Regression alanysis of airline costs

Jeremy Baffou, Valeriia Timonina, Ali Saadat

08/04/2022

Introduction

The purpose of this analysis is to find variables which affect airline costs (Operating Costs per revenue ton-mile). The effect of seven factors is studied: 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). Regression based on natural logarithms of all factors, except load factor, is performed.

Explanatory Data Analysis (EDA)

First, we transform the following variables into log-scale because they have a wide range of variability: *LengthOfFlight, SpeedOfPlane, DailyFlightTime, PopulationServed, AvailableCapacity , TotalAssets, Investments, NetAssets.* We did not change Ton-Mile load factor into log scale because it is between 0 and 1, and taking a log would change it to big negative values. Then we summarize each column as a univariate. The distribution of each univariate is as follow:

Table 1: Summary Statistics for the Air Dataset

Measure	*Min*	*Q1*	*Median*	*Mean*	*Q3*	*Max*
LengthOfFlight	3.81	4.26	4.61	4.71	5.16	5.68
SpeedOfPlane	4.75	4.95	5.01	5.07	5.20	5.38
DailyFlightTime	0.85	1.77	1.89	1.83	1.98	2.25
PopulationServed	5.21	7.82	8.78	8.81	9.86	10.95
TotalOperatingCost	42.30	50.80	75.40	113.51	120.75	820.90
RevenueTons	0.07	0.80	1.19	1.73	2.68	4.30
TonMileLoadFactor	0.17	0.40	0.50	0.48	0.57	0.69
AvailableCapacity	-0.86	0.75	0.88	1.06	1.53	2.02
TotalAssets	0.71	2.58	3.07	3.84	5.12	7.27
Investments	-4.61	-3.71	0.75	-0.32	1.78	5.24
NetAssets	0.34	2.58	2.95	3.79	5.07	7.15

It could be more informative to look at the boxplot of each univariate (Figure 1).

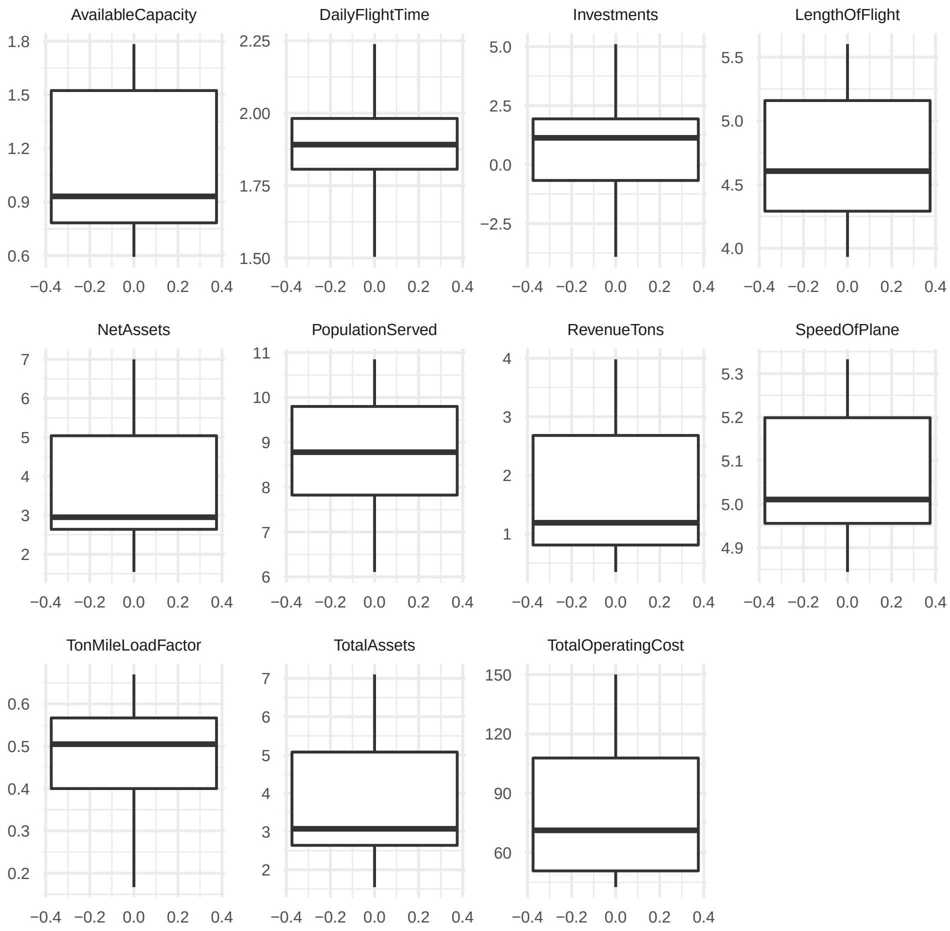

Figure 1: Boxplot of univariates (outliers are removed)

