Assignment 1: Linear Regression

Nonhlanhla Luphade
Student Number: LPHNON003

June 8, 2020

Contents

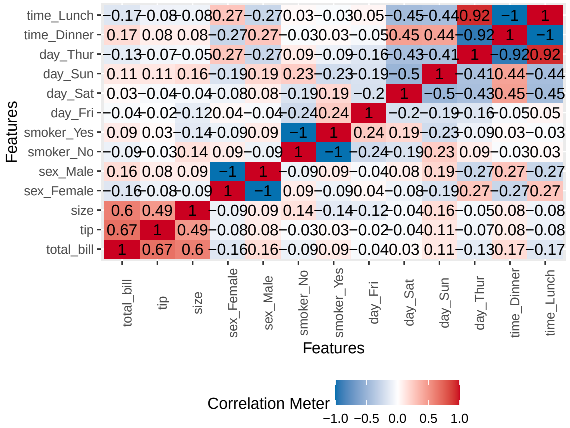
1 Introduction

The goal of this assignment is to build a linear model for predicting the average amount of tip in dollars a waiter can expect to earn from the restaurant given the predictor variables.The tip data is used for this analysis, this data set consists of 200 observations and 7 variables after the data split process.The tip data has seven variables, three numerical variables (i.e. total bill,tip and size) and four categorical variables (i.e. sex, smoker, time and day). For this analyses tip is the response variable and the other 6 variables are predictors.

1.1 Exploratory Data Analysis

Firstly exploratory data analysis is done as it is critical to any project related to Machine Learning. It is an approach to understand and summarize the main characteristics of a given data. For this assignment, a check for missing values, outliers and correlation within the variables was done. Further on we obtained a summary of the data and correlation plot.

Figure 1: Correlation plot



	Total_bill	Tip	Size
Min	3.070	1.000	1.000
1st Qu.	13.66	2.000	2.000
Median	17.81	3.000	2.000
Mean	19.89	3.034	2.588
3rd Qu.	24.18	3.575	3.000
Max	48.33	9.000	6.000

Table 1: continuous variables summary

Variable	**Description**
Smoker	Yes: 72 No: 127
Sex	Female: 72 Male: 127
Day	Thur: 52 Fri: 14 Sat: 68 Sun: 65
Time	Dinner: 143 Lunch: 56

Table 2: categorical variables summary

The data has many outliers, however one outlier value is very influential, this value is removed for the analysis as it is affecting the results of the analysis. The correlation plot shown in 1, it is observed that the variables are not that correlated to each other. Total bill and size are correlated to some extent but its not that bad. Tip and total bill are also correlated to some extent. The data set is clean and contains no missing values.

1.2 Data Partitioning

Before the analysis is done, the tip data is divided randomly into 70:30 split of train and test set as shown in 3. The 70:30 split is the most common and is mostly used during the training phase. 70% of the data is used for building our model (i.e. training), and the rest 30% is used for evaluating model performance (i.e. testing).

	Dimensions	
Training Data	139	7
Testing Data	40	7

Table 3: Dimensions of training and testing set.

2 Question 1: Multiple Linear Regression

To begin the analysis, a multiple linear regression model is fitted using all the predictors to assess if there is a relationship between the response and the predictors. Table 4 gives the output of the fitted model coefficients and table 5 gives the model evaluation metrics.

Coefficients	**Estimate**	**Std. Error**	**t value**	**Pr(>$\|t\|$)**	**Signif. code**
(Intercept)	0.67901	0.485668	1.398	0.164	
total_bill	0.10986	0.01280	8.584	2.41e-14	***
sexMale	-0.08961	0.18484	-0.485	0.629	
smokerYes	-0.27735	0.19787	-1.402	0.163	
daySat	0.06574	0.40833	0.161	0.872	
daySun	0.23608	0.41838	0.564	0.574	
dayThur	-0.71287	0.60856	-1.171	0.244	
timeLunch	0.80702	0.69929	1.154	0.251	
size	0.06898	0.12185	0.566	0.572	

Table 4: Coefficients of the full model

F-statistic	DF	P-value	R^2	Adjusted R^2	Training MSE	Testing MSE
17.38	8,130	<2.2e-16	0.517	0.487	0.924	1.126

Table 5: Model evaluation metrics of the full model

2.1 Is there any relationship between the response and predictors?

The null hypothesis states that there is no relationship between any of the predictors and the response, which can be tested by computing the F statistic. The p-value of F statistic can be used to determine whether the null hypothesis can be rejected or not. A high value of F statistic, with a very low p-value, implies that the null hypothesis can be rejected. In this model, $F = 17.38$ which is fairly high and the p-value ($< 2.2e - 16$) is very small, this suggests that at least one of the predictors must be related to the response.

2.2 Which of the predictor variables are significant?

To answer this, the p-value associated with each coefficient is used. From table 4 we observe that total bill is the only predictor variable with a significant p-value ($2.41e - 14$). Total bill is to be the only statistically significant predictor variable in this analysis. The total bill coefficient suggests that for every \$1000 increase in the total bill, an increase of 114 tip units while holding the other predictors constant.

2.3 Is this model fit?

To evaluate the model, R^2, the adjusted R^2, training mean squared error (MSE) and the testing MSE. R^2 measures the proportion of variability in the outcome that can be explained by the model, however, R^2 increases as number of predictors increase. Adjusted R^2 avoids this as it increase only if the new term improves the model. In this analysis the adjusted R^2 has a value of (0.487) as show in table 5, this shows that approximately 48% of the variance in the data is being explained by this model. 48% is not close to one ,therefore, the model is not a good fit. Another way to assess model accuracy is using the training and testing error. For good model we require our testing error to be small. Taking into consideration the bias-trade, a good model usually has a testing error that is greater than the training error (i.e. not under fitting) and the difference between the training error and the testing error should not be too big (i.e. not over fitting). In our case the training mse (0.924) and the testing mse (1.126) indicate that that the model is neither over fitting nor under fitting. Therefore, this model's performance is fair.

3 Question 2: Variable Selection

When we have a high dimensional data set, it would be highly inefficient to use all the variables since some of them might be imparting redundant information. It is important select the right set of variables which give an accurate model which is able to explain the dependent variable well and give good predictions. Variable selection is the process of selecting the best set of features that give an better model. There are many variable selection methods such as forward selection, backward selection, best subset selection, lasso regularization and many more. For this analysis, best subset selection ,lasso and elastic net were used. The table 6 gives a summary of the results obtained from these methods. The three techniques were used to select which variables subsets are best and linear models of those variable subsets were fitted to find our best model.

Variable Selection Method	R^2	**Adjusted-R^2**	**Testing MSE**	**Important variables**
Best Subset Selection	0.510	0.488	1.118	totalbill,sex,smoker,day
Lasso	0.509	0.490	1.110	totalbill,smoker,day
Elastic Net	0.507	0.495	1.114	totalbill,sex,smoker

Table 6: Model evaluation metrics after variable selection

Since the main aim of this analysis is prediction. The best model is one that generalises unseen data well and this model usually has high adjusted R^2 and a low testing mse. The subset of variables selected from all the variable selection methods were subsets of variables that would decrease the testing error of the full model and increase the variation of the response that would be explained by the model. The 3 models in table 6 were made from the best subsets for the variable selection methods.

3.1 Interaction term

Interaction terms are adding when the effect of one predictor variable on the the response variable depends on another. Adding interaction terms to model is a useful technique as enhances knowledge about the response variables. However, adding interaction terms makes the model more complex and hard to interpret,thus, it is important to consider the goal of the analysis before adding interaction terms. In this analysis the goal is to develop a model that predicts the response as accurately as possible. Therefore, it would be advantageous to add an interaction term to the final model to increase the predictive power of the final model.

