Ryan Chan & Sangyoon Park

Prof. Ming-Wen An

Mathematics 242

28 October 2015

Distance, Willingness to Pay, and Itinerary Region Type May Be Predictors of Airfare: Multiple Linear Regression Modeling of the U.S. Domestic Flights

Abstract

In this study, we examined potential predictors of airfare. Multiple linear regression modeling was performed on a sample of 400 U.S. domestic flights. The result indicates that market distance, willingness to pay, and itinerary region type are significant predictors of airfare in the U.S. The result also shows that, contrary to common expectation, layovers and the time of the year are not significant predictors of airfare. Implications and limitations of the study are discussed.

If you ever flew across the sky to get somewhere far, chances are you must have wondered at some point, "What is going on with the airfare?" One day, it is high; next day, it is low; and on the day you decide to finally purchase it, it is high again! This is the result of the practice commonly known as "variable pricing." Variable pricing is the strategy airlines and hotels employ to maximize their revenue against fixed sunk costs (i.e. flying costs, room maintenance costs). But does this mean that the airfare is completely random? Our common sense does not quite yield to such a claim. At least the flying distance should relate to the fare. And direct flights should be more expensive than those with layovers. But are/do they? We address this issue in the current analysis and probe: 1) whether airfare has certain predictors; and 2) if so, what these predictors might be. For this aim, we

examined a sample of 400 U.S. domestic flights, drawn from the Bureau of Transportation Statistics. The dataset contained 12 variables including airport, state, distance, fare, and layover(s).

Methods

We employed box plot and histogram to examine the distribution of airfare in the dataset. We also reordered the dataset by airfare and compared different parts of it to find any potential patterns. We then examined scatter plots to identify potential predictor variables for airfare. For non-continuous variables, such as the quarter of the year in which each flight took place, we compared the mean airfare across different groups (e.g., 1st vs. 2nd vs. 3rd vs. 4th quarter) and used the degree of group differences to evaluate each variable as a potential predictor of airfare. We also performed simple linear regression of airfare on each potential predictor variable and examined the result through the F-test.

Some non-continuous variables indicated good candidacy for the predictor variable but required better categorization. Hence, when necessary, we derived new variables from the existing ones and evaluated/used the former in the latter's instead. For instance, the given binary classification of the flight's geographical type (contiguous vs. non-contiguous) showed a high level of group difference in the mean airfare, suggesting finer categorization may also result in group differences. However, the relevant existing variables such as the flight's origin/destination state generate too many levels to be useful for creating a model. We hence created a new variable called "REGION," under which all itineraries were categorized into 5 different regional groups ("Northeast," "Midwest," "South," "West," and "Non-contiguous"; based on the U.S. Census Bureau's regional division). We also observed a high level of group difference in the mean airfare when we categorized itineraries into two groups: 1) those with one passenger who paid the same price as the itinerary's market fare (i.e. low willingness to buy the ticket at the market fare) and 2) those with more than one such passengers (i.e. high willingness to buy the ticket at the market fare). For this reason, we derived the binary variable "WILLINGNESS" from the existing variable "PASSENGERS."

Based on this preliminary evaluation process, we prioritized several potential predictors of airfare including: distance, willingness to pay, itinerary region type, quarter of the year, and number of layovers. We then created and tested our (multiple) linear regression model using the forward selection procedure. We performed the F-test on the reduced vs. full model to check whether each newly added predictor variable carries statistical significance. After adopting our final model, we diagnosed it by examining the residual plot. Finally, we calculated the model's variance inflation factor to check if it is subject to multicollinearity.

Results

The analysis of the box plot indicates that the dataset contains about 20 potential outliers with the fare higher than $600. However, even without these extreme observations, the distribution of airfare in the dataset displays a heavy positive skew (Figure 1; outliers excluded for illustrative purpose only). The minimum fare is $72; the maximum $3451; the median $223; and the mean $290.5. New York is the origin or destination for all itineraries in the dataset, two of which have New York as both origin and destination.

The forward selection procedure resulted in a multiple linear regression model with 3 predictor variables: distance, willingness to pay, and itinerary region type. The model has 7 coefficient terms as summarized in Table 1 [$F(6, 393) = 8.26$, $p < .001$, $R^2 = .11$, $R^2_{Adjusted} = .09$, $SE_{Residual} = 270.5\ (df = 393)$].

