STAT480_Homework_2

Bin Feng

```
#include library
library(reshape2)
library(ggplot2)
library(biganalytics)
```

```
## Loading required package: bigmemory

## Loading required package: foreach

## Loading required package: biglm

## Loading required package: DBI
```

```
library(foreach)
require("knitr")
```

```
## Loading required package: knitr
```

```
#set working directory
opts_knit$set(root.dir = "~/Stat480/RDataScience/AirlineDelays")
```

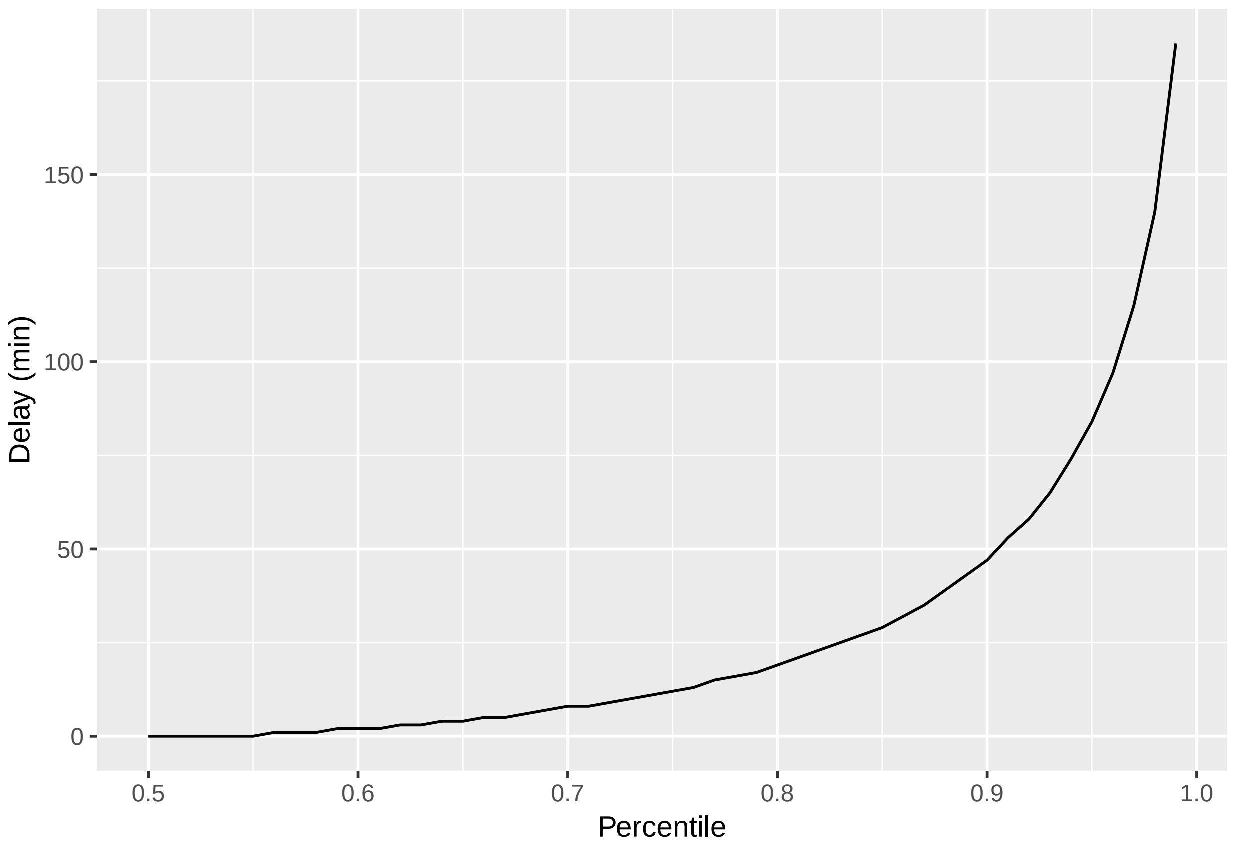
Question 1

A traveler is planning a trip for July 2009 and wonders about the amount of departure delay they might encounter. They have the data from 2007 and 2008 and want to look at delays that are at least of median length. Obtain the 50th through 99th percentiles for July data in those years and interpret what the results tell us about magnitudes and frequency of delayed departures in July during those two years.

```
# Attach the same big matrix to flight0708 using the descriptor file without
# creating any new large matrix
flight0708 <- attach.big.matrix("air0708.desc")
# Extrack July data from 0708
Month07 <- flight0708[(flight0708[,"Month"] == '7'), ]
# Create a variable to hold the quantile probabilities.
myProbs <- seq(0.5, 0.99, 0.01)
# foreach loop to calculate the quantile
delayQuantiles.July <- foreach(month = list(1:nrow(Month07)), .combine=cbind) %do% {
