**Module Assignment**

**Module 6**

**QMB 6304 Analytical Methods for Business**



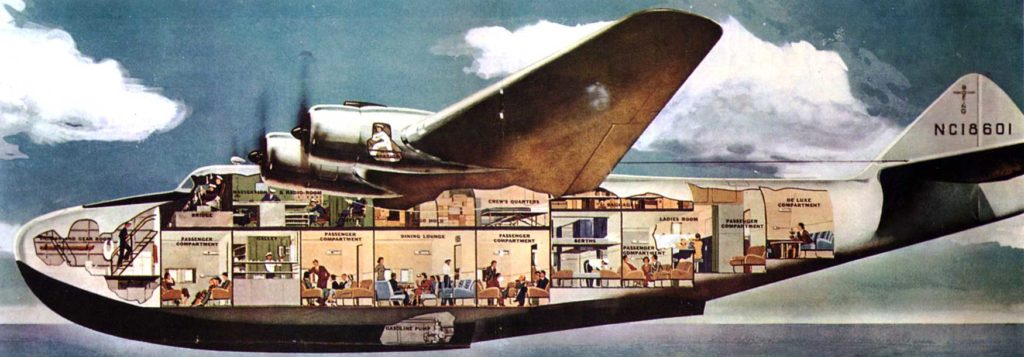
**Preprocessing**

1. Load the file “6304 Module 6 Assignment Data.xlsx” data set into R. This data shows airfares and passengers for certain U.S. Domestic Routes for the 4th quarter of 2002. This is not an exhaustive list of all U.S. cities or flights.
2. Create a second dataframe which will be a subset of the original and include only destination cities (airports) of IAD, IAH, LAS, LAX, LGA, MSP, PHX, SAN, SEA, SFO, STL, and TPA.
3. R likely sees certain variables as character (chr) variables. Convert these to factor variables. Recall the *str()* and *as.factor()* commands.
4. Take a random selection of this data of size n=50. Use the numerical portion of your U number as a random number seed for the selection.

**Analysis Using Your Sample (n=50) Data**

1. Show a scatterplot matrix of the continuous variables only.
2. Show two correlation matrices of the continuous variables, one in a numbers form and one in the form of correlation ellipses.
3. Using your correlation analysis and plots as a guide, find a regression model with the best fit to y=price using any or all of the remaining continuous variables. Show the R summary of the model output file and briefly explain why you have selected your independent variables. Give a verbal interpretation of the impact on price of each continuous variable included in your model.
4. Drawing on Step 3, determine whether your model meets the LINE assumptions of regression.
5. Drawing on Step 3, evaluate whether multicollinearity appears to be present in your model. If so, state which independent variables you believe are affected. Show and explain how you have arrived at this conclusion.
6. Drawing on Step 3, determine if adding the variable “destination” improves model fit. State whether you believe adding this factor variable to your model is an improvement.
7. Drawing on Step 6, assume airport MSP is significant in your model (whether it is or isn't). Show the beta coefficient for MSP and state the interpretation of the beta coefficient.

Your deliverable will be a single MS-Word file showing 1) the R script which executes the above preprocessing and analysis instructions and 2) the results of those instructions and needed written interpretations. The first line of your script file should be a “#” comment line showing your name as it appears in Canvas. Results should be presented in the order in which they are listed here. Deliverable due time will be announced in class and on Canvas. **This is an individual assignment to be completed before you leave the classroom. No collaboration of any sort is allowed on this assignment.**



**The Boeing 314 "Flying Boat"**

**Preprocessing:**

**#remove all the variables in Environment window**

**rm(list=ls())**

**#Import the required libraries**

**library(rio)**

**#1**

**#Load the data into R**

**my\_data=import('6304 Module 6 Assignment Data.xlsx')**

**colnames(my\_data)=tolower(make.names(colnames(my\_data)))**

**#2**

**#Create 2nd data set with selected Destinations**

**selected\_data=subset(my\_data,(destination=="IAD"|destination=="IAH"|**

**destination=="LAS"|destination=="LAX"|**

**destination=="LGA"|destination=="MSP"|**

**destination=="PHX"|destination=="SAN"|**

**destination=="SEA"|destination=="SFO"|**

**destination=="STL"|destination=="TPA"))**

**#3**

**str(selected\_data)**

**selected\_data$origin=as.factor(selected\_data$origin)**

**selected\_data$destination=as.factor(selected\_data$destination)**

**selected\_data$market.leading.airline=as.factor(selected\_data$market.leading.airline)**

**selected\_data$low.price.airline=as.factor(selected\_data$low.price.airline)**

**attach(selected\_data)**

Output:

