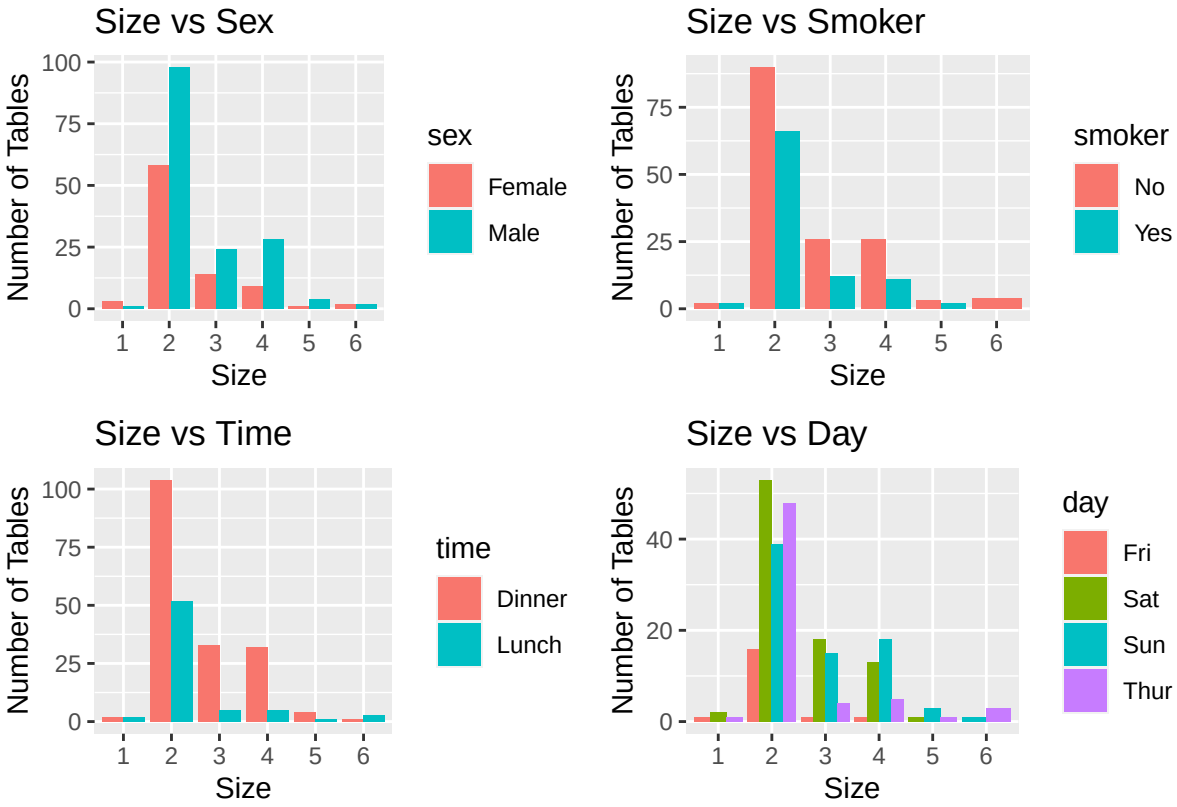



```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  df$time and df$sex
## X-squared = 9.3438, df = 1, p-value = 0.002237

##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  df$time and df$smoker
## X-squared = 0.50537, df = 1, p-value = 0.4771

##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  df$time and df$time
## X-squared = 239.05, df = 1, p-value < 2.2e-16

##
##  Pearson's Chi-squared test
##
## data:  df$time and df$day
## X-squared = 217.11, df = 3, p-value < 2.2e-16
```

- Sembra che gli uomini tendano a pagare di più a cena (ma questo non significa che lasceranno più mancia).

- Essere un fumatore non ha un impatto nella scelta tra pranzo o cena. Come si vede dalle proporzioni.
- Il giovedì si preferisce il pranzo, mentre nel weekend si preferisce la cena.
- A cena è più probabile trovare tavoli da 3 o più persone rispetto che a pranzo.

Vediamo adesso la relazione tra la il numero di persone al tavolo e le altre variabili categoriali:



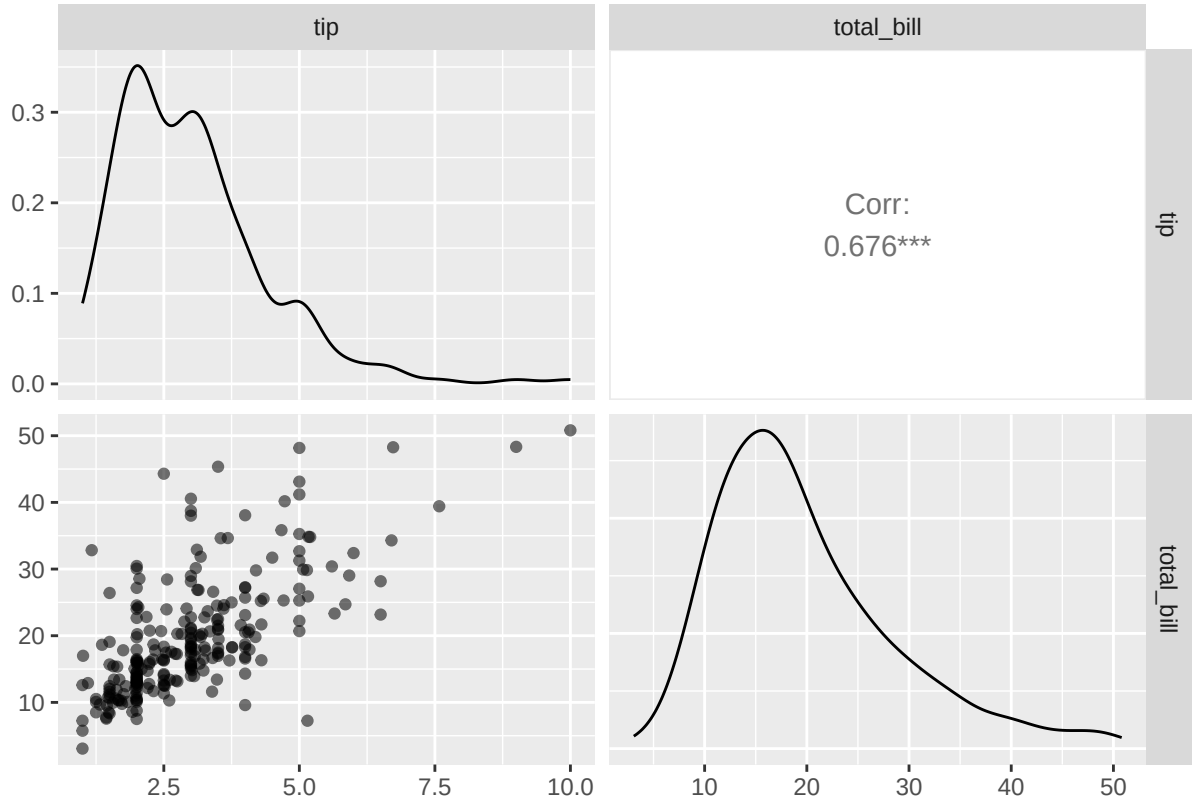

```
##
##  Pearson's Chi-squared test
##
## data:  df$size and df$sex
## X-squared = 5.8437, df = 5, p-value = 0.3217

##
##  Pearson's Chi-squared test
##
## data:  df$size and df$smoker
## X-squared = 5.6645, df = 5, p-value = 0.3402

##
##  Pearson's Chi-squared test
##
## data:  df$size and df$time
## X-squared = 15.75, df = 5, p-value = 0.007595

##
```

```
##  Pearson's Chi-squared test
##
## data:  df$size and df$day
## X-squared = 29.633, df = 15, p-value = 0.01332
```

- E' semplice intuire che il sesso e l'essere fumatori o meno non abbiano alcuna influenza sulla scelta del tavolo.
- Mentre sembra che si preferisca scegliere i tavoli da due persone a cena.
- Sembra esserci una relazione tra la grandezza del tavolo e il giorno della settimana.

Analisi delle variabili qualitative

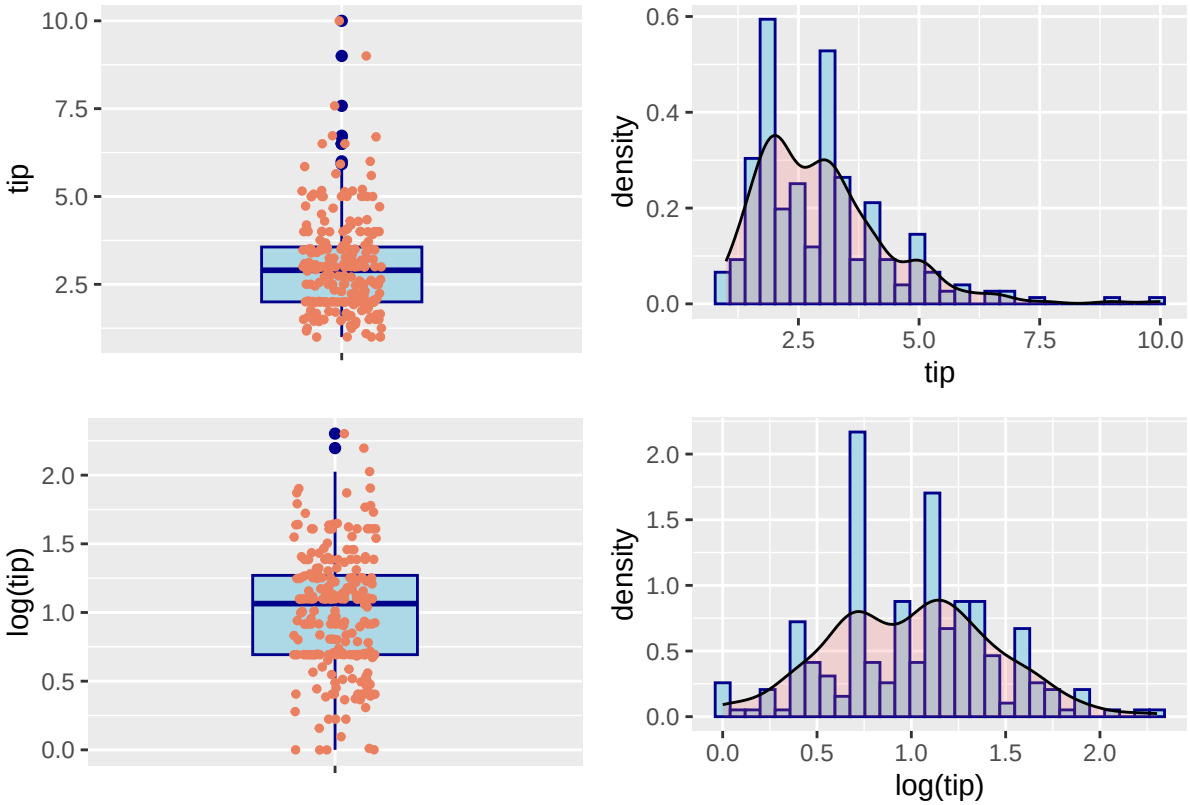
Continuiamo l'analisi osservando la distribuzione delle variabili quantitative del dataset e la loro correlazione:



E' facile intuire che la mancia debba essere correlata positivamente con il conto, ma anche con la grandezza del tavolo.

