Analyzing Key Determinants of Airfare Pricing Between 1997 and 2000

Datasci 203 Team 5: Ameya Chander, Daniel Wang, Bonnie Yang

Contents

1 Abstract

This study investigates the relationship between airfare and key factors such as flight distance, market share, and passenger volume using a large-sample linear regression model. The dataset includes ~4,600 observations collected from 1997 to 2000, providing a historical perspective on airfare determinants. Analysis reveals that flight distance is the most significant predictor of airfare, with a strong positive association.

2 Introduction

In today's digital, post-COVID world, the complexities and costs of everyday life have risen sharply, and airfare is no exception. Airline pricing has become increasingly opaque, with varying prices across websites, fluctuating rates depending on the day, and fares that often seem arbitrary. This leads us to our research question:

> *How is airfare between 1997 and 2000 affected by flight distance, market presence, and passenger volume?*

We selected this question specifically to examine the foundational factors driving airline pricing during a period that predates major industry shifts, such as the post-9/11 restructuring and the rise of airline monopolies. Recent studies have highlighted the role of models like "continuous revenue management," which assigned variable prices to seats on the same flight[1]; however, focusing on this historical period allows us to explore the impacts of demand, competition, and distance without the added complexities of modern tools like online price comparison websites or loyalty programs. Ultimately, understanding these historical dynamics lays the groundwork for analyzing how these factors continue to shape airfare pricing today.

3 Description of Data Source

We will use Wooldridge's airfare dataset provided by the Department of Economics of Michigan State University. The data points are drawn from the Domestic Airline Consumer Report by the U.S. Department of Transportation. There are a total of 4596 observations from 1997 to 2000. We are interested in 3 features – airline market concentration, flight distance, and average passenger counts – and how they influence airfare within this time period.

4 Data Wrangling

Upon reviewing the dataset, it appears to be clean and well-organized. The data is structured by years, with four data points recorded for each route ID per year. Importantly, there are no missing (NULL) values, as shown in Table 1. The dataset contains 14 columns, including six raw data fields and eight derived columns, such as boolean indicators for different years and log transformations of airfare, distance, and passenger volume.

However, limited documentation of the variables required further exploration to clarify certain relationships. For instance, we investigated whether multiple unique distances were associated with a single route ID. By removing duplicates based on route and distance pairs, we confirmed a one-to-one relationship between these variables. In addition, the purpose of the "biggest carrier market share" feature ("bmktshr") was initially unclear. Through exploration, we observed that this value varies by route and year, as shown in Table 2, indicating that the dataset captures the market share of the *dominant airline* on a given route for a specific year. All in all, the dataset is well-prepared for analysis, with minimal need for additional cleaning or transformations, making it a convenient resource for our study.

5 Operationalization

The three variables identified for this analysis – flight distance, market presence, and passenger volume – are the primary metrics provided by the dataset. All observations are included, as the dataset is already cleaned and all data points are relevant to the study. A brief definition of each concept is provided below:

- **Airfare**: The average one-way price per passenger on a specific route.
- **Flight distance**: The distance in miles between two locations.

[1] Walsh, D. (2023, October 20). Research shows how airline pricing really works. Haas News | Berkeley Haas.

Table 1: Checking for NULL Values in the Dataset

year	id	dist	passen	fare	bmktshr	ldist	y98	y99	y00	lfare	ldistsq	concen	lpassen
0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 2: Data for One Unique Flight ID

year	id	dist	passen	fare	bmktshr
1997	1	528	152	106	0.8386
1998	1	528	265	106	0.8133
1999	1	528	336	113	0.8262
2000	1	528	298	123	0.8612

- **Market presence**: The largest market share held by a specific carrier on a given route. This value changes yearly based on the dominant airline, but the specific airline details are not provided and are not relevant to this analysis.
- **Passengers**: The average number of passengers per day for the route.

There are several potential approaches to answering the research question. One option we considered was comparing coefficients across the four years to identify trends and variations, which could reveal insights into evolving pricing strategies. For example, a consistent increase in the coefficient for market presence might suggest growing price-setting power among dominant airlines. However, the limited data points (each route ID only has four observations) made it challenging to conduct year-to-year analyses. Figure 2 in the Appendix shows only minimal year-to-year changes in average airfare prices, with a gradual upward trend. Ultimately, we decided to apply a model that averages the values across all four years to assess the overall impact during the time period studied.

