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 was taken from Jesse W. Proctor et al. [A Regression Analysis of Airline Costs, 21 J. AIR L. & COM. 282 (1954)] and 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 (RTM) divided by the ton-mile load factor (LF), and the adjusted assets are equal to the difference between the total assets (TA) and the investments and special funds (I). Therefore, due to these dependencies, RTM, LF, TA and I were removed for the following analysis. Only six independent variables (FL, SoP, DFT, PS, C, AA) were analyzed to predict the total operating cost (TOC).

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; there are very different magnitude values between each factor. This causes an issue when building a

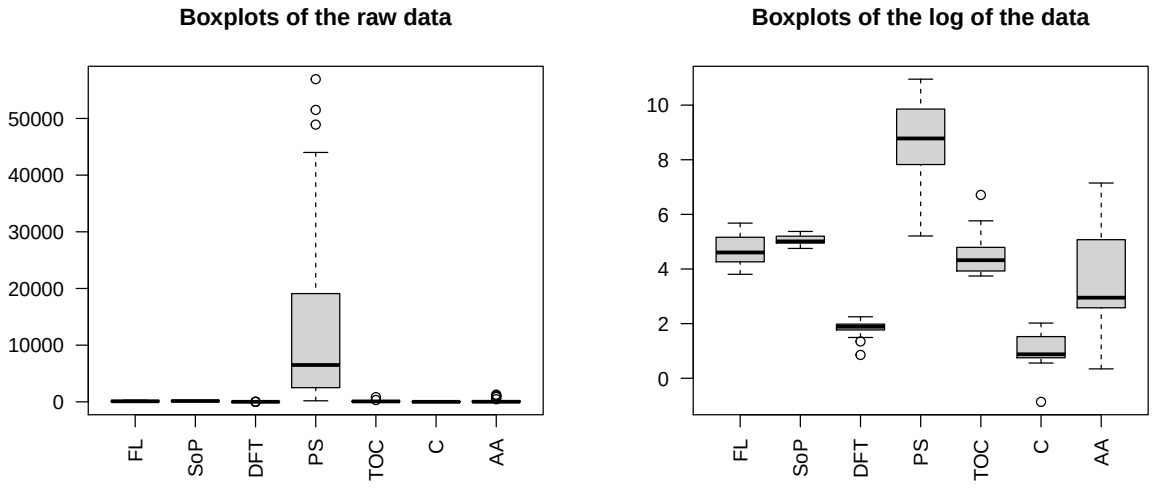

Figure 1: Boxplots of the raw data and the logarithm of the data.

model because some factors, due to the fact that they have bigger values, can hold a non-realistic weight which would 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. From here onward, each time the paper will cite a factor's name, it will in fact be the logarithm of said factor.

Table 2: 5-numbers summary of the data: Min, 1st Quartile, Median, 3rd Quartile and Max of the log variables.

	FL	SoP	DFT	PS	TOC	C	AA
Min	3.81	4.75	0.85	5.21	3.74	-0.86	0.34
1st Q	4.26	4.95	1.77	7.82	3.93	0.75	2.58
Median	4.61	5.01	1.89	8.78	4.32	0.88	2.95
3rd Q	5.16	5.20	1.98	9.86	4.79	1.53	5.07
Max	5.68	5.38	2.25	10.95	6.71	2.02	7.15

Univariate numerical

Table 3: Mean, standard deviation (SD) and median absolute deviation (MAD) of the logarithm 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)

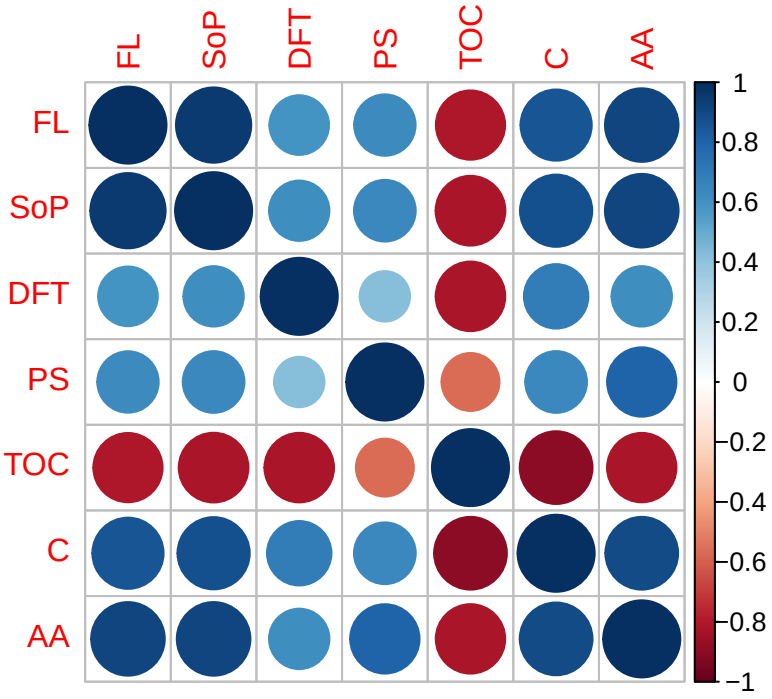

Figure 2: Correlation values between each feature.

