Neemias Moreira

Professor Rob Poulton

Introduction of Economic Data Analysis

04/15/2025

Regression Analysis Project: Airline Fare

The goal of this project is to determine the most suitable airport for travelers flying from Lamoni, Iowa. When booking a flight, individuals often face the challenge of choosing between Des Moines International Airport and Kansas City International Airport, as both offer various advantages. However, comparing ticket prices between these two airports can be time-consuming due to fluctuating airfares influenced by factors such as seasonality, demand, and airline competition. To address this issue, this study conducts a regression analysis to evaluate domestic flight prices from both airports to major destinations across the United States. By identifying key factors that affect airfare, the analysis aims to streamline the decision-making process, making flight planning more efficient. This data-driven approach will help determine which airport consistently provides the most cost-effective travel options, ultimately saving travelers both time and money.

To conduct this analysis, data was collected from Kaggle, specifically from the dataset provided by Amitzala, which contains historical flight fare information spanning from 1993 to 2024. However, for this study, only flights from the year 2023 were considered to ensure relevance to current pricing trends.

The data preprocessing involved several key steps:

1. Filtering the data: I selected only flights departing from Des Moines International Airport and Kansas City International Airport to national destinations within the United States.
2. Selecting relevant variables: I removed columns that were not necessary for my regression analysis, keeping only the variables that could influence flight prices.
3. Creating dummy variables: Since seasonality can impact prices, I created dummy variables for the quarter of departure, with the fourth quarter serving as the baseline category.
4. Checking for multicollinearity: I conducted a correlation matrix analysis to ensure that the independent variables were not highly correlated with each other, which could otherwise distort the regression results.
5. Building the regression model: After cleaning and preparing the data, I ran a multiple linear regression analysis with flight price as the dependent variable and five independent variables.

Before interpreting the results of the regression model, it is essential to establish hypotheses to guide the statistical analysis. By defining the null and alternative hypotheses, this study aims to determine whether there is a statistically significant difference in flight prices between the two airports.

Null Hypothesis (H_0): There is no significant difference in average flight prices between Des Moines International Airport (DSM) and Kansas City International Airport (MCI). Any observed differences in prices are attributed to random variation.

Alternative Hypothesis (H_1): There is a significant difference in average flight prices between Des Moines International Airport (DSM) and Kansas City International Airport (MCI), suggesting that one airport consistently offers lower fares than the other.

Only using location as my variable is my resultant graph:

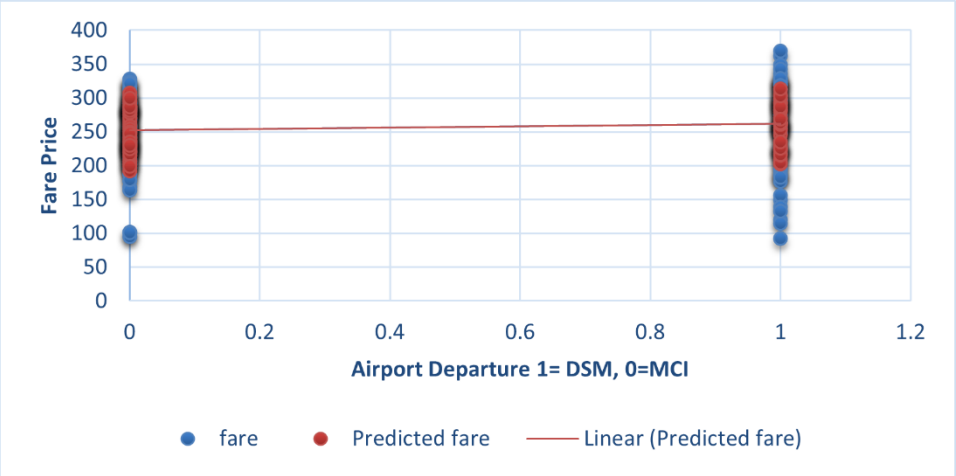

Figure 1: Fare Price Where 0 is Kansas City International Airport and 1 is Des Moines International Airport.

While the hypothesis test focuses on comparing average flight prices between the two airports, the regression model incorporates additional variables that may also influence ticket prices. Factors such as seasonality, distance, passenger volume, and airline market share contribute to fluctuations in airfare. By including these variables in the model, this analysis aims to assess their impact and determine whether airport location is the primary driver of price differences or if other factors play a more significant role.

