SAS Data Analysis Course Project

By Ari Akerstein, Gabriel Mohanna, and Tina Park

Contents

Abstract

The purpose of this project is to examine the major factors affecting the cost of flying a passenger per mile (CPM). We examined 6 different independent variables and 24 derived variables and finally settled on 3 different OLS models based on length of flight and the size of the aircraft.

We found that the most important factors in the data are the load factor and the number of seats per air craft in various interactions with themselves and other variables. Below is a summary of our study.

Executive Summary

After careful consideration, we concluded that for the purpose of finding the best fitting models, the data needed to be broken into three different groups, based on type of flight (long range or short range) and the size of the plane.

Long-Range Flights (ASL ≥ 1,200 Miles)

Adj R-Sq	F-ratio	Pr > F	# Obs	Parameters:	Intercept	utl_alf	invspa
0.9427	99.64	<.0001	13	Coefficients	5.74931	-0.95781	0.14205
				P-value	<.0001	<.0001	0.0009
				Vif		1.01722	1.01722

Short-Range Flights (ASL < 1,200 Miles) & Small Air Plans (SPA < 0.202)

R-Sq	F-ratio	Pr > F	# Obs	Parameters:	Intercept	lspa_alf
0.9019	73.51	<.0001	10	Coefficients	-5.94547	-3.28214
				P-value	0.0005	<.0001
				Vif		1

Short-Range Flights (ASL < 1,200 Miles) & Large Airplanes (SPA ≥ 0.300)

R-Sq	F-ratio	Pr > F	# Obs	Parameters:	Intercept	lspa_alf
0.707	14.48	0.0089	8	Coefficients	-1.6152	-2.19638
				P-value	0.22	0.0089
				Vif		1

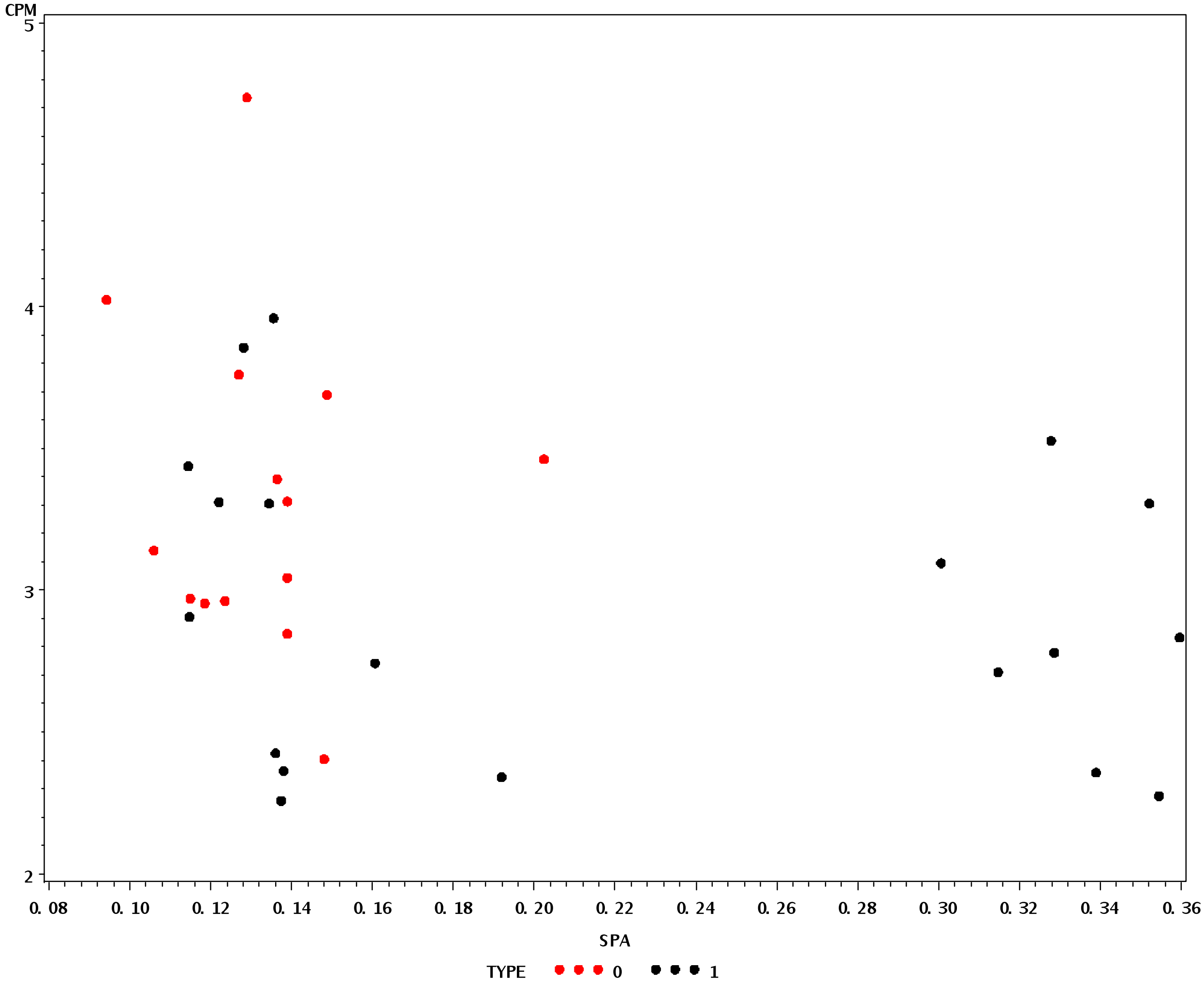
There are three main factors in long-range flight; namely, number of hours the aircraft is in the air, the load factor, and number of seats the aircraft has, which is an indication of how large the plane is. The model above represents the best combination of these three variables in a way to minimize the possible interactions and maximize the model performance. The model above says that the longer the aircraft is in the air, the less the cost per mile is. This is the same for the load factor and size of the plane. We did notice, however, that the data suggest high interaction between UTL and ALF and, hence, we used their interaction variable (UTL_ALF = UTL x ALF).