Model	R^2	**Adjusted-R^2**	**Testing MSE**
$lm(tip \sim totalbill + day + totalbill : day)$	0.527	0.502	0.986
$lm(tip \sim totalbill + smoker + totalbill : smoker)$	0.574	0.563	0.963
$lm(tip \sim totalbill + sex + totalbill : sex)$	0.490	0.483	1.013

Table 7: Model evaluation metrics after adding interaction term

To select the best interaction term, the testing mse and the adjusted R^2 were consider to decide which model is best. The variables obtained from the variable selection procedure were considered, namely, total_bill,smoker,day and sex. Since total_bill is the most significant variable it was the first choice for all the models and the other terms were used to see which combination gives the best predictions. Table 7 gives a summary of the models considered and the testing mean squared errors obtained. The second model is the best and has the lowest testing mse compared to the other models, furthermore, it has the highest adjusted R^2 which means that the predictions are fairly good.

3.2 Final Model

The final model selected is the second model in table 7, the summary and the model evaluation metrics of this model are shown in table 8 and 9 respectively.

Coefficients	**Estimate**	**Std. Error**	**t value**	$Pr(> \lvert t \rvert)$	**Signif code**
(Intercept)	0.22019	0.24534	0.897	0.371055	
total_bill	0.14825	0.01173	12.642	<2e-16	***
smokerYes	1.40043	0.40786	3.434	0.000791	***
total_bill:smokerYes	-0.08402	0.01806	-4.653	7.73e-06	***

Table 8: Summary table for final model

F-statistic	DF	P-value	R^2	Adjusted R^2	Training MSE	Testing MSE
60.621	3,135	<2.2e-16	0.574	0.564	0.815	0.963

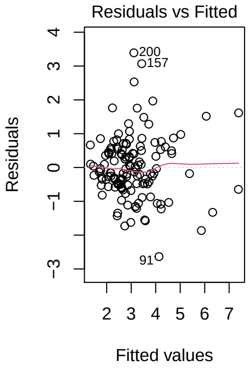
Table 9: Model evaluation metrics of the final model

The final model is better than the full model. The adjusted R^2 (0.564) has improved and the testing mse (0.963) is smaller. Total bill, smokerYes and the interaction term are significant and the model is simpler as it has less variables. The final generalises the unseen data well and the predictions are more precise compared to the full model. In the next section, the diagnostics of the final model are assessed.

4 Question 3: Residual Diagnostics

4.1 Diagnostic 1: Linearity

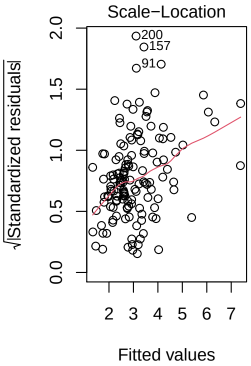
Figure 2: Residuals vs Fitted plot



The Residuals vs Fitted plot is used to check the linear relationship assumptions. A horizontal red line without distinct patterns is an indication for a linear relationship. In this analysis, there is no distinct pattern in the residual plot and the red horizontal line is approximately at zero. This suggests that we can assume a linear relationship between the predictors and the response.

4.2 Diagnostic 2: Homoscedasticity.

Figure 3: Scale-Location plot



Scale-Location plot is used to assess the homogeneity of variance of the residuals (homoscedasticity). Horizontal line with equally spread points is a good indication of homoscedasticity. In this analysis the scale-location plot has a horizontal line with equally spread points. Therefore we can assume homoscedasticity.