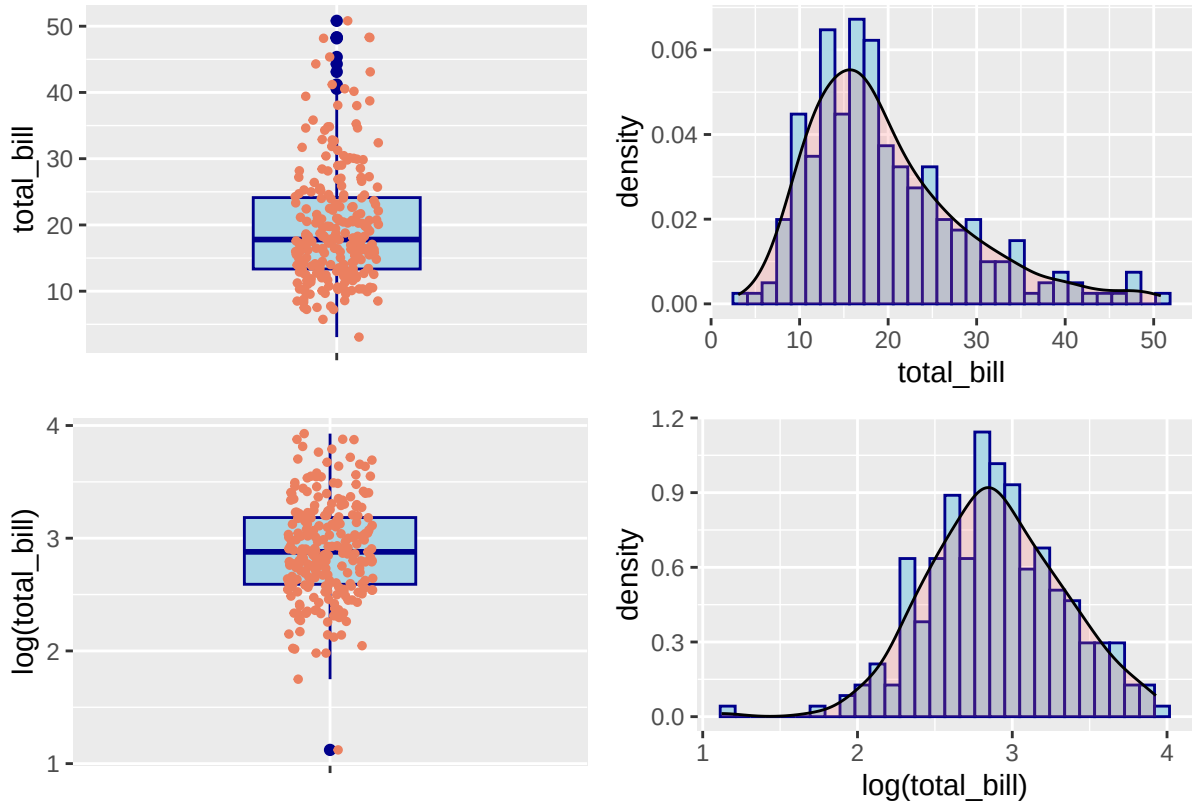
Distribuzione di tip e log(tip)

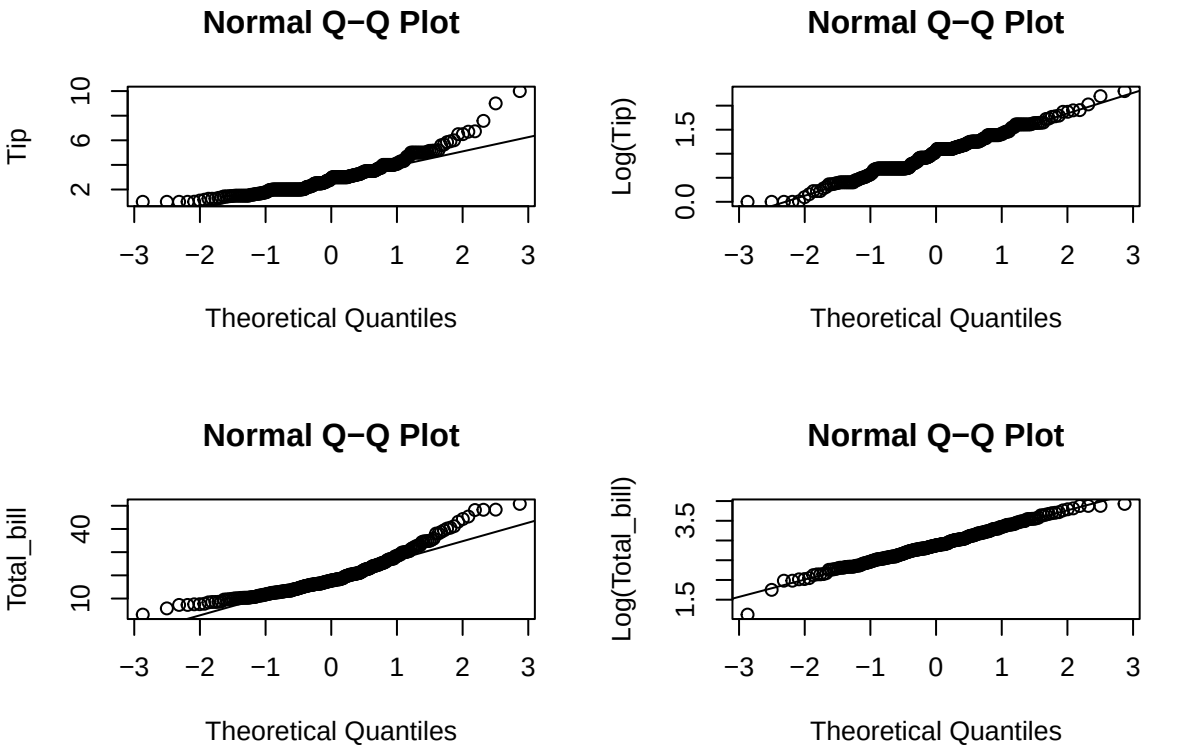
Analizziamo la distribuzione della variabile risposta e di una sua eventuale trasformazione logaritmica.



Distribuzione di total_bill e log(total_bill)

Analizziamo la distribuzione di "*total_bill*", la variabile qualitativa maggiormente correlata alla variabile risposta.



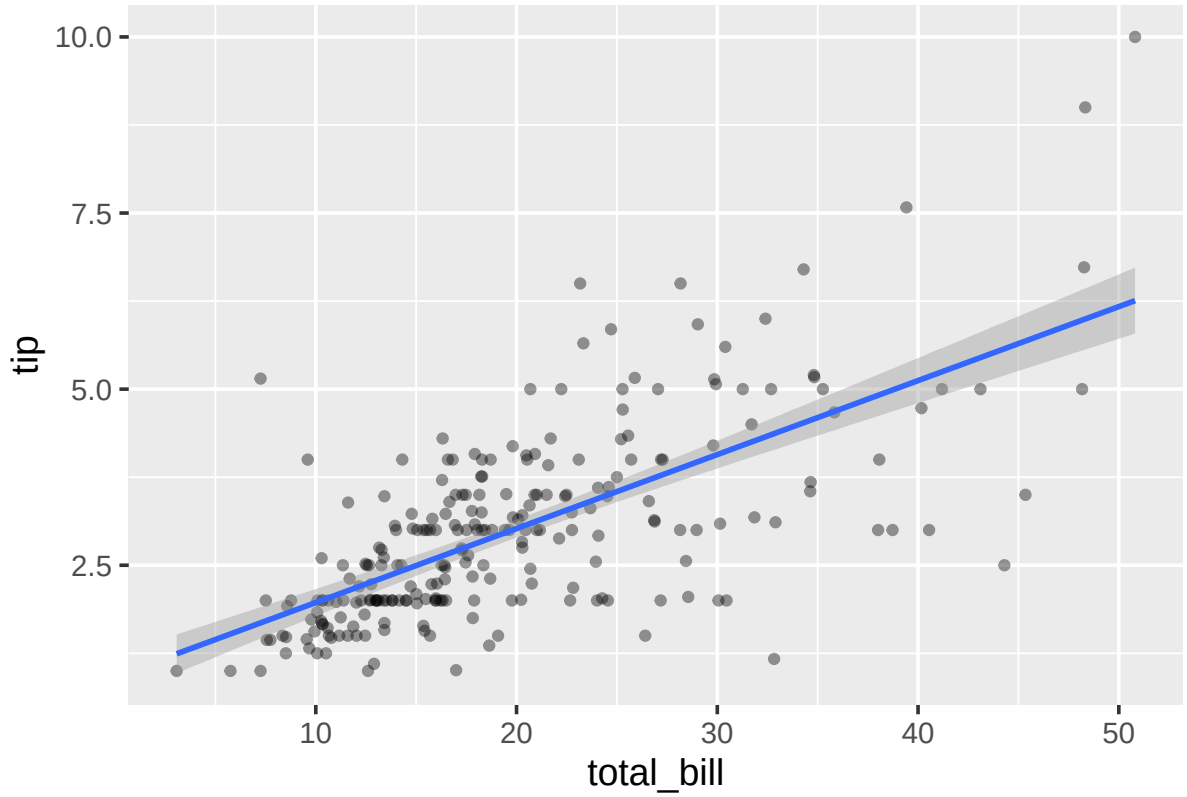


Sembra ragionevole prendere in considerazione una trasformazione logaritmica delle variabili "*tip*" e "*to-*

tal_bill", anche se ciò potrebbe comportare ad una perdita in termini di interpretabilità.

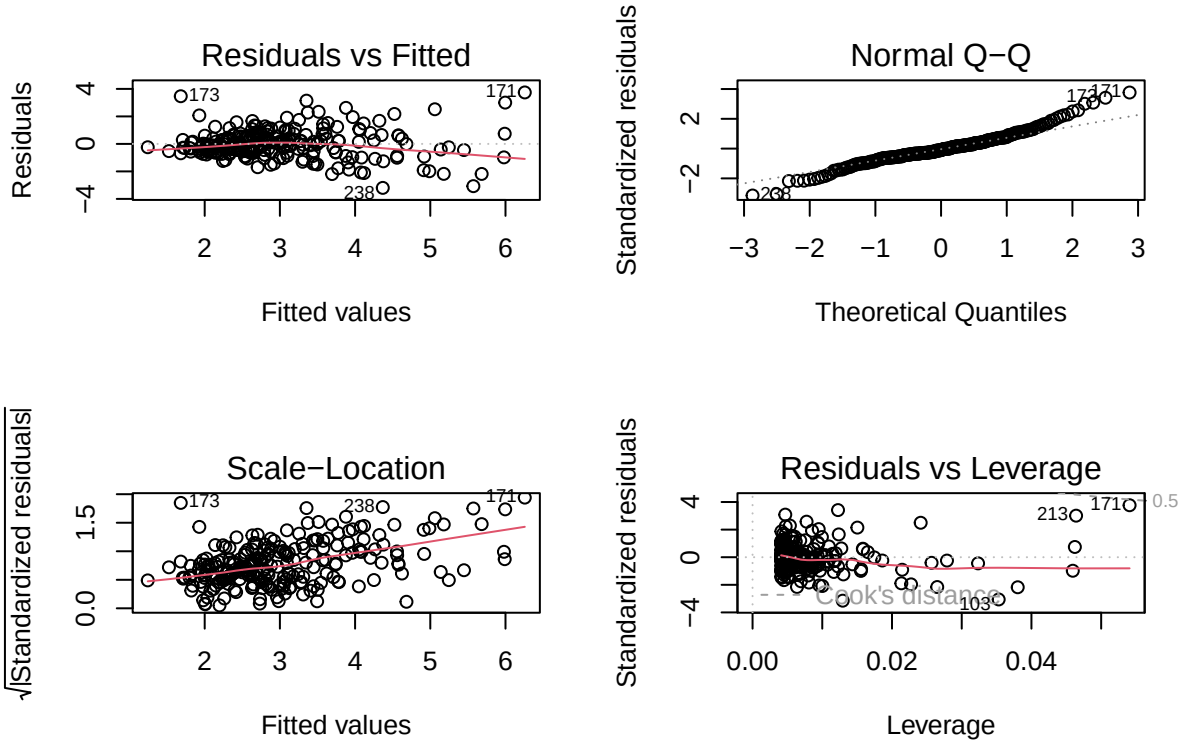
Modelli di regressione lineare multipla

Creiamo innanzitutto un semplice modello di regressione lineare tra mancia e conto da pagare. Data la correlazione ci si aspetta un coefficiente angolare positivo.




```
##
## Call:
## lm(formula = tip ~ total_bill, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1982 -0.5652 -0.0974  0.4863  3.7434
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.920270   0.159735   5.761 2.53e-08 ***
## total_bill  0.105025   0.007365  14.260  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.022 on 242 degrees of freedom
## Multiple R-squared:  0.4566, Adjusted R-squared:  0.4544
## F-statistic: 203.4 on 1 and 242 DF,  p-value: < 2.2e-16
```

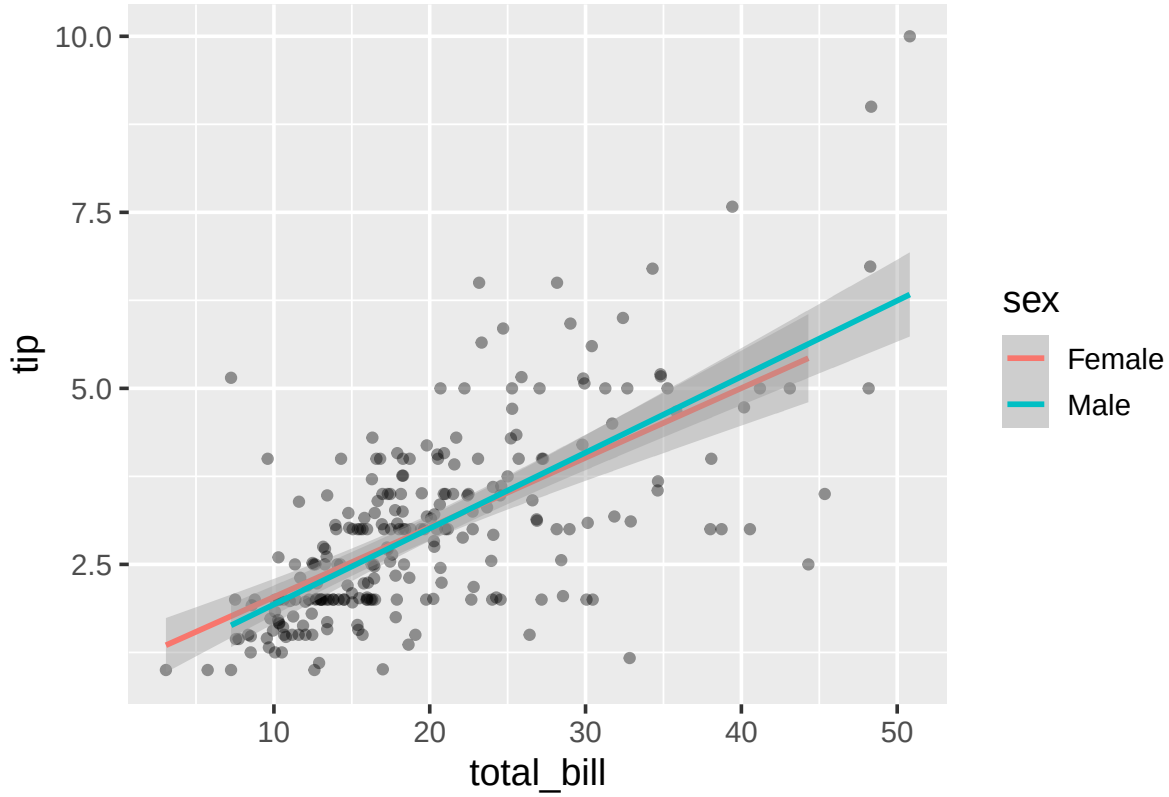
- Per ogni dollaro in più sul conto la mancia aumenta di 10 centesimi.
- L'errore standard è relativamente contenuto.
- Possiamo scartare l'ipotesi di nullità del coefficiente legato alla variabile "*total_bill*".
- Con solamente la variabile "*total_bill*" riusciamo a spiegare il 45% della variabilità di "*tip*".



Si nota come la linearità dei residui sia sufficientemente soddisfatta, mentre la normalità sembra perdersi nelle code della distribuzione dei residui. Inoltre sembra esserci una certa correlazione tra i residui e i valori teorici.

Esaminiamo il contributo che potrebbe portare l'aggiunta di un'altra variabile al modello, considerando le informazioni che abbiamo ottenuto dall'analisi univariata e bivariata delle variabili esplicative.

Effetto del sesso sulla relazione tra mancia e conto pagato

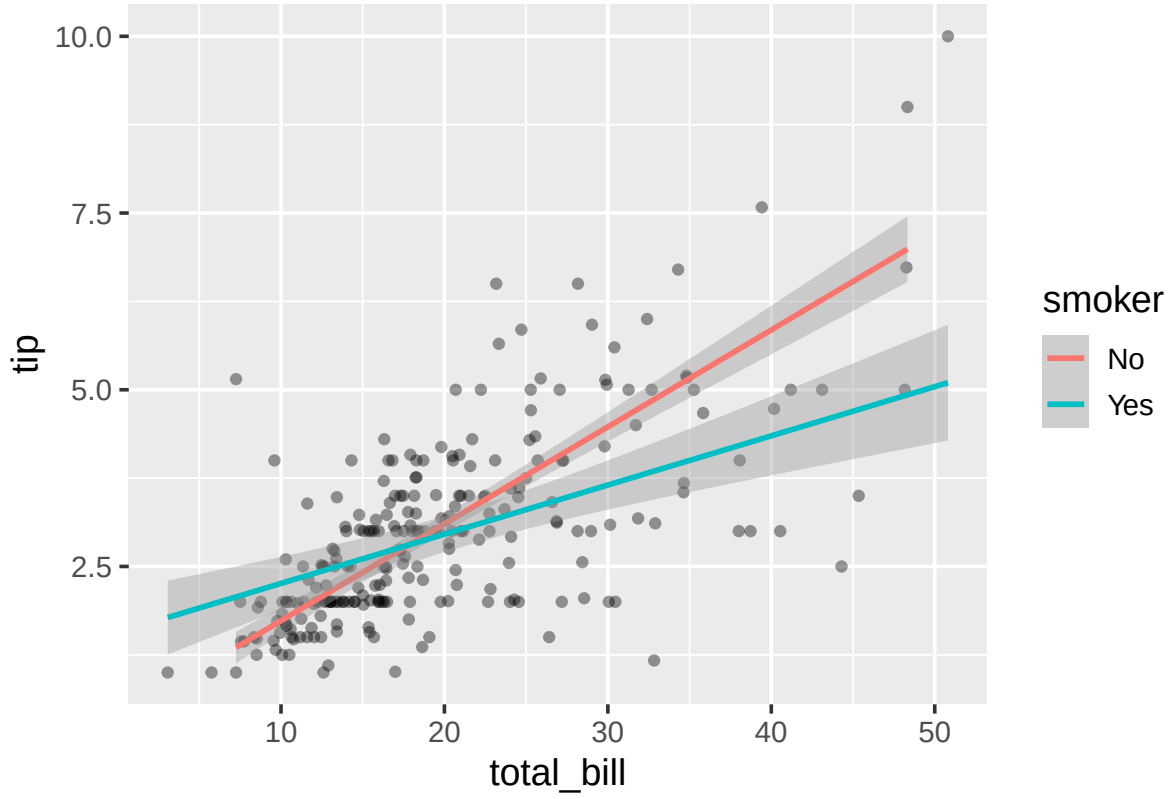


- IL grafico suggerisce che gli uomini tendano a pagare meno mancia delle donne fino a 20$.

```
##
## Call:
## lm(formula = tip ~ total_bill + sex, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1914 -0.5596 -0.0875  0.4845  3.7465
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.933278   0.173756   5.371 1.84e-07 ***
## total_bill   0.105232   0.007458  14.110  < 2e-16 ***
## sexMale     -0.026609   0.138334  -0.192    0.848
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.024 on 241 degrees of freedom
## Multiple R-squared:  0.4567, Adjusted R-squared:  0.4522
## F-statistic: 101.3 on 2 and 241 DF,  p-value: < 2.2e-16
```

- Il sesso non influisce sulla relazione tra mancia e conto pagato, e non aggiunge informazione al modello

Effetto dell'essere un fumatore sulla relazione tra mancia e conto pagato



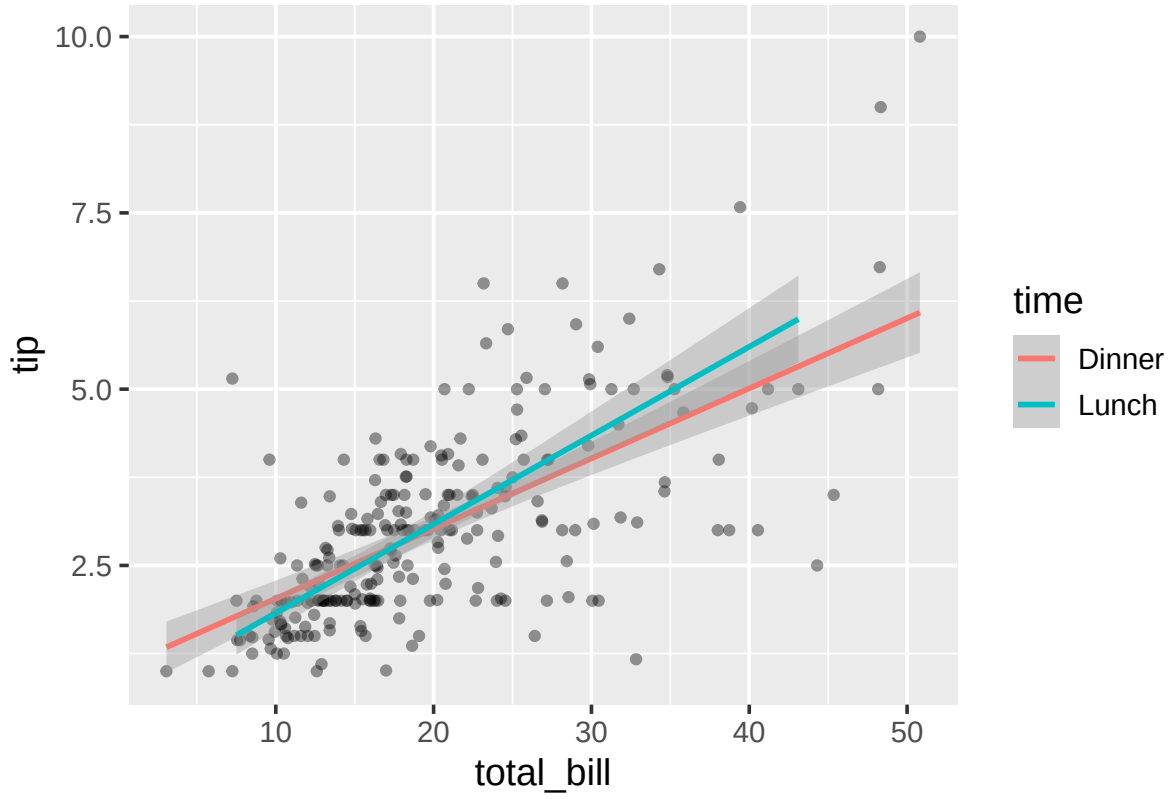
- Sembra che i fumatori paghino meno mancia per conti superiori a 20$.

```
##
## Call:
## lm(formula = tip ~ total_bill + smoker, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1152 -0.5884 -0.0812  0.4978  3.8139
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.963224   0.164354   5.861 1.51e-08 ***
## total_bill   0.105722   0.007389  14.309  < 2e-16 ***
## smokerYes   -0.148924   0.135159  -1.102    0.272
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.022 on 241 degrees of freedom
## Multiple R-squared:  0.4593, Adjusted R-squared:  0.4549
## F-statistic: 102.4 on 2 and 241 DF,  p-value: < 2.2e-16
```

- Il coefficiente di regressione per i fumatori è di -0.148924, il che significa che se una persona è un fumatore ci si aspetta che la mancia sia inferiore di 0.148924 rispetto a un non un fumatore. Tuttavia,

questo coefficiente ha un P-value relativamente elevato di 0.272, il che suggerisce che questa relazione non sia significativa e possa essere dovuta al caso.

Effetto di time sulla relazione tra mancia e conto pagato

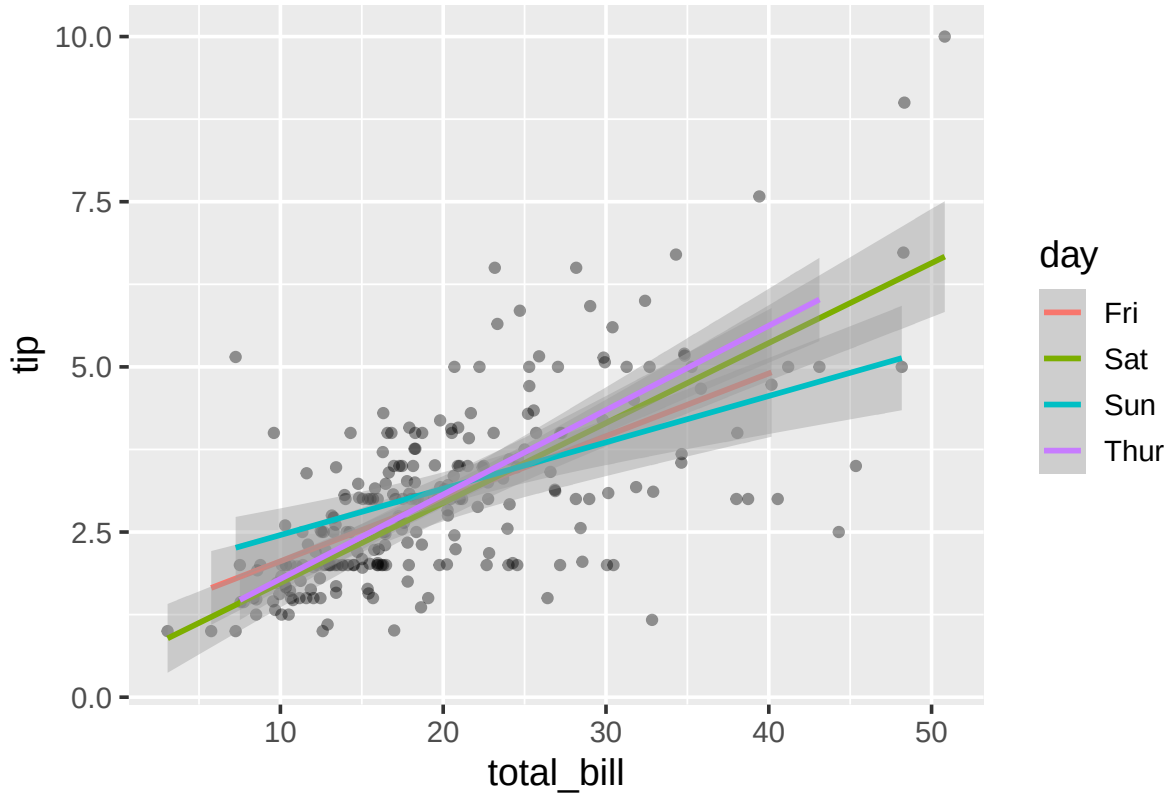


- Sembra che all'aumentare del conto a pranzo si lasci più mancia

```
##
## Call:
## lm(formula = tip ~ total_bill + time, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1972 -0.5689 -0.0954  0.4884  3.7434
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.917167   0.174167   5.266 3.09e-07 ***
## total_bill  0.105087   0.007507  13.999  < 2e-16 ***
## timeLunch   0.006723   0.148751   0.045    0.964
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.024 on 241 degrees of freedom
## Multiple R-squared:  0.4566, Adjusted R-squared:  0.4521
## F-statistic: 101.3 on 2 and 241 DF,  p-value: < 2.2e-16
```

- Ma anche questo effetto non sembra essere significativo.

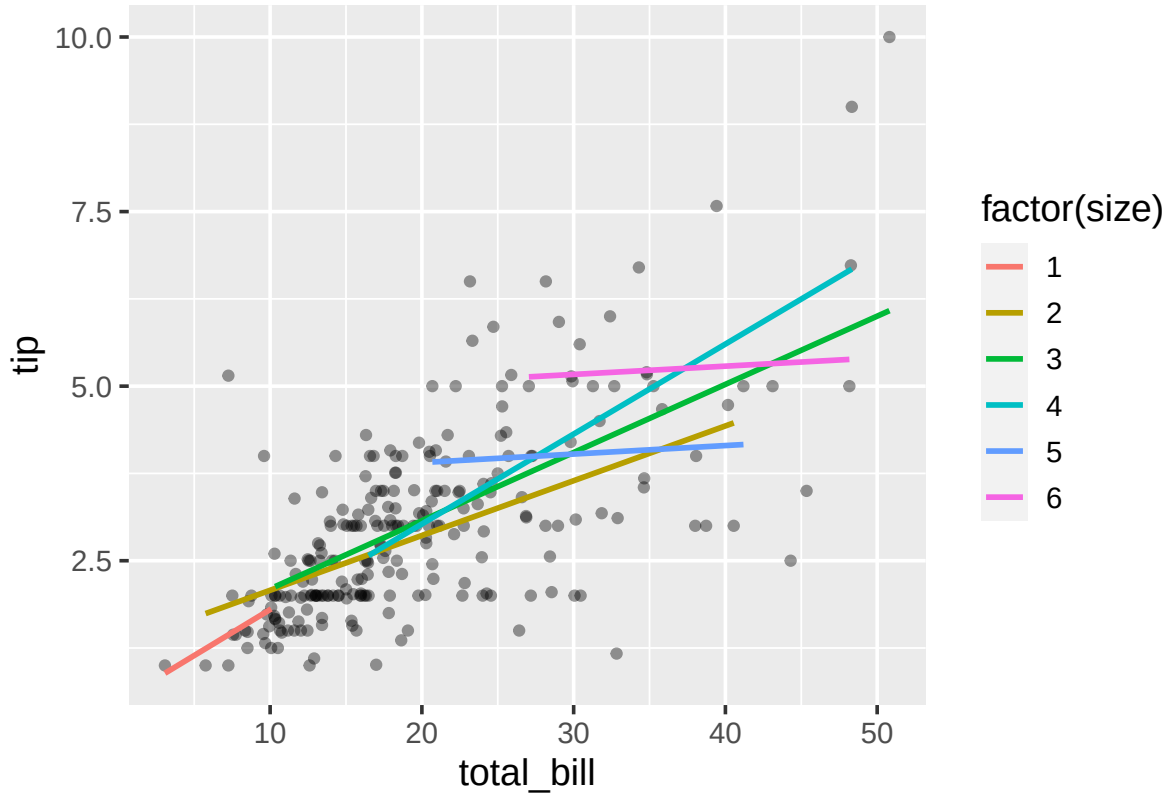
Effetto del giorno sulla relazione tra mancia e conto pagato




```
##
## Call:
## lm(formula = tip ~ total_bill + day, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1199 -0.5368 -0.0907  0.5054  3.8281
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.939433   0.268482   3.499 0.000557 ***
## total_bill   0.104673   0.007522  13.915  < 2e-16 ***
## daySat      -0.085986   0.261067  -0.329 0.742169
## daySun       0.074654   0.265183   0.282 0.778557
## dayThur     -0.018884   0.269150  -0.070 0.944125
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.026 on 239 degrees of freedom
## Multiple R-squared:  0.4589, Adjusted R-squared:  0.4498
## F-statistic: 50.67 on 4 and 239 DF,  p-value: < 2.2e-16
```

- Il giorno della settimana non ha un effetto statisticamente significativo.

Effetto delle dimensioni del tavolo sulla relazione tra mancia e conto pagato

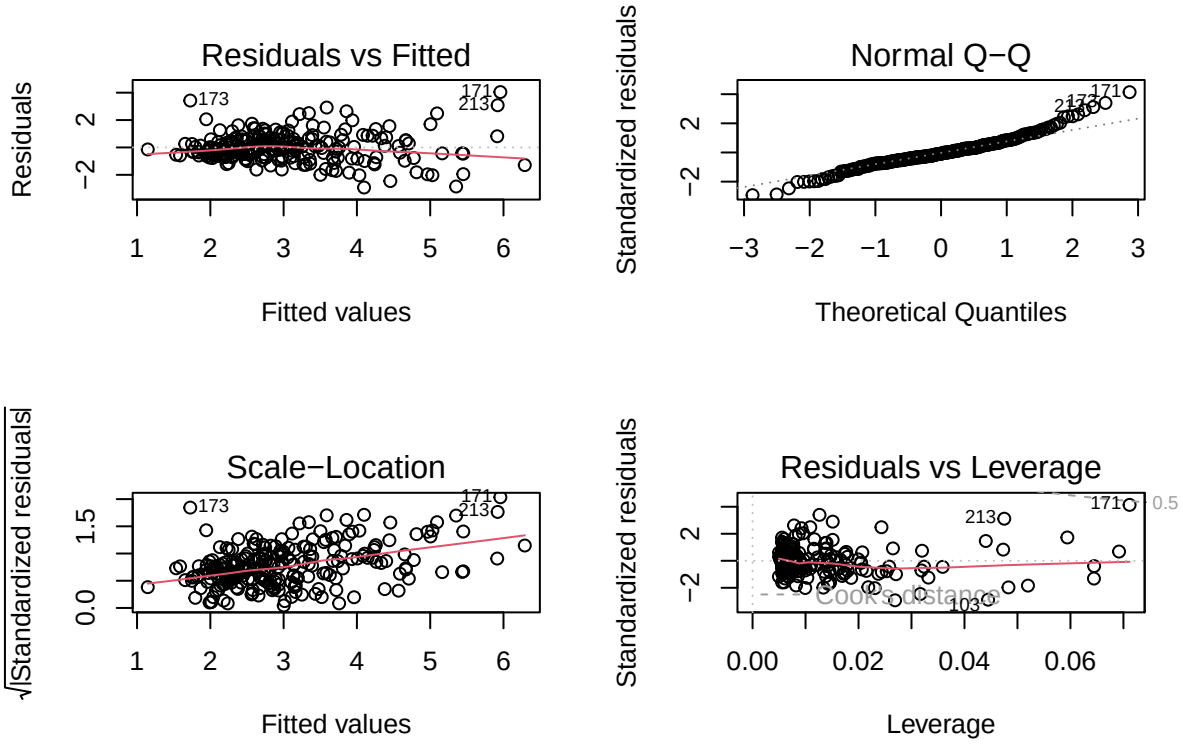


- I tavoli da 5 e 6 persone tendono a lasciare meno mancia all'aumentare dello scontrino
- I tavoli che lasciano più mancia in media sono quelli da 3 e 4 persone.

```
##
## Call:
## lm(formula = tip ~ total_bill + size, data = df)
##
## 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
```

- per ogni aumento di 1 dollaro sul conto, la mancia aumenterà di 9 centesimi. E per ogni persona in più al tavolo, la mancia aumenterà di 19 centesimi a parità di scontrino.

- Si può concludere che il numero di persone al tavolo è statisticamente significativo e che il 46,79% della varianza della mancia può essere spiegata da questo modello.



Considerando le informazioni dedotte dal dataset si ritiene che le variabili esplicative "*total_bill*" e "*size*" siano le uniche che diano un contributo significativo nel prevedere la mancia.

AIC

Vediamo se si possono migliorare i risultati con un approccio automatico per la ricerca delle variabili più significative. Si decide di considerare inizialmente un modello che include tutte le variabili.

```
fitA = lm( tip ~ ., data = df)
summary(fitA)
```

```
##
## Call:
## lm(formula = tip ~ ., data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8475 -0.5729 -0.1026  0.4756  4.1076
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.803817   0.352702   2.279   0.0236 *
## total_bill   0.094487   0.009601   9.841   <2e-16 ***
```

```
## sexMale     -0.032441   0.141612  -0.229   0.8190
## smokerYes   -0.086408   0.146587  -0.589   0.5561
## daySat      -0.121458   0.309742  -0.392   0.6953
## daySun      -0.025481   0.321298  -0.079   0.9369
## dayThur     -0.162259   0.393405  -0.412   0.6804
## timeLunch    0.068129   0.444617   0.153   0.8783
## size         0.175992   0.089528   1.966   0.0505 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.024 on 235 degrees of freedom
## Multiple R-squared:  0.4701, Adjusted R-squared:  0.452
## F-statistic: 26.06 on 8 and 235 DF,  p-value: < 2.2e-16
```

```
require(MASS)
```

```
## Caricamento del pacchetto richiesto: MASS
```

```
fitB = stepAIC(fitA, direction = "both")
```

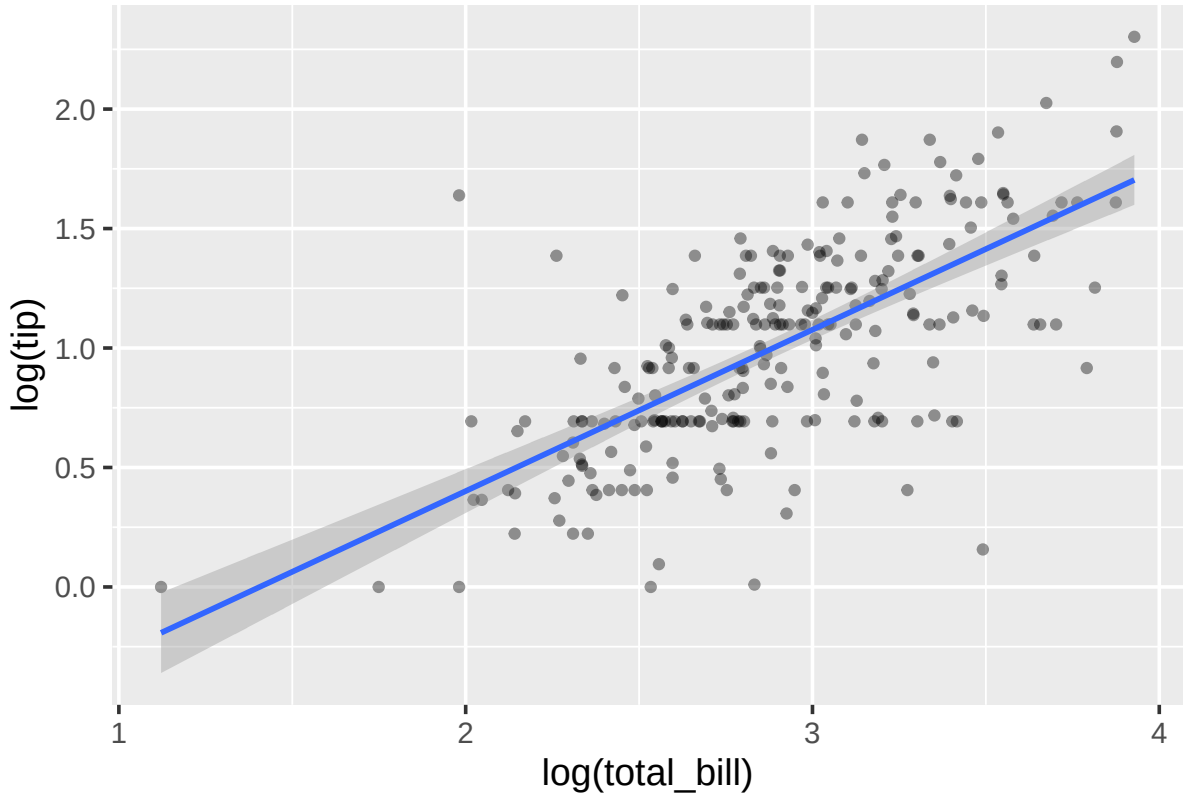
```
## Start:  AIC=20.51
## tip ~ total_bill + sex + smoker + day + time + size
##
##              Df Sum of Sq    RSS     AIC
## - day         3     0.609 247.14  15.116
## - time        1     0.025 246.55  18.538
## - sex         1     0.055 246.58  18.568
## - smoker      1     0.365 246.89  18.874
## <none>                    246.53  20.513
## - size        1     4.054 250.58  22.493
## - total_bill  1   101.595 348.12 102.713
##
## Step:  AIC=15.12
## tip ~ total_bill + sex + smoker + time + size
##
##              Df Sum of Sq    RSS    AIC
## - time        1     0.001 247.14 13.117
## - sex         1     0.042 247.18 13.157
## - smoker      1     0.380 247.52 13.490
## <none>                    247.14 15.116
## - size        1     4.341 251.48 17.365
## + day         3     0.609 246.53 20.513
## - total_bill  1   101.726 348.86 97.232
##
## Step:  AIC=13.12
## tip ~ total_bill + sex + smoker + size
##
##              Df Sum of Sq    RSS    AIC
## - sex         1     0.041 247.18 11.157
## - smoker      1     0.379 247.52 11.491
## <none>                    247.14 13.117
## + time        1     0.001 247.14 15.116
## - size        1     4.342 251.48 15.366
```

```
## + day         3     0.586 246.55 18.538
## - total_bill  1   103.327 350.46 96.350
##
## Step:  AIC=11.16
## tip ~ total_bill + smoker + size
##
##              Df Sum of Sq    RSS    AIC
## - smoker      1     0.376 247.55  9.528
## <none>                    247.18 11.157
## + sex         1     0.041 247.14 13.117
## + time        1     0.000 247.18 13.157
## - size        1     4.344 251.52 13.408
## + day         3     0.571 246.61 16.592
## - total_bill  1   104.263 351.44 95.029
##
## Step:  AIC=9.53
## tip ~ total_bill + size
##
##              Df Sum of Sq    RSS    AIC
## <none>                    247.55  9.528
## + smoker      1     0.376 247.18 11.157
## + sex         1     0.038 247.52 11.491
## + time        1     0.001 247.55 11.527
## - size        1     5.235 252.79 12.634
## + day         3     0.594 246.96 14.942
## - total_bill  1   106.281 353.83 94.685

##
## Call:
## lm(formula = tip ~ total_bill + size, data = df)
##
## 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
```

