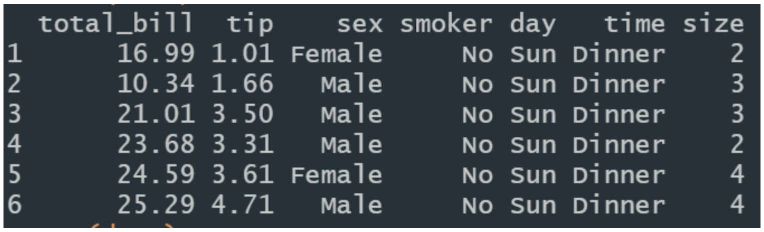
Exam Project Statistical Models

Cazzolato Kevin SM3201245

Year: 2024

1 Problem Description and Dataset Description

We want to study the ”Tips” dataset, and in particular the variable ”tip” is of interest. Therefore, we want to determine if an adequate linear regression model can be established to estimate the ”tip” variable. After importing the dataset, the function head() is used to determine which variables are present in it. These variables turn out to be:

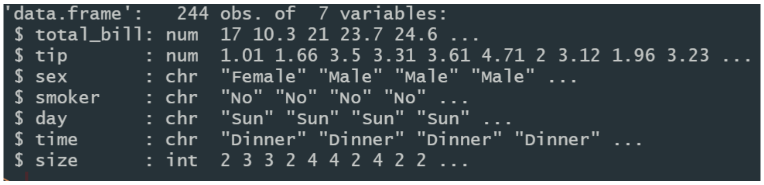


```
  total_bill  tip    sex smoker day   time size
1      16.99 1.01 Female     No Sun Dinner    2
2      10.34 1.66   Male     No Sun Dinner    3
3      21.01 3.50   Male     No Sun Dinner    3
4      23.68 3.31   Male     No Sun Dinner    2
5      24.59 3.61 Female     No Sun Dinner    4
6      25.29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

We observe that the variables in the dataset are:

- total_bill: the total bill for the i-th observation
- tip: the tip left (target of our analysis)
- sex: gender of the customers
- smoker: if the customer is a smoker
- day: day indicated (Mon, Tue,..)
- time: lunch or dinner
- size: number of people at the table

We can also verify the types of the variables using the function 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 ...
```

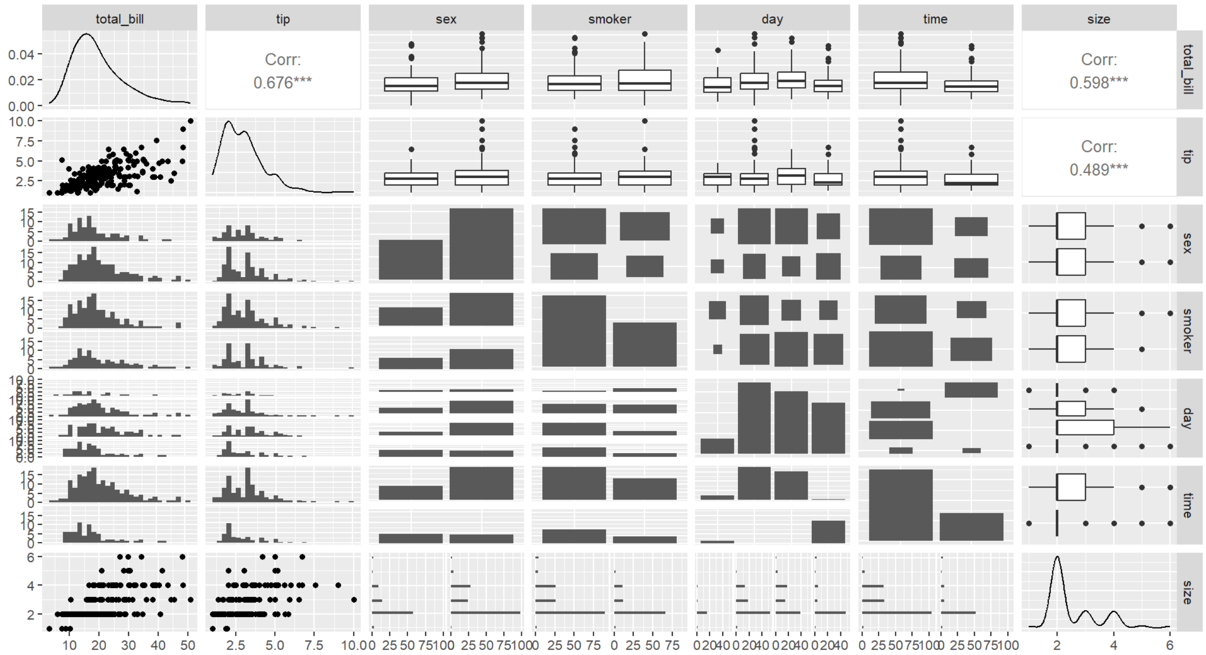
We note that the dataset consists of 244 total observations, and that there are 3 numerical variables: total_bill, tip, and size, while the remaining variables can be transformed into factors for simplicity and convenience. Thus, we transform the categorical variables as follows:

```
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"))
```

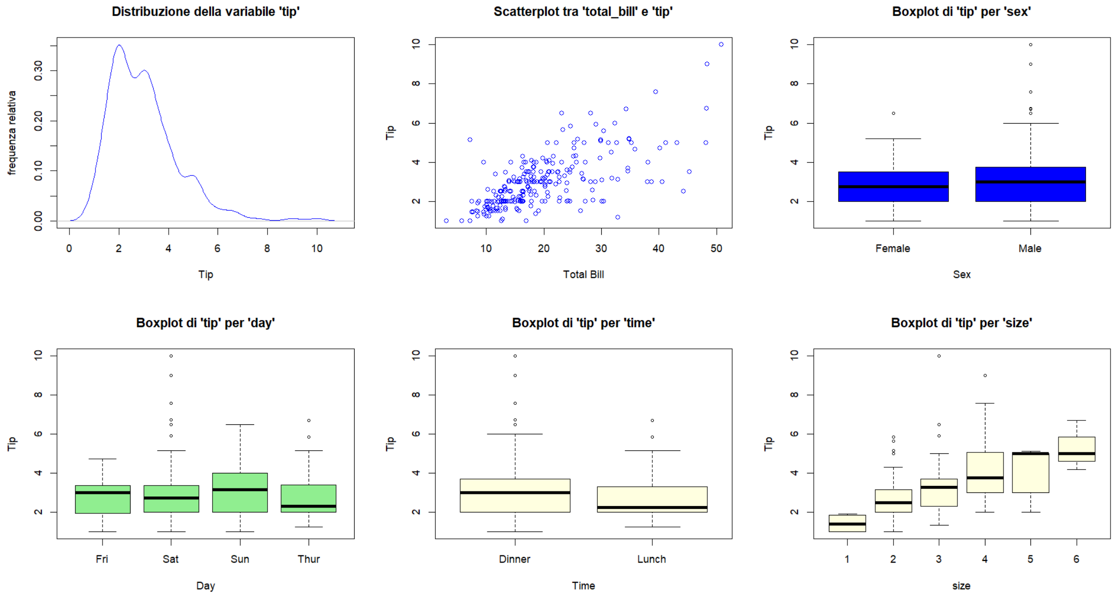
Using the function any(is.na()), we can check if there are any null values that need to be handled, but none are present in the dataset. Moreover, all values are homogeneous, so we have no syntax discrepancies such as ”Yes” and ”yes”, or unexpected values within the numerical variables (e.g., negative or null values). We then proceed with exploratory analysis.

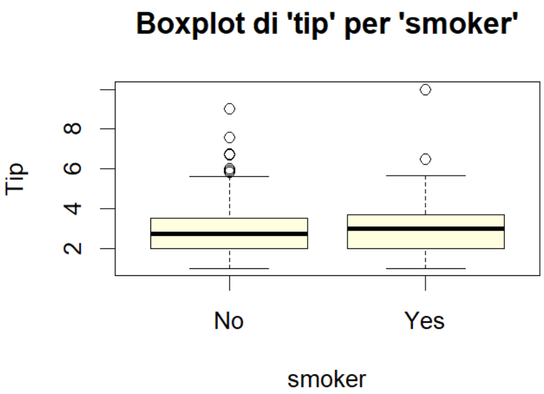
2 Exploratory Analysis

Using the GGally library and the ggpairs function, we perform an initial exploratory analysis of the data:



We can observe that total_bill has a strong correlation with the variables tip and size, thus likely to be of interest for our analysis. Since our analysis focuses on the tip variable, we can immediately concentrate on the relationships between it and the other variables. The following graphs are analyzed:





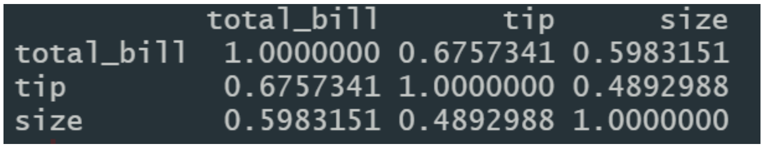
From the first graph, we can observe that the tip variable seems to follow a chi-squared distribution, but this does not affect our analysis, as the Gaussianity hypothesis is required for the model $Y = \alpha + \beta x_i$ and the residuals, not for the response variable. Furthermore, from the second graph, it seems that the tip variable is linearly related to the Bill variable, although there is an asymmetry in the data quantity (so in a future model, if the total bill variable is considered, it would be a good idea to normalize the data). We can also observe heteroscedastic variance, particularly with increasing x, the variance appears to increase (another reason to normalize). The Sex variable does not seem to influence our response variable, as the distributions of Female and Male seem to be equivalent, with a greater quantity of data for Male. The Day variable also does not appear to be significant. There seems to be a slight difference between the Dinner and Lunch distributions for the Time variable, so further analysis is needed to verify a possible correlation with the Tip variable. Finally, the Size variable also seems to influence the Tip variable, as the tip increases with the number of customers at the table, although growth decreases for groups larger than 5 (so it may follow a logarithmic relationship).

2.1 Analysis of Dependencies with the Response Variable

In this section, we want to analyze the possible dependencies and independencies previously identified graphically with the response variable.

2.1.1 Quantitative Variables

To conduct this analysis with quantitative variables, we use the cor() function, which indicates the linear relationship between variables. Thus, we have:



```
           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
```

These results were obtained using the formula:

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

where X_i and Y_i are the values of the i-th observations and $\bar{X}$ and $\bar{Y}$ are the sample means of the variables X and Y, respectively.
Therefore, we observe that the response variable is strongly (positively) correlated with the total bill variable and also, although to a lesser extent, with the size variable. Additionally, the size variable is also strongly correlated (positively) with the total bill variable. Thus, a good model might be one that considers these two variables, given the strong linear relationship.

2.1.2 Categorical Variables

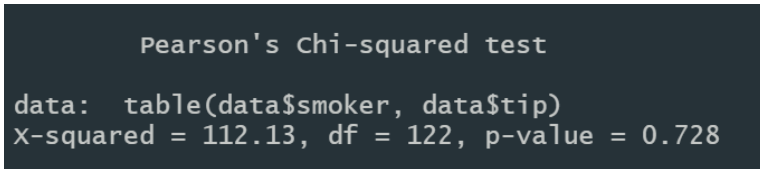
For categorical variables, we can use the **chi-squared test** (we perform it only on categorical variables because it is assumed that X, Y can take on only a finite number of values). This involves constructing a contingency table and testing the null hypothesis:

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

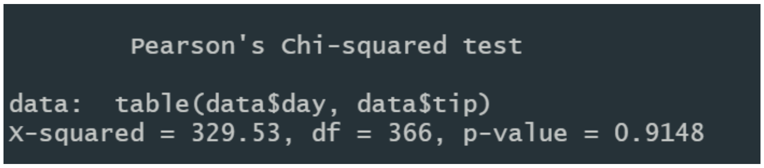
$\forall(i, j)$, where p_{ij} are the theoretical joint probabilities (alternatively, we can use the MLEs: empirical joint frequencies), while $p_{i.}, p_{.j}$ are the theoretical marginal probabilities. In R, we can use the chisq.test() function, which provides the χ^2 statistic, where:

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

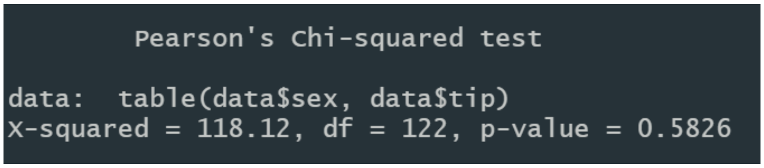
with its degrees of freedom and the p-value. We reject the null hypothesis at a significance level below 0.05. Thus, we obtain the following results:



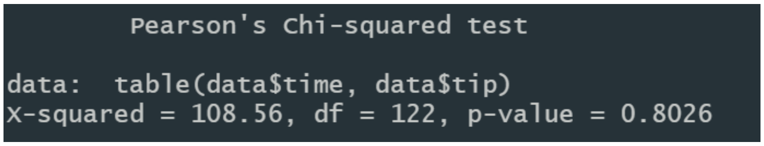
```
        Pearson's Chi-squared test

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
```

Therefore, we do not reject the null hypothesis of independence between these variables and the response variable, and moreover, given the very high p-value, we have strong doubts that they are independent.

3 Model Creation

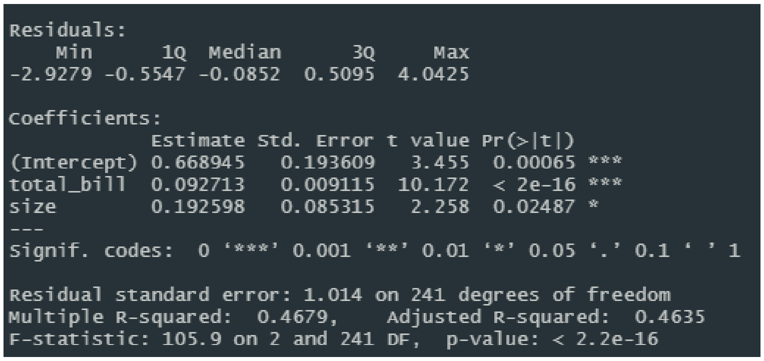
We can now proceed with creating an initial linear regression model: given the small dataset size, we can apply a stepwise algorithm starting with all variables in the dataset. Therefore, using the AIC criterion, defined as: $AIC = 2k - 2ln(L)$ (where k is the number of parameters and L is the maximized value of the likelihood function of the estimated model), the stepAIC() function of the MASS library will return the model with the lowest AIC. In our case, the best model is found to be the one with the variables total bill and size. However, we might consider that another valid model could be the one that considers the relationship between total bill and size (which we saw had a strong linear relationship), hence we can use the following model:

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

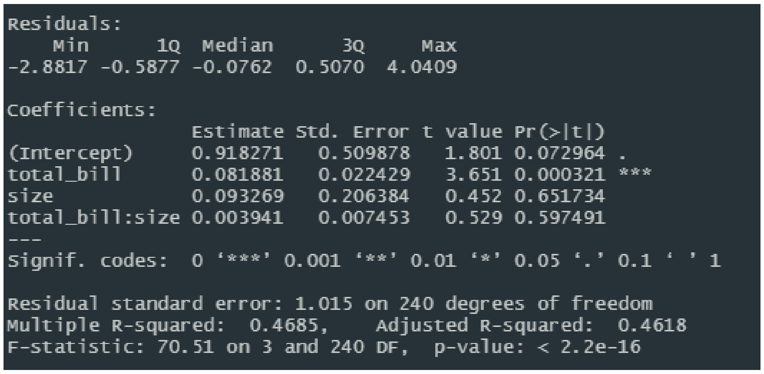
We notice that the first model has a slightly lower AIC compared to the second (703 vs. 705), so we seek further confirmation using the BIC criterion, defined as $BIC = -2ln(L) + kln(n)$. We obtain, respectively, the following results: (717,723), providing further confirmation that the first model might be more adequate. To further confirm this, we use the k-fold cross-validation method to estimate the errors. Using the following code:

```
k_fold <- function(set, k, modelli, norm) {
  errori_complessivi = vector("list", length(modelli)) #list of lists
  fold_size= nrow(set) %/% k #floor of the modulo

  for (i in 1:k) {
    # Select test and training data
    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)
    }
  }

  # Calculate mean errors for each model
  medie = sapply(errori_complessivi, mean)

  return(medie)
}

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

We obtain that the mean error of the first model is 0.7515757 while for the second model it is 0.7595539. Finally, we analyze the summary of both models:



```
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
```

We can observe that all the coefficients in the first model are significant, unlike the second one where only the total bill is significant. We can note, in addition to a higher F-statistic in the first model compared to the second, that the adjusted R-squared value, defined as:

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

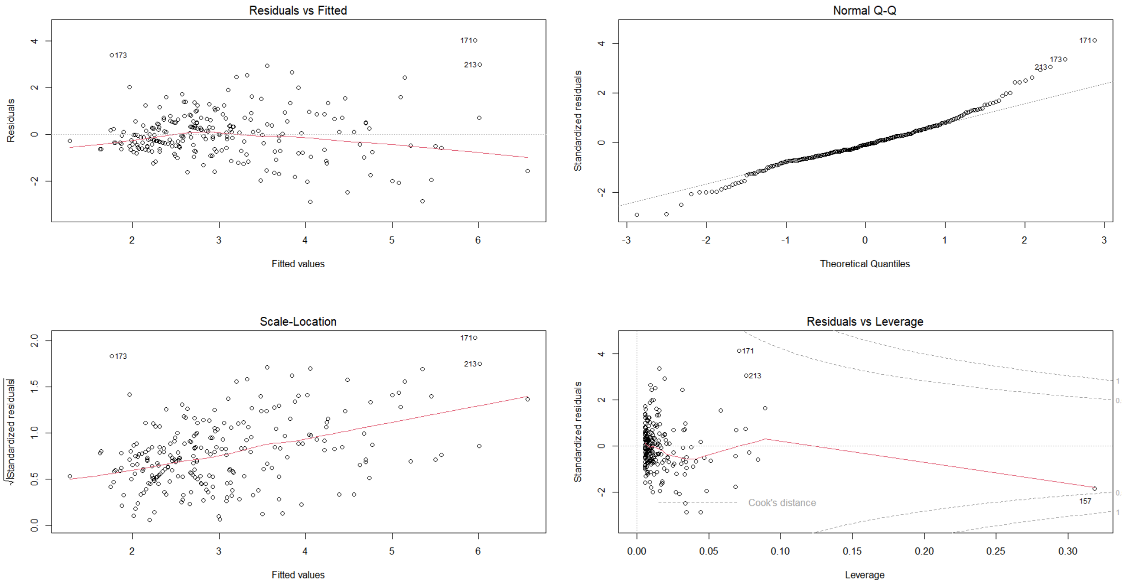
with

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

is higher in the first model. Therefore, we decide to adopt the first model as our reference model.

4 Residual Analysis

Having adopted the previous model, we now want to conduct a residual analysis $(y_i - \hat{y}_i)$ to verify the hypotheses of linearity, normality of errors, and homoscedasticity. We can then plot the following graphs:



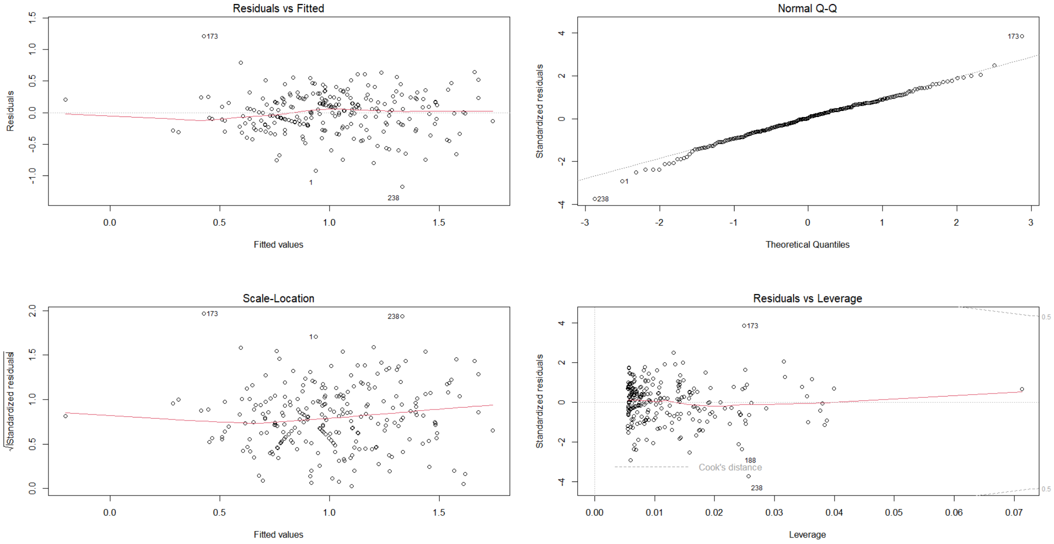
We observe the absence of a pattern in the first graph, which suggests that the linearity hypothesis is respected. Furthermore, from the first graph, it appears that the residuals are heteroscedastic, as the variance appears to increase for higher x values. There also seems to be strong asymmetry in the observations along the y-axis, while symmetry with respect to the X-axis appears to be satisfied. From the second graph, we note that the behavior is quite consistent with that of a Gaussian distribution, except for the tails, which deviate slightly from the line. Finally, in the third graph, there appears to be a pattern of linearity. Thus, the model does not seem to be fully efficient as is, and it is necessary to apply normalization of the data to achieve unique variance and improve the hypotheses of normality and linearity.

4.1 Normalization

We apply various normalizations to the data: Z-score, min-max, and finally a logarithmic transformation. We obtain 3 models trained on 3 different datasets and repeat the steps previously performed to choose the model:

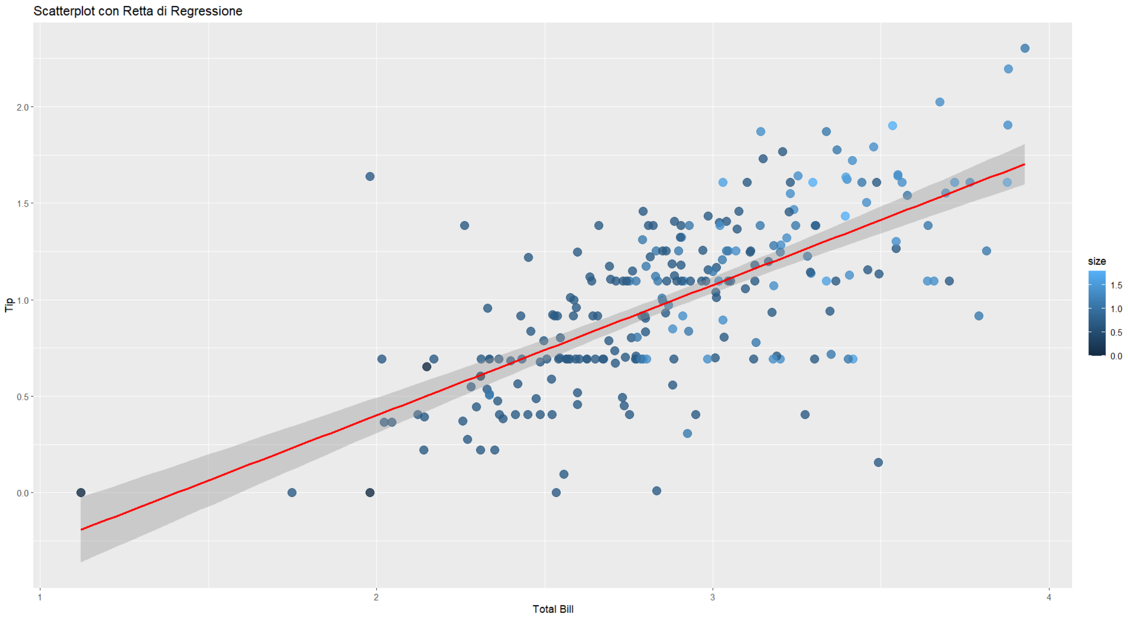
	Mean Error	AIC	BIC
Z score	4.28	545	559
Min Max	0.3847693	368	354
Log	0.2484026	139	153

Thus, without further investigation, we can conclude that logarithmic normalization is the most effective for our dataset. Indeed, we can now observe from these residuals:



that normality appears to be respected, as do Gaussianity and, finally, the homoscedasticity hypothesis.

5 Model Representation


Scatterplot con Retta di Regressione