STA 141C Final Project R Markdown

Yutian Yang(915327218) & Xin Ye(915263534) & Xinyue Wang (915059201) & Falak Shah (914663151)

3/10/2020

Introduction

Our analysis utilizes a dataset of “Metro Instate Traffic Volume” consisting of parameters for hourly Minneapolis-St Paul, MN traffic volume for westbound I-94 from 2012-2018. For our project, we took on the primary task of using the bag of little bootstraps to predict the confidence interval for the correlation between traffic volume and other numerical parameters via parallel processing. In order to accomplish this goal, we, preliminarily, needed to create a model of deterministic and significant predictors that contributed to the general explanation of the generation of instate traffic volume on I-94 from Minneapolis-St.Paul during 2012-2018. Furthermore, we can use this model to verify its certain characteristics such as differences in traffic volume at different times or on holidays along with using it to predict the confidence interval for the traffic volume using BLB.

Data Set

The dataset contains 48,204 observations with 9 different attributes. The numerical parameters include temperature (measured in kelvin), amount of rainfall at hourly intervals (measured in mm), amount of snowfall at hourly intervals (measured in mm), percentage of cloud cover, date, time of the data collected and finally, our parameter of interest - Traffic volume at hourly intervals. The categorical variables include holiday (including state holidays), short description of the weather (Eg: “Clouds”, “Clear”) and longer weather descriptions (Eg: “Sky is clear”, “Overcast clouds”, “Scattered clouds”, “light rain”, “light intensity drizzle”, etc).

Statistical Questions of Interest

- What is an appropriate linear model to predict traffic volume for our data set ?
- What is the Confidence Interval for traffic volume, intercept, Temperature and Cloud coverage using BLB ?

Methods of Analysis

A.1 Study design:

First, we build a linear regression model for traffic volume after. We then further our examination by selecting the most appropriate model for our ariable of interest and identify significant attributes. After identifying these attributes and selecting the best model, we use bag of little bootstraps to help construct a 95% confidence interval for traffic volume along with its correlation with other variables.

We first visually explore the data to examine the assumptions that validate our regression methods and chosen classification techniques. Given the nonnormality of our parameter of interest, the number of shares (as shown below), we utilize a Poisson regression to explain the correlations between our significant regressors to our variable of interest.

A.2 Statistical Analysis

A.2.1 Visual Exploration

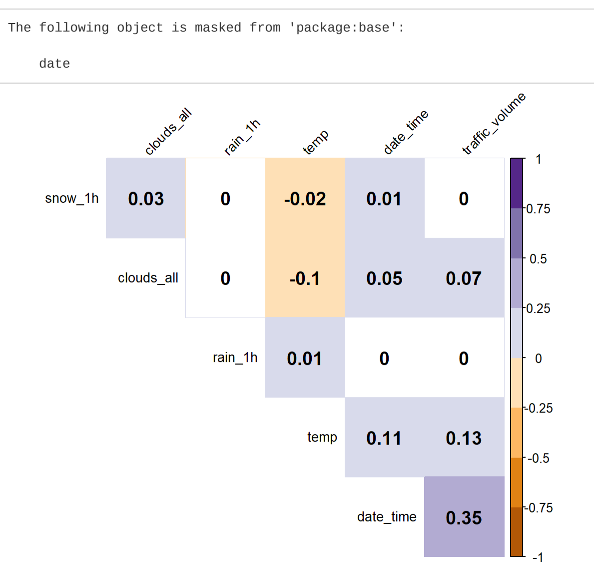
In this section, we collect and visualise some statistics of our dataset. Based on our visual analysis, we make the following observations about our dataset:

1. From the first plot, we make initial prognosis about the relationship among various numeric variables through the correlogram. Initial observation of the correlogram suggests that traffic volume has a significant relationship only with the date_time variable.
2. The second plot depicts the distribution of our variable of interes - Traffic volume. We can observe from the histogram that the distribution of traffic volume is not perfectly symmetric, it is slightly skewed to the left.
3. The third plot depicts the qq plot of traffic volume, from which we can observe that it does not follow normality.
4. The next four plots check the variation of traffic volume on various holdiays, different weather conditions, hours and years respectively, through conditional boxplots. The most traffic volume on holidays is observed on New Years Day and the least volume is observed on Columbus Day, whereas on all the other days, the traffic volume is in mid-range. Traffic volume does not vary much with the type of weather conditions and it is similar throughout all the six years. As far as traffic volume across various hours is concerned, as expected, it increases as we move from mornng to afternoon and decreases as we proceed from late evening to night.

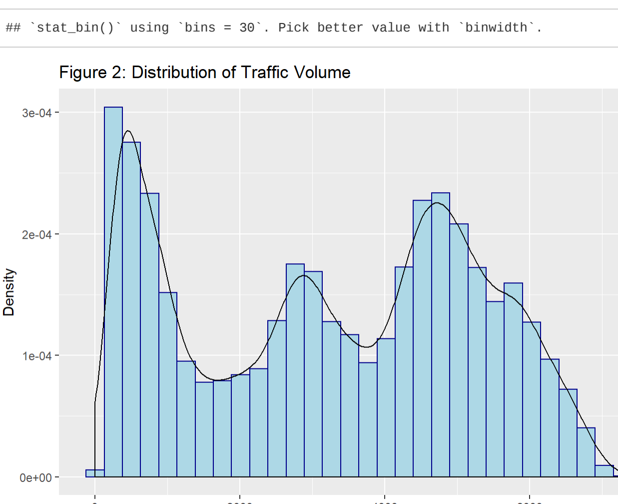
```
## corrplot 0.84 loaded
```

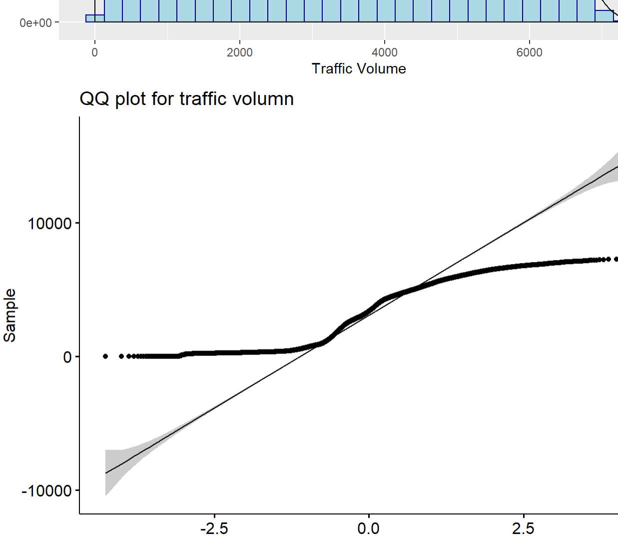
```
##
## Attaching package: 'lubridate'
```

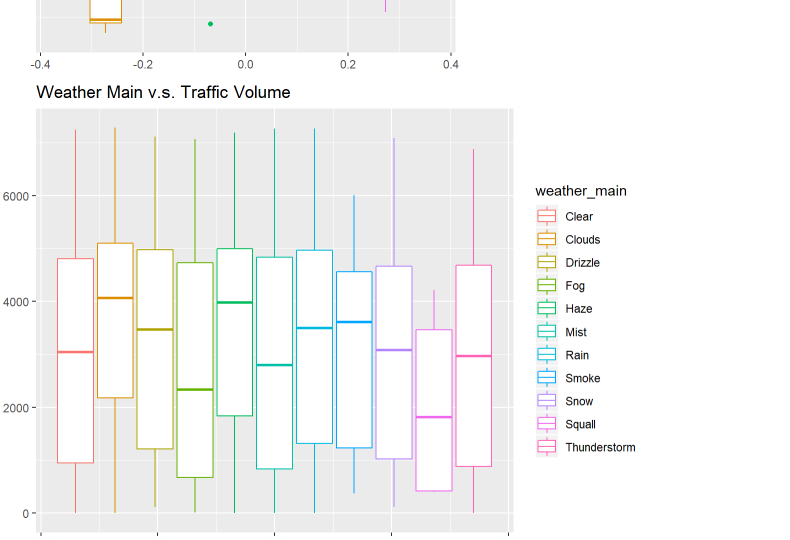
```
## The following object is masked from 'package:base':
##
##     date
```

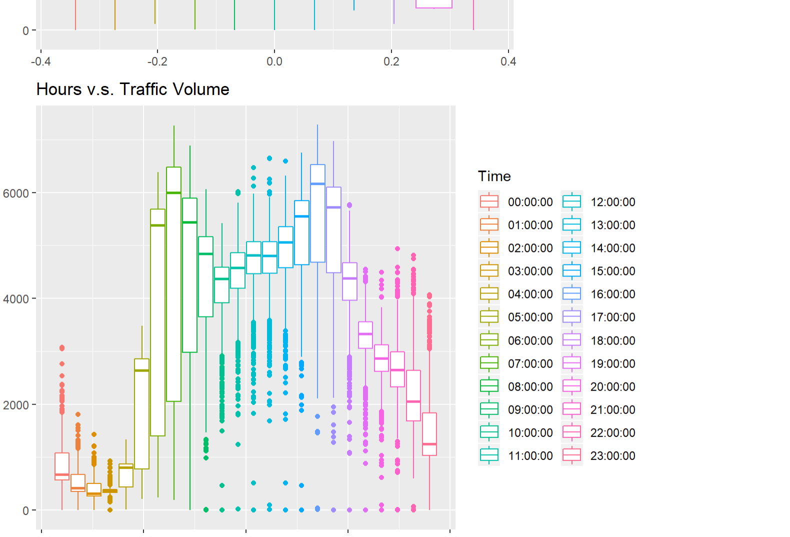


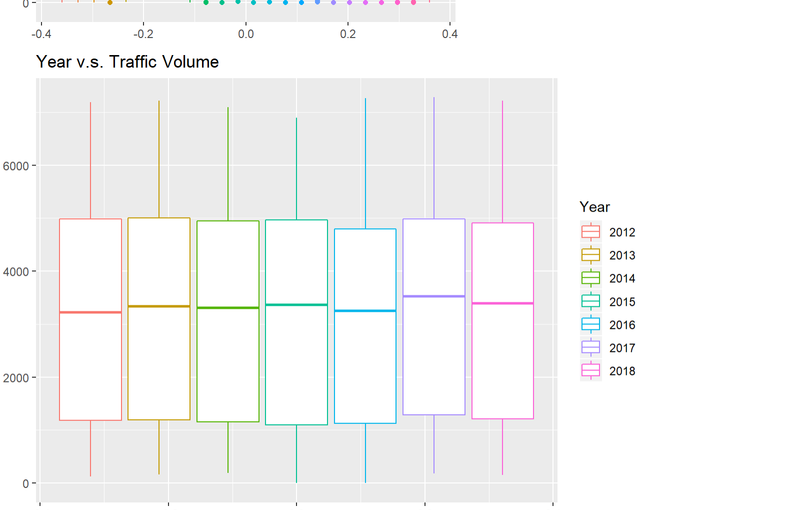

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

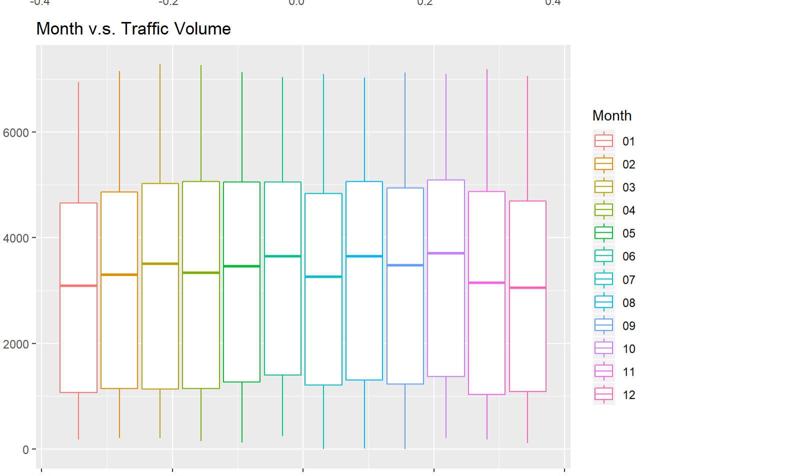












Main Analysis

Model Building

Considering the variables of interest,traffic volume , we decided to model our data using linear regression. In total, we developed five models : The first three models consist of experimentation with a variety of variables. For instance, the first model finds time as an important factor. In the fourth and fifth model, we demean traffic volume and weather respectively. We then calculate and compare AIC, BIC and R^2 for each model.

AIC (Akaike Information Criterion) is calculated as: -2(log-likelihood + k/n), where k represents the number of parameters.

BIC (Bayesian Information Criterion) is calculated as: -2(log-likelihood + log(n)k), where k represents the number of parameters.

We then evaluated each model and found their Null and Residual Deviances. We use the McFadden Pseudo R^2 for poisson regression, which the simple formula below describes

Below are the summary statistics for all the five linear models

$$\begin{bmatrix} ModelNumber & AIC & BIC & R^2 \\ 1 & 867035.7 & 867158.7 & 0.0392 \\ 2 & 796897.4 & 797134.6 & 0.7758 \\ 3 & 796749 & 797074 & 0.7766 \\ 4 & 796530.2 & 796653.1 & 0.01385 \\ 5 & 796575.6 & 796610.7 & 0.006542 \end{bmatrix}$$

Interpretation

We interpret each R^2 as the amount of variation in the traffic volume explained by the respective model. For example, the first model explains about 3.92% of the variation in the traffic volume. We notice also that the R^2 values indicate explanation power for our five models.

The third model has the highest R^2 value (0.7766), which becomes our model of choice.

Best Linear Regression Model for predicting traffic volume:

traffic_volume ~ temp + clouds_all + as.factor(time) + as.factor(weather_main)

As we can see from our regression model, the linear regression procedure deemed 4 of our variables as significant to predicting traffic volume, with two variables being categorical. The R^2 as mentioned above is 0.7766

Bag of little Bootsraps

As the assumptions of the Central Limit Theorem are violated i.e. the data points do not exhibit same variance, we cannot use the classical method to calculate the confidence interval. Therefore, we implement the bag of little bootstraps to compute the Confidence Interval.

BLB is a procedure which incorporates features of both the bootstrap and subsampling to yield a robust, computationally efficient means of assessing the quality of estimators.

It includes the following steps:

- sample without replacement the sample s times into sizes of b
- for each subsample
- resample each until sample size is n, r times
- compute the bootstrap statistic (e,g., the mean) for each bootstrap sample
- compute the statistic (e.g., confidence interval) from the bootstrap statistics
- take the average of the statistics

In other words, the bag of little bootstraps = subsample + bootstrap. However, for each bootstrap, we sample n from b with replacement instead of sample b from b as in oridinary bootstrap.

We implement the above mentioned procedure to caclculate the 95% confidence interval for traffic volume, the intercept, coefficient of temperature and coefficient of cloud coverage.

We use a multi-core, parallel implementation in which we read in the data to each worker to speed up the process and make it more efficient as compared to a single core implementation.

We consider B = 1000 in each of the four cases to compute 95% CI using BLB and the obtained 95% confidence intervals are summarized below:

$$\begin{bmatrix} Parameter & Upper\ Bound & Lower\ Bound & Confidence\ Interval \\ Traffic\ Volume & 2526.835 & 3750.232 & (2526.835, 3750.232) \\ Intercept & -3540.370 & -2805.921 & (-3540.370, -2805.921) \\ Tempreature & 20.81797 & 23.40296 & (20.81797, 23.40296) \\ Cloud\ Coverage & 3.745826 & 4.676337 & (3.745826, 4.676337) \end{bmatrix}$$

Conclusion

We begin the procedure with initial data visualization to get a graphical idea of the variation of traffic volume based on the various holidays, weather types, hours and years. After initial graphical diagnosis, we used linear regression to find the best model to predict traffic volume. After testing five different models and comparing their respective AIC, BIC and R^2 values, we find that the third model, with an R^2 = 0.7766 is the most robust model. Our results indicate a moderately strong inferential power to predict traffic volume. Finally, we use BLB coupled with multi core parallel implementation to compute a robust confidence interval as well as speed up the process.

Appendix: All code for this report

```r
library(car)
library(GGally)
library(dplyr)
library(ggpubr)
library(tidyr)
library(tidyverse)
library(ggplot2)
##Import dataset and split date into month, day, and time
Metro <- read_csv("Metro_Interstate_Traffic_Volume.csv")
New_Metro = separate(Metro, date_time, c("date","time"), sep = " ", remove = TRUE,
  convert = FALSE, extra = "warn", fill = "warn")
New_Metro = separate(New_Metro, date, c("year","month","day"), sep = "-", remove = TRUE,
  convert = FALSE, extra = "warn", fill = "warn")
summary(New_Metro)
## plot the correlation bewteen variables to see if there are interactions.
#ggcorr(New_Metro)


library(corrplot)
library(RColorBrewer)
library(lubridate)
data = Metro
#extract hour from date_time
data$date_time <- hour(data$date_time)

#correlation coefficients
M <-cor(data[,-c(1,6,7)])
corrplot(M, type="upper",method="color",
         addCoef.col = "black", order="hclust", number.cex= 7/ncol(data[,-c(1,6,7)]),
         tl.col="black", tl.srt=45,tl.cex = .85,
         p.mat = cor.mtest(data[,-c(1,6,7)])$p, sig.level = 0.1, insig = "blank",
         diag=FALSE, col=brewer.pal(n=8, name="PuOr"))

#distribution of y
ggplot(data= New_Metro, aes(x = traffic_volume)) +
  geom_histogram(aes(y=..density..),color="darkblue", fill="lightblue") +
  geom_density()+
  labs(title="Figure 2: Distribution of Traffic Volume",x="Traffic Volume", y = "Density")
ggqqplot(Metro$traffic_volume,main = "QQ plot for traffic volumn")
#how holiday influence y
ggplot(data = New_Metro[New_Metro$holiday!='None',], aes(group = holiday, y=traffic_volume,color=holiday )) +
  geom_boxplot()+
  labs(title= "Figure 3:Holiday v.s. Traffic Volume",y="Traffic Volume", color = "Holiday")

#how weather influence y
ggplot(data = New_Metro, aes(group = weather_main, y=traffic_volume,color=weather_main )) +
  geom_boxplot()+
  labs(title="Weather Main v.s. Traffic Volume",y="Traffic Volume", color = "weather_main")

#how hour influence y
ggplot(data = New_Metro, aes(group = time, y=traffic_volume,color=time )) +
  geom_boxplot()+
  labs(title="Hours v.s. Traffic Volume",y="Traffic Volume", color = "Time")

#how year influence y
ggplot(data = New_Metro, aes(group = year, y=traffic_volume,color=year )) +
  geom_boxplot()+
  labs(title="Year v.s. Traffic Volume",y="Traffic Volume", color = "Year")

#how month influence y
ggplot(data = New_Metro, aes(group = month, y=traffic_volume,color= month )) +
  geom_boxplot()+
  labs(title="Month v.s. Traffic Volume",y="Traffic Volume", color = "Month")
##Linear Model Selection
lm1 = lm(traffic_volume~ temp + clouds_all + as.factor(weather_main),data = New_Metro)
### Combining AIC, BIC and linear model selection to select the prediction model.
AIC(lm1)
BIC(lm1)
summary(lm1)
###Factor Time is significant
lm2 = lm(traffic_volume~ temp + clouds_all + as.factor(time),data = New_Metro)
summary(lm2)
AIC(lm2)
BIC(lm2)
#plot(lm2)
#hist(lm2$residuals)
### Another possible model
lm3 = lm(traffic_volume~ temp + clouds_all + as.factor(time) + as.factor(weather_main),data = New_Metro)
AIC(lm3)
BIC(lm3)
summary(lm3)


## Demean the traffic volume since it is the dominant factor
Metro_without_time = New_Metro %>%
  group_by(time)%>%
  filter(temp != 0.00) %>%
  mutate( new_traffic_volume = traffic_volume - mean(traffic_volume))


## New linear Model
lm4 = lm(new_traffic_volume~ temp + clouds_all + as.factor(weather_main),data = Metro_without_time)
AIC(lm4)
BIC(lm4)
summary(lm4)
# lm4$coefficients
# plot(lm4)


## Demean the weather factor
Metro_without_time_weather = Metro_without_time%>%
  group_by(weather_main)%>%
  mutate(new_traffic_volume2 = new_traffic_volume - mean(new_traffic_volume))

lm5 = lm(new_traffic_volume2~ temp + clouds_all,data = Metro_without_time_weather)
summary(lm5)
AIC(lm5)
BIC(lm5)
#plot(lm5)
#hist(lm5$residuals)

write_csv(Metro_without_time_weather,"Metro_without_time_weather.csv")

## Using parallelization to bootstrap for CI of traffic volume
library(parallel)
cl = makeCluster(4)
B = 1000
predict_newdata <- function(data){
  model = lm(data$traffic_volume~data$temp+data$clouds_all)
  coefficient = model$coefficients
  y = coefficient[[1]] + coefficient[[2]]* data$temp + coefficient[[3]]* data$clouds_all
}
singleBoots <- function(i){
  index = sample(x = seq_len(n), size = n, replace = TRUE)
  data_star = data[index,]
  predict_newdata(data_star)
}
clusterEvalQ(cl,{
  library(tidyverse)
  data <- read_csv("Metro_without_time_weather.csv")
  n <- length(data$traffic_volume)
  predict_newdata <- function(data){
  model = lm(data$traffic_volume~data$temp+data$clouds_all)
  coefficient = model$coefficients
  y = coefficient[[1]] + coefficient[[2]]* data$temp + coefficient[[3]]* data$clouds_all
}
})
predict_newdata = parSapply(cl, seq_len(B),singleBoots)
predict_newdata %>% quantile(c(0.025,0.975))
stopCluster(cl)

## Using parallelization to bootstrap for CI of intercept
cl = makeCluster(4)
B =1000
coef_beta0 <- function(data){
  model = lm(data$traffic_volume~data$temp+data$clouds_all)
  coefficient = model$coefficients[[1]]
}
singleBoots <- function(i){
  index = sample(x = seq_len(n), size = n, replace = TRUE)
  data_star = data[index,]
  coef_beta0(data_star)
}
clusterEvalQ(cl,{
  library(tidyverse)
  data <- read_csv("Metro_without_time_weather.csv")
  n <- length(data$traffic_volume)
  coef_beta0 <- function(data){
  model = lm(data$traffic_volume~data$temp+data$clouds_all)
  coefficient = model$coefficients[[1]]
}
})
beta0_ci = parSapply(cl, seq_len(B),singleBoots)
beta0_ci %>% quantile(c(0.025,0.975))
stopCluster(cl)

## Using parallelization to bootstrap for CI of coefficient of temperature
cl = makeCluster(4)
B =1000
coef_beta1 <- function(data){
  model = lm(data$traffic_volume~data$temp+data$clouds_all)
  coefficient = model$coefficients[[2]]
}
singleBoots <- function(i){
  index = sample(x = seq_len(n), size = n, replace = TRUE)
  data_star = data[index,]
  coef_beta1(data_star)
}
clusterEvalQ(cl,{
  library(tidyverse)
  data <- read_csv("Metro_without_time_weather.csv")
  n <- length(data$traffic_volume)
  coef_beta1 <- function(data){
  model = lm(data$traffic_volume~data$temp+data$clouds_all)
  coefficient = model$coefficients[[2]]
}
})
beta1_ci = parSapply(cl, seq_len(B),singleBoots)
beta1_ci %>% quantile(c(0.025,0.975))
stopCluster(cl)

## Using parallelization to bootstrap for CI of coefficient of clouds coverage.
cl = makeCluster(4)
B = 1000
coef_beta2 <- function(data){
  model = lm(data$traffic_volume~data$temp+data$clouds_all)
  coefficient = model$coefficients[[3]]
}
singleBoots <- function(i){
  index = sample(x = seq_len(n), size = n, replace = TRUE)
  data_star = data[index,]
  coef_beta2(data_star)
}
clusterEvalQ(cl,{
  library(tidyverse)
  data <- read_csv("Metro_without_time_weather.csv")
  n <- length(data$traffic_volume)
  coef_beta2 <- function(data){
  model = lm(data$traffic_volume~data$temp+data$clouds_all)
  coefficient = model$coefficients[[3]]
}
})
beta1_ci = parSapply(cl, seq_len(B),singleBoots)
beta1_ci %>% quantile(c(0.025,0.975))
stopCluster(cl)
```