PDA Report – Airlines Model: PDA

Section 56

Winter Quarter

****

School of Continuing Studies

Northwestern University

****

Daniel Prusinski

In compliance for Master of Science Predictive Analytics

Bachelor of Science Business Marketing

****

Program Analyst

Wooddale Church

6630 Shady Oak Road

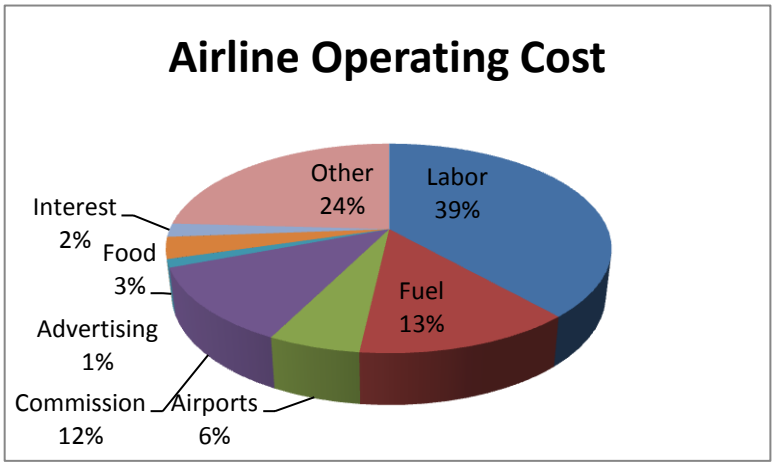
Eden Prairie, MN 55344

Executive Summary

In the current economic climate, keeping total cost down is a matter of survival for any competitive airline industry. Through the initial exploratory data analysis (EDA), it was found that the dataset was optimal for Panel Data analysis. After analyzing Pooled, Fixed Effect, and Random Effects regression techniques, it was concluded from the Lagrange Multiplier test that Random Effects regression was the best model. It should be noted that all three models had strong predictive qualities as well as statistically valid output. Specifically, Fixed Effects produced a solid second place model utilizing airline ID as the dummy subject variable. This analysis will equip executives to better understand its bottom line and how to remain profitable.

Introduction

In 1978, the US government deregulated the airline industry and as a result over $60 billion has been lost to-date through airlines filing for bankruptcy (Npr.org & Severin Borenstein). For the deregulated airline industry the game is quite simple, cover your total costs or cease to exist in your business model. In order for an airline to stay profitable, it must understand the dynamics between its total cost (C-dependent variable) with revenue (Q), price of fuel (PF), and capacity utilization (LF) -independent variables. When studying profitability in the airline industry, time (T) and airline companies (I) are additional important variables that add depth and breadth to the EDA.



The pie chart to the left was created by Charles Najda, from the Department of Economics at Stanford University, and visually breaks down airline operating

costs. Fuel only represents 13 percent of operating costs and capacity utilization does not encompass operating costs, thus I suspect neither of these will have a strong correlative relationship with total cost. For this report, revenue is expressed as follows: Revenue Per Passenger Miles, and can be understood as the more miles a passenger accumulates, the greater the total cost for the airline. Revenue per passenger miles as a variable encompasses all the operational expenses of an airline and I suspect it will have a strong correlative relationship with total cost. I expect time and total cost to have a positive relationship, such that as time increase cost rises as well. Airline ID is a rather arbitrary variable, and I expect specific ID's to follow an inclusive pattern with total cost. A further analysis of these dynamics will aid airline executives in better understanding its bottom line and how to remain profitable.

Analysis

In order to meet the objective of exploring the relationship between the dependent variable and independent variables, an exploratory data analysis must be conducted. I will be utilizing the EDA paradigm and structure put forth by Bruce Ratner found in his book *Statistical and Machine-Learning Data Mining*:

> Problem/Objective: Explore the relationship between Total Cost (C), and the independent variables Revenue Per Passenger Miles (Q), Price of Fuel (PF) Time (T), Airline (I), and Capacity Utilization for load factor (LF). While exploring the relationship, I will conduct both fixed and random panel data analysis on this dataset.
>
> Data: The data has been aggregated and has been supplied from management.
>
> Analysis: Scatter plots and correlation coefficients will be used to study the nature of the relationships between the independent variables and their relation to the dependent variable. I will conduct a pooled analysis and briefly comment on the overall findings.
>
> Model: After assessing the data, a model will be used. Management has recommended using a regression model, but the standard OLS assumptions will need to be validated. In addition, least squares dummy variable (LSDV), fixed and random panel models will be used to draw further conclusions from the data.

