A REGRESSION ANALYSIS OF AIRLINE COSTS

Odysseas Chaliotis
odysseas.chaliotis@epfl.ch

Maksim Kriukov
maksim.kriukov@epfl.ch

January 10, 2022

1 Introduction

The analysis of the variables that influence airline costs present significant interest, especially due to their extensive implications in flight- and cost-related policies [1]. This report describes the analysis of the following variables, along with its potential influence on the flights' operational costs: (1) length of flight, (2) speed of plane, (3) daily flight time per aircraft, (4) population served, (5) ton-mile load factor, (6) available tons per aircraft mile, and (7) firms net assets. We present the linear regression analysis that explores the potential relations of each of the variables mentioned above and their respective significance in deducing the operational costs of a flight.

2 Data and Method

We are provided with a data set from J.W. Proctor and J.S. Duncan (1954) [1] that contains the following sample variables and their respective units: Airline names, Length of flight (miles), Speed of Plane (miles per hour), Daily Flight Time per plane (hours), Population served (1000s), Total Operating Cost (cents per revenue ton-mile), Revenue Tons per Aircraft mile, Ton-Mile load factor (proportion), Available Capacity (tons per mile), Total Assets ($100,000s), Investments and Special Funds ($100,000s) and Adjusted Assets ($100,000s).

2.1 Selection of variables

Proctor and Duncan were guided in their selection according to the variables' implications on air transport policies [1]. Therefore, the available capacity of the plane is an important aspect of the operational cost, in the sense that the purchase and operation of a large and a small plane cost almost comparable amounts. The influence of the length of the flight on the operational cost appears to be prominent, according to the authors [1]. Additionally, the time that a plane is in operation is of significance when deducing the operational costs, in the sense that greater utilization yields lower unit costs. Furthermore, the airplane's speed is argued to be an important aspect of airline travel costs [1].

As suggested in [1], for the purpose of linearization of explanatory variables the natural logarithms of each of the variables were applied, with the exception of the ton-mile load factor. A preliminary inspection of the underlying correlations between the attributes of this problem is provided in the pairwise plot below (fig. 1). We can observe that Adjusted_Assets feature is linearly dependent on the Total_Assets and Funds in the sense that:

$$\text{Adjusted\_Assets} = \text{Total\_Assets} - \text{Funds} \tag{1}$$

hence we chose to eliminate Total_Assets before the regression analysis.

The second variable that needed to be excluded is Revenue. The first reason is that Revenue has suspiciously high negative Pearson correlation value with Cost ($r_{\text{Cost}}^{\text{Revenue}} = -0.97$). We assume that Revenue compensate for the operational costs, and, therefore, should be inversely proportional to the variable of interest. Since

