Engsci 211 Task 3: Airline Delays

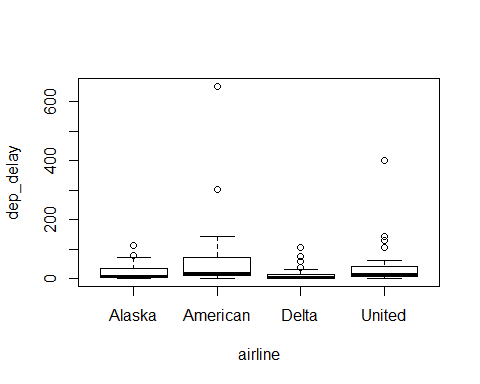
Alex Verkerk

library(s20x)  
delays.df = read.table("delays.txt", header = TRUE)  
head(delays.df)

## airline dep\_delay  
## 1 American 652  
## 2 American 4  
## 3 American 142  
## 4 American 302  
## 5 American 1  
## 6 American 57

### Exploratory Analysis

plot(dep\_delay~airline,data = delays.df)

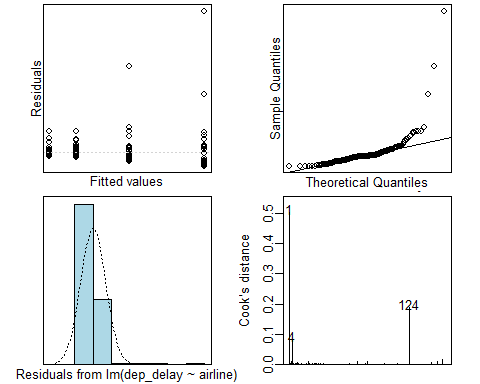


summaryStats(dep\_delay~airline,data = delays.df)

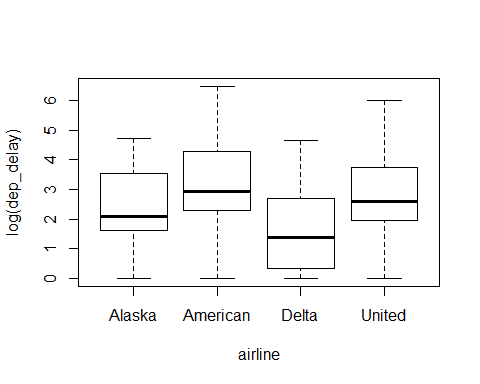
## Sample Size Mean Median Std Dev Midspread  
## Alaska 40 21.375 8.0 25.00635 28.00  
## American 40 57.875 18.5 112.76779 61.25  
## Delta 40 13.850 4.0 21.70259 12.75  
## United 40 36.425 13.5 67.72641 35.25

* Centre: The data for each airline appears to be bunched up toward the bottom. This makes the mean a poor measure of centrality, as the mean for all the airlines is significantly larger than any of the medians. Delta has the lowest mean & median values, with the mean & median for Alaska second lowest. United has a lower mean & median than American.
* Spread: American is the most spread, United second most spread, Alaska third, and Delta least spread. The spreads differ relatively significantly between all airlines.
* Skew: All 4 samples appear to show evidence of right skew, as evidenced by the normality check below.

delays.fit = lm(dep\_delay~airline, data = delays.df)  
modcheck(delays.fit)



boxplot(log(dep\_delay)~airline, data = delays.df)

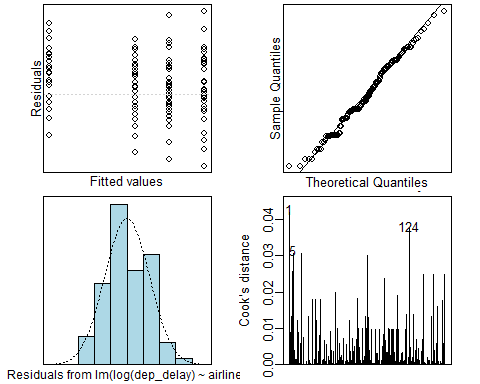


summaryStats(log(dep\_delay)~airline, data = delays.df)

## Sample Size Mean Median Std Dev Midspread  
## Alaska 40 2.406188 2.079442 1.204843 1.885744  
## American 40 2.982998 2.917405 1.533585 1.963591  
## Delta 40 1.690666 1.386294 1.400752 2.152580  
## United 40 2.692084 2.602003 1.375397 1.797642

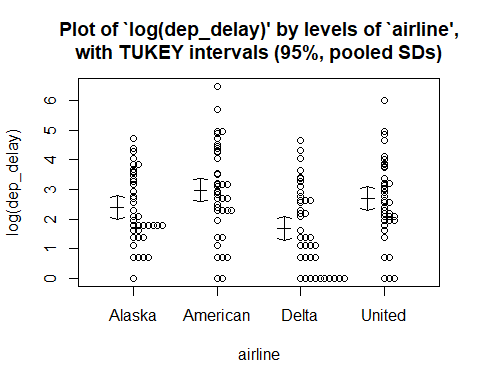
* Centre: Upon log transformation, the median delay for all of the airlines appear to be within about 1 unit of each other - with Delta having a slightly lower median delay than the other airlines.
* Spread: The spread of all samples is now very similar with a midspread of around 2 for all four.
* Skew: There is no evidence to suggest any significant skew remains after the data has been transformed. Delta could possibly be considered slightly right skewed but not significantly enough to affect our analysis.

delays.fit2 = lm(log(dep\_delay)~airline,data=delays.df)  
modcheck(delays.fit2)



summary1way(delays.fit2)

## ANOVA Table:  
## Df Sum Squares Mean Square F-statistic p-value   
## Between Groups 3 36.84011 12.28004 6.41477 4e-04   
## Within Groups 156 298.63672 1.91434   
## Total 159 335.47683   
##   
## Numeric Summary:  
## Sample size Mean Median Std Dev Midspread  
## All Data 160 2.44298 2.44140 1.45256 2.03104  
## Alaska 40 2.40619 2.07944 1.20484 1.88574  
## American 40 2.98300 2.91741 1.53359 1.96359  
## Delta 40 1.69067 1.38629 1.40075 2.15258  
## United 40 2.69208 2.60200 1.37540 1.79764  
##   
## Table of Effects: (GrandMean and deviations from GM)  
## typ.val Alaska American Delta United   
## 2.44298 -0.03680 0.54001 -0.75232 0.24910



multipleComp(delays.fit2)

## Estimate Tukey.L Tukey.U Tukey.p  
## Alaska - American -0.5768106 -1.3803 0.2266 0.2477  
## Alaska - Delta 0.7155220 -0.0879 1.5190 0.0995  
## Alaska - United -0.2858957 -1.0893 0.5175 0.7920  
## American - Delta 1.2923326 0.4889 2.0958 0.0003  
## American - United 0.2909149 -0.5125 1.0944 0.7832  
## Delta - United -1.0014177 -1.8049 -0.1980 0.0080

exp(multipleComp(delays.fit2))[,1:3]

## Estimate Tukey.L Tukey.U  
## Alaska - American 0.5616869 0.2515031 1.2543280  
## Alaska - Delta 2.0452540 0.9158525 4.5676553  
## Alaska - United 0.7513410 0.3364519 1.6778278  
## American - Delta 3.6412703 1.6305217 8.1319439  
## American - United 1.3376507 0.5989962 2.9873897  
## Delta - United 0.3673583 0.1644909 0.8203699

#reversing order of comparisons in order to compare United to Delta  
exp(-1\*multipleComp(delays.fit2))[,1:3]

## Estimate Tukey.L Tukey.U  
## Alaska - American 1.7803512 3.9760943 0.7972396  
## Alaska - Delta 0.4889368 1.0918789 0.2189307  
## Alaska - United 1.3309537 2.9721928 0.5960087  
## American - Delta 0.2746294 0.6133007 0.1229718  
## American - United 0.7475793 1.6694596 0.3347404  
## Delta - United 2.7221383 6.0793635 1.2189624

### Methods and Assumption Checks

The airline delays are numerical measurements made across 4 different groups of airlines. This suggests that we should use a one-way ANOVA or a linear regression model.

Assuming that our observations are independent, we initially tested our assumption of equality of variance with a linear model, which appears unsatisfied. There appears to be an influential observation at observation 1.

The Q-Q plot of our sample showed evidence of relatively significant right skewing of our data. This implies that a multiplicative model should be used, as the median will provide a better measure of centrality than the mean.

After applying a log transformation, the equality of variance and Normality assumptions appear to be satisfied. The influential observation at observation 1 appears to no longer be influential in this transformation.

Only the delays for American vs Delta and Delta vs United had a p-value of less than 0.05 meaning only those two comparisons have statistical significance. Those two p-values are 0.0003 and 0.008 respectively.

The model fitted is $\log({\tt dep\\_delays})\_{i} = \beta\_0 + \beta\_1 \times {\tt America\_{i}} + \beta\_2 \times {\tt Delta\_{i}} + \beta\_3 \times {\tt United\_{i}} + \varepsilon\_{i}$ where $\tt America\_i$ is 1 when looking at the delays for America airlines, $\tt Delta\_i$ is 1 when looking at the delays for Delta airlines, $\tt United\_i$ is 1 when looking at the delays for United airlines, and Alaska airlines is the baseline median; and .

Alternatively the model can be described as $\log({\tt dep\\_delays})\_{ij} = \mu+\alpha\_i+\epsilon\_{ij}$ where is the effect of the delay of any of the airlines, and .

### Executive Summary

The delays for 4 different airlines were recorded in order to compare which airlines had significant differences in delay times.

In order to analyse the data, we found we had to transform our observations. This resulted in the analysis for the delay comparison for the four airlines to be in relation to the median delay difference, and is representative of a multiplicative relationship.

There is no evidence in a difference in the median delays between Alaska & American, Alaska & Delta or Alaska & United. There is also no evidence of any difference between the median delays between American & United.

We have observed that for the median difference in delays between American & Delta airlines, delays for American will be between 1.63 to 8.13 times as long as those for Delta.

We have also observed that for the median difference in delays between United & Delta airlines, delays for United will be between 1.22 to 6.08 times as long as delays for Delta.