Flight Ticket Price Analysis with Linear Regression

Jun Cai, Vikas Boddu

February 11, 2016

This document describes an implementation of Mapreduce instance with Hadoop doing ticket price analysis on the OTP (On-Time Performance) dataset. The purpose of the application is to perform a simple linear regression on the ticket price. The features we used are distance and air time. The label is average ticket price. For each carrier whic is active in 2015, we finished two simple linear regression for it. One between distance and average ticket price. The other one between air time and average ticket price.

Implementation

Data Normalization

In order to get a good regression results, the features should be normalized. We normalized the data so that they have zero mean and one variance. A R script is used to go over all the data and find out the mean and standard deviation of the distance as well as the air time. Then these information is provided to Mapper along with the dataset. Mapper will normalize the features in map phrase.

Model and Normal Equation

For a simple linear regression, there is only one explanatory variable which in our case is either distance or air time. So the model will simply be

$$y = \theta_0 + \theta_1 x$$

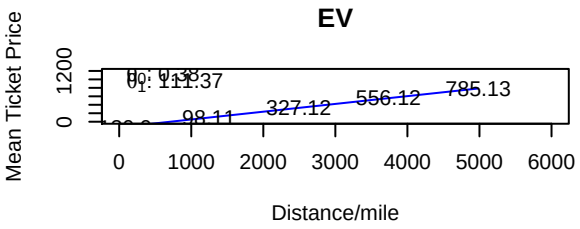
where y is the average ticket price and x is distance or air time of the flight. θ_0 and $\theta 1$ are two model parameters which can be estimated from linear regression. In this implementation, we use normal equation to solve the linear regression problem.

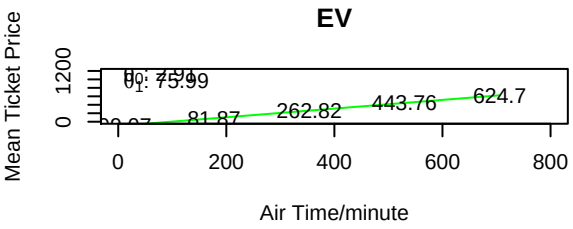
$$\theta = (X^T X)^{-1} X^T Y$$

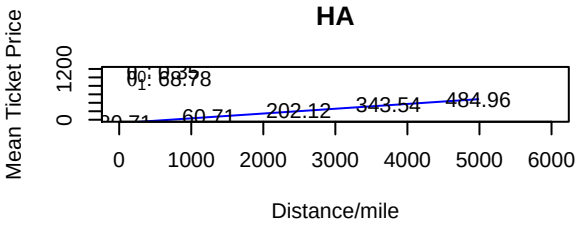
In order to calculate the normal equation in Mapreduce, we divide the computation into three steps: 1. Compute $X_i^T X_i$ and $X_i^T Y_i$ for each observation and combine the results for each carrier in the mapper. 2. Compute the actual $X^T X$ and $X^T Y$ in the reducer by summing up all $X_i^T X_i$ and $X_i^T Y_i$ for each carrier. 3. Compute $\theta = (X^T X)^{-1} X^T Y$ for each carrier.

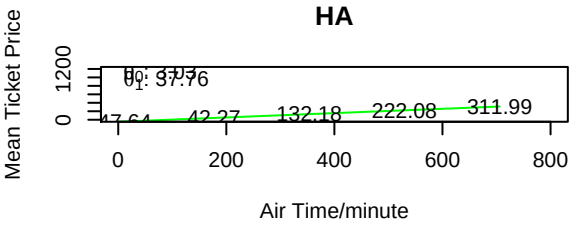
Regression Results

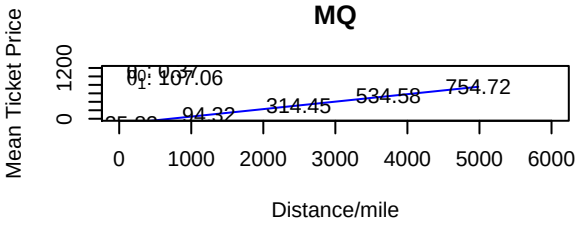
Following is the regression results created by R.