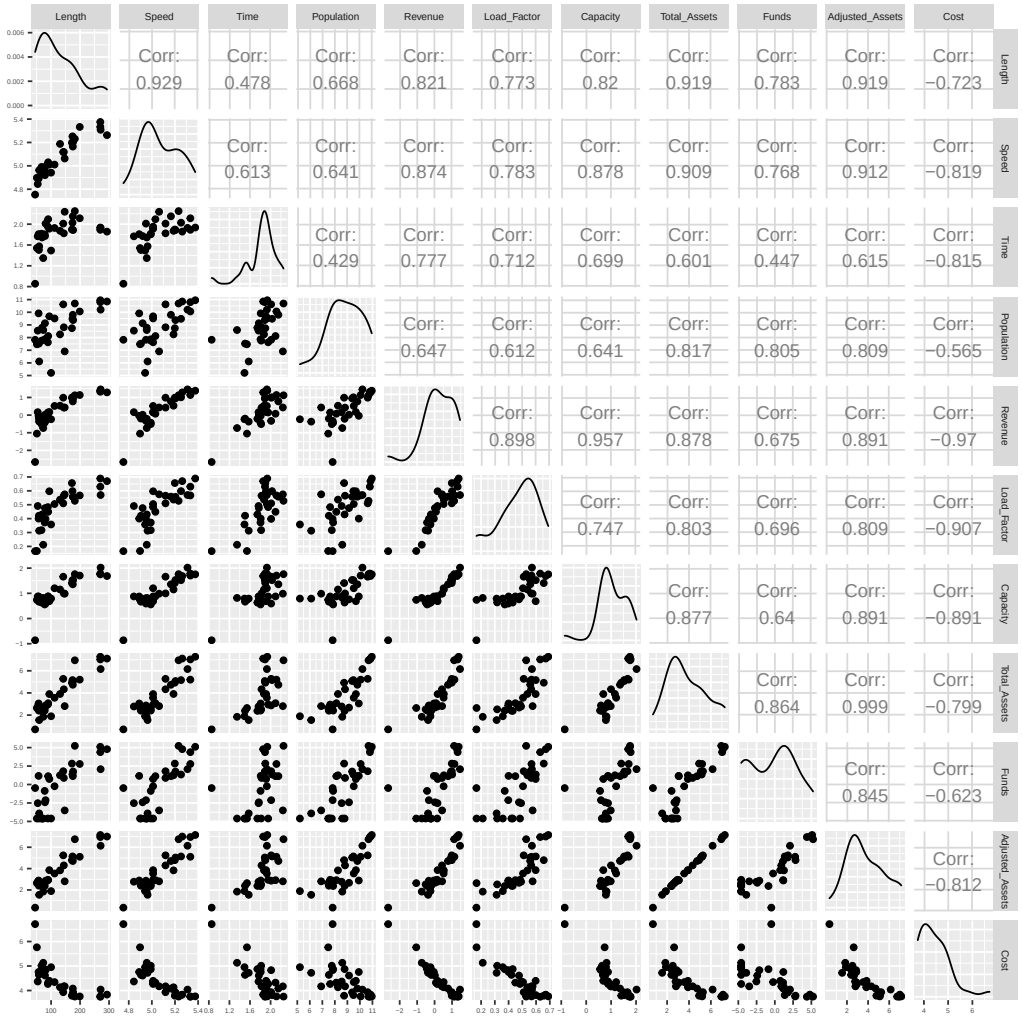

Figure 1: Pairwise plots for all input variables. The histogram distribution are shown on the diagonal of matrix plot, the scatter plots for all pairs of variables are presented in the lower triangle and their Pearson correlation values are presented in the upper triangle. All the variables except for the Load_Factor were log-transformed for better representation.

revenue cannot be known before deducing the operational costs it is neglected from the preliminary regression. Moreover, [1] suggests that Revenue is dependent on the Load_Factor and Capacity variables as follows:

$$\text{Capacity} = \frac{\text{Revenue}}{\text{Load\_Factor}} \tag{2}$$

All these dependencies are additionally represented in the heat map plot (fig. 2). We can note that Total_Assets, Revenue and Length are accordingly highly correlated with Adjusted_Assets, Capacity and Speed. The latter association could be explained by the fact that long flights have higher average flight speed as the plane is cruising longer at high altitudes.

As previously discussed, to construct an accurate linear prediction model one should exclude all dependencies in the input variables. After applying this filtering, our preliminary model would include the following features:

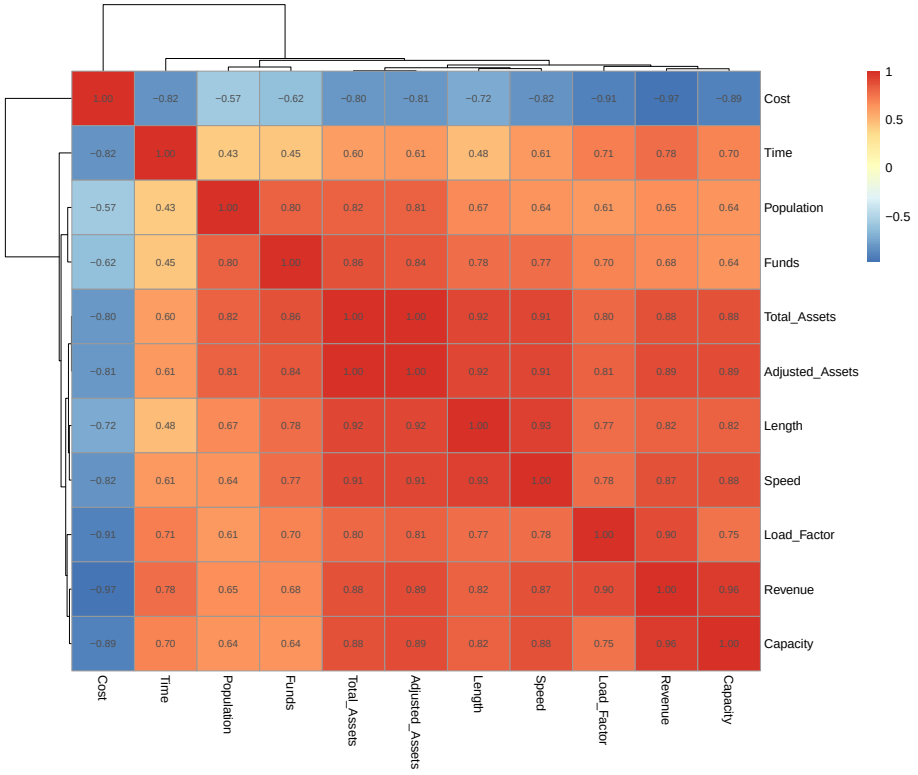

Figure 2: Heat map plot indicating the Pearson correlation between the pairs of variables. All the highly-correlated cells were clustered together for better representation. All the variables except for the Load_Factor were log-transformed for better representation.

- $x_1 = \ln(\text{Operational\_Cost})$
- $x_2 = \ln(\text{Capacity})$
- $x_3 = \text{Load\_Factor}$
- $x_4 = \ln(\text{Length})$
- $x_5 = \ln(\text{Speed})$
- $x_6 = \ln(\text{Time})$
- $x_7 = \ln(\text{Adjusted\_Assets})$
- $x_8 = \ln(\text{Funds})$

2.2 Preliminary regression model

After the pre-processing step and the explanatory data analysis, our data was fitted to multiple linear regression model, which is defined as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i \qquad i = 1, \ldots, n, \tag{3}$$

where $x_{i1}, \ldots x_{ip}$ are p regressors for each statistical unit i, y_i - the predicted variable that is linearly dependent on the regressors, ε_i - error variable which adds "noise" to the linear equation.

Given that $\varepsilon_1, \ldots, \varepsilon_n \sim$ i.i.d $N(0, \sigma^2)$ we could perform a statistical test and identify the confidence intervals for the coefficients.

The overall summary of the fitted preliminary model is described in table 1. The results indicate that Length, Load_Factor, Capacity and Intercept variables are most significant with low standard error and low p-value for the statistical t-test. The majority of previous studies [1] highlights the importance of load factor as airlines would make more profit with fully occupied airplanes. The preliminary model supports this notion by having the highest negative estimate of the Load_Factor coefficient. The Length and Time variables have higher standard errors and p-values, while Speed, Population, Funds and Adjusted_Assets variables are not significant and should be excluded in the final model. Indeed, increased utilization of the airplanes (Time variable) could lead to a more optimized use of airplanes and lower cost, while length of flight (Length

Table 1: The summary of the preliminary linear model.

COEFFICIENTS:	ESTIMATE	STD. ERROR	t VALUE	Pr(> |t|)	LEVEL OF SIGNIFICANCE
(Intercept)	8.698	2.599	3.346	2.930 e-03	Moderate
Length	0.385	0.181	2.124	4.520 e-02	Low
Speed	-0.779	0.602	-1.293	2.094 e-01	Not significant
Time	-0.313	0.148	-2.115	4.603 e-02	Low
Population	0.036	0.037	0.968	3.435 e-01	Not significant
Load_Factor	-3.023	0.395	-7.651	1.231 e-07	High
Capacity	-0.757	0.132	-5.726	9.260 e-06	High
Funds	-0.017	0.018	-0.903	3.763 e-01	Not significant
Adjusted_Assets	0.094	0.065	1.458	1.589 e-01	Not significant

Residual standard error:	0.129 on 22 degrees of freedom		
Multiple R-squared:	0.972	Adjusted R-squared:	0.961
F-statistic:	93.670 on 8 and 22 DF	p-value:	3.416 e-15

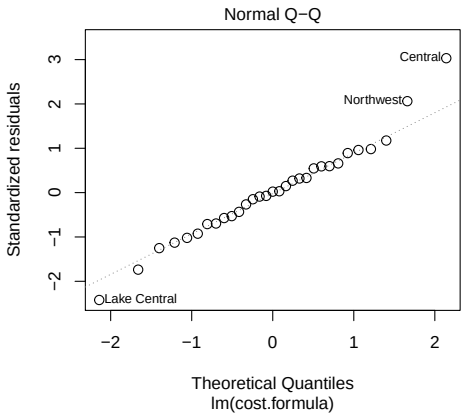

Figure 3: Quantile-quantile plot which compares the data distribution with normal distribution. Each dot represents an airline. In this case The Lake Central, Northwest and Central airlines are outliers.