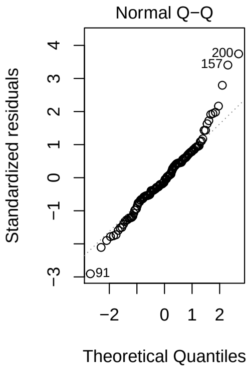
4.3 Diagnostic 3: Independence

	DW	p-value
Value	2.114	0.750

To check for independence the Durbin-Watson test is used. The null hypothesis states that the error terms are independent, the results of the test as shown in the table above suggest that there error terms are independent. The p-value (0.750) which is greater than 0.05, therefore, we conclude that the error terms are independent.

4.4 Diagnostic 4: Normality

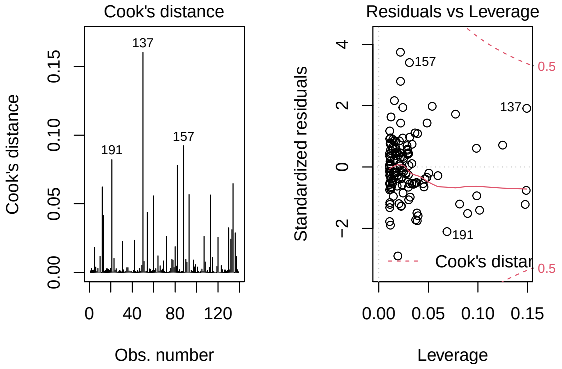
Figure 4: Normal Q-Q plot



Normal Q-Q plot Used to examine whether the residuals are normally distributed. It's good if residuals points follow the straight dashed line. In our analysis, all the points fall approximately along the reference line, therefore we can assume normality.

4.5 Diagnostic 5: Outliers and high levarage points

Figure 5: Normal Q-Q plot



Outliers are points that have an extreme outcome variable values. The presence of outliers may affect the interpretation of the model because it increases the RSE. Influential points are points which inclusion or exclusion can alter results of regression analysis. Not all outliers are influential. The Residuals vs Leverage plot is used to identify influential cases. In our plots, there are three extreme values. The three points seem to be well inside of the Cook's distance lines.

5 Appendix

5.1 Question 1

Listing 1: Question 1

```
#load libraries
library(DataExplorer )
library(leaps)
library(glmnet)
library("lmtest")

#load the data
setwd('C:/Users/user/Desktop/supervised Learning/Assignment1')
tip_data <- read.csv("my_tipdata.csv")[-1]
tip_data$sex <- as.factor(tip_data$sex)
tip_data$smoker <- as.factor(tip_data$smoker)
tip_data$time <- as.factor(tip_data$time)
tip_data$day <- as.factor(tip_data$day)
summary(tip_data)

#exploring data set
plot_missing(tip_data)
#corr plot
plot_correlation(tip_data, type = "all")
#
#outliers
mod <- lm(tip~.,data=tip_data)
cooksd <- cooks.distance(mod)
influential <- as.numeric(names(cooksd)[(cooksd > 4*mean(cooksd,
   na.rm=T))])
outlier <- car::outlierTest(mod)

#new tip data without the outlier
tip_data[19,] <- NA
tip_data <- na.omit(tip_data)

#Data Partioning for building model
set.seed(1234)
rand_sel <- sample(nrow(tip_data), 0.7*nrow(tip_data))
train <- tip_data[rand_sel,]
```

```
test <- tip_data[-rand_sel,]

#Build full Linear model
model1 <- lm(tip~.,data=train)
(model1_summary <- summary(model1))

#model Accuracy function
modelfit <- function(sm,model,traindat,testdat){
  trainpred <- predict(model,newdata=traindat[-2])
  testpred <- predict(model,newdata=testdat[-2])
  train_mse <- mean((trainpred-traindat$tip)^2)
  test_mse <- mean((testpred-testdat$tip)^2)
  rsq <- sm$r.squared
  adj_rsq <- sm$adj.r.square
  F_statistic <- sm$fstatistic
  return(cbind(F_statistic[1],F_statistic[2],F_statistic[3],rsq,adj_rsq,train_m
}

modelfit(model1_summary,model1,train,test)
```

