University of Cape Town

5092

STA5076Z

Supervised Learning Assignment 01

Author:
Dibanisa Fakude

Student Number:
FKDDIB001

April 22, 2024

Contents

Authorship Declaration

I, Dibanisa Fakude, declare that:

1. This research report and the work presented in it, is my own.
2. I know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is one's own.
3. These calculations/report/plans are my own work.
4. I have not allowed and will not allow anyone to copy my work with the intention of passing it off as his or her own work.

Signature: D. Fakude

1

1.1 Introduction

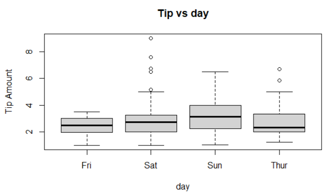
The assignment titled "Predicting Restaurant Tips using Multiple Linear Regression: A Model Comparison Approach" conducted at the University of Cape Town, Department of Statistical Sciences, aimed to analyze a dataset created by a waiter recording information about tips received over a period in a California restaurant in 1995. With 244 observations and 7 variables, including tip amount, total bill, customer's sex, smoking status, day of the week, time of day, and party size, the assignment sought to understand factors influencing tip amounts. The primary objective was to build a linear regression model predicting average tip amounts based on predictor variables. The tasks involved exploratory data analysis (EDA) to understand data distribution and relationships, model building with a full regression model, variable selection techniques like stepwise regression and regularization methods (Ridge and Lasso), and evaluating models based on mean squared error (MSE). Furthermore, the assignment included incorporating interaction terms into the final model and conducting residual diagnostics to check model assumptions. The conclusion emphasized practical implications of findings and interpretations of model outcomes, with a strict report format guideline for submission.

1.2 Data Wrangling

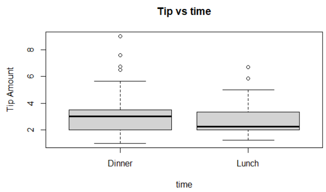
Data wrangling involved preparing the dataset for analysis by ensuring its cleanliness, consistency, and compatibility with statistical models. Initially, the full dataset, consisting of 244 observations, was read into R. To facilitate analysis, a subsample of 200 observations was randomly selected using the 'sample()' function with a specified seed to ensure reproducibility. This subsample was stored as 'my_data'. The structure of 'my_data' was examined using the 'str()' function to understand variable types and identify any inconsistencies. Categorical variables sex", "smoker" , "day" , "time" and "size" were encoded as factors using the 'mutate()' function from the 'tidyverse' package. Subsequently, a summary of the dataset was generated to gain insights into variable distributions and identify potential outliers or missing values.

1.3 Question 1

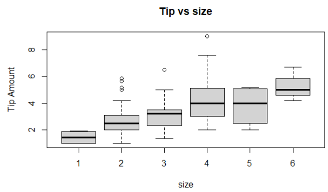
Exploratory data analysis (EDA) was conducted to visualize relationships between predictor variables (total_bill, sex, smoker, day, time, and size) and the target variable (tip amount). A custom function was created to generate scatter plots for each predictor against the tip amount, allowing for visual inspection of potential associations. The plots were then generated and examined to identify any trends or patterns. This process provided an initial understanding of the data and informed subsequent model building and analysis.