In order to compare the distribution of univariates, we look at their QQ-plot (Figure 2).

Next, we explore the pairs of the variables (i.e. bivariates) which can provide new information. Therefore, we draw the pairwise scatterplot for bivariates, and calculate the pairwise correlation coefficient (Figure 3).

Model Fitting

To perform our regression we will be using a linear model based initially on seven variables: *LengthOfFlight,SpeedOfPlane,DailyFlightTime,PopulationServed,AvailableCapacity,NetAssets* and *TonMileLoadFactor*. We decided to keep only the raw variables and to not consider the possible interactions terms because:

1. Our preliminary exploration showed that interaction terms only make the model more complex without any significant
2. We want to keep the model as simple as possible following the Occam razor principle.

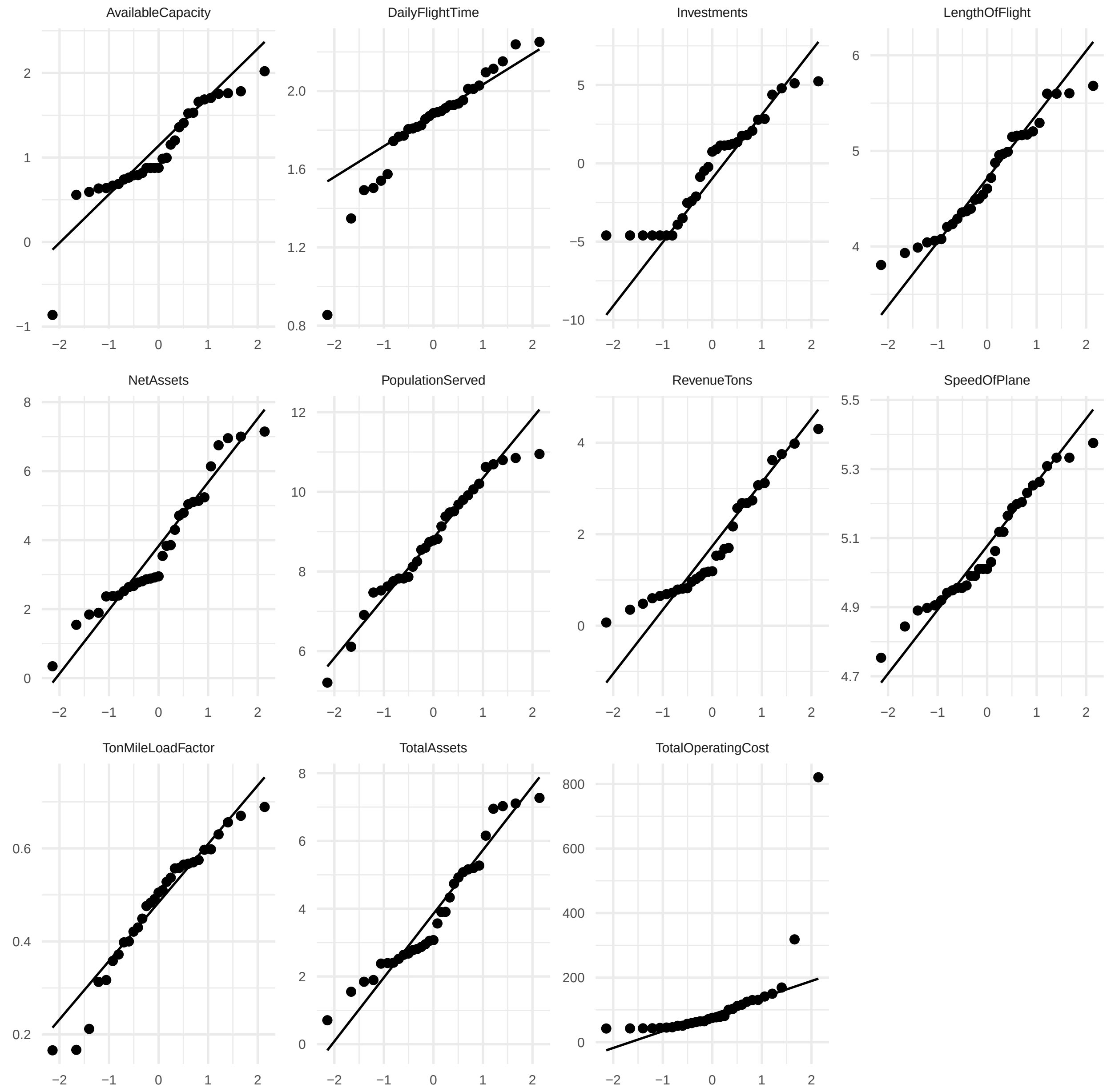

Figure 2: QQ-plot for each univariates

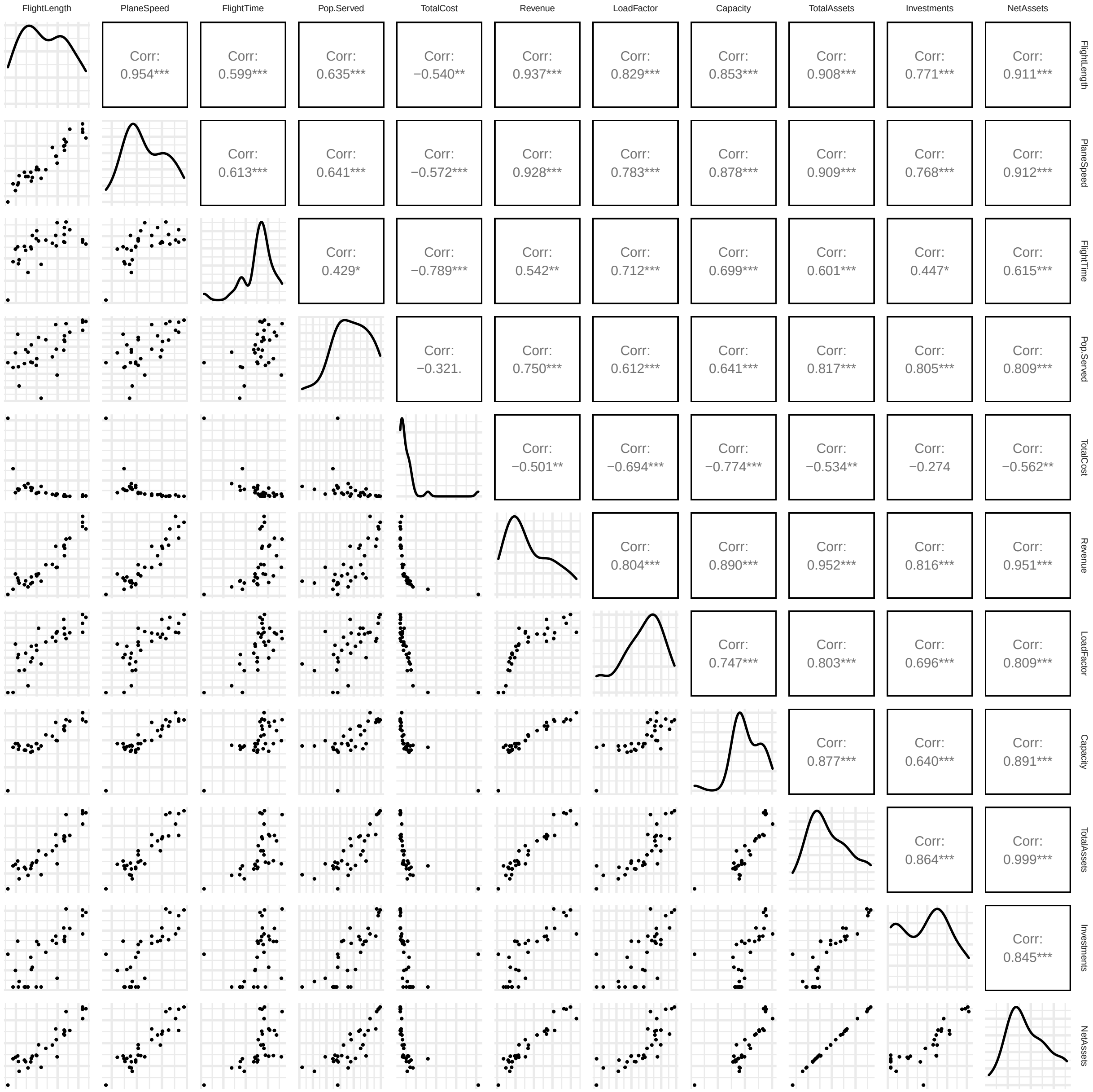

Figure 3: Bivariate plots; upper pannel shows the pairwise correlation, lower pannel shows that pairwise scatterplot, and diagonal pannel illustrates the univariate distribution

Thus, our initial model is as follow:

$$\begin{aligned}\log(\text{TotalOperatingCost}/\text{RevenueTons}) = {} & \beta_0 + \beta_1(\text{LengthOfFlight}) + \\ & \beta_2(\text{SpeedOfPlane}) + \beta_3(\text{DailyFlightTime}) + \\ & \beta_4(\text{PopulationServed}) + \beta_5(\text{AvailableCapacity}) + \\ & \beta_6(\text{NetAssets}) + \beta_7(\text{TonMileLoadFactor}) + \\ & \epsilon \end{aligned} \tag{1}$$

