Report 1: Airline Costs analysis

Introduction/background

This project aims to explore the relationships between airline costs and multiple factors enumerated down below. This will be done through regression analysis. The data set consists of 11 variables extracted from 31 airlines. The dependent variable y is the total operating costs (TOC). Its unit of measurement is cents per revenue ton-mile. In order to model the TOC, the following 10 variables were analyzed: the length of the flight, the speed of the plane, the daily flight time, the metropolitan population served, the revenue tons per aircraft mile, the ton-mile load factor, the available capacity, the total assets, the investments and special funds, and the adjusted assets. Two variables were determined by other variables: the available capacity corresponds to the revenue tons per aircraft mile divided by the ton-mile load factor, and the adjusted assets are equal to the difference between the total assets and the investments and special funds.

Table 1: Abbreviations of the variables.

FL	Length of flight (miles)
SoP	Speed of plane (miles/hr)
DFT	Daily flight time (hrs)
PS	Population served (thousands)
TOC	Total operating costs (cents per revenue ton-mile)
RTM	Revenue tons per aircraft mile
LF	Ton-mile load factor (proportion)
C	Available capacity (tons per mile)
TA	Total assets ($100,000s)
I	Investments and special funds ($100,000s)
AA	Adjusted assets ($100,000s)

Exploratory data analysis

Univariate graphical

We can see in the left boxplot that the raw data is unequally distributed; it does not have the same magnitude of values. This causes an issue when building a model because some factors, due to the fact that they have bigger values, can hold a non-realistic weight and thereby lead to a biased regression. In order to get the data in the same order of magnitude, we decided to take the logarithm of each variable. Moreover, due dependencies explained in the introduction, RTM, LF, TA and I are removed for the following analysis. In this manner, six independent variables are used to explained the total operating cost.

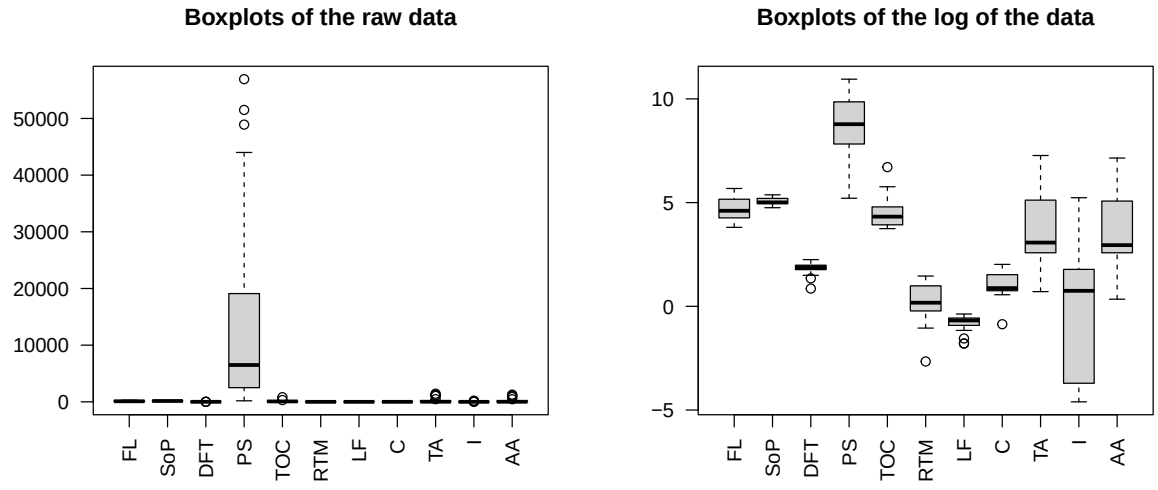

Figure 1: Boxplots of the raw data and the logarithm of the data.

Univariate numerical

Table 2: Mean, standard deviation (SD) and median absolute deviation (MAD) of the variables.

	FL	SoP	DFT	PS	TOC	C	AA
Mean	4.71	5.07	1.83	8.81	4.43	1.06	3.79
SD	0.56	0.16	0.28	1.43	0.66	0.57	1.78
MAD	0.78	0.16	0.18	1.51	0.64	0.41	1.57

Bivariate numerical (correlations)

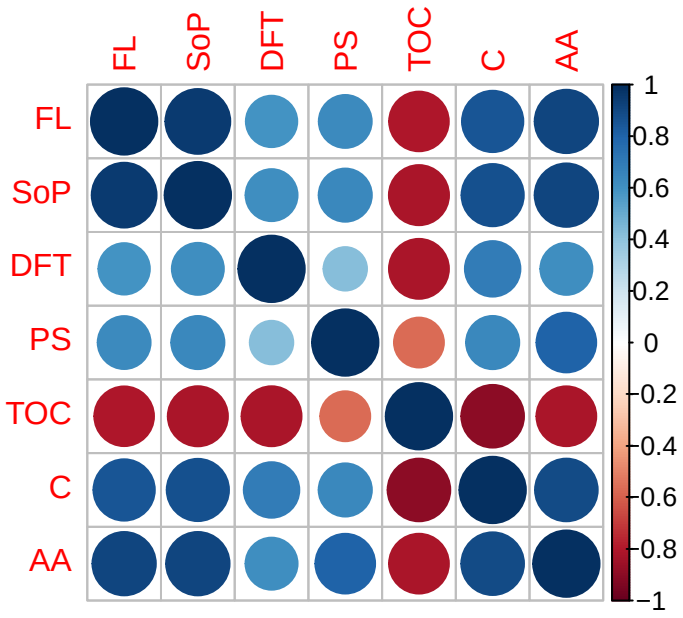

Figure 2: Correlation values between each feature.