  quantile(Month07[month, "DepDelay"], myProbs, na.rm = TRUE)
}
delayQuantiles.July
```

```
## 50% 51% 52% 53% 54% 55% 56% 57% 58% 59% 60% 61% 62% 63% 64% 65% 66% 67%
##   0   0   0   0   0   0   1   1   1   2   2   2   3   3   4   4   5   5
## 68% 69% 70% 71% 72% 73% 74% 75% 76% 77% 78% 79% 80% 81% 82% 83% 84% 85%
##   6   7   8   8   9  10  11  12  13  15  16  17  19  21  23  25  27  29
## 86% 87% 88% 89% 90% 91% 92% 93% 94% 95% 96% 97% 98% 99%
##  32  35  39  43  47  53  58  65  74  84  97 115 140 185
```

```
# plot
qplot(myProbs, delayQuantiles.July, data = as.data.frame(delayQuantiles.July),
      geom = "line", xlab = "Percentile", ylab = "Delay (min)")
```

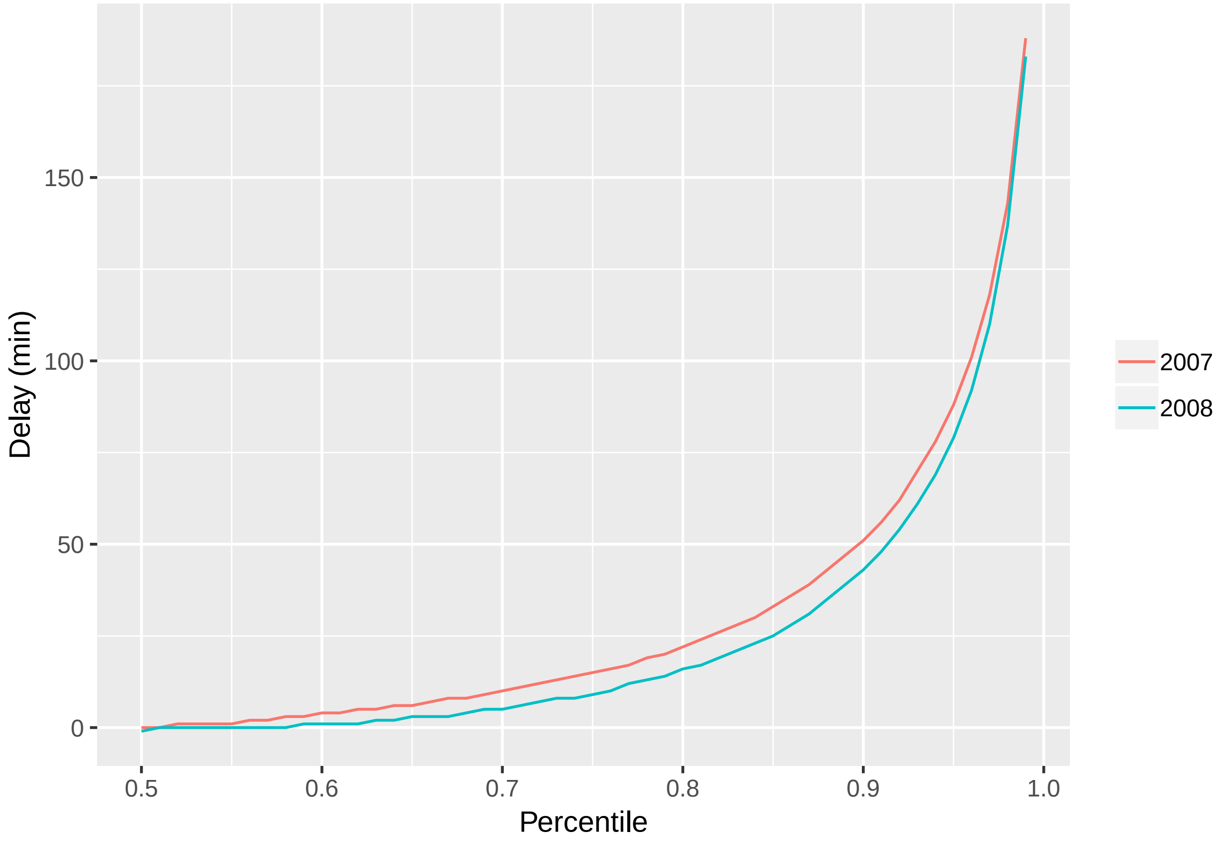


Based on the results, as the percentile increases (meaning the frequency of encountering such event decreases), the magnitude of delay time also increases. For the 50th percentile, the delay time is 0 minute, meaning the flight will depart on time. For the 80th percentile, the delay time are 19 minutes. One may also be interested in the 90th percentile since it is still likely to happen. The delay time are 47 minutes, which is also acceptable. If we look at the 99th percentile, 185 minutes of delay will occur for those unfortunate 1 percent passengers.

