Cost Data for U.S. Airlines (1970-1994)

Beatriz Santos (20230521), Daniela Camilato (20221641), Emília Gonçalves (20230446), Ricardo Kayseller (20230450).

Master's in Data Science and Advanced Analytics
Statistics For Data Science

Introduction

This poster presents a focused analysis aimed at identifying the primary factors influencing operating costs within the U.S. airline industry, spanning data from 1970 to 1984. Sourced from NYU Stern School of Business and originally compiled by Christensen Associates, the dataset encompasses 90 observations from six major U.S. airlines. Our research is centered on unraveling the answer to a critical question: "What are the primary contributors to the costs of U.S. airlines?" A key finding of our study is the identification of three statistically significant factors that impact total costs: output in revenue passenger miles (Q), fuel price (PF), and Load Factor (LF), which is the average capacity utilization of the fleet. These elements are pivotal in understanding and managing costs in the highly competitive and constantly evolving airline sector. The insights gained are not only valuable for airline companies in strategizing and optimizing operations but also provide crucial information for policymakers and analysts interested in the economic and operational dynamics driving the industry.

Research Question

What are the factors that impact the most the total cost relative to the U.S. Airlines?

Methodology

In order to conduct our research, we used a dataset with 90 observations on 6 airline companies, which was collected between 1970 and 1984, i.e., for 15 years. The team run an analysis on a panel dataset using the plm (panel linear model) tool in R. Panel data refers to the collection of data that includes observations on several subjects or individuals over multiple time periods. The analysis encompasses the estimate of various models, diagnostic tests, and comparison tests. To summarize, we made a brief exploration of the data, estimated different (Pooled OLS, Robust Pooled OLS, Random Effects, Fixed Effects and Robust Fixed Effects) and run some test, for instance, Hausman Test, Breusch-Pagan Test, F-Test, among others.

Results

Given the following p-values (all above 0.05) in all models, we state that all variables are statistically significant and have importance in these. Therefore, the Total Cost depends on Q (output in revenue passenger miles), PF (fuel price) and LF (average capacity utilization of the fleet).

Interpreting the R-Squared and having in mind that for Robust models it was not obtained, it is possible to say that the model that had the highest R-Squared was the Pooled OLS, explaining 94.6% of y's (Total Cost) variability. However, due to the Pooled OLS model's inability to account for unobserved heterogeneity and its assumption of homogeneity across all airlines, we don't consider this model as the most appropriate one.

Model	R-Squared	Q p-value	PF p-value	LF p-value
Pooled OLS	0.94609	< 2.2e-16 ***	< 2.2e-16 ***	3.058e-05 ***
Robust Pooled OLS	-	< 2.2e-16 ***	0.0007077 ***	0.0033838 **
Random Effects	0.91129	< 2.2e-16 ***	< 2.2e-16 ***	2.126e-05 ***
Fixed Effects	0.92937	< 2.2e-16 ***	9.698e-12 ***	2.375e-08 ***
Robust Fixed Effects	-	< 2.2e-16 ***	0.006701 **	0.030521 *

Through the Breusch-Pagan test, the existence of heteroskedasticity was verified, i.e., the errors variance is not constant.

According to Wooldridge test, there is serial correlation or autocorrelation in the data, meaning that the errors on regression are dependent on each other.

Lastly, we conducted the Pesaran CD test to test whether there was cross-sectional dependence. The p-value was above 0.05, so we concluded that there was cross-sectional dependence among individual units in the panel, i.e. individual airline effects are not influencing each other significantly.

Test	H0	p-value	Result
Hausman	RE is the preferred model	3.832e-13	FE is the best fit
Robust Hausman	RE is the preferred model	3.832e-13	FE is the best fit
F-test	Pooled OLS	3.467e-10	FE is the best fit
Breusch-Pagan	Homoskedasticity	6.094e-05	Heteroskedasticity
Wooldridge	There is no serial correlation	1.795e-13	There is no serial correlation
Pesaran CD	There is no cross-sectional dependence among the individual units in the panel	0.1052	There is no cross-sectional dependence among the individual units in the panel

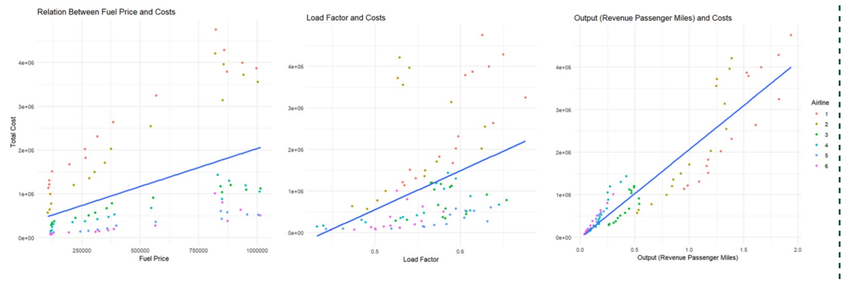

Figure 1: Comparative Analysis of Airline Operating Costs in Relation to Fuel Prices, Load Factors, and Output

The plot above illustrates the impact of fuel prices, load factors, and passenger output on the operating costs of six airlines, with each factor showing a potential to increase costs when elevated and highlighting a positive correlation.

Next Steps

Moving forward, it's important to explore if factors like the number of flights, fuel prices, and how full the flights are might be hiding other important influences on airline costs. Future research could look into things like how efficiently airlines run, changes in the airline market, new rules and technologies, and how big economic events affect airlines. This would help get a clearer picture of what really drives costs in the airline industry, useful for both airlines and those making policies or studying the industry.

Conclusion

The conclusions were derived from the statistical significance of coefficients, the adequacy of model assumptions, and the outcomes of diagnostic tests. We meticulously analysed and made sense of these findings within the framework of the dataset and research inquiries.

The Fixed Effects model was the best fitted one, according to Hausman and Robust Hausman tests, however, its interpretation needs to be cautious due to the heteroskedasticity detected. Our decision to favour the Fixed Effects model over the Pooled OLS (highest R-Squared) was based on its ability to account for unobserved airline-specific heterogeneity, which is crucial in the context of airline cost analysis. Furthermore, the analysis took into consideration both domain knowledge and theoretical factors while drawing findings, and finally we concluded that all variables are statistically significant.

All of these inferences are vital for airlines in strategizing cost management and operational efficiency, offering a data-driven pathway to enhance profitability and sustainability in a highly competitive industry.

References

De Haan, M., Stock, W., & Chapter. (n.d.). ECON4150 -Introductory Econometrics Seminar 6. Retrieved January 13, 2024, from https://www.uio.no/studier/emner/sv/oekonomi/ECON4150/v14/

Explain Serial Correlation and How It Affects Statistical Inference. (2022, December 21). CFA, FRM, and Actuarial Exams Study Notes. https://analystprep.com/study-notes/cfa-level-2/quantitative-method/explain-serial-correlation-and-how-it-affects-statistical-inference/

Panel Dataset / Cost Data of U.S. Airlines. (n.d.). Www.kaggle.com. https://www.kaggle.com/datasets/sandhyakrishnan02/paneldata

Qin, Y. (n.d.). Research Guides: Panel Data Using R: Fixed-effects and Random-effects. Libguides.princeton.edu. https://libguides.princeton.edu/R-Panel