STA5076X - Supervised Learning Assignment 1

ALEX MIRUGWE - MRGALE005

07-June-2020

Contents

Declaration

Department of Statistical Sciences Plagiarism Declaration form

A copy of this form, completed and signed, to be attached to all coursework submissions to the Statistical Sciences Department. Submissions without this form will not be marked.

COURSE CODE: **STA5076Z**

COURSE NAME: **SUPERVISED LEARNING**

STUDENT NAME: **Alex Mirugwe**

STUDENT NUMBER: **MRGALE005**

TUTORS NAME: **Juwa Nyirenda**

PLAGIARISM DECLARATION

1. I know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is one's own.
2. I have used a generally accepted citation and referencing style. Each contribution to, and quotation in, this tutorial/report/project from the work(s) of other people has been attributed, and has been cited and referenced.
3. This tutorial/report/project is 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.
5. I acknowledge that copying someone else's assignment or essay, or part of it, is wrong, and declare that this is my own work.

Note that agreement to this statement does not exonerate you from the University's plagiarism rules (http://www.uct.ac.za/uct/policies/plagiarism_students.pdf).

Signature: Mirugwe Date: June 8, 2020

1 Abstract

The goal of the exercise is to build a linear model for predicting the average amount of tip in dollars a waiter is expected to earn from the restaurant given the predictor variables i.e. total bill paid, day, the gender of the customer (sex) time of the party, smoker, and size of the party. And this was achieved through the use of the Linear Regression method.

The dataset of 200 observations and 7 variables was divided into training and testing sets in a ratio of 8:2 respectively. The model was fitted using the *lm()* function of R on the train set and tested on the testing set using *predict()* function. And the model fitness was deeply analyzed to understand how well it fits the data.

Using Lasso regularization approach, the model was improved and this helped to identify the most important predictors in estimating the amount of tip received by the waiter. And also an interaction of size and smoker was included in the final model which greatly improved its data fitness.

2 Introduction

2.1 Dataset Description

The model was fitted using the dataset of 200 observations and 7 variables. Of the 7 variables, 3 numerical were variables and these are total bill, tip amount, and size of the party. 4 were categorical variables and that is sex(with 2 levels Female and Male), smoker (Yes or No), day(Thursday, Friday, Saturday, and Sunday), and time (Lunch and Dinner).

Table 1: Of the 200 bill payers, 66 were female 134 were male. This means 67% of the total bill payers were male and 33% were females.

Female	**Male**
66(33%)	134(67%)

Table 2: The table shows that 123 Customers who attended were not smoking and 77 were not. Most of the restaurant's customers were not smoking.

Smoker	**Non-Smoker**
77(38.5%)	123(61.5%)

Table 3: Most of the restaurant parties were organized on Saturdays followed by Sunday, then Thursday and Friday had the fewest.

Thursday	**Friday**	**Saturday**	**Sunday**
45(22.5%)	14(7.0%)	77(38.8%)	64(32.0%)

Table 4: The Restaurant's parties mostly took place during lunchtime than dinner.

Lunch	Dinner
151(75.5%)	49(24.5%)

Table 5: The table shows that the $20.06 was the average total bill paid and $3.07 and $50.81 were the lowest and highest total bill respectively that the restaurant received. And on average a waiter received a tip of $3.065. The lowest tip that was given to the waiter is $1 and the highest was $10.

Totalbill($)	Tip ($)
Mean :20.06	Mean :3.065
Min. : 3.07	Min. :1.000
Max. :50.81	Max. :10.00
Median :17.8	Median :2.960

And also because the mean total bill is greater than the median bill, this implies that the distribution of the total bill is right-skewed also confirmed by the histogram below.

2.2 Graphical relationship between response and predictors

2.2.1 Tip Amount Vs Total Bill paid

The figure shows that there is some positive linear relationship between the total bill paid by the restaurant customers and the tip amount given to waiters. A more interesting finding from the graph above is the fact more customers paid smaller bills (most total bills are below $25) and also most waiters received lower tips.

In the graphs above,

Graph A shows that men attendance was quite high compared to that of females on every day of the week. Over the days of the week, non-smokers were many apart from only Friday which had more of the smokers than the non-smokers and this is demonstrated by graph B. Graph C shows that they were no parties organized during lunchtimes on Sunday and Saturday. On Thursday parties were organized more during lunchtime than the Dinner time and Friday Dinner had more parties.

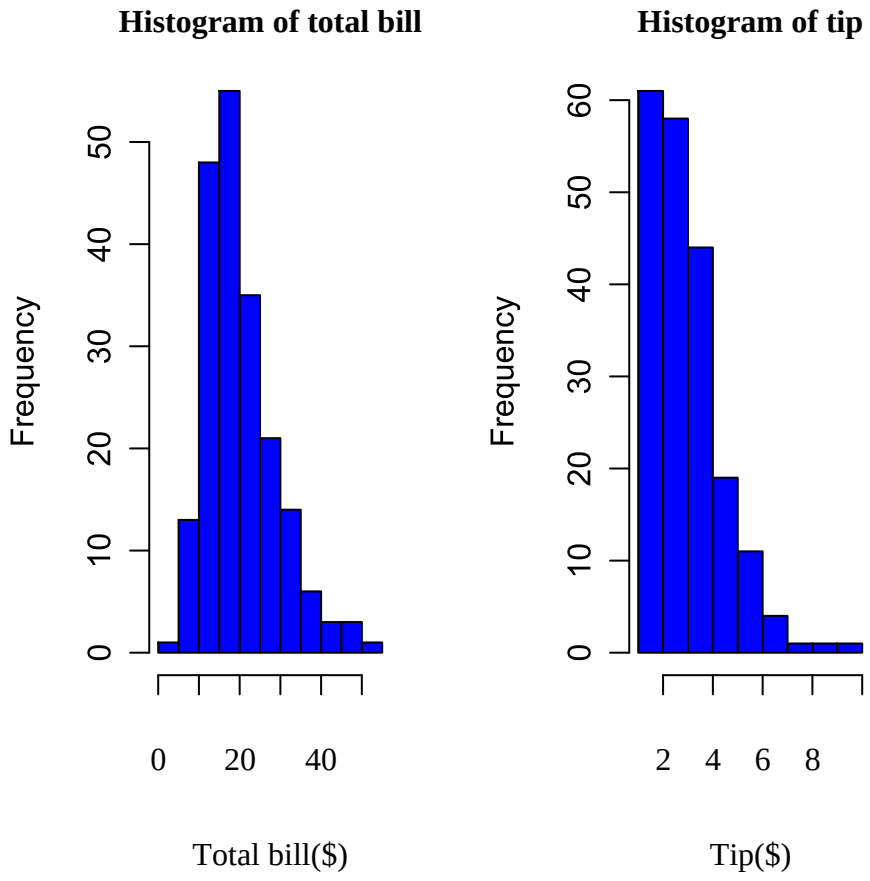


Figure 1: The figures show that majority of the restaurant customers paid a bills below $25 and also most waiters received tip amount less than $4.

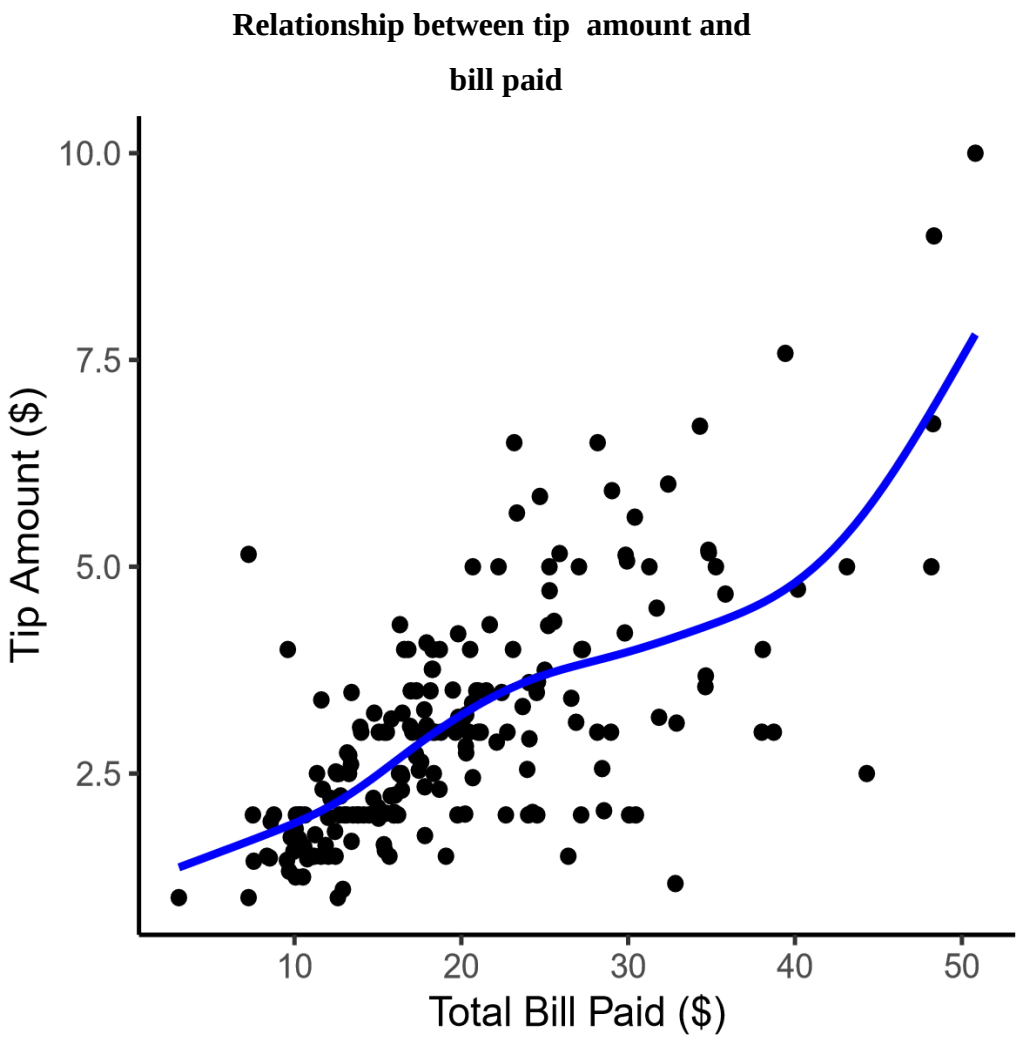


Figure 2: Relationship between the tip received by the waiter and the total bill paid by the customer

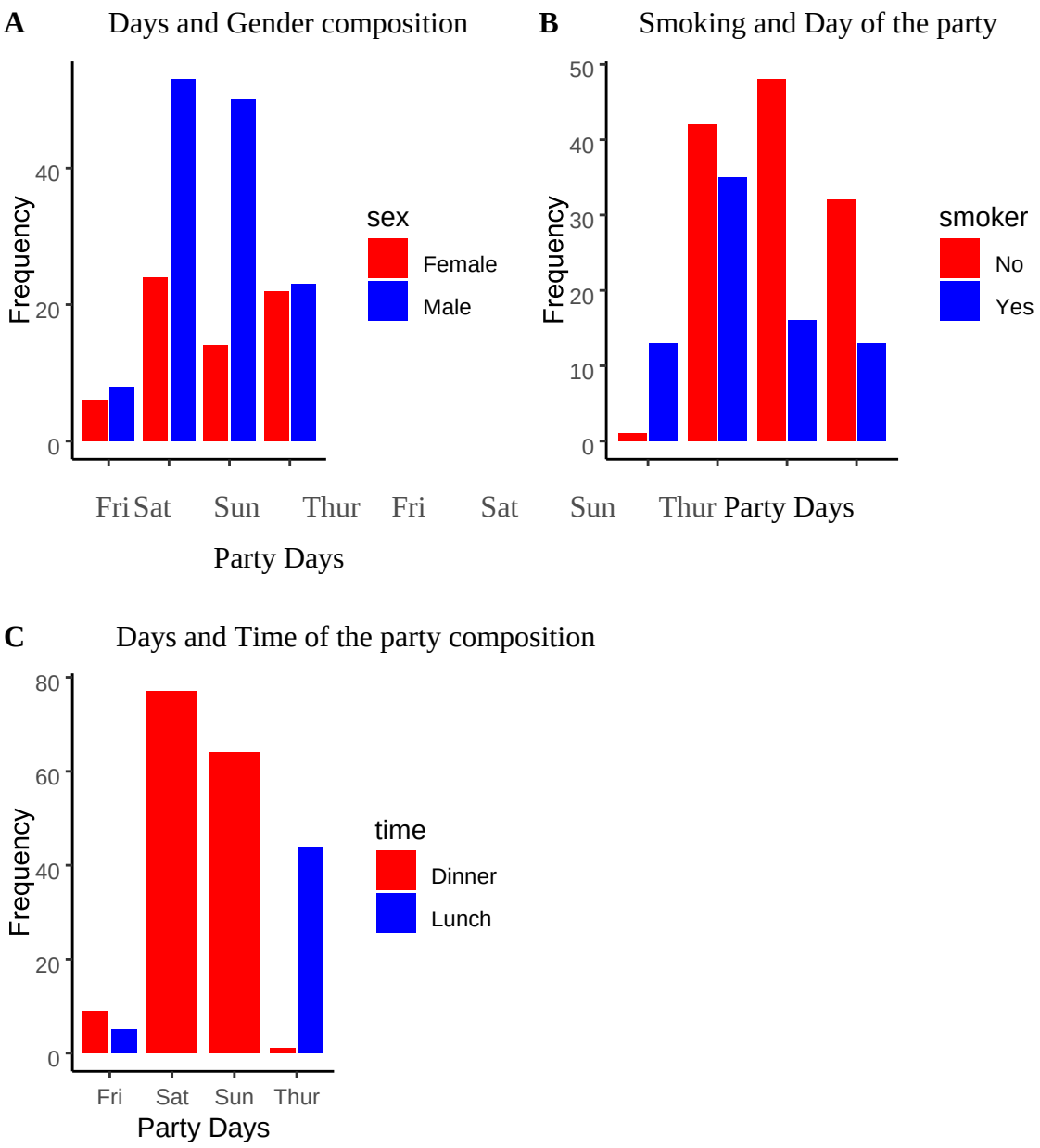


Figure 3: The figures above show realtionship between the day of the party and different variable like sex,time of the party and the smoking status of the customer

3 Linear Model

3.1 Model predictors' significance

The model obtained from the regression analysis is;

Tip_amount = 0.77083 + 0.09192total_bill - 0.05458sex + 0.15709smoker 0.10900time + 0.15676size- 0.21973daySat + 0.01615daySun + 0.16276dayThur

The restaurant waiter receives $0.77083 as a tip whenever a null hypothesis for all the predictors (total_bill, sex, smoker, time, size, and day) cannot be rejected. This means when there is no relationship between the tip and all the predictors. The level of statistical significance of a predictor is often expressed as a p-value and it's between 0 and 1. The smaller the p-value, the stronger the evidence that you should reject the null hypothesis[1]. If the p-value is less than 0.05, then it's statistically significant. It indicates strong evidence against the null hypothesis, as there is less than a 5% probability the null is correct or at least 95% confidence interval in the predictor. Therefore, we reject the null hypothesis and accept the alternative hypothesis.

Total bill is the most statistically significant predictor in all the variables with level three significance and 99.99% confidence interval. For other variables is quite hard to reject the Null hypothesis due to high p-values. Predictors like SmokerMale and Size also have relatively small p-values though still greater than the significance level of 0.05. And for the smoke predictor, the model shows statistical evidence of a difference in the average tip amount received by the waiter based on the smoking status of the customer.

On the other hand, Sex, Time, and day predictors have extremely large p-values and because of that, their null hypothesis cannot be rejected. The impact of these variables on the tip received by the restaurant waiters seems negligible. A change in these three variables literally don't impact on the tip amount. The model shows no statistical evidence of a difference in average tip received by the waiter between the days of the week but there exists some statistical evidence in the average tip given to the waiter based on the time of the party and the gender of who is paying the bill.

Finally, we can conclude that the total bill is the most important tip predictor in the model followed by the size of the party.

[1] A situation where there is no relationship between the predictor and response.

Table 6: *Coefficients and p-values of the model predictors*

Predictors	**Coefficient Estimate**	**p-values**
total bill	0.09192	8.87e-11
SexMale	-0.05458	0.768
SmorkerYes	0.15709	0.436
TimeLunch	-0.10900	0.846
Size	0.15676	0.155
daySat	-0.21973	0.609
daySun	0.01615	0.972
dayThur	0.16276	0.752

Table 7: *The table below shows that a waiter receives a tip amount of $0.77083 whenever the null hypothesis of all predictors cannot be rejected*

Intercept
0.77083

The estimated predictor coefficients shown in the table indicate the increase in the tip amount, for example, a unit dollar increase on the total bill, increases the tip received by the waiter by $0.09191 keeping other predictors constant. Similarly, the size of the party increases the tip amount by $0.15676 whenever an additional person attends the party.

It's noticed from the table above that although the model was fitted on six (6) predictors, the model expression has eight (8) coefficients. This has happened because of the factors variables (sex, day, time, and smoker). The linear model has applied a dummy[2] code to each factor predictor. For example, the smoker variable has two categories, Yes and No. The model has split this into smokers and smokers. The model as assigned 0 whenever the customer's smoking status is No and 1 for smokers. Sex has been split into sexMale and sexFemale and model assigned sexFemale with 0 and sexMale with 1. Similarly, the day variable is also split into dayLunch and dayDinner with

[2] A dummy variable is a variable that represents a categorical variable.

assigned numerical values of 1 and 0 respectively. The four-level day predictor has been split into dayThur, dayFri, daySat, and daySun.

Therefore we can say that if a bill is paid by a male customer, the waiter's tip decreases by $0.05458 and it neither increases nor decreases when a bill is paid by a female customer. And the amount of $0.15709 for every additional increase on the smoking customers and an additional increase in the non-smoking customers, the tip amount doesn't change. And whenever the bill is paid by a male customer, the waiter's tip decrease by $0.05458 than what he/she gets when the bill paid by a female customer.

The time of the party also affects the average amount of tip received by the waiter in the restaurant whereby the tip decreases by $0.10900 whenever an additional party is organized during lunchtime and does not affect the tip amount received by the waiter for parties organized during lunchtime.

The overall effect of the day on the tip received by the waiter is ($0.01615+$0.16276-$0.21973) = -$0.04082. Though individual days on which the party is organized affect the tip amount independently. A party on Sunday increases the tip amount by $0.01615 and $0.16276 on Thursdays but decreases by $0.21973 for a unit increase in Saturday numbers. The model also that Friday seems not to affect the tip amount.