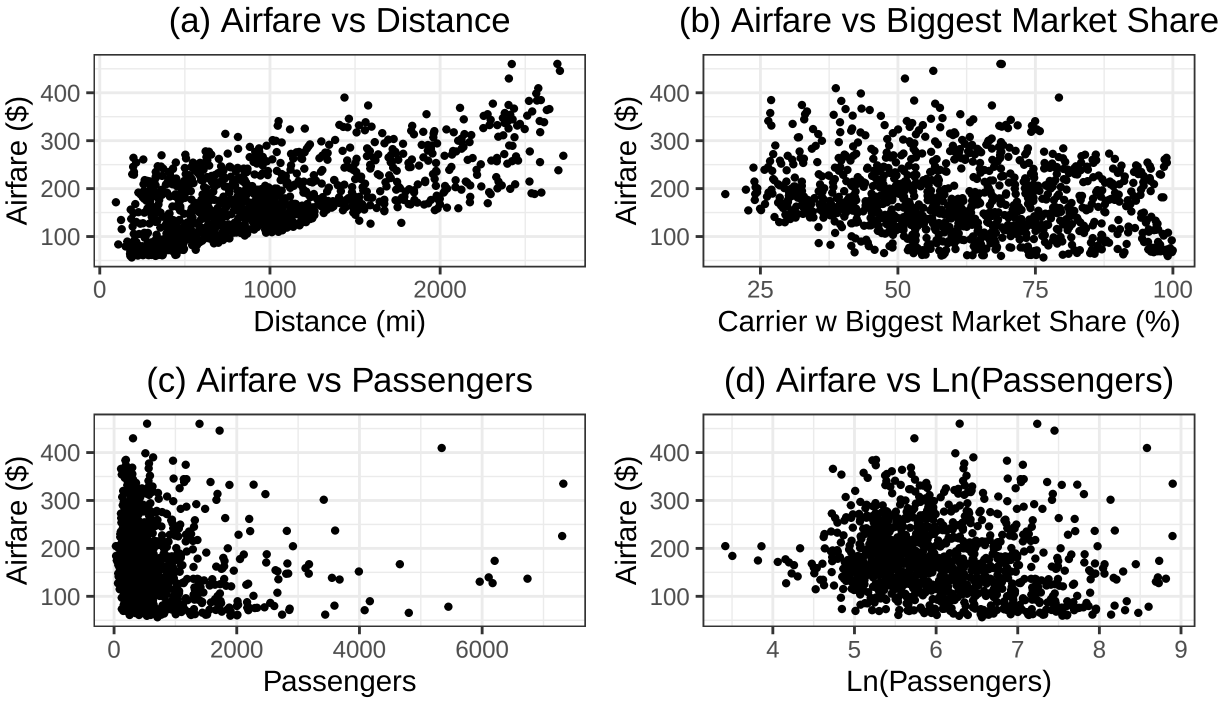

Figure 1: Comparing airfare with (a) distance, (b) market share, (c) passengers, and (d) ln(passengers).

6 Model Specification

To begin, we explored the relationship between airfare and distance through a simple linear regression, as these two variables are most commonly linked. This analysis revealed a positive linear relationship, where longer distances correlate with higher airfares (Fig 1a). The coefficient for distance was highly significant, with a p-value of 2.2e-16, well below the 0.05 threshold. The coefficient estimate of 0.076 indicates that for every additional mile flown, the price increases by approximately 7.6 cents. To put it into relative terms, for every 100 miles flown (roughly the distance from LA to San Diego), the price increases by $7.63. These findings provide strong evidence of a significant relationship between distance and airfare.

Next, we examined whether the other key factors – passenger volume and market share – also influence airfare. Visual exploration suggested that market share has a roughly linear relationship with airfare (Fig 1b), requiring no transformation. However, passenger volume showed a strong skew (Fig 1c), with most of the data concentrated on the left-hand side. This provided justification for a log transformation, which improved the distribution but still indicated a weak correlation with airfare (Fig 1d).