Question 2

The traveler is also curious about differences in departure delay percentiles for July during those two years. Compute and compare the 50th through 99th percentiles for July 2007 and July 2008. Provide an informative visualization along with interpretation of similarities and differences in the delay quantiles.

```r
# Create a variable to hold the quantile probabilities.
myProbs <- seq(0.5, 0.99, 0.01)
# split the data into two parts based on years, obtain the index
year.index <- split(1:nrow(Month07), Month07[,"Year"])
# foreach loop to calculate the quantile
delayQuantiles.July <- foreach( year = year.index, .combine=cbind) %do% {
  quantile(Month07[year, "DepDelay"], myProbs, na.rm = TRUE)
}
# Clean up the column names.
delayQuantiles.July <- cbind(myProbs, delayQuantiles.July)
colnames(delayQuantiles.July) <- c("Percentile", "2007", "2008")
# plot
ggplot(as.data.frame(delayQuantiles.July), aes(x=Percentile)) +
  geom_line(aes(y = delayQuantiles.July[,"2007"], colour = "2007")) +
  geom_line(aes(y = delayQuantiles.July[,"2008"], colour = "2008")) +
  xlab("Percentile") + ylab("Delay (min)") + theme(legend.title=element_blank())
```



Plot is shown above. Comparing the delay quantiles between 2007 and 2008, we note that they have the similar trend: as the percentile increases, the magnitude of delay time also increases. But we also notice that the plotted line for 2008 is always below the line of 2007, meaning the delay time is less if we are comparing the same percentile.

Question 3

Consider month and day of week as continuous linear predictors for departure delay. Obtain a linear regression model for departure delay as a function of month and day of week using the 2007-2008 data. Interpret what the model suggests about the relationship between delay time and the day of week and month. Comment on the usefulness of the model and any issues with using this model.

```
# use biglm.big.matrix() here. But since the data isn't very big, direct using
# lm() is acceptable.
delay.model01 <- biglm.big.matrix(DepDelay ~ Month + DayOfWeek, data = flight0708)
summary(delay.model01)
```

```
## Large data regression model: biglm(formula = formula, data = data, ...)
## Sample size =  14165949
##                Coef    (95%     CI)     SE p
## (Intercept) 11.2030 11.1478 11.2582 0.0276 0
## Month       -0.1914 -0.1970 -0.1858 0.0028 0
## DayOfWeek    0.1884  0.1788  0.1979 0.0048 0
```

```
summary(delay.model01)$rsq
```

```
## [1] 0.0004415451
```

Looking at the summary for this model, it suggests that as the month number incrases, the delay time will decrease. On the opposit, as the day of week increase, the delay time will increase. However, this model is quite useless because the residual sum of square for this model is only 0.0004415451, which indicates that even though these two variables ara statisticaly significant, they have only explained a very tiny amount of the data. Therefore, using this model will cause large inaccuracy.

Question 4

Rather than a straight linear trend, it is suggested that delays might be much worse in winter and not as bad in summer. Likewise, it is suggested that delays might get increasingly worse as the week goes on.

```
# use biglm.big.matrix() here. But since the data isn't very big, direct using
# lm() is acceptable.
delay.model02 <- biglm.big.matrix(DepDelay ~ I((Month - 6)^2) + I((DayOfWeek)^2),
                                  data = flight0708)
summary(delay.model02)
```

```
## Large data regression model: biglm(formula = formula, data = data, ...)
## Sample size =  14165949
##                    Coef   (95%    CI)     SE p
## (Intercept)      9.8811 9.8450 9.9172 0.0180 0
## I((Month - 6)^2) 0.0313 0.0295 0.0330 0.0009 0
## I((DayOfWeek)^2) 0.0235 0.0223 0.0246 0.0006 0
```

```
summary(delay.model02)$rsq
```

```
## [1] 0.0002049134
```

Looking at the summary for this model, it suggests that as the month moving towards winter season, the delay time increases. Also, as the day of week increase, the delay time will increase. However, this model is also quite useless because of its residual sum of square, which is even lower than the value in calculated for the linear model discussed in question 3. Comparing this two models, variable “Day of Week” has the same effect to the delay time in both models; variable “Month” has a negative effect to the delay time in the linear model which “(Month - 6)^2” has a postive effect to the delay time in the quatratic model. But since both of the models have very low residual sum of square, neither of them should be used for any interpretation or prediction without further modification.