Name : Saurabh Rao Donthineni

Course: STAT 515

Instructor: Ran Ji

PART A

What route has the highest average fare?

Boston(MA) to San Jose(CA) has the highest average fare , 402 $.

What route is the busiest route (in terms of total number of flights)?

The busiest routes are : Chicago,IL to New York/Newark,NY and New York/Newark,NY to Washington,DC

What route is the least favorite route (in terms of total number of passengers)?

Baltimore to Providence is the least favorite route.

What route has the shortest distance?

Los Angeles to San Diego route has the shortest distance of 114 miles.

What proportion of vacation routes originated in DC?

Proportion of vacation routes originating in DC is 0.012

PART B

Correlation table between fare and other numeric predictors

COUPON 0.49653696 NEW 0.09172969 HI 0.02519492 S_INCOME 0.20913485

S_POP 0.14509708 E_POP 0.28504299 **DISTANCE 0.67001599** PAX -0.09070541

FARE 1.00000000

As can be seen, distance has the highest correlation with fare with a value of 0.67001599.

Attachments : 1b_correlationplot.jpg , 1b_numericdataplot.jpg

PART C

The mean difference of fare according to each category is mentioned below.

```
> # mean difference
> # for vacation
> abs(diff(vacation[,2]))
[1] 47.57162
> # for southwest
> abs(diff(sw[,2]))
[1] 89.80052
> #for slot
> abs(diff(slot[,2]))
[1] 35.23372
> #for gate
> abs(diff(gate[,2]))
[1] 40.03308
```

We can see that SW (Whether Southwest Airlines serves that route (Yes) or not (No)) has the largest difference in mean FARE values between qualitative levels.

PART D

After running the model, the plot of residual vs. predicted values is generated. On observing the plot, it is observed to be funnel shaped and it therefore violates the constant variance assumption. The residual value is lower for the smaller values of x and the value of variance increases as x increases.

- Parameter estimates of regression output :

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 115.766413   5.034756   22.99   <2e-16 ***
SWYes       -69.153977   5.372827  -12.87   <2e-16 ***
DISTANCE      0.069949   0.003898   17.94   <2e-16 ***
```

- R squared value : 0.6239413
- BIC value : 4042.047
- Attachments : 1d_prediresid.jpg

PART E

- The parameter estimates in the final model of the regression output is nothing but the coefficients , which are :

```
coefficients(back)
(Intercept)   VACATIONYes         SWYes            HI      E_INCOME         S_POP 
 4.373489e+01 -3.684492e+01 -4.246592e+01  1.010253e-02  8.799114e-04  4.726780e-06 
        E_POP      SLOTFree      GATEFree      DISTANCE           PAX 
 3.826093e-06 -2.074825e+01 -1.946100e+01  8.014906e-02 -7.337664e-04 
```

- The R squared value is : 0.80790
- The BIC value is 3833.064

PART F

```
AIC(trainingfit)
[1] 4026.266
> AIC(back)
[1] 3785.719
```

When we compare the AIC values for the models that have been developed in part (d) and part (e), we can conclude that the backward regression model (part e) is better because it has a lower AIC value.

PART G

- The average fare on the route with the mentioned characteristics is 255.5089 $
- The average fare on the route if Southwest Airlines operates is 213.043 $
- The reduction in average fare on the route if Southwest decides to cover this route is 42.46592 $
- The regression coefficient of SW is -4.246592e+01
  Interpretation : For every 1 dollar increase in price , the SW_Yes value decreases by -4.246592e+01 units.

PART H

The following factors will not be typically available for predicting the average fare from a new airport (i.e., before flights start operating on those routes)
Coupon
Vacation
Hi
Distance
Pax

PART I

The values for the models are :

	new model	best model
r sq.	0.411	0.8079
AIC	4209.64	3785.71
BIC	4249.11	3833.06

- The r squared value for the best model is close to 81 % , which is significantly higher when compared to 41.1 % achieved by the new model. The best model is better when compared to the new model.
- The AIC and BIC values follow a similar trend. The best model is better when compared to the new model.

Appendix

```
#remove everything from global environment

rm(list=ls())



# set working directory

setwd("C:/Users/SOURAV/Desktop/stat")



#read data using read.csv

data<-read.csv("C:/Users/SOURAV/Desktop/stat/data.csv")



#1.a.