The regression model produced the following general equation:

$$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \beta_5X_5 + \beta_6X_6 + \beta_7X_7$$

Substituting the coefficients and variable names, the equation is:

Flight price = 335.43 - 11.69(Quarter 1) + 2.05(Quarter 2) - 13.39(Quarter 3) - 11.21(Airport Des Moines) + 0.0259(distance) - 0.0824(passengers) - 125.77(large market share)

The regression analysis provides valuable insights into the factors influencing domestic flight prices from Des Moines International Airport and Kansas City International Airport. The intercept of the model is 335.43, which represents the base flight price when all independent variables are zero. This means that if all other factors in the model were absent or neutral, the expected flight price would be approximately 335.43 dollars.

The model incorporates dummy variables for different quarters of the year to account for seasonal variations in flight prices. The results indicate that flights in the first quarter (January to March) and the third quarter (July to September) have negative coefficients of -11.69 and -13.39, respectively. This suggests that flights during these periods tend to be slightly cheaper compared to the fourth quarter (October to December), which serves as the baseline category. However, the p-values for these coefficients are not statistically significant, indicating that there is insufficient evidence to conclude that these seasonal trends consistently affect flight prices.

The airport variable for Des Moines has a coefficient of -11.21, suggesting that flights departing from Des Moines International Airport are, on average, $11.21 cheaper than those from Kansas City International Airport. However, this result is not statistically significant, indicating that the observed price difference may be due to random variation rather than a consistent pricing pattern.

The distance variable, which represents the number of miles traveled, has a small positive coefficient of 0.0259. This implies that for every additional mile traveled, the flight price increases by approximately 2.6 cents. While this finding aligns with the expectation that longer flights tend to be more expensive, the small magnitude of the coefficient suggests that distance alone has a minimal impact on flight prices.

Among all variables in the model, the number of passengers and airline market share were the only statistically significant factors influencing flight prices. The passenger variable, with a coefficient of -0.0824, shows a negative relationship with flight prices, indicating that as the number of passengers increases, ticket prices tend to decrease. This trend may be attributed to higher demand leading to more competitive pricing or economies of scale that allow airlines to offer lower fares.

The market share variable, with a coefficient of -125.77, has the most substantial impact on ticket prices. The negative coefficient suggests that airlines with a larger market share tend to offer significantly lower fares, averaging $125.77 less per ticket. This result is statistically significant, highlighting the crucial role of airline dominance in determining flight costs. Larger airlines may be able to reduce fares due to operational efficiencies, increased competition, or established customer loyalty programs.

While some factors like seasonality and departure airport show slight trends in flight pricing, the strongest predictors of lower flight prices are a higher number of passengers and airlines with a large market share.

The model's R-squared value is 0.259, indicating that approximately 25.9% of the variation in flight prices is explained by the included variables. While this suggests moderate fit, additional factors such as airline promotions, fuel costs, and broader economic conditions could further enhance predictive accuracy. Furthermore, the model's Significance F value of 7.65716E-07 suggests that the overall regression model is statistically significant. This low p-value indicates that at least one of the independent variables meaningfully contributes to explaining variations in flight prices, reinforcing the model’s validity despite its moderate explanatory power.

The regression analysis provided valuable insights into flight pricing patterns from Des Moines International Airport and Kansas City International Airport. The results suggest that while flights from Des Moines tend to be slightly cheaper, the difference is not statistically significant. Additionally, market share plays a crucial role in determining flight prices, with larger airlines offering significantly lower fares.

For future work, I would love to incorporate more detailed information about the timing of flight purchases, such as the number of days until departure and the specific days of the week when tickets are purchased. These factors could significantly influence ticket prices, as airlines often adjust fares based on booking windows and peak travel days. Additionally, integrating flight fare data from 2024 would be valuable, given the recent surge in prices across various sectors in the United States. Including these elements in the analysis could provide a more comprehensive understanding of flight pricing patterns and improve the ability to predict the most cost-effective times and locations for booking flights.

Source

Amitzala. "US Airline Flight Routes and Fares (1993-2024)." Kaggle, https://www.kaggle.com/datasets/amitzala/us-airline-flight-routes-and-fares. Accessed 25 of March 2025.