Comparison of Regression-based methods

Rosario Urso
e-mail address: rosario_urso@hotmail.it

Contents

1 Introduction

The dataset (downloaded from Kaggle[1] examined concerns the tips received by a waiter over several months of work in a restaurant and consists of **244** records on which **6** variables, both quantitative and categorical in nature, were recorded. The variables analyzed are as follows:

- **total bill**, which is the total amount paid, expressed in dollars;
- **sex**, defined as the gender of the bill payer;
- **smoker**, a dichotomous variable to indicate whether there were smokers in the party;
- **day of the week**, from Thursday to Sunday;
- **time**, whether it was lunch or dinner;
- **size**, meaning table size, which ranges from 1 to 6.
  It was decided to make this variable a dichotomous variable, indicating *"Small"* tables with up to 2 eaters and *"Large"* tables with more than 2 eaters.

2 Exploratory Data Analysis

The goal is to, after accurate variable selection using shrinkage and best subset selection methods, compare, interpret and analyze different regression-based methods and select the best method for the examined dataset (comparing the root mean square error of each model).
Before proceeding to select the variables and subsequently to estimate the different models, it was appropriate to analyze the dependent variable **Tip**, expressed in dollars. Since the distribution of the variable appears to be positively skewed, the **Box-Cox transformation** (1964) was applied to choose the lambda value that maximizes the Likelihood Function and ensure that the distrubution can approximate to that of a Normal Random Variable and assume Homoschedasticity (constant variance).
For each $Y_i > 0$, the Box-Cox transformation is defined by:

$$Z_i^{(\lambda)} = \begin{cases} \frac{Y_i^{\lambda}-1}{\lambda}, & \text{if } \lambda \neq 0; \\ \log\left(Y_i\right), & \text{if } \lambda = 0; \end{cases} \quad i = 1, 2, \ldots, n \tag{1}$$

$$\tag{2}$$

Since $\lambda = -0.1$, the logarithmic transformation was applied to the dependent variable, which made the distribution more symmetrical than the untransformed variable, as shown below:

[1] https://www.kaggle.com/datasets/jsphyg/tipping/data and included in the book "Bryant, P. G. and Smith, M (1995) Practical Data Analysis: Case Studies in Business Statistics. Homewood, IL: Richard D. Irwin Publishing"