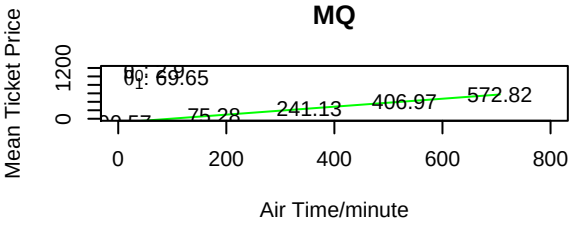

EV
Mean Ticket Price
0
1200
0: 111.37
1: 98.11
327.12
556.12
785.13
0
1000
2000
3000
4000
5000
6000
Distance/mile

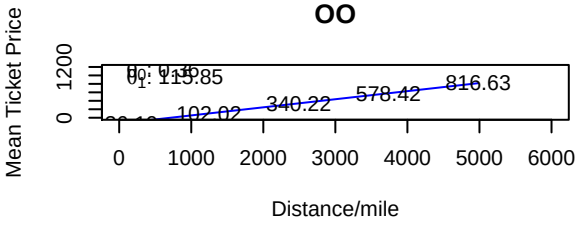

EV
Mean Ticket Price
0
1200
0: 75.99
81.87
262.82
443.76
624.7
0
200
400
600
800
Air Time/minute

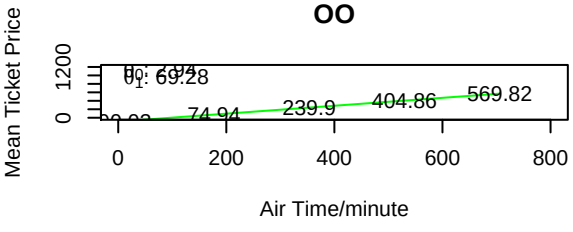

HA
Mean Ticket Price
0
1200
0: 60.78
60.71
202.12
343.54
484.96
0
1000
2000
3000
4000
5000
6000
Distance/mile

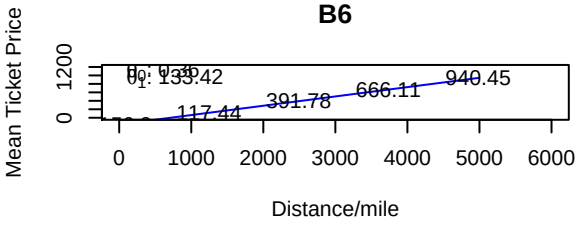

HA
Mean Ticket Price
0
1200
0: 37.76
42.27
132.18
222.08
311.99
0
200
400
600
800
Air Time/minute

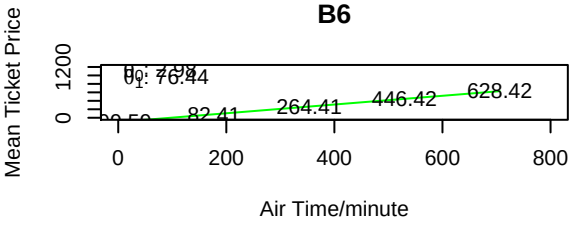

MQ
Mean Ticket Price
0
1200
0: 107.06
94.32
314.45
534.58
754.72
0
1000
2000
3000
4000
5000
6000
Distance/mile

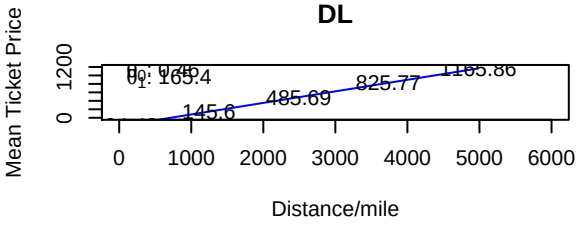

MQ
Mean Ticket Price
0
1200
0: 69.65
75.28
241.13
406.97
572.82
0
200
400
600
800
Air Time/minute

