Assignmnet #3

Saeid Hosseinipoor

10/7/2016

Problem 1

1a)

Some tecniques were applied to investigate the distribution, and the relationship between the predictor variables. The following figures shows that there are not exist strong linear relationship between the predictors except ***Refractive Index*** and ***Calcium***. This implies that calcium and reflactive index are correlated and we may use one of them or combination of them as a single predictor which one is easier.

The weak linear relationship between predictors never deny existance of nonlinear correlations. Visualization helps us to realize whether any pattern is available in plotted data. By looking at the plots, I would say some nonlinear correlations could be discovered between ***Aluminum*** and ***Calcium*** or ***Sodium***.

Moderate linear correlations either negative or positive are aviable between these pairs: *RI/Si*, *Mg/Al*, *Mg/Ba*, *Ba/Al*, and *RI/Al*.

Some of ions like *Mg*, *Ba*, and *Fe* have values of zero. Looking into chemistry of the glasses releavs that these ions are not essential part of the glasses, but they carries some identifications such as color of galss. Therefore; they would be very useful to classify the glasses as we have a variable as type.

Adjusted box plots shows there is some outliers in data set.

```
Ion = c("RI", "Na", "Mg", "Al", "Si", "K", "Ca", "Ba", "Fe")    # Ion vector
par(mfrow=c(3,3))
for (i in Ion){
  adjbox(data=Glass, RI ~ Type, xlab="Type", ylab=i, main="Adjusted")}
```



```
par(mfrow=c(1,1))               # Reset display
```

Skewness for predicors were calculated. It shows that some predictors are highly skewed.

```
apply(Glass[,1:9], 2, skewness)
```

```
##         RI         Na         Mg         Al         Si          K
##  1.6254305  0.4541815 -1.1525593  0.9072898 -0.7304472  6.5516483
##         Ca         Ba         Fe
##  2.0470539  3.4164246  1.7543275
```

1b)

The skewness calculations show that ***Potasium***, ***Barium***, and ***Calcium*** are three most skewed values in glass data set. Iron and Refractive index are also highly skewed. *symbox* and *boxcox* fuctions confirmed the findings as well.

