# Predicting Airline Operating Costs

Homer Kay

## INTRODUCTION

Airline ticket pricing is one of many variables that can lead to the success or failure of an airline corporation in the long run. Underlying to the evaluation of pricing is (or should be) partly the operating costs of an airline. Without the assessment of operating costs a business would be missing an essential part in estimating their prices. Today's modern airlines use many variables to set ticket prices of flights: flight hours, time of day, air miles, number of seats, and fuel prices just to name a few. Although commercial airliner operations have gotten larger, sophisticated, and more complicated over the years (with larger airliners offering over 5,000 flights per day) the underlying fundamental costs associated with operating airplanes has stayed the same.

Working at Rockwell Collins, a supplier for large commercial airplane manufacturers (Boeing & Airbus), data that is within the aerospace industry interests me. As I am currently in the engineering side of the organization, I enjoy and prefer the difference in looking at decisions from a business perspective, rather than design, manufacturing, or other engineering factors. Whereas engineering problems can typically be solved by applying solid fundamentals to concepts and developing the detailed solution, I find business problems combine logic (stats and numbers), past experience, and usually some psychology to approach the best solution. To me business problems are inherently much more fascinating to solve because there is rarely a clear path to a correct answer.

## STATEMENT OF BUSINESS PROBLEM

The purpose of the analyses that follow is to create models for predicting airplane operating costs as a function of numerical inputs. Both continuous and binary methods will be

contrasted and compared to investigate which might be more appropriate for the application. Ultimately the goal is for a business to accurately predict operating costs.

## DATA SOURCES & DATA DESCRIPTION

The data is sourced from the University of Florida statistics database, specifically from "Airline Costs as a Function of Operating Variables" set. Operating variables are: length of flight (LOF), speed of plane (SOP), daily flight time per aircraft (DFT), population served (PS), ton-mile load factor (TMLF), revenue tons per aircraft mile (RTPAM), available capacity (AC), total assets (TA), investments and special funds (ISF), and adjusted assets (AA). Also note that operating costs is denoted as TOC throughout the paper and analysis.

For clarification:

- Ton-mile load factor - the ratio of revenue passenger miles to the available seat miles.
- Revenue ton per aircraft mile – single ton of goods that is transported for one mile.

Due to the fact that the dataset is quite small, n = 31, I will not be able to utilize a Training vs. Test set to validate the predictions calculated. Predictions will be compared using $R^2$, confusion matrices, and further insights.

## EXPLORATORY DATA ANALYSIS

Exploring the mean of each variable, population served is the largest at 14,000 (14 million people served), while the smallest is the ton mile load factor at just .14. Due to the variance in the scale of the numbers this indicates normalizing the variables could be useful. The standard deviations of many variables vary largely, indicating the groups are not necessarily well defined. Looking at a histogram of operating costs it is clear the data is skewed and that there are clear outliers on the opposite spectrum.