We extended our analysis with two multiple linear regression models: one including distance, passenger volume, and market share as predictors, and another substituting passenger volume with its natural logarithm (ln(passengers)). The results of these models are summarized in Table 3 below:

Table 3: Regression Table for Airfare Analysis

	Dependent variable:		
	Airfare ($)		
	(1)	(2)	(3)
Distance (mi)	0.076***	0.089***	0.088***
	(0.003)	(0.004)	(0.004)
Passengers		−0.004	
		(0.002)	
Ln Passengers			−7.233***
			(1.819)
Market Share (%)		0.772***	0.763***
		(0.108)	(0.106)
Constant	103.261***	46.183***	88.463***
	(3.301)	(9.311)	(15.435)
Observations	1,149	1,149	1,149
R^2	0.412	0.444	0.450
Adjusted R^2	0.411	0.443	0.448
Residual Std. Error	55.844 (df = 1147)	54.328 (df = 1145)	54.072 (df = 1145)
F Statistic	803.145*** (df = 1; 1147)	305.152*** (df = 3; 1145)	311.672*** (df = 3; 1145)

Note: *p<0.05; **p<0.01; ***p<0.001

Column 1 (Model Short): Includes just distance as a predictor.

Column 2 (Model A): Includes all predictors without transformations.

Column 3 (Model B): Replaces average passengers with log-transformed passengers.

7 Model Assumptions

Due to the sample size, the large-sample linear model assumptions apply:

1. Independent and Identically Distributed (IID) Data:

- While the data may be identically distributed if collected properly by the authors, independence is likely violated as airfare is influenced by market conditions, which depend on other flights. Additional information on geographic clustering for airlines, routes, or connecting flights could provide better insights into potential dependencies. Nevertheless, despite the likely violation of IID, we consider this dataset suitable for analysis, assuming the authors accounted for the spatial clustering effects. Temporal clustering is likely not a concern, as we averaged data over the four-year timeframe and observed minimal variation.

2. Unique BLP Exists:

- A BLP **exists** when $\text{cov}[X_i, X_j]$ and $\text{cov}[X_i, Y]$ are finite (no heavy tails): Although most of the variables do not exhibit heavy tails, the "passengers" variable shows a noticeable right skew (Fig 1c), raising questions about this condition. However, with a large sample size, variances are likely still finite. To address the skewness, we applied a log transformation to the "passengers" variable, which effectively mitigated this condition.
- A BLP is **unique** when there is no perfect collinearity ($E[X^T X]$ is invertible): This condition is likely fulfilled as our model did not drop any variables, indicating that no X_i can be written as a linear combination of the other X's (and therefore the X's have unique variation).

8 Model Results and Interpretation

Adding more variables – biggest market share and average passengers per day – yields a different perspective. The coefficient for biggest market share is statistically significant (p-value = 1.23e-12) with an estimate of 0.772, meaning that for every additional percent of market share, airfare increases by \$0.77. Given that market share ranges from 16% to 100% (0.16 to 1.0), the maximum possible impact of this variable is \$64 (indicating a full monopoly on a particular route), assuming all other factors remain constant. However, such extreme scenarios are rare, and in most cases, the effect of market share on airfare is considerably smaller compared to the influence of distance.

In contrast, average passengers per day was not significant in the untransformed model (p-value = 0.07), which was expected as seen in Figure 1c. After applying the log transformation, the coefficient for log passengers becomes significant (p-value = 7.473e-05), with an estimate of -7.23. This indicates that a small $\alpha\%$ increase in passengers leads to a decrease in airfare of approximately 7.23α. For example, a 10% increase in passengers leads to a \$0.72 decrease in price. While statistically significant, practically speaking, this effect is nearly negligible compared to the influence of distance.

F-tests confirm that adding biggest market share, log passengers, or both, all significantly improve the model (p-values of 4.202e-14, 2.146e-06, and 2.2e-16, respectively). This indicates that these variables enhance the model's explanatory power beyond the simple regression of airfare on distance. Nevertheless, while adding these coefficients is statistically significant, their practical significance is minimal compared to the dominant impact of distance on airfare.