As Table 1 shows, the model predicts the fare of the imaginary flight with zero market distance, low willingness to pay, and itinerary within New York to be about $213. With the same market distance and itinerary region type, an itinerary with high willingness to pay is predicted to have a fare about $87 higher than that of an itinerary with low willingness to pay. With the same market distance and willingness to pay, an itinerary between New York and a Midwestern state is predicted to be about $46 more expensive than an itinerary within New York; an itinerary between New York and a Southern state to be about $5 more expensive; an itinerary between New York and a

Western state to be about $153 more expensive; and an itinerary between New York and a non-contiguous state/territory to be about $107 more expensive. Finally, with the same willingness and itinerary region type, the model predicts the fare to increase by about 5 cents for every 1-mile increase in the market distance.

The residual plot shows asymmetric clustering/concentration below zero and towards lower fitted values. Variance inflation factors of the model are all less than 1.5.

Discussion

The dataset displays several unusual features that call for discussion. First, as noted above, the airfare distribution is heavily skewed. This is clearly a violation of the normality assumption, which should call into question the validity of the linear regression method employed here. However, it need be pointed out that the dataset also displays another unusual feature: every itinerary in the dataset either departs from New York or arrives in New York (or both). In other words, the dataset is biased towards itineraries involving New York and is hence not very representative of all domestic flights in the U.S. Were the dataset randomly drawn from all itineraries in the U.S., the airfare distribution may have been less skewed or even normal.

But this limitation of the current dataset does not necessarily justify the validity of the linear regression method we employed here. We would rather like to find in it the room for our exploratory aim and endeavor. That is, though we do not think our linear model is appropriate for the current dataset, we do think that it can nevertheless provide some interesting speculations about the nature of airfare.

As explained above, our model advocates distance, willingness to pay, and itinerary region type as significant predictors of airfare. The statistical significance the model assigns to these variables is somewhat surprising when we consider the corresponding scatter plots that do not show much meaningful pattern. For instance, the scatter plot of market fare vs. distance hardly supports a linear relationship that the model predicts (Figure 2; outliers excluded for illustrative purpose only).

Though the model explains relatively little of the observed variability (11% at best), it nonetheless suggests that these variables do have real impact on airfare. Future research is called upon to further examine this suggestion, preferably with samples that are more representative of the U.S. domestic flight population.

Finally, it is noteworthy that our model ruled out the number of layovers and the quarter of the year from significant predictors of airfare. This result is contradictory to our initial expectation that airfare will be subject to seasonal factors as well as to layovers. Again, future research is called upon to examine whether this finding still holds for more representative flight samples.

Conclusion

In this study, we examined potential predictors of airfare based on a sample of 400 U.S. domestic flights. We employed multiple linear regression for our analysis. The result indicates that distance, willingness to pay, and itinerary region type are significant predictors of airfare in the U.S., while layovers and the time of the year are not significant predictors of airfare. Despite assumption violations and the low level of explanatory power, our model at its very least suggests that airfare is not entirely subject to random fluctuations due to such factors as variable pricing. Furthermore, it specifies a concrete group of predictors, which enables us to decide on the focus and direction of our further inquiry into airfare. Yet, due to the study's limitations, its findings and suggestions are at best preliminary and further research is needed to test them.