As for the short-range flight the derived variable SPA_ALF explains that the data in the best possible way. SPA_ALF = SPA x ALF stands for the occupation % of each aircraft. This simply means that the more occupied an airplane is, the lower the cost per mile per passenger and this makes intuitive sense. We do want to point out that large airplanes have slightly different coefficients than small airplanes. The rate of decrease of CPM for small airplane is steeper than large airplanes. Our recommendation here is for airlines to use small airplanes as it decrease the cost per passenger per mile more than large airplanes. This makes sense since large airplanes probably consume more fuel than small ones.

All the models above and their coefficients are significant to the 99% level. We did exclude two points from the models and we'll explain how and why in project detailed work below.

Project Detailed Work

First, it is very important to show why we split the data into three groups. The following scatter plot shows the different groups of long-range flights (red) and their short-range counter parts (black). Notice how there are two different groups for short-range flights suggesting a different model for each.



After making this discovery, we ran proc means on all the variables to see if there is a difference in their means especially when it comes to the CPM. We then ran proc insight to quickly graph how different variables affect the CPM. The following variables stood out as possible interactions variables:

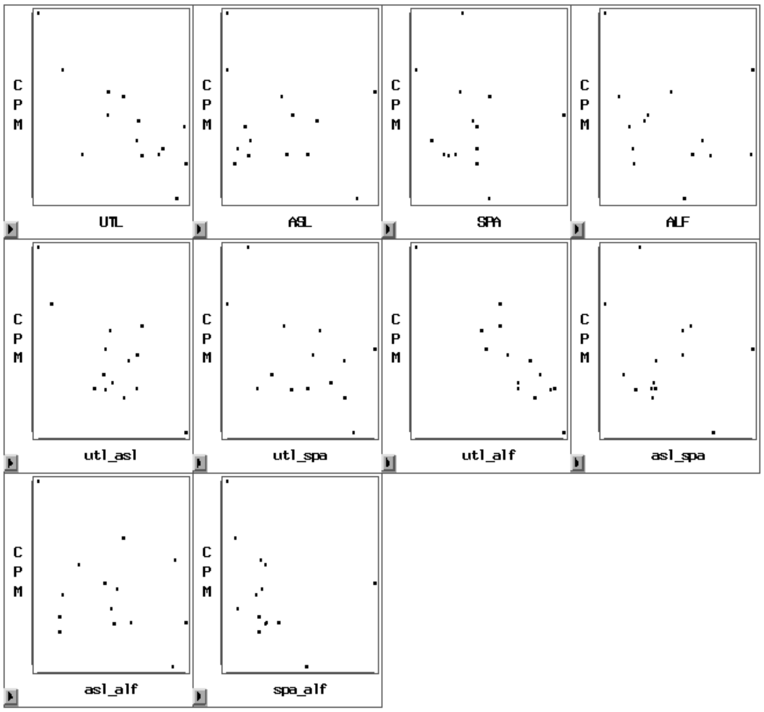
Type=0, Size=0: UTL, UTL_ALF and SPA_ALF