Model Selection

1. **Multicollinearity**

As the first step of model selection. we look for multicollinearity. The use of this method makes sense in the context of our data because several variables are clearly correlated (high correlation between some variables in Figure 3). To check for possible multicollinearity we used the Variance Inflation Factor (VIF) metric. The higher the VIF is the higher the collinearity. A commonly used threshold is 10, i.e. that any variable with a VIF greater than 10 should be removed from the model. We thus first compute the VIF for each variable in our initial model:

Table 2: Initial model VIF metrics

	VIF
LengthOfFlight	15.44
SpeedOfPlane	14.23
DailyFlightTime	2.60
PopulationServed	3.76
AvailableCapacity	6.95
NetAssets	18.01
TonMileLoadFactor	4.59

We remove *NetAssets* since it has the largest VIF above threshold. We calculate VIF for the remaining variables:

Table 3: VIF metrics after removing NetAssets

	VIF
SpeedOfPlane	13.57
DailyFlightTime	2.58
PopulationServed	1.89
AvailableCapacity	5.67
TonMileLoadFactor	4.48
LengthOfFlight	14.30

We remove *LengthOfFlight* because it has a VIF >10. After calculating VIF for the 5 remaining variables, we observe that all of them have small VIFs.

Table 4: VIF metrics after removing LengthOfFlight

	VIF
SpeedOfPlane	5.45
DailyFlightTime	2.46
PopulationServed	1.89
AvailableCapacity	5.60
TonMileLoadFactor	3.52

Thus at this step we have a model with five variables: *SpeedOfPlane, DailyFlightTime, PopulationServed, AvailableCapacity,* and *TonMileLoadFactor*.

