University of Cape Town

Course Code

STA 5076Z

Predicting Restaurant Tips using Multiple Linear Regression: A Model Comparison Approach

Author:
Khuliso Mmbi

Student Number:
MMBKHU001

April 22, 2024

Contents

1 Plagiarism Declaration

I, Khuliso Mmbi, hereby declare that the work on which this document is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university. I authorize the University to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.
Signature: Khuliso Mmbi
Date: 21 April 2024

2 Introduction

This study was done based on 200 observations created by one waiter who recorded information about each tip she received over a period of a few months working in one restaurant in California.

The primary objective of this study is to analyze the relationship between these factors and the tip amount received by the waiter. By examining this relationship, the study aims to uncover insights into customer tipping behavior, contributing to a better understanding of the dynamics at play in the service industry.

3 Explanatory data analysis

3.1 Data description

The data set includes the following explanatory variables for predicting tip amounts:

- **Tip:** Amount in dollars
- **Total bill:** Amount in dollars
- **Size:** size of the party
- **Sex:** Gender of the bill payer (Female or Male)
- **Smoker:** Smoking status of the party (Smoker or Non-Smoker)
- **Day:** Day of the week (Thursday, Friday, Saturday, or Sunday)
- **Time:** Time of day (Lunch or Dinner)

The tables below (Table 1 to Table 6) show the summary statistics of the above-mentioned explanatory variables.

Table 1: Summary Statistics for Total Bill

Statistic	**Value**
Minimum	3.07
1st Quantile	13.24
Median	17.80
Mean	19.54
3rd Quartile	24.34
Maximum	50.81

Table 1 shows the total bill that the customer paid. The lowest amount paid is \$3.07 and the highest being \$50.81 while the average is \$17.80

Table 2: Counts for Sex of the Bill Payer

Category	**Count**
Female	72
Male	128

Table 2 shows the number of customers from the 200 observations that are female and the number of customers that are males.

Table 3: Counts for Smoker Status

Category	**Count**
Smoker	79
Non-Smoker	121

Table 3 presents the distribution of customers based on their smoking status, with 79 individuals identified as smokers and 121 individuals categorized as non-smokers.

Table 4: Counts for Day of the Week

Day	**Count**
Friday	16
Saturday	74
Sunday	59
Thursday	51

Table 4 displays the counts of customers visiting the restaurant on different days of the week, with Saturday being the busiest day.

Table 5: Counts for Time of Day

Time	**Count**
Lunch	56
Dinner	144

Table 5 shows the counts of customers visiting during lunch and dinner hours, with dinner attracting a significantly higher number of patrons.

Table 6: Counts for Size of the Party

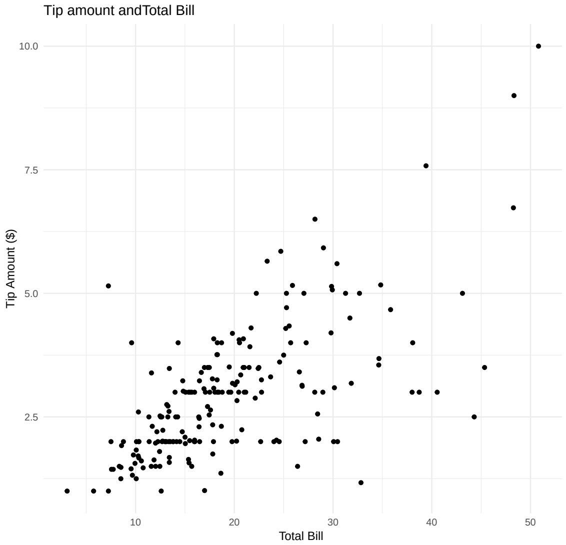
Size	**Count**
1	4
2	136
3	27
4	28
5	3

Table 6 presents the distribution of party sizes among customers, with parties of 2 people being the most common.

3.2 Data Visualization

Visualizing relationships between predictors and the target variable

Figure 1: Demonstrates a positive linear relationship between the total bill paid and the tip received, indicating that as the total bill increases, so does the tip amount.
Figure 2: Indicates that smaller parties, particularly those with only two people, tend to leave higher tip amounts compared to larger parties.