# this part of the question has been finished using tableau and excel.
```

```
#1.b.

# open csv file and see what all coulmns have numeric data and store in an array

integercoulmns <- c(5,6,9,10,12,13,16,17,18)

#use the data of those coulmns

numericdata<-data[,integercoulmns]

#correlation

x <- cor(numericdata,data$FARE)

#distance is the best predictor, obtained from correlation table

x

#plot the correlation to indicate best predictor graphically

plot(x,main=" best single predictor of FARE")

#scatter plots

plot(numericdata,pch="*",bg='black',col='black')
```

#1.c

```
#computing mean value of fare according to each category

vacation<- aggregate(FARE~VACATION, data=data, FUN=mean)

sw<- aggregate(FARE~SW, data=data, FUN=mean)

slot<- aggregate(FARE~SLOT, data=data, FUN=mean)

gate<- aggregate(FARE~GATE, data=data, FUN=mean)

# mean value of fare according to each category

# for vacation

vacation

# for southwest

sw

#for slot

slot
```

```
# for gate

gate



# mean difference

# for vacation

abs(diff(vacation[,2]))

# for southwest

abs(diff(sw[,2]))

#for slot

abs(diff(slot[,2]))

#for gate

abs(diff(gate[,2]))




# 1.d
```

```
#to get the number of rows

dim(data[2])

set.seed(12345)

# 60 % of 638 is 382

trainingindex <- sample(638, 382, replace=FALSE)

var <- c(8,16,18)

#data model consisting of the variables

datamodel <- data[,var]

# training set data

training = datamodel[trainingindex,]

# validation set data

validation = datamodel[-trainingindex,]

#find out the dimensions of the training and the validation set data

dim(training)
```

```
dim(validation)

# build the linear model

trainingfit <- lm(FARE~SW+DISTANCE, data= training)

summary(trainingfit)

BIC(trainingfit)

#plot for the predicted and residual values

predi <- fitted(trainingfit)

resid<- residuals(trainingfit)

plot(predi,resid)

abline(h=0,v=175)

#coefficients are parameter estimates




# 1.e.
```

```
# generate for backware regression

back<-step(lm(FARE~COUPON+NEW+VACATION+SW+HI+S_INCOME+E_INCOME+S_POP+E_POP+SLOT+GATE+DISTANCE+PAX,data = data[trainingindex,]),direction = 'backward')

#Coefficient values for the above model

coefficients(back)

#BIC value for the above model

BIC(back)

#Summary values for the above model

summary(back)


# 1.f

# findx the AIC values for both the models

AIC(trainingfit)

AIC(back)
```

```
#1.g.

gdata=data.frame(COUPON=1.202,NEW=3,VACATION="No",
SW="No",HI=4442.41,S_INCOME=28760,E_INCOME=27664,
S_POP=4557004,E_POP=3195503,SLOT="Free",GATE="Free",PAX=12782,DISTANCE=1976)

pred_val<-predict(back,gdata,se.fit = TRUE,terms=NULL,scale=NULL)

#average fare of model

pred_val$fit

gdata_SW_YES=data.frame(COUPON=1.202,NEW=3,VACATION="No",
SW="Yes",HI=4442.41,S_INCOME=28760,E_INCOME=27664,
S_POP=4557004,E_POP=3195503,SLOT="Free",GATE="Free",PAX=12782,DISTANCE=1976)

pred_val_SW_YES<-predict(back,gdata_SW_YES,se.fit = TRUE,terms=NULL,scale=NULL)

#average fare of model if route is covered by southwest airlines
```

```
pred_val_SW_YES$fit

#reduction in fare when southwest operates

y <- pred_val$fit - pred_val_SW_YES$fit

y

# to get the regression coefficient of sw

coefficients(back)



# part i

# exclude Coupon  , Vacation , Hi , Distance , Pax

trainingfit_i=lm(FARE~S_INCOME+E_INCOME+S_POP+E_POP+SLOT+GATE+NEW+SW,
data = data[trainingindex,])

summary(trainingfit_i)

summary(back)

AIC(trainingfit_i)

AIC(back)
```

```
BIC(trainingfit_i)

BIC(back)
```