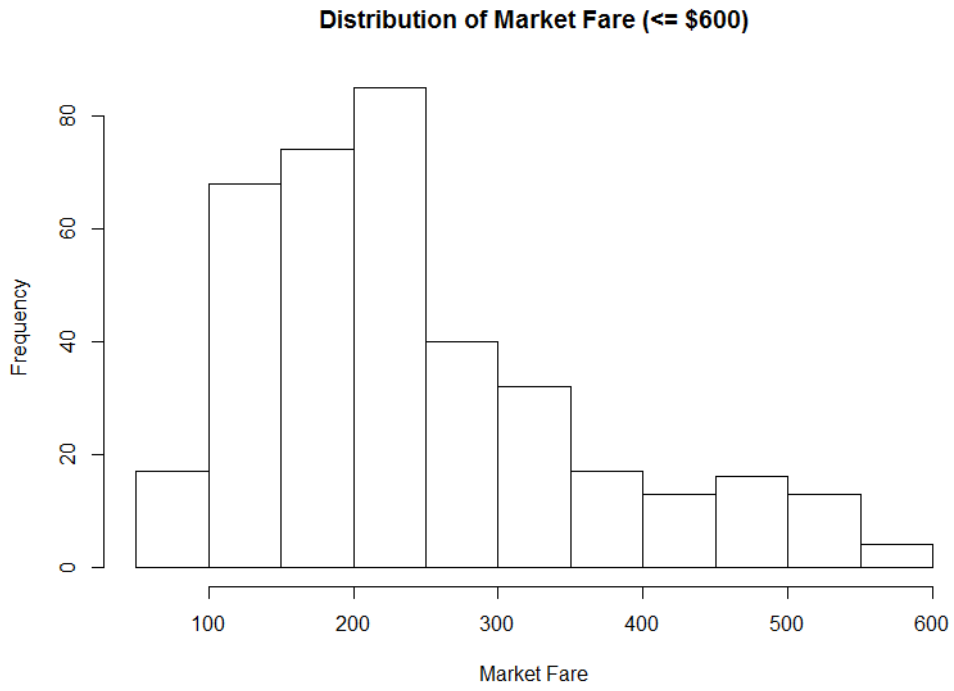

Figure 1. Distribution of Airfare

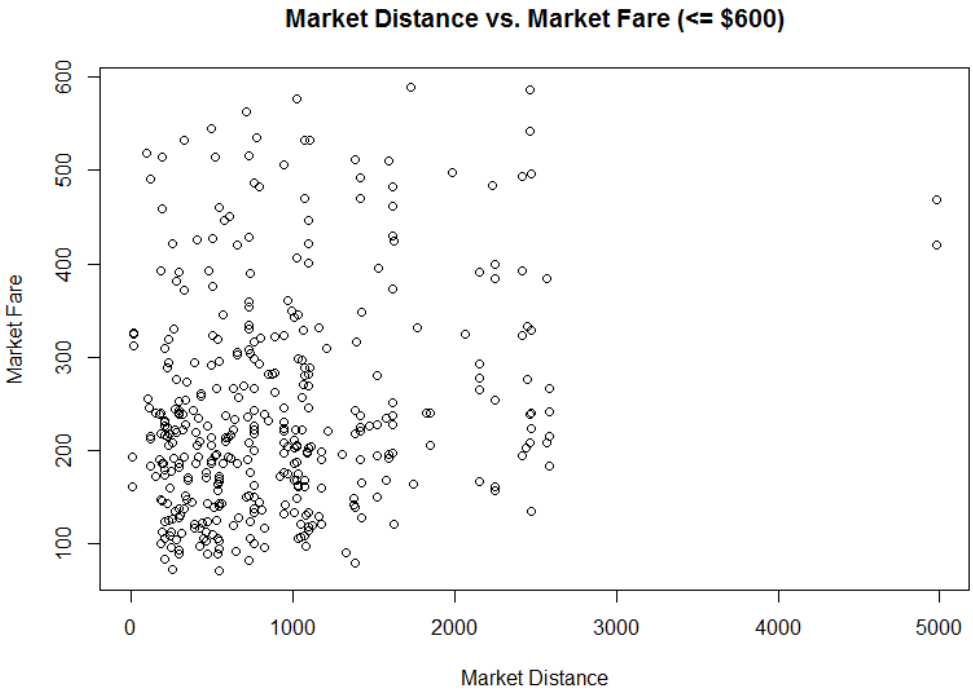

Figure 2. Flight Distance vs. Airfare

Table 1. Summary of the Multiple Linear Regression Model

Coefficient	Estimate	Standard Error	t-value	Probability (> | t |)
Intercept†	213.41168	58.23146	3.665	0.000281***
Distance	0.04955	0.02377	2.085	0.037741*
High Willingness to Pay	-87.15491	33.75489	-2.582	0.010184*
Midwestern Itinerary	45.87703	67.77106	0.677	0.498841
Southern Itinerary	4.93810	61.74517	0.080	0.936297
Western Itinerary	153.17561	71.16316	2.152	0.031970*
Non-contiguous Itinerary	106.63750	96.48249	1.105	0.269726

† Reference group: Low willingness to pay and Northeastern itinerary
** p < .05; ** p < .01; *** p < .001*