Lab 01

Yixuan Li - yil845

31-07-2021

The Data Set and purpose of this study

The data are counts of vehicles at different locations on state highways across New Zealand. A statistical approach was attempted to find the relationship of square root of counts (denoted as *scount*) in relation with vehicle type (denoted as *class*) and the passing location (denoted as *siteRef*) in data file.

Linux commands

The following actions were done; 1, created a directory called lab01;

```
mkdir lab01
head lab01.csv
```

```
## mkdir: cannot create directory ‘lab01’: File exists
## class,siteRef,startDatetime,endDatetime,direction,count
## L,01600024,07-AUG-2020 03:00,07-AUG-2020 03:15,2,1.5
## H,01600024,07-AUG-2020 06:00,07-AUG-2020 06:15,1,4
## L,01600024,07-AUG-2020 09:00,07-AUG-2020 09:15,1,195.5
## L,01600024,07-AUG-2020 10:00,07-AUG-2020 10:15,2,280
## L,01600024,07-AUG-2020 11:45,07-AUG-2020 12:00,1,275
## L,01600024,07-AUG-2020 19:30,07-AUG-2020 19:45,2,107
## H,01600024,08-AUG-2020 13:15,08-AUG-2020 13:30,2,6
## H,01600024,08-AUG-2020 14:45,08-AUG-2020 15:00,2,5
## L,01600024,08-AUG-2020 19:00,08-AUG-2020 19:15,2,149
```

2, copied data file in directory lab01;

```
cp lab01.csv lab01/lab01.csv
```

3, navigated to the directory lab01;

```
cd lab01
```

4, total number of rows of lab01.csv was shown;

```
wc -l lab01.csv
```

```
## 429689 lab01.csv
```

5, the number of rows that contain 08-AUG-2020 is 141714.The total difference of rows from 08/aug/2020 is 287975 less than the original.

```
grep -i "08-AUG-2020" lab01.csv | wc -l
```

```
## 141714
```

6, the number of rows for which the start date is August 8th 2020 is 140153. The difference with result of V is 141714-140153= 1561. An example is shown for the start date is not August 8 but the end date is August 8.

```
awk -F',' '$3 ~ /08-AUG-2020/ {print $0}' lab01.csv | wc -l
awk -F',' '$3!~/08-AUG-2020/ && $4~/08-AUG-2020/' lab01.csv | sed 1q
```

```
## 140152
## L,01N00526,07-AUG-2020 23:45,08-AUG-2020 00:00,2,7.5
```

7, the unique values of column class are shown. They are H and L only.

```
awk -F',' '{print($1)}' lab01.csv | sort -u
```

```
## class
## H
## L
```

8, A file named August8.csv is generated which of the start date is August 8th 2020.

```
head -1 lab01.csv > August8.csv
awk -F',' '$3 ~ /08-AUG-2020/ {print $0}' lab01.csv >> August8.csv
```

Import, tidy and data manipulation

1, The following code reads the CSV file into R to produce a data frame.

```
august8 <- read.csv("August8.csv")
dim(august8)
```

```
## [1] 140152      6
```

2, The column “startDatetime” was converted from format of character to date/time format, and assigned to a variable august$start.

```
august8$start <- as.POSIXlt(august8$startDatetime,format="%d-%b-%Y %H:%M")
table(august8$start$mday)
```

```
## 
##      8 
## 140152
```

3, The following code subsets the data to eliminate rows where the start time is before 8:00 or after 18:00.

```
august8day <- subset(august8, start$hour >= 8 & start$hour <= 18)
dim(august8day)
```

```
## [1] 64240     7
```

4, A new variable scount which is the square root of the count was created to minimize the skewed distribution explained in previous study.

```
august8day$scount <-sqrt(august8day$count)
```

Modelling scount vs. class

5, The following code splits the data into training (90%) and test (10%) sets.

```
index <- sample(rep(1:10, length.out=nrow(august8day)))
train <- august8day[index > 1, ]
test <- august8day[index == 1, ]
```

6,A RMSE() function was defined.

```
RMSE <- function(obs, pred) {
    sqrt(mean((obs - pred)^2))
}
```

