|  |  |  |
| --- | --- | --- |
| **Logo  Description automatically generated** | **University of New Haven**  **Tagliatela College of Engineering** | **Logo  Description automatically generated** |

**COURSE:**

INDE 6645 — Data Analytics  
Individual Project: Multiple Linear Regression

**COURSE INSTRUCTOR:**

**Narjes Sadeghiamirshahidi**

**STUDENT INFORMATION:**

Sai Teja Potluri (00826412) fall 2023

Table of Contents

**1.Project description 3**

**2.Outputs and interpretation4**

Part 14

Part 26

Part 38

# **Project Description**

The United States airlines are facing issues with airport congestion and they are thinking about using low‐fare carriers such as Southwest (SW) to compete on existing routes and offer nonstop service on routes that previously lacked it. The aviation consulting firm would like to predict the fares in the event a new carrier is brought into service.

Goal:

The goal is to build a Predictive Model to predict the average fare (FARE) on a new route based on information such as the average number of nonstop, one‐stop, and two‐stop flights (COUPON), number of new carriers entering a route (NEW), whether the rout is a vacation route (VACATION), etc. The file Flightfares.csv includes information on the United States airlines.

Dataset:

The dataset contains records of 638 with 18 variables.

Descriptions of variables are presented in Table 1.

Table 1: Description of Variables

|  |  |
| --- | --- |
| **Variables** | **Description** |
| **S\_CODE** | Starting airport’s code |
| **S\_CITY** | Starting city |
| **E\_CODE** | Ending airport’s code |
| **E\_CITY** | Ending city |
| **COUPON** | Average number of coupons for that route:   * a one‐coupon flight is a nonstop flight * a two‐coupon flight is a one‐stop flight  etc. |
| **NEW** | Number of new carriers entering that route between Q3—1996 and Q2—1997 |
| **VACATION** | Whether (Yes) or not (No) a vacation route |
| **SW** | Whether (Yes) or not (No) Southwest Airlines serves that route |
| **HI** | Herfindahl index: the measure of market concentration |
| **S\_INCOME** | Starting city’s average personal income |
| **E\_INCOME** | Ending city’s average personal income |
| **S\_POP** | Starting city’s population |
| **E\_POP** | Ending city’s population |
| **SLOT** | Whether or not either endpoint airport is slot‐controlled (this is a measure of airport congestion) |
| **GATE** | Whether or not either endpoint airport has gate constraints (this is another measure of airport congestion) |
| **DISTANCE** | Distance between two endpoint airports in miles |
| **PAX** | Number of passengers on that route during the period of data collection |
| **FARE** | The average fare on a route |

# Output and Interpretations:

# Part 1:

h) Create a new Data Frame for Numerical Variables  
• Show the output

A screenshot of a computer

Description automatically generated

i) Create a new Data Frame for Categorical Variables  
• Show the output

A screenshot of a computer program

Description automatically generated

j) For the Numerical Variables: create a Correlation Table to explore the relationship between  
Numerical Variables and Response Variable 'FARE'  
• Which variable has the highest correlation with Response Variable 'FARE'? What does that  
highest correlation mean?

A table with numbers and letters

Description automatically generated

Variable with the Highest Correlation with 'FARE':Distance

Correlation Value: # Interpretation of correlation:0.67

This correlation indicates that as distance increases the fare also increases.

The correlation value ranges from -1 to 1.

A correlation close to 1 indicates a strong positive relationship (as one variable increases, the other tends to also increase).

A correlation close to -1 indicates a strong negative relationship (as one variable increases, the other tends to decrease).

A correlation close to 0 indicates a weak or no linear relationship.

k) For the Categorical Variables: create a Pivot Table with the average fare in each Category.  
• Show the output for each Categorical Variables

A white screen with black text

Description automatically generated

l) Convert all the Categorical Variables into Dummy Variables.  
• Show the output: show the Data Frame with converted Categorical Variables.