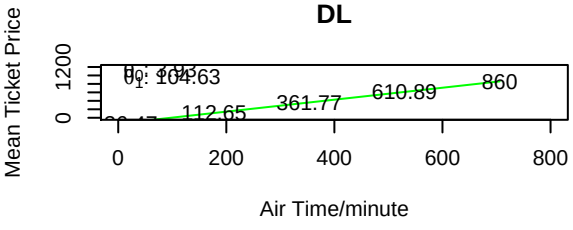

OO
Mean Ticket Price
0
1200
0: 115.85
102.02
340.22
578.42
816.63
0
1000
2000
3000
4000
5000
6000
Distance/mile

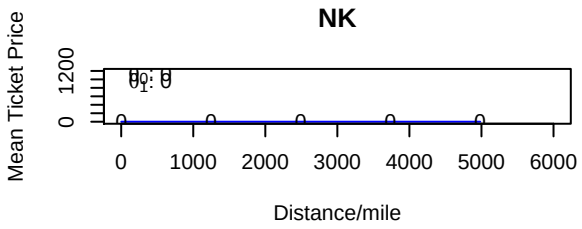

OO
Mean Ticket Price
0
1200
0: 69.28
74.94
239.9
404.86
569.82
0
200
400
600
800
Air Time/minute

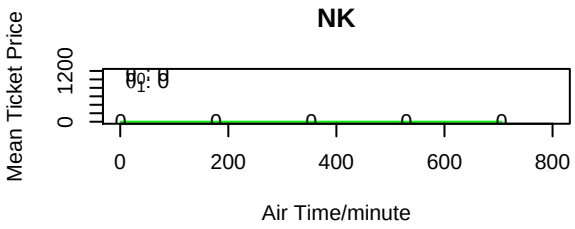

B6
Mean Ticket Price
0
1200
0: 133.42
117.44
391.78
666.11
940.45
0
1000
2000
3000
4000
5000
6000
Distance/mile

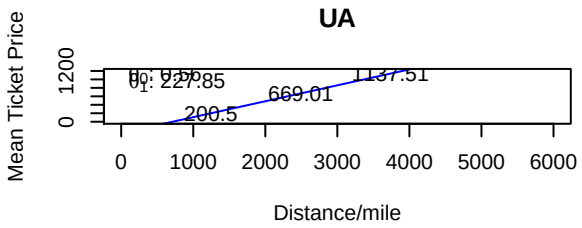

B6
Mean Ticket Price
0
1200
0: 76.44
82.41
264.41
446.42
628.42
0
200
400
600
800
Air Time/minute

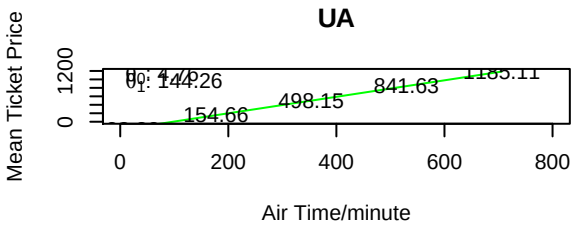

DL
Mean Ticket Price
0
1200
0: 165.4
145.6
485.69
825.77
1165.86
0
1000
2000
3000
4000
5000
6000
Distance/mile

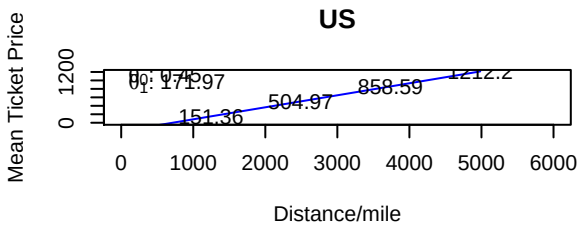

DL
Mean Ticket Price
0
1200
0: 104.63
112.65
361.77
610.89
860
0
200
400
600
800
Air Time/minute