This diagram demonstrates the correlation between each feature. The bluecolor illustrates a positive correlation, whereas the red color 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 compared to the rest of the variables.

Bivariate graphical

This graphics demonstrates shows the scatter plots between each variable. We can find some clear linear relationships between certain features such as between FL and SoP or FL and AA. This can be concluded because we can observe see that the points almost form a straight line between these two factors. We can observe that between TOC and other factors, there are not any straight line; it means that we will need more than one factor to best predict TOC linearly.

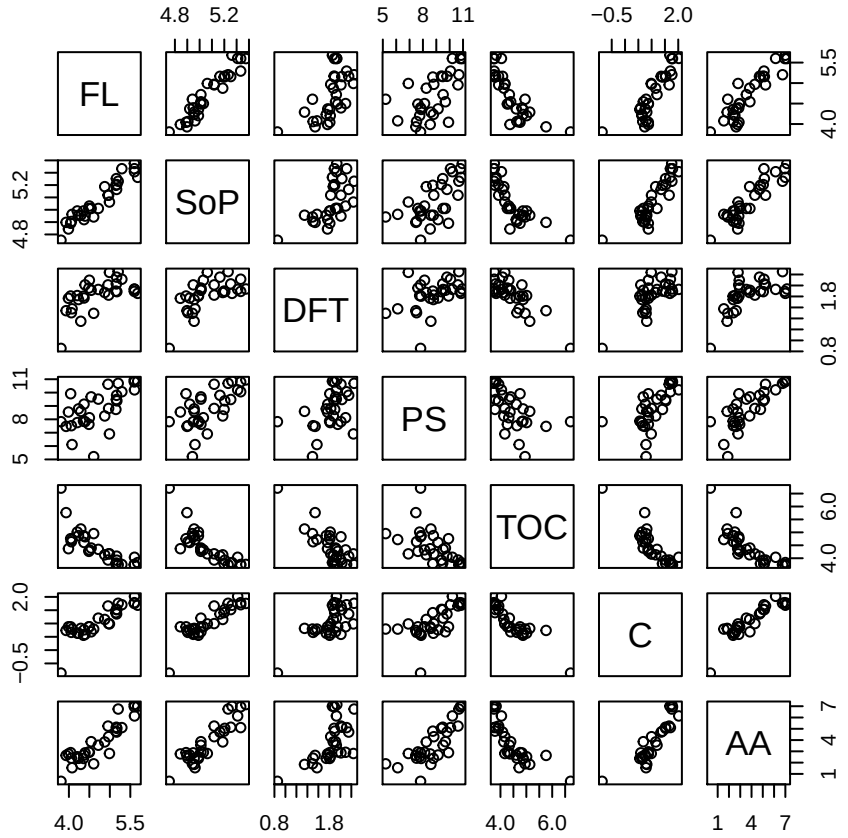

Figure 3: Bivariate scatter plots for each possible interaction

Model fitting

According to the summary of the linear model, 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 4: P-values for the linear model

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

Two factors show significant results: the daily fight time (DFT) and the available capacity (C), which have a p-value equal to 0.0011 and 0.0163, respectively. Additionally, both variables have an high correlation with the total operating cost, as shown in the figure 2.

A step-wise model selection by AIC was also performed on the previous linear model. The best model found through this method depends on the two variables: daily flight time (DFT) and available capacity (C). This correlates to what has been found previously with the linear regression. Here is the model:

$TOC = \hat{\beta}_0 + \hat{\beta}_1 \times DFT + \hat{\beta}_2 \times C$

$TOC = 6.83 + (-0.73 \times DFT) + (-0.88 \times C)$

The R-squared values are $R^2 = 0.87$ and adjusted $R^2 = 0.86$

From the initial hypotheses ($H_0 : \beta_0 = \beta_1 = \beta_2 = 0$ and H_1 : at least one non-zero parameter), the p-value is equal to $5.8824096 \times 10^{-13} < \alpha = 0.05$. The null hypothesis is therefore rejected and we know that at least one coefficient does play a role in predicting the dependent variable.

The p-values of the different features for our final model are smaller, they are close to 0. It confirms the hypothesis that the final model give better results than the one taking all variables into account ; it has smaller p-values, a greater adjusted R-squared, a lower AIC.