Overall, our analysis revealed that **distance** remains the strongest predictor of average fare, likely due to the costs of fuel and other expenses associated with longer routes. It would be interesting to investigate how market share influences fare in the modern era. Future analyses could explore time-based trends, as our dataset's limitations (such as averaging data across four years) may obscure temporal effects. Additional variables – such as specific route distances, locations/popularity of routes, airport characteristics, and booking factors – would also provide valuable insights into airfare determinants.

9 Appendix

Link to data source: http://fmwww.bc.edu/ec-p/data/wooldridge2k/AIRFARE.DES

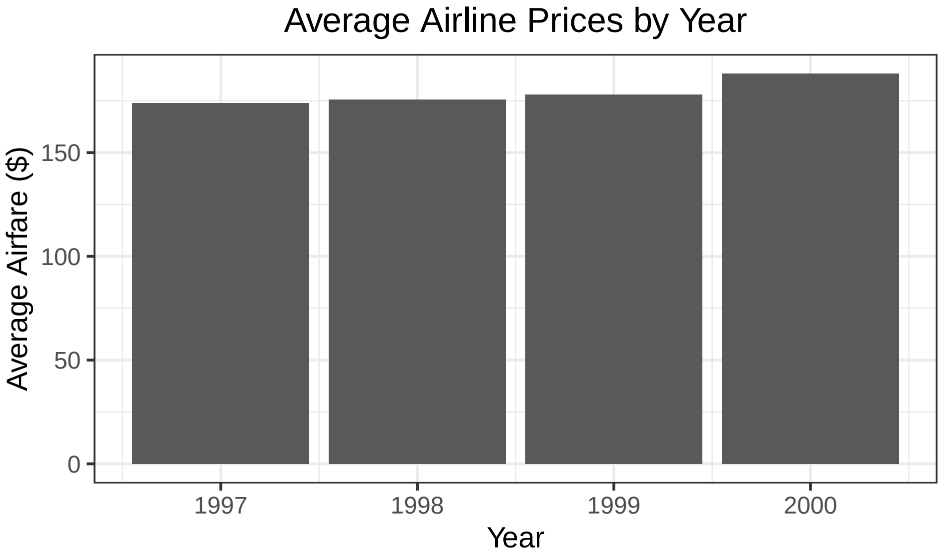

Figure 2: Average airfare prices over the years.

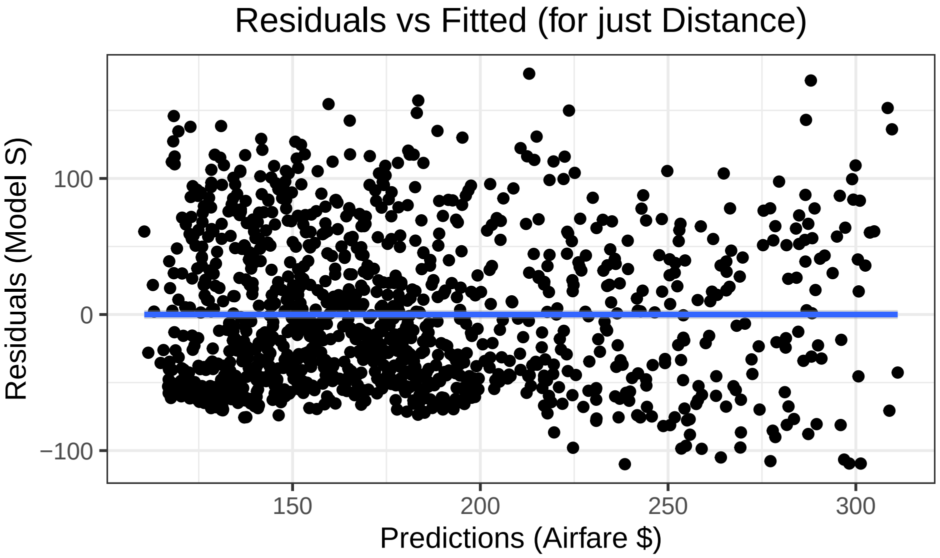

Figure 3: Comparing residuals vs predictions for single-variable (distance) model.