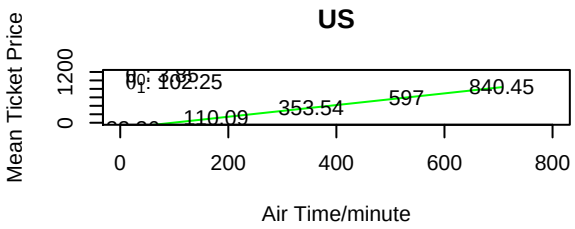

NK
Mean Ticket Price
0
1200
0
1000
2000
3000
4000
5000
6000
Distance/mile

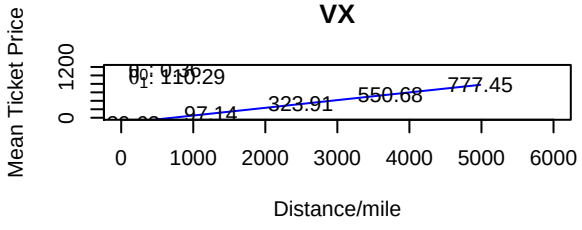

NK
Mean Ticket Price
0
1200
0
200
400
600
800
Air Time/minute

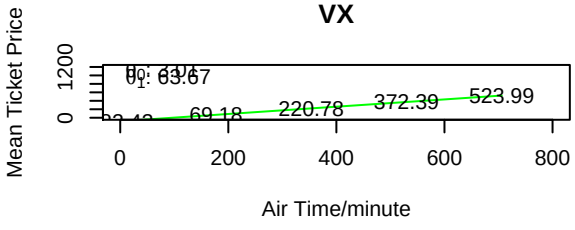

UA
Mean Ticket Price
0
1200
227.85
669.01
200.5
0
1000
2000
3000
4000
5000
6000
Distance/mile

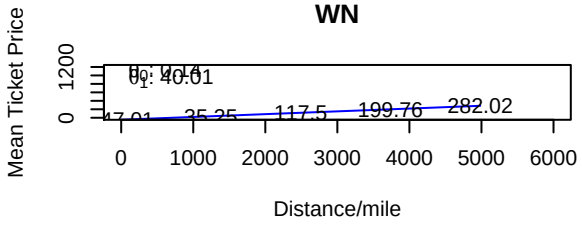

UA
Mean Ticket Price
0
1200
144.26
498.15
154.66
841.63
0
200
400
600
800
Air Time/minute

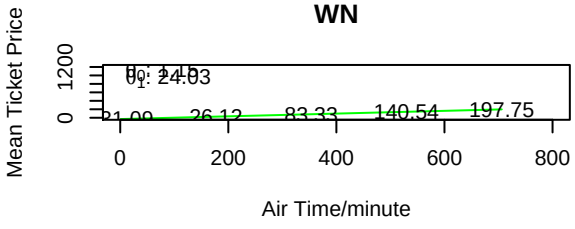

US
Mean Ticket Price
0
1200
171.97
504.97
151.36
858.59
0
1000
2000
3000
4000
5000
6000
Distance/mile

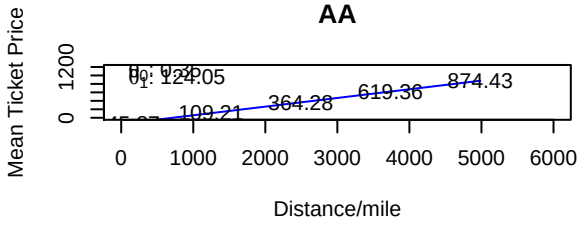

US
Mean Ticket Price
0
1200
102.25
353.54
110.09
597
840.45
0
200
400
600
800
Air Time/minute