Model assessment

From the different plots seen in figure 4, the residuals can be evaluated. However, many assumptions must be controlled. Firstly, the normal QQ-plot illustrates whether our data follows a normal distribution. In this case, despite the central airlines' data, the standardized residuals are mostly aligned on a straight line. This confirms that the errors follow a normal distribution.

The second assumption is that the errors have a mean equal to zero. This can be observed on the residuals vs. fitted plot. The data of our residual values are located around 0 with a small variation. This observation means that no non-linear relationship are present in the fitted model. From the same plot, uncorrelated residues can be assumed because the values appear to be randomly plotted, the red line is horizontal, and no trend is observable.

Concerning the homoscedastic property, the scale-location plot helps us determine it. Indeed, a horizontal line proves the homoscedasticity of the residues and determines whether non-linearity is present. In this model, despite the last point which causes a descending slope, the points are generally randomly spread and form an almost horizontal line. To

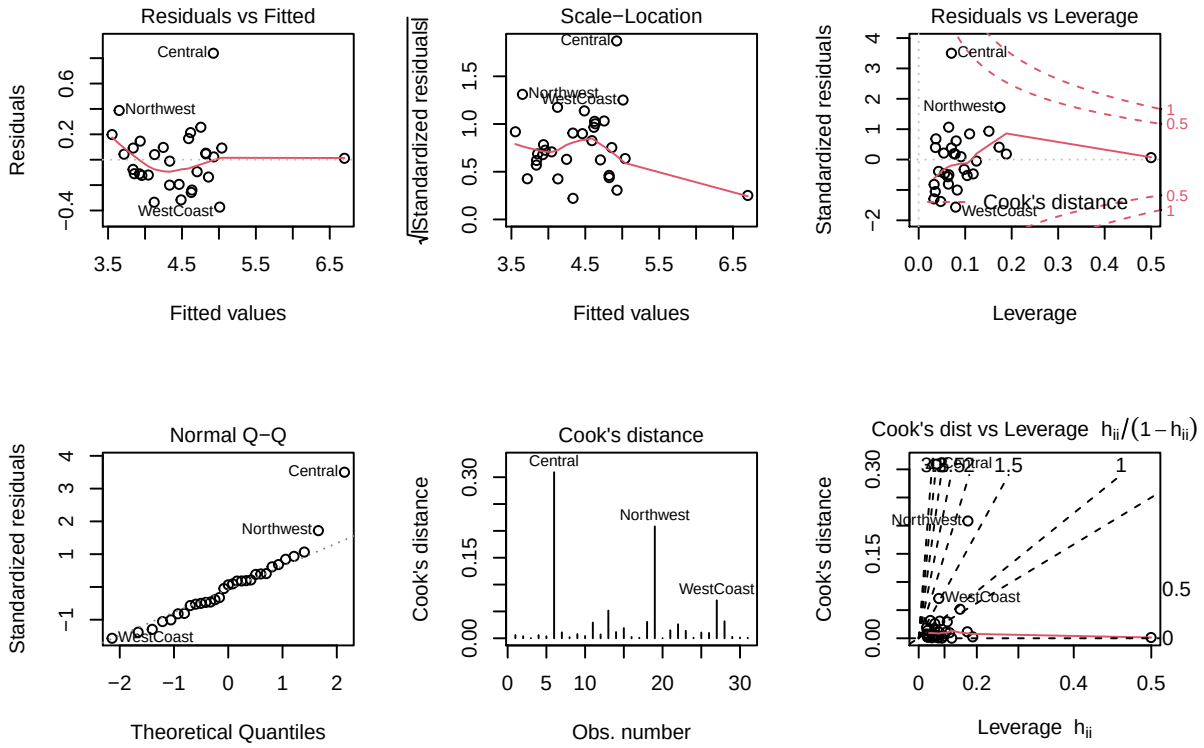

Figure 4: Residual analysis.

confirm the homoscedascity, the Breusch-Pagan test can be used. The p-value being $0.5127 > 0.05$, it appears to be non-significant: therfore, the null hypothesis suggesting homoscedacity cannot be rejected.

Finally, the residuals vs. leverage plot and the Cook's distance plot allow to identify potential outliers. No airline errors have a Cook's distance higher than 0.4. This means that no outliers are observable in this model.

Conclusion

In conclusion, to get the best linear model, we first need to remove the factors that are dependent from each other. Then, given that our data lies on a great magnitude, we take the log of the data to avoid bias in the regression. We find the best model by performing an AIC selection. The model is further confirmed by applying a linear model on the dataset and by verifying whether the variables with a significant p-value correlates with the AIC model. To finish, the usual statistic assumptions are verified.

Our final model to predict TOC only depends on DFT and C; it is summarized here. $TOC = 6.83 + (-0.73 \times DFT) + (-0.88 \times C)$ with a p-value equal to $5.8824096 \times 10^{-13} < \alpha = 0.05$ and an adjusted $R^2 = 0.86$.