3.2 Model fitness

Table 8: *More details on the least-squares model for the regression of the tip amount received by the waiter on the different model predictors.*

Quantity	**Value**
Residual standard error	1.042
*R*2	0.5248
F-statistic	23.88
Adjusted R^2	0.5028
Overall p-value	0.5028

It's observed from the fitted model output which is summarised in table 7 above that the model Residual standard error (RSE) is 1.042. This means that the actual tip received by the wait

deviates by $1.042 from the true value, on average. And from this, we can say that the model doesn't fit the data quite well due to a relatively large RSE.

The R2 value shows that there is 52.5% less variation around the model than the mean. And this indicates that the model did not explain approximately 47.5% of the variability in the tip response. The fitted model only explains 52.5% of the variability in the amount of tip received by the waiter which is still small.

The F-statistic of 23.88 is far from 1, which informs us that the null hypothesis(the situation where is there is no relationship between tip amount and its predictors is zero) of the model can be rejected. And based on this, we can conclude that at least one of the predictors is statistically significant in determining the amount of tip received by the waiter.

Table 9: The table shows the model predictors' confidence intervals.

	2.5%	97.5%
(Intercept)	-0.18691464	1.7285755
total_bill	0.06612763	0.1177213
sexMale	-0.42010801	0.3109394
smokerYes	-0.24093097	0.5551136
daySat	-1.06663114	0.6271670
daySun	-0.87938937	0.9116989
dayThur	-0.85284335	1.1783588
timeLunch	-1.21781402	0.9998048
size	-0.06008829	0.3736033

3.3 Tip prediction and RSE

The Root Mean Squared Error(RMSE) of the model is 1.197048 and this indicates that the average deviation of the predicted tip from an actual tip received by the waiter is $1.197048. This error is extremely large making the model less suitable to determine how much tip a waiter receives.

The Model's Mean Square of 1.432924 is quite large indicating also a bigger deviation from the actual amount of tip hence making the model less efficient in predicting the amount of tip that should be given to the waiter.

4 Model Improvement

Model improvement was achieved through the use of lasso regression.

Table 10: *The table shows the improved model RMSE alongside the RMSE of the previously fitted model*

Lasso model RMSE	Linear Model RMSE
1.194968	1.197048

Comparing the models, the regularized model gives a very small RMSE value of $1.194968 compared to the previously fitted model. The lasso model has also reduced the coefficients of the less important predictors to zero as shown in the table and graph below.

Table 11: *Coefficients of the regularised model*

Predictors	Coefficient Estimate
total bill	0.08610221
SexMale	0
SmorkerYes	0
TimeLunch	0
Size	0.09892514
daySat	-0.03508312
daySun	0
dayThur	0

Four out of the six predictors are shrunk to zero. From the table above, the total bill and size are the most determinants of tip given to waiters. We exclude the day of the week because it's only Saturday which contributes and there is no way it can only be included without other days.

Also, the Backward selection approaches indicate that total bill and size are the most important predictors in determining the tip amount.

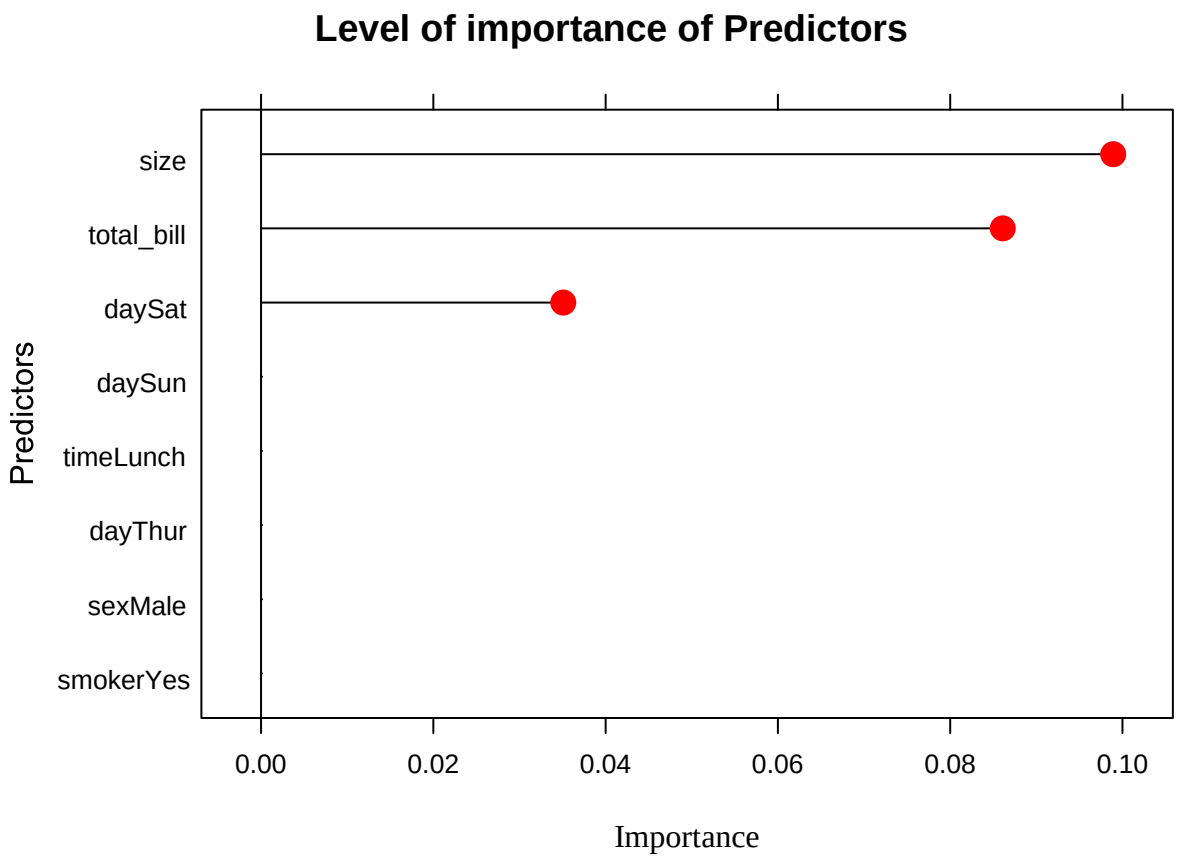


Figure 4: The figure above shows the most important predictors in estimating the amount of tips received by the waiter in the restaurant. The graph indicates that the size of the party is most contributors to tip prediction followed by total bill paid and then party day Saturday.

4.1 Final Model

The final model is,

$$Tip = 0.38277 + 0.09194totalbill + 0.26966size + 0.98489smoker - 0.45443(size * smoker)$$

The inclusion of an interaction between size and smoker makes the smoker predictor statistically significant which is not the case when the interaction is excluded. And also the interaction between the two is statistically significant.

And also the size-smoker interaction increases the model R-squared value from 0.4832 to 0.4887, therefore, making it a better interaction in the model.

Table 12: *Coefficients and p-values of the final model*

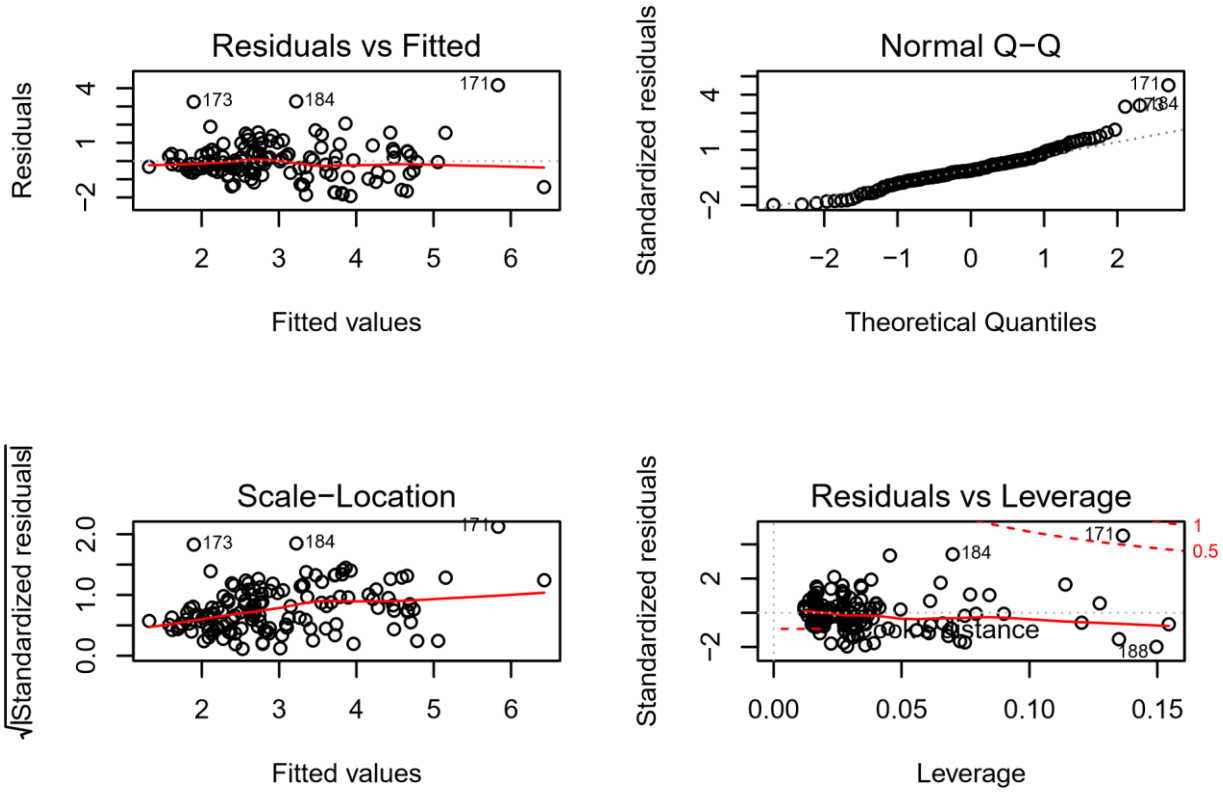
Predictors	**Coefficient Estimate**	**p-values**
total bill	0.09194	1.8e-11
Smorker	0.98489	0.0455
Size	0.26966	0.0228
size:smokerYes	-0.45443	0.0546

From the table above, the total bill is the most important predictor of the tip amount and a unit increase on the total bill increases the tip amount by $0.09194 and if smoking is allowed at the party the tip increases by $0.98489, each additional person added on the party also increases the tip by $0.26966 and finally if the additional person added on the party smokes, then tip amount decreases by $0.45443.

The improved model gives a slightly smaller Root Mean Square Root of 1.181519 compared to that of the initial fitted model of 1.197048. This means that the final model's predictions deviate on average from the actual tip received by the waiter by $1.181519. And therefore, we can conclude that the model ***Tip = 0.38277 + 0.09194totalbill + 0.26966size + 0.98489smoker -0.45443(size*smoker)*** is more suitable and efficient for estimating the amount of tip given to the wait because it reduces the RMSE of the initial model by 0.015529 (1.55%).

5 Checking Model Assumptions with Residual Plots

The residual plots below were used to evaluate the assumptions of normality of residuals, Homoscedasticity[3], and independence. These assumptions were made while fitting the model.



The Plot of Residuals versus Fitted values shows that the variability is not the same across the board where the variance of the residuals increases with fitted values and this violates the assumption of constant variance.

We can see from the Normal Quantile-Quantile (Q-Q) plot that most points do lay along the diagonal red line which means that residuals are approximately normally distributed and therefore this validates our use of the parametric method (i.e. Linear Regression model) in designing a model to predict the tip amount. Overall, from the above plots we can say that there is a moderate linear relationship between the tip variable and the predictors.

[3] Tendance of the variance of residuals remaining constant regardless of changing fitted values

In conclusion, total bill paid by the customer is the most determinant of the tip received by waited followed by the size of the party. And the interaction between size and smoker predictors makes the model more efficient in estimating the tip amount.

6. Appendix

```
knitr::opts_chunk$set(echo = TRUE)
options(tinytex.verbose = TRUE)
#Loading neccessary libraries
library(caTools)
library(ggplot2)
library(dplyr)
library(glmnet)
library(caret)
library(tinytex)
library(gridExtra)
library(cowplot)
library(kableExtra)
```

Data Analysis

```
# Reading the datasets of years 2015 to 2019
sl_data <- read.csv("G:/UCT-MSc. Data Science/Semester 1/Supervised Learning/Assignment/a
ssignment1/tipdata.csv",stringsAsFactors = TRUE,header = TRUE)

set.seed(39)
sl_data <- sl_data[sample(1:nrow(sl_data), 200, replace=FALSE), ]
write.csv(sl_data, 'my_tipdata.csv', row.names = FALSE)

#Changing the size dataset fron integer to numeric
sl_data$size <- as.numeric(sl_data$size)

sl_data <- sl_data[-1]

#structure if the dataset
str(sl_data)

#Viewing the summary sheet of the dataset
summary(sl_data$total_bill)

#sex distribution numbers
table(sl_data$sex)
#Percentage of male and female
round((prop.table(table(sl_data$sex))*100),2)

#number of different categories in smoker variable
table(sl_data$smoker)
#probability appearance of different categories
round((prop.table(table(sl_data$smoker))*100),2)

#Days' numbers
table(sl_data$day)
#probability of a party being organized on a particular day
```

```r
round((prop.table(table(sl_data$day))*100),2)

#number of times a party happening on different times of the day
table(sl_data$time)
#probability of a party occuring on lunch or dinner time
round((prop.table(table(sl_data$time))*100),2)

summary(sl_data[1:2])

#Ploting total bill and tip histgram alongside each other
par(mfrow = c(1,2))
hist(sl_data$total_bill,main = "Histogram of total bill",xlab = "Total bill($)", col = "b
lue")
hist(sl_data$tip,main = "Histogram of tip",xlab = "Tip($)", col = "blue")

#Relationship between total bill and tip
ggplot(data = sl_data,aes(total_bill,tip)) +
  geom_point() +
  theme_classic() +
  geom_smooth(method = "gam",se = F,color = 4, formula = y ~s(x)) +
  labs(x = "Total Bill Paid ($)", y="Tip Amount ($)" ,title = "Relationship between tip \
n amount and  bill paid") +
  theme(plot.title = element_text(hjust = 0.5,face = "bold",family = "Times",size = 15),
        strip.background = element_blank())
#relationships between variables using histograms
plot1 <- ggplot(data = sl_data,aes(day, fill =sex)) +
  geom_bar(position = "dodge2") +
  theme_classic()+
  labs(title = "Days and Gender composition" , x = "Party Days" , y = "Frequency") +
  theme(plot.title = element_text(hjust = 0)) +
    scale_color_manual(values = c("Red", "blue"))+ #coloring bins
  scale_fill_manual(values = c("Red", "Blue"))

plot2 <- ggplot(data = sl_data,aes(day, fill =smoker)) +
  geom_bar(position = "dodge2") +
  theme_classic()+
  labs(title = "Smoking and Day of the party" , x = "Party Days" , y = "Frequency") +
  theme(plot.title = element_text(hjust = 0)) +
   scale_color_manual(values = c("Red", "blue"))+
  scale_fill_manual(values = c("Red", "Blue"))

Plot3 <- ggplot(data = sl_data,aes(day, fill = time)) +
  geom_bar(position = "dodge2") +
  theme_classic()+
  labs(title = "Days and Time of the party composition" , x = "Party Days" , y = "Frequen
cy") +
  theme(plot.title = element_text(hjust = 0)) +
    scale_color_manual(values = c("Red", "blue"))+
  scale_fill_manual(values = c("Red", "Blue"))
```

```r
#Plotting histgrams alongside each other
plot_grid(plot1, plot2, Plot3,labels = "AUTO")
```

Splitting Data

```r
#Dividing the dataset into the train and test sets in a ratio of 8:2
linear_set <- sample.split(sl_data, SplitRatio = .8)

#Train set
train_set <- subset(sl_data, linear_set == "TRUE")
dim(train_set)

#Test set
test_set <- subset(sl_data, linear_set == "FALSE")
dim(test_set)

names(train_set)
```

Linear Model – Question 1

```r
#fitting a linear model
linear_model <- lm(tip~.,data = train_set)
summary(linear_model)

#mean of the tip amount in the train set
mean(train_set$tip)

#obtaing model's confidance intervals
confint(linear_model)

#numerical values assigned to categorical variables (smoker,sex,day,time)
contrasts(train_set$smoker)
contrasts(train_set$day)
contrasts(train_set$sex)
contrasts(train_set$time)
#Model's confidance intervals
confint(linear_model)
#Using the fitted model to predict the tip
model_predict <- predict(linear_model,newdata = test_set)

model_predict1 <- cbind(Actual = test_set$tip, Predicted = model_predict)

model_predict2 <- as.data.frame(model_predict1)

model_predict2 <-  cbind(model_predict2, error= c(model_predict2$Actual - model_predict2$
Predicted))

mean((model_predict2$error)^2)
#Mean Root square Error
```

```r
sqrt(mean((model_predict2$error)^2))
```

Lasso Regression – Question 2

```r
#Compute Lasso regression:
lambda <- 10^seq(-1, 1, length = 100)
# Build the model

#Compute fitness level using lasso regression:
set.seed(39)
model_lasso <- train(
  tip ~., data = train_set, method = "glmnet",
  trControl = trainControl("cv", number = 10),
  tuneGrid = expand.grid(alpha = 1, lambda = lambda) #alpha = 1, because we are fitting a
 lasso model
  )

# Model coefficients
coef(model_lasso$finalModel, model_lasso$bestTune$lambda)
# Make predictions
predictions <- model_lasso %>% predict(test_set)
# Model prediction performance
data.frame(
  RMSE = RMSE(predictions, test_set$tip),
  Rsquare = R2(predictions, test_set$tip)
)
#Lasso Plot
plot(varImp(model_lasso,scale = F),main = "Level of importance of Predictors",ylab = "Pre
dictors",col = "red",
     pch = 21, bg = "lightgray",lwd = 0.9, cex = 1.5, fill = "red")


#Using backward selection method to select most important predictors
step( object = linear_model,
       direction = "backward"
 )



final_model <- lm(tip ~ total_bill + size + smoker + size * smoker, data = train_set)
summary(final_model)



#Using the final model to predict the tip
model_final_predict <- predict(final_model,newdata = test_set)

model_final_predict1 <- cbind(Actual = test_set$tip, Predicted = model_final_predict)
```

```r
model_final_predict2 <- as.data.frame(model_final_predict1)

model_final_predict2 <-  cbind(model_final_predict2,
                               error= c(model_final_predict2$Actual -model_final_predict2
$Predicted))

#RME of the model
(mean((model_final_predict2$error)^2))

#Mean Root square Error
sqrt(mean((model_final_predict2$error)^2))

#RMSE Error reducation
(1.197048 - 1.181519)

#Model's diagnostic plots
par(mfrow = c(2,2)) #2*2 plots display
plot(final_model) #final model diagnostic plots
```