7, I calculated predictions for the test set from a simple mean model and a linear regression model with terms of the vehicle class (both fit on the training set).

```
obs <- test$scount
predMean <- mean(train$scount)
lmfit <- lm(scount ~ class, train)
predLM <- predict(lmfit, test)
```

8, The following code compared the performance of these two models in terms of RMSE.

```
RMSE(obs, predMean)
```

```
## [1] 6.830848
```

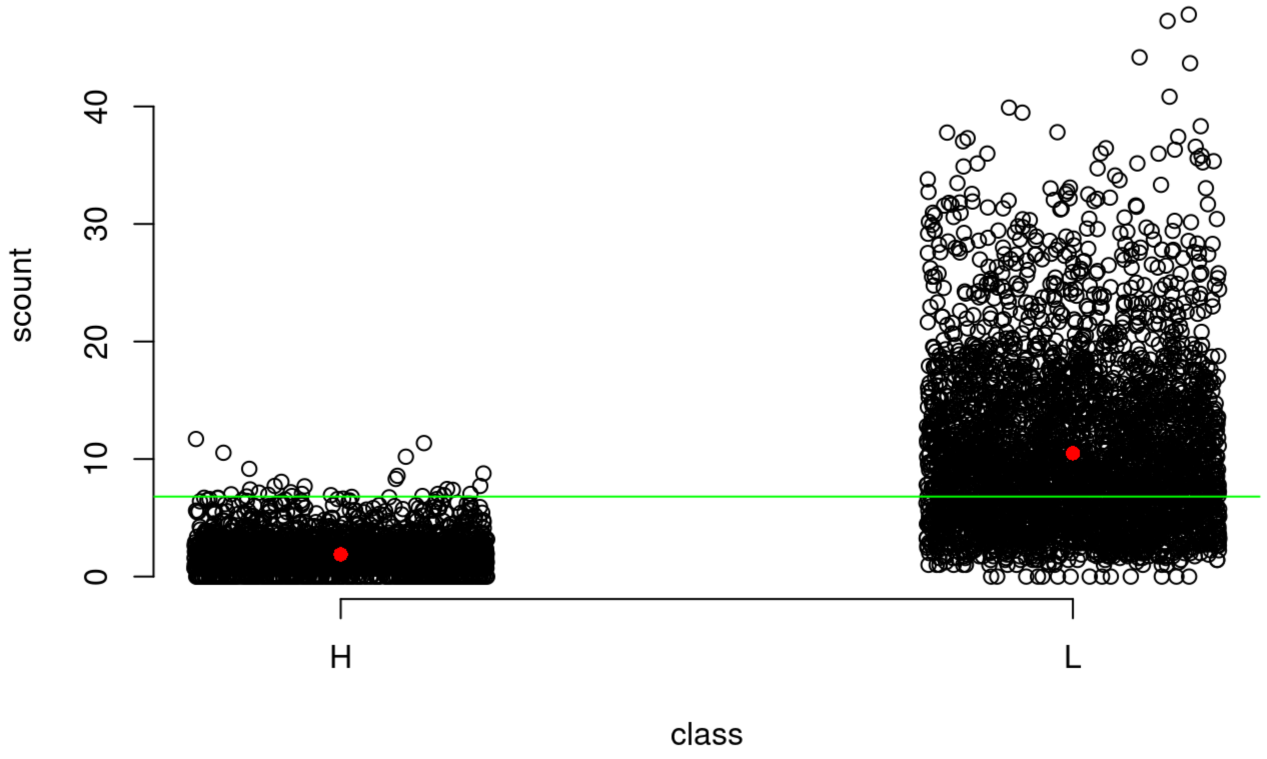
```
RMSE(obs, predLM)
```

```
## [1] 5.310705
```

Linear model that predicts scount in relationship with class H and L

This model did not predict the value very well.It shows the enormous amount of variability that is not captured by the linear model. It also shows that the simple overall mean over-predicted for almost all heavy (H) vehicle counts. While another group of data - scount of L vehicle class, was predominantly under the average predicted mean.

```
plot(scount ~ jitter(as.numeric(factor(class))), test,
     xlab="class", axes=FALSE)
axis(2)
axis(1, at=as.numeric(unique(factor(test$class))),
     label=unique(factor(test$class)))
abline(h=predMean, col="green")
points(as.numeric(unique(factor(test$class))),
       predict(lmfit, data.frame(class=unique(factor(test$class)))),
       pch=16, col="red")
```



##

Modeling scount in relationship of vehicle class and siteRef In order to find a model that had better predictive performance, I attemped the following approach.

