Comparison of Regression-based methods

Erasmus+ - Blended Intensive Program (BIP)

Rosario Urso
rosario_urso@hotmail.it

University of Naples Federico II

Thursday 19th October, 2023

Dataset Structure

The dataset concerns the tips received by a waiter over several months of work in a restaurant and consists of **244** records on which **6** variables (*total bill*, *sex*, *smoker*, *day of the week*, *time* and *size*)

What about the outcome?

Since the distribution of the variable appears to be positively skewed, the **Box-Cox transformation** (1964) was applied

For each $Y_i > 0$, the Box-Cox transformation is defined by:

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

Since $\boldsymbol{\lambda} \simeq \mathbf{-0.1}$, the logarithmic transformation was applied to the dependent variable.

Dataset Structure

Before performing the following analysis, it was appropriate to split the dataset, and the approach used was **k-fold cross validation**.

The dataset was divided into $\mathbf{80\%}$ training test and $\mathbf{20\%}$ test set.
Then, $\mathbf{k = 10}$ folds were created on the training set, and at each iteration, one of the k subsets is used as the validation set, while the other **k-1** subsets are used as the training set, as shown below:

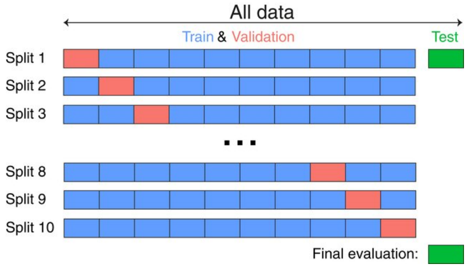

Figure 1: K-Fold Cross Validation Approach (Smirnov, 2020)

Exploratory Data Analysis

A scatter plot was used to determine the linear relationship ($\rho = \mathbf{0.676}$) between the response and total bill, discriminating by meal type .

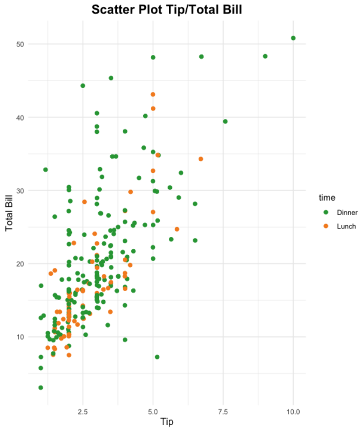

Figure 2: Scatter Plot Tip — Total Bill

Exploratory Data Analysis

The same analysis was conducted in Figure 3 compared to the day of the week using Raincloud Plot.

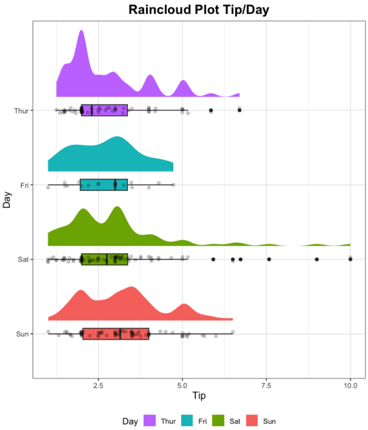

Figure 3: Raincloud Plot Tip — Day

Methods

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

- **Shrinkage Methods** (Lasso and Ridge Regression).
  The purpose of Lasso and Ridge Regression is to set constraints and regularize the coefficient estimates, in order to increase the performance.
- **Subset Selection** (Best Subset Selection)

What's the difference?

The difference is that while Shrinkage Methods include all p predictors in the model, Best Subset Selection includes only a subset of $m < p$ predictors.

The common goal is to reduce the variance of estimates, and thus increase the performance of the model.

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**.
The coefficient estimates $\hat{\beta}^R$ are the values that minimize this function:

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

The drawback of using Ridge Regression is that, introducing all p predictors into the model, does not set any estimated coefficient exactly to 0.
The alternative to this approach is the Lasso Regression.

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:

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

Lasso regression has the effect of forcing coefficient estimates to exactly 0 when λ is sufficiently large.

Best Subset Selection

Best Subset Selection evaluates all possible combinations of the p predictors, trains and evaluates a separate model for each combination, and then selects the best one based on an evaluation criterion.

The algorithm run the following steps:

- It starts with the null model M_0;
- for each $k = 1, 2, \ldots, p$, it estimates all $\binom{p}{k}$ models that contain k predictors. The best one (M_k) is chosen based on a evaluation criterion.

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.

Number of predictors

Both approaches provide the same results, as shown below:

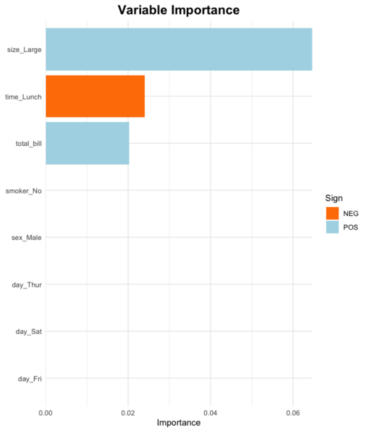

Figure 4: Variable Selection using Shrinkage Methods

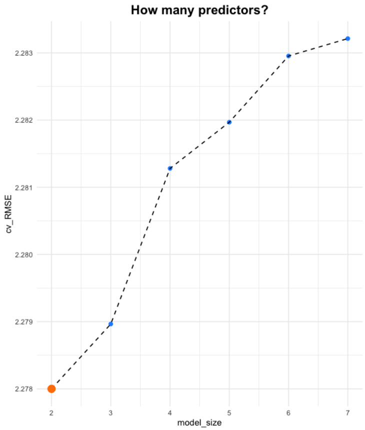

Figure 5: Variable Selection using Best Subset Selection

Multiple Linear Regression

A multiple linear regression model 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:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$

The coefficients $\beta_1, \beta_2, \ldots, \beta_p$ 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.

Polynomial Regression

A Polynomial Regression is used to approximate the relationship among the variables to describe a non-linear relationship between dependent variable and predictors.
Model specification is shown below:

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

where the parameter d represents the degree of the polynomial and results to be the tuning parameter.

Polynomial Regression

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

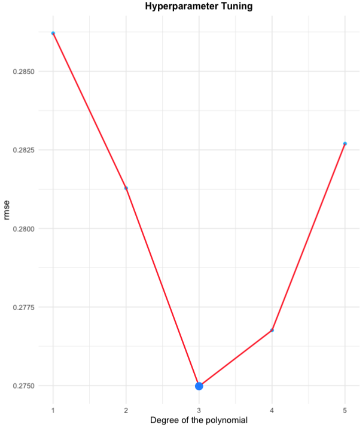

Figure 6: Hyperparameter Tuning for Polynomial Regression

Empirical Analysis

In order to get an accurate result on the influence of the variables, it is necessary to calculate the exponential of the estimated coefficients.

	Coefficients	**Exp Coefficients**
Intercept	0.3768	1.4577
Total Bill	0.0290	1.0295
SizeLarge	0.0998	1.1049

Table 1: Multiple Lineare Regression Results

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.

Results

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

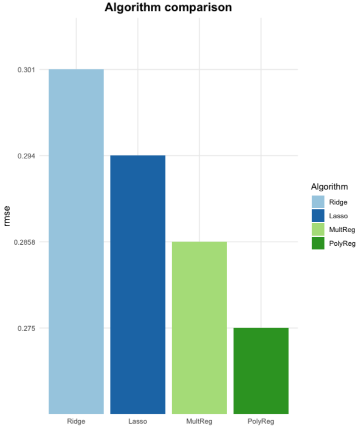

Figure 7: Algorithm Comparison

Conclusions

The model that performed best appears to be the **polynomial regression model**.

It could be supposed that there is a non-strictly linear relationship between the response and the predictors.

Any drawbacks?

- 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.
- The interpretation of the coefficients is not always so easy, compared with the coefficients of the linear regression.

References

[1] Alan Agresti. *Categorical data analysis*. Vol. 792. John Wiley & Sons, 2012.

[2] Gareth James et al. *An introduction to statistical learning*. Vol. 112. Springer, 2013.

[3] D. Piccolo. *Statistica*. Strumenti / il Mulino. Il Mulino, 2010. ISBN: 9788815139023. URL: https://books.google.it/books?id=q7YNSQAACAAJ.

Thank you for your attention!

☺