**Data Science Math Final, CUNY Summer 2016 Bridge Program**

**Todd Weigel**

**8/11/16**

## Summary

I have used the included hFlights package to do the various data and statistical analyses required for the project. Types of analysis done, include probability, descriptive/inferential statistics, correlation and linear algebra, and calculus based probability and statistics. The hFlights package is a large (>200k) dataset containing flight arrival and departure times, inbound/outbound delays, flight numbers, etc., for domestic flights that departed Houston TX in 2011. In the analysis below, I will concern myself with only the arrival and departure delays.

Note, the dataset as a whole has approximately 227k rows, but there are a number of rows with missing data (i.e., NA’s) for some rows in either or both the arrival and departure delays column. These rows have been excluded from any analysis, so we actually have about 223k rows to work with.

## Probability

The first analysis undertaken was on probabilities of certain subsets of data related to the arrival and departure delays. The data subsets and variables were based on the following:

X is the departure delays (in minutes).

x is the 3rd quartile of X (i.e., departure delay values between 50% and 75%)

Y is the arrival delay in minutes.

y is the 2nd quartile of Y (i.e., arrive delay values between 25% and 50%)

1. P(X>x|Y>y) = .4804153, this is the probability of a flight having departed later than the 3rd quartile, **given** that the associated flight arrival time was later than the 2nd quartile.
2. P(X>x, Y>y) = .2294416, this is the probability of a flight both haven departed later than the 3rd quartile **and** having arrived later than the 2nd quartile.
3. P(X<x| Y>y) = .2061448, this is the probability of a flight having departed earlier than the 3rd quartile **given** that the associated arrival flight was later than the 2nd quartile.

The below chart gives total counts of flights for the year based upon the appropriate x and y quartiles. Also we are defining new variables, A and B. A is the flights that departed later than 3rd quartile (i.e., 109574) and B is the flights that arrived later than the 2nd quartile (i.e., 106920).

|  |  |  |  |
| --- | --- | --- | --- |
| x/y | <=2d quartile | >2d quartile (B) | Total |
| <=3d quartile | 113141 | 55554 | 168695 |
| >3d quartile (A) | 3813 | 51366 | 55179 |
| Total | 116954 | 106920 | 223874 |

Splitting the data in the above fashion helps to visual the data, but it does not make the values independent.

P(A|B) = 55179/106920 = .5160774

P(A) = 109574/223874 = .48944

P(B) = 106920/223874 = .47759

P(A)P(B) = .48944 \* .47759 = .2337541

One should note that P(A|B) P(A)P(B). This is shown in the above calculations, but intuitively it should be apparent as well, since they are describing two different things. P(A|B), says give me the chance of getting an “A” value from the “B” values, which were sub-setted. P(A)P(B), essentially says, give me the chance of both “A” and “B” being found in the aggregate set of data.

Running a ‘chisq.test’ in ‘R’ gives the following. Note the P value is very small so the correlation does exist.

> mABTots <- matrix(c(55554, 3813,3813,51366), nrow=2)

> chisq.test(mABTots)