1, I created training and testing data for class “H” and “L” respectively, named *Htrain*, *Ltrain*, *Htest* and *Ltest*.

```
Htrain <- subset(train, class == "H")
Ltrain <- subset(train, class == "L")
Htest <- subset(test, class == "H")
Ltest <- subset(test, class == "L")
```

2, I calculated predicted means for the test data (*Htest* & *Ltest* respectively) from a simple mean model and a linear regression model with a term for the vehicle siteRef, which were fitted on training set (Htrain & Htest respectively).

```
obsH <- Htest$scount
obsL <- Ltest$scount
predMeanH <- mean(Htrain$scount)
predMeanL <- mean(Ltrain$scount)
lmfitH <- lm(scount ~ siteRef, Htrain)
lmfitL <- lm(scount ~ siteRef, Ltrain)
predLMH <- predict(lmfitH, Htest)
predLML <- predict(lmfitL, Ltest)
```

3, The RMSEs of linear regression model of train and test set were compared for class H & L respectively.

```
RMSE(obsH, predMeanH)
```

```
## [1] 1.444369
```

```
RMSE(obsL, predMeanL)
```

```
## [1] 6.905828
```

```
RMSE(obsH, predLMH)
```

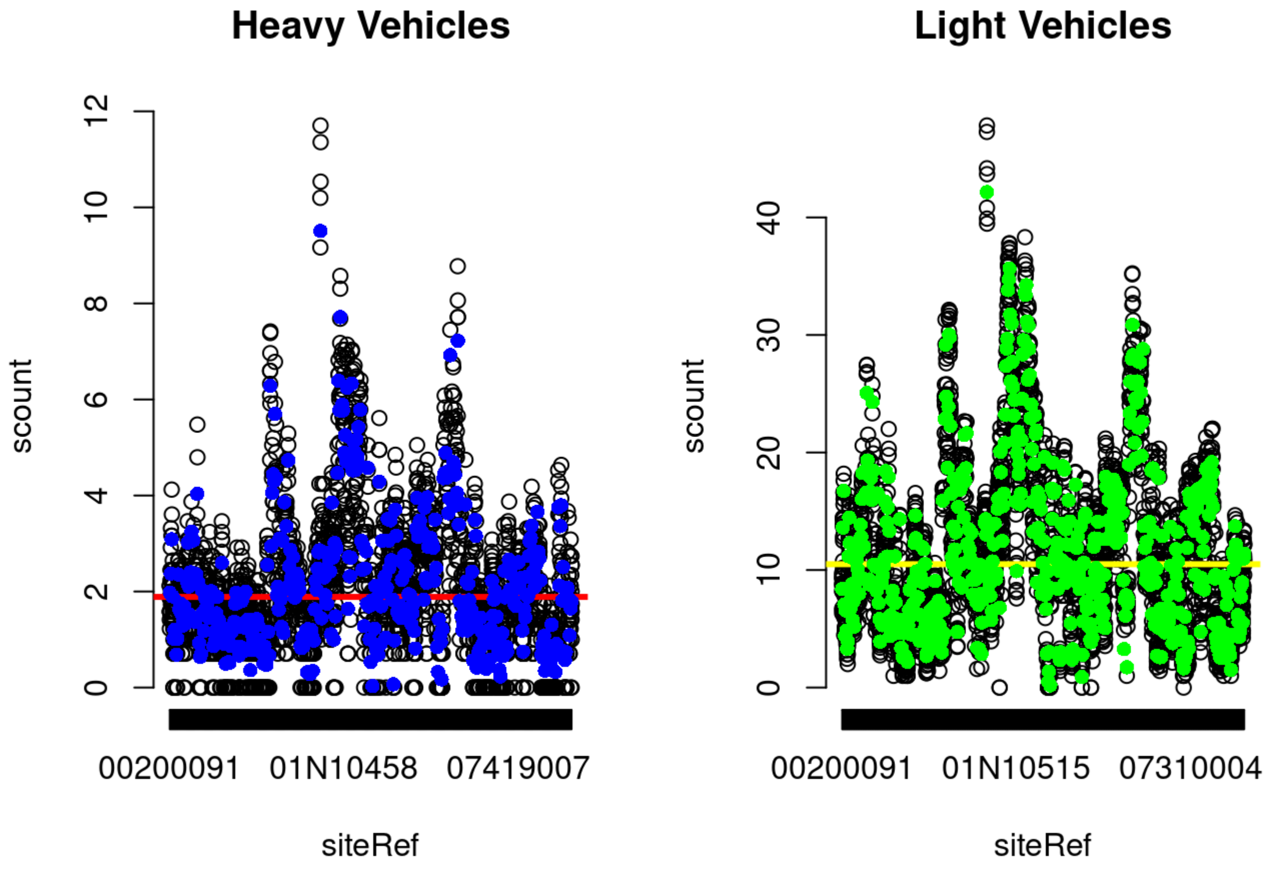
```
## [1] 0.6143512
```

```
RMSE(obsL, predLML)
```

```
## [1] 1.543638
```

