Latent Class Modeling of Passenger Airfares in the U.S. Airline Industry

Neela D. Manage
Florida Atlantic University, 777 Glades Road, Boca Raton, FL 33431

Abstract
This paper analyzes the determination of domestic passenger airfares for U.S. airlines. The sample contains quarterly panel data from 2011:Q1 to 2016:Q1 which spans a period of airline mergers as well as entry of new firms and exit of existing firms in the industry. A hedonic regression model of airfares is estimated using a latent class modeling framework in order to examine the effects of route distance, market structure, and measures of airline quality.

Key Words: Latent class models, passenger airfares, hedonic regression, panel data

1. Introduction

The U.S. airline industry has experienced several major changes in its industry structure since the deregulation of the industry in the late 1970s. Following the recessions in the early 1980s, several airlines either merged with other airlines or declared bankruptcy and ceased operations. This resulted in increased concentration in the industry in the years that followed and several studies have analyzed the effects on competition, pricing behavior, and price discrimination in the industry (Borenstein (1985, 1989, 1991, 1992), Borenstein and Rose (1994), Stavins (2001)). During the last two decades, the U.S. airline industry has continued to go through major changes. The entry of new firms, especially low-cost, discount airlines and their pricing strategies have altered the industry in a significant manner (Alves and Barbot (2009)). Also, there have been several mergers and acquisitions in the industry, some of which involved large carriers in the industry. Market concentration and price discrimination continue to have important effects on airfares and the quality of air travel (Lewis (2021) and Gill and Kim (2021)).

This study estimates a hedonic regression model of airfares using a latent class modeling framework in order to examine if there is heterogeneity in the effects of route distance, market structure, and quality of air travel on passenger airfares.

2. Model, Methodology, and Data

2.1 Model of Passenger Airfares

The airfare for the ith city-pair at time t is given by the following hedonic regression model:

$$fare_{it} = \beta_0 + \beta_1 dist_{it} + \beta_2 passengers_{it} + \beta_3 share_{it} + \sum_{j=1}^{k} \gamma_j CS_{jit} + \varepsilon_{it} \quad (1)$$

In the above equation, *fare* refers to the one-way fare for a given route, *dist* is the nonstop distance in miles between the origin city and destination city of the route, *passengers* refer to the total passengers per day on that route for all carriers serving that route, *share* is the market share of the carrier corresponding to the indicated fare in that city-pair market, CS_j, are variables that reflect convenience of air travel for customers and hence are indicators of quality of the service provided to the passenger.

Expected signs of coefficients:
$\beta_1 > 0$, $\beta_2 < 0$, $\beta_3 > 0$
$\gamma_j > 0$ if CS_j increases the convenience of air travel
$\gamma_j < 0$ if CS_j decreases the convenience of air travel

Airfares are expected to be positively associated to distance and inversely related to the number of passengers. It is assumed that an increase in the quality of a service will have a positive impact on the price. The quality variables capture the extent of flight delays, oversales resulting in bumped passengers, mishandled baggage, and consumer complaints. The fare is expected to be positively associated with travel features that improve the quality of the travel service and inversely related to features that result in more inconvenience and reduce the quality of air travel.

2.2 Methodology

This paper adopts a latent class framework to model the determination of domestic passenger airfares in the U.S. airlines industry. In this approach it is assumed that data come from distinct, but unobserved populations (called classes). Mixtures of probabilities or regression models are used to model the dependent variable.

2.2.1 Standard Linear Regression Model

Consider a regression model given by

$$y_i = x_i'\beta + \epsilon_i \qquad i = 1, 2, \ldots, n \tag{2}$$

where y is the dependent variable, x is a vector of one or more independent variables, β is a vector of regression parameters, and ϵ is a random error term.

Assume the errors are normally distributed ($\epsilon \sim N[0, \sigma^2]$) and denote $z_i = x_i'\beta$. The regression model in equation (2) is estimated by maximum likelihood based on the normality assumption.

The density function upon which the likelihood is based is given by the equation

$$\varphi(y_i; z_i, \sigma^2) = \frac{\exp[-\frac{1}{2}(y_i - z_i)^2/\sigma^2]}{\sigma\sqrt{2\pi}} \tag{3}$$

2.2.2 Latent Class Model

Consider a latent class model with two classes. The density function for a latent class model with two classes is given by

$$\gamma_1.\varphi_1(y_i; z_{i1}, \sigma_1^2) + \gamma_2.\varphi_2(y_i, z_{i2}, \sigma_2^2) \tag{4}$$

where,
$z_{ij} = x_i^{'}\beta_j$ for $j = 1, 2,$

φ_j (.) is the conditional probability density function for the observed response in the j^{th} class model, and

$0 \leq \gamma_j \leq 1$ and $\Sigma\, \gamma_j = 1$.

It is assumed that there is a true proportion γ_j of individuals in the population that are in class j. The values of γ_j are to be estimated using sample observations along with the regression estimates of β_j and σ_j (Greene (2018)).