2. **Goodness of Fit**

Here we use Akaike Information Criterion (AIC) to check the fittness of our model considering its complexity. We use stepAIC (backward selection) which starts from an initial complex model and try to reduce its complexity by removing at each step the variable that contributes the less to the reduction in AIC. Based on the stepAIC results (Table 5), *SpeedOfPlane* and *PopulationServed* have a negligible contribution to the the reduction of AIC and their statistical significance were low (pvalue $>$ 0.05). We thus stop with a model with three parameters and no interaction term: *DailyFlightTime, AvailableCapacity, TonMileLoadFactor*.

Table 5: Step AIC output

	Df	Sum of Sq	RSS	AIC	F Value	Pr(F)
			0.95	-96.16		
SpeedOfPlane	1	0.1	1.04	-95.09	2.6	0.12
DailyFlightTime	1	0.31	1.26	-89.31	8.25	0.01
PopulationServed	1	0.15	1.10	-93.60	3.96	0.06
AvailableCapacity	1	4.83	5.78	-42.06	127.7	0
TonMileLoadFactor	1	4.42	5.36	-44.38	116.68	0

Model Assessment

Before assessing the model we will recall the different assumptions that we made about the model and the data which are Normal Theory Assumptions (NTA). The NTA suppose that all errors are independently normally distributed with mean 0 and common variance σ. As the variance is fixed, NTA also implies homoscedicity of the residuals. To assess if our assumptions were right and thus if our model is correct we will proceed to a graphical inspection of our residuals.