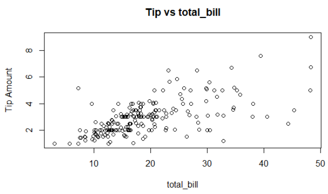
(a) Tips and the day



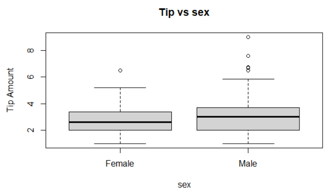
(b) Tips versus the time of the day



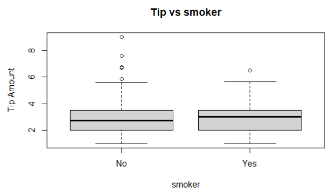
(c) Tips vs the size of the party



(d) Tips vs total bill relationship



(e) Tips vs the gender of the customers



(f) Tips vs the customer is a smoker or not

Figure 1: Visualizing relationships between predictors and the target variable

The provided visualizations offer a comprehensive exploration of the relationships between various predictors and the target variable, likely tip amounts. Each visualization encapsulates distinct insights into tipping behavior. The box plot depicting "Tips vs Day" illustrates the distribution of tips across different days, revealing outliers on Saturday and Thursday, and a notably higher median tip on Sundays. In "Tips vs Time," the box plot discerns higher median tips during dinner compared to lunch. "Tips vs Size of Party" presents a boxplot showcasing a peculiar pattern with tip amounts increasing as the size of the party increases ,with size 2,3,4 having outliers in tip amounts. The scatter plot "Tips vs Total Bill" suggests a positive relationship between tip amounts and total bills, with clustered data points between 10 and 20 and occasional outliers. "Tips vs Gender" reveals through box plots that male customers tend to give higher tips, with notable outliers among males. Finally, "Tips vs Smoker" demonstrates similar tip distributions between smokers and non-smokers, albeit with more outliers among non-smokers and a slightly higher mean tip for smokers. These visualizations collectively provide valuable insights into tipping behaviors across different variables, offering avenues for further statistical analysis and interpretation.

2 Model Building

2.1 Fitting a Full Model

I have created a simple linear regression analysis using the dataset my_data. Initially, the first column "X" doesn't have a significant relationship with the response variable,so it is excluded from the dataset. Using the caTools library, the data is split into training and testing sets using a random sampling technique called "sample.split()". The training set, x_train, contains 80% of the data, while the remaining 20% is stored in the test set, x_test.

Subsequently, a linear regression model, regressor, is created using the lm() function, where the target variable is "tip" and all other variables in the dataset are used as predictors. Below is the output summary of the model

```
Call:
lm(formula = tip ~ ., data = x_train)

Residuals:
    Min      1Q  Median      3Q     Max
-2.2328 -0.5010 -0.0846  0.4664  3.2196

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.75747    0.67005   1.130   0.2601
total_bill   0.07875    0.01169   6.739 3.38e-10 ***
sexMale     -0.06443    0.16267  -0.396   0.6926
smokerYes   -0.02642    0.16991  -0.155   0.8766
daySat      -0.01127    0.39100  -0.029   0.9770
daySun       0.13220    0.40294   0.328   0.7433
dayThur     -0.20670    0.42100  -0.491   0.6242
timeLunch    0.20748    0.49208   0.422   0.6739
size2        0.56066    0.56846   0.986   0.3256
size3        0.52424    0.60589   0.865   0.3883
size4        1.17169    0.64272   1.823   0.0703 .
size5        0.84376    0.76248   1.107   0.2703
size6        1.72315    0.80453   2.142   0.0339 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9374 on 147 degrees of freedom
Multiple R-squared:  0.4723,Adjusted R-squared:  0.4292
```

```
F-statistic: 10.96 on 12 and 147 DF,  p-value: 2.181e-15)
```

In this output, 'total_bill' and 'size' appear to be statistically significant predictors of tip amounts, as indicated by their low p-values. The other predictor variables ('sex', 'smoker', 'day', 'time') do not appear to be statistically significant based on their p-values.

2.2 Testing Model

After creating the linear regression model, it was then utilized to predict tip amounts for new unseen data. This process involved applying the model to a separate dataset ('x_test'), resulting in the generation of predicted tip amounts for the observations within this dataset. After conducting model testing, the next step involved computing the Mean Squared Error (MSE) of the linear regression model. The MSE provides a measure of the average squared difference between the actual tip amounts in the test dataset (x_test$tip) and the predicted tip amounts (y_pred).

The computed MSE value obtained from this model is 1.479537. This result indicates that, on average, the squared difference between the actual and predicted tip amounts is approximately 1.48. A lower MSE value suggests better accuracy of the model predictions, while a higher MSE value indicates poorer performance.

