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.

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. 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. increases the sample size and includes unobserved heterogeneity effects in the model. uld be unwise to conduct

Fixed effects (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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| LSDV Coefficients and Intercept | | | | |
| **Para** | **Estimate** | **SE** | **t Value** | **Pr > |t|** |
| **Inter** | 9.793 | 0.264 | 37.14 | <.0001 |
| **I 1** | -0.087 | 0.084 | -1.03 | 0.3042 |
| **I 2** | -0.128 | 0.076 | -1.69 | 0.0941 |
| **I 3** | -0.296 | 0.050 | -5.92 | <.0001 |
| **I 4** | 0.097 | 0.033 | 2.95 | 0.0041 |
| **I 5** | -0.063 | 0.024 | -2.64 | 0.01 |
| **I 6** | 0.000 | . | . | . |
| **LnQ** | 0.919 | 0.030 | 30.76 | <.0001 |
| **LnPF** | 0.417 | 0.015 | 27.47 | <.0001 |
| **LF** | -1.070 | 0.202 | -5.31 | <.0001 |
| Pooled Model | | | | |
| **Inter** | 9.517 | 0.229 | 41.51 | <.0001 |
| **log\_q** | 0.883 | 0.013 | 66.60 | <.0001 |
| **log\_pf** | 0.454 | 0.020 | 22.36 | <.0001 |
| **LF** | -1.628 | 0.345 | -4.71 | <.0001 |

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 below. 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. But, 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 6 headed snake with one body. If one were to zoom in on the first plot, there would be 7 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 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 |

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 6 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.