Results/Interpretation: Once the model has been validated and iterations complete, a recommendation will be written to management in regard to the relational dynamics amongst the variables listed above. In addition, the Beusch and Pagan Lagrange Multiplier test or the Hausman Specification test will be used to determine which model should be used.

A properly executed EDA for management must reflect that the data was the driving force behind constructing the model. The steps outlined above are ordered such that the data drives which model is used, and the analyst's personal bias is mitigated.

Data

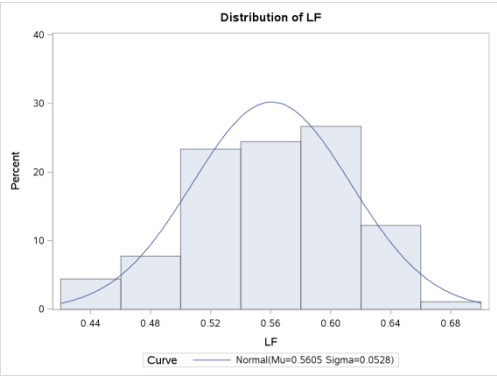
Following the outline above, exploring the data is the next step for the EDA. There are a total of 90 observations with 0 missing values in the data set for each variable. The response variable along with two of the independent variables requires a data transformation in order to better study the variables. Utilizing log transformations alter the data to a fairly standard shape (Ajmani). Specifically, this technique is used for positively skewed data and the result of the log transformation moves the majority of the data such that it follows a normal distribution.

Each variable has its own descriptive breakdown explained in a subsection below.

Capacity Utilization (LF): This variable did not require a data transformation and represents the utilization of overall capacity for the airplane load factor.

Descriptive Stats for Variable: Capacity Utilization as LF							
N	Miss	Minimum	Maximum	Median	Mean	Variance	Std Dev
90	0	0.432	0.676	0.566	0.560	0.003	0.053

LF is the easiest variable to understand given that the minimum and maximum values are less than one and the difference is .244. In addition, the mean and median are relatively close to one another which would lead me to believe there is a small standard

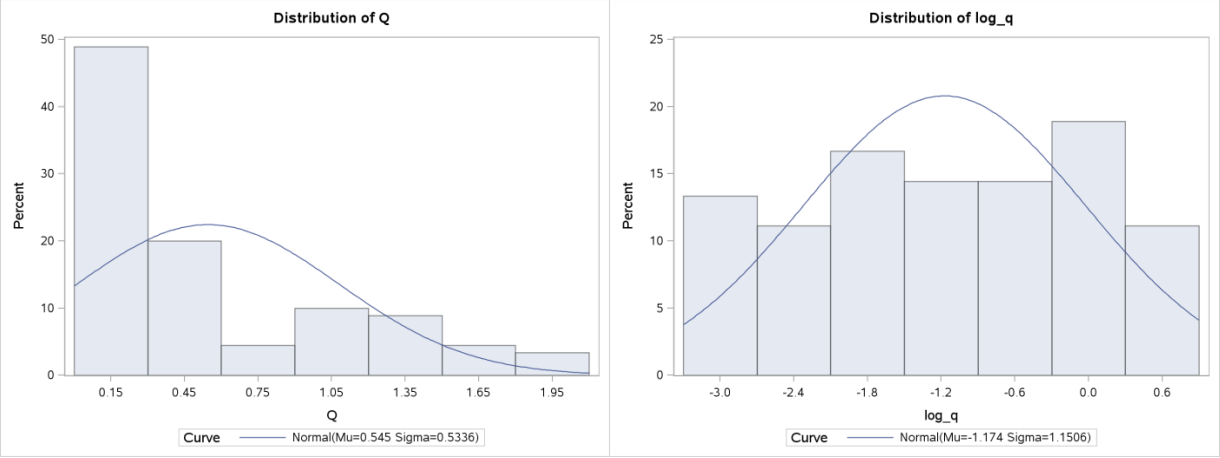


deviation (SD). The variance is small, of which the SD is based. The visual demonstration, via the histogram to the right, reveals exactly what would be expected from the table above. This variable is slightly negatively skewed, but overall is an excellent variable to conduct analysis.

Revenue Passenger Miles (Q and LogQ): Variables often need transformation in order to be better understood and presented in a form that is conducive to iterative analysis. In this analysis, variable Q needed a log transformation.