3 Model Improvement

After testing the initial full linear model, further improvements were sought through variable selection techniques, including Backwards Stepwise Regression based on AIC and BIC criteria, RIDGE regularization, and LASSO regularization. The Backwards Stepwise Regression technique was applied to select subsets of predictors based on AIC and BIC. The resulting models were then evaluated for their Mean Squared Error (MSE) on the test data.

3.1 Stepwise Regression

The interpretation of the AIC model reveals that the "total_bill" variable exhibits a noteworthy positive influence on the "tip" amount, implying that as the total bill escalates, tips tend to increase correspondingly. Furthermore, the variable "size," denoting the size of the dining party, demonstrates significant effects for specific categories, such as size6, indicating that larger party sizes may lead to higher tip amounts.. The significance tests (t-tests) reveal that predictors total bill and size6 party are statistically significant at the 0.05 level, indicating that they have a significant impact on tip amounts. The adjusted R-squared value of 0.45 suggests that this model explains 45% of the variability in tip amounts, which is a moderate level of explanatory power. Additionally, the mean squared error (MSE) for this model is 0.81334, indicating that, on average, the model's predictions are off by approximately 0.81334 units squared. Overall, the stepwise regression approach resulted in a simplified model that retains two significant predictors, total_bill and size, while still achieving a much bettwer level of predictive accuracy compared to the full model. Below is the summary of the retained parameters.

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

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.74224    0.53763   1.381   0.1694
total_bill   0.07863    0.01094   7.189  2.7e-11 ***
size2        0.56613    0.54918   1.031   0.3042
size3        0.56315    0.58834   0.957   0.3400
size4        1.18601    0.61762   1.920   0.0567 .
size5        0.89713    0.73918   1.214   0.2267
size6        1.74405    0.76822   2.270   0.0246 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

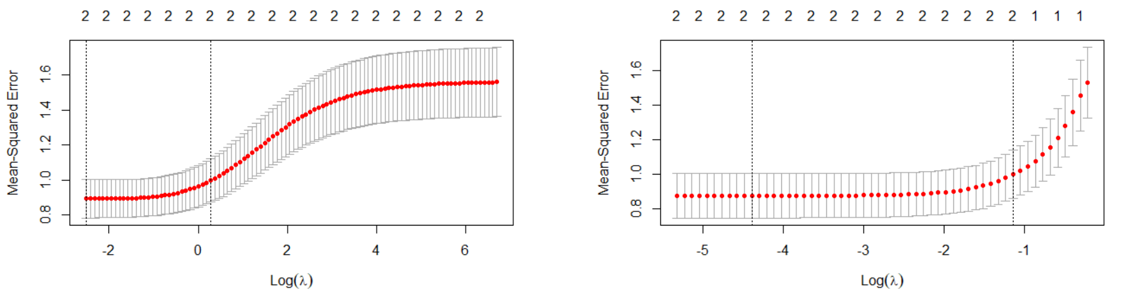
```
Residual standard error: 0.9223 on 153 degrees of freedom
Multiple R-squared:  0.4684,Adjusted R-squared:  0.4475
F-statistic: 22.47 on 6 and 153 DF,  p-value: < 2.2e-16

"MSE for model selected by AIC: 0.813342668228686"
```

Secondly the BIC regression model excluded all the other predictors and predicted tip amounts based solely on the total_bill predictor using the training dataset, and yielded insightful results. The model indicated that for every one-unit increase in the total bill, the estimated tip amount increases by approximately 0.096 units, with a statistically significant intercept of around 1.072. Both coefficients show strong significance, supported by extremely low p-values. The model exhibits a reasonable fit, explaining approximately 43.23% of the variability in tip amounts, as evidenced by the multiple R-squared value. The BIC had a mean square error of about 0.869 which is slighty higher than the AIC but less than the full model. Below is the summary of the BIC

```
Call:
lm(formula = tip ~ total_bill, data = x_train)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8357 -0.5398 -0.0552  0.4186  3.3802

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.072021   0.181567   5.904  2.1e-08 ***
total_bill  0.096245   0.008775  10.969  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9379 on 158 degrees of freedom
Multiple R-squared:  0.4323, Adjusted R-squared:  0.4287
F-statistic: 120.3 on 1 and 158 DF,  p-value: < 2.2e-16

"MSE for model selected by BIC: 0.868570874510884"
```

