US Commercial Airline Revenue: A Case Study

Problem Description

Extreme downward pressure on commercial US airline revenue has led to many airlines reviewing their long-established management practices. The US airlines are broadly divided into two categories i) sub-network carriers and ii) low-cost carriers (LCC's). Rapid growth in LCCs and competitive pricing coming with internet distribution channels, led to airline revenues falling behind their historic upward relationship with the US GDP, since late 2000s. Through our study we seek to answer the key determinants of US total operating commercial airline revenue (USD bn) for both sub-network carriers' and LCCs. We examine the following factors and asses their impact on total aircraft revenue.

1)*Supply-side factors*

- Airfare yield (passenger yield cents/ revenue passenger mile)
- Fuel costs (USD mn)
- Labour costs (USD bn)
- Total expenses (USD bn)
- Available seat miles (mn)

2) *Demand-Side factors*

- Passenger load factor (%)
- US GDP (USD bn)

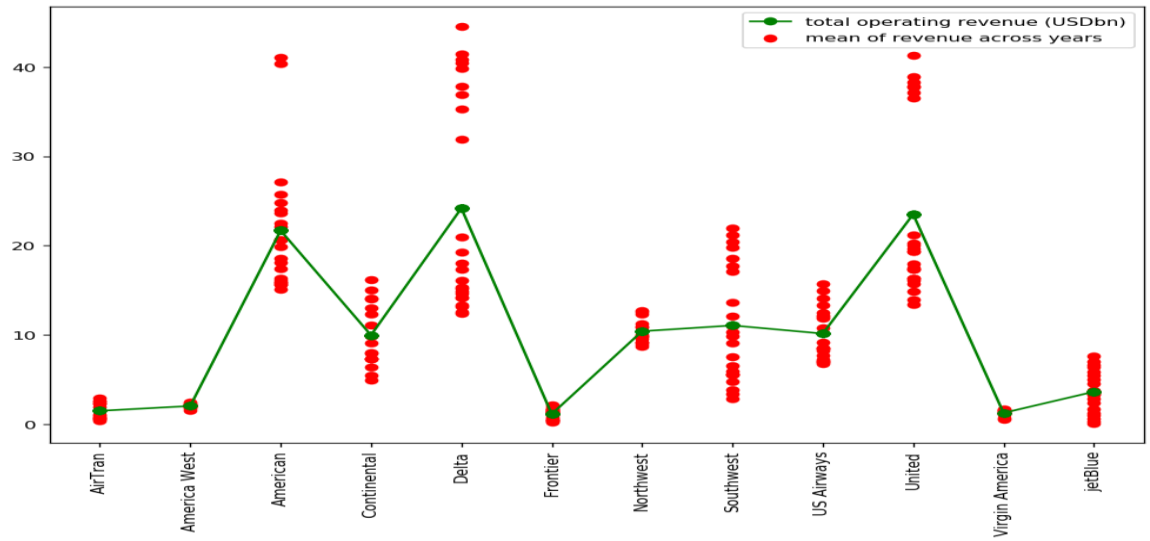
Data description and processing

We have used the data from the Airline Data Project established by the MIT Global Airline Industry Program in order to examine the total operating revenue (USDbn) of the U.S. commercial airline industry from 1995-2018. For the purpose of our study, longitudinal data was prepared by merging different variables together basis airline and year from MIT website, and missing values and outliers for each individual airline were cleaned and removed. Our dataset has both a cross-sectional and a time series dimension, whereby each airline (consisting of 7 LCCs and 5 sub network airlines) consists of a panel, spanning across a particular time dimension, unique to each airline.

Methodology

We postulate a panel data model for our study. Panel data models allow us to control for individual airline behaviour that is not quantifiable (that may be airport-invariant or time-invariant), but may affect airline revenue (like cultural factors or difference in business practices across airlines). This individual airline heterogeneity, cannot be controlled for by a simple cross sectional or times series analysis (Fig1).

Fig 1: Heterogeneity across airlines



$Y_{it}= \beta_0 + u_i + \beta_1(Fuel\_ cost)_{it} + \beta_2(labour\_cost)_{it} + \beta_3(airfare)_{it} + \beta_4 (US\_GDP)_t + \beta_5 (passenger\_load\_factor)_{it} + \beta_6 (airfare)_{it} + \varepsilon_{it}$

- For, i = 1,2....12 airlines (comprising of 7 sub-network-carriers and 5 LCCs) & t = 1995-2018;
- Y_{it}= Airline revenue (USD bn) for airline i in period t
- u_i= Individual airline heterogeneity that's specific to an airline, but constant over time
- USD_GDP: constant for each airline but changing over time
- The rest of the explanatory variables (including the error term), have both time and airline specific dimension.