**> str(selected\_data)**

**'data.frame': 360 obs. of 9 variables:**

**$ origin : chr "ALB" "ALB" "ALB" "ABQ" ...**

**$ destination : chr "LAS" "LAX" "TPA" "IAH" ...**

**$ average.fare : num 153 191 134 154 114 ...**

**$ distance : num 2237 2467 1130 767 487 ...**

**$ avg.weekly.passengers : num 237 192 203 373 621 ...**

**$ market.leading.airline: chr "WN" "DL" "US" "WN" ...**

**$ route.market.share : num 59.9 17.9 35.4 50.5 93.9 ...**

**$ low.price.airline : chr "WN" "US" "DL" "WN" ...**

**$ price : num 149 174 125 152 114 ...**

**#4**

**#Get random sample from the split data**

**set.seed(97)**

**my\_sample = selected\_data[sample(1:nrow(selected\_data),50,replace=FALSE),]**

Using rm statement, we can clear all the variables in the environment window. All the useful libraries are imported and using the import statement the data is loaded into R. As per the given step 2 a subset for the given Destinations is split into a new data frame. As per the step 3 using the str() statement in R, the structure of data frame is given and the chr variables as converted into vectors using as.factor() statement. As per step 4 a seed is set with my last 2 digits of U number and a random sample of 50 observations are taken for the analysis part. Proceeding to the analysis part.

**Analysis:**

1. Show a scatterplot matrix of the continuous variables only.

Rcode:

**#1**

**# Copy the continuous variables to a new data object.**

**my\_sample\_con=subset(my\_sample,select=c("average.fare","distance","avg.weekly.passengers","route.market.share","price"))**

**attach(my\_sample\_con)**

**#Scatterplot Matrix**

**plot(my\_sample\_con,pch=19,**

**main="ScatterPlot Matrix for continuous variables")**

**#Correlation analysis of the continuous variables.**

**cor(my\_sample\_con)**

**round(cor(my\_sample\_con),3)**

Result in console window:

**> #Scatterplot Matrix**

**> plot(my\_sample\_con,pch=19,**

**+ main="ScatterPlot Matrix for continuous variables")**

**> #Correlation analysis of the continuous variables.**

**> cor(my\_sample\_con)**

**average.fare distance avg.weekly.passengers route.market.share price**

**average.fare 1.0000000 0.67758078 0.17272593 -0.1978658 0.9078426**

**distance 0.6775808 1.00000000 0.09202736 -0.3762117 0.7248841**

**avg.weekly.passengers 0.1727259 0.09202736 1.00000000 -0.1109178 0.1395003**

**route.market.share -0.1978658 -0.37621166 -0.11091783 1.0000000 -0.2792413**

**price 0.9078426 0.72488408 0.13950031 -0.2792413 1.0000000**

**> round(cor(my\_sample\_con),3)**

**average.fare distance avg.weekly.passengers route.market.share price**

**average.fare 1.000 0.678 0.173 -0.198 0.908**

**distance 0.678 1.000 0.092 -0.376 0.725**

**avg.weekly.passengers 0.173 0.092 1.000 -0.111 0.140**

**route.market.share -0.198 -0.376 -0.111 1.000 -0.279**

**price 0.908 0.725 0.140 -0.279 1.000**

A subset for all the continuous variables are subsetted into a new data frame and used this subset for scatterplot matrix.

A picture containing qr code

Description automatically generated

From this matrix graph, we see there is some linear relation between ‘average.fare’, variable and ‘price’ variable and ‘average.fare’ variable and ‘distance’ shows some linearity and finally ‘price’ and ‘distance’ variables also shows some linearity . And for remaining variables, there is no relation between the other variables.

1. Show two correlation matrices of the continuous variables, one in a numbers form and one in the form of correlation ellipses.

