STAT480 Homework2

Chenz Zhang, NetID chenziz2

2/6/2019

Contents

Exercise 1

```
library(biganalytics)
```

```
## Loading required package: bigmemory

## Loading required package: foreach

## Loading required package: biglm

## Loading required package: DBI
```

```
setwd("~/Stat480/RDataScience/AirlineDelays")
x <- attach.big.matrix("air0708.desc")  #We got this desc from homework1
JulyInds <- which(x[,"Month"] == 7)
Probs <- seq(0.50,0.99,0.01)
delayQuantiles <- quantile(x[JulyInds, "DepDelay"], Probs,
                           na.rm = TRUE)
print(delayQuantiles)
```

```
## 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
```

Magnitudes and frequency of delayed departures:

- The median of delay time is 0 and over half of the flights didn't delay. (frequency)
- 23% of the fights had serious delay which means they delayed over 15 mins. (magnitudes)

Exercise 2

```
y <- x[JulyInds,]
YearInds <- split(1:length(JulyInds),y[,"Year"])
```

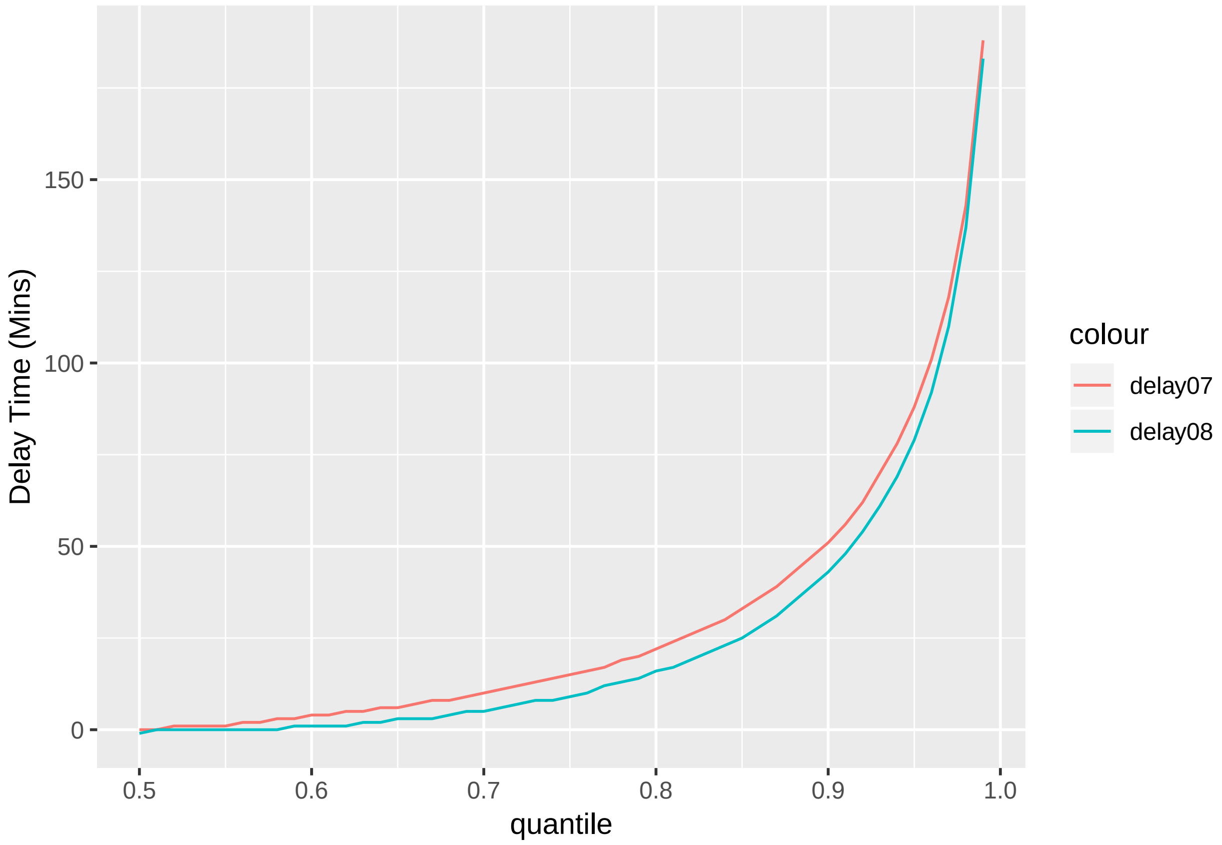
```r
delayQuantiles2 <- foreach( year = YearInds, .combine=cbind) %do% {
  quantile(y[year, "DepDelay"], Probs, na.rm = TRUE)
}

print(delayQuantiles2)
```

```
##     result.1 result.2
## 50%        0       -1
## 51%        0        0
## 52%        1        0
## 53%        1        0
## 54%        1        0
## 55%        1        0
## 56%        2        0
## 57%        2        0
## 58%        3        0
## 59%        3        1
## 60%        4        1
## 61%        4        1
## 62%        5        1
## 63%        5        2
## 64%        6        2
## 65%        6        3
## 66%        7        3
## 67%        8        3
## 68%        8        4
## 69%        9        5
## 70%       10        5
## 71%       11        6
## 72%       12        7
## 73%       13        8
## 74%       14        8
## 75%       15        9
## 76%       16       10
## 77%       17       12
## 78%       19       13
## 79%       20       14
## 80%       22       16
## 81%       24       17
## 82%       26       19
## 83%       28       21
## 84%       30       23
## 85%       33       25
## 86%       36       28
## 87%       39       31
## 88%       43       35
## 89%       47       39
## 90%       51       43
## 91%       56       48
## 92%       62       54
## 93%       70       61
## 94%       78       69
## 95%       88       79
## 96%      101      92
## 97%      118     110
## 98%      143     137
## 99%      188     183
```