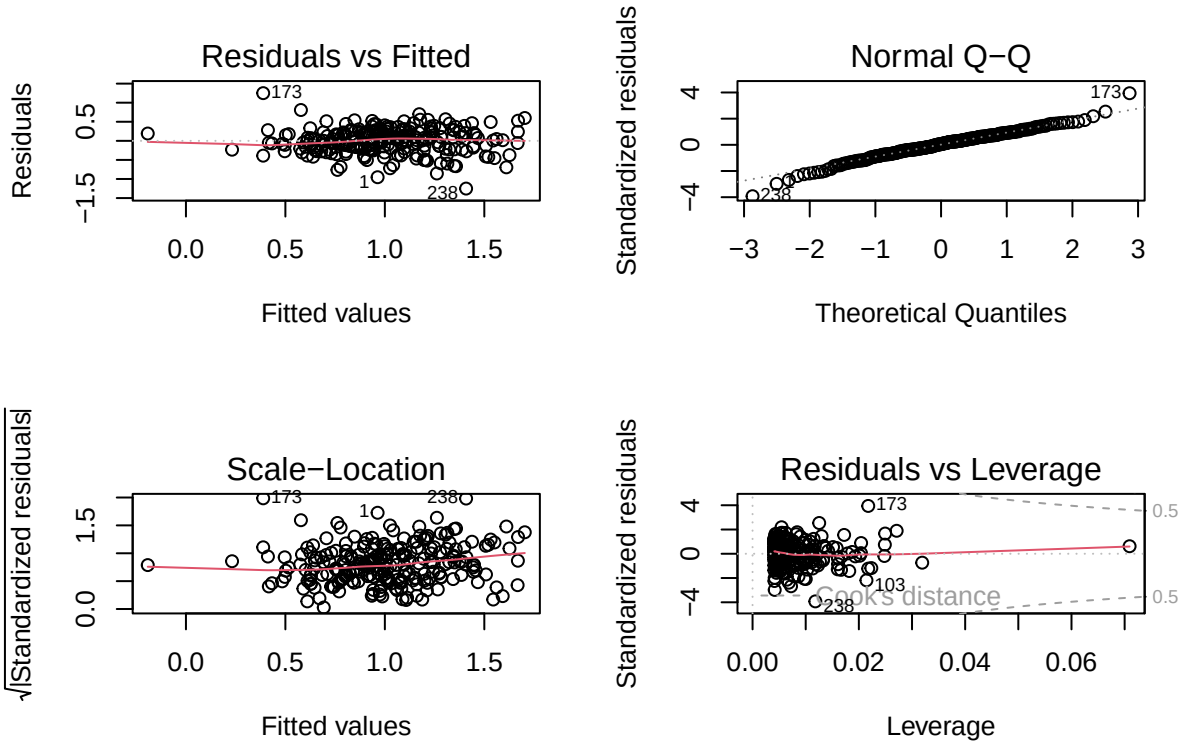
Si è giunti alla stessa conclusione del modello dedotto in precedenza.

Trasformazione logaritmica

Proviamo ad applicare una trasformazione logaritmica:

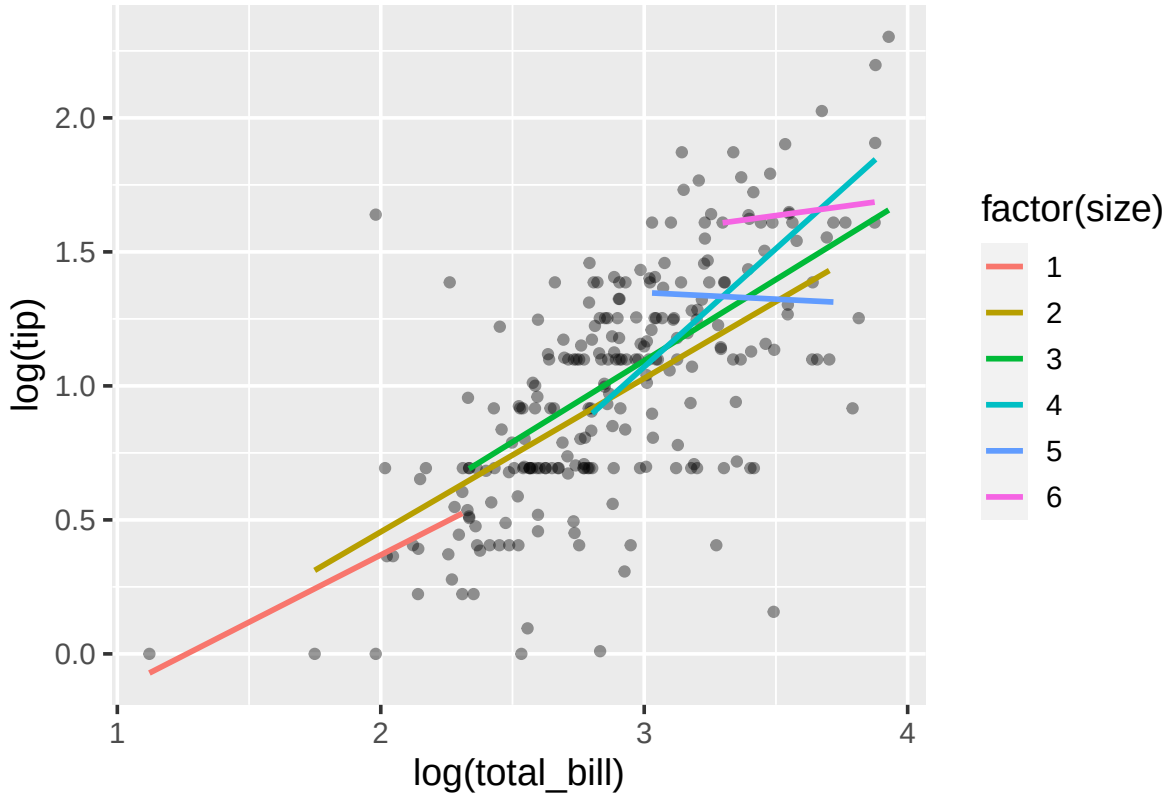



```
##
## Call:
## lm(formula = log(tip) ~ log(total_bill), data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.25145 -0.19246  0.01746  0.20360  1.25058
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -0.94948    0.13701   -6.93 3.78e-11 ***
## log(total_bill)  0.67537    0.04687   14.41  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3206 on 242 degrees of freedom
## Multiple R-squared:  0.4618, Adjusted R-squared:  0.4596
## F-statistic: 207.7 on 1 and 242 DF,  p-value: < 2.2e-16
```

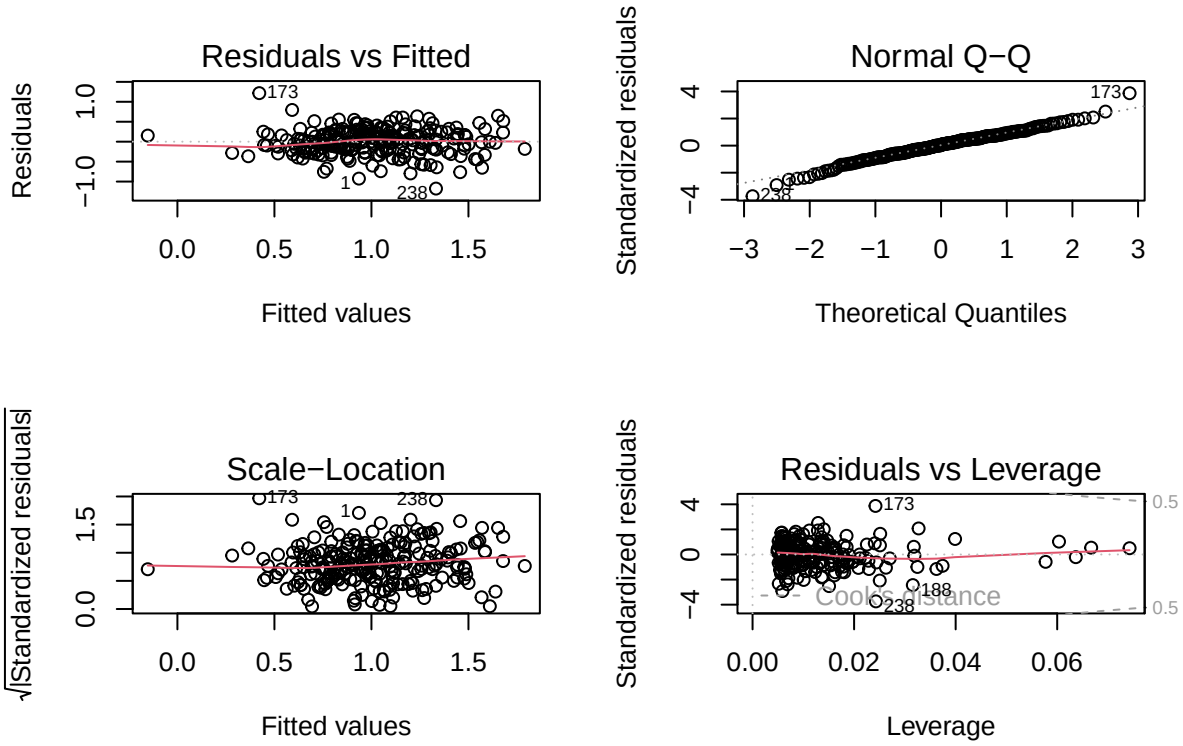


- La percentuale di variabilità spiegata è leggermente migliorata
- L'analisi dei residui è soddisfacente

Analizziamo anche la relazione con "*size*" includendola nel modello:




```
##
## Call:
## lm(formula = log(tip) ~ log(total_bill) + size, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.17575 -0.18745  0.01035  0.21012  1.21706
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -0.88589    0.13926  -6.362 9.96e-10 ***
## log(total_bill)  0.60305    0.05761  10.468  < 2e-16 ***
## size             0.05659    0.02658   2.129   0.0343 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3183 on 241 degrees of freedom
## Multiple R-squared:  0.4717, Adjusted R-squared:  0.4674
## F-statistic: 107.6 on 2 and 241 DF,  p-value: < 2.2e-16
```



- In conclusione abbiamo migliorato leggermente il modello in termini di variabilità spiegata e analisi dei residui, ma abbiamo perso in interpretabilità pratica dei coefficienti del modello.

AIC

```
fitlogA = lm( log_tip ~ ., data = dflog)
summary(fitlogA)
```

```
##
## Call:
## lm(formula = log_tip ~ ., data = dflog)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.13425 -0.19344  0.00314  0.19068  1.20235
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.851236   0.168959  -5.038 9.38e-07 ***
## log_total_bill  0.613450   0.060091  10.209  < 2e-16 ***
## sexMale        -0.025278   0.044343  -0.570    0.569
## smokerYes      -0.007530   0.045537  -0.165    0.869
## daySat         -0.070371   0.096903  -0.726    0.468
## daySun          0.001543   0.100506   0.015    0.988
```

```
## dayThur        -0.070385   0.123159  -0.571    0.568
## timeLunch       0.031746   0.139156   0.228    0.820
## size            0.051950   0.027599   1.882    0.061 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3204 on 235 degrees of freedom
## Multiple R-squared:  0.478,  Adjusted R-squared:  0.4602
## F-statistic:  26.9 on 8 and 235 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = log_tip ~ log_total_bill + size, data = dflog)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.17575 -0.18745  0.01035  0.21012  1.21706
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.88589    0.13926  -6.362 9.96e-10 ***
## log_total_bill  0.60305    0.05761  10.468  < 2e-16 ***
## size            0.05659    0.02658   2.129   0.0343 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3183 on 241 degrees of freedom
## Multiple R-squared:  0.4717, Adjusted R-squared:  0.4674
## F-statistic: 107.6 on 2 and 241 DF,  p-value: < 2.2e-16
```

E se considerassi la mancia percentuale?

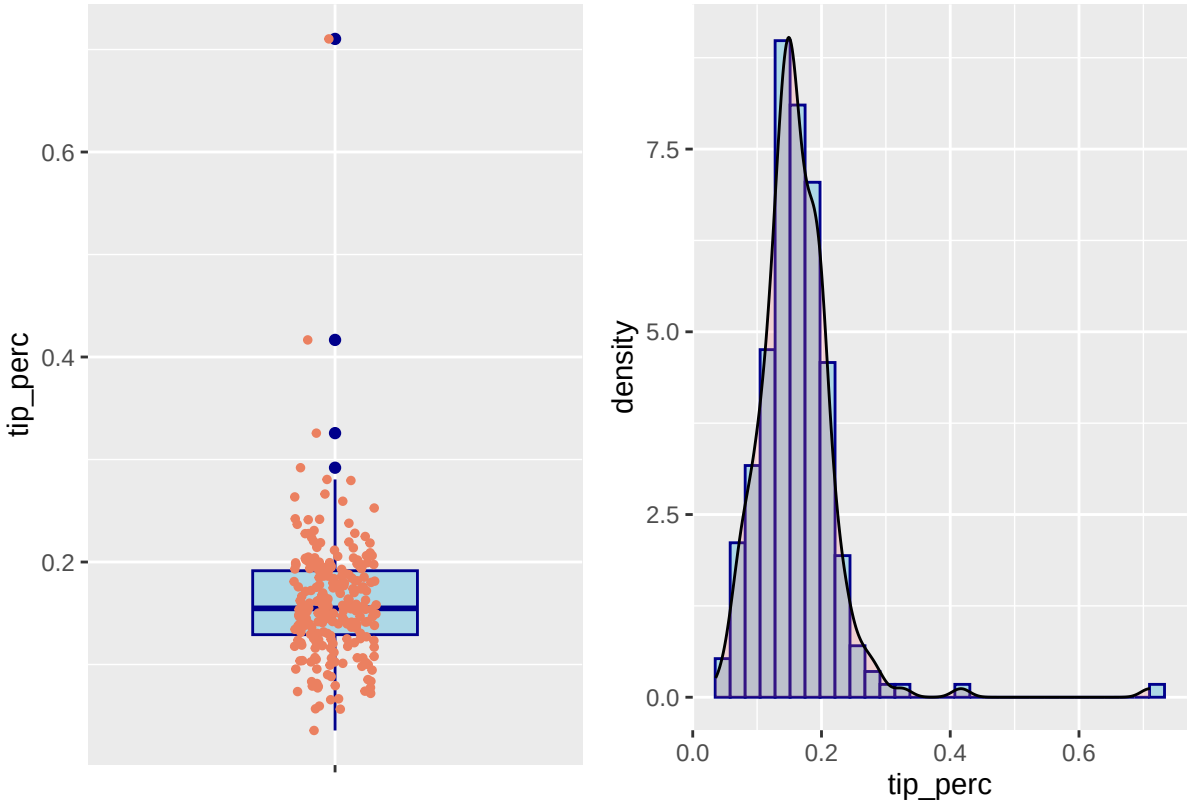
Ha senso chiedersi come varia la mancia in termini percentuali rispetto al conto pagato, e se in questo caso si riscontra una tendenza diversa rispetto al modello precedente.

```
##   total_bill    sex smoker day   time size   tip_perc
## 1      16.99 Female     No Sun Dinner    2 0.05944673
## 2      10.34   Male     No Sun Dinner    3 0.16054159
## 3      21.01   Male     No Sun Dinner    3 0.16658734
## 4      23.68   Male     No Sun Dinner    2 0.13978041
## 5      24.59 Female     No Sun Dinner    4 0.14680765
## 6      25.29   Male     No Sun Dinner    4 0.18623962
```

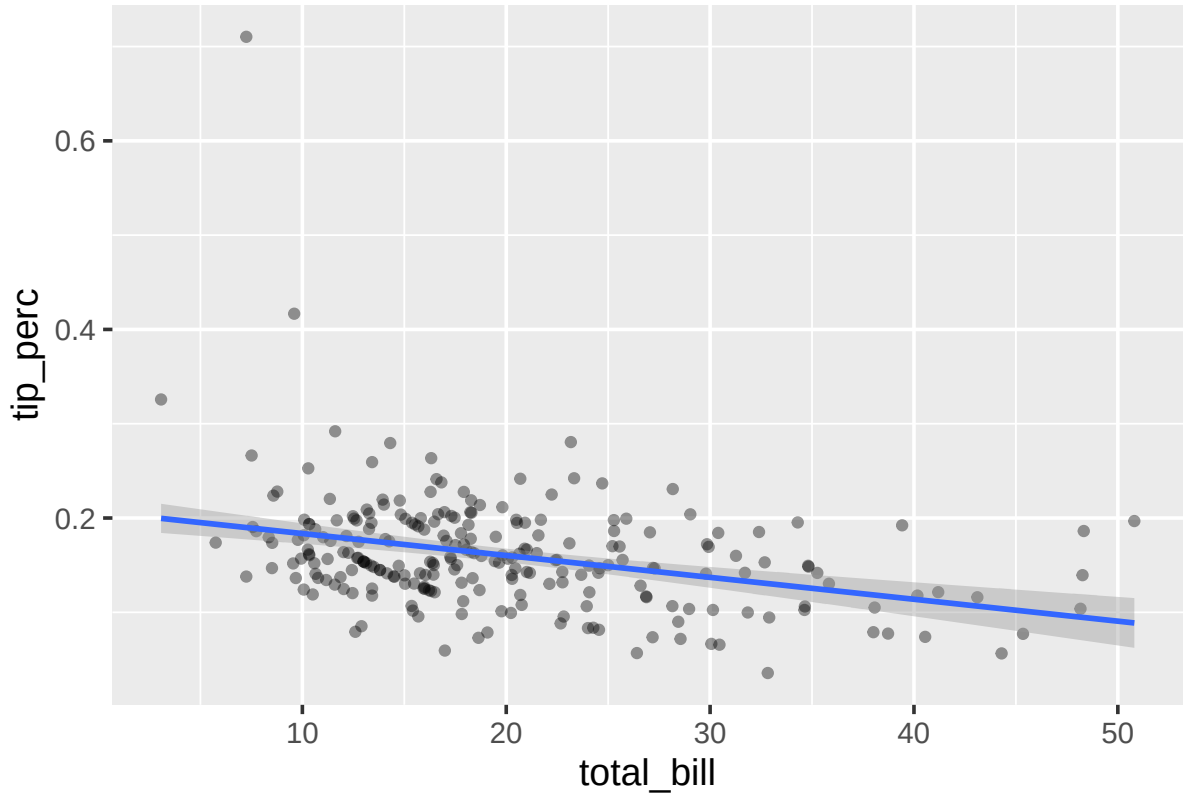
```
##     total_bill    sex smoker day   time size  tip_perc
## 173       7.25   Male    Yes Sun Dinner    2 0.7103448
## 179       9.60 Female    Yes Sun Dinner    2 0.4166667
## 68        3.07 Female    Yes Sat Dinner    1 0.3257329
## 233      11.61   Male     No Sat Dinner    2 0.2919897
## 184      23.17   Male    Yes Sun Dinner    4 0.2805352
## 110      14.31 Female    Yes Sat Dinner    2 0.2795248
```

- Sono presenti degli outlier che possono essere dei punti leva per il modello. Si può pensare di toglierli successivamente.

Analizziamo la distribuzione della variabile risposta:

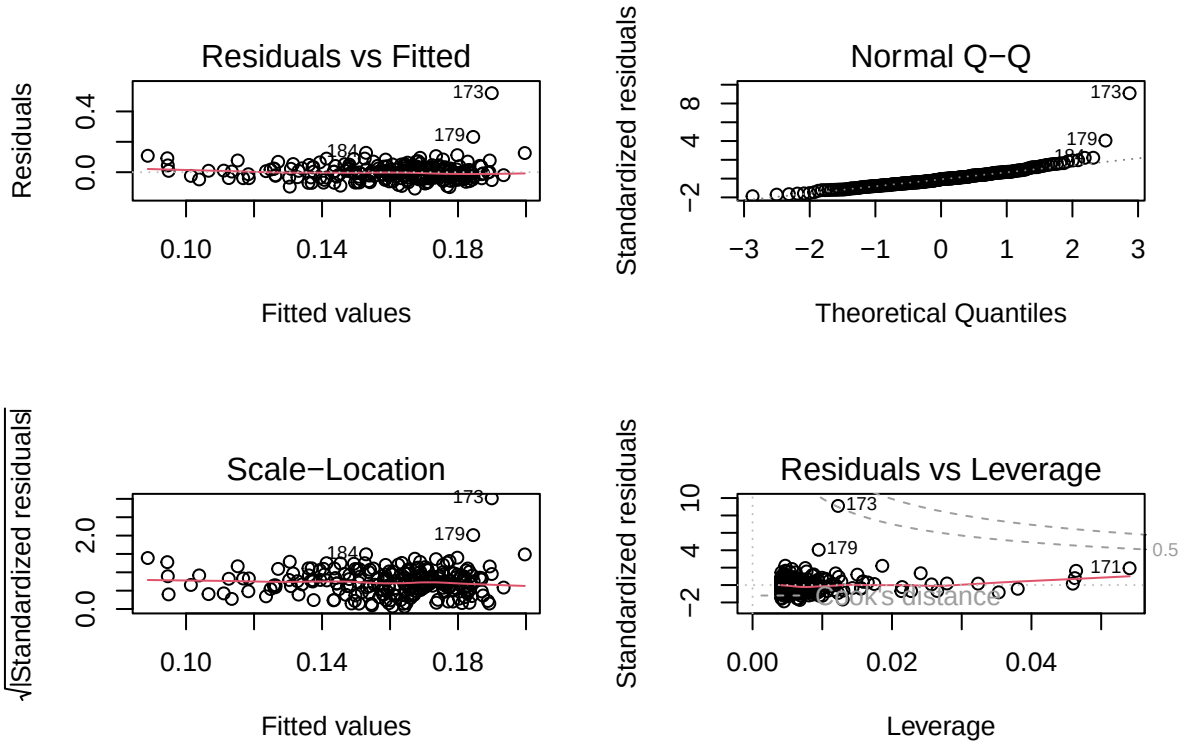


Analizziamo la relazione tra mancia percentuale e conto pagato:

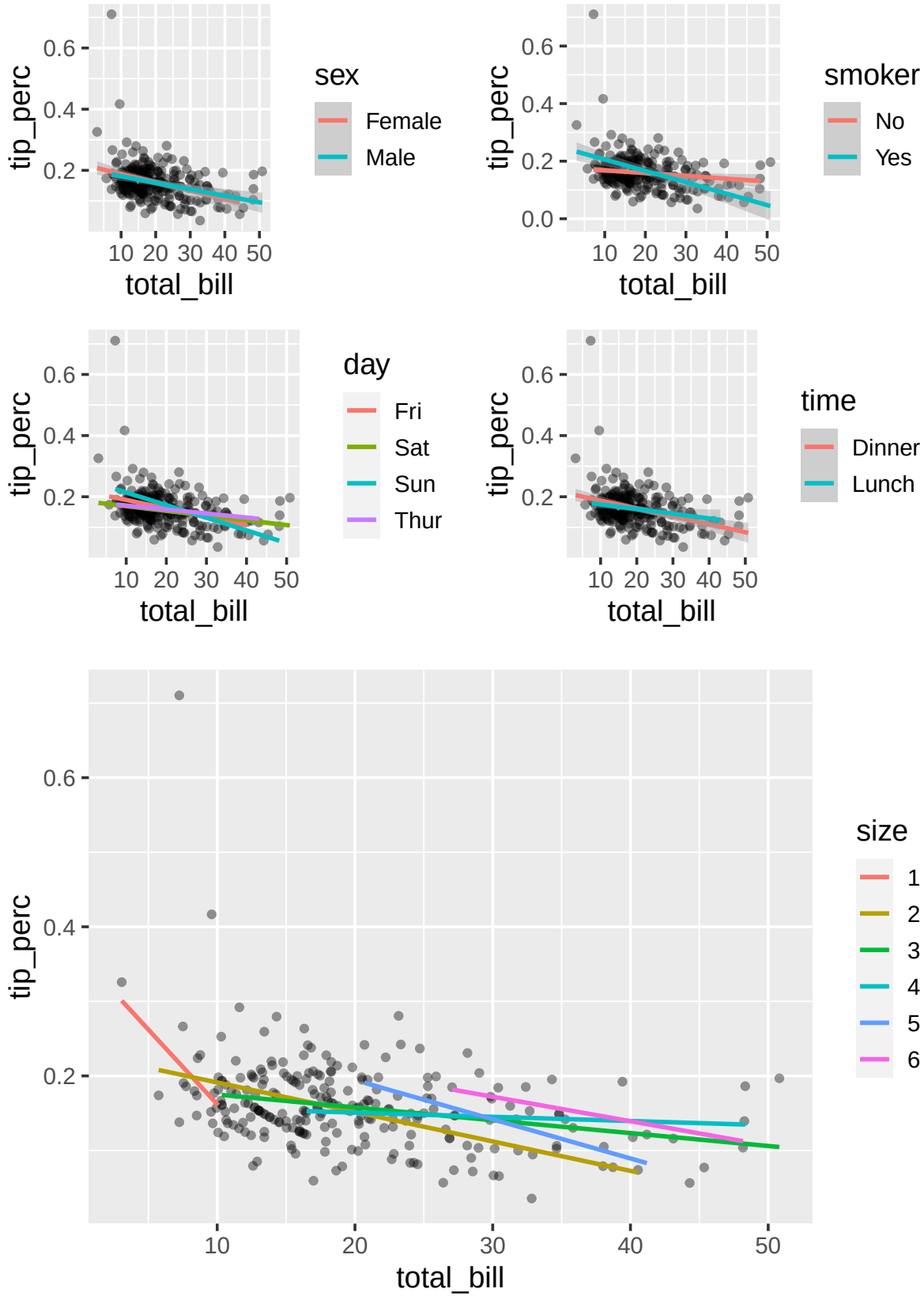


- C'è una correlazione NEGATIVA: all'aumentare del conto da pagare diminuisce la mancia percentuale per il cameriere.

```
##
## Call:
## lm(formula = tip_perc ~ total_bill, data = mydata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.10785 -0.03331 -0.00332  0.02437  0.52042
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.2067658  0.0089995  22.975  < 2e-16 ***
## total_bill  -0.0023230  0.0004149  -5.599 5.85e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05758 on 242 degrees of freedom
## Multiple R-squared:  0.1147, Adjusted R-squared:  0.111
## F-statistic: 31.34 on 1 and 242 DF,  p-value: 5.848e-08
```



- Per ogni dollaro in più possiamo sostenere che la mancia diminuisca del 2%
- La variabilità spiegata dal modello è molto bassa. Circa l'89% della variabilità dipende da altri fattori.



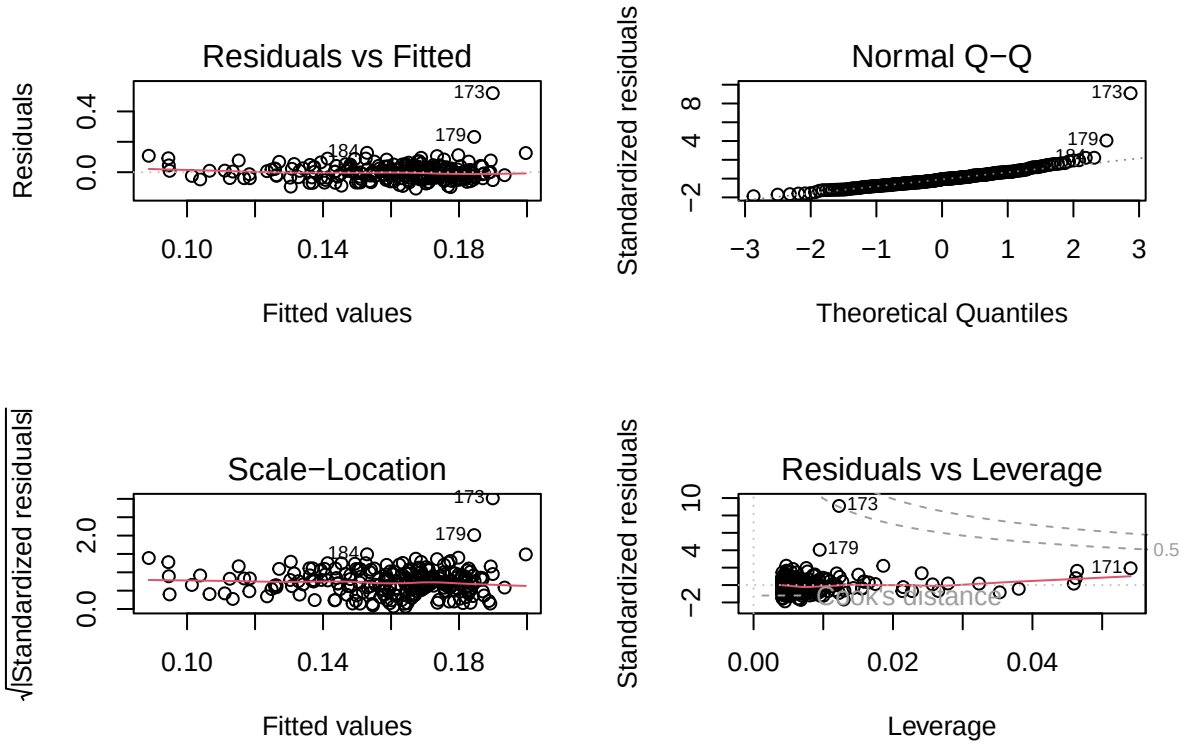
Usiamo un approccio automatico per la ricerca delle variabili esplicative cdi cui disponiamo per provare a

spiegare meglio la varianza della mancia.

```
fitpercA = lm( tip_perc ~ ., data = mydata)
summary(fitpercA)
```

```
##
## Call:
## lm(formula = tip_perc ~ ., data = mydata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.11446 -0.03299 -0.00463  0.02465  0.50003
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.2331193  0.0342788   6.801 8.82e-11 ***
## total_bill  -0.0027879  0.0005461  -5.105 6.92e-07 ***
## sexMale     -0.0036617  0.0080419  -0.455    0.649
## smokerYes    0.0129146  0.0083107   1.554    0.122
## daySat      -0.0032160  0.0176028  -0.183    0.855
## daySun       0.0152658  0.0182169   0.838    0.403
## dayThur     -0.0043163  0.0224666  -0.192    0.848
## timeLunch    0.0030178  0.0253811   0.119    0.905
## size2       -0.0271107  0.0301534  -0.899    0.370
## size3       -0.0222788  0.0320271  -0.696    0.487
## size4       -0.0145897  0.0331143  -0.441    0.660
## size5       -0.0182917  0.0412284  -0.444    0.658
## size6        0.0191991  0.0443782   0.433    0.666
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05778 on 231 degrees of freedom
## Multiple R-squared:  0.1491, Adjusted R-squared:  0.1049
## F-statistic: 3.373 on 12 and 231 DF,  p-value: 0.0001485
```

```
##
## Call:
## lm(formula = tip_perc ~ total_bill, data = mydata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.10785 -0.03331 -0.00332  0.02437  0.52042
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.2067658  0.0089995  22.975  < 2e-16 ***
## total_bill  -0.0023230  0.0004149  -5.599 5.85e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05758 on 242 degrees of freedom
## Multiple R-squared:  0.1147, Adjusted R-squared:  0.111
## F-statistic: 31.34 on 1 and 242 DF,  p-value: 5.848e-08
```



A parte la relazione tra mancia percentuale e conto da pagare, non sembra esserci un contributo significativo delle altre variabili. Il numero di persone al tavolo non sembra influire sulla mancia percentuale.

Conclusioni:

- Le variabili scelte per creare il dataset non sembrano molto utili per spiegare il comportamento di questo fenomeno. Probabilmente una scelta di variabili più legate al servizio del cameriere o del ristorante avrebbero potuto dare un contributo maggiore.
- L' importo della mancia dipende da quanta gente c'è al tavolo e da quanto spendono come era logico immaginare.
- Le persone lasciano meno mancia in percentuale all'aumentare dello scontrino.
- In base alle informazioni ottenute dai dati, il cameriere deve sperare che si presentino più tavoli da 3 e 4 persone perché in termini di frequenza e mancia media sono i più profittevoli per il cameriere.