5.2 Question 2

Listing 2: Question 2

```
#Variable selection
set.seed(1234)
x_vars <- model.matrix(tip˜.,data= train)[,-1]
y_var <- train$tip
x_vars1 <- model.matrix(tip˜. , test)[,-1]
y_var1 <- test$tip
‘%ni%‘<-Negate(‘%in%‘)

#Best subset selection
regfit_full = regsubsets(tip ˜ ., data = train)
(reg_summary <- summary(regfit.full))
plot(reg_summary$adjr2)

#lasso
cv_lasso <- cv.glmnet(x_vars,y_var,type.measure = ’mse’,nfolds =
   10,alpha = 1)
plot(cv_lasso)
c1_lasso<-coef(cv_lasso,s=’lambda.min’,exact=TRUE)
c2_lasso<-coef(cv_lasso,s=’lambda.1se’,exact=TRUE)
inds0<-which(c1_lasso!=0)
inds00<-which(c2_lasso!=0)
variables_lasso1<-row.names(c1_lasso)[inds0]
(variables_lasso1<-variables_lasso1[variables_lasso1 %ni%
   ’(Intercept)’])
variables_lasso2<-row.names(c2_lasso)[inds00]
(variables_lasso2<-variables_lasso2[variables_lasso2 %ni%
   ’(Intercept)’])

#elastic Net
cv_elastic <- cv.glmnet(x_vars,y_var,type.measure = ’mse’,nfolds
   = 10,alpha = 0.5)
plot(cv_elastic)
c1_elastic<-coef(cv_elastic,s=’lambda.min’,exact=TRUE)
c2_elastic<-coef(cv_elastic,s=’lambda.1se’,exact=TRUE)
inds1<-which(c1_elastic!=0)
inds11<-which(c2_elastic!=0)
variables_elastic1<-row.names(c1_elastic)[inds1]
```

```
(variables_elastic1<-variables_elastic1[variables_elastic1 %ni%
   ’(Intercept)’])
variables_elastic2<-row.names(c2_elastic)[inds11]
(variables_elastic2<-variables_elastic2[variables_elastic2 %ni%
   ’(Intercept)’])

#testing the models fit for models produced by our variable
   selections
model2 <- lm(tip~total_bill+smoker+day,data = train)
model4 <- lm(tip~total_bill+sex+smoker,data = train)
model3 <- lm(tip~total_bill+sex+smoker+day,data =train)

#model fit
modelfit(summary(model2),model2,train,test)
modelfit(summary(model3),model3,train,test)
modelfit(summary(model4),model4,train,test)

#Interactions
#total_bill is the most significant term to find the best
   interaction
#i investigated on how well total_bill interacts with smoker,day
   and sex
model5 <- lm(tip~total_bill+smoker+total_bill:smoker,data = train)
model6 <- lm(tip~total_bill+sex+total_bill:sex,data = train)
model7 <- lm(tip~total_bill+day+total_bill:day,data =train)

#model fit
modelfit(summary(model5),model5,train,test)
modelfit(summary(model6),model6,train,test)
modelfit(summary(model7),model7,train,test)
```

5.3 Question 3

Listing 3: Question3

```
#Final model
final_model <- lm(tip˜total_bill+smoker+total_bill:smoker,data =
   train)
(SM <- summary(final_model))

#model accuracy
modelfit(SM,final_model,train,test)

#residual dignostics
par(mfrow=c(3,3))
plot(final_model,which=1:5)
#test for independence
dwtest(final_model)
```

5.4 References

All the information in the assignment made us of the reference listed below:

- title= An introduction to statistical learning,
- author = James, Gareth and Witten, Daniela and Hastie, Trevor and Tibshirani, Robert

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: **NONHLANHLA LINDA LUPHADE**

STUDENT NUMBER: **LPHNON003**

TUTORS NAME: ______ TUT. GROUP #: ______

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: ______ Date: 4/6/2020