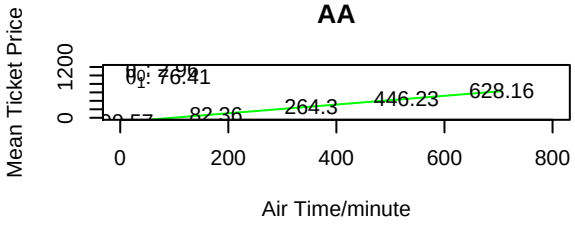

VX
Mean Ticket Price
0
1200
110.29
323.91
97.14
550.68
777.45
0
1000
2000
3000
4000
5000
6000
Distance/mile

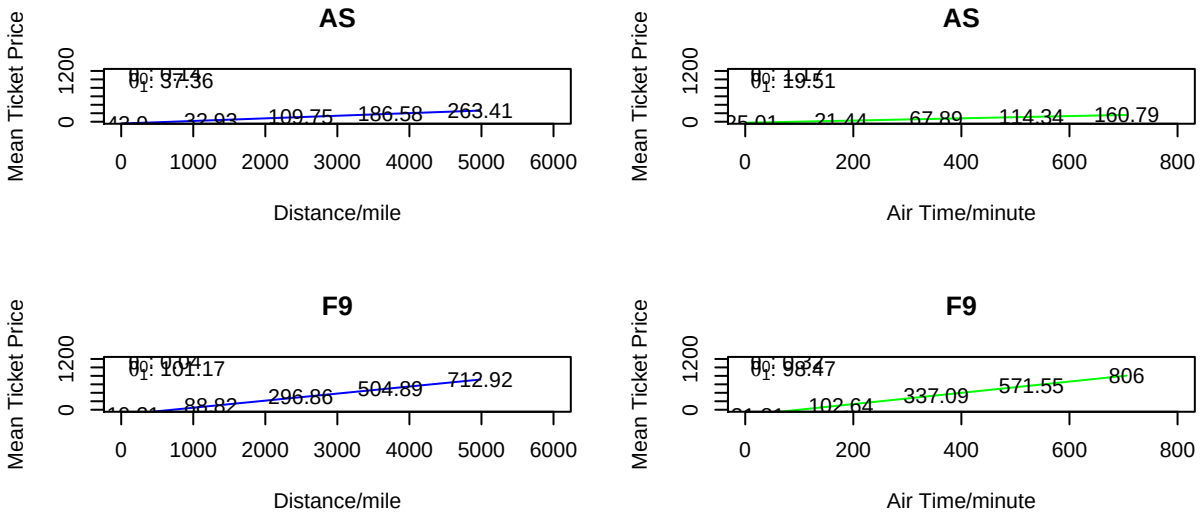

VX
Mean Ticket Price
0
1200
63.67
220.78
69.18
372.39
523.99
0
200
400
600
800
Air Time/minute


WN
Mean Ticket Price
0
1200
40.01
117.5
35.25
199.76
282.02
0
1000
2000
3000
4000
5000
6000
Distance/mile


WN
Mean Ticket Price
0
1200
24.03
83.33
26.12
140.54
197.75
0
200
400
600
800
Air Time/minute


AA
Mean Ticket Price
0
1200
124.05
364.28
109.21
619.36
874.43
0
1000
2000
3000
4000
5000
6000
Distance/mile


AA
Mean Ticket Price
0
1200
76.41
264.3
82.36
446.23
628.16
0
200
400
600
800
Air Time/minute



Where the title of each plot is the carrier's name. The left column is for the regression using distance as the feature and the right column is for that using air time. Model parameters for each regression are also given in the plots as θ_0 and θ_1.

Conclusions

Comparing the Features

We think the distance is a better feature than air time when predicting the ticket price. From all the coefficients above, we can see that θ_1 is always much larger then the θ_0 which means the features always contribute much more to the result rather than the contstant does. The other observation is that, the θ_1 for distance is always larger than θ_1 for air time model; and the θ_0 for distance model is always smaller than θ_0 for the air time model from which we conclude that distance is a better explanatory varialble here.

Cheapest Carrier

We think the cheapest carrier is AS since it has the smallest model parameters among all the carriers.