The normality and homoscedasticity assumptions seem to hold as the QQ-plot (Figure 4)

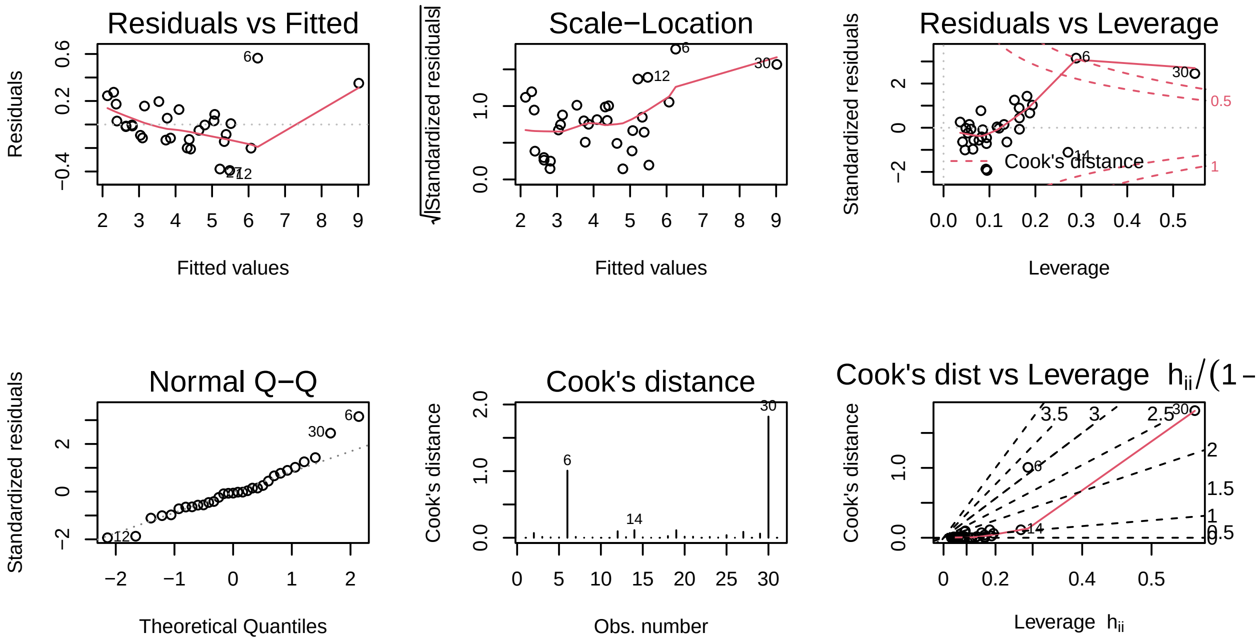

Figure 4: Model assessment

lies approximately on $y = x$ and the spread of the errors does not seem to vary as we move among the fitted values. We can also remark that some points have a high Cooks distance (6,30). Thus that they are data points with high leverage (such a point will have a stronger effect on the model than the rest of the points, inducing the risk of over fitting). When we look at these data points, we can see that they are outliers in the predictive variable. On such a small dataset and without field expertise we cannot remove these outliers but they may affect the model fitting if they are artifacts.

Final Model

$$\begin{aligned}\log(\text{TotalOperating}\widehat{\text{Cost}}/\text{RevenueTons}) = 9.14 - 0.7(\text{DailyFlightTime}) - \\ 1.43(\text{AvailableCapacity}) - 4.58(\text{TonMileLoadFactor})\end{aligned} \tag{2}$$

Conclusion

Linear regression was performed to investigate which factors affect airline costs. We can see that all coefficients are negative, thus an increase in any of them would reduce the ratio. Which is a wanted thing for the airline as they wish to have the smallest operating cost per unit of revenue ton-mile. The exact amplitude of the influence is not intuitive due to the log scale on the parameters.