A screenshot of a computer code

Description automatically generated

# Part 2 :

a) Fit a Multiple Linear Regression Model to the average fare on a route (FARE) as a function of  
13 predictors.  
♣ Fit the regression model y on X  
♣ Compute & Print Intercept & Coefficients  
• Show the output

A screenshot of a computer

Description automatically generated

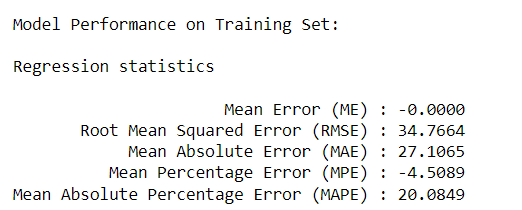
♣ Predict average fare on a route (FARE) using the MLR Model with 13 Predictors  
Meaning: Predict average fare on Validation set, print first few predicted & actual values, and residuals  
• Show the output & explain the result

A screenshot of a computer

Description automatically generated

**Output table shows: A sample of Predicted Prices & their Residuals (Errors) (relative to Actual Prices) for 20 flights in Validation Set, using MLR Model.**

b) Compute the performance of the MLR Model on the Training set and Validation set.  
♣ Compute & Print Performance (Accuracy) Metrics (on Training Set)  
• Show the output

****

♣ Compute & Print Performance (Accuracy) Metrics (on Validation Set)  
• Show the output

**A screenshot of a computer error

Description automatically generated**

**Interpretation:**

**Comparing errors for Training Set & Validation Set, shows that Model with 13 Predictors performs well**

# Part 3:

a) Exhaustive Search Method  
• Show the output & interpret (i.e., what are the best predictors suggested by Exhaustive Search Method)

**A screenshot of a computer

Description automatically generated**

**Output of Exhaustive Search:AIC decreases until ten predictors are used and then slowly increases.**

**#𝑹\_𝒂𝒅𝒋^𝟐 increases until eleven predictors are used and then slowly decreases.**

**#It indicates that a Model with ten Predictors is the best Model.**

**#Dominant Predictor in all Models is distance, with Whether Southwest Airlines or not serves that route playing important role as well.**

**#Best Predictors suggested by exhaustive method are 10 variables as listed below:**

**#1)DISTANCE**

**#2)SW\_Yes**

**#3)VACATION\_Yes**

**#4)HI**

**#5)GATE\_Free**

**#6)SLOT\_Free**

**#7) PAX**

**#8)E\_POP**

**#9)S\_POP**

**#10)E\_INCOME**

b) Three Partial Search Methods  
• Backward Elimination  
• Show the output & interpret (i.e., what are the best predictors suggested by Backward  
Elimination)

A white background with black text

Description automatically generated

**Interpretation: Best Predictors suggested by Backward Elimination Method are 10 variables as listed below:**

**#1)HI**

**#2)E\_INCOME**

**#3)S\_POP**

**#4)E\_POP**

**#5)DISTANCE**

**#6)PAX**

**#7)VACATION\_Yes**

**#8)SW\_Yes**

**#9)SLOT\_Free**

**#10)GATE\_Free**

• Forward Selection  
• Show the output & interpret (i.e., what are the best predictors suggested by Forward  
Selection)

**A white background with black text

Description automatically generated**

**Interpretation: Best Predictors suggested by Forward Selection Method are 10 variables as listed below:**

**#1)DISTANCE**

**#2)SW\_Yes**

**#3)VACATION\_Yes**

**#4)HI**

**#5)GATE\_Free**

**#6)SLOT\_Free**

**#7)PAX**

**#8)E\_POP**

**#9)S\_POP**

**#10)E\_INCOME**

**#We also see “10-predictor” suggested by Forward Selection is identical to “10-predictor” suggested by Exhaustive Search and backward elimination.**

• General Stepwise  
• Show the output & interpret (i.e., what are the best predictors suggested by General  
Stepwise)

**A white background with text

Description automatically generated**

**Interpretation: Best Predictors suggested by General Stepwise Method are 10 variables as listed below:**

**#1)DISTANCE**

**#2)SW\_Yes**

**#3)VACATION\_Yes**

**#4)HI**

**#5)GATE\_Free**

**#6)SLOT\_Free**

**#7) PAX**

**#8)E\_POP**

**#9)S\_POP**

**#10)E\_INCOME**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Predictors Name** | **Exhaustive Search** | **Forward Selection** | **Backward Elimination** | **General Stepwise** |
| **COUPON** |  |  |  |  |
| **DISTANCE** | ✔ | ✔ | ✔ | ✔ |
| **E\_INCOME** | ✔ | ✔ | ✔ | ✔ |
| **E\_POP** | ✔ | ✔ | ✔ | ✔ |
| **GATE\_Free** | ✔ | ✔ | ✔ | ✔ |
| **HI** | ✔ | ✔ | ✔ | ✔ |
| **NEW** |  |  |  |  |
| **PAX** | ✔ | ✔ | ✔ | ✔ |
| **SLOT\_Free** | ✔ | ✔ | ✔ | ✔ |
| **SW\_Yes** | ✔ | ✔ | ✔ | ✔ |
| **S\_INCOME** |  |  |  |  |
| **VACATION\_Yes** | ✔ | ✔ | ✔ | ✔ |
| **S\_POP** | ✔ | ✔ | ✔ | ✔ |

c) Compare methods of reducing the number of Predictors (i.e., Exhaustive Search and all three  
Partial Search Methods) in terms of the best subset of predictors they suggested.  
• Show the comparison in a table (as shown below)  
• Interpret the result

**Interpretation: Best Predictors suggested by General Stepwise Method , forward selection, backward selection and exhaustive search are the same as per the comparison table and represents the best subset of predictors as follows:**

**#1)DISTANCE**

**#2)SW\_Yes**

**#3)VACATION\_Yes**

**#4)HI**

**#5)GATE\_Free**

**#6)SLOT\_Free**

**#7) PAX**

**#8)E\_POP**

**#9)S\_POP**

**#10)E\_INCOME**

e) Fit the Model with the Best Subset of Predictors  
♣ Define Predictors & the Outcome  
♣ Fit the regression model  
♣ Compute & print coefficients  
• Show the output & interpret the result

**Predictors = ['VACATION', 'SW', 'SLOT', 'GATE','HI', 'E\_INCOME', 'S\_POP', 'E\_POP', 'DISTANCE', 'PAX']**

**Outcome = 'FARE'**

**A screenshot of a computer

Description automatically generated**

**Interpretation:**

**The linear regression equation predicts the airline fare (`Fare`) based on several factors. Among them, the distance (`DISTANCE`) has the highest positive impact, while being a vacation route (`VACATION\_Yes`), Southwest Airlines operating on the route (`SW\_Yes`), and having no gate constraints (`GATE\_Free`) significantly lower the fare. Passenger count (`PAX`) and being a slot-controlled route (`SLOT\_Free`) have negative impacts on fare, indicating lower fares with more passengers and absence of slot control. Other factors like household income (`HI`), external income (`E\_INCOME`), and population metrics (`S\_POP`, `E\_POP`) also contribute to fare variations.**

**Fare=52.89100015641333+HI(0.007188)+ E\_INCOME(0.001148)+ S\_POP(0.000004) +E\_POP(0.000004) + DISTANCE(0.075558)+ VACATION\_Yes (-35.865596) + SW\_Yes(-13.915304)+ GATE\_Free (-21.410803)+PAX (-0.000829)+SLOT\_Free (-13.915304)**

Compute & Print Performance (Accuracy) Metrics (on Training Set)  
♣ Compute & Print Performance (Accuracy) Metrics (on Validation Set)  
♣ Compare the Performance Metrics on Training and Validation Sets for the MLR Model with  
the Best Subset of Predictors  
• Show the outputs & interpret the result.

