Benjamin Straus (bstraus7)
Prediction of Airline Safety
BDS Project

Introduction/Problem

How active is an airline? Is safety an indicator of activity? The goal of this analysis is to see whether we can predict the number of seat kilometers per week flown by airlines (i.e. number of seats times kilometers flown per week) by the amount of incidents of two time spans and fatal accidents of an airline. The reason for this guiding question is that *if* we can predict this reliably, then we can answer whether *safety of the airline is an indicator of how active an airline is.*

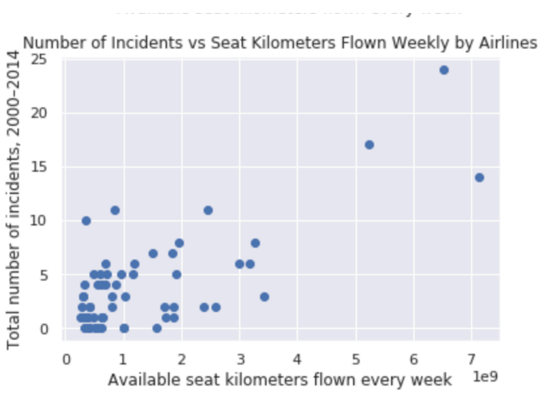


Figure: Example of Basic Data Plot

Methods (Data & Model Fitting)

This analysis was done using Google Collaboratory and Python. Statsmodel was used to run the ordinary least squares multiple linear regression, Pandas was used to import the data, and Numpy was used for data manipulation.

Data was provided by FiveThirtyEight.com (see sources) and was a table of airlines data including incident and accidents counts over different time spans and number of seat km flown per week. Upon basic plotting of this data, we see data that seems like linear fits would work well (see one plot of such data in the figure to the right).

The data fitting used was a multivariable linear model. Performance metrics used were the r-squared value and a comparison between training and testing data set's mean-squared-error (of predictions vs known seat km flown per week).

Avoiding overfitting was done in two ways. First, overfitting with linear regression is a lot less likely than with neural networks as the number of layers drastically increases the number of coefficients. So, linear regression with 3 regressors (and an intercept) is a relatively safe number for around 60 rows of data. Second, the mean-square-error was compared between the training and testing data sets. It was found to be very similar (within 6%) meaning that the predictions for the test set were just as good as those for the training set.

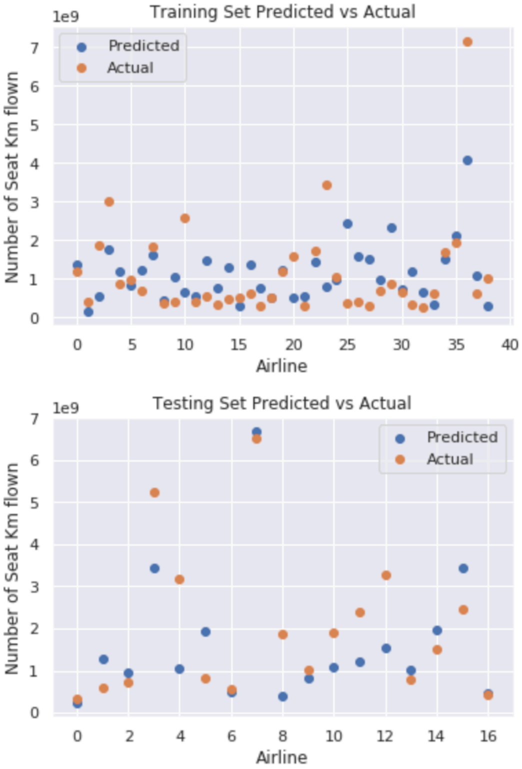


Figure: Plots of Predictions and Actual Values

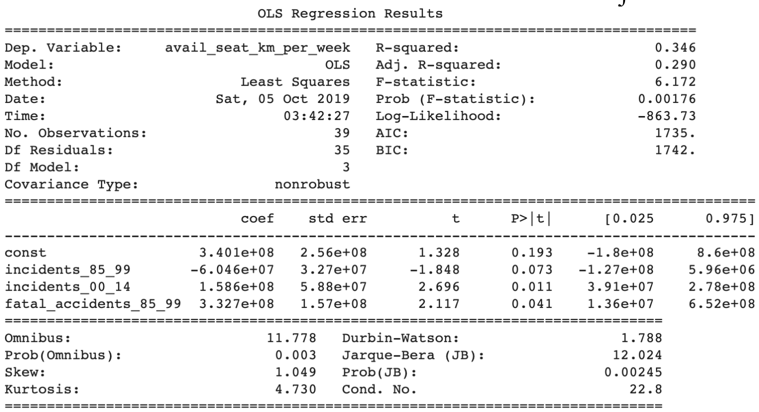
Interpretation of Results & Covariate Coefficients