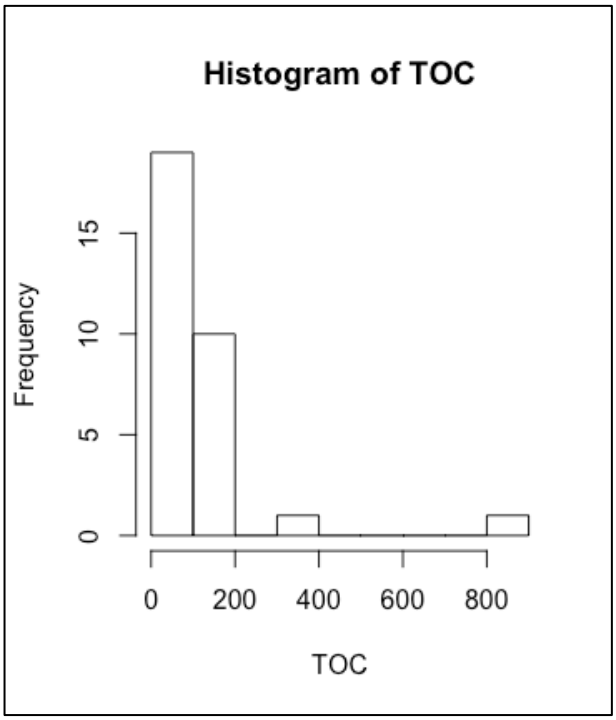


**Figure 1) Histogram of Operating Costs Data**

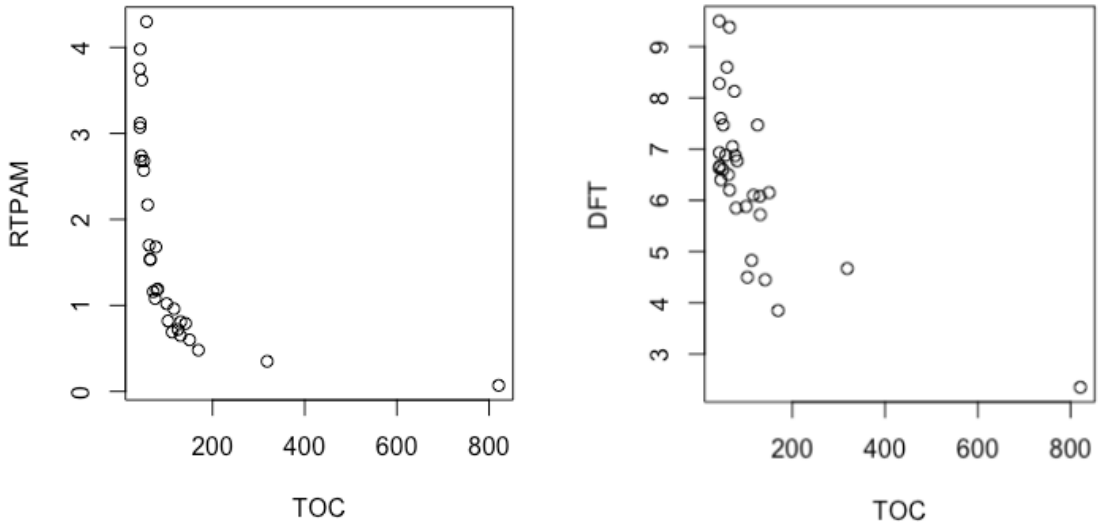


**Figure 2) Plots of DFT and RTPAM against Operating Cost**

From Figure 2 it is clear the inverse relationship both daily flight time and revenue tons per aircraft mile have on the operating costs. This is sensible as flight time/day increases the amount of revenue will increase from the ticket sales/plane increase, similarly as revenue tons per aircraft mile increases this will positively impact the bottom line, reducing operating costs.

## STATISTICAL ANALYSIS

### Part 1: Continuous Models

Due to the fact that each of the variables is numeric and the output variable, operating costs, is continuous, a linear regression model is appropriate. To begin with a linear regression is fitted with all variables attached. Immediately noticed was the adjusted assets (AA) variable was not calculated due to "singularities". By running a correlation matrix of all the numeric variables it is clear that there are multiple relationships with high correlations (Correlation > 0.9): LOF-SOP, LOF-RTPAM, SOP-RTPAM, LOF-AC, SOP-AC, PS-TA, AC-RTPAM, AA-PS, AA-TA, AA-ISF. The AA-TA relationship has a correlation of 99.87% indicating it is the leading case of multicollinearity. To fix the issue, the variable adjusted assets was deleted from the analysis. After resolving this issue the results for a full model linear regression are as follows:

Referring to Table 1, The high P-Values (>.05) for variables LOF, SOP, DFT, and PS indicate they are statistically insignificant in the regression model. Including a confidence interval of 90% would show that LOF and DFT both cross 0 in their intervals, and thus further indicate their insignificance. The $R^2$ value for the function is .92 indicating that 92% of operating costs is explained by the regression model variables; this indicates high confidence in the regression model's ability to predict operating costs.

After reducing the model by eliminating insignificant variables, the Adj$R^2$ reduced by 3% indicating that the prior model was superior although some variables were insignificant. However, a 3% delta indicates that the reduced model performs nearly as well as the full model. If the dataset was much larger, it would be preferred to proceed with the reduced model for decreased processing times and fewer sources of error. Also notice the intercept change from 1433 to 1163 as well as a reduction in each of the variable coefficient's; indicating improved predictability compared to the full model. The reduced multiple linear regression implies that as the ton mile load factor, available capacity, and total assets increase, the total operating costs decrease.

Due to the fact that variables vary in scale dramatically, from many thousands in size to just tenths it seems appropriate to apply a standardization of variables prior to running a regression. After applying a beta regression technique, looking at Table 3, one can see that RTPAM, TMLF, and AC are anywhere from 2-11 times more important as TA and ISF at predicting operating expenses. Considering RTPAM, TMLF, and AC are more like variables that effect the bottom line and TA and ISF are more like functions of those and other variables, it makes sense that those would have a greater impact to operating costs.

Applying polynomial regression was considered, although after plotting TOC against available variables, there is nothing suggesting it is obvious that including a polynomial term will be an improvement in the analysis.

**Part 2: Binary Models**

To test the ability of functions that deal with binary outcomes on this dataset I defined a new variable that split the dataset into "high" (TOC>110) and "low" (TOC<=110) operating cost outcomes.

After defining a new operating cost variable as 1 or 0 a logistic regression is performed on the 5 statistically significant variables (same as reduced models in Part 1). Due to the nature of the dataset the Logit's resulted in extremely large or extremely small values. These values were large enough that when $e^{Logit}$ was calculated errors of INF were thrown in cases where the probability approached 100%. Checking an example manually the Logit's still calculate the correct Operating Cost outcome, the results are just extreme. Notice in Table 4 that coefficients are very small or very large and that RTPAM, TMLF, and ISF are negative. Thus, the higher those values associated with those variables are the lesser the logarithmic odds of having high operating costs will be. P Values were extremely small indicating high significance for all 5 variables in the logistic regression model.

The linear discriminant analysis proved to output fairly nominal values in contrast to the large logistic regression results. I believe this to be due to the fact that n (number of data points) is small in this sample set of Airline data, thus potentially causing the logistic regression to be unstable. I also believe that by defining a new output variable and making the center the mean (110) this made the logistic regression and linear discriminate analysis, "to good", similar to how highly correlated datasets in continuous models will have multicollinearity issues. However, the application and results are still relevant and have value in the purpose of this analysis.

Comparing the confusion matrices in Table 5 it's clear that both the logistic regression and linear discriminant analysis prove to be highly accurate in predicting airline operating costs. The LDA performs slightly superior to the logistic model and both models are in the high 90% accuracy ranges.

## RESULTS OF ANALYSIS

| Linear Regression 1 (Full Model) | | Intercept | LOF | SOP | DFT | PS | RTPAM | TMLF | AC | TA | ISF |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Intercept | Length of Flight (miles) | Speed of Plane (mph) | Daily flight time per plane (Hours) | Population Served (1000s) | Revenue Tons per aircraft mile | Ton mile load factor (proportion) | Available Capacity (Tons per mile) | Total Assets ($100,000s) | Investments & Special funds ($100,000s) |
| | Coefficient | 1433.00 | 7.52E-01 | -2.253 | 3.317 | 3.11E-03 | 723.20 | -2175.00 | -382.10 | -0.52 | 1.96 |
| | P-Value | 3.90E-09 | 0.24371 | 0.06176 | 0.74932 | 0.06354 | 5.70E-08 | 1.23E-09 | 9.75E-09 | 1.19E-03 | 1.29E-02 |
| | R^2 | 0.9213 | | | | | | | | | |
| | AdjR^2 | 0.8875 | | | | | | | | | |
| | Std. Error | 1.49E+02 | 0.6273 | 1.142 | 1.03E+01 | 1.59E-03 | 8.84E+01 | 2.12E+02 | 4.20E+01 | 1.40E-01 | 7.21E-01 |
| 90% CI | High | 1.68E+03 | 1.79E+00 | -3.69E-01 | 2.02E+01 | 5.72E-03 | 8.69E+02 | -1.83E+03 | -3.13E+02 | -2.93E-01 | 3.15E+00 |
| | Low | 1.19E+03 | -2.83E-01 | -4.14E+00 | -1.36E+01 | 4.89E-04 | 5.77E+02 | -2.52E+03 | -4.51E+02 | -7.54E-01 | 7.71E-01 |

**Table 1) Linear Regression Full Model**

| Linear Regression 2 (Reduced Model) | | Intercept | LOF | SOP | DFT | PS | RTPAM | TMLF | AC | TA | ISF |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Intercept | Length of Flight (miles) | Speed of Plane (mph) | Daily flight time per plane (Hours) | Population Served (1000s) | Revenue Tons per aircraft mile | Ton mile load factor (proportion) | Available Capacity (Tons per mile) | Total Assets ($100,000s) | Investments & Special funds ($100,000s) |
| | Coefficient | 1163 | | | | | 688 | -2099 | -362 | -0.3383 | 1.475 |
| | P-Value | 5.65E-13 | | | | | 7.47E-09 | 4.59E-11 | 3.03E-09 | 4.00E-03 | 3.12E-02 |
| | R^2 | 0.883 | | | | | | | | | |
| | AdjR^2 | 0.8596 | | | | | | | | | |
| | Std. Error | 86 | | | | | 80.86 | 190.9 | 40.65 | 0.1067 | 0.6463 |
| 90% CI | High | 1304.9 | | | | | 821.419 | -1784.015 | -294.9275 | -0.162245 | 2.541395 |
| | Low | 1021.1 | | | | | 554.581 | -2413.985 | -429.0725 | -0.514355 | 0.408605 |

**Table 2) Linear Regression Reduced Model**

| Beta Regression | | Intercept | LOF | SOP | DFT | PS | RTPAM | TMLF | AC | TA | ISF |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Intercept | Length of Flight (miles) | Speed of Plane (mph) | Daily flight time per plane (Hours) | Population Served (1000s) | Revenue Tons per aircraft mile | Ton mile load factor (proportion) | Available Capacity (Tons per mile) | Total Assets ($100,000s) | Investments & Special funds ($100,000s) |
| | Coefficient | 0 | | | | | 5.747 | -2.047 | -4.294 | -0.9545 | 0.5041 |
| | P-Value | 1.00E+00 | | | | | 7.47E-09 | 4.59E-11 | 3.03E-09 | 4.00E-03 | 3.12E-02 |
| | R^2 | 0.883 | | | | | | | | | |
| | AdjR^2 | 0.8596 | | | | | | | | | |
| | Std. Error | 6.73E-02 | | | | | 6.75E-01 | 1.86E-01 | 4.81E-01 | 3.01E-01 | 2.21E-01 |

**Table 3) Beta Regression Reduced Model**

| Logistic | | Intercept | LOF | SOP | DFT | PS | RTPAM | TMLF | AC | TA | ISF |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Intercept | Length of Flight (miles) | Speed of Plane (mph) | Daily flight time per plane (Hours) | Population Served (1000s) | Revenue Tons per aircraft mile | Ton mile load factor (proportion) | Available Capacity (Tons per mile) | Total Assets ($100,000s) | Investments & Special funds ($100,000s) |
| | Coefficient | 2.66E+15 | | | | | -3.31E+15 | -5.93E+14 | 1.32E+14 | 6.09E+12 | -1.53E+13 |
| | P-Value | 2.00E-16 | | | | | 2.00E-16 | 2.00E-16 | 2.00E-16 | 2.00E-16 | 2.00E-16 |

**Table 4) Logistic Regression Reduced Model**

| Confusion Matrices (High/Low Operating Costs) | | | | | |
|---|---|---|---|---|---|
| | | Logistic Prediction | | LDA Prediction | |
| | | Low | High | Low | High |
| Actual | Low | 20 | 1 | 21 | 0 |
| | High | 1 | 9 | 1 | 9 |
| Accuracy | | 94% | | 97% | |

**Table 5) Confusion Matrices**

## CONCLUSION

Airlines can accurately predict their operating costs using continuous regression modeling techniques. It is also possible to predict outcome costs by grouping the dependent variable using binary modeling techniques such as logistic regression and linear discriminate analysis.