The, classical linear regression model that uses OLS in order to estimate the impact of the independent variables on the dependent variable, fails in the case of a panel data model as:

1)*Exogeneity:* Each individual airline has unobserved heterogeneity which is consistent over time but specific to the airline(u_i). Since $u_{(i)}$ is unobserved, and likely to be correlated with the independent variables, the OLS estimator of β is biased and inconsistent as a consequence of an omitted variable bias. In order to address exogeneity, a fixed effects model is proposed that eliminates u_i by demeaning the variables using the *within* transformation:

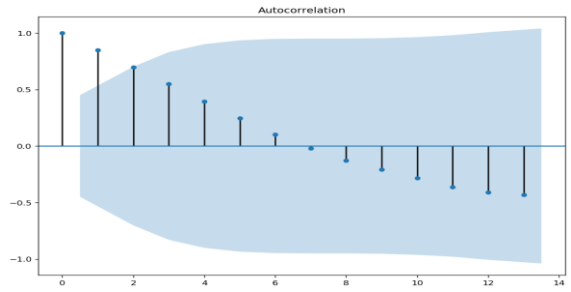
$$If,\ Y_{it} = \beta_0 + (u_i) + \beta_1 (X_{it}) + (\varepsilon_{it});$$

Within Fixed Effects Transformation: $(Y_{it} - \bar{Y}_i) = \beta_0 + (u_i - \bar{u}_i) + \beta_1 (X_{it} - \bar{X}_i) + (\varepsilon_{it} - \bar{\varepsilon}_i)$

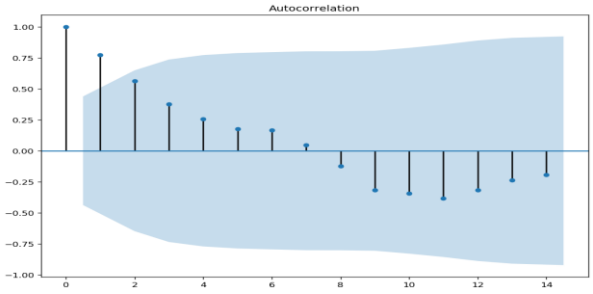
2) ***Autocorrelation in the errors:*** Since each individual airline i is repeatedly observed over time (contrary to pooled data), it is likely that $E(u_{is}, u_{it}/X) \neq 0$ (for $t \neq s$); henceforth OLS estimation of will be β inefficient. When autocorrelation function (ACF) for absolute revenue of each airline was studied, we noticed a correlation at lag1 for most airlines (Fig2-shown for 1 LCC and 1 sub network carrier). Thus, to account for autocorrelation, the growth rate (% change)

of the dependent on independent variables was considered, as opposed to absolute values. This eliminated the first period autocorrelation (Fig-2).

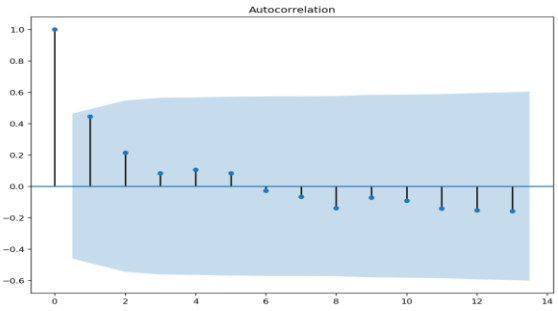
Fig2: Analysing autocorrelation of the dependent variable



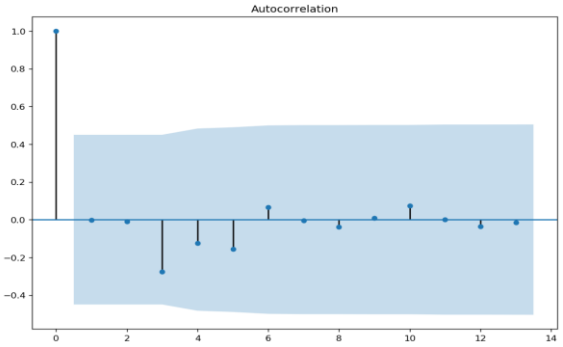
ACF for revenue of US Airways shows correlation at lag 1