Type=1, Size=0: UTL, SPA, ASL_ALF and SPA_ALF

Type=1, Size=1: ALF, UTL_ASL, UTL_SPA, UTL_ALF, ASL_SPA, ASL_ALF and SPA_ALF

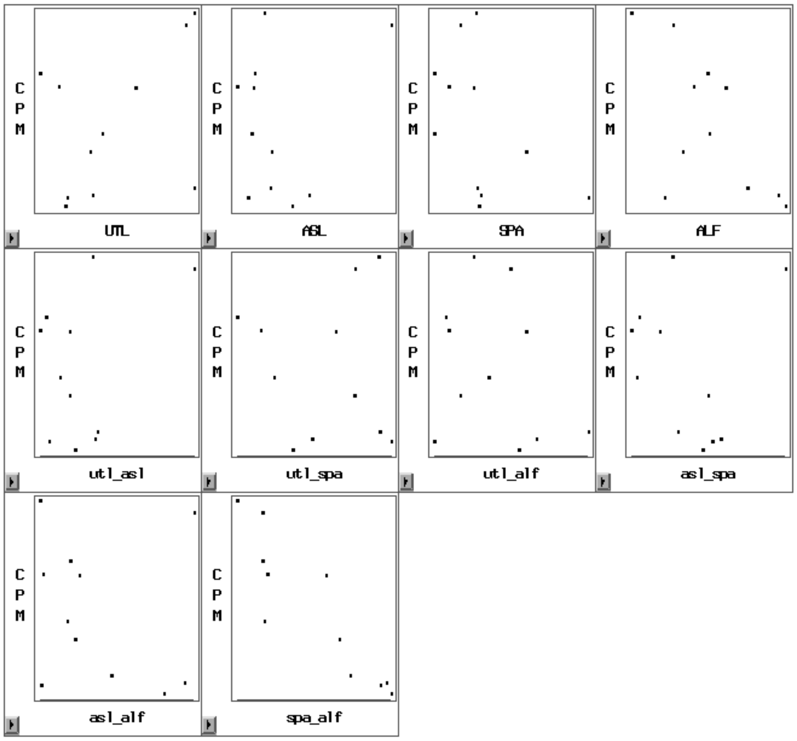
TYPE = 0

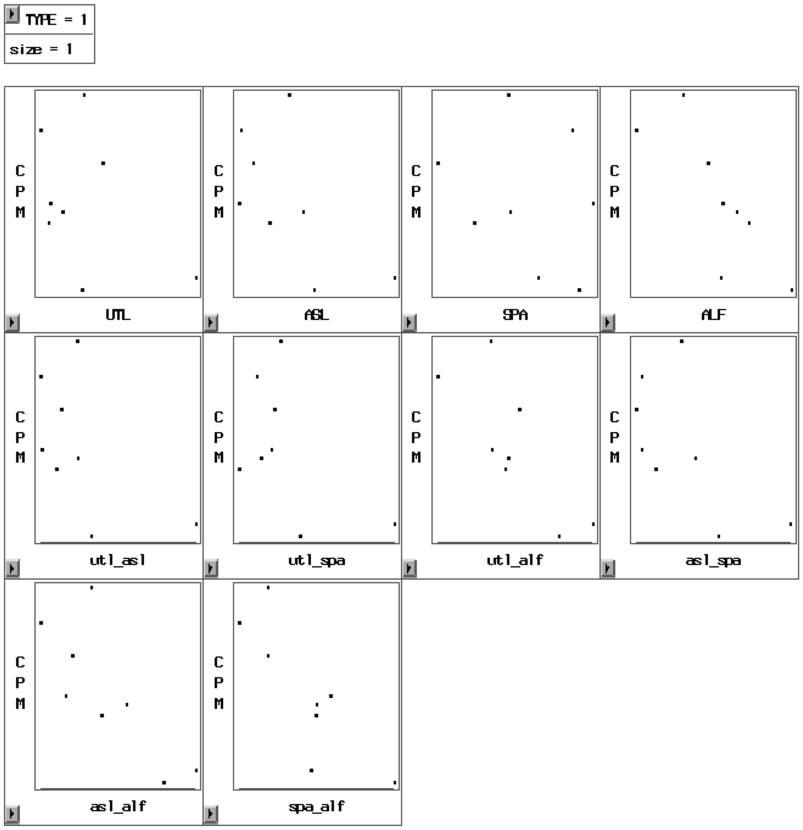
size = 0



TYPE = 1

size = 0





We only ran proc insight to get a good idea what to focus on. However, most of our work depended on stepwise selection along with diagnostic checks to determine possible outliers.

Model Diagnostic Checks

We relied on five diagnostics to check our models:

1. F-ratio to see how good our model predictive power is
2. P-value of parameter estimates
3. Variance inflation factors
4. Student Residuals
5. Cook's D
6. Adjusted R squared

The first two variables are explained above and they are all valid to the 99% confidence level.

The variance inflation factors also look very favorable as they are all close to 1.

Below is a list of Cook's D values before and after high affecting observations were removed. We used a rule of thumb of 4/n, n is # of observations, to remove certain observations. We noticed that all high residuals disappear when we follow that process.

Long-Range Flights (ASL ≥ 1,200 Miles)	Long-Range Flights (ASL ≥ 1,200 Miles)
All Data Included	Observation 5 Excluded

```
                      Cook's                                     Cook's
  Obs   -2-1 0 1 2            D            Obs   -2-1 0 1 2            D

    1 |    *|     |     0.045                1 |    *|     |     0.110
    2 |    *|     |     0.019                2 |   **|     |     0.052
    3 |     |     |     0.006                3 |     |     |     0.000
    4 |     |*    |     0.056                4 |     |*    |     0.062
    5 | ****|     |     0.231                5 |   **|     |     0.070
    6 |    *|     |     0.029                6 |    *|     |     0.055
    7 |     |     |     0.002                7 |     |***  |     0.137
    8 |     |***  |     0.087                8 |     |     |     0.001
    9 |     |     |     0.000                9 |     |*    |     0.130
   10 |     |     |     0.038               10 |     |     |     0.001
   11 |     |     |     0.000               11 |     |**   |     0.047
   12 |     |**   |     0.035               12 |     |     |     0.046
   13 |     |*    |     0.232               13 |    *|     |     0.292
   14 |     |     |     0.037
```

Short-Range Flights (ASL < 1,200 Miles) & Small Air Plans (SPA < 0.202)