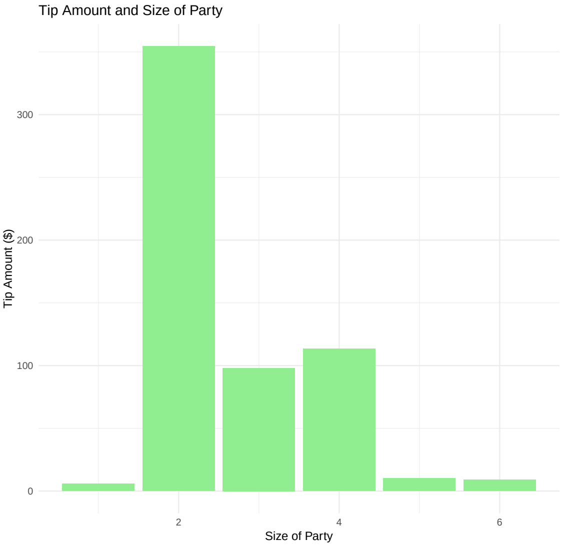
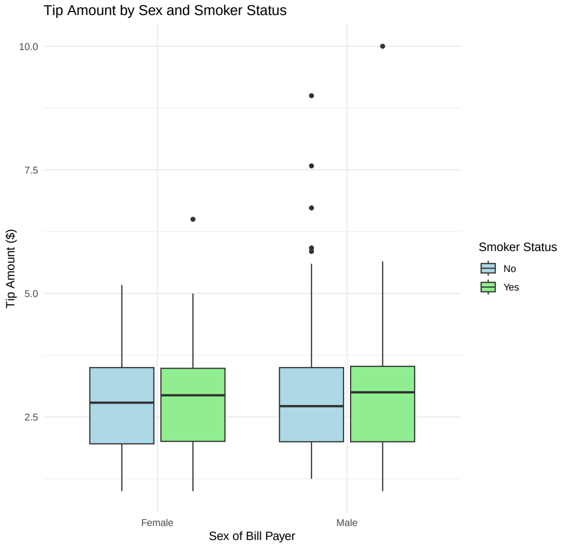


Figure 3: Suggests that there's little variation in tip amounts based on the sex of the customer or their smoking status, as the tip amounts appear similar regardless of these factors.
Figure 4: Highlights that tips are generally higher on weekends (Saturday and Sunday) compared to weekdays (Thursday and Friday), possibly due to increased leisure spending during weekends.



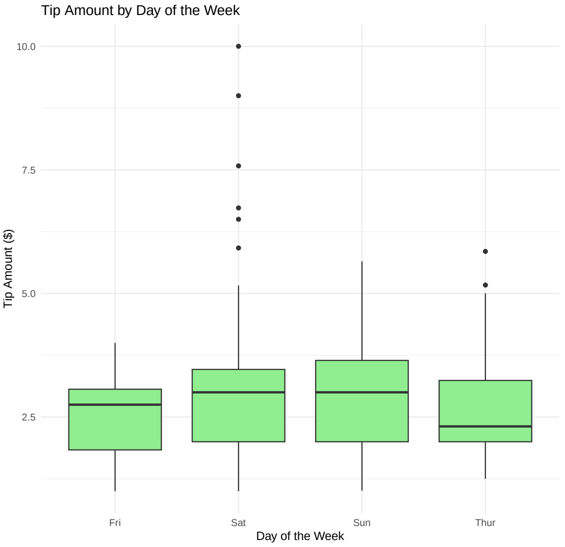
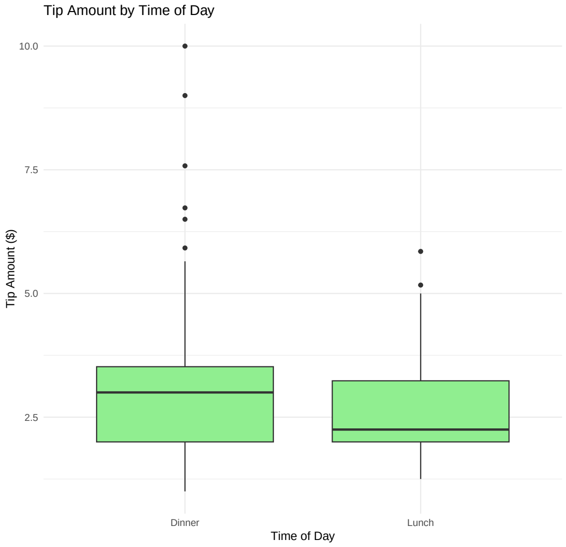


Figure 5: Illustrates that, on average, tip amounts are higher during dinner hours compared to lunchtime, indicating that customers may be more generous during dinner service.
Figure 6: Indicates that the most common tip amount falls around $5, suggesting that customers often leave tips within this range.



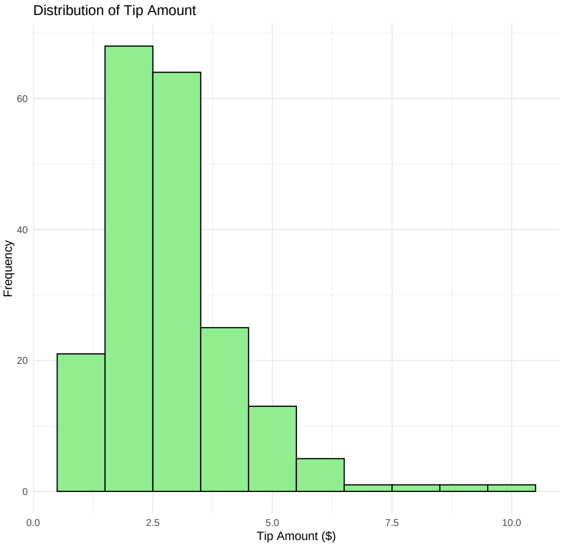
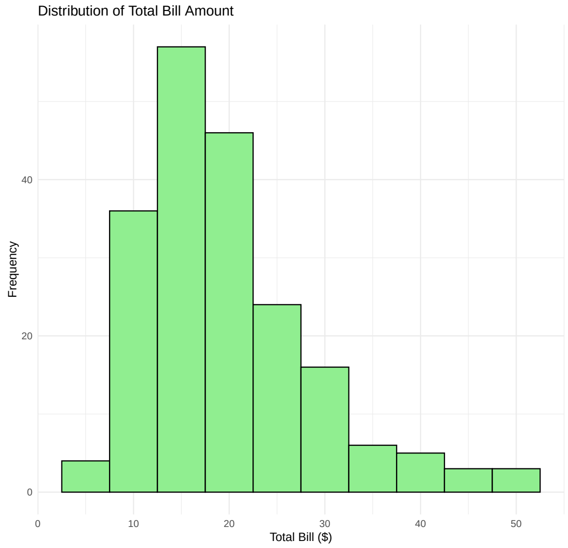
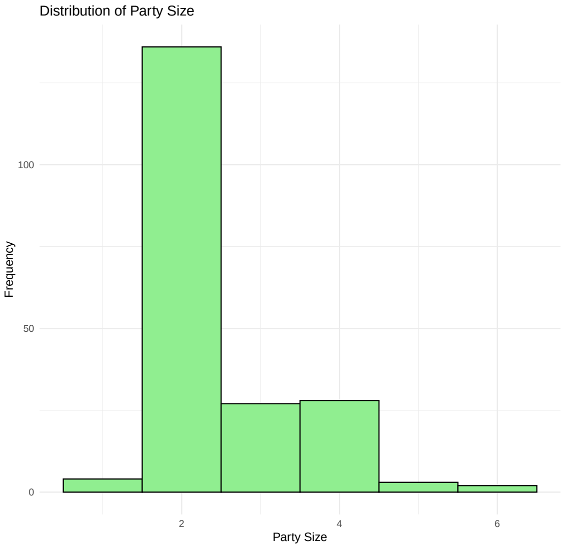


Figure 7: Shows that most total bill amounts fall within the range of $10 to $30, indicating the typical expenditure of customers at the restaurant.
Figure 8: Indicates that most parties consist of two people, suggesting that this is the most common group size among customers at the restaurant.





3.3 Outliers

Outliers were noticed in the box plot that shows tip amounts based on the bill payer's sex and whether they smoke, especially among male customers who smoke. These anomalies point to possible variation in tipping practices within this demographic subset and may call for more research into the variables affecting male smokers' tipping practices. There are also some outliers in the day of the week being Saturday, this could be because of it being month end or the restaurant might have hosted a special occasion. The outliers were not removed from the data as they did not have significant impact on the model

4 Model Building

4.1 Full Linear Model

The model above assumes that there is a linear relationship between the variables. If there is a small p- value, we can infer that there is a relationship between the predictor and the response variable. Therefore, we reject the null hypothesis (James et al., 2017, p. 68). Therefore, for the purposes of this study, the cutoffs for rejecting the null hypothesis will be 5%.

Table 7: Linear Regression Model: `tip` .

Variable	**Estimate**	**Std. Error**	**t value**	**Pr(> $\|t\|$)**
(Intercept)	0.94603	0.45196	2.093	0.0382*
total_bill	0.08959	0.01262	7.100	6.29×10^{-11}***
sexMale	-0.12259	0.19766	-0.620	0.5362
smokerYes	-0.27455	0.19944	-1.377	0.1709
daySat	0.06017	0.39598	0.152	0.8794
daySun	0.07266	0.41125	0.177	0.8600
dayThur	0.00580	0.66435	0.009	0.9930
timeLunch	0.08881	0.72482	0.123	0.9027
size	0.13122	0.12041	1.090	0.2777
Signif. codes: 0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1				
Residual standard error: 1.084 on 136 degrees of freedom Multiple R-squared: 0.4032, Adjusted R-squared: 0.3681 F-statistic: 11.48 on 8 and 136 DF, p-value: 2.291×10^{-12}				

From the model above, while keeping all other predictor variables constant, the average tip that the waiter can expect is 0.946029.

1. **Relationship between tip and total bill:**
   On average, a $1 increase in the total bill will lead to a $0.089568 increase in the tip amount received by the waiter. The p-value for the total bill is 6.29×10^{-11}, which is statistically significant at the 5% level of significance. Therefore, we can reject the null hypothesis and infer that there is some relationship between the tip amount and the total bill paid.

2. **Relationship between tip and other explanatory variables:**
   From the above summarized model output, when a customer is male, the tip amount will be $0.11854 less than when a customer is female, on average. Additionally, the tip amount will be lower when a customer smokes than when a customer does not smoke. The size of the party, on the other hand, indicates that for every one-person increase in the party size, the total tip will increase by $0.14223, on average.

3. **Model fitness:**
   - **Residual Standard Error:**
     The residual standard error is 1.084, indicating that on average, the actual tip amount and the predicted tip amount deviate by approximately 1.084. This suggests some variability in the model's predictions compared to the actual data points.

     The percentage error is calculated as:

$$\text{Percentage Error} = \left(\frac{\text{RSE}}{\text{Mean of Response Variable}} \right) \times 100$$

Substituting the values, we get:

$$\text{Percentage Error} \approx \left(\frac{1.084}{2.9561}\right) \times 100 \approx 36.66\%$$

This indicates that, on average, the predictions made by the model have an error of approximately 36.66% relative to the mean value of the response variable.

- **Multiple R-squared and Adjusted R-squared:**
  The multiple R-squared for the model is 0.4032, suggesting that 40.32% of the variability in the tip amounts can be explained by the predictor variables. The adjusted R-squared is 0.3681, indicating that 36.81% of the variability in the tip amounts can be explained by the predictor variables, accounting for the model's complexity.

- **F-Statistic:**
  The entire model is statistically significant, as indicated by the F-statistic of 11.48 and the extremely small p-value (2.291×10^{-12}), suggesting that at least one of the predictors has an impact on the tip amount that is not zero.

In conclusion, The R-squared of 0.4032 indicates that the predictors account for about 40.32% of the variability in tip amounts—the comparatively high residual standard error of 1.084 indicates that there is still a significant amount of variability in the data that cannot be explained. As a result, even though the model might offer some insights, its predictive power might be restricted.

4.2 Prediction and MSE Calculation:

Our model's test mean squared error (MSE) is 0.8598, meaning that the squared difference between the testing set's actual tip amounts and the amounts predicted by the model is 0.8598

To demonstrate how the predictions were calculated in R and to show the evaluation of the model's performance using the Mean Squared Error (MSE), we can use the following R code:

```
# a. Predictions
predicted <- predict(full_model, newdata = test_data)
summary(predicted)

# b. Mean Squared Error (MSE)
actual <- test_data$tip
mse <- mean((actual - predicted)^2)
print(paste("Test Mean Squared Error (MSE):", mse))
```

Explanation:

- In part (a), we use the `predict()` function to generate predictions using the `full_model` on the `test_data`.
- In part (b), we calculate the Mean Squared Error (MSE) by computing the mean of the squared differences between the actual tip amounts (`actual`) and the predicted tip amounts (`predicted`).
- Finally, we print out the test MSE to evaluate the model's performance on the test data. The lower the MSE value, the better the model's performance.

5 Model Improvement

As we are working with data that has a high predictors, the stepwise method is best as we can easily avoid overfitting and high variance of the coefficient estimates. We wish to have a model that has the smallest test error

5.1 Backwards stepwise

Table 8: Linear Regression Model: `tip` .

Variable	**Estimate**	**Std. Error**	**t value**	**Pr($>|t|$)**
(Intercept)	0.94603	0.45196	2.093	0.0382*
total_bill	0.08959	0.01262	7.100	6.29×10^{-11}***
sexMale	-0.12259	0.19766	-0.620	0.5362
smokerYes	-0.27455	0.19944	-1.377	0.1709
daySat	0.06017	0.39598	0.152	0.8794
daySun	0.07266	0.41125	0.177	0.8600
dayThur	0.00580	0.66435	0.009	0.9930
timeLunch	0.08881	0.72482	0.123	0.9027
size	0.13122	0.12041	1.090	0.2777
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
Residual standard error: 1.084 on 136 degrees of freedom Multiple R-squared: 0.4032, Adjusted R-squared: 0.3681 F-statistic: 11.48 on 8 and 136 DF, p-value: 2.291×10^{-12}				

comment - We start with the full model, where tip is regressed on all predictors (total_bill, sex, smoker, day, time, size). The final model only selects two predictor variables, namely total_bill and smokerYes and regress it against tip.

Model Interpretation:

- **Total bill:** On average, a $1 increase in the total bill will result in a $0.0896 increase in the tip amount received. With a p-value of 6.29e-11, we reject the null hypothesis and assume that there is a significant relationship between the total bill paid and the tip amount.
- **SmokerYes:** On average, the tip amount paid will be $0.2746 lower when a customer smokes compared to when a customer does not smoke. However, the p-value of 0.1709 is not statistically significant, and therefore, we cannot conclude that the tip amount differs when a customer smokes compared to when they do not smoke.

5.2 AIC Subset selection

Table 9: Linear Regression Model: `tip  total_bill + smoker`

Variable	**Estimate**	**Std. Error**	**t value**	**Pr(> $\|t\|$)**
(Intercept)	1.15413	0.21980	5.251	5.41×10^{-7}***
total_bill	0.09619	0.01001	9.610	$< 2 \times 10^{-16}$***
smokerYes	-0.32836	0.18520	-1.773	0.0784 .
Signif. codes: 0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1				
Residual standard error: 1.067 on 142 degrees of freedom Multiple R-squared: 0.3954, Adjusted R-squared: 0.3868 F-statistic: 46.42 on 2 and 142 DF, p-value: 3.066×10^{-16}				

Yields the same results as the Backwards Stepwise regression. Our model did not improve.

5.3 BIC Subset selection

The following is the summary of the linear regression model:

Table 10: Linear Regression Model: `tip  total_bill`

Variable	**Estimate**	**Std. Error**	**t value**	**Pr(> $\|t\|$)**
(Intercept)	1.073	0.217	4.954	2.02×10^{-6}***
total_bill	0.094	0.010	9.401	$< 2 \times 10^{-16}$***
Signif. codes: 0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1				
Residual standard error: 1.075 on 143 degrees of freedom Multiple R-squared: 0.382, Adjusted R-squared: 0.3776 F-statistic: 88.38 on 1 and 143 DF, p-value: $< 2.2 \times 10^{-16}$				

Our final model for BIC only had the total bill as the predictor variable, which is also statistically significant at a 5% level of significance with a p-value of 2×10^{-16}.

5.4 Ridge Regularization

- Lambda controls the shrinkage applied to the coefficients when fitting the model. In this model, a lambda of 0.2 was used. A lower lambda value was chosen to ensure less regularization, aiming to strike a balance between reducing overfitting and preserving important predictor variables.
- The model has a deviance of 97.9
- There are 9 parameters in the model, as indicated by the degrees of freedom (DF), representing the explanatory variables included in our model.
- Overall, the model appears to perform well, as it captures a significant amount of variability in the tip amounts.

5.5 Lasso Regularization

- A lambda of 0.2 was used, aiming to penalize less important predictors more heavily and effectively performing variable selection.

- The model has a deviance of 97.83
- The model has 1 degree of freedom, indicating that only one independent parameter is estimated. This is because the lasso model performs variable selection by shrinking coefficients towards 0.

5.6 Prediction and MSE Calculation

The Mean Squared Error (MSE) for each model is calculated as follows:

- Full Model: MSE = 0.8597
- Stepwise Model: MSE = 0.8449
- RIDGE Model: MSE = 0.0323
- LASSO Model: MSE = 0.0417

These MSE values represent the average squared difference between the predicted tip amounts and the actual tip amounts for each respective model.

In summary, the Full model has an MSE of 0.8597, indicating that, on average the the squared difference between the predicted and actual tip amounts is 0.8598.

The Stepwise Model performs slightly better with an MSE of 0.8449

The ridge model on the other hand show a big improvement in the model with an MSE on=f 0.0323, which suggests that it has significantly reduced the overfitting problem and improved the predictive theory.

The Lasso Model also has a lower MSE of 0.0417 as compared to the Full and Stepwise model

5.7 Final Model

The inclusion of the interaction term improves the explanatory power of the model. An interaction term between smoker. The choice to use this model was based on a trial and error method of testing all the interactions and choosing the one with the most significant variables.

The following is the summary of the linear regression model:

Table 11: Linear Regression Model with Interaction Term: `tip . + total_bill * smoker`

Variable	**Estimate**	**Std. Error**	**t value**	**Pr(> $\|t\|$)**
(Intercept)	-0.0551	0.4742	-0.116	0.9077
total_bill	0.1467	0.0171	8.587	1.9×10^{-14}***
sexMale	-0.0387	0.1853	-0.209	0.8347
smokerYes	1.5506	0.4372	3.547	0.0005***
daySat	0.2339	0.3713	0.630	0.5298
daySun	0.2826	0.3863	0.732	0.4657
dayThur	0.3333	0.6238	0.534	0.5940
timeLunch	0.0308	0.6763	0.046	0.9637
size	0.0004	0.1158	0.003	0.9974
total_bill:smokerYes	-0.0921	0.0200	-4.614	9.1×10^{-6}***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
Residual standard error: 1.011 on 135 degrees of freedom Multiple R-squared: 0.4844, Adjusted R-squared: 0.4501 F-statistic: 14.09 on 9 and 135 DF, p-value: 7.717×10^{-16}				

Our final model has 9 parameters. This is expected as the Ridge Model was chosen as the best method. The ridge method doesn't remove parameters, but rather, it penalizes the coefficients of the regression model by adding a penalty term to the least squares objective.

From the results above, the multiple R-squared increased to 0.4844 which is the highest of all the models tested. Total bill amount and smokerYes are statistically significant at a 5% level of significance, which is an improvement because in the other models, we only had one variable which was significant.

The final model has an MSE of 0.8856, which is higher than all the other models and all though the most of the parameters are still statistically insignificant, we can base our conclusion on this model (the best model).

6 Residual Diagnostics

Assumptions

Errors terms are assumed to be normally distributed and have the same variance at every X value(Homoscedastic). They are also assumed to be independent of each other.

From the below graphs,

The residuals are not randomly scattered around the fitted values, suggesting that the assumption of Linearity is not met. This can further mean that the relationship betweeen tip and the explanatory variables is not linear. The homoscedastic assumption seems to be met as the spread of the residuals remains constant across all predictor variables. The residuals' histogram displays a distribution that is almost symmetric, and the Q-Q plot indicates that the residuals largely follow the diagonal line. This suggests that the residuals are roughly normally distributed, indicating that the assumption of normalcy is probably met.

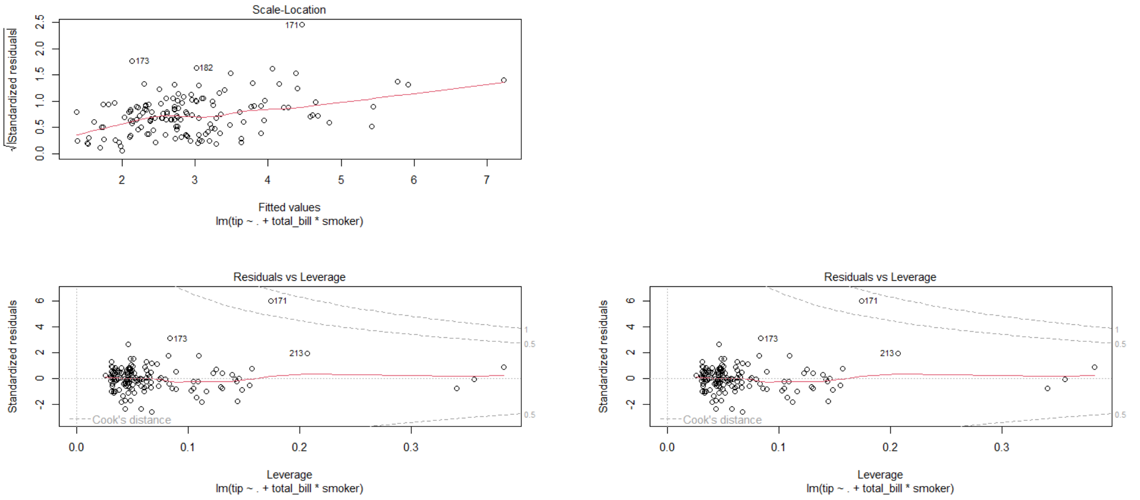


Figure 1: Graphs for Linearity

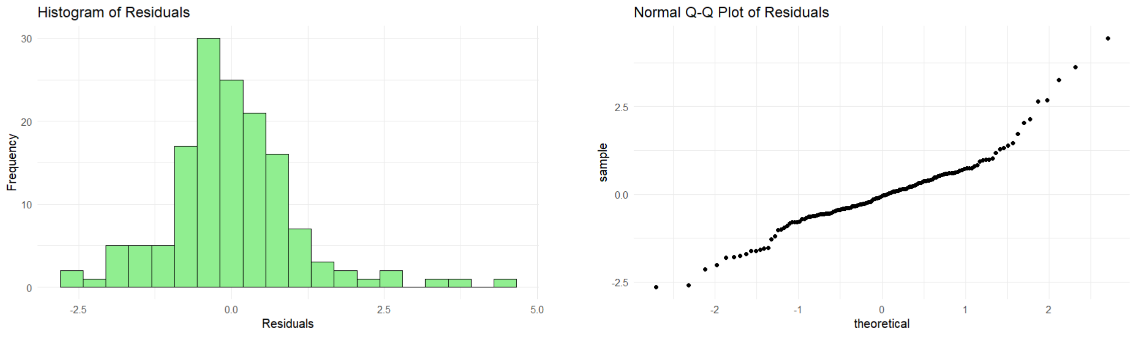


Figure 2: Graphs for Normality and Homoscedasticity

7 Conclusion

The overall paper shows that the total bill amount is the most significant explanatory variables. It is evident that as the total bill increase, the tip amount also increase. Furthermore, with the inclusion of the interaction between smoker and the total bill, our model improved significantly to include smoker as a one of the variables that are statistically significant. .

Tips seem to vary depending on the day of the week and the time of day, with bigger tips given on weekends and during dinner service. This could mean that patrons are more giving when they are having fun or going out to eat. It can also suggest that during the weekend, that is when the restaurant has more patrons.

All things considered, this study advances our knowledge of tipping behavior and provides operators of restaurants with useful information to improve both their bottom line and patron happiness. Establishments can better meet client expectations and encourage good eating experiences while also enhancing financial outcomes by acknowledging the significance of total bill amounts and temporal factors.

8 Reference list

1. James, G., Witten, D., Hastie, T., Tibshirani, R. (2017). An Introduction to Statistical Learning with Applications in R (2nd ed.). Springer.