Linear Mixed Effect Model for Air Flight Fare Data

Jamel Thomas, Spencer Bassett, Yiming Chen

San Diego State University

December 11, 2017

Abstract

In this paper, we obtain a longitudinal linear mixed effects model for the average airfare price of an airline in the United States. We take a randomly selected subset of Routes and consider this covariate to be a random effect. The random slope effect for time, Quarter, was found not to be statistically significant, and we exclude it from the final model. We also consider interactions and they are found to not be practically significant. The chosen model obtains the smallest residual error of all models tested. All covariates are significant.

I. Introduction

In the United States, various factors affect the airfare prices that consumers see. An issue is that a single linear model may not fit the data, as prices can vary drastically across different routes. In this paper, we obtain Quarterly US Air Fare and Volume data from Professor Larry Winner's miscellaneous datasets. The original dataset contains 108,602 observations and 14 attributes and there are no missing values. For the sake of simplicity, in this paper, we only consider a random sample of 50 routes as the original dataset contains 4177 routes. Moreover, for every route, there are 26 quarters for which the average fare price is taken. Therefore, we simplify the problem dramatically with the exclusion of most of the routes. Our objective is to build a linear mixed effects model to predict the average fare price for each route.

II. Methods

It is unknown how the data was collected, but it is assumed that the data is obtained from one airline company. The response variable is the Average Fare. The covariates include the route number, the quarter, average weekly passengers, departing airport, arrival airport, the population of the departing city, the population of the arrival city, the distance traveled, the angle of the route, and the longitude and latitude of both departing and arriving airports. As stated in the introduction, we only took a random sample of 50 routes. Table 1 shows the summary statistics for the subsetted dataset.

In Table 1 we see the mean, standard deviation, minimum and maximum of the continuous variables. We immediately notice that the response variable, Average Fare, is skewed right. The mean fare over the 50 randomly selected routes is about $213.64. The standard deviation is about $70, with a minimum of $57 and maximum of $517.

Variable	Mean	SD	Min	Max
Average Fare	213.64	69.25	57.94	517.63
Average Passenger	174.86	272.74	11.97	1562.17
Population Departing (Mil)	2.83	4.17	.0148	19.87
Poulation Arriving (Mil)	26.34	3.80	.0182	19.88
Distance	10345.70	621.55	67.00	2687.00
Angle	5.29	39.63	-89.80	89.70

Table 1: *Summary statistics for continuous variables*

In this paper, we considered a linear mixed effects model with a random intercept route term for the randomly selected routes. This is fitted on log-transformation of average fares and we consider all covariates without interaction. In the appendix, we justify using a log-transformation rather than a box-cox transformation of the data. The essence is that there is not much gained from the box-cox transformation over the logarithm. Moreover, the Normal Q-Q plot for the log-transformed response appears to have about the same skewness as the Box-Cox.

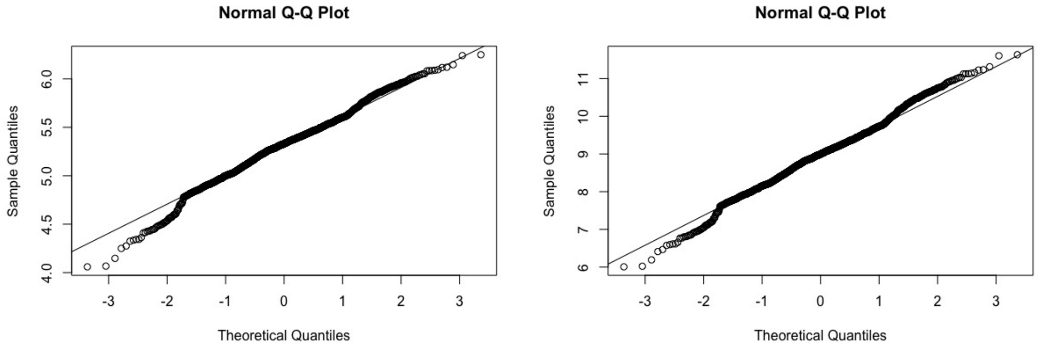

Figure 1: *A justification on the log-transformation over the box-cox. The left is the log-transformation and the right is a Box-Cox transformation.*

We attempted to aggregate the data by combining all common quarters together. This means that the data was divided into 4 quarters in terms of the business quarters for each year. The leftover 2 quarters were included in this dataset. When we attempted to fit a model to this data it did not yield better results than the original time-scale. We did the same thing with a yearly time-scale but were met with the same results.

III. Results

Figure 2 indicates the change in distribution for the Average-Fare price before and after we take a logarithm. All of the following models are built on the log-scale. Moreover, the correlation plot seen in the appendix indicates little to no correlation among all covariates.

One thing to note is that the relationship between average passenger and average fare is not perfectly linear. This nonlinearity is enhanced when we consider the log-scale for the response

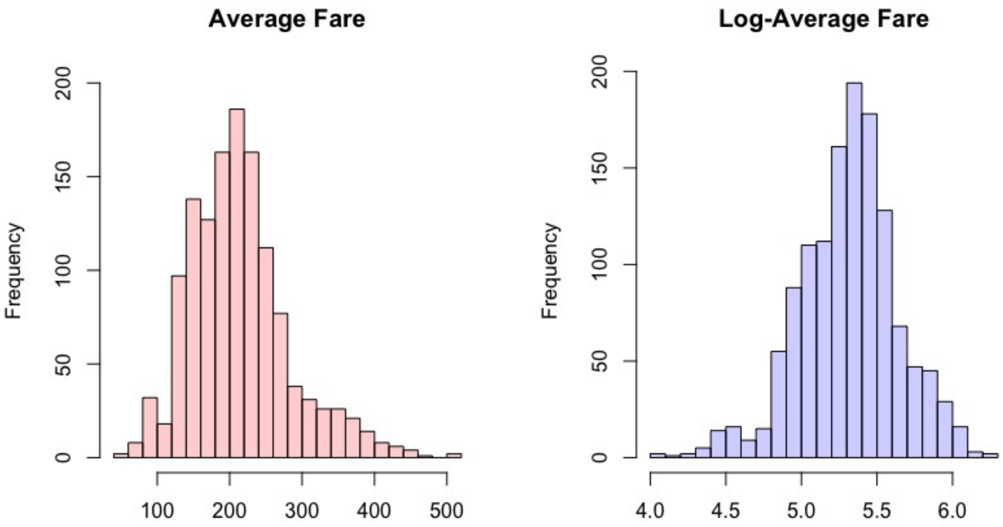

Figure 2: *Histograms of the original response and log-response*

variable. In order to coerce a more linear relationship between these two covariates, we consider the log-transformation of average passenger. This relationship is seen in the appendix, and has a better linear relationship than before.

Figure 3 indicates the average fare price over all 26 quarters. These fair prices are partitioned out by the random route number. We can see that there is not a single pattern of linearity among all routes. The average fare prices can vary widely depending on the route taken. However, the overall trend of each route tends to follow a linear trend. Therefore, we will consider a random route effect on the linear mixed effects model. Moreover, there appears to be a quadratic trend among the covariates to the response variable.

Building this model, we considered a random slope for routes. When the analysis was run the random slope was found to be statistically significant but it was not practically significant. Therefore, the random slope was dropped from the model. We also considered a quadratic term for average passengers. This was also found to be statistically significant but not practically significant. Therefore, it was excluded from the model.

The final model contains the covariates quarter, log-average passenger, quarter squared and a random intercept effect for route.

$$log(Avg.Fare)_{ij} = \beta_0 + \beta_1 Quarter + \beta_2 Quarter^2 + \beta_3 log(Avg.Passenger) + u_{ij} + \epsilon_{ij}$$

Where $u_{ij} \sim N(0, \sigma^2_{route})$ and $\epsilon_{ij} \sim N(0, \sigma^2)$. The ϵ_{ij} are normally distributed random errors and u are the route random effect. $i = 1, \ldots, 26, j = 1, \ldots, 50$.

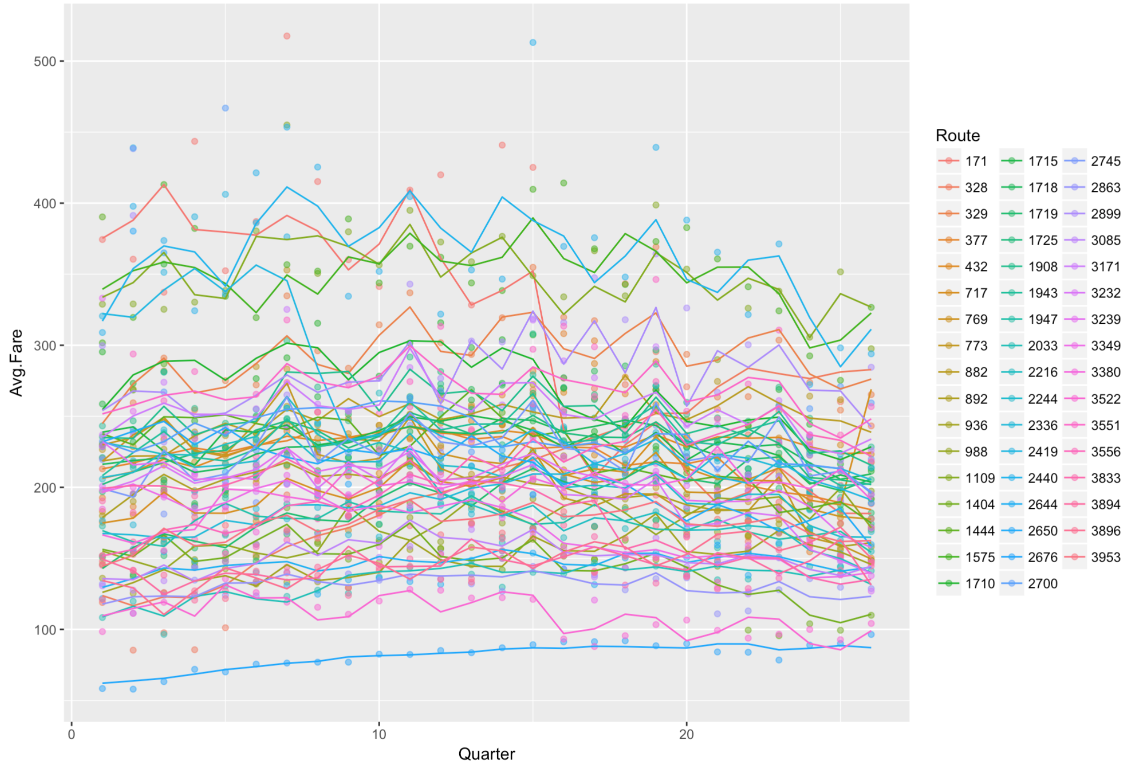

Figure 3: *This figure indicates the model predictions for Average Fare over 26 Quarters seperated out by route.*

Figure 3 shows the model fits with the random intercept Route. The random slope effect was tested and was found to not be statistically significant. The fit overall obtains a Q-Q residual plot that mostly follows the normal quantile assumption. This plot can be found in the appendix.

This model yields final predicitons of the form

$$Avg.Fare_{ij} = EXP\{\beta_0 + \beta_1 Quarter + \beta_2 Quarter^2 + \beta_3 log(Avg.Passenger) + u_{ij} + \epsilon_{ij}\}$$

for the same domains for i and j.

Table 2 indicates the model point estimates, and they are all significant. Due to the fact that we are only interested in the accuracy of the model, we build the model on the log-scale in order to improve the normality of the response. The table is still on the log scale.

	Value	SE	DF	P-Value	C.I
$(Intercept)$	6.52	0.079	1247	0	(6.36, 6.67)
$Quarter$	0.02	0.0019	1247	0	(0.02, 0.03)
$Quarter^2$	-0.001	0.0000	1247	0	(-0.001, -0.001)
$log(Avg.Pass)$	-0.30	0.012	1247	0	(-0.33, -0.28)
$SD(Route)$	0.408	-	-	0	(0.33, 0.50)
$Residuals$	0.125	-	-	0	(0.12, 0.13)

Table 2: *This table assesses the significance of the coefficients and gives their confidence intervals on the log scale.*

IV. Conclusion

In conclusion, we obtain a model with linear mixed effects on Route. We began with exploratory analysis of the response variable and found the best, simple, model has a quadratic term of Quarter and log-Average Passengers. For every one unit increase in Quarter, we obtain about a 2% increase in Average Fare, assuming all other covariates are fixed. Moreover, For every one unit increase in Average Passengers per week, we obtain about a 25.9% decrease in Average Fare, assuming all other covariates are fixed. This may be because an increase in passengers allows the airline to lower fare prices. The standard deviation is nearly 0.4 on the log-scale, with residuals around 0.12. None of the confidence intervals contain zero, which verifies their significance in the model.

The model may not be the best fit, as it obtains an MSE for observed vs predicted of 792.5. There may be a better model for this dataset, however, the simple model is the most convenient. To improve efficiency more covariates may be needed. Finally, since the model was built on a subset based on routes and it is not perfect, the model does not generalize well to the entire dataset.