**A white background with black text

Description automatically generated**

**A white background with black text

Description automatically generated**

**Interpretation:**

**Comparing errors for Training Set & Validation Set, shows that Model with 1O predictors performs well**

♣ Predict average fare (FARE) using the MLR Model with the Best Subset of Predictors  
Meaning: Predict average fare, print the first few predicted values, actual values, and  
residuals  
• Show the output & explain the result.

**A screenshot of a computer

Description automatically generated**

**Interpretation:**

**Output Table Shows:**

**A sample of Predicted Prices & their Residuals (Errors) (relative to Actual Prices) and average fare for 20 flights in Validation Set, using MLR Model.**

f) Write the Equation for predicting average fare (FARE) based on the final MLR Model with the  
Best Subset of Predictors.  
• Write the MLR Equation based on the name of predictors and the outcome variable  
(Please do not use generic terms Y and X in the MLR Equation)  
o In the MLR Equation, show the coefficient values up to 6 digits

**MLR Equation for predicting average fare (FARE) based on the final MLR Model with the Best Subset of Predictors.**

**Fare=52.89100015641333+HI(0.007188)+ E\_INCOME(0.001148)+ S\_POP(0.000004) +E\_POP(0.000004) + DISTANCE(0.075558)+ VACATION\_Yes (-35.865596) + SW\_Yes(-13.915304)+ GATE\_Free (-21.410803)+PAX (-0.000829)+SLOT\_Free (-13.915304)**

g) Use the final MLR Model with the Best Subset of Predictors to predict the average fare on a  
route with the following characteristics:  
COUPON = 1.202, NEW = 3, VACATION = No, SW = No, HI = 4442.141, S\_INCOME = 28,760,  
E\_INCOME = 27,664, S\_POP = 4,557,004, E\_POP = 3,195,503, SLOT = Free, GATE = Free,  
PAX = 12,782, DISTANCE = 1976 miles  
• Show the outputs & interpret the result

****

COUPON = 1.202, NEW = 3, VACATION = No, SW = No, HI = 4442.141, S\_INCOME = 28,760,E\_INCOME = 27,664, S\_POP = 4,557,004, E\_POP = 3,195,503, SLOT = Free, GATE = Free, PAX = 12,782, DISTANCE = 1976 miles

substituting in MLR equation we get

Fare=52.89100015641333+(4442.141×0.007188)+(27764×0.001148)+(4557004×0.000004)+(3195503×0.000004)+(1976×0.075558)+0×(−35.865596))+(0×(−13.915304))+1×(−21.410803))+(12782×−0.000829)+(1×(−13.915304))

Fare=250.96963266441333

h) Use the final MLR Model with the Best Subset of Predictors to predict the average fare on a  
route if Southwest decides to cover the route in part g.  
• What is the reduction or increase in average fare on the route?



COUPON = 1.202, NEW = 3, VACATION = No, SW = Yes, HI = 4442.141, S\_INCOME = 28,760,E\_INCOME = 27,664, S\_POP = 4,557,004, E\_POP = 3,195,503, SLOT = Free, GATE = Free, PAX = 12,782, DISTANCE = 1976 miles

substituting in MLR equation we get

Fare=52.89100015641333+(4442.141×0.007188)+(27764×0.001148)+(4557004×0.000004)+(3195503×0.000004)+(1976×0.075558)+(0×(−35.865596))+(1×(−13.915304))+(1×(−21.410803))+(12782×−0.000829)+(1×(−13.915304)

Fare=237.0543286644133

Using the final MLR Model with the Best Subset of Predictors to predict the average fare on a route if Southwest decides to cover the route in part g.

The reduction in the average fare on the route if southwest operates is 13.915304.

Therefore, the interpretation is that, on average, the fare on a route is expected to decrease by approximately 13.92 units when Southwest decides to cover that route. This reduction in average fare is a consequence of the negative coefficient associated with the Southwest variable in the regression model, indicating a downward impact on fares when Southwest operates on a given route.