4, The following graphs showed linear fitting of scount with siteRef for different class H (left) & L (right) respectively. Although data were highly dispersed, the predicted mean fitted in middle of both data in H and L.The tested data fitted nicely in model. The obvious difference of average scount of class H and L could be seen.

```
par(mfrow=c(1,2))
plot(scount ~ jitter(as.numeric(factor(siteRef))), Htest,
     xlab="siteRef", axes=FALSE,main="Heavy Vehicles")
axis(2)
axis(1, at=as.numeric(unique(factor(Htest$siteRef))),
     label=unique(factor(Htest$siteRef)))
abline(h=predMeanH, col="red",lwd=3)
points(as.numeric(unique(factor(Htest$siteRef))),
       predict(lmfitH, data.frame(siteRef=unique(factor(Htest$siteRef)))),
       pch=16, col="blue")
plot(scount ~ jitter(as.numeric(factor(siteRef))), Ltest,
     xlab="siteRef", axes=FALSE,main="Light Vehicles")
axis(2)
axis(1, at=as.numeric(unique(factor(Ltest$siteRef))),
     label=unique(factor(Ltest$siteRef)))
abline(h=predMeanL, col="yellow",lwd=3)
points(as.numeric(unique(factor(Ltest$siteRef))),
       predict(lmfitL, data.frame(siteRef=unique(factor(Ltest$siteRef)))),
       pch=16, col="green")
```



The significance test on H & L

In order to understand if there was a significant difference of two datasets of class H & L, I did a simple t test based on two paired data of train and test. The results showed **p value <2.2e^{-16}**, which showed a significant difference for class H & L.

```
t.test(Htrain$scount,Ltrain$scount,paired=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  Htrain$scount and Ltrain$scount
## t = -224.73, df = 36674, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.679560 -8.529469
## sample estimates:
## mean of x mean of y
##  1.888132 10.492646
```

```
t.test(Htest$scount,Ltest$scount,paired=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  Htest$scount and Ltest$scount
## t = -74.141, df = 4102.6, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.914050 -8.454756
## sample estimates:
## mean of x mean of y
##  1.944937 10.629341
```

Summary

In this study, I analyzed the traffic flow counted as square root of counts (*scount*) in relation to the vehicle type (*class*) and the passing site (*siteRef*). In order to analyze the daytime traffic flow, we extracted dataset that the starting time from 8 to 18 o'clock. I used a square-root transform on the counts to get a less skewed variable scount.

Firstly I tried a linear regression model to correlate scount with vehicle class based on a set of train data. I tested a set of test data on the model.However, majority of data from class H fall under mean which shows the limitation of this model.

In order to explore a better model, I seperated train and test data based on vehicle type H and L and created four dataset which are Htrain, Ltrain and Htest, Ltest. The linear models were done on scount vs. siteRef again on these new data sets. The results showing the vehicle count vs siteRef based on different vehicle class performed better than a model based on the overall mean count. To test difference of two datasets based on H and L, I did a simple t test, which shows **p value <2.2e^{-16}**, a significant difference between H and L datasets. According to model during the day time (8am - 6pm) on August 8 2020, the average count of light vehicles at different site over New Zealand is about **100** in contrast to the count of heavy vehicles ~ **4**.

Due to highly scattered data sets and skewed probability distribution, a better model probably exists.