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 taken in 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: Abreviations 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 the 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 therefore, it leads 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, except for the LF variable. This variable already has a low magnitude (mean=0.48) and is therefore within the range of magnitude of the transformed variables.

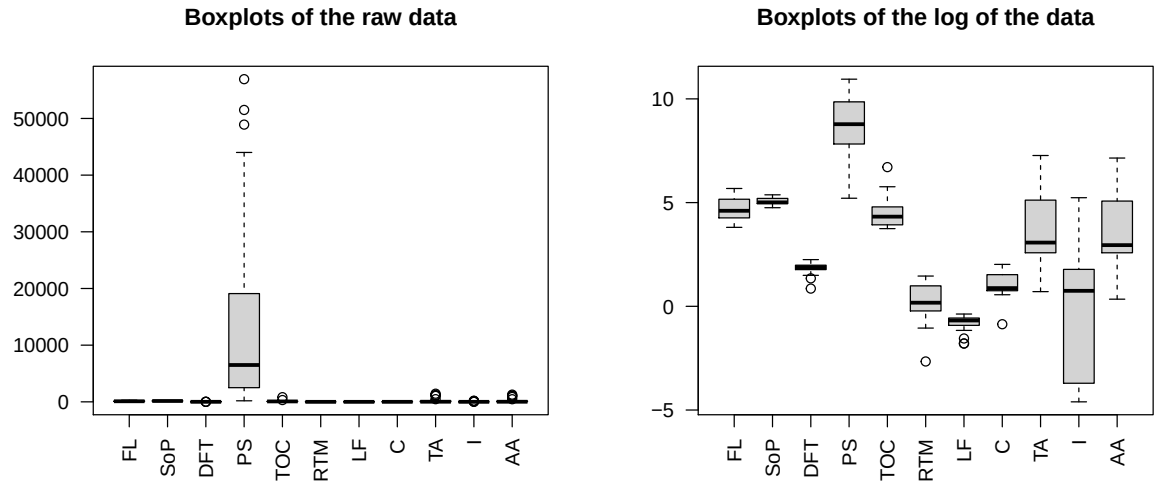

Figure 1: Boxplots of the raw data and the logarithm of the data.

Univariate numerical

Table 2: Mean, standard deviation and median absolute deviation of the variables.

	FL	SoP	DFT	PS	TOC	RTM	LF	C	TA	I	AA
Mean	4.71	5.07	1.83	8.81	4.43	0.26	0.48	1.06	3.84	-0.32	3.79
SD	0.56	0.16	0.28	1.43	0.66	0.87	0.14	0.57	1.79	3.27	1.78
MAD	0.78	0.16	0.18	1.51	0.64	0.89	0.12	0.41	1.75	4.25	1.57

Bivariate numerical (correlations)

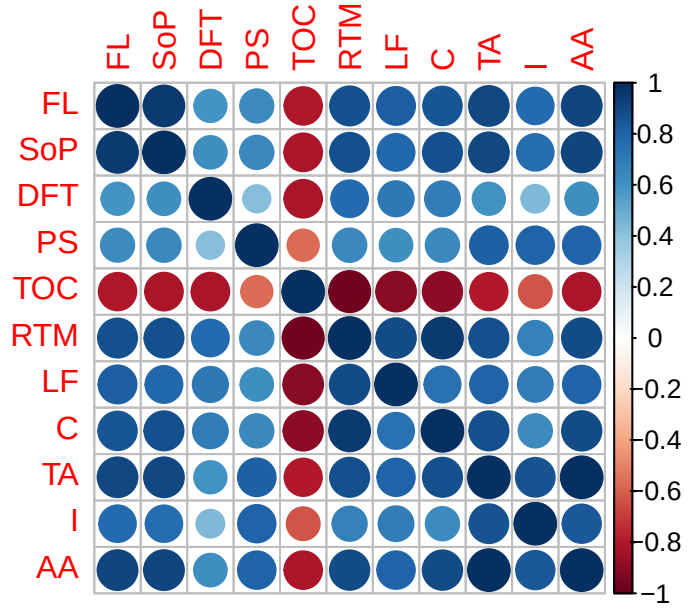

Figure 2: Correlation values between each feature.

Model fitting and model assessement

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. In the summary of the linear regression, we can see that the variable RTM has the lowest p-value (=0.040 < 0.05). RTM seems to be the variable that can predict the TOC the best. If we look at figure 2, we can confirm that RTM is one of the most important variable to predict TOC because it has the most prominent colored point indicating a strong correlation between the two variables.

Model assessment

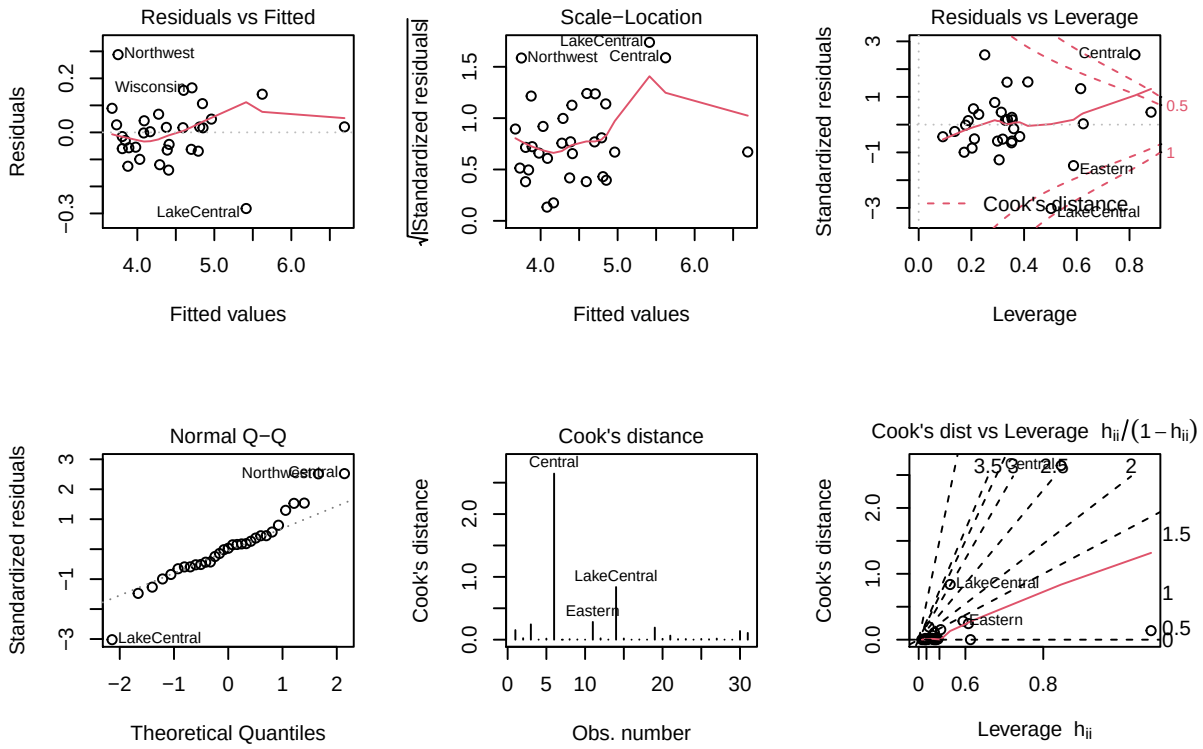

Figure 3: Residual analysis.

Assumptions (from report criteria): 1. errors have mean 0 2. errors are homoscedastic (same variance) 3. errors are uncorrelated 4. errors are normally distributed

FOR LOG DATA: Normal QQ plot shows that absolute values of standardized residuals reach approximately 3 at most, and that no particular deviation from theoretical normal distribution (data follow the theoretical straight line).

From Residuals vs. fitted plot, we can observe that residuals are spread around a horizontal line at approx. 0. No non-linear relation are present in the model.

Scale-location allows to observe homoscedasticity. Until fitted value 5, the points are randomly spread and the line is horizontal which affirms the homoscedasticity. After 5, there is some variances, however in a general view the line is quite horizontal.

Residuals vs. leverage allows to identify potential influencor outliers. We can observe that NorthEast and lake central are out of the cook distance (cook's distance higher than 1 (cook's distance plot)) and have influence on the regression.

https://data.library.virginia.edu/diagnostic-plots/ (explication comment interpréter les graphs)

Final estimated model

Conclusions