MBA6693

Airquality

GS Kumbhare
g.kumbahre@unb.ca

Objective

Objective is to find best classification model that can fit air quality data set. We also took relationship between error rates and model complexity. We also investigate relationship between several predictor and response variable.

Data description

The data set we have is obtained from R directory of datasets. We obtained the Airquality data for a period of 5 month. In total we have 153 instances in the dataset. In this the class is months number.

Attributes of data

1. Ozone level
2. Solar radiation
3. Windspeed
4. Temperature
5. Month
6. Day

Libraries

Libraries needed for classification model

```
library(ggthemes)

library(ggplot2)

library(caret)

library(ggiraphExtra)

library(ggplot2)

library(broom)

library(readr)

library(MASS)

library(e1071)

library(nnet)

library(corrplot)

library(tidyverse)

library(car)
```

Load Data

datasets::airquality

Once the library is loaded we find the data sets structure using

```
> str(airquality)
'data.frame':	153 obs. of  6 variables:
 $ Ozone  : int  41 36 12 18 NA 28 23 19 8 NA ...
 $ Solar.R: int  190 118 149 313 NA NA 299 99 19 194 ...
 $ Wind   : num  7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...
 $ Temp   : int  67 72 74 62 56 66 65 59 61 69 ...
 $ Month  : int  5 5 5 5 5 5 5 5 5 5 ...
 $ Day    : int  1 2 3 4 5 6 7 8 9 10 ...
```

We observe that there are NA values in the dataset

We need to clean it,

We use na.omit() function to omit NA

```
> str(na)
'data.frame':	111 obs. of  6 variables:
 $ Ozone  : int  41 36 12 18 23 19 8 16 11 14 ...
 $ Solar.R: int  190 118 149 313 299 99 19 256 290 274 ...
 $ Wind   : num  7.4 8 12.6 11.5 8.6 13.8 20.1 9.7 9.2 10.9 ...
 $ Temp   : int  67 72 74 62 65 59 61 69 66 68 ...
 $ Month  : int  5 5 5 5 5 5 5 5 5 5 ...
 $ Day    : int  1 2 3 4 7 8 9 12 13 14 ...
```

After cleaning up data we summarise the remaining datset

```
> summary(na)
     Ozone           Solar.R     
 Min.   :  1.0   Min.   :  7.0  
 1st Qu.: 18.0   1st Qu.:113.5  
 Median : 31.0   Median :207.0  
 Mean   : 42.1   Mean   :184.8  
 3rd Qu.: 62.0   3rd Qu.:255.5  
 Max.   :168.0   Max.   :334.0  
      Wind            Temp      
 Min.   : 2.30   Min.   :57.00  
 1st Qu.: 7.40   1st Qu.:71.00  
 Median : 9.70   Median :79.00  
 Mean   : 9.94   Mean   :77.79  
 3rd Qu.:11.50   3rd Qu.:84.50  
 Max.   :20.70   Max.   :97.00  
     Month            Day       
 Min.   :5.000   Min.   : 1.00  
 1st Qu.:6.000   1st Qu.: 9.00  
 Median :7.000   Median :16.00  
 Mean   :7.216   Mean   :15.95  
 3rd Qu.:9.000   3rd Qu.:22.50  
 Max.   :9.000   Max.   :31.00  
```

We split the dataset into training and testing sets with an ratio of 80:20. Our training set is divided on the basis of month number.

```
#we will split the dataset into subset of 80:20(trainging:validation)
split_data <- createDataPartition(na$Month, p=0.8, list=FALSE)
testset <- na[-split_data,]
trainset <- na[split_data]
```

Summarizing the training dataset

```
> summary(trainset)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1.00   18.00   29.50   39.67   57.25  168.00 
> 
```

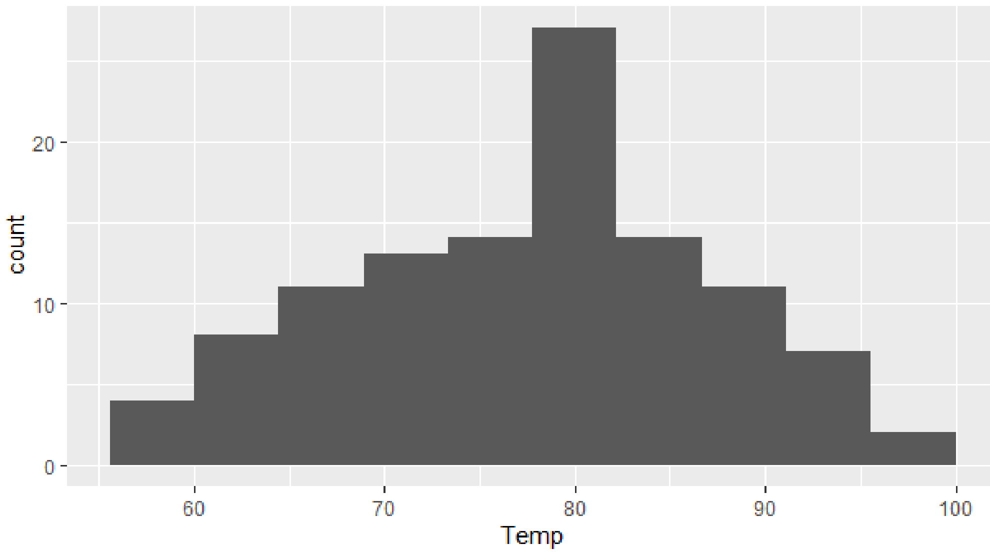
Charts

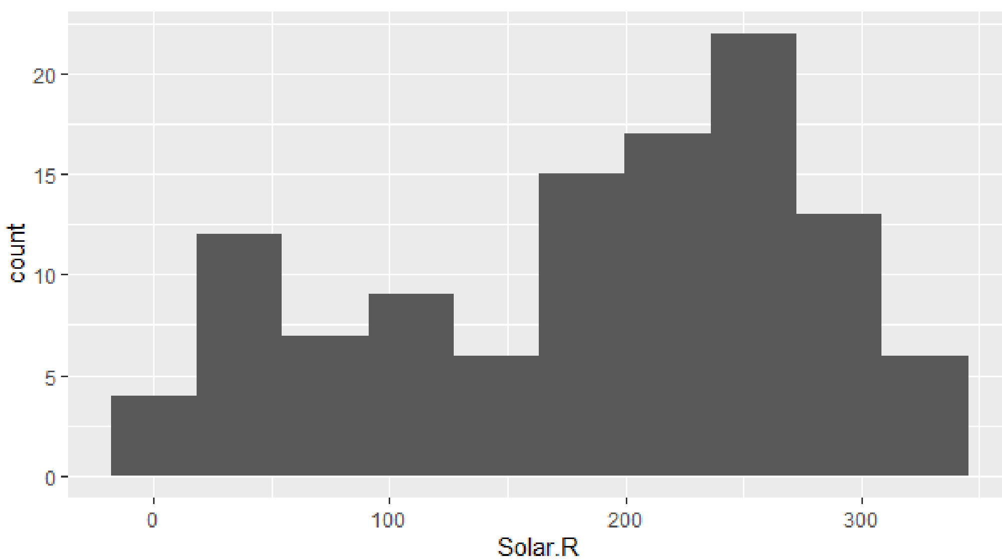
A. Histogram

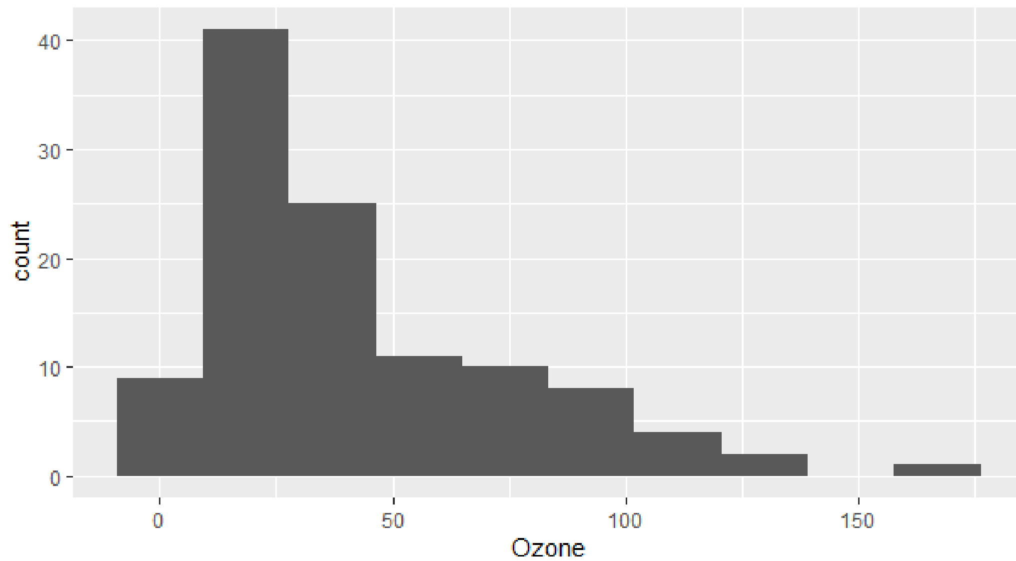
Histogram for each 4 predictor and their distribution

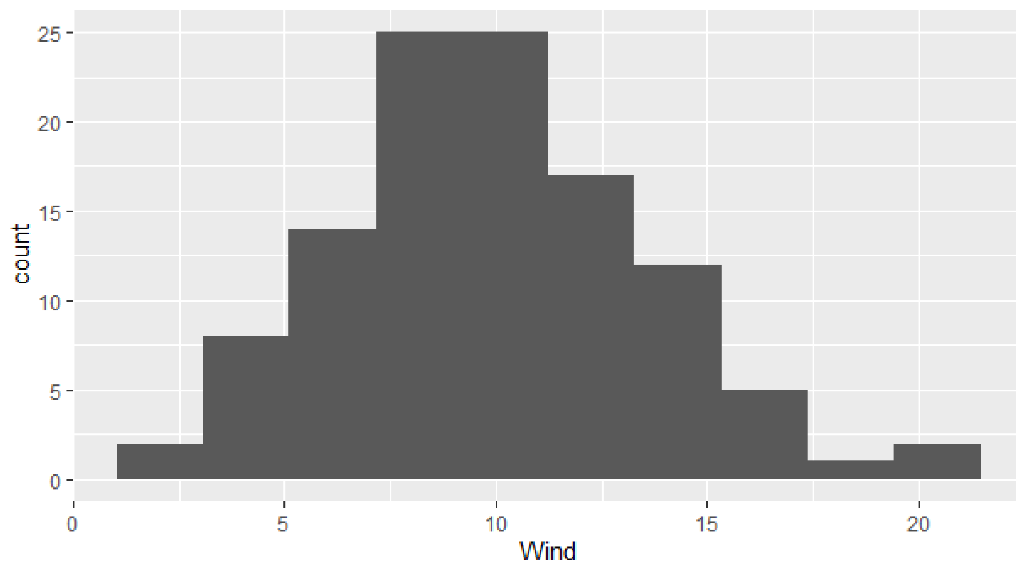
```
#We plot histogram
ggplot(data = na,mapping =  aes(Ozone))+geom_histogram( bins = 10)
ggplot(data = na,mapping =  aes(Temp))+geom_histogram(bins = 10)
ggplot(data = na,mapping =  aes(Wind))+geom_histogram(bins = 10)
ggplot(data = na,mapping =  aes(Solar.R))+geom_histogram(bins = 10)
```

Histogram of Temperature



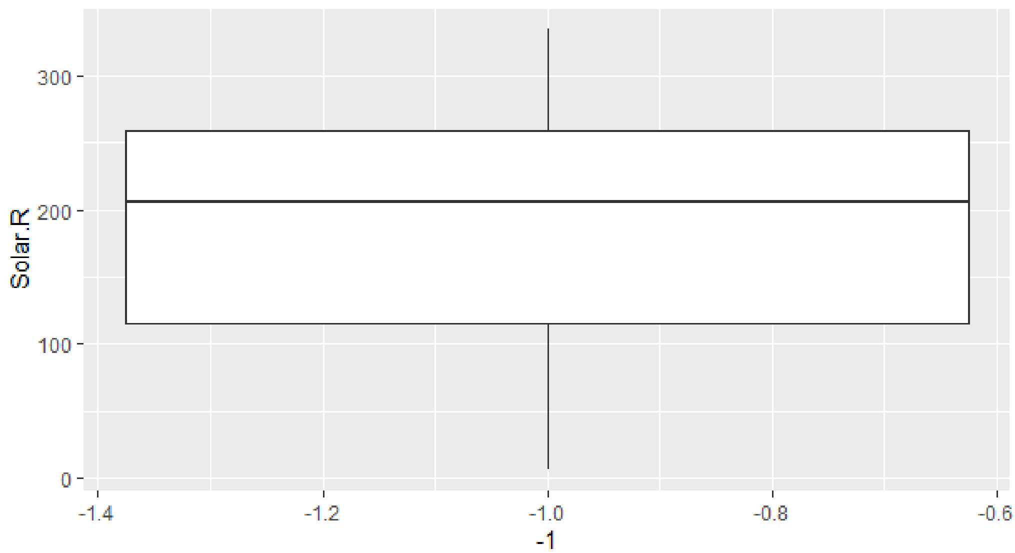




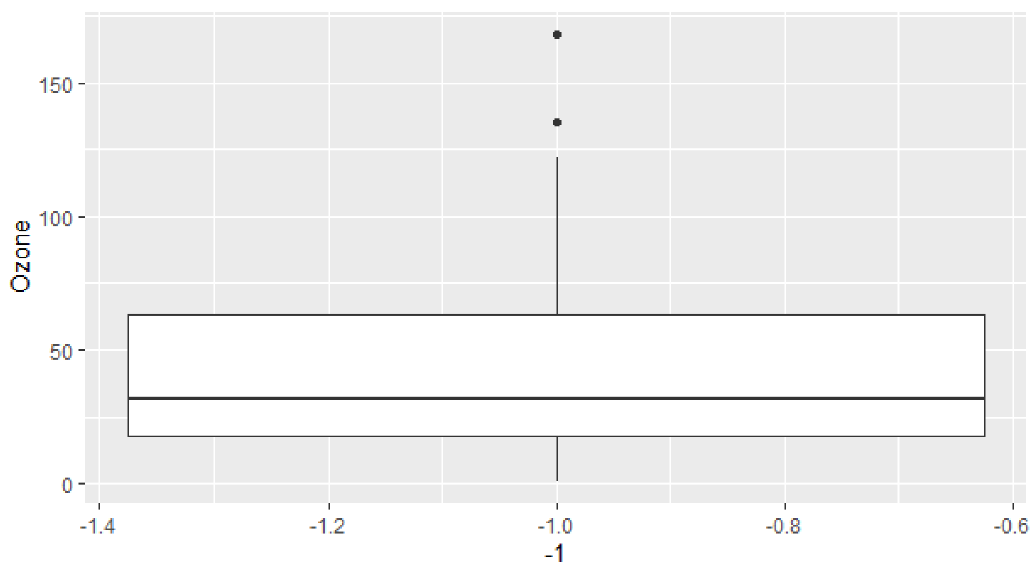


B. Boxplot for each variables

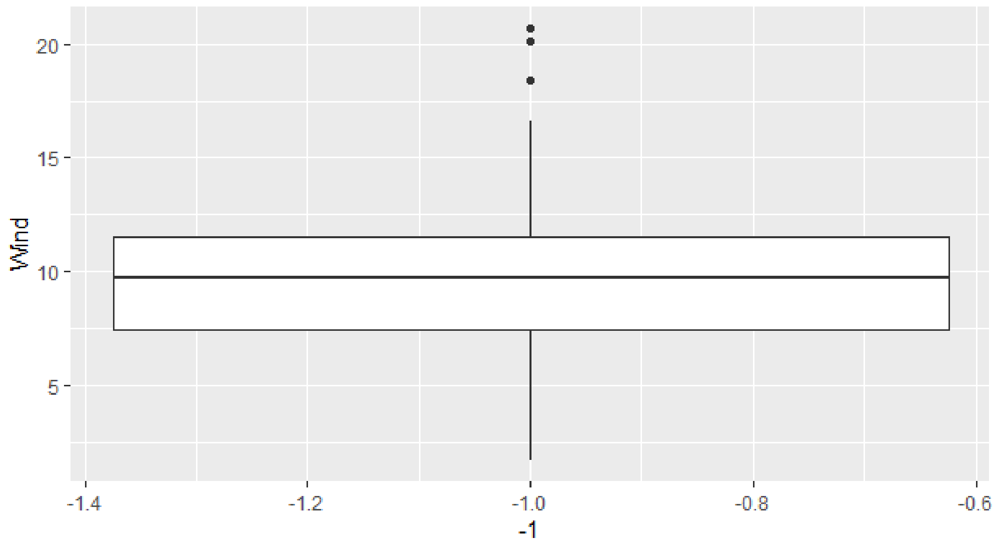
```
#We plot boxplpot
ggplot(data = airquality,mapping =  aes(-1,Ozone))+geom_boxplot()
ggplot(data = airquality,mapping =  aes(-1,Temp))+geom_boxplot()
ggplot(data = airquality,mapping =  aes(-1,Wind))+geom_boxplot()
ggplot(data = airquality,mapping =  aes(-1,Solar.R))+geom_boxplot()
```

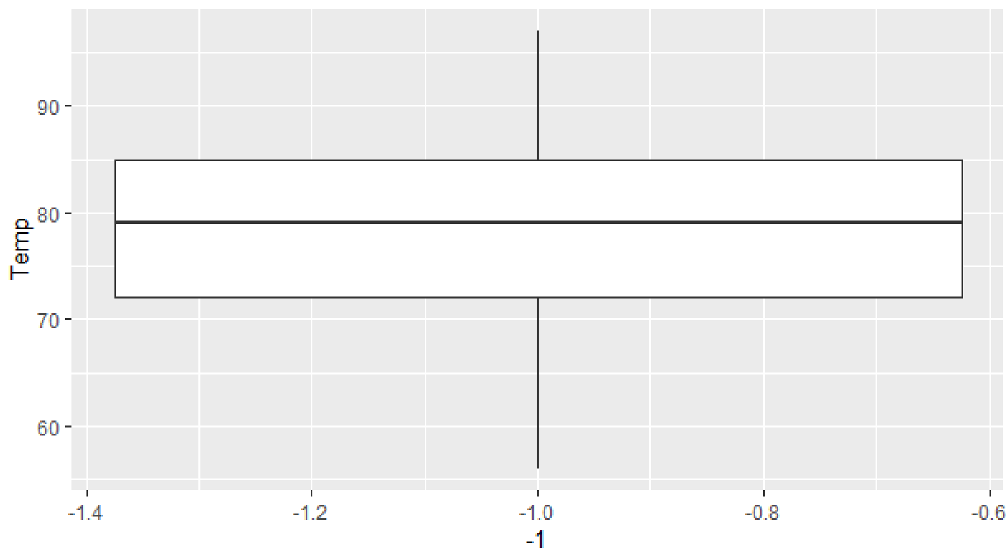


300
200
100
0
Solar.R
-1.4
-1.2
-1.0
-0.8
-0.6
-1



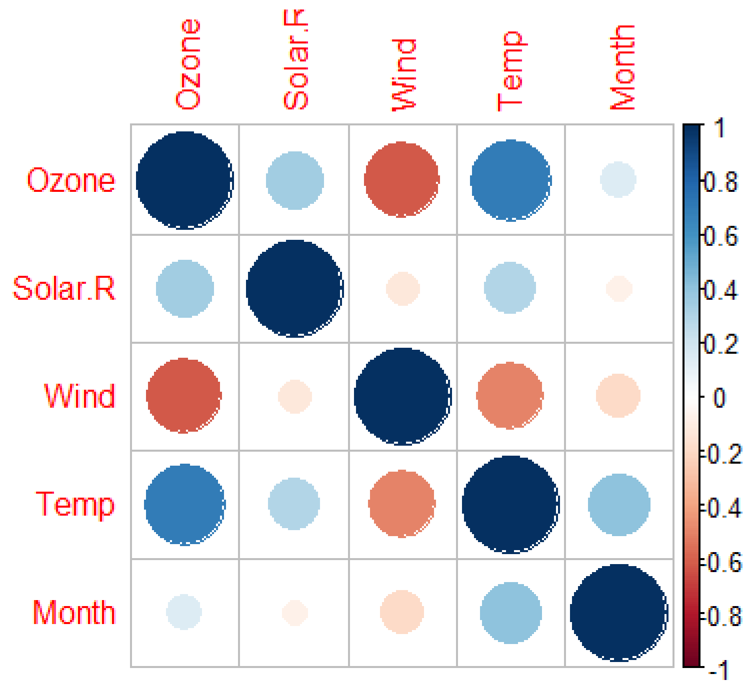
150
100
50
0
Ozone
-1.4
-1