Based on equation (4), the log likelihood for a sample of n observations is given by

$$\ln \mathrm{L} = \sum_{i=1}^{n} \ln[\gamma_1.\varphi_1(y_i; z_{i1}, \sigma_1^2) + \gamma_2.\varphi_2(y_i, z_{i2}, \sigma_2^2)] \tag{5}$$

The likelihood function in equation (4) can be estimated using the expectations maximization (EM) algorithm of Dempster, Laird, Rubin (1977).

The regression model given by equation (2) is a special case of a latent class model with a single class.

2.3 Data and Variable Definitions

The data is obtained from the U.S. Department of Transportation's Office of Aviation Analysis. For each quarter from 2011:Q1 to 2016:Q1, the fare data covers the 1000 largest domestic city-pair markets within the 48 contiguous states. For each city-pair market (route), data included the one-way nonstop distance, quarterly average of one-way passenger trips per day for all airlines, average quarterly one-way fare, market share and average fares for the airline with the largest market share and for the airline with the lowest average fare. The data on flight delays, mishandled baggage, oversales, and consumer complaints were monthly data and were averaged over each quarter.

The list of the 1000 city-pairs identified as the top 1000 markets is not constant from one quarter to the next. The sample period for this study (21 quarters) includes 21,005 observations. The descriptive statistics for the variables used in the empirical estimation are given in Table 1.

Table 1: Descriptive Statistics

Variable	*Mean*	*Std. deviation*	*Min*	*Max*
fare	225.27	68.30	69.16	559.85
nsmiles	1065.74	608.20	129	2724
npass	843.31	1333.02	148.44	21383.8
m_share	0.548	0.178	0.156	1
ot_arrival	78.00	13.95	65.3	88.35
oversales	0.78	0.46	0.005	2.07
mis_bag	3.04	0.97	0.874	5.71
ccomp	1.17	1.18	0.253	12.14

where,
fare = average quarterly one-way fare on a given route
nsmiles = nonstop distance (in miles) between the origin and destination (O&D) cities for each route
npass = quarterly average of the total number of O&D passengers per day for all airlines for each route
m_share = measure of market share for the airline whose fare is indicated. It is equal to the total O&D passengers by airline on a given route divided by the total O&D passengers on that route.
ot_arrival = the quarterly average on-time percentage for the indicated airline over all routes
oversales = indicated airline's average number of confirmed passengers per quarter who are involuntarily denied boarding due to oversales. It is measured per 10,000 passengers enplaned over all routes.
mis_bag = indicated airline's average number of reports that were filed by passengers (per quarter over all routes) for baggage loss, delay, damage, or pilferage. It is measured per 1000 passengers enplaned
ccomp = average number of complaints filed with the DOT (per quarter over all routes) per 100,000 enplaned passengers.

3. Econometric Results

3.1 Ordinary Least Squares Estimates

Prior to estimating a latent class model, the theoretical model in equation (1) was estimated by ordinary least squares (OLS). The dependent variable was *ln fare* and, in addition to the variables defined above, year dummy variables (Y2012 – Y2016) were also included as explanatory variables. The OLS regression estimates are reported in Table 2.

Table 2: OLS Estimation

Variable	Est. Coefficient
Intercept	4.422***
nsmiles	0.0003***
npass	-0.00002**
m_share	-0.058***
ot_arrival	0.007***
oversales	-0.164**
mis_bag	0.072***
ccomp	0.014***
Y2012	0.074***
Y2013	0.129***
Y2014	0.093***
Y2015	0.064***
Y2016	0.017**

*** denotes .01 level of significance
** denotes .05 level of significance
* denotes .10 level of significance

The OLS regression results in Table 2 provide estimated coefficients for a single-class model. The results suggest that fares are positively related to the distance on the flight and to on-time performance of the airline. Fares are inversely related to the number of passengers and to oversales. These results are consistent with the model described by equation (1) and the coefficients were statistically significant. However, the OLS estimation suggests that the impact of market share, mishandled baggage and consumer complaints are the opposite of their impacts outlined in the theoretical model. All the year dummy variables were positive and statistically significant.

3.2 Latent Class Model

The selection of the number of classes is based upon commonly used information criteria such as Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC). This paper uses maximum likelihood to estimate a latent class model with four groups (classes). The dependent variable was *ln fare* and year dummy variables (Y2012 – Y2016) were also included as explanatory variables. The estimated coefficients for each class and the class probabilities are reported in Table 3.