This diagram demonstrates the correlation between each feature. The color blue illustrates a positive correlation, whereas the color red shows a negative correlation. Negative correlations are visible between the total operating costs and the six other variables. The TOC seems to decrease when the other factors increase. The correlation between the total operating cost and the population served is lower.

Model fitting

According to the summary, the equation of our regression model is the following:

$TOC = \hat{\beta}_0 + \hat{\beta}_1 \times FL + \hat{\beta}_2 \times SoP + \hat{\beta}_3 \times DFT + \hat{\beta}_4 \times PS + \hat{\beta}_5 \times C + \hat{\beta}_6 \times AA$

$TOC = 8.13 + (-0.18 \times FL) + (-0.15 \times SoP) + (-0.88 \times DFT) + (0.01 \times PS) + (-0.56 \times C) + (0 \times AA)$

$R^2 = 0.87$ and adjusted $R^2 = 0.84$

From the initial hypotheses ($H_0 : \beta_0 = ... = \beta_6 = 0$ and H_1 : at least one non-zero parameter), the p-value is equal to $1.0973324 \times 10^{-9} < \alpha = 0.05$. The null hypothesis is therefore rejected and we know that at least one coefficient does play a role in predicting the dependent variable. The values of the R^2 and adjusted R^2 are close to 1, which means that the variables are able to closely predict the dependent variable TOC.

The p-values of the different features are the following:

Table 3: P-values

Intercept	FL	SoP	DFT	PS	C	AA
0.09	0.56	0.89	0	0.87	0.02	0.98

Two features show significant results. These are the daily fight time (DFT) and the available capacity (C), which have a p-value equal to 0.0011 and 0.016, respectively. Additionaly, both variables have an high correlation with the total operating cost, as shown in the figure 2.

Model assessment

From these different plot, the residuals can be evaluated. Many assumptions must be controlled. Firstly, the Normal QQ plot illustrates the normal distribution. In this case, despite the central airlines data, the standardized residuals are mostly aligned on the theoretical straight line. This confirms that the errors follow a normal distribution.

The second assumption is that the errors have a mean equal zero. This can be observed on the residuals vs. fitted plot. Indeed, the residual values are located around 0 with small variation, less than 0.4. This observation means that no non-linear relation are present in the fitted model. From the same plot, the uncorrelated residues can be assumed because the values appears randomly plotted and no trend is observable.

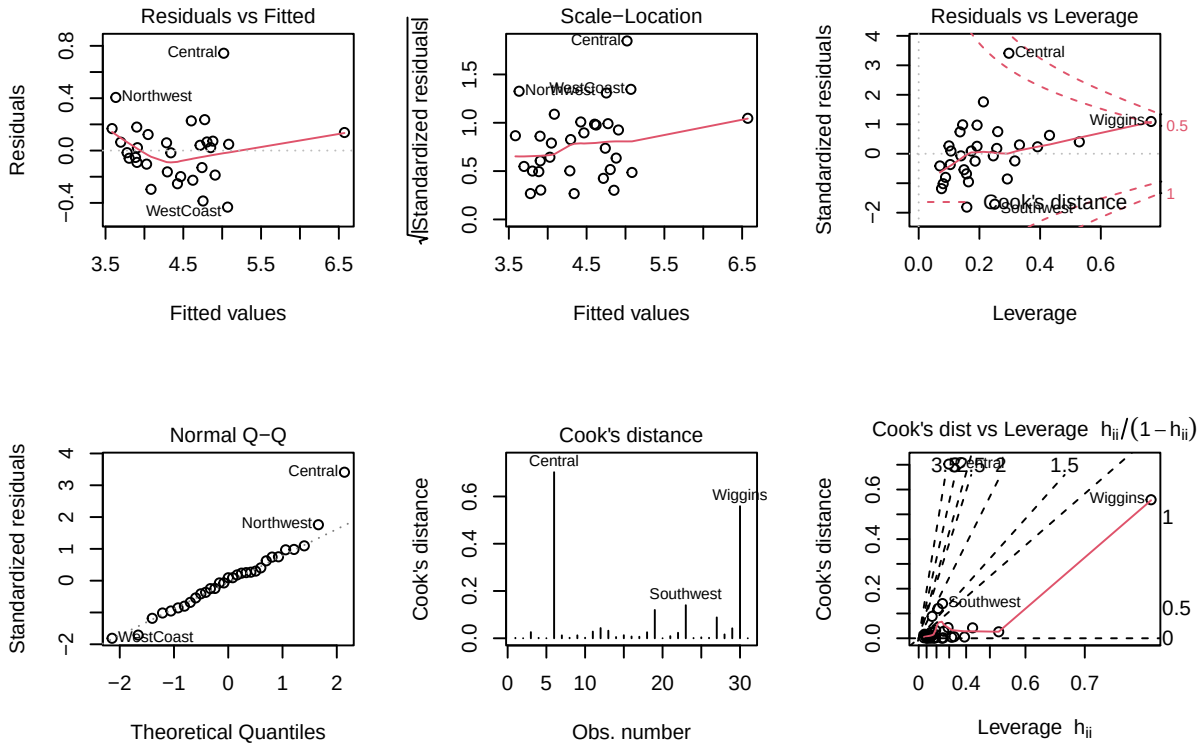

Figure 3: Residual analysis.

Concerning the homoscedastic property, the scale-location plot helps to determine it. Indeed, a horizontal line affirms the homoscedasticity of the residues. In this model, the points are effectively randomly spread and form a horizontal line.

Finally, residuals vs. leverage and cook's distance plots allow to identify potential influencor outliers. No airline errors are higher than 1. Central data seems to have a small deviation taht can influence the regression, but it stays reasonable.

Final estimated model

Conclusions