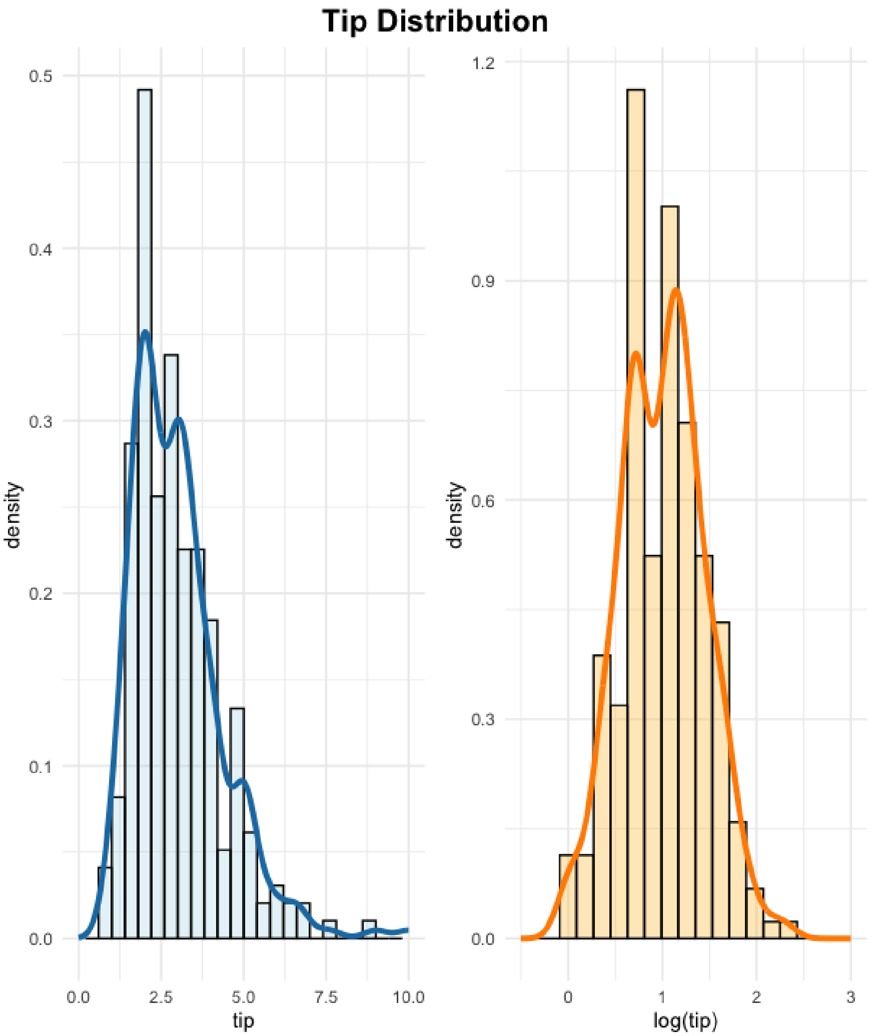

Figura 1: Tip Distribution before and after Box-Cox Transformation

From Figure 1, it can be observed that the average tip is about \$3 and that only 18 people out of the 244 in the dataset leave a tip of an amount greater than \$5.

Hereafter, it was considered appropriate to observe the behavior of the dependent variable regards to the total bill paid using a scatter plot and the results are as follows:

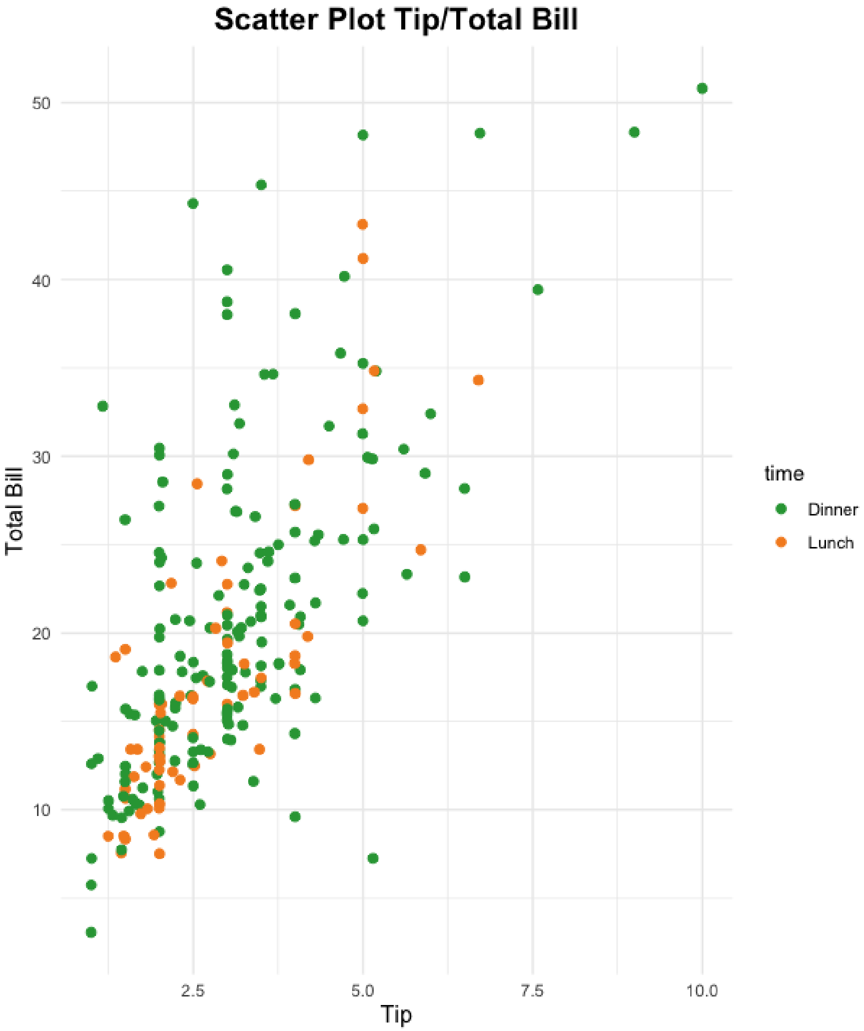

Figura 2: Scatter Plot Tip — Total Bill

From the graph in Figure 2, it can be seen that, as intuitively imagined, as the total bill paid increases, so does the tip given. This is also confirmed by the linear correlation coefficient, which is found to be **0.676**.

However (even though those who lunch are less than 80 , while those who dine are about 180), another interesting fact emerges from the graph, namely that those who dine tend to pay a higher bill than those who lunch. The same is true for tipping, as previously reported.

Then, another graph was made by relating the tip and the sex of the person paying the bill using a *Violin Plot*.

The graph shows that male people tend to tip higher than female people ($3.09 of men compared to $2.83 of women) and that women pay a tip that does not exceed $7, as illustrated in Figure 3.

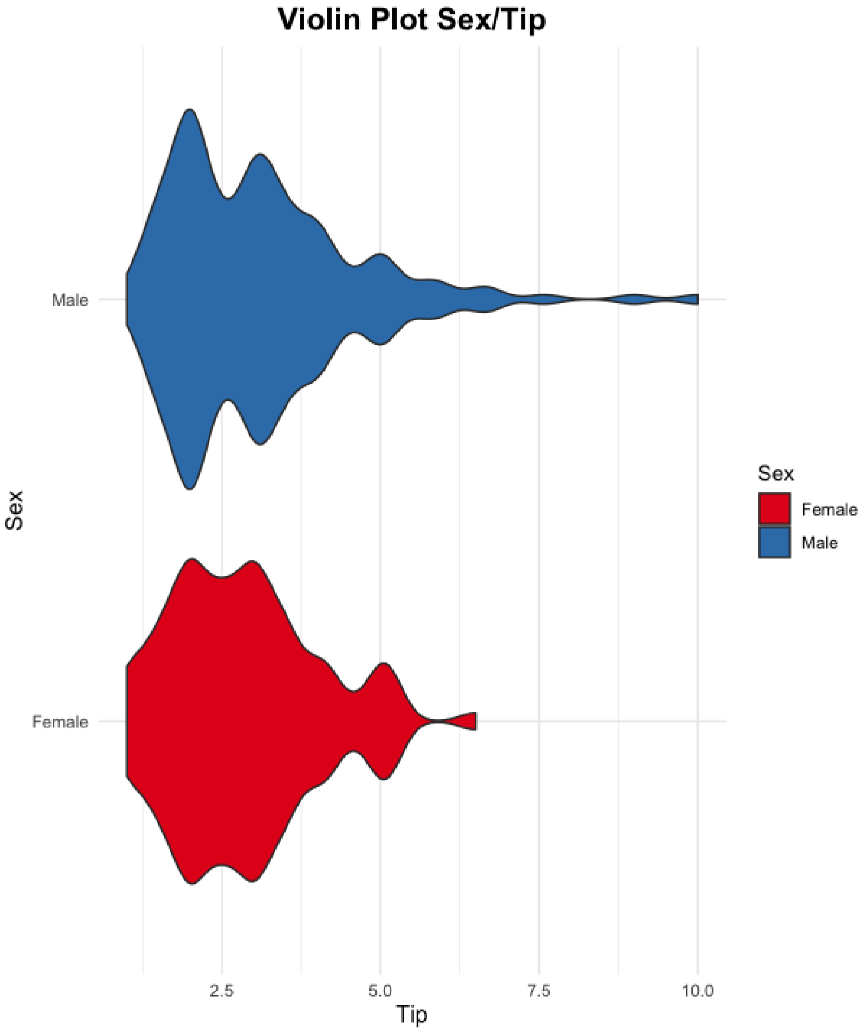

Figura 3: Violin Plot Tip — Total Bill

The same analysis was conducted in Figure 4 compared to the day of the week using the *Raincloud Plot*, which is an intuitive and robust data visualization tool.
From that graphical representation, it can be deduced how those who frequent the restaurant on Sundays tend to leave a higher tip, which amounts to $3.26.
Furthermore, as evidenced by the tails of the graph, the highest tips are recorded on Saturday, followed by Thursday.
However, as evident from the cloud of points (representing observations), only 18 observations were recorded on Friday, so it is difficult to generalize what was stated previously.

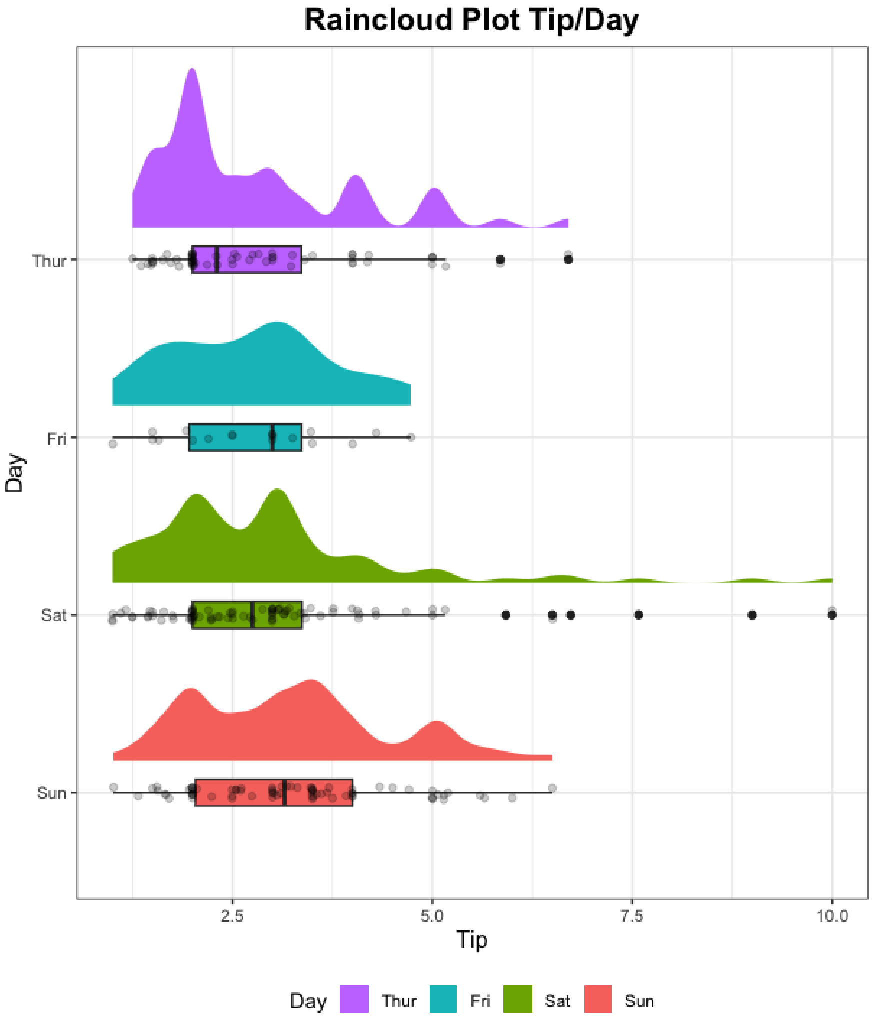

Figura 4: Raincloud Plot Tip — Day

To complete the exploratory analysis performed, the relationship between the response variable **Tip** and all other variables was analyzed, but these did not return significant results, except for the variable **Size**, (which, as mentioned, was made dichotomous). From the point-biserial correlation calculation, which computes correlation between a dichotomous and a continuous variable, a correlation coefficient of about **0.44** resulted, which means that as the size of the table increases, the amount of tip also increases.

3 Dataset Structure

Before proceeding to the various model estimates, in order to obtain more reliable estimation and avoid running into the problem of overfitting, the **v-fold cross validation** approach was used.
The dataset was splitted into 80% training set and 20% test set, then set the number of folds equal to 10. This means that the training set was divided into 10 equal parts, and then the model was trained and estimated 10 times, using each of the 10 parts as a test set once at a time.

4 Variable Selection

For the selection of variables, two different methodologies were used:

- **Shrinkage Methods** (Lasso and Ridge Regression)
- **Subset Selection** (Best Subset Selection)

4.1 Shrinkage Methods

The goal of shrinkage methods (whether Lasso or Ridge regression) is to increase the performance of the model by working on the estimated coefficients and thus going to reduce the variance of the estimates. This is achieved by setting constraints and regularizing the coefficient estimates.
Differently from subset selection, as will be seen in the next section, all **p** predictors are included in the model, and not just a subset.

4.1.1 Ridge Regression

Ridge Regression (also called $\ell 2$ Regularization) is a regularization method that is very similar to least squares, except for a term called the **shrinkage penalty**, which has the effect of forcing the coefficient estimates β_j towards 0.
The coefficient estimates $\hat{\beta}^R$ are the values that minimize this function:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=\text{RSS}+\lambda\sum_{j=1}^{p}\beta_j^2 \tag{3}$$

where **RSS** is the Residual Sum of Squares, which is a statistical technique used to measure the amount of variance in a dataset that is not explained by a regression model itself and $\lambda \geq 0$ is the **tuning parameter**.
$\lambda \sum_{j=1}^{p} \beta_j^2$ is called the **shrinkage penalty**.
When the tuning parameter λ is set exactly 0, the ridge regression produces the same results as the least squares estimates, so no shrinkage penalty is applied. Conversely, when $\lambda \rightarrow \infty$, the impact of the shrinkge penalty grows and the coefficient estimates will tend to 0.
The drawback of using Ridge Regression is that, introducing all p predictors into the model, does not set any estimated coefficient exactly to 0, so in cases where there are many predictors there may be difficulties in the interpretation of the model. The alternative to this approach is the Lasso Regression.

4.1.2 Lasso Regression

The Lasso Regression (also called $\ell 1$ Regularization) is an alternative to the Ridge Regression that ”solves” the issue of estimates that are not set to be exactly 0.
The coefficient estimates $\hat{\beta}_\lambda^L$ are the values that minimize this function:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|=\text{RSS}+\lambda\sum_{j=1}^{p}|\beta_j| \tag{4}$$

As the other shrinkage method, λ is the parameter to tune and in this case it changes the penalty parameter which result to be $\lambda \sum_{j=1}^{p} |\beta_j|$.
However, Lasso regression has the effect of forcing coefficient estimates to exactly 0 when λ is sufficiently large, so it performs like a variable selection method.

4.2 Best Subset Selection

Best subset selection evaluates all possible combinations of the p predictors, trains and evaluates a separate model for each combination, and finally selects the best one based on an evaluation criterion (e.g., R$^{\mathbf{2}}$, Akaike Information Criterion (**AIC**), Bayesian Information Criterion (**BIC**)).
The drawback is that it is computationally very cumbersome if the dataset contains too many predictors, since the algorithm proceeds to estimate 2^p models.
Then the algorithm run the following steps:

- It starts with the null model M_0, which contains no predictors;
- for each $k = 1, 2, \ldots, p$, it estimates all $\binom{p}{k}$ models that contain k predictors. From these estimated models, the best one is chosen based on a evaluation criterion, such as the lowest RSS or the highest R$^{\mathbf{2}}$, called M_k.

The variables included in this model will form the optimal subset of variables to use.
Once the best subset of variables has been selected, it is possible to interpret the coefficients of the resulting model to understand the relationships between the response variable and the selected predictors.

4.3 Empirical Analysis

In the case under consideration, both a Lasso and a Ridge Regression were estimated.
However, since the dataset contained no large number of variables, both shrinkage methods returned almost the same results, considering the rmse criterion.
The values of parameter λ subjected to tuning are reported below in Table 1:

	Ridge	**Lasso**
λ	0.0596	0.0373

Tabella 1: Tuning Parameter for Ridge and Lasso Regression

Using Lasso Regression, but the same in this case happens at Ridge, the best subset of predictors was selected using a graphical tool representing **variable importance**, both positive and negative, as shown in Figure 5:

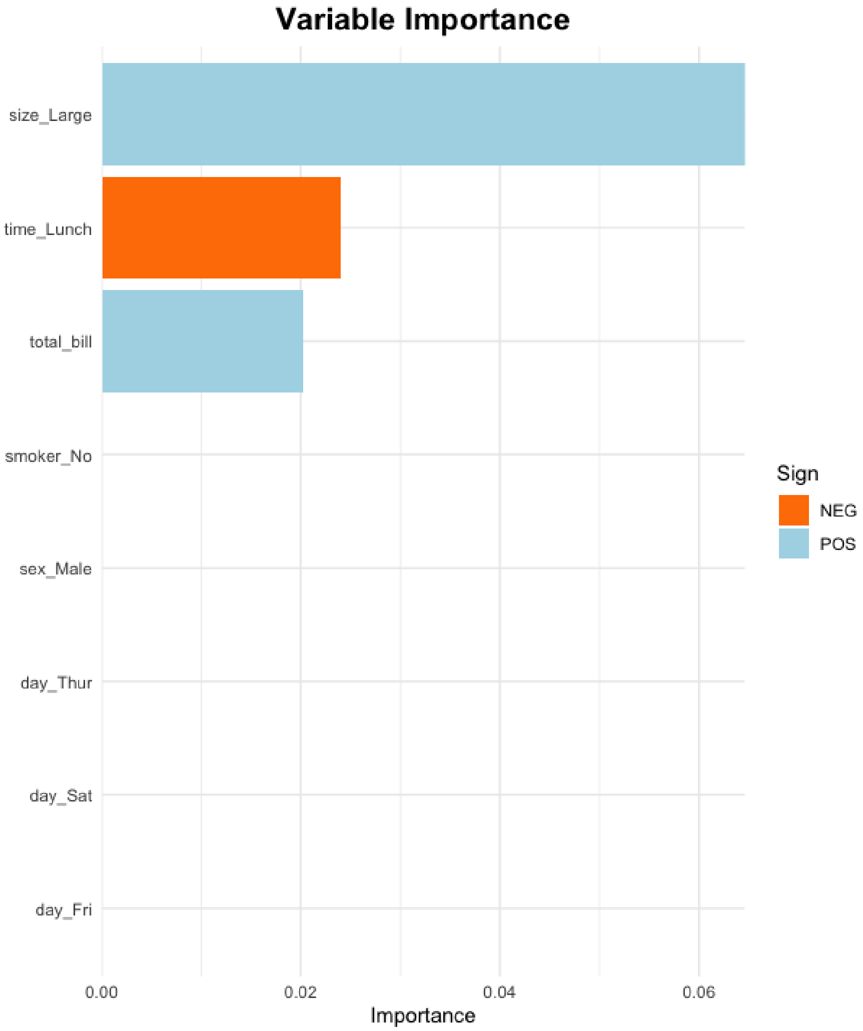

Figura 5: Variable Importance using Shrinkage Methods

The variables used to explain the response variable are Size and Total Bill, which give positive contribution, while negative importance values (Lunch) suggest that the variable can have a detrimental impact on the classification. Same result is obtained by Best Subset Selection, where the algorithm leads to the choice of 2 predictors, which turn out to be, as for the shrinkage methods, Size and Total Bill, as shown next:

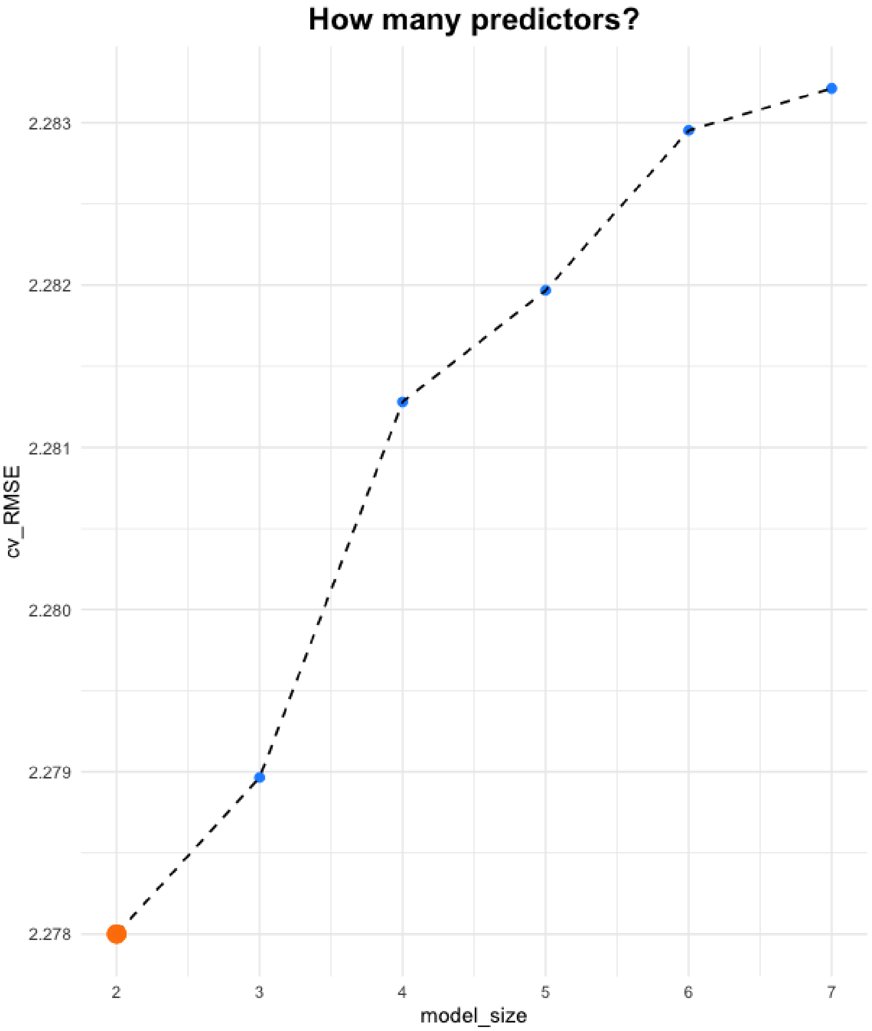

Figura 6: Number of predictors using Best Subset Selection

In Figure 6, the ideal predictor number is reported using best subset selection, using the number of explanatory variables that makes the root mean square error minimum (computed on cross validation approach) as criterion.

5 Model Implementation

5.1 Multiple Linear Regression

A multiple linear regression model, as discussed in previous sections, is a statistical model used to analyze the relationship between the response variable and multiple explanatory variables that may influence the dependent variable.
The model is specified as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon \tag{5}$$

where Y is the dependent variable, β_0 is the intercept, which represents the value of Y when all predictors are equal to zero, $\beta_1 X_1, \beta_2 X_2, \ldots, \beta_p X_p$ measure the effect of the explanatory variables $X_1, X_2, \ldots, X_p$, on the dependent variable and ϵ is the error term.
The coefficients $\beta_1 X_1, \beta_2 X_2, \ldots, \beta_p X_p$ are estimated using least squares method, where the aim is to minimize the sum of squares of the residuals between the observed values of Y and the predicted values from the model.

5.2 Polynomial Regression

To describe a non-linear relationship between dependent variable and predictors, instead of using a straight line (as in linear regression), a polynomial is used to approximate the relationship between the variables.

$$Y = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \ldots + \beta_d X_i^d + \epsilon_i \tag{6}$$

where the notation is the same as the notation used for multiple linear regression: the only difference is in the parameter d, which represents the degree of the polynomial and results to be the tuning parameter (usually no greater than 3/4).
Again, coefficient estimation is calculated by the least squares method, since this regression is simply a regression model as seen above but with predictors $x_i, x_i^2, \ldots, x_i^d$

5.3 Empirical Analysis

Estimation of the multiple linear regression model helps to understand how the predictors affect the response variable.
Results are shown in the table below:

	Intercept	**Total Bill**	**SizeLarge**
Coeff	0.37687885	0.02909711	0.09981901
ExpCoeff	1.457728	1.029525	1.104971

Tabella 2: Multiple Linear Regression Coefficients

Keeping in mind that a logarithmic transformation was made to the dependent variable, in order to get an accurate result on the influence of the variables, it is necessary to calculate the exponential of the estimated coefficients.
However, the positive sign of the coefficients suggests that:

- - as the **Total Bill** increases, so does the tip;
- - as the **Size** of the table increases, a bigger tip is received.

The precise interpretation of the results is as follows:

- for variable **Total Bill**, it can be deduced that, for each additional dollar spent in the restaurant, the tip is expected to increase by about 2.9525%, keeping the other variables constant;
- for variable **Size**, it is observed, that the tip will be about 10% higher if the table is large rather than small.

Finally, considering the estimation of a polynomial regression model, the tuning parameter was appropriately chosen, which is shown in the Figure 7:

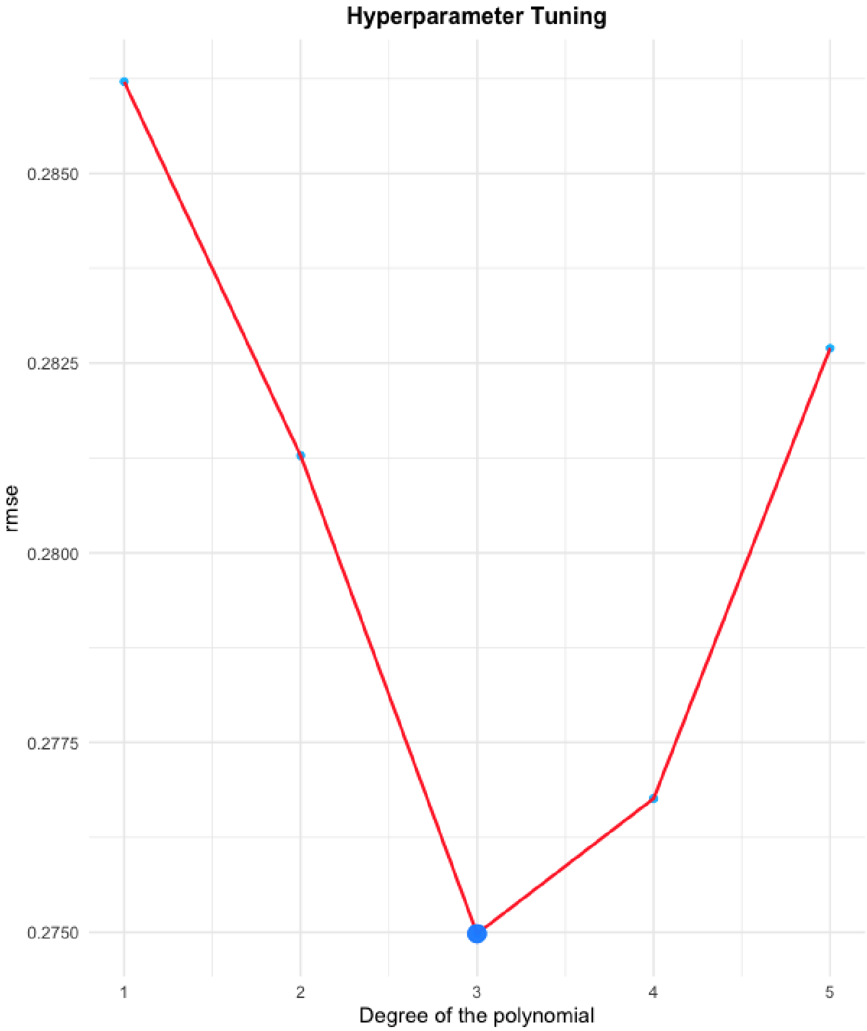

Figura 7: Hyperparameter Tuning for Polynomial Regression

The algorithm thus recommends a polynomial degree of 3, corresponding to the minimum value of rmse.

Results are shown below:

	Intercept	**Total Bill1**	**Total Bill2**	**Total Bill3**	**SizeLarge**
Coeff	0.9064	3.0130	-0.8310	0.8356	0.0736

Tabella 3: Polynomial Regression Coefficients

6 Results and Conclusion

To compare all the methods used, the measurement used is rmse:

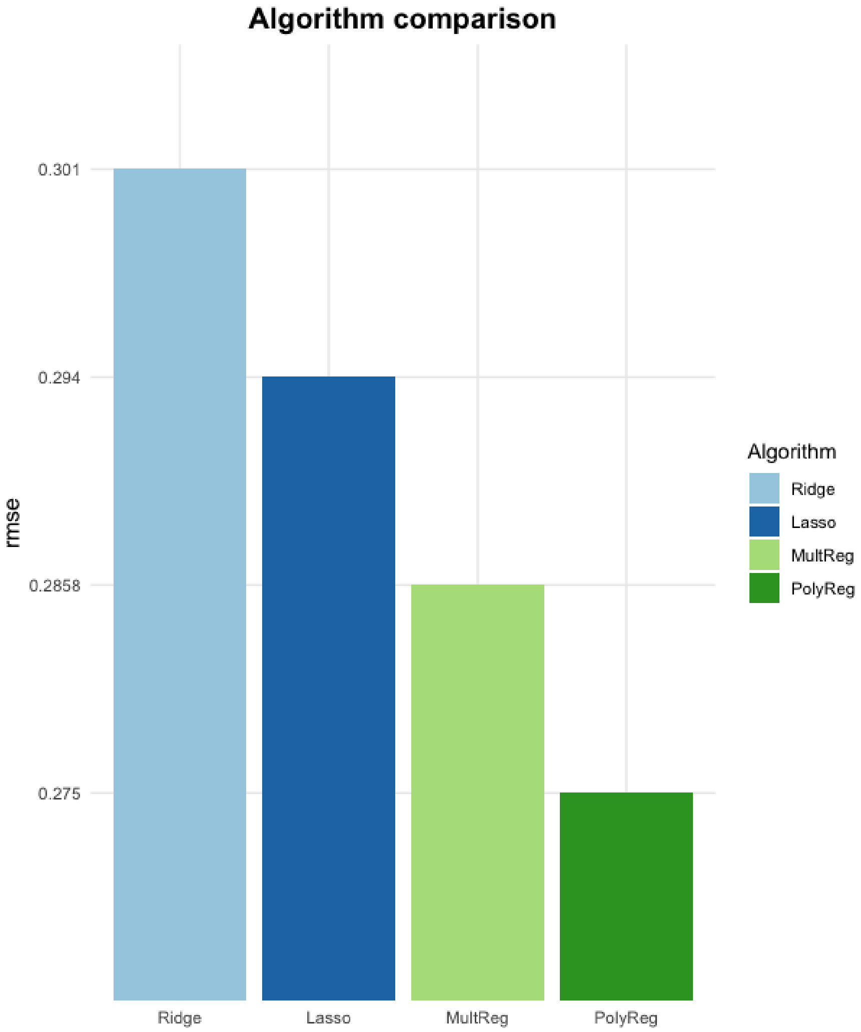

Figura 8: Algorithm Comparison

From the graph in Figure 8, it can be deduced that Lasso and Ridge performed quite similarly, which could be due to the few variables available in the dataset.

Furthermore, the model that performed best appears to be the **polynomial regression model**, whose rmse results equal to0.2750.

In conclusion, it could be said that there is a non-strictly linear relationship between the response and the predictors.

However, it should be pointed out that by increasing the degree of the polynomial (3 in this case) it is possible to model more complex relationships between the variables, but overfitting could be incurred, since a polynomial with too high a degree may fit too much of the training data and not generalize well to new data.

7 References

- Piccolo Domenico, *Statistica.* Il Mulino, 2010.

- Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. *An Introduction to Statistical Learning : with Applications in R.* New York Springer, 2013.

- Alan Agresti, *Categorical Data Analysis.* Vol792, John Wiley Sons, 2012.