All Data Included

```
                     Cook's
 Obs   -2-1 0 1 2             D

   1 |      |      |     0.005
   2 |      |      |     0.004
   3 |      |      |     0.002
   4 |     *|      |     0.044
   5 |      |      |     0.000
   6 |  ****|      |     0.449
   7 |      |***   |     0.154
   8 |      |      |     0.019
   9 |      |      |     0.004
  10 |      |**    |     0.199
  11 |      |*     |     0.094
```

Short-Range Flights (ASL < 1,200 Miles) & Small Air Plans (SPA < 0.202)

Observation 6 Excluded

```
                     Cook's
 Obs   -2-1 0 1 2             D

   1 |      |      |     0.006
   2 |      |      |     0.006
   3 |      |      |     0.005
   4 |    **|      |     0.100
   5 |      |      |     0.000
   6 |      |****  |     0.240
   7 |    **|      |     0.157
   8 |     *|      |     0.097
   9 |      |**    |     0.220
  10 |      |      |     0.001
```

Short-Range Flights (ASL < 1,200 Miles) & Large Airplanes (SPA ≥ 0.300)

All Data Included

```
                     Cook's
 Obs   -2-1 0 1 2             D

   1 |      |      |     0.027
   2 |   ***|      |     0.250
   3 |      |      |     0.003
   4 |      |      |     0.001
   5 |      |*     |     0.072
   6 |      |      |     0.006
   7 |     *|      |     0.152
   8 |      |***   |     0.383
```

Appendix

Variable name:

Long-Range Flights (Type=0)

We find two predictive variables for modeling long-range flights, interaction variable UTL_ALF and SPA. UTL-ALF describes the average load factor crossed by the day use hours per aircraft. This can be interpreted as load factor hours. The other variable, SPA, is the size of the aircraft.

Objective
The purpose of this project is to examine the major factors affecting the cost of flying a passenger per mile (CPM). The source of the data is from the Civil Aeronautics Board Report, Aircraft Operating Costs and Performance Report (August 1972) provided by Jianmin Liu.

Variables
The variables we examined were the following:

- Average hours per day use of aircraft (UTL)
- Average length of nonstop flights (ASL) per 1000 miles
- Average number of seats per aircraft per 100 seats (SPA)
- Average load factor - defined as the percentage of seats occupied by passengers (ALF)
- An indicator variable of whether the average length of a nonstop flight is greater or equal than 1200 miles (Type=0) or less than 1200 miles (Type=1)

Furthermore, we examined the interactions between the above variables to rule out any collinearity and transformed the data by taking the log and inverse of each variable and interaction variable for a total of 30 different variables.

The first group is Type=0, or greater than 1200 miles. The second group is less than 1200 miles. In our analysis, we noticed that in short-range flights two distinct clusters of data emerged, with respect to the average number of seats per aircraft (SPA), an indicator of how large the aircraft is. The explanation for this is that shorter range flights have a greater cost (com) volatility than do the longer flights. Upon further examination, we decided to further break up the short-range flights according to this pattern - size=0 describes smaller aircrafts of approximately 200 seats or less among short-range flights, and size=1 describe the larger aircrafts of approximately 300 seats and greater among short-range flights. Thus we conclude that the best methodology for describing the CPM is with three distinct OLS models. There are no large aircrafts that fly short-range flights.

P-Values
For UTL-ALF, the p-value is <0.0001. For inverse of SPA, the p-value is 0.0009

Short-Range Flights (Type=1)
We find that the variable with greatest predictive power on CPM is the interaction variable SPA-ALF. This effectively means that the interaction between the average load factor and the size of the aircraft determines the CPM above all other variables considered. While the variable used is the same, our results suggest that there is a three-fold magnitude increase on smaller planes than on larger planes, with respect to this variable.

P-Values

Size 0: for SPA-ALF, the p-value is
Size 1: for SPA-ALF, the p-value is

Model Type

The dependent variable has a continuous outcome, therefore we attributed linear regression as the best regression analysis.

Outliers

NA

Equations of Models

Per our summary, the following table illustrates the equations of the models we obtained: