Question 1: Predicting Airfares on New Routes

Contents

[Multiple Linear Regression 2](#_Toc414354463)

[1. Numerical Predictors vs Fare 2](#_Toc414354464)

[2. Categorical Predictors vs Fare 2](#_Toc414354465)

[3. Predicting Fare 5](#_Toc414354466)

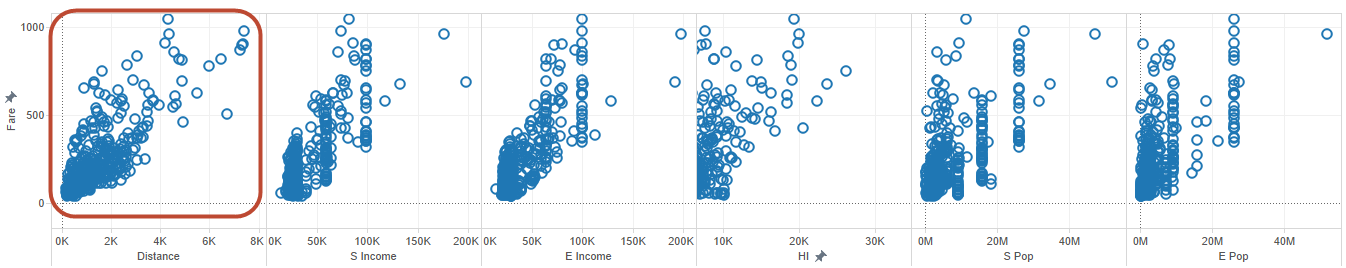
[4. Impact of SWA on Airlines Industry 8](#_Toc414354467)

# Multiple Linear Regression

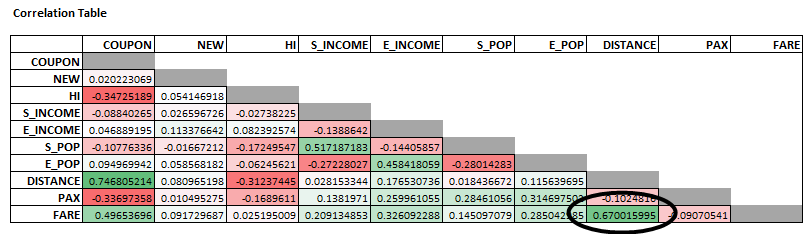
## Numerical Predictors vs Fare

The following scatter plot represents the relationship between the numerical predictors used in estimating the airlines fare and the actual fare. From this plot it is evident that **DISTANCE** is the best single numerical predictor for fare.

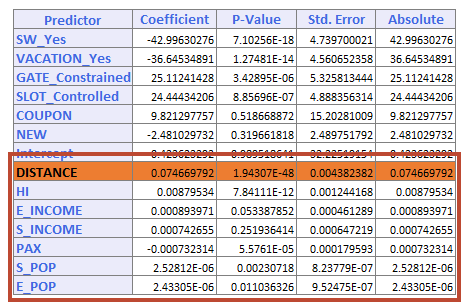
In this plot, fare seems to be directly proportional to distance unlike other numerical predictors where a definite relationship seems uncanny.



Also on creating the correlation table, the distance has the maximum influence on fare when compared to other numerical predictors.

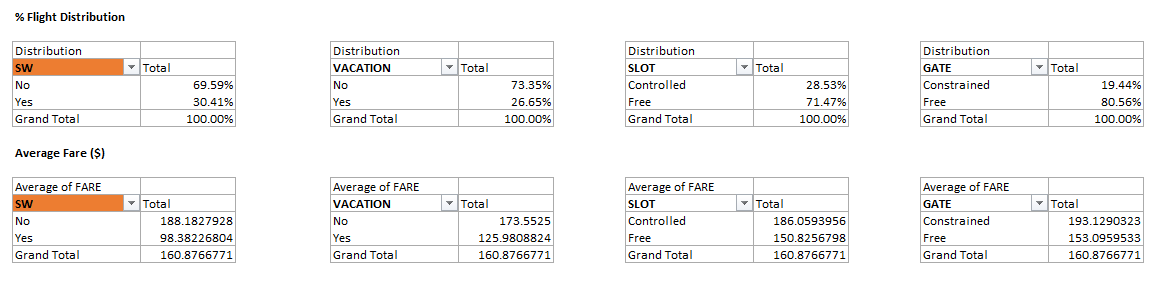


Also on performing a detailed analysis using MLR, our interpretation from the scatter plot is supported by the fact that distance has the least **P-Value** making it best numerical predictor for fare.

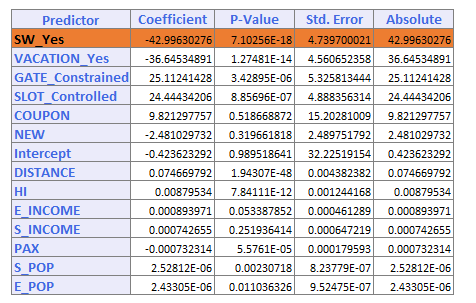


## Categorical Predictors vs Fare

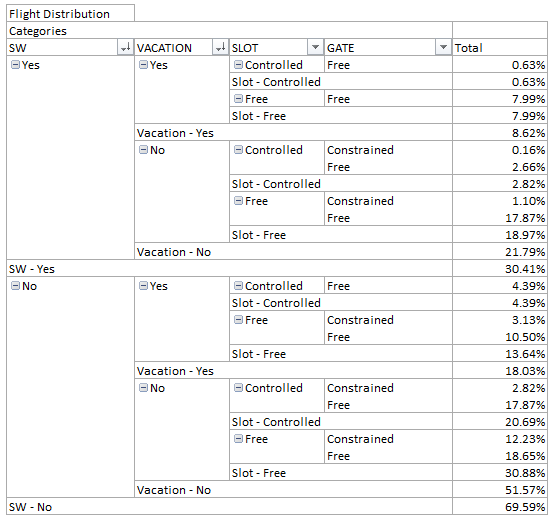
Analysis was performed on the airlines data using pivot tables and the following were identified. The first row represents the % Flight distribution across each of the categorical variables (SW, Vacation, Slot and Gate) and the second row represents the average fare in $.



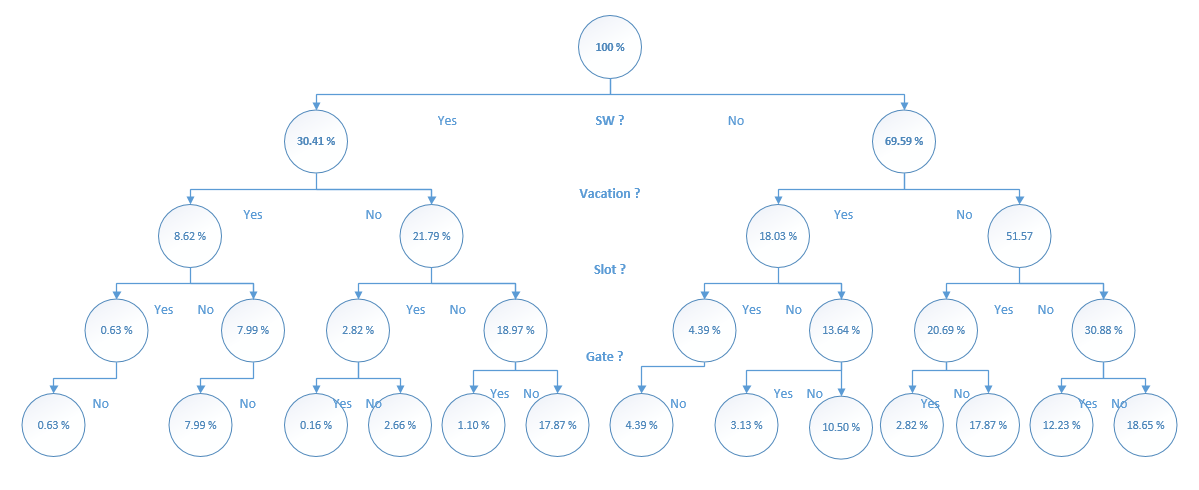
On analyzing the data, it was identified that **SW** presencehad the most impact on average Fare. Detailed analysis for performed using MLR and our interpretation was supported by the fact that SW variable has the highest coefficient and least **P-Value** making it the best categorical predictor for fare.



The following pivot table represents the hierarchical distribution of % flights with SW as root node followed by Vacation, Slot and Gate

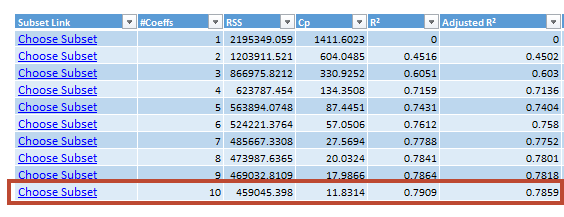


Graphical representation of the flight distribution hierarchy



## Predicting Fare

1. The categorical variables have been converted to dummy variables and the data has been partitioned (60-40) into training and validation sets
2. On performing **stepwise regression**, it was identified that the 10 variable model containing the predictors listed below the table is best suitable for estimating the fare as it has highest adjusted r^2, with Cp (11.83) closest to P (1+10 = 11)



The variables that are being considered are **HI, S\_POP, E\_POP, DISTANCE, PAX, VATION, SW\_YES, SLOT\_CONTROLLED, GATE\_CONSTRAINED and Intercept**

1. The **best fit** model uses 14 predictors these are inclusive of the 10 predictors identified in stepwise regression with addition of the predictors

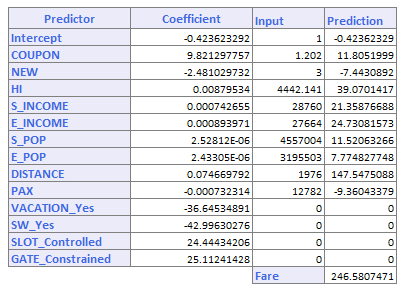
Additional variables under consideration: **COUPON, NEW, S\_INCOME, E\_INCOME**

1. Stepwise vs Best Fit

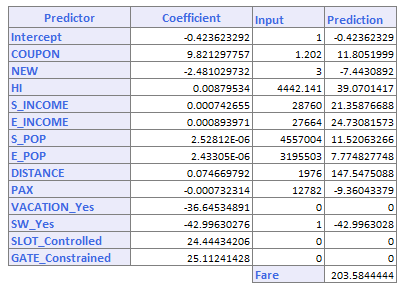
The RMS error value of best fit model is slightly better than that of model chosen using step wise regression as shown in the table where as the average error of the model chosen using stepwise regression has better average error values. Also there aren’t alarming differences when the lift charts of these models are compared.

|  |  |
| --- | --- |
| *Stepwise Regression – Selected Model* | *Best Fit Model* |
| C:\Users\Bala\AppData\Local\Temp\SNAGHTML261656e.PNG | C:\Users\Bala\AppData\Local\Temp\SNAGHTML260c2c6.PNG |
|  |  |
|  |  |

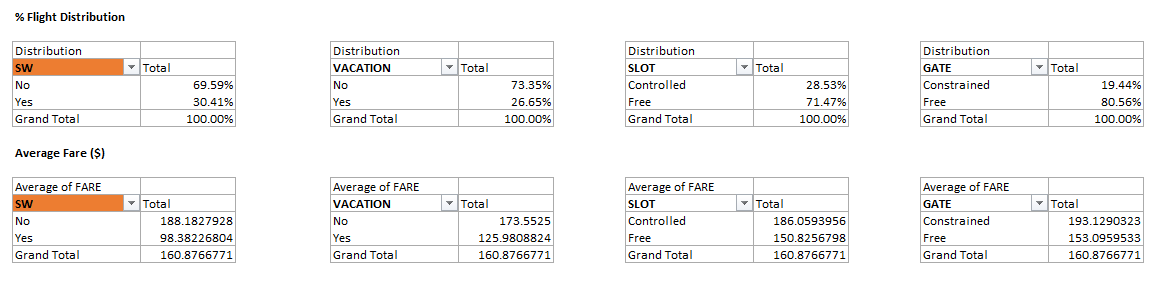
1. Prediction



1. Price Reduction if SW covers a specific route (Difference = - 42.99)

The fare falls by 42.99$ if SW decides to cover the route mentioned in (v).

Calculation of average price Reduction if SW covers all the routes

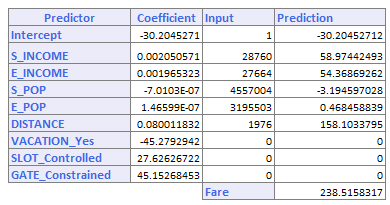


* 1. Current Average Price = 160.87 (as shown in the image)
  2. Average if SW covers all the routes = 130.84 (Estimated using exhaustive search)
  3. The average price reduction is $30.03

1. In general, the following variables will not be available in advance: **Coupon, New, SW, HI and PAX**.

The factors that can be estimated are **Vacation, S\_INCOME, E\_INCOME, S\_POP, E\_POP, SLOT, GATE and DISTANCE**. These are geographic/ economic factors that are independent of our analysis.

1. The model with variables available for prediction has been created.
2. The predicted result is using this limited data model is



1. *Less Variable Model vs Best Fit Model*

|  |  |
| --- | --- |
| *Less Variable Model* | *Best Fit Model* |
| C:\Users\Bala\AppData\Local\Temp\SNAGHTML27da7ce.PNG | C:\Users\Bala\AppData\Local\Temp\SNAGHTML260c2c6.PNG |

It is **worthwhile** reevaluating the model since the unknown variables have a considerable impact on the ticket price

## Impact of SWA on Airlines Industry

When analyzing the impact of southwest airlines on the industry, the categorical variable SW will play a vital role in data analysis. The data will have to be **partitioned** based on the presence of SWA. Relationships (say using scatter plot/ identifying avg. using pivot tables) between fare and predictor variables will have to be identified for the data partitions created (SW = Yes, SW = No).

The relationships identified for predictor variables in a data partition will have to be compared with the results of another.

*For Ex, Consider the predictor variable DISTANCE*

* *For SWA = No, Assuming that Co-efficient (fare/ mile) identified using MLR for distance = 1.2*
* *For SWA = Yes, Assuming that Co-efficient (fare/ mile) identified using MLR for distance = 1*

*When the relationships identified are compared, we get to know the airlines are forced to reduce their price/ mile by .2 (1.2-1) when SWA is present in a specific route.*

Also we could consider HI variable in routes that SWA covers to understand the distribution of business (whether it is a monopoly/ equally distributed).

Such an analysis will provide a detailed report on the impact SWA is having on each of the predictor variables which when aggregated provides the impact on the airlines industry. This would give an insight into the SW airlines pricing strategy and would help understand the factors that have positive/ negative impact on pricing based on presence of SWA.