The results for Class 1 – Class 4 are reported in columns 2-5, respectively, in Table 3. The nonstop distance variable, *nsmiles*, and the number of passengers, *npass*, are statistically significant at α =.01 level and have the expected positive signs in all four classes. The market share variable, *m_share*, has different impacts on airfare across the four classes. The coefficient of *m_share* is not statistically different from

Table 3: Estimation Results for a Model with Four Latent Classes

Variable	Class 1	Class 2	Class 3	Class 4
Intercept	4.461***	4.143***	4.782***	4.227***
nsmiles	0.00032***	0.00027***	0.0001***	0.0003***
npass	-5.17e-06**	-0.0001***	-0.022***	-0.00001***
m_share	-0.147***	-0.029	-0.022	0.137***
ot_arrival	0.0015***	0.008***	0.009***	0.007***
oversales	-0.028**	-0.597***	-0.174***	-0.017
mis_bag	0.140***	0.229***	-0.049***	-0.022***
ccomp	-0.039***	-0.156***	0.125***	0.007***
Y2012	0.025**	0.287***	0.092***	0.029***
Y2013	0.048***	0.422***	0.199***	0.101***
Y2014	0.050***	0.266***	0.135***	0.112***
Y2015	0.003	0.220***	0.106***	0.088***
Y2016	-0.027	0.012***	0.198***	0.072***
Class Probability	0.192	0.045	0.263	0.500

*** denotes .01 level of significance
** denotes .05 level of significance
* denotes .10 level of significance

zero in Class 2 (which represents 5 per cent of the sample) and in Class 3 (which represents 26 per cent of the sample). The variable *m_share* has a positive and statistically significant coefficient in Class 4 which indicates that a higher market share is associated with higher fares for 50 per cent of the sample. The variable *m_share* has a negative and statistically significant coefficient in Class 1 which indicates that a higher market share is associated with lower fares for 19 per cent of the sample.

Among the airline quality variables, on-time arrival has a consistent impact on airfares across all classes. The coefficient of ot_arrival is positive and statistically significant in all four classes. The coefficient of oversales is negative, as expected, in all four classes but statistically significant only in Class 1, Class 2, and Class 3. This implies that 50 per cent of the sample is characterized by no relationship between oversales and airfares. The coefficient of mis_bag is statistically significant in all classes but has the expected negative sign only in Class 3 and Class 4. This suggests that in 76 percent of sample observations the inconvenience associated with mishandled baggage reduces the quality of air travel and is associated with lower airfares. The coefficient of ccomp is also statistically significant in all classes but has the expected negative sign only in Class 1 and Class 2. This suggests that in 24 per cent of the sample observations the inconvenience reflected in consumer complaints reduces the quality of air travel and is associated with lower airfares.

4. Summary and Conclusions

The results of the latent class modeling of domestic passenger airfares in the U.S. airline industry suggests that all the independent variables specified in the model have significant explanatory power for explaining passenger airfares. Some explanatory variables, such as the route distance, number of passengers, and on-time arrival had the expected impact on the dependent variable in all the four classes. But the signs and significance of other independent variables, such as market share and some measures of quality of air travel, suggested heterogeneity in the relationship between airfares and some of its determinants. First, the coefficient of the market share variable was positive in some classes and negative in other classes. Fares charged by an airline are likely to be positively associated with the share of the market controlled by the airline as higher concentration could result in increased pricing power. On the other hand, if a low-cost carrier has a large share in a city-pair market it could result in lower fares. Second, oversales, mishandled baggage, and consumer complaints also revealed varying impacts on airfares across the four classes. These variables reflect the quality of air travel and the associated convenience experienced by passengers.

References

Alves, C.F. and Barbot, C. (2009). “Price Discrimination Strategies of Low-cost Carriers,” *Journal of Transport Economics and Policy*, 43(3), 345-363.

Borenstein, S. (1985). “Price Discrimination in Free-Entry Markets,” *RAND Journal of Economics*, 16 (3), 380-397.

Borenstein, S. (1989). “Hubs and High Fares: Dominance and Market Power in the U.S. Airline Industry,” *RAND Journal of Economics*, 20 (3), 344-365.

Borenstein, S. (1991). “The Dominant-Firm Advantage in Multi-Product Industries: Evidence from the U.S. Airlines,” *Quarterly Journal of Economics*, 106 (4), 1237-1266.

Borenstein, S. (1992). “The Evolution of U.S. Airline Competition,” *Journal of Economic Perspectives*, 6 (2), 45-73.

Borenstein, S. and Rose, N. (1994). “Competition and Price Dispersion in the U.S. Airline Industry,” *Journal of Political Economy*, 102(4), 653-683.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). “Maximum Likelihood Likelihood Estimation from Incomplete Data via the EM Algorithm,” *Journal of the Royal Statistical Society, Series B*, 39 (1), 1-22.

Gill, R. and Kim, M. (2021). “Does Competition Increase Quality? Evidence from the U.S. Airline Industry,” *International Journal of Industrial Organization*, 77 (2021) 102742.

Greene, W.H. (2018). *Econometric Analysis*, 8th ed. (New York: Pearson)

Lewis, M. S. (2021). “Identifying Airline Price Discrimination and the Effect of Competition,” *International Journal of Industrial Organization*, 78 (2021) 102761.

Stavins, J. (2001). “Price Discrimination in the Airline Market: The Effect of Market Concentration,” *The Review of Economics and Statistics*, 83 (1), 200-202.

U.S. Department of Transportation, Office of Aviation Analysis.