Descriptive Stats for Variable: Log_Q and Q									
Variable	**Label**	**N**	**Miss**	**Minimum**	**Maximum**	**Median**	**Mean**	**Variance**	**Std Dev**
log_q		90	0	-3.279	0.661	-1.187	-1.174	1.324	1.151
Q	Q	90	0	0.038	1.936	0.305	0.545	0.285	0.534

At first glance, this variable might appear to not need a transformation based on the descriptive statistics, the min, max, variance and SD all look fine. The red flag that caught my eye was the difference between the median and mean, which suggests that the observations are not normally distributed. After the log transformation, the mean and median are much closer.

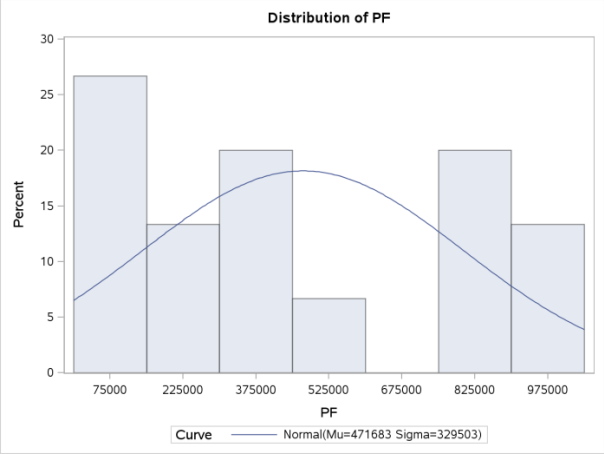


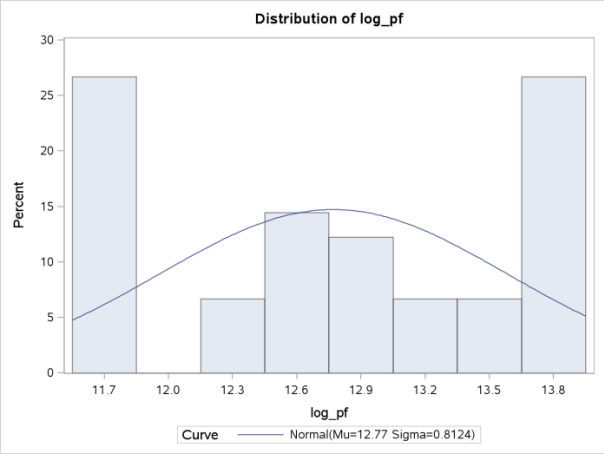
The histogram reveals the severely skewed variable Q, which is why visual statistics are an important asset to EDA. After the log transformation, variable Q follows a normal distribution with only a slight negative skew of -.1.

Price of Fuel (PF and LogPF): Similar to variable Q, variable PF needed a log transformation.

Descriptive Stats for Variable: Log PF and PF									
Variable	Label	N	Miss	Minimum	Maximum	Median	Mean	Variance	Std Dev
log_pf		90	0	11.550	13.831	12.787	12.770	0.660	0.812
PF	PF	90	0	103795.000	1015610.000	357433.500	471683.011	108572166191	329502.908

In its original form, variable PF is very hard to understand. Grasping the variance and (SD) is rather trivial given the sheer size of the numbers. In addition, one should note the difference between mean and median. After the log transformation, the min and max are not far apart. The median and mode would lead me to believe there is a relatively normal distribution, and the variance/SD fits the variable.



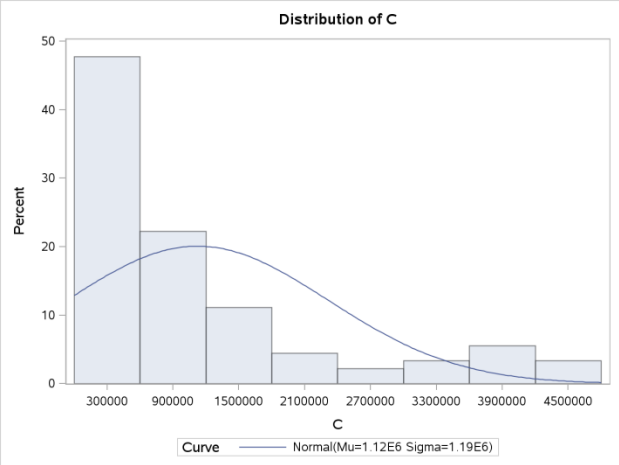


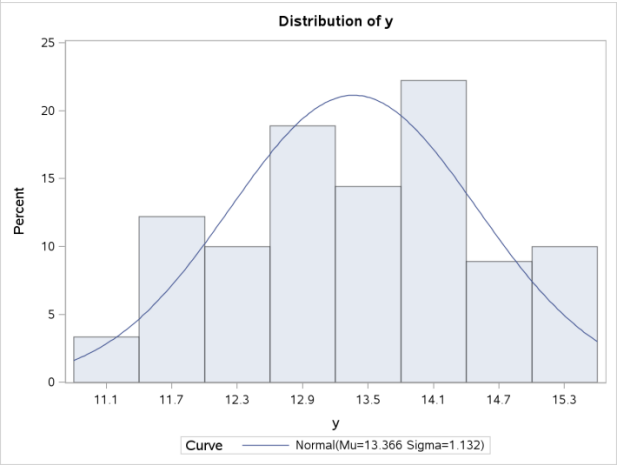
Variable PF in its original form is positively skewed .40, and appears to not follow a normal distribution. After the log transformation, the skew is only -.14, and the distribution allows one to conduct further analysis.

Total Cost (C and LogC expressed as y): This variable represents the dependent variable, and is expressed in millions of dollars.

Variable	Label	N	Miss	Minimum	Maximum	Median	Mean	Variance	Std Dev
y		90	0	11.142	15.373	13.365	13.366	1.281	1.132
C	C	90	0	68978.000	4748320.000	637001.000	1122523.833	1.4210421E12	1192074.704

Variable C is very similar to variable PF in that its large numbers are hard to understand. In addition, the large difference between median and mean suggest that a log transformation is necessary. After the log transformation, expressed as y, the numbers are perceivable and the median and mean fall close to one another.





Before the log transformation, C had a massive positive skew of 1.53 and would be a difficult variable to analyze. After the log transformation, the skew is only -.10 and the variable is easier to understand.

Airline Companies (I): In this data set, there are a total of six airline companies being studied. As demonstrated in the table to the left, each airline company accounts for 16.67 percent of the total data being studied. In addition, each airline has 15 individual observations that represent a year. If one were to conduct an individual analysis on each

I = Airline Companies				
I	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	15	16.67	15	16.67
2	15	16.67	30	33.33
3	15	16.67	45	50.00
4	15	16.67	60	66.67
5	15	16.67	75	83.33
6	15	16.67	90	100.00

airline, the results would not be statistically significant, but pooling the airlines together creates 90 data points, which is statistically more significant.

Time (T): This is the first EDA that is analyzing time in conjunction with other independent/dependent variables. The time variable is split into 15 years. It can be seen that each year accounts for 6.67 percent of the total time variable. In addition, each airline company is represented once though each year.

T = Time				
T	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	6	6.67	6	6.67
2	6	6.67	12	13.33
3	6	6.67	18	20.00
4	6	6.67	24	26.67
5	6	6.67	30	33.33
6	6	6.67	36	40.00
7	6	6.67	42	46.67
8	6	6.67	48	53.33
9	6	6.67	54	60.00
10	6	6.67	60	66.67
11	6	6.67	66	73.33
12	6	6.67	72	80.00
13	6	6.67	78	86.67
14	6	6.67	84	93.33
15	6	6.67	90	100.00

The data has been analyzed in its original form and transformed for better analysis. From analyzing variable I, it can be seen that an individual airline only produces 15 data points. It would be unwise to conduct a thorough EDA on an individual airline because the sample dataset is very small. Thus, using a method that utilizes panel data is desired since this method increases the sample size and includes unobserved heterogeneity effects in the model.

Results

Unobserved heterogeneity is a new concept that I would like to further explain. When building an OLS model, issues/model biases can arise from including or excluding important variables from the model that affect how one interprets the results. Excluding an important variable, either knowingly or unknowingly, to the model is called unobservable heterogeneity (Ajmani 2013). Utilizing panel data models allows one to better control unobserved

heterogeneity as well as increase the sample size. Pooled, fixed effects, and random effects are specific model techniques that are used to analyze panel data.

From the data analysis, management has encouraged initially using a pooled regression model utilizing Ordinary Least Squares (OLS) to estimate the parameters for the data. Through using this model and method, a brief overview will be given for the overall findings.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	112.70545	37.56848	2419.34	<.0001
Error	86	1.33544	0.01553		
Total	89	114.04089			
Root MSE	0.12461	R-Square		0.9883	
Dependent Mean	13.36561	Adj R-Sq		0.9879	
Coeff Var	0.93234				

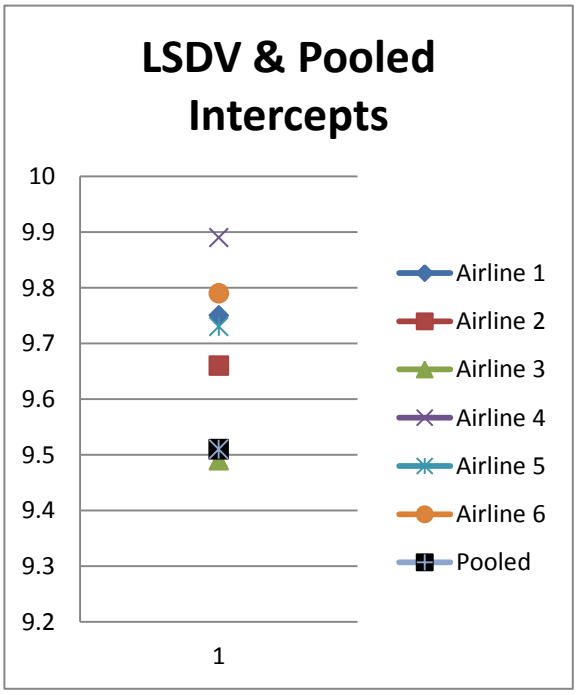
Preliminarily, the model has a strong R-squared but the Adj R-squared is preferred given that the model has more than one variable. The F-value is very significant based on three degrees of freedom. This can be interpreted as at least one variable is explanative of the dependent variable in the model.

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	9.51692	0.22924	41.51	<.0001	0
log_q	1	0.88274	0.01325	66.60	<.0001	1.33304
log_pf	1	0.45398	0.02030	22.36	<.0001	1.55936
LF	1	-1.62751	0.34530	-4.71	<.0001	1.90468

Statistically the variables are significant, and Log Q has the largest r-squared. The variance inflation factors (VIFs) do not warrant concern for multi-collinearity. The overall model has strong predictive qualities and is statistically significant, but in order to use this model the OLS assumptions need to be validated. Load factor is the only variable to have an inverse relationship, while LogQ and LogPF have positive relationships with total cost given the caveat than when interpreted the other variables are held constant.

Fixed effect (FE) models allow for different intercepts per subject (airline) in the data set, but assumes that the slopes are constant, ie parallel to one another. Least squares dummy variables (LSDV) model is a technique within FE. LSDV captures unobserved heterogeneity through creating dummy variables for each subject (airline) in the model. For this particular dataset there are only 6 subjects being studied, but in other datasets where there might be more subjects creating individual dummy variables becomes cumbersome and convoluted. Given that each subject has its own dummy variable and subsequent intercept, there is no constant intercept and one avoids a "dummy-variable trap" (Ajamani 2013). My interpretation of the LSDV model is that it draws on the benefit of a larger sample size, ie the coefficients will be more statistically sound, but also gives each subject (airline) the autonomy of its own intercept.

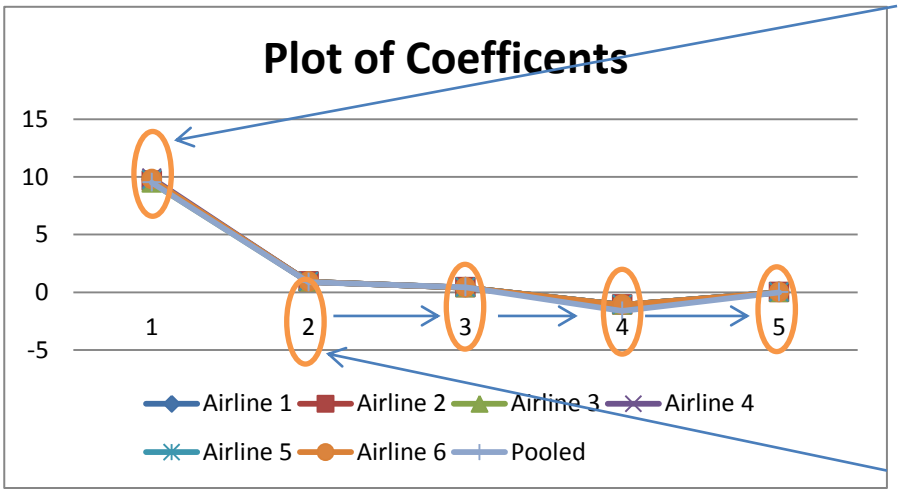
After fitting the LSDV model to the airlines data running the appropriate SAS code, the dummy intercept variables along with the pooled intercept are graphically displayed to the left. I created this graphic to visually demonstrate the 6 different intercepts along with the intercept for the pooled model. 5 of the 6 airlines all have intercepts that are greater than the pooled intercept. Each airline has its own intercept, and subsequently the coefficient of determination is greater, as a result of the precise intercepts, than the pooled model and the mean square error is smaller. Highlighting the intercepts in the LSDV model is the key differentiator compared to pooled OLS regression. After highlighting the autonomous intercepts, reviewing the variable coefficients are rather monotonous since they



remain

the same throughout the model. However, in order to spice up the interpretation a comparison of the pooled regression model will be referenced. All the subject intercepts are compared to airline 6, which can be quite confusing to interpret. Airlines one and two do not have statistically strong estimates. Revenue per passenger (Q) grew slightly stronger in its size in the LSDV model. Load factor dropped substantially as a coefficient. Graphically the model has the appearance of a six- headed snake with one body. If one were to zoom in on the first plot, there would be seven individual points. Notice how I have included the pooled regression coefficients as well. After the first intercept, all the coefficients for the LSDV model remain the same.

LSDV Coefficients and Intercept				
Para	**Estimate**	**SE**	**t Value**	**Pr > |t|**
Inter	9.793	0.26	37.14	<.000
I 1	-0.087	0.08	-1.03	0.304
I 2	-0.128	0.07	-1.69	0.094
I 3	-0.296	0.05	-5.92	<.000
I 4	0.097	0.03	2.95	0.004
I 5	-0.063	0.02	-2.64	0.01
I 6	0.000	.	.	.
LnQ	0.919	0.03	30.76	<.000
LnPF	0.417	0.01	27.47	<.000
LF	-1.070	0.20	-5.31	<.000
Pooled Model				
Inter	9.517	0.22	41.51	<.000
log_q	0.883	0.01	66.60	<.000
log_pf	0.454	0.02	22.36	<.000
LF	-1.628	0.34	-4.71	<.000



Looking at the model diagnostics one can see that the LSDV model has a better fit. The coefficient of determination is higher for the LSDV model, which is result of the six individual dummy intercepts. A topic I would like to further investigate is how the LSDV model takes into

account over fitting. It would seem that this approach would be prone to some of the pitfalls surrounding over fitting a model. The LSDV model captures the robust nature of pooled data, but has the flexibility of individual subject intercepts.

LSDV Model				
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**
Model	8	113.748	14.22	3935.8
Error	81	0.293	0.00	P-Value
Correct Total	89	114.04089		<.0001
Pooled Model				
Model	3	112.465	14.22	3935.8
Error	86	1.335	0.00	P-Value
Correct Total	89	114.04089		<.0001

Another approach to analyzing the airline data utilizing the fixed method is to make Time the subject such that the data is divided up by year. Given that there are 15 time periods, there will be 15 dummy variable intercepts for this model. The highlighted variables lack statistical significance to warrant using in a final model. Notice how the variable LnPF is now statistically insignificant. Bear in mind, the variable coefficients in the model will remain the same since we are drawing on pooled data, which is the same approach used in the example above. The FE technique does not have a shared intercept, thus SAS uses the last subject as the intercept. Time period 15 is the intercept and the other subject dummy variables are compared to its value. For example, the intercept for T1 can be interpreted as having an intercept of 22.537-2.041= 20.496 and this applies to the other subject dummy variables but not to the variable coefficients. This is the same approach for interpreting the

FE Time				
Para	**Estimate**	**Std Error**	**t Value**	**Pr > |t|**
Inter	22.537	4.941	4.56	<.0001
T 1	-2.041	0.735	-2.78	0.007
T 2	-1.959	0.723	-2.71	0.008
T 3	-1.881	0.720	-2.61	0.011
T 4	-1.796	0.699	-2.57	0.012
T 5	-1.337	0.506	-2.64	0.010
T 6	-1.125	0.409	-2.75	0.008
T 7	-1.033	0.376	-2.75	0.008
T 8	-0.883	0.326	-2.71	0.009
T 9	-0.707	0.295	-2.4	0.019
T 10	-0.423	0.167	-2.54	0.013
T 11	-0.071	0.072	-1	0.323
T 12	0.115	0.098	1.16	0.248
T 13	0.080	0.084	0.95	0.348
T 14	0.015	0.073	0.21	0.832
T 15	0.000	This is the intercpet		
LnQ	0.868	0.015	56.32	<.0001
LnPF	-0.484	0.364	-1.33	0.188
LF	-1.954	0.442	-4.42	<.0001

coefficients that was used when airline ID was the subject. I've created the chart to the right in order to highlight a few of the summary numbers behind the different FE models along with the pooled model. I would have expected Time to be a better fit than ID based on the amount of dummy subject variables in the model. When analyzing Time and ID together, there is almost a near perfect fit. Bear in mind there are 21 dummy subject variables in addition to three fixed variables. In my novice opinion, one would really have to gird against over fitting the model.

Model	Time	ID	Pooled	T & ID
R-Squ	0.990	0.997	0.988	0.998
Coe Var	0.920	0.450	0.932	NA
R MSE	0.123	0.060	0.125	0.051

How does one gauge whether or not adding a variable/s is helpful for the model? One approach to answering this question is to analyze the sum of squares in conjunction with each new variable being added to the model as well as the variable standing alone with just the intercept. My favorite FE model was ID and my least favorite model was Time. The Type 1 test to the left is measuring how much the residual sums of squares (RSS) is reduced by just adding the specific variable and the constant. The Time model is okay in my opinion. Given how many dummy subject variables are in the model, I was hoping the T variable would reduce the RSS in a greater capacity. In addition, not how LnPF is not a statistically significant variable. The ID model paints a better picture. Here, ID is substantially contributing to the reduction of RSS, and all the variables play well together. The ID model is parsimonious

Time Model			
Source	Type I	F Value	Pr > F
T	37.307	176.31	<.0001
LnQ	75.303	4982.42	<.0001
LnPF	0.048	3.16	0.0797
LF	0.295	19.52	<.0001
ID Model			
ID	74.680	4134.390	<.0001
LnQ	36.333	10057.300	<.0001
LnPF	2.634	729.000	<.0001
LF	0.102	28.170	<.0001

.

The type 3 test measures how much the RSS is reduced when the specific variable is added to the model with all the other variables in the model. I prefer this test when dealing with tawdry variables, as it measures how much they are really contributing as a team. As expected, the time model is a mess. LnQ is the only major contributor and Time/LnPF are not statistically valid. On the other hand, the ID model shows a team of variables that are contributing rather evenly to the reduction of the RRS and all the variables are statistically significant.

Time Model			
Source	**Type III**	**F Value**	**Pr > F**
T	0.247	1.170	0.318
LnQ	47.933	3171.480	<.0001
LnPF	0.027	1.770	0.188
LF	0.295	19.520	<.0001
ID Model			
ID	1.043	57.730	<.0001
LnQ	3.417	945.900	<.0001
LnPF	2.726	754.500	<.0001
LF	0.102	28.170	<.0001

In addition to the type 1 and 3 tests, the F test for no fixed value measures the FE ability. The F-test reflects the EDA thus far. Time does not pass the F-test for fixed effects, but this is not surprise given the results thus far in the EDA. The F-test reflects that ID and the Joint T&I model are better than the pooled model.

F Test for No Fixed Effects ID			
Num DF	**Den DF**	**F Value**	**Pr > F**
5	81	57.73	<.0001

F Test for No Fixed EffectsTime			
Num DF	**Den DF**	**F Value**	**Pr > F**
14	72	1.17	0.3178

F Test for No Fixed EffectsT&I			
Num DF	**Den DF**	**F Value**	**Pr > F**
19	67	23.10	<.0001

From the airlines data set, FE modeling has been demonstrated, coefficients have been interpreted, and different subject variables have been utilized. FE modeling was appropriate

since the subjects were parametric shifts in the model. It is my opinion that the ID FE model is solid statistically as well as explanatory.

Random Effects (RE) Model

FE modeling has quite a few benefits, but one of the major weaknesses is the assumption that the subjects represent the entire population. From my experience, one rarely is able to capture an entire population, thus the conclusions drawn from FE can only be inferred on the data studied. RE modeling assumes random distribution based on the differences from the subjects. This assumption broadens the scope drawn from the model to include larger populations than just the studied data. The caveat for this model is the assumption that the unobserved heterogeneity is independently distributed from the subjects, which is very hard to satisfy in reality. In essence, the RE model is very similar to the pooled model except that an analysis is conducted between the error terms. Based on the error term, one will discern whether or not to utilize the FE model or the pooled model. The error term is focused specifically on variance between subjects, assumes constant intercepts and slopes. The model for RE is displayed below and it is statistically significant and predictive. At this point, one needs to ascertain which method to utilize. Rather than show the output for both scenarios, Time and ID, let's focus on the tests that discern which modeling method should be used.