Pearson's Chi-squared test with Yates' continuity correction

data: mABTots

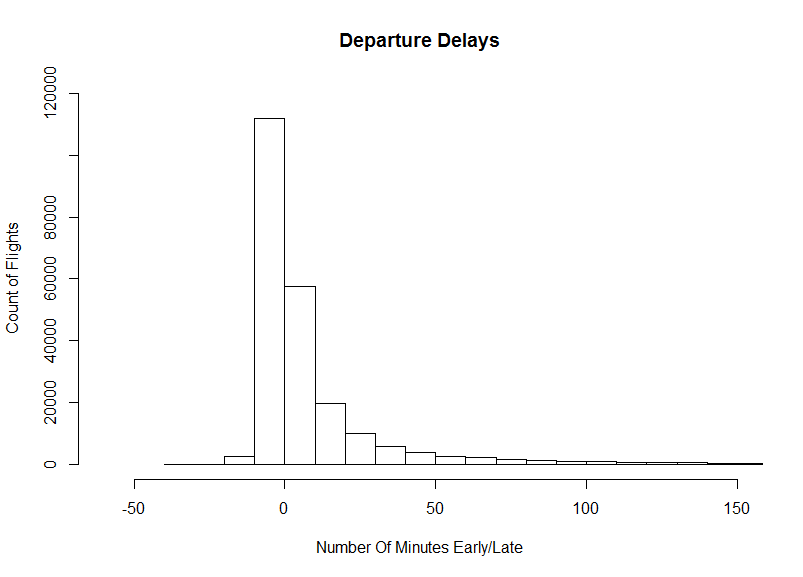
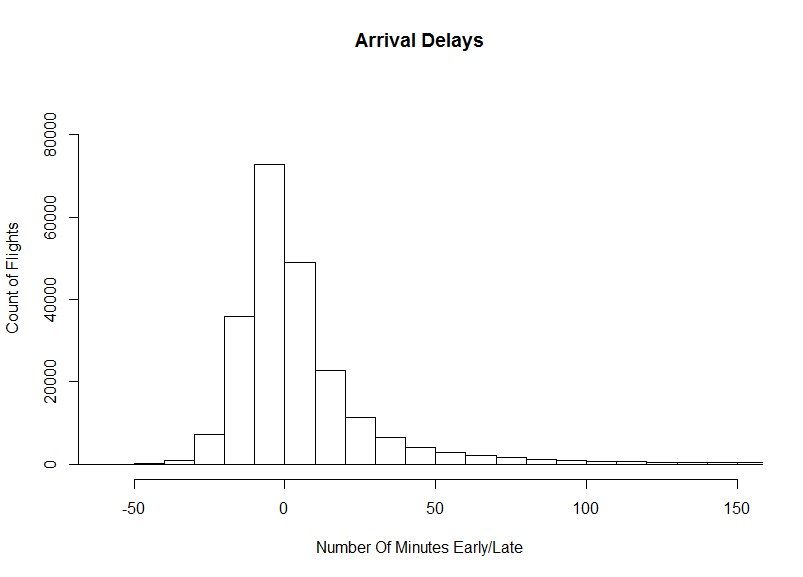
X-squared = 86034, df = 1, p-value < 2.2e-16

## Descriptive and Inferential Statistics

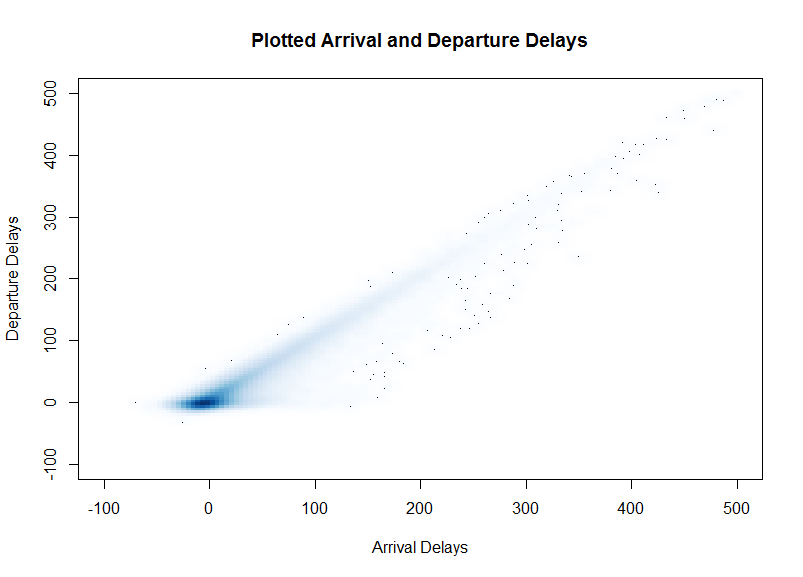
For the arrival and departure delay data, some basic univariate statistics are below:

|  |  |  |
| --- | --- | --- |
| Statistic | Arrival Delays | Departure Delays |
| Minimum Delay | -70 | -33 |
| Maximum Delay | 978 | 981 |
| Median | 0 | 0 |
| Mean | 7.09 | 9.42 |
| SD | 30.71 | 28.74 |
| Variance | 943.01 | 823.40 |

Below are some histograms of the arrival and departure delays. As can be noticed easily, the plots follow similar patterns, with most flights arriving or departing basically on time. Both charts show a right skewing of the data, with the arrival delays being slightly more pronounced then the departure delays, indicating delays within flight can occur. While both have a long right skew (the histograms actually don’t plot all the way to right, to make the concentrated data points more visible), it can be seen that after about 50 minutes, the number of delays is rather miniscule (for the number of flights as a whole).



Below is a scatterplot comparing Arrival and Departure Delays. Note the strong correlation, but also note arrival delays can be seen as more skewed to the right, just as we can observe with the two histograms.



The mean of arrival delays is 7.09 and departure delays is 9.42. A 95% confidence interval for each is as follows:

* 95% CI for Arrival Delays is: 6.967129 < < 7.22154
* 95% CI for Departure Delays is: 9.295924 < < 9.534043

For the difference of the mean, the value is 2.320649, and the 95% confidence interval is:

A correlation matrix for the arrival and departure delays is:

* 95% CI for Arrival Delays is: 2.28871 < < 2.352588

|  |  |  |
| --- | --- | --- |
| Correlation Matrix | Arrival Delay | Departure Delay |
| Arrival Delay | 1.0 | 0.9292181 |
| Departure Delay | 0.9292181 | 1.0 |

Running a correlation test in R, returns the following results:

> cor.test(dfDelaysClean[,1], dfDelaysClean[,2], conf.level = 0.99) ## correlation @ 99% confidence, and tests alternative hypothesis, p-value <.05 actually really small so high confidence correlation

Pearson's product-moment correlation

data: dfDelaysClean[, 1] and dfDelaysClean[, 2]

t = 1189.8, df = 223870, p-value < 2.2e-16

alternative hypothesis: true correlation is not equal to 0

99 percent confidence interval:

0.9284710 0.9299578

sample estimates:

cor

0.9292181

This results tells us a few things. First, with a p-value so small, it is exceptionally unlikely the data as a whole occurs because of chance. Next the correlation between the arrival delays and departure delays is very strong, showing, for example, that a departure delay is likely to be accompanied by an arrival delay, or leaving on time, tends to be associated with arriving on time. Further the 99% correlation confidence intervals are very close given correlation of .9292, meaning that the true correlation is very likely extremely near that value. And so obviously with the values returned from the ‘cor.test’ function in R, it is high unlikely the correlation is zero.

## Linear Algebra and Correlation

Inverting the correlation matrix from above, gives the following precision matrix:

|  |  |  |
| --- | --- | --- |
| Inverse Correlation Matrix | Arrival Delay | Departure Delay |
| Arrival Delay | 7.323130 | -6.804785 |
| Departure Delay | -6.804785 | 7.32313 |

Multiplying the precision matrix by the original matrix gives the following identity matrix:

|  |  |  |
| --- | --- | --- |
| Precision %\*% Orig. Correlation | Arrival Delay | Departure Delay |
| Arrival Delay | 1.0 | 0.0 |
| Departure Delay | 0.0 | 1.0 |

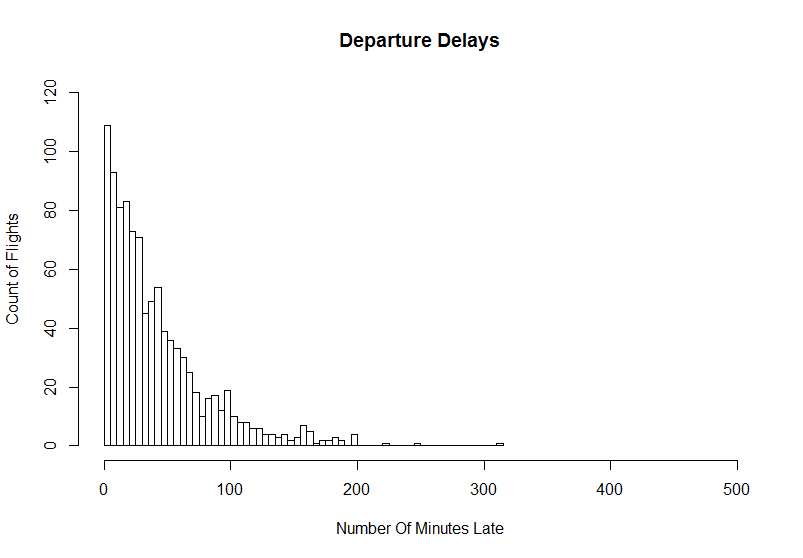
Not surprisingly since the prior multiplication gave us the identity matrix, reversing the matrices when multiply will still give the same identity matrix:

|  |  |  |
| --- | --- | --- |
| Orig. Correlation %\*% Precision | Arrival Delay | Departure Delay |
| Arrival Delay | 1.0 | 0.0 |
| Departure Delay | 0.0 | 1.0 |

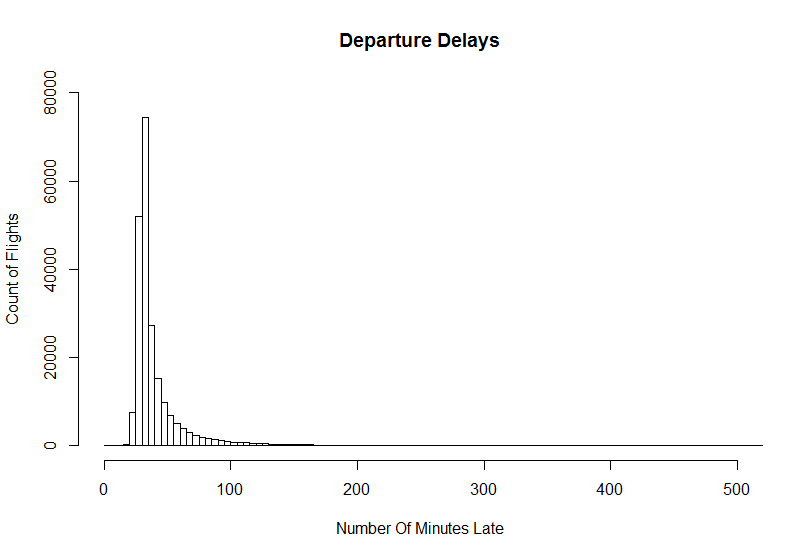
The diagonal on the inverse correlation matrix shows how tightly the variance around the mean is and the values in this case show that they are close. The multiplication of the original and the inverse giving the identity matrix just confirms that the two matrixes were in fact inverses of each other. Often, multiplying matrix AxB and then BxA would not give same product, but with an identity matrix they do.

## Calculus Based Probability and Statistics

The fitDistr function was run on the departure delay values shifted to the right by 33, so that the minimum value in the values was zero. The value of was 2.357657E-02. Next a random exponential distribution of a 1000 values was computed using that lambda value (expDist <- rexp(1000, 2.357657e-02)). This distribution’s histogram looks like the following:



The histogram of the departures delays (still shifted to the right so no zero values) is below. It was slightly factored from the original departure histogram to make the number of x and y axes more sensible to the 1000 value exponential function’s plot above.



The two histograms look somewhat similar and show that PDF giving by the fitDistr function is not that far off from the actual distribution that was being modeled.

The 5th and 95th percentiles of the random exponential distribution are:

* 5%: 2.330871
* 95%: 123.4885

The 95% confidence interval of the mean is 38.16333 < < 43.0291 (from mean(expDist) qnorm(.975)\* sd(expDist) / sqrt(1000). As for comparison the mean of the empirical data at 95% confidence is 42.29592 < < 42.53404

The empirical 5th and 95th percentiles of the departure delays (still shifted to the right to avoid negative numbers) is:

* 5%: 26
* 95%: 89

From the analysis of the exponential distribution as compared to the empirical shifted data, it can be seen that the exponential equation is not a perfect fit for actual model, but nonetheless, it is not widely different. The distribution of the empirical data tends towards a normal distribution, albeit quite right skewed, while obviously the empirical distribution is not that way. The 5th and 95th percentiles of the random distribution surround the empirical data’s percentiles, showing that the empirical data is not so tightly clustered. Lastly the means of the data are very close, so in terms of that, the results match similarly.