3.2 Ridge and Lasso

Before fitting the Ridge and Lasso regularization models, a preliminary step involved performing 10-fold cross-validation to determine the optimal values of the regularization parameter (lambda). For Ridge regression, the optimal lambda was found to be approximately 0.081 , while for Lasso regression, it was approximately 0.012. figure2(a) and 2(b) below shows the cross-validation plots, the dashed vertical lines on either side of the solid line represent the standard deviation of the mean error and the width between the lines indicates the variability or uncertainty in the estimated error ,notable figure 2(b) has a larger variability . These optimal lambda values were obtained using the 'glmnet' package in R. Following this, the Ridge and Lasso models were fitted using the identified optimal lambda values. The Ridge model, with a lambda of 0.08, achieved a percentage deviance explained of 45.58% , while the Lasso model, with a lambda of 0.02, achieved a slightly higher percentage deviance explained of 45.69% with 2 degrees of freedom, meaning only 2 effective number of parameters in both of these models were left.These findings suggest that both regularization techniques were able to effectively penalize coefficients to prevent overfitting and improve model performance.



(a) Cross Validation for Ridge Regularization model

(b) Cross Validation for Lasso Regularization model

Figure 2: Finding the optimal lambda for the regularization models

3.3 Prediction and MSE

After determining the optimal lambda values and fitting the Ridge and Lasso regression models on the training data, the next step involved applying these models to the testing dataset. To achieve this, the testing dataset was converted into a matrix format, excluding the response variable. Subsequently, the testing data was fitted to both the Ridge and Lasso regressors using the 'predict()' function. The predictions generated by each model were stored as 'ridge_pred' and 'lasso_pred', respectively.

Following this, the mean squared error (MSE) was calculated for each model to evaluate their predictive performance on the testing dataset. The MSE for the Ridge model was found to be 1.504276, while the MSE for the Lasso model was slightly higher at 1.50661. These findings provide insights into the accuracy of the Ridge and Lasso models in predicting tip amounts on unseen data, with the Ridge model demonstrating slightly better performance compared to Lasso but worse compared to Full, AIC,and BIC models in terms of MSE.

3.4 Comments on Models Perfomance

Starting with the Full model, which incorporated all predictors, it exhibited an MSE of 1.479537. Despite including multiple predictors, only 'total_bill' and 'size6' were found to be statistically significant in predicting tip amounts, suggesting potential overfitting and inefficiency in utilizing all variables.

In contrast, the Stepwise regression model, which selected a subset of predictors based on AIC and BIC criteria, retained only 'total_bill' and 'size'and achieving a a very low MSE of 0.813 and 0.869, this model offered a more better solution, emphasizing the significance of 'total_bill' and 'size' in predicting tips.

Following this, Ridge and Lasso regularization techniques were employed to prevent overfitting and improve model generalization. After performing 10-fold cross-validation to determine optimal lambda values, Ridge regression with lambda = 0.08 achieved a deviance explained of 45.58%, while Lasso regression with lambda = 0.02 achieved a slightly higher deviance explained of 45.66%. Despite these promising results, both Ridge and Lasso models exhibited higher MSE values compared to the Full and Stepwise models, with Ridge yielding an MSE of 1.504276 and Lasso an MSE of 1.50661.

Ultimately, considering the MSE as the primary criterion for model selection, the Stepwise regression model specifically the AIC appeared to be the most suitable choice among the evaluated models, as it achieved a relatively lower MSE while maintaining a simpler model structure with only two predictors ('total_bill' and 'size'). This model strikes a balance between predictive accuracy and model complexity, making it the preferred option for predicting tip amounts in this context. Table below shows the comparison of all the models MSE.

