Flight Delays and Causes

Michael Hendrix

Department of Mathematics and Statistics
Coastal Carolina University

Introduction

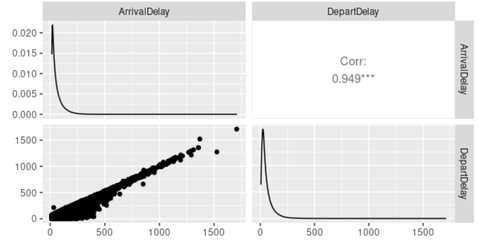
HAS your flight ever been late or delayed, sitting in the plane waiting for the flight to take off can be so painful. So I wanted to try and predict how long a typical flight will be delayed. In this study I used multiple linear regression to predict arrival delays for flights based on multiple variables. In this set there are 20 variables that create some type of delay like weather, time of day, security reasons and many more. However, not all of them will be used because of variable selection. Variable selection is the process of discarding of variables that are either redundant or non-significant. The quantitative variables are:

- Local Depart Time
- Scheduled Arrival Time
- Estimated Total Flight Time
- Time In Air
- Distance
- Taxi Out
- Weather Delay Minutes
- Security Delay Minutes
- Local Arrival Time
- Total Flight Time
- Air Time
- Departure Delay
- Taxi In
- Carrier Delay Minutes
- NAS Delay Minutes
- Late Aircraft Delay Minutes

Data Processing

DATA processing is a series of steps that helps us understand relationships between variables, identify and remove missing observations, transforms the data to normality, and may reduce some of the variables considered for the model. For our project the steps included:

1. **Histograms** give us an idea of the distribution of the variables.We can get a sense if the variables are normally distributed or not.
2. **Correlation Matrix** allows us to see how well a predictor predicts the response variable. In this part I reduced our number of predictors by only keeping those with a correlation either below -0.1 or above 0.1. An example of the correlation matrix can be seen below:

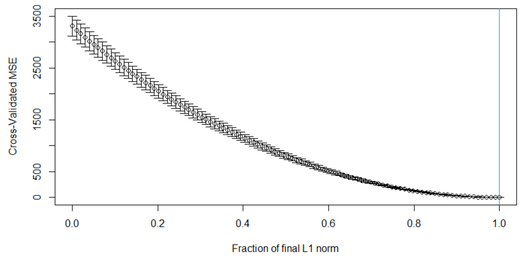


3. **Transforming data** towards Normality helps us meet the assumptions of linear regression. I did this through an R function `powerTransform()` which outputs an exponent to transform the data with. Ithen fit simple linear regression models with each of the transformed predictor variables.
4. **VIF** stands for Variance Inflation Factor which measures how much a predictor can be modeled by the others. After performing our tests the DepartDelay was the variable with the highest VIF of 26, thus it was discarded.
5. **Transformed Model Diagnostics** before continuing with model selection ensures assumptions for multiple linear regression are met. I constructed graphs that show the residuals vs fitted points, normal Q-Q plot, scale-location, and residuals vs leverage.

Methods

IN addition to the full model I will also use LASSO, AIC, and BIC to select other models. I used R Studio to perform all of our analysis and make all of our graphs and models. Due to limitations of processing power, I took a random sample of 10,000 from 484,551.

1. **Full Linear Model:** Our final linear model consists of the variables not discarded through data processing.
2. **LASSO Selection and Model:** LASSO selection is a variable selection method that uses regularized regression through cross validation between the variables' Mean Squared Error.



3. **AIC:** AIC stands for Akaike Information Criterion. This is a variable selection method that penalizes models for excess terms in the variables in the model.

$$AIC=2K-2\ln(L)$$

4. **BIC:** BIC stands for Bayesian Information Criterion. Much like AIC this variable selection method penalizes models for excess terms in the model, but the penalty is stronger than AIC due to the K value.

$$BIC=K\ln(n)-2\ln(L)$$

Results

TO determine the best performing model Iused an independent test data set of 3,000 observations and computed R^2_{adj} and root means square error (RSME) values of each model.

Model	Full	LASSO	AIC	BIC
RMSE	0.2657	0.2660	0.2657	0.2656
R^2_{adj}	0.8644	0.8642	0.8644	0.8645

Model	Variables Selected
Full	LocalDepartTime, DepartDelay, LocalDepartTime, TaxiIn, TaxiOut, CarrierDelay, WeatherDelay, NASDelay, LateAircradtDelay, TimeInAir
LASSO	DepartDelay, TaxiIn, TaxiOut, CarrierDelay, WeatherDelay, NASDelay, LateAircradtDelay
AIC	LocalDepartTime, DepartDelay, LocalDepartTime, TaxiIn, TaxiOut, CarrierDelay, WeatherDelay, NASDelay, LateAircradtDelay, TimeInAir
BIC	LocalDepartTime, DepartDelay, LocalDepartTime, TaxiIn, TaxiOut, CarrierDelay, WeatherDelay, NASDelay, LateAircradtDelay

Discussion

MY goal was to fit the best model to predict ArrivalDelay. However what is best? Is it the model with the lowest RMSE, most predictors? Well it's a little of both. The rule of parsimony says that if some models tell basically the same information, then choose the one with less terms as to not over complicate things. Choosing between the models, there really isn't much of a difference between the RMSE and R^2_{adj}, George Box describes this situation well with his famous quote: "All models are wrong, but some are useful". When comparing the models, LASSO does have the lowest R^2_{adj} and the full and AIC have the highest R^2_{adj}. However it really isn't too big to make a difference due the the observations and the small differences in the RMSE and R^2_{adj} values. All of these models are fine, it comes down to cost and what the customer wants in the model.

References

[1] RStudio Team (2022). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

[2] Sheather, Simon. Modern Approach to Regression with R. Springer Nature, 2009.

[3] Trivedi, Pawan. "Flight Delay and Causes." Kaggle, May 1, 2021. https://www.kaggle.com/datasets/undersc0re/flight-delay-and-causes.