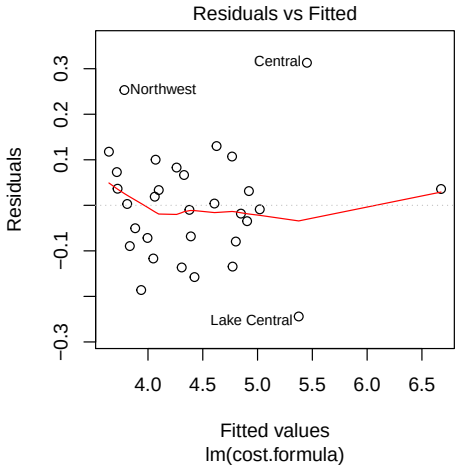

Figure 4: The residuals plot which highlights difference between real and predicted data points. Each dot represents an airline. In this case The Lake Central, Northwest and Central airlines are outliers.

variable) is positively associated with gasoline consumption and, therefore, with the operating cost. It was shown in the literature, that the size of the airline is unimportant (Funds and Adjusted_Assets variables) [1], and Speed and Population variables could differ greatly between airlines and not reliable for predicting the cost.

Interestingly, our regression model suggests that the length of the flight truly is a significant factor of operational costs, contrary to the results of Proctor and Duncan who excluded it from their final models, despite their initial hypothesis [1]. At that point, this contradiction was mainly justified as a lack of the sufficient number of data points and exclusion of more technical attributes of the planes and flight firms. Now, this claim -coupled with our findings- may also be elucidated by the evolution in regression algorithms that has led to fewer data points yielding more representative models of their respective data sets.

The total evaluation of the model fitness is done by calculating R-squared, adjusted R-squared metrics and statistical F-test, which unanimously show good accuracy of the model. To identify the presence of outliers, we plotted normal QQ plot with the plot that shows the relationship between residuals and fitted points 4. We can observe the correctness of the previous assumption that the data is normally distributed. The plots also indicate the presence of 3 outliers: The Lake Central, Northwest and Central airlines. After careful inspection of the raw data, we decided to keep these airlines in the final model.

3 Final regression model

As described, our preliminary analysis concluded on four variables majorly determining the operational cost of flights, namely Capacity, Length, Time and Load_Factor. In the previous section, we showed that negative

correlations between Costs and either of these variables are expected in a sense that the greater these attributes, the more cost-effective the flight [1].

After excluding insignificant variables, another linear model was fitted from which we get the metrics and results of Table 2. This model corresponds to the following logarithmic expression of operational flight costs:

$$\ln(\text{Cost}) = 5.794 - 0.633 \cdot \ln(\text{Capacity}) - 2.731 \cdot \text{Load\_Factor} - 0.382 \cdot \ln(\text{Time}) + 0.279 \cdot \ln(\text{Length}) \quad (4)$$

Table 2: The summary of the final linear model.

COEFFICIENTS:	ESTIMATE	STD. ERROR	t VALUE	Pr($>\|t\|$)	LEVEL OF SIGNIFICANCE
(Intercept)	5.794	0.498	11.644	8.150 e-12	High
Capacity	-0.633	0.105	-6.010	2.400 e-06	High
Load_Factor	-2.731	0.407	-6.717	3.980 e-07	High
Time	-0.382	0.157	-2.430	2.230 e-02	Low
Length	0.279	0.119	2.343	2.700 e-02	Low

Residual standard error:	0.151 on 26 degrees of freedom		
Multiple R-squared:	0.954	Adjusted R-squared:	0.947
F-statistic:	134.600 on 4 and 26 DF	p-value:	2.200 e-16

Our simplified model shows comparable results with statistical significance for all variables $\alpha = 0.03$ and satisfactory overall statistics. The advantage of the final model is its simplicity and interpretability of the dependencies between variables. Again we may note the crucial importance of load factor for predicting the operating cost of the airplane flights. These results could find various applications in the optimization of aircraft utilization and profit increasing strategies.

References

[1] J.W. Proctor and J.S. Duncan. A Regression Analysis of Airline Costs. In *Journal of Air Law and Commerce, Vol.21, 3*, pages 282–292, 1954.