3.5 Best Model Interaction

The inclusion of interaction terms in a linear regression model allows for the exploration of potential relationships between predictors that may not be adequately captured by considering them individually. In my model, I've included interaction

Table 1: Mean Squared Error (MSE) comparison

Model	MSE
Full Model	1.4795370
Stepwise AIC	0.8133427
Stepwise BIC	0.8685709
LASSO	1.5066103
RIDGE	1.5042761

terms such as total_bill*smoker, which captures the interaction between the total bill amount and the smoking status of the customer.

```
best_model_aic <- lm(formula = tip ~ total_bill*smoker + size, data = x_train)
```

The choice of interaction is based on reasoning, it's reasonable to hypothesize that the effect of the total bill on the tip amount might vary depending on whether the customer is a smoker or not.

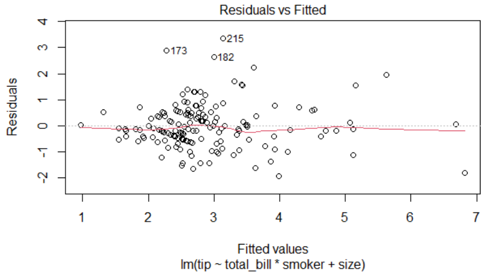
3.6 Prediction and MSE Calculation

After identifying the best model, which in this case is the Stepwise regression model selected based on AIC criteria, predictions were obtained for the test data using this model. The predict() function was utilized to generate predictions (best_predictions_aic) based on the best model (best_model_aic) and the test dataset (x_test). These predictions represent the estimated tip amounts for each observation in the test dataset.

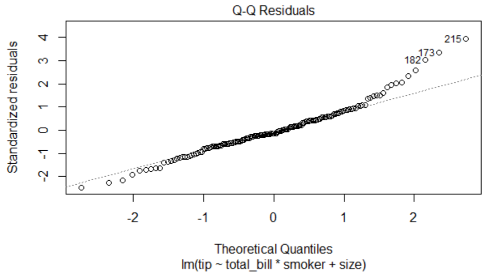
Following the prediction step, the Mean Squared Error (MSE) was calculated for the model with interaction terms included. An interaction term captures the combined effect of two predictor variables on the response variable, offering a more nuanced understanding of the relationship between predictors and the outcome.

However the computed MSE value for the model with interaction terms was found to be 1.245, which is a which is worse that the model without interaction, which might potential means the interaction inclusion resulted in overfitting or noise amplification in the model.

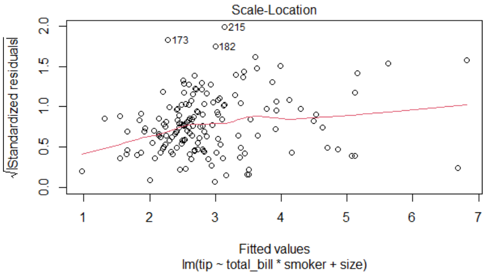
4 Residual Diagnostics



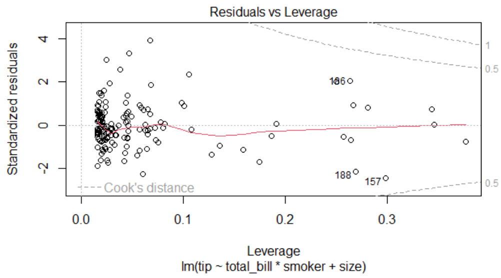
(a) Scatterplot of residuals vs fitted values



(b) Q-Q Plot (Quantile-Quantile Plot)



(c) Scale-Location (or Spread-Location) Plot



(d) Residuals vs. Leverage Plot

Figure 3: Residuals plots

Based on the interpretation of the residual plots provided, several insights can be gleaned about the regression model's performance and the characteristics of the data. Firstly, in the Residuals vs. Fitted plot, the spread of residuals around the line at zero suggests that the assumption of homoscedasticity is generally met. However, the clustering of residuals between 2 and 3 on the x-axis, with more dispersion as you move further away, indicates a potential violation of homoscedasticity. This suggests that the variability of the residuals might not be consistent across all levels of the fitted values, potentially indicating heteroscedasticity.