**APPENDIX**

**Data Sources:**

http://www.stat.ufl.edu/~winner/data/airline_costs.txt

http://www.stat.ufl.edu/~winner/data/airline_costs.dat

**Code:**

```
## Airline  Data Analysis
##### Part 1 ######
attach(Airline_Cost_Data_1)
summary(Airline_Cost_Data_1)
library(psych)
TABLE_DES<- describe(Airline_Cost_Data_1)
### Linear Regression Model 1 - All Variables
REGMODEL1 <- lm(TOC~LOF+SOP+DFT+PS+RTPAM+TMLF+AC+TA+ISF+AA)
summary(REGMODEL1)
## Issues with Multicollinearity
## Check Correlation  Matrix
REDUCED <- Airline_Cost_Data_1[,3:13]
cor(REDUCED)
## Reduced Full Model
RED_REGMODEL1 <-lm(TOC~LOF+SOP+DFT+PS+RTPAM+TMLF+AC+TA+ISF)
RED_REGMODEL1
summary(RED_REGMODEL1)
## Reducing again due to insignificant  variables
RED_REGMODEL2 <-lm(TOC~RTPAM+TMLF+AC+TA+ISF)
summary(RED_REGMODEL2)
## Predictions
PREDICION_RED_RM2 <- predict(RED_REGMODEL2)
RESIDUALS_RM2 <- TOC - PREDICION_RED_RM2
RESIDUALS_RM2

##PLOTS  to check if Polynomial  Regression is Suitable.  None appear to be.
plot(RTPAM,TOC)
plot(TMLF,TOC)
plot(AC,TOC)
plot(TA,TOC)
plot(ISF,TOC)

## Beta Regression
DATASET1  <- data.frame(TOC,RTPAM,TMLF,AC,TA,ISF)
Z_DATA1  <- data.frame(scale(DATASET1))
describe(Z_DATA1)
```

```
## Mean = 0, SD = 1
Z_REGMODEL2 <- lm(TOC~RTPAM+TMLF+AC+TA+ISF, data=Z_DATA1)
summary(Z_REGMODEL2)


######### PART 2 ##########
## Define New Variable #### 1 = High OC, 0 = Low OC
TOC_NEW <- ifelse(TOC>110,1,0)
### Logistic Regression

LOGITMODEL <- glm(TOC_NEW~RTPAM+TMLF+AC+TA+ISF, family=binomial())
summary(LOGITMODEL)
library(DiscriMiner)
logLik(LOGITMODEL)
predict(LOGITMODEL)
LOGITS <- data.frame(predict(LOGITMODEL))
ODDS <- exp(LOGITS)
PROBABILITIES <- ODDS/(ODDS+1)
describe(PROBABILITIES)
### INF = Probability of 1
### 0 = Proability of 0
PREDICTION_LOG_REG <- ifelse(ODDS == 0,0,1)
PREDICTION_LOG_REG
data.frame(PREDICTION_LOG_REG)
COMPARISON <- data.frame (TOC_NEW,PREDICTION_LOG_REG)
TABLE <- table(COMPARISON)
TABLE

#### Linear Discriminant Analysis (LDA)

LDA_MODEL <- linDA((Airline_Cost_Data_1)[,8:12],TOC_NEW)
LDA_MODEL
LDA_MODEL$functionsL
LDA_MODEL$scores
LDA_MODEL$classification
PREDICTION_LDA <-data.frame(LDA_MODEL$classification)
COMPARISON_2 <- data.frame(TOC_NEW,PREDICTION_LDA)
TABLE_2 <- table(COMPARISON_2)
TABLE_2
```