Rcode:

**#2**

**# Presenting correlation graphically.**

**# First put a correlation matrix into an object.**

**library(corrplot)**

**gcorr=cor(my\_sample\_con)**

**gcorr**

**corrplot(gcorr,method="ellipse")**

**corrplot(gcorr,method="number")**

**# Correlation matrix with p values.**

**corr\_with\_p=Hmisc::rcorr(as.matrix(my\_sample\_con))**

**corr\_with\_p**

Result in console window:

**> #2**

**> # Presenting correlation graphically.**

**> # First put a correlation matrix into an object.**

**> library(corrplot)**

**> gcorr=cor(my\_sample\_con)**

**> gcorr**

**average.fare distance avg.weekly.passengers route.market.share price**

**average.fare 1.0000000 0.67758078 0.17272593 -0.1978658 0.9078426**

**distance 0.6775808 1.00000000 0.09202736 -0.3762117 0.7248841**

**avg.weekly.passengers 0.1727259 0.09202736 1.00000000 -0.1109178 0.1395003**

**route.market.share -0.1978658 -0.37621166 -0.11091783 1.0000000 -0.2792413**

**price 0.9078426 0.72488408 0.13950031 -0.2792413 1.0000000**

**> corrplot(gcorr,method="ellipse")**

**> corrplot(gcorr,method="number")**

**> # Correlation matrix with p values.**

**> corr\_with\_p=Hmisc::rcorr(as.matrix(my\_sample\_con))**

**> corr\_with\_p**

**average.fare distance avg.weekly.passengers route.market.share price**

**average.fare 1.00 0.68 0.17 -0.20 0.91**

**distance 0.68 1.00 0.09 -0.38 0.72**

**avg.weekly.passengers 0.17 0.09 1.00 -0.11 0.14**

**route.market.share -0.20 -0.38 -0.11 1.00 -0.28**

**price 0.91 0.72 0.14 -0.28 1.00**

**n= 50**

**P**

**average.fare distance avg.weekly.passengers route.market.share price**

**average.fare 0.0000 0.2303 0.1684 0.0000**

**distance 0.0000 0.5250 0.0071 0.0000**

**avg.weekly.passengers 0.2303 0.5250 0.4432 0.3339**

**route.market.share 0.1684 0.0071 0.4432 0.0495**

**price 0.0000 0.0000 0.3339 0.0495**

From the above correlation p-values, average.fare and distance has good p-value, saying that we are rejecting the null hypothesis and accepting the alternative hypothesis that the value is not equal to ‘0’. And average.fare and price has good p-value. Distance with route.market.share and price variables has good p-value. Avg.weekly.passengers doesn’t show good p-values with other variables. Route.market.share shows good p-value with price.

Chart, bubble chart

Description automatically generated

A picture containing timeline

Description automatically generated

From the correlation matrix, we can see that, price and average.fare has highest linear relationship with 91% , price and distance has some linear relationship with 72% , average.fare and distance also has some linear relationship with 68%.

1. Using your correlation analysis and plots as a guide, find a regression model with the best fit to y=price using any or all of the remaining continuous variables. Show the R summary of the model output file and briefly explain why you have selected your independent variables. Give a verbal interpretation of the impact on price of each continuous variable included in your model.

Rcode:

**#3**

**# Conducting a Regression -- with Continuous Variables Only**

**output=lm(price~.,data=my\_sample\_con)**

**summary(output)**

**output\_3=lm(price~average.fare+distance ,data=my\_sample\_con)**

**summary(output\_3)**

Result in console window:

**> #3**

**> # Conducting a Regression -- Continuous Variables Only**

**> output=lm(price~.,data=my\_sample\_con)**

**> summary(output)**

**Call:**

**lm(formula = price ~ ., data = my\_sample\_con)**

**Residuals:**

**Min 1Q Median 3Q Max**

**-66.063 -10.581 3.093 10.360 66.376**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 28.8286711 17.3385131 1.663 0.1033**

**average.fare 0.6761717 0.0691820 9.774 1.06e-12 \*\*\***

**distance 0.0149716 0.0071631 2.090 0.0423 \***

**avg.weekly.passengers -0.0008777 0.0028490 -0.308 0.7595**

**route.market.share -0.2138823 0.2201437 -0.972 0.3365**

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**Residual standard error: 23.56 on 45 degrees of freedom**

**Multiple R-squared: 0.8498, Adjusted R-squared: 0.8364**

**F-statistic: 63.63 on 4 and 45 DF, p-value: < 2.2e-16**

The Adjusted R-squared value is very good for this model. We got 83% which is very good value.

The Intercept is 28.82.

The **average.fare** variable is a significant value and the p-values is very good, where we reject the null hypothesis and accept the alternative hypothesis saying that the average.fare value is not equals to ‘0’. So, for 1 value increase for average.fare, there is an increase of 0.68 cents of price.

The **distance** variable is a significant value and the p-values is good, where we reject the null hypothesis and accept the alternative hypothesis saying that the distance value is not equals to ‘0’. So, for 1 value increase for distance, there is an increase of 0.015 cents of price.

The **avg.weekly.passengers** and **route.market.share** variables are not significant.

Doing the linear model only for significant variables.

**> output\_3=lm(price~average.fare+distance ,data=my\_sample\_con)**

**> summary(output\_3)**

**Call:**

**lm(formula = price ~ average.fare + distance, data = my\_sample\_con)**

**Residuals:**

**Min 1Q Median 3Q Max**

**-69.429 -7.960 2.435 10.113 71.431**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 14.73671 9.72938 1.515 0.1366**

**average.fare 0.66738 0.06733 9.912 4.26e-13 \*\*\***

**distance 0.01739 0.00666 2.611 0.0121 \***

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**Residual standard error: 23.31 on 47 degrees of freedom**

**Multiple R-squared: 0.8464, Adjusted R-squared: 0.8399**

**F-statistic: 129.5 on 2 and 47 DF, p-value: < 2.2e-16**

Doing the linear model for only significant variables doesn’t increase the Adjusted R-squared value much. There is a little improvement in the R-value. We got 84% of Adjusted R-squared value.

1. Drawing on Step 3, determine whether your model meets the LINE assumptions of regression.

Rcode:

**#4**

**# Assumptions of Regression**

**par(mfrow=c(2,2))**

**# Linearity**

**plot(my\_sample\_con$price,output$fitted.values,**

**pch=19,main="Actuals v. Fitted Values, Price")**

**abline(0,1,col="red",lwd=3)**

**# Normality**

**qqnorm(output$residuals,pch=19,**

**main="Normality Plot, Price")**

**qqline(output$residuals,lwd=3,col="red")**

**hist(output$residuals,col="red",**

**main="Residuals, Price",**

**probability=TRUE)**

**curve(dnorm(x,mean(output$residuals),**

**sd(output$residuals)),**

**from=min(output$residuals),**

**to=max(output$residuals),**

**lwd=3,col="blue",add=TRUE)**

**# Equality of Variances**

**plot(output$fitted.values,rstandard(output),**

**pch=19,main="Equality of Variances, Price")**

**abline(0,0,lwd=3,col="red")**

**par(mfrow=c(1,1))**

Result in console window:

**> #4**

**> # Assumptions of Regression**

**> par(mfrow=c(2,2))**

**> # Linearity**

**> plot(my\_sample\_con$price,output$fitted.values,**

**+ pch=19,main="Actuals v. Fitted Values, Price")**

**> abline(0,1,col="red",lwd=3)**

**> # Normality**

**> qqnorm(output$residuals,pch=19,**

**+ main="Normality Plot, Price")**

**> qqline(output$residuals,lwd=3,col="red")**

**> hist(output$residuals,col="red",**

**+ main="Residuals, Price",**

**+ probability=TRUE)**

**> curve(dnorm(x,mean(output$residuals),**

**+ sd(output$residuals)),**

**+ from=min(output$residuals),**

**+ to=max(output$residuals),**

**+ lwd=3,col="blue",add=TRUE)**

**> # Equality of Variances**

**> plot(output$fitted.values,rstandard(output),**

**+ pch=19,main="Equality of Variances, Price")**

**> abline(0,0,lwd=3,col="red")**

**> par(mfrow=c(1,1))**

Chart

Description automatically generated

From the linearity graph, we can say that the data follow the linearity, as all the data points are close to the standard linear red line. But there are some outlines after $250 price. The actual values are less than predicted values. This might be because as the flight distance increases, there might be a significant reduce in the price of the flight.

From the Normality graph, we can say that the data follows Normality, as all the points are close to the standard line, but at the end of tail there are some outliers. So, overall we can say the data follows Normaility.

From the Histogram graph, the data follow the normal distribution curve or bell curved shape.

From the Equality graph, we see that equality is not good, because it follows Heteroscedasticity. Because, the data at the low price values are close to the line as the price increases the data is spread out far from the line and some outliers are at 3rd standard deviation. So, if we remove the 3rd standard deviation values, then the data follow equality of variances.

1. Drawing on Step 3, evaluate whether multicollinearity appears to be present in your model. If so, state which independent variables you believe are affected. Show and explain how you have arrived at this conclusion.

Rcode:

**#5**

**# Evidence of Multicollinearity.**

**plot(my\_sample\_con,pch=19,**

**main="Continuous Variables")**

**my\_cor=cor(my\_sample\_con)**

**corrplot(my\_cor,method="number")**

**corrplot(my\_cor,method="ellipse")**

**library(car)**

**vif(output)**

Result in console window:

**> library(car)**

**> vif(output)**

**average.fare distance avg.weekly.passengers route.market.share**

**1.909999 2.092827 1.041734 1.184178**

Generally, a VIF value above 4 indicates that multicollinearity might exist. Some sources say the VIF values should be less than 5 that indicates that multicollinearity doesnot exist. When VIF is higher than 10, there is significant multicollinearity exists.

The values for all the continuous variables are less than 4 or 5. Clearly indicates that there is no multicollinearity with price variable.

1. Drawing on Step 3, determine if adding the variable “destination” improves model fit. State whether you believe adding this factor variable to your model is an improvement.

Rcode:

**#6**

**# Adding Destination binary variables.**

**output2=lm(price~average.fare+distance+avg.weekly.passengers+route.market.share**

**+destination,data=my\_sample)**

**summary(output2)**

**output2=lm(price~average.fare+distance,data=my\_sample)**

**summary(output2)**

**output2=lm(price~average.fare+distance+destination,data=my\_sample)**

**summary(output2)**

Result in console window:

**> #6**

**> # Adding Destination binary variables.**

**> output2=lm(price~average.fare+distance+avg.weekly.passengers+route.market.share**

**+ +destination,data=my\_sample)**

**> summary(output2)**

**Call:**

**lm(formula = price ~ average.fare + distance + avg.weekly.passengers +**

**route.market.share + destination, data = my\_sample)**

**Residuals:**

**Min 1Q Median 3Q Max**

**-60.293 -9.647 1.834 6.777 66.579**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 32.623433 27.451843 1.188 0.2429**

**average.fare 0.638443 0.099577 6.412 2.53e-07 \*\*\***

**distance 0.019603 0.009138 2.145 0.0392 \***

**avg.weekly.passengers 0.001761 0.004791 0.368 0.7155**

**route.market.share -0.320828 0.272854 -1.176 0.2478**

**destinationIAH 21.280594 21.403223 0.994 0.3271**

**destinationLAS -9.431758 19.668028 -0.480 0.6346**

**destinationLAX -19.548333 19.393992 -1.008 0.3206**

**destinationLGA -12.260483 22.421383 -0.547 0.5881**

**destinationMSP 3.009807 16.614994 0.181 0.8573**

**destinationPHX 3.534338 15.769284 0.224 0.8240**

**destinationSAN 4.622353 28.001388 0.165 0.8699**

**destinationSEA 9.982081 21.376257 0.467 0.6435**

**destinationSFO 3.688047 15.207851 0.243 0.8098**

**destinationSTL 4.731599 14.251432 0.332 0.7419**

**destinationTPA -3.048972 17.823525 -0.171 0.8652**

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**Residual standard error: 25.65 on 34 degrees of freedom**

**Multiple R-squared: 0.8654, Adjusted R-squared: 0.8061**

**F-statistic: 14.58 on 15 and 34 DF, p-value: 1.219e-10**

I have tried to add the destination binary variable to the linear model with all the continuous variables. The Adjusted R-Squared value in 80% which is less than the previous model with all the continuous variables model.

And, none of the destinations are significant in my model, all the destinations have high p-values.

**> output2=lm(price~average.fare+distance+destination,data=my\_sample)**

**> summary(output2)**

**Call:**

**lm(formula = price ~ average.fare + distance + destination, data = my\_sample)**

**Residuals:**

**Min 1Q Median 3Q Max**

**-63.439 -7.726 1.997 7.440 71.271**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 8.691223 18.330666 0.474 0.6383**

**average.fare 0.644560 0.095234 6.768 6.63e-08 \*\*\***

**distance 0.022219 0.008491 2.617 0.0129 \***

**destinationIAH 24.552165 21.031538 1.167 0.2507**

**destinationLAS -7.277739 19.028669 -0.382 0.7044**

**destinationLAX -10.933104 17.860734 -0.612 0.5443**

**destinationLGA -0.859408 14.870411 -0.058 0.9542**

**destinationMSP 7.463458 16.064667 0.465 0.6450**

**destinationPHX 5.601767 15.431742 0.363 0.7187**

**destinationSAN 9.936814 27.427033 0.362 0.7192**

**destinationSEA 14.496735 20.816571 0.696 0.4906**

**destinationSFO 6.629561 14.703761 0.451 0.6548**

**destinationSTL 6.044169 14.095390 0.429 0.6706**

**destinationTPA 2.654727 17.007359 0.156 0.8768**

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**Residual standard error: 25.45 on 36 degrees of freedom**

**Multiple R-squared: 0.8598, Adjusted R-squared: 0.8091**

**F-statistic: 16.98 on 13 and 36 DF, p-value: 1.375e-11**

Here, I have removed the non significant continuous variables, to check whether there is an improvement in the model. But there is a very slight change in the Adjusted R-Squared value. There is not much improvement from the previous model.

1. Drawing on Step 6, assume airport MSP is significant in your model (whether it is or isn't). Show the beta coefficient for MSP and state the interpretation of the beta coefficient.

Rcode:

**#7**

**my\_sample$msp=NA**

**for(i in 1:length(my\_sample$destination)){**

**if(my\_sample$destination[i]=="MSP"){**

**my\_sample$msp[i]="MSP"}**

**else{**

**my\_sample$msp[i]="Other"**

**}**

**}**

**output2=lm(price~average.fare+distance+avg.weekly.passengers+route.market.share**

**+msp,data=my\_sample)**

**summary(output2)**

**output2=lm(price~average.fare+distance**

**+msp,data=my\_sample)**

**summary(output2)**

Result in console window:

**> #7**

#my\_sample.relevel=my\_sample

> #my\_sample.relevel$destination=relevel(my\_sample.relevel$destination,"MSP")

> my\_sample$msp=NA

> for(i in 1:length(my\_sample$destination)){

+ if(my\_sample$destination[i]=="MSP"){

+ my\_sample$msp[i]="MSP"}

+ else{

+ my\_sample$msp[i]="Other"

+ }

+ }

> output2=lm(price~average.fare+distance+avg.weekly.passengers+route.market.share

+ +msp,data=my\_sample)

> summary(output2)

Call:

lm(formula = price ~ average.fare + distance + avg.weekly.passengers +

route.market.share + msp, data = my\_sample)

Residuals:

Min 1Q Median 3Q Max

-65.950 -10.759 3.157 10.436 65.850

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 29.4013812 20.4104615 1.441 0.1568

average.fare 0.6752996 0.0717472 9.412 4.18e-12 \*\*\*

distance 0.0150878 0.0075476 1.999 0.0518 .

avg.weekly.passengers -0.0008649 0.0028906 -0.299 0.7662

route.market.share -0.2126052 0.2238395 -0.950 0.3474

mspOther -0.7110502 12.9712063 -0.055 0.9565

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23.83 on 44 degrees of freedom

Multiple R-squared: 0.8498, Adjusted R-squared: 0.8327

F-statistic: 49.77 on 5 and 44 DF, p-value: < 2.2e-16

The MSP destination is not significant in my model and the Beta coefficient is

0.7110502.

I have tried to add the destination MSP variable to the linear model with all the continuous variables. The Adjusted R-Squared value in 83% which is less than the previous model with all the continuous variables model.

> output2=lm(price~average.fare+distance

+ +msp,data=my\_sample)

> summary(output2)

Call:

lm(formula = price ~ average.fare + distance + msp, data = my\_sample)

Residuals:

Min 1Q Median 3Q Max

-69.032 -8.053 2.144 10.368 69.717

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 16.758156 15.367590 1.090 0.2812

average.fare 0.664954 0.069495 9.568 1.62e-12 \*\*\*

distance 0.017697 0.006969 2.539 0.0145 \*

mspOther -2.176929 12.719706 -0.171 0.8649

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23.55 on 46 degrees of freedom

Multiple R-squared: 0.8465, Adjusted R-squared: 0.8365

F-statistic: 84.59 on 3 and 46 DF, p-value: < 2.2e-16

Here, I have removed the non significant continuous variables along with MSP destination, to check whether there is an improvement in the model. But there is a very slight change in the Adjusted R-Squared value. There is not much improvement from the previous model.