Secondly, in the Q-Q plot, the majority of the points closely follow the diagonal line, suggesting that the residuals are approximately normally distributed. However, the deviation of points in the far upper right from the line, where they diverge and move upwards, indicates potential departures from normality in the extreme upper tail of the distribution.Moving to the Scale-Location plot, the non-horizontal line suggests a lack of constant variance in the residuals, which is indicative of heteroscedasticity.

The upward angle of the line indicates increasing variability in residuals as the fitted values increase, further supporting the presence of heteroscedasticity.

Finally, in the Residuals vs. Leverage plot, the clustering of points within the four dotted lines suggests that influential observations are not overly influencing the regression model. However, the clustering of points in the far left of the plot suggests a concentration of observations with potentially higher leverage values.

In summary, while the residual plots provide some indication of adherence to regression assumptions such as homoscedasticity and normality, there are also indications of potential violations, particularly in terms of heteroscedasticity and deviations from normality in the upper tail of the distribution. These findings suggest potential areas for further investigation and refinement of the regression model to improve its reliability and validity.

5 Question 5

5.1 practical implications

Factors Contributing to Higher Tips: Through the analysis, it was observed that the total bill amount and the size of the dining party were significant predictors of tip amounts. Specifically, as the total bill increases, tip amounts tend to increase as well. Additionally, larger party sizes were associated with higher tip amounts, especially for parties of size 4 and above. These findings suggest that providing excellent service to larger groups and ensuring customer satisfaction with their dining experience, particularly in situations where the bill is higher, can lead to higher tips for waitstaff.

Gender and Smoking Status: Interestingly, gender and smoking status did not exhibit significant effects on tip amounts in the analyzed dataset. This suggests that factors such as the quality of service provided, the overall dining experience, and the size of the bill may have a more substantial influence on tipping behavior compared to demographic factors like gender or smoking status.

Time of Day and Day of the Week: The analysis also explored the impact of the time of day and the day of the week on tip amounts. While there were some variations in tip amounts across different times of the day and days of the week, these factors did not emerge as significant predictors in the final model. However, further investigation into the specific dynamics of tipping behavior during peak hours or on weekends could provide additional insights for restaurant management.

Model Interpretation and Application: The stepwise regression model selected based on AIC criteria emerged as the most suitable model for predicting tip amounts, offering a balance between predictive accuracy and model simplicity. This model highlighted the importance of considering only essential predictors, such as the total bill amount and the size of the dining party, for accurately estimating tip amounts. By focusing on these key factors, restaurants can better tailor their service strategies to maximize tip earnings while maintaining operational efficiency.

5.2 interaction Impact

The inclusion of interaction terms in the regression model allowed for the exploration of nuanced relationships between predictors. However, the model with interaction terms did not significantly improve predictive performance compared to the model without interactions. This suggests that while interactions may exist between certain predictors, they may not have a substantial impact on overall tipping behavior in this context.

6 Appendix

6.1 Question 2-Code

````
```{r}
library(caTools)
#Exluding the first column
datasets<- my_data[-1]

#Splitting the datastes
split<-sample.split(datasets$tip,SplitRatio = 0.8)

#Subsetting the data acording to the split ratio
x_train<- subset(datasets,split == TRUE)
x_test<- subset(datasets,split == FALSE)

#Creating the regressor for the datasets
regressor<- lm(formula =tip ~ ., data = x_train )
#Checking the models performance
summary(regressor)
```


```{r}
#Predicting tip amount using the created model
y_pred<- predict(regressor,newdata = x_test)
y_pred
```

```{r}
#Computing the MSE of the model
# Calculate Mean Squared Error (MSE) of the lm model
mse <- mean((x_test$tip - y_pred)^2)
mse
```
````

6.2 Question 3-Code

```{r}
# Perform Backwards Stepwise Regression based on AIC
backward_aic <- step(regressor, direction = "backward", k = 2, trace = 0)
# Print summary of models
summary(backward_aic)
```

```{r}
# Perform Backwards Stepwise Regression based on BIC
r<-log(nrow(x_train)) # for BIC
backward_bic <- step(regressor, direction = "backward", k = r, trace = 0)

summary(backward_bic)
```

```{r}
# Predicted values
predicted <- predict(backward_aic, newdata = x_train)

# Calculate squared errors
squared_errors <- (x_train$tip - predicted)^2

# Calculate MSE
mse_AIC <- mean(squared_errors)

# Print the MSE
print(paste("MSE for model selected by AIC:", mse_AIC))
```

```{r}
# Predicted values
predicted_bic <- predict(backward_bic, newdata = x_train)

# Calculate squared errors
squared_errors_bic <- (x_train$tip - predicted_bic)^2
```

```
# Calculate MSE
mse_bic <- mean(squared_errors_bic)

# Print the MSE
print(paste("MSE for model selected by BIC:", mse_bic))
```

```{r}
# Load the glmnet package (if not already installed)
# install.packages("glmnet")
library(glmnet)

# Convert the data to matrix format
x_train_matrix <- as.matrix(x_train[, -2])  # Exclude the target variable
y_train_vector <- x_train$tip

# Perform cross-validation for RIDGE
cv_ridge <- cv.glmnet(x_train_matrix, y_train_vector, alpha = 0)

# Optimal lambda value for RIDGE
opt_lambda_ridge <- cv_ridge$lambda.min
print(paste("Optimal lambda for RIDGE:", opt_lambda_ridge))

# Perform cross-validation for LASSO
cv_lasso <- cv.glmnet(x_train_matrix, y_train_vector, alpha = 1)

# Optimal lambda value for LASSO
opt_lambda_lasso <- cv_lasso$lambda.min
print(paste("Optimal lambda for LASSO:", opt_lambda_lasso))
```

```{r}
#Plotting the cross validation graphs
plot(cv_lasso)
plot(cv_ridge)
```

```
```{r}
# Perform ridge regression
ridge_model <- glmnet(x_train_matrix, y_train_vector, alpha = 0, lambda = 0.08)

# Print summary of the ridge model
print(ridge_model)
```
```

```
```{r}
# Perform ridge regression
lasso_model <- glmnet(x_train_matrix, y_train_vector, alpha = 1, lambda = 0.01)

# Print summary of the lasso model
print(lasso_model)
```
```

```
```{r}
#converting the datasets to a matrix
x_test_matrix<- as.matrix(x_test[,-2])
#fitting ot to the ridge regressor
ridge_pred<- predict(ridge_model,newx = x_test_matrix)
ridge_pred
#Fitting on the lasso regression
lasso_pred<-predict(lasso_model, newx = x_test_matrix)
lasso_pred
```
```

```
```{r}
#Calculation the mean square error for the ridge regression
# Calculate Mean Squared Error (MSE)
mse_ridge <- mean((x_test$tip - ridge_pred)^2)
mse_ridge

#Calculation the mean square error for the lasso regression
# Calculate Mean Squared Error (MSE)
mse_lasso <- mean((x_test$tip - lasso_pred)^2)
```
```

````
mse_lasso
```

```{r}
#Calculating the models MSE
# Fit the selected models to the training data
model_AIC <- lm(formula = tip ~ total_bill*smoker + size, data = x_train)
```


```{r}
# Obtain predictions from the models on the test data
best_predictions_aic <- predict(model_AIC, newdata = x_test)
best_predictions_aic
```


```{r}
# Extract the actual values of the target variable from the test data
best_actual_values <- x_test$tip

# Calculate squared differences and MSE for the model selected by AIC
best_mse_aic <- mean((best_actual_values - best_predictions_aic)^2)


# Print the MSE for each model
print(paste("MSE for model selected by AIC(stepwise):", best_mse_aic))
```


```{r}
#Plotting the residuals plot
plot(best_model_aic)
```
````
````