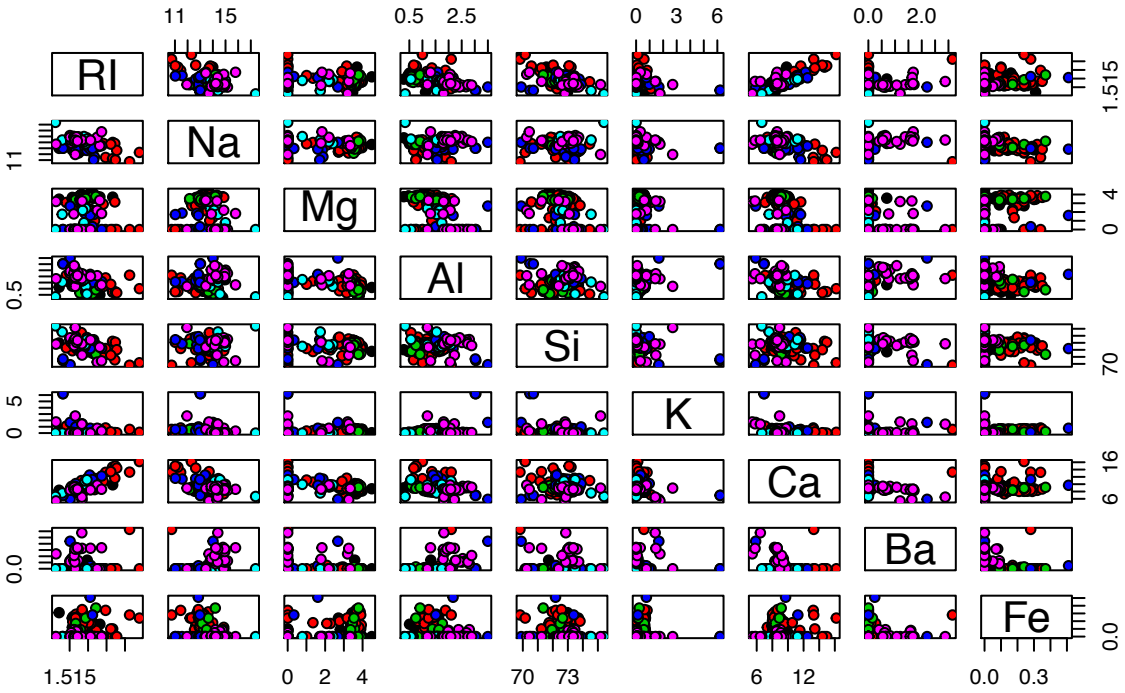

Figure 1: Glass Data Visualization

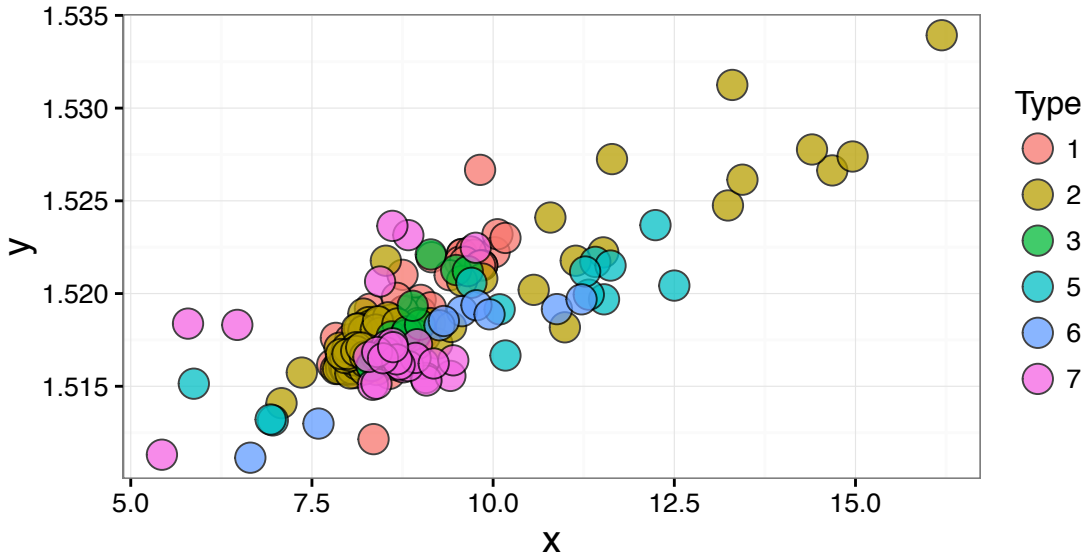

Figure 2: Relationship Between Reflactive Index and Calcium Ion.