Parameter Estimates Ran One						
Variable	**DF**	**Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Label**
Intercept	1	9.709637	0.3521	27.58	<.0001	Intercept
LnQ	1	0.918714	0.0289	31.83	<.0001	
LnPF	1	0.417726	0.0147	28.38	<.0001	
LF	1	-1.06998	0.1959	-5.46	<.0001	LF

Fit Statistics			
SSE	0.2933	**DFE**	86
MSE	0.0034	**Root MSE**	0.0584
R-Square	0.9926		

Breusch and Pagan developed the Lagrange Multiplier test, which assesses multiple regressors. In my opinion, this test has a similar framework to White's test in that it regresses the squared residuals against independent variables testing for significant variations (MTSU.edu). This test is validated through a hypothesis test, and assumes ac hi-squared distribution with *k* degrees of freedom based on the observations in the model. The test statistic generated needs to cross the appropriate threshold given the degrees of freedom. I am testing to verify that the errors between subjects are equal to zero. I prefer this testing method for the airlines data. The generated results for the residuals are 1.3354. Using that value in the LM formula, one gets a value of 334.85, which is substantially higher than the needed chi-squared table of 3.84. From the LM test, it can be concluded that the null hypothesis should be rejected leading us to conclude that the RE model should be utilized. The Hausman Test validates whether or not to use FM or RM. This test is distributed as a chi-squared random variable. The results show that the null hypothesis cannot be rejected, which is interpreted as the unobserved heterogeneity subject-specific effects are not correlated with any of the explanatory variables. Thus, both FE and ME are consistent indicators, but given the LM test, the RE model is preferred.

Hausman Test for Random Effects

DF	m Value	Pr > m
3	0.01	0.9999

Parameter Estimates Ran 2

Variable	DF	Estimate	Standard Error	t Value	Pr > |t|	Label
Intercept	1	9.362676	0.2440	38.38	<.0001	Intercept
LnQ	1	0.866448	0.0255	33.98	<.0001	
LnPF	1	0.43616				
LF	1	-0.9805				

Fit Statistics

SSE	0.2322	DFE	86
MSE	0.0027	Root MSE	0.0520
R-Square	0.9829		

Given that both models, ran one and ran two are strong models, this RE modeling is preferred over the FE modeling.

Pooled regression has many benefits for better understanding the relationship with panel

data. This study highlights the tip of the iceberg when it comes utilizing different modeling techniques.

Future Work

Further recommendations on how this study can be improved upon are the following:

- It was rather unclear as to the origins of the data. Logically, I could not ascertain if this was a sample population or a completed population, which made it rather difficult to process FE and RE modeling.
- Expanding on the actual years would allow for additional variables to be added into the model. For example, economic data would be helpful if the year was known.
- Additional information on over fitting models would be helpful as new modeling techniques are explored.
- After a model is found to be adequate, it would be helpful to review the validation techniques for goodness-of-fit.

Through this initial EDA, coupled with the future work recommendations, total cost can be reduced by focusing on maximizing value on specific variable outputs.

References

Borenstein, Severin. Phone interview. 16 Dec. 2011.

Ajmani, Vivek B.. *Applied Econometrics Using the SAS System*. Hoboken, NJ: Wiley, 2008. Print.

Borenstein, Severin. *Why Airlines Keep Going Bankrupt*. Washington DC: Interview - NPR, 2011. Print.

Chatterjee, Samprit, and Ali S. Hadi. *Regression Analysis by Example*. Fifth ed. Hoboken, New Jersey: Wiley, 2012. Print.

Cody, Ronald P.. *SAS Statistics by Example*. Cary, N.C.: SAS Pub., 2011. Print.

Greene, William H.. *Econometric Analysis*. 7th ed. Upper Saddle River, N.J.: Prentice Hall, 2012. Print.

Kenney, Caitlin. "Why Airlines Keep Going Bankrupt : Planet Money : NPR." *NPR : National Public Radio : News & Analysis, World, US, Music & Arts : NPR*. N.p., n.d. Web. 8 Jan. 2013. <http://www.npr.org/blogs/money/2011/12/16/143765367/why-airlines-keep-going-bankrupt>.

"Multicollinearity." *Statistics Solutions*. N.p., n.d. Web. 12 Jan. 2013. <http://www.statisticssolutions.com/resources/dissertation-resources/data-entry-and-management/multicollinearity>.

Najda, Charles. "Low-Cost Carriers and Low Fares: Competition and Concentration in the U.S. Airline Industry." *Stanford University Theses* 1 (2003): 9. *Department of Economics Stanford University*. Web. 8 Jan. 2013.

Ratner, Bruce. *Statistical and Machine-Learning Data Mining* . 2nd ed. Boca Raton, FL: CRC Press, 2012. Print.