We want to see if there's more to the story than just the distance shaping the airfare, which leads to the graphs below:

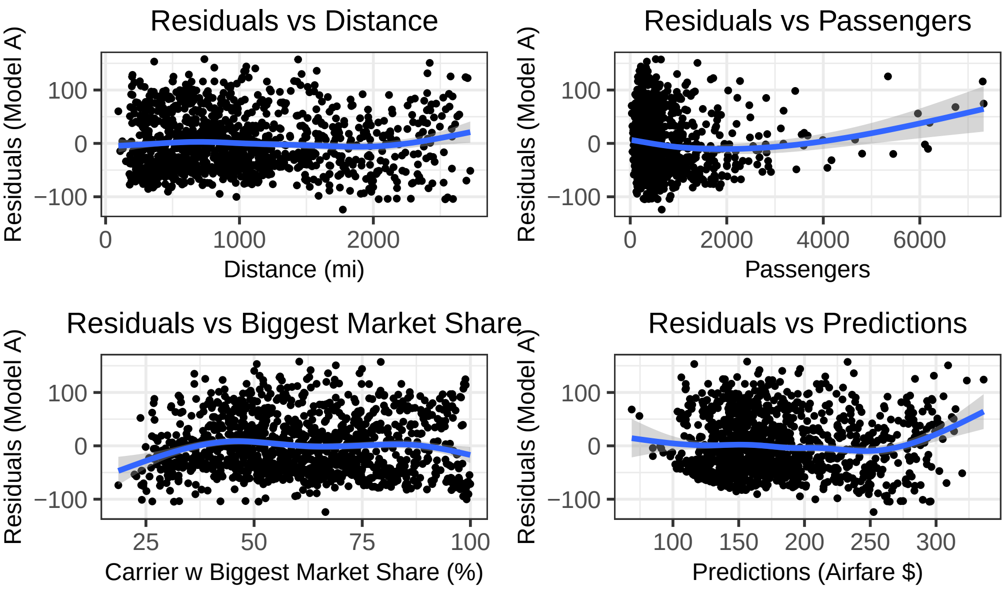

Figure 4: Comparing residuals vs fitted and independent variable plots without transformations.

And lastly, we want to analyze the final, transformed model (with log of passengers).

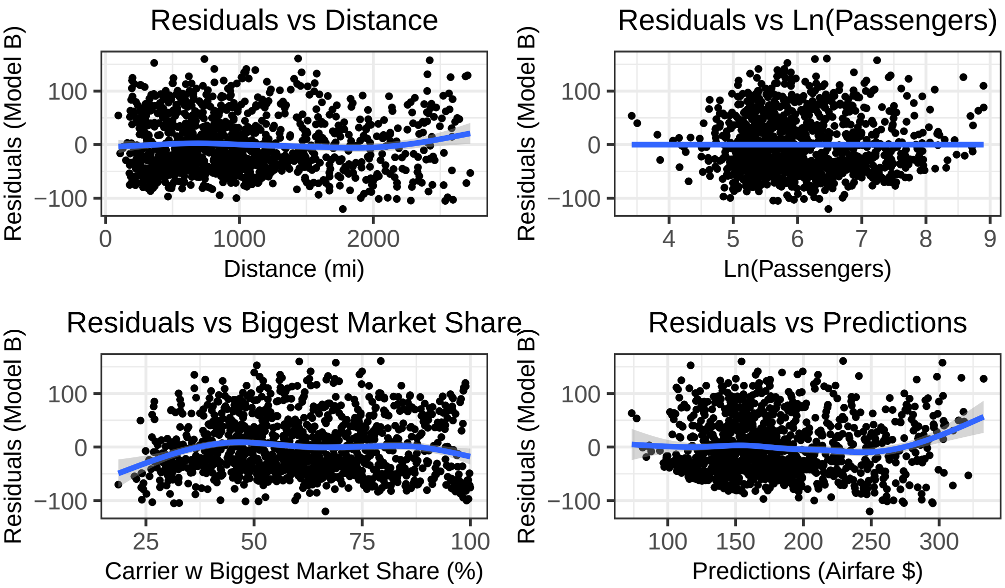

Figure 5: Comparing residuals vs fitted and independent variable plots with log transform on passengers.