ACF for revenue of JetBlue shows correlation at lag 1



ACF for revenue (%) of US Airways shows correlation at lag 0

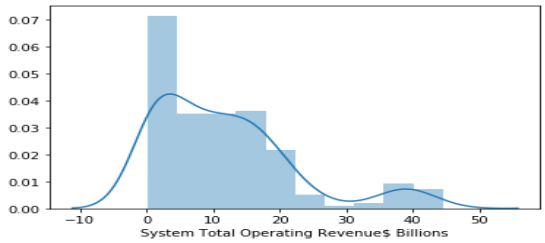


ACF for revenue (%) of JetBlue shows correlation at lag 0

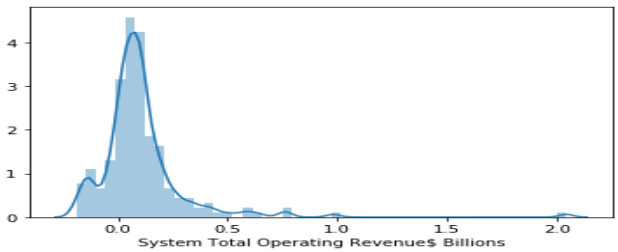
Analytical Process

- When we examine the dependent variable -System total operating airline revenue (USD bn), we see evidence of a multimodal distribution. Next, we visualise total operating revenue (% growth), whereby, while variance reduces, multimodal nature of the distribution still persists. However, upon the within fixed-effects transformation (outlined above), the variable displays normality and this validated by the Q-Q plot.

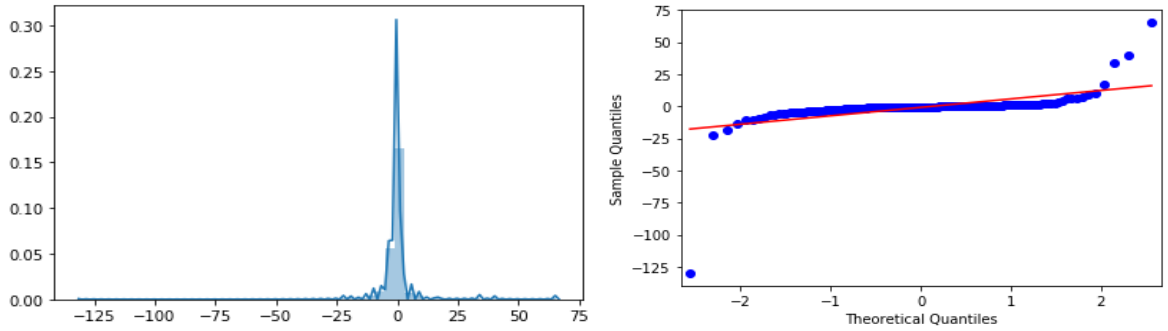
Fig3: Analysing the distribution of the dependent variable



Aircraft Revenue shows multimodal distribution

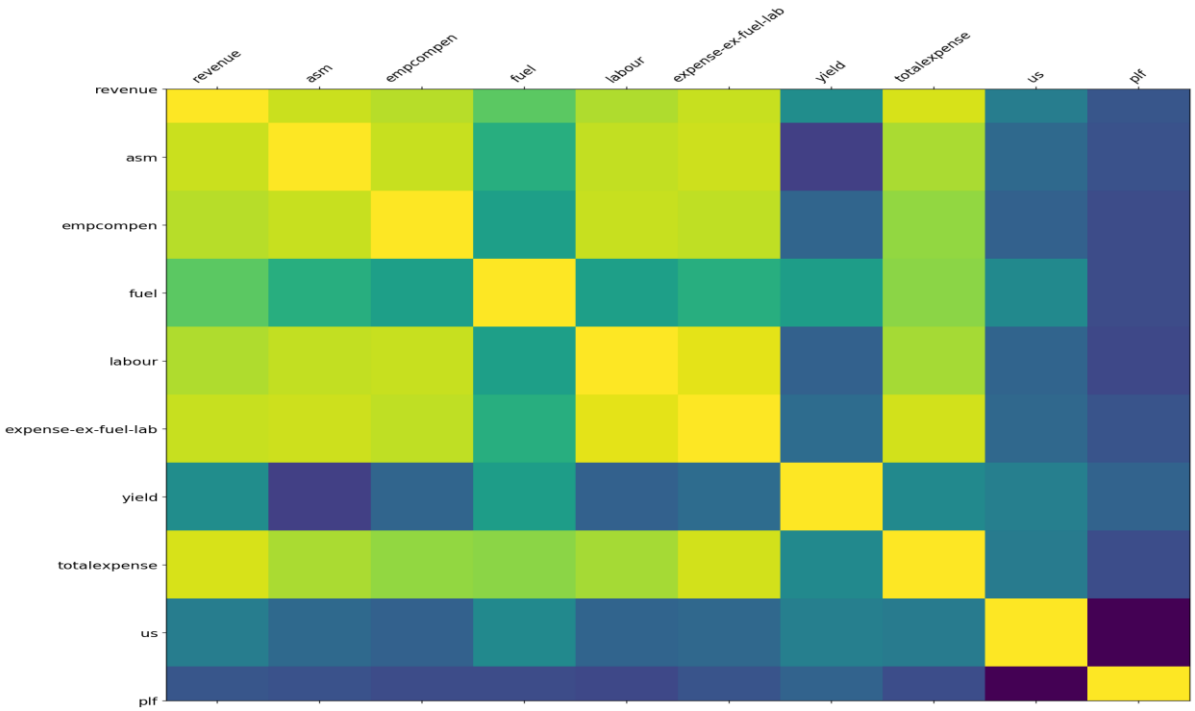


With Aircraft Revenue growth while variance of aircraft revenue reduces, multimodal nature persists



Post fixed effect transformation, KDE shows the dependent variable tends to normality, validated by the Q-Q plot

Fig4: Correlation matrix of variables



- All independent variables apart from '*passenger load factor growth'* and '*yield growth'* showed a high level of correlation with airline revenue growth. '*No. of Employees'* and '*labour wages'* showed high correlation, thus to avoid collinearity, only no. of employees (one with higher correlation with revenue growth) was chosen. '*Available Seat Miles (ASM)'* showed high collinearity with all expense related variables, likely suggesting the direct impact of operating expenses on ASM, and thus was dropped to avoid collinearity. *However, while correlation matrix gives the degree of linear dependence between two variables, it doesn't help us solve the problem of multicollinearity, which can emerge even when three or more variables are highly correlated. The issue of multicollinearity is considered later in the model.*

- *Feature Creation:* In order to estimate impact of *no. of employees* on the growth rate of airline revenue, the '*square of the no. of employees*' was chosen, since it displayed

higher correlation with revenue growth (validated by quadratic relationship between two seen in the scatter plot). A new feature the' *total expense excluding fuel and labour costs'* is crafted, in order to study the remaining impact of expenses on growth rate, post accounting for fuel costs and labour wages (these two variables already considered).

Findings

OLS Regression Results

Dep. Variable:	revenue	R-squared:	0.900
Model:	OLS	Adj. R-squared:	0.900
Method:	Least Squares	F-statistic:	4413.
Date:	Sun, 15 Dec 2019	Prob (F-statistic):	0.00
Time:	23:06:03	Log-Likelihood:	3339.6
No. Observations:	2460	AIC:	-6667.
Df Residuals:	2454	BIC:	-6632.
Df Model:	5		
Covariance Type:	nonrobust		

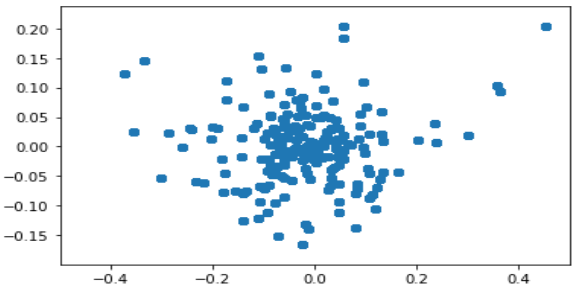
	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0094	0.001	-7.310	0.000	-0.012	-0.007
fuel	0.1316	0.005	25.779	0.000	0.122	0.142
expense-ex-fuel-lab	0.6588	0.012	55.949	0.000	0.636	0.682
yield	0.3481	0.016	21.479	0.000	0.316	0.380
us	1.1929	0.081	14.747	0.000	1.034	1.352
employeesq	0.2200	0.006	39.252	0.000	0.209	0.231

Omnibus:	79.353	Durbin-Watson:	1.668
Prob(Omnibus):	0.000	Jarque-Bera (JB):	122.151
Skew:	0.304	Prob(JB):	2.99e-27
Kurtosis:	3.907	Cond. No.	64.5

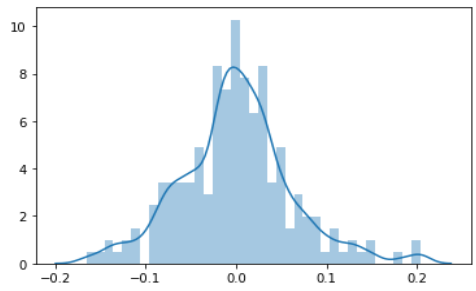
1) Model Validation

- The scatter plot of residuals against the predicted values are randomly scattered around 0, indicating our model's predictions are correct on an average (Fig 5).
- The distribution and Q-Q plot of residuals shows evidence of normality (Fig 5).
- Both Breusch Pagan and White test show a test statistic > 0.05, thus we fail to reject the null hypothesis of homoscedasticity
- The Durbin Watson test statistic of 1.7~2, shows evidence of no-autocorrelation in error terms
- Adjusted R^2 shows a value of 90% showing a good fit to the model (as more predictors are added to the model R^2 will always rise, where as adjusted R^2 will rise or fall depending on goodness of fit of the model)
- Low VIF of the predictors <3, indicates multicollinearity was not a problem.

Fig5: Model Validation

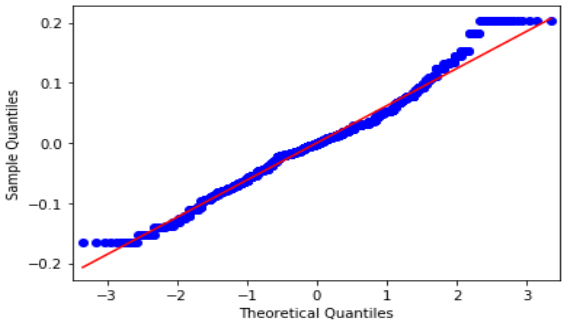


Scatter plot of residuals plotted against the fitted values of Y is random



Distribution of Residuals showing normality

Fig5: Model Validation



Q-Q plot of residuals showed evidence of normal distribution

	VIF Factor	features
0	1.0	const
1	1.6	fuel
2	2.1	expense-ex-fuel-lab
3	1.3	yield
4	1.2	us
5	1.8	employeesq

Low VIF of the predictors, indicates multicollinearity was not a problem

2) **Interpreting the Results**

- We used Recursive Feature Elimination to select variables. All supply-side variables showed a positive impact on airline revenue growth and were statistically significant at 0.01% level of significance (indicated by the p values)
- In line with expectations (as seen by the correlation plot), passenger load factor wasn't statistically significant and hence was excluded from the model.
- *The demand side factor 'US GDP growth' is expected to have the largest impact on airline revenue growth, ceteris paribus, with a 1% rise in US GDP showing a 1.19% rise in aircraft revenue;* ***displaying the importance of an uptick in consumer demand to propel airline revenue growth.***
- *Among supply side factor, 'total expenses excluding- fuel and- labour', had the largest impact on airline revenue growth, ceteris paribus; with a 1% rise in expense showing a 0.65% rise in aircraft revenue;* ***displaying the importance of investing in general and administrative expenses on airline revenue growth*****.**

Reflections

- **Panel data gives more variability, less collinearity among the variables, more degrees of freedom, and greater efficiency.** Time series data is expected to be plagued with multicollinearity. However, adding the cross-sectional dimension along with time series in panel data; adds a lot of variability, adding more informative data on explanatory variables that are expected to impact commercial airline revenue growth (Baltagi, 2008).
- **Panel data is better able to study the dynamics of adjustment over time of a unit.** Cross-sectional distributions camouflage an assembly of changes over time, however, panel data shows what proportion of factors are varying over time for each particular airline, thus highlighting both intra and inter temporal variations (Baltagi, 2008).
- **Panel data assumes Cross-sectional independence.** However, accounting for cross-sectional dependence is crucial, as it is likely to bias our hypothesis (Baltagi, 2008).
- We used recursive feature elimination in order to study the variables impacting airline revenue growth, and as such, a manual approach may not always be optimal. While at this juncture, it seems rational to resort to a method like PCA to eliminate collinearity and institute feature selection; this would have a big loophole in the case of a panel dataset. PCA would

just capture covariance between variables for feature selection, but not time and individual heterogeneity (u_i) specific to airline, which was taken care of by fixed effects regression.

- Another approach would be to use penalising models like lasso or ridge regression; however, lasso would arbitrarily select one feature from a group of corelated features. This arbitrary approach having low interpretability (likely to lead to losing of important information on features) isn't best in case in our case *where the goal is inference of variables, and not predictability.* Further, in case of ridge regression all variables will be retained, which would undermine the key objective of our exercise, interpretability, i.e. finding the determinants and their impact on airline revenue growth.

References

- Econometric Analysis of Panel Data, Baltagi, 2008
- An Econometric Dynamic Model to estimate passenger demand for air transport industry, Benítez & Miranda, 2016
- Productivity Trends in the US Passenger Airline Industry 1978-2010; Belobaba, Jenkins, Powell, Swelbar, 2011