Virtualize data from 2007 to 2008:

```
library(ggplot2)
df <- data.frame(delay07 = delayQuantiles2[,1], delay08 = delayQuantiles2[,2],
                 quantile = seq(0.50,0.99,0.01))
plot1 <- ggplot(df,aes(quantile)) +
         geom_line(aes(y = delay07, color = "delay07")) +
         geom_line(aes(y = delay08, color = "delay08")) +
         ylab("Delay Time (Mins)")
plot1
```



- **Difference**:
  - Delay time in 2007 has larger magnitudes than delay time in 2008 when they are under the same quantile.

  - 25% of the fights in 2007 and 20% of the fights in 2008 had serious delay which means they delayed over 15 mins.

  - The median of delay time in 2007 is 0, but the median of delay time in 2008 is -1.

- **Similarity**:
  - Delay time in 2007 and 2008 have the same trend (increasing) when quantile increases.

- About half of the fights in 2007 and 2008 delayed.

Exercise 3

```
fit1 <- biglm.big.matrix(DepDelay ~ Month + DayOfWeek, data = x)
sumfit1 <- summary(fit1)
summary(fit1)$rsq
```

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

```
sumfit1
```

```
## 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
```

The model suggests the relationship between DepDelay and the DayOfWeek and Month should be:

$$DepDelay = 11.2030 - 0.1914 \times Month + 0.1884 \times DayOfWeek$$

From the r-square result which is 0.0004415451 (extremely small), we can see the model lacks goodness of fit.
Two predictors are individually significant according to small p-value.
Month has negative effect on delay time value.(positive effect on delay condition)
DayOfWeek has positive effect on delay time value.(negative effect on delay condition)

- Delay time could be influenced by a lot of extra factors sach as location rather than date.
- Lack of goodness of fit means that Month and DayOfWeek cannot explain delay time by a linear model.
- Straight linear trend of Month and DayOfWeek cannot meet the real condition.

Some issues will appear by using this straight linear model:

- When Month = DayOfWeek = 0, delay time still exists as 11.2030.
- When the Month increases and DayOfWeek doesn't change, delay time will decrease.
  This is different from the reality. And it cannot explain seasonal weather change's influence.

Exercise 4

```
fit2 <- biglm.big.matrix(DepDelay ~ I((Month-6)^2) + I(DayOfWeek^2), data = x)
sumfit2 <- summary(fit2)
summary(fit2)$rsq
```

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

```
sumfit2
```

```
## 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
```

Interpret:

- Intercept: When Month = 6 and DayOfWeek = 0, delay time is 9.8811.

- Delays become worse in winter and better in summer. Comparing delay time on the same DayOfWeek, delay in June is the lowest.

- Quadratic DayOfWeek's coefficient is 0.0235. Delays on weekends is much worse than those on weekdays.

From the r-square result which is 0.0002049134 (extremely small), we can see the model lacks goodness of fit.
Two quadratic form predictors are individually significant according to small p-value.
I((Month - 6)ˆ2) and I(DayOfWeekˆ2) both have positive effect on delay time value.(negative effect on delay condition)

This model is even worse than the model in Exercise 3 for smaller r-square. It might because the model in this form is still uncorrect.

Therefore, sometimes the model which seems match the real situation performs worse than the original usless model.