Overall, this study is inconclusive as to whether fatal accidents, number of incidents between 1985-1999, and number of incidents between 2000-2014 are a reliable predictor of seat km flown per week. This is because the mean-square-error of both the test predictions and the training predictions was high (on the order of 10^18) and the r-squared value rather low (0.346). As well, the plot (to the right) shows that while some predictions were correct, others were far off. So, we cannot say for certain that there exists correlation and would

	Training Data	Test Data
Mean Squared Error	$1.0096*10^{18}$ (Seat km/week)^2	$1.0643*10^{18}$ (Seat km/week)^2

need more data or regressions to say for sure (see Future Improvements section for suggested improvements).

However, we note that the mean-squared-error of the testing and training datasets is very comparable (within 6%). This means that the model predicted test and training results equally well. So, while the model may not very well predict seat km flown per week, the fitting method and amount of regressors used were good as it doesn't seem that the data have been overfit.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:     avail_seat_km_per_week   R-squared:                       0.346
Model:                                OLS   Adj. R-squared:                  0.290
Method:                     Least Squares   F-statistic:                     6.172
Date:                    Sat, 05 Oct 2019   Prob (F-statistic):            0.00176
Time:                            03:42:27   Log-Likelihood:                -863.73
No. Observations:                      39   AIC:                             1735.
Df Residuals:                          35   BIC:                             1742.
Df Model:                               3
Covariance Type:                nonrobust
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
const                3.401e+08   2.56e+08      1.328      0.193    -1.8e+08     8.6e+08
incidents_85_99     -6.046e+07   3.27e+07     -1.848      0.073   -1.27e+08    5.96e+06
incidents_00_14      1.586e+08   5.88e+07      2.696      0.011    3.91e+07    2.78e+08
fatal_accidents_85_99 3.327e+08   1.57e+08      2.117      0.041    1.36e+07    6.52e+08
==============================================================================
Omnibus:                       11.778   Durbin-Watson:                   1.788
Prob(Omnibus):                  0.003   Jarque-Bera (JB):               12.024
Skew:                           1.049   Prob(JB):                      0.00245
Kurtosis:                       4.730   Cond. No.                         22.8
==============================================================================
```

Figure: OLS Regression Results

Despite the inaccuracy of the model, the coefficients as the result of the linear regression can still be interpreted:

- For every additional incident between 1985 and 1999, the estimated number of seat kilometers flown per week decreases by 60,000,000 with all other variables held constant.
- For every additional incident between 1999 and 2014, the estimated number of seat kilometers flown per week increases by 159,000,000 with all other variables held constant.
- For every additional fatal accident between 1985 and 1999, the estimated number of seat kilometers flown per week increases by 333,000,000 with all other variables held constant.
- The intercept represents the expected number of seat kilometers flown per week when an airline had 0 fatal accidents between 1985-1999, 0 incidents between 1985-1999, and 0 incidents between 2000-2014. So, in this case, when all regressors are 0, we expect the number of seat km flown per week to be 340,000,000.
- The r-squared value represents the proportion of variation in the data that is explained by the model. In this case, only a small portion (0.346) of variation in the data was explained by the model.

In summary, though we cannot conclude that incidents and accidents well predict the activity of an airline, we can say that linear regression created a model that predicted both the test and training data with similar loss. So, the method clearly works and the right number of regressors was chosen as to not overfit the data. But, if we are to interpret our results despite the high error using our model coefficients, we can say that more incidents between 1999 and 2014 and more fatal accidents between 1985 and 1999 will increase the predicted number of seat km flown by an airline, while number of incidents between 1985 and 1999 will decrease it.

Future Improvements

Though this study demonstrated the use of linear regression in data analysis, the data set chosen was very small (only 56 rows). Future studies should also obtain more data so that neural networks could be fit (without risk of overfitting) and the MSE could be lowered (i.e. performance increased). Examples of more data could be simply data about more airlines or could be additional data about the given airlines like the revenue of the airline, the number of airplane meals served, or the amount paid to captains by an airline.

Sources

Data was provided by the FiveThirtyEight.com Github Repository