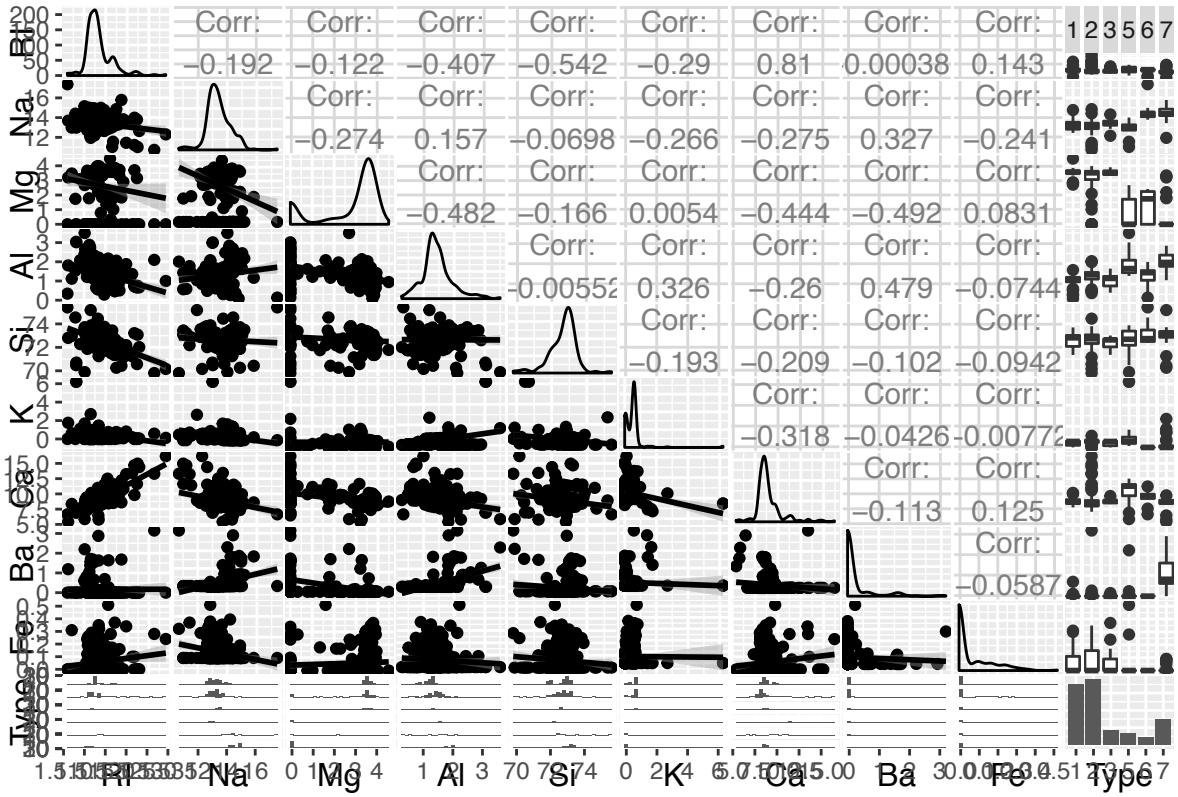

Figure 3: Predictor Correlations

```
lambda.opt
```

```
##          RI          Na          Mg          Al          Si           K
##  1.00000000  0.05045353  5.06060897  0.48445306  9.66790955  0.40864153
##          Ca          Ba          Fe
## -0.85935912  0.24494315  0.60028054
```

- Optimum values of lambda are numbers with decimal points. It is very important to select a rounded number which could describe the model physically. Noises and erorrs involved in data collection and registration may suggest this values mathematically but a good analyzer always select more logical and simple model.

1c)

Pricipal component analysis shows that, we can reduce the dimension from 9 variables into 5 variables keeping about 90% of the variance of data. More investigation on the data and PCs by *ggbiplot* indicates that ***Silisium***, ***Calcium***, and ***Refractive Index*** narates smae story and could be collaped on a single component. They may have different direction but are in same line.

1d)

PCA is an unsupervised method which reduces the problem dimensions keeping data variance as much as possible. On the other hand, LDA is a supervised method which uses the given data to produce a simpler

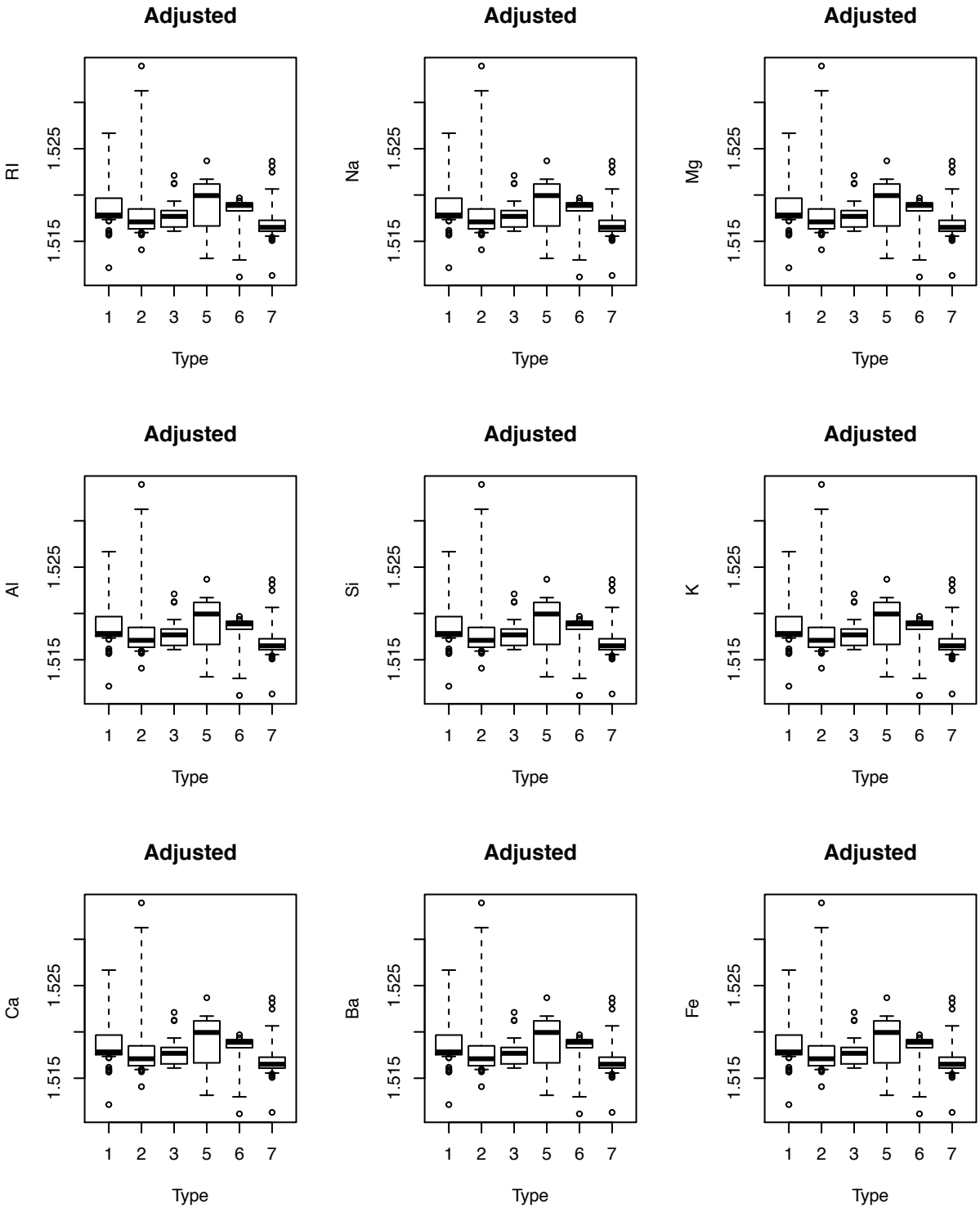

Figure 4: Adjusted Boxplots for Glass Data

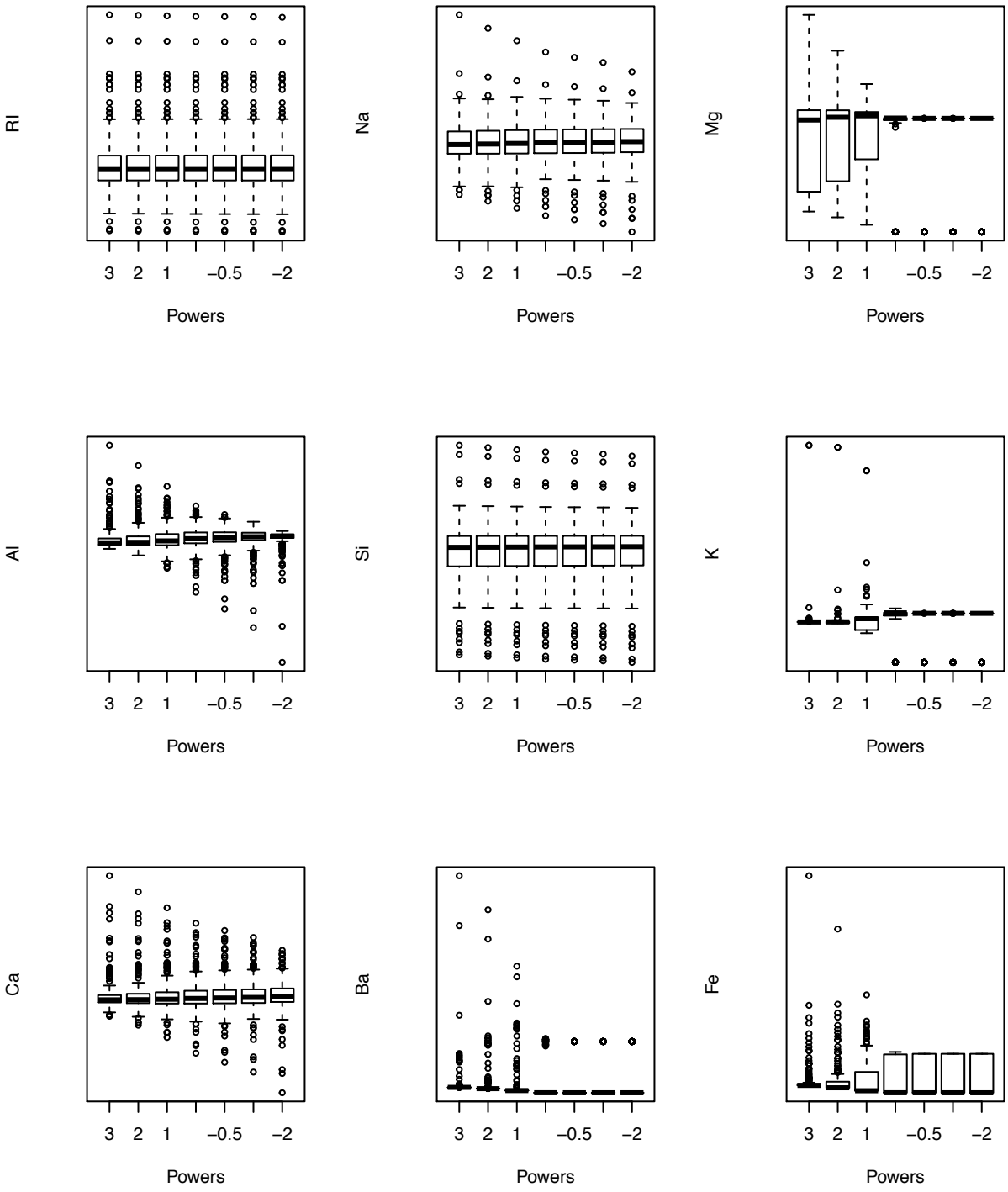

Figure 5: Symbox Function Results on Galss Data

and more accurate model. Here, in our data, table shows that LDA did a very good job in classification of group 7, but its performance was tribble in group 3. For other gruops, the result could acceptale depending on our desire.

Problem 2

2a)

Here is the code for regression using listwise deletetion. The results and outputs are nor shown but are available in R code attached to this write down.

```
freetrade.LD <- na.omit(freetrade)  #listwise deletion
freetrade.LD.fit <- lm(data = freetrade.LD,
                   tariff~year+country+polity+pop+gdp.pc+intresmi+signed+fiveop+usheg)
```

2b)

Following code was provided to perform regression after using mean imputation:

```
freetrade.mimp <- freetrade
freetrade.mimp[is.na(freetrade.mimp$tariff), "tariff"] <- mean(freetrade.mimp$tariff,na.rm=T)
freetrade.mimp[is.na(freetrade.mimp$polity), "polity"] <- mean(freetrade.mimp$polity,na.rm=T)
freetrade.mimp[is.na(freetrade.mimp$intresmi), "intresmi"] <- mean(freetrade.mimp$intresmi,na.rm=T)
freetrade.mimp[is.na(freetrade.mimp$signed), "signed"] <- mean(freetrade.mimp$signed,na.rm=T)
freetrade.mimp[is.na(freetrade.mimp$fiveop), "fiveop"] <- mean(freetrade.mimp$fiveop,na.rm=T)

freetrade.mimp.fit <- lm(data=freetrade.mimp,
                         tariff ~ year + country + polity + pop + gdp.pc
                         + intresmi + signed + fiveop + usheg)
```

2c)

Following code was provided to perform regression after using multiple imputation. Different methods such as *mean*, *rf*, *sample*, and *cart*. The lastest one has been used to perform regression. More details are availabel in R script.

```
freetrade.MI <- mice(freetrade, m=5, maxit=10, method="cart", printFlag = FALSE)
freetrade.MI.copmlete <-complete(freetrade.MI, "long")   #Complete Imputed data

freetrade.MI.fit <- with(freetrade.MI,
                         lm(tariff ~ year + country + polity +
                              pop + gdp.pc + intresmi + signed + fiveop + usheg))
```

2d)

As can be seen by using mean imputation the coeficient would not be changed. However, the coeficient of mice is highly related to the method we are choosing. Imputations based on the modeling use the model as a prediction. They may add some noises to the predicted value, but the reality is that those values are not the

missing valuse. If the missing values were missed in the heart of data set and much points are surounded the missing values, we have a lot of chance to catch a very close value to the missed one by using the models. If the missing values are part of a cut, we can not say if the predicted model on part of data works for another part in darkness. We can not say that model could be extrapolated.

```
predictorMatrix<-freetrade.MI$predictorMatrix #Extract matrix from earlier
predictorMatrix[,5] <- 0

# single imputation on different variables with different methods
freetrade$polity <- as.factor(freetrade$polity)    # Converting to factor
freetrade$signed <- as.factor(freetrade$signed)    # Converting to factor
freetrade.SI <- mice(freetrade, method=c("","","norm","polr","","","norm","logreg","norm",""),
                     predictorMatrix = predictorMatrix, printFlag = FALSE)
freetrade.SI.fit <- with(freetrade.SI,
                         lm(tariff~year+country+polity+pop+gdp.pc+intresmi+signed+fiveop+usheg))


data.frame(coef.a[,1], coef.b[,1])
```

```
##                       coef.a...1.   coef.b...1.
## (Intercept)        -2.650433e+02  1.633387e+03
## year                3.580765e-01 -7.938926e-01
## countryIndonesia   -1.900660e+02 -4.620179e+01
## countryKorea       -2.254931e+02 -5.937894e+01
## countryMalaysia    -2.318437e+02 -5.531281e+01
## countryNepal       -2.270878e+02 -4.631776e+01
## countryPakistan    -1.616933e+02 -1.440892e+01
## countryPhilippines -2.103454e+02 -5.033981e+01
## countrySriLanka    -2.168838e+02 -4.536998e+01
## countryThailand    -2.014832e+02 -4.141807e+01
## polity             -1.902494e-01 -2.111236e-01
## pop                -2.111286e-07 -2.628999e-08
## gdp.pc              2.910265e-04  5.922484e-04
## intresmi            2.929493e-01 -6.674644e-01
## signed             -1.288913e+00  2.872480e+00
## fiveop             -1.579368e+01  2.254838e+00
## usheg               9.582074e+00 -1.988981e+01
```

2e)

The listwise deletion produces best fit for the data though with very low degrees of freedom.For pooled regression using multiple imputation, the countries Korea and Pakistan, Polity and gdp.pc have the most significant coefficients.For pooled regression using single imputation, the countries Pakistan and Thailand, Polity and fiveop have the most significant coefficients.

2f)

I would say listwise deletion showed the best results amoung the other methods. Therefore, I would suggest to use this method istead of imputation.

Problem 3

3a)

The goal of this research is to identify a vehicle automatically based on its weight, number of axels, aerodynamics etc. Sensors on the brige record the vibrations in a very small time snaps. There are several thought here may help to understand and simplify the problem: * Looking into the available daya shows taht when a vehicle passes the brigde, we can distinguish two frequency peaks. They could be representative of vehicle's axels. Because the axels and attached tires are the points that touch the bridge and make the vibration. * The frequency is also a charateristics which depends on the vehicle's weight and speed. The hevier vehicle will generate stronger waves. It also could be an indicator for speed of a specific vehicle. * The difference between the start and end time of a wave may imply on the length and speed of the vehicle. * The frequency itself could be related to a specific vehicle. It could be a characteristic signal of a vehicle. * The ratio of two peaks also might be important.

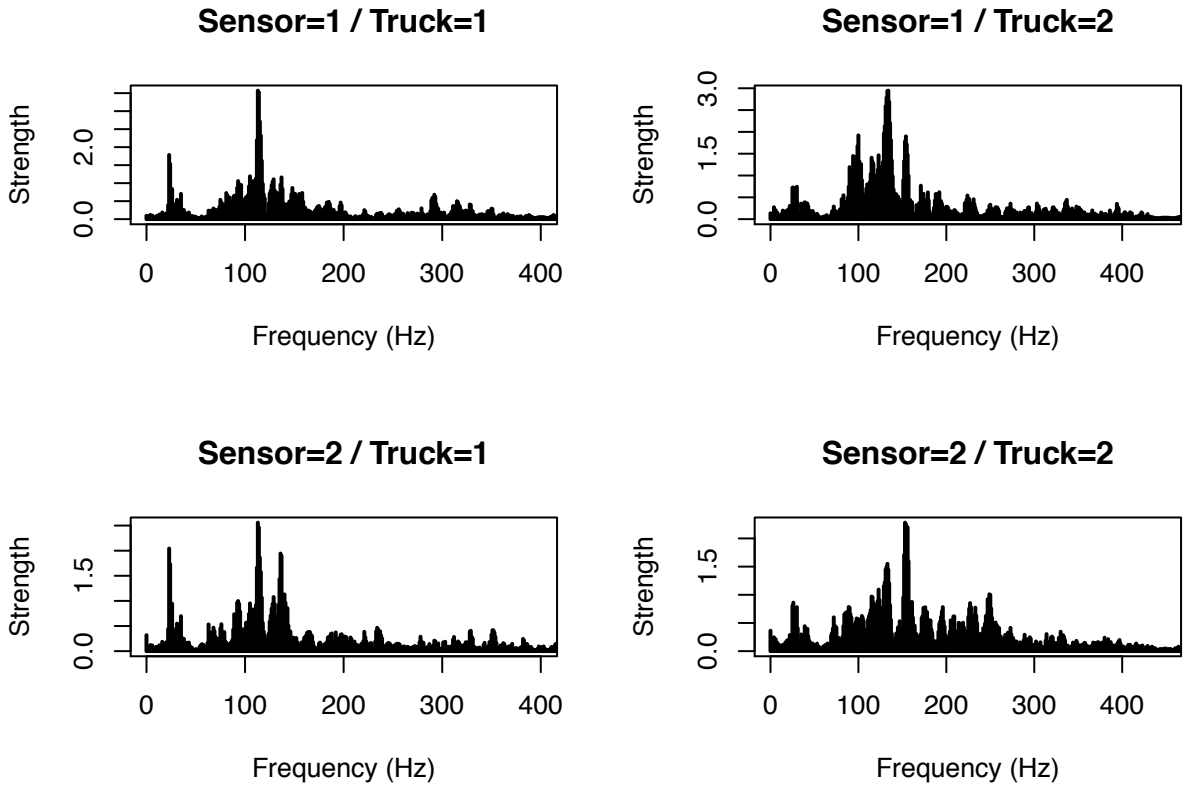

Figure 6: Wave Strength vs. Frequency

3b)

```
#1.
#Sensor 1 / Truck 1
max(abs(Sensor11$Sensor1))
#Sensor 1 / Truck 2
max(abs(Sensor12$Sensor1))
#Sensor 2 / Truck 1
max(abs(Sensor21$Sensor2))
```

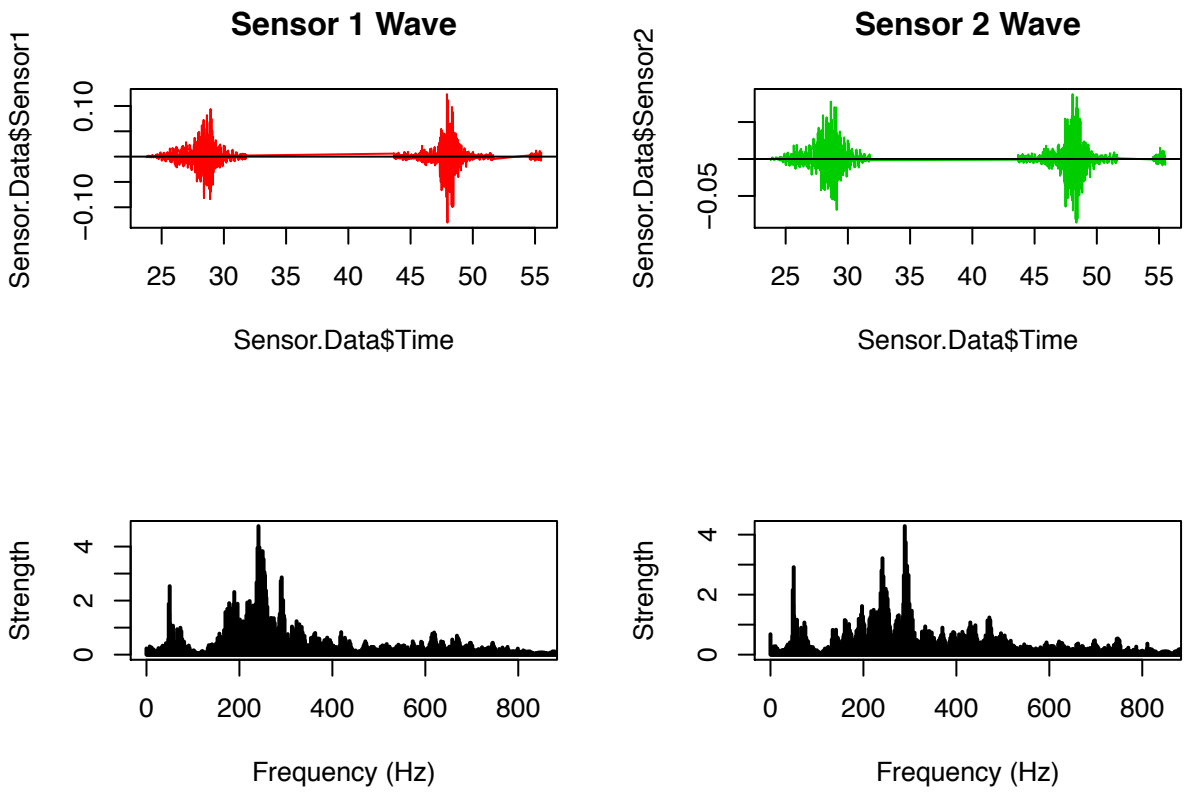

Figure 7: Wave Signals

```
#Sensor 2 / Truck 2
max(abs(Sensor22$Sensor2))

#from the maximum absolute value above it can be understand that the Truck 1 has less weight than Truck

#2.
#Sensor 1 / Truck 1
max(PS11[,2])
#Sensor 1 / Truck 2
max(PS12[,2])
#Sensor 2 / Truck 1
max(PS21[,2])
#Sensor 2 / Truck 2
max(PS22[,2])

#Using the maximum value of the furior the same result can be drawn which Truck 1 is heavier than Trcuk

#3.
#Sensor 1 / Truck 1
which.max(PS11[1:(length(PS11[,2])/2),2])
#Sensor 1 / Truck 2
which.max(PS12[1:(length(PS12[,2])/2),2])
#Sensor 2 / Truck 1
which.max(PS21[1:(length(PS21[,2])/2),2])
#Sensor 2 / Truck 2
```

```r
which.max(PS22[1:(length(PS22[,2])/2),2])

#4.
Sensor.Data$strength1 <- Mod(sens1fft) #strength1 for sensor 1
Sensor.Data$strength2 <- Mod(sens2fft) #strength2 for sensor 2
Sensor.Data$ang1 <- Arg(sens1fft) #fft angle for sensor 1
Sensor.Data$ang2 <- Arg(sens2fft) #fft angle for sensor 2
#interaction between strength and frequency
Sensor.Data$str.fre1 <- (Sensor.Data$Time-Sensor.Data$Time[1])*Sensor.Data$strength1*100
Sensor.Data$str.fre2 <- (Sensor.Data$Time-Sensor.Data$Time[1])*Sensor.Data$strength2*100
head(Sensor.Data,7)
```

3c)

The bigest problem in this part was lack of information. We just had two observations where collected from an unknown bridge. For the data analysis problems which follows the statistical rules and techniques, we need more observations and data records to build a model and analysis data.

Data shortage hindered us to find a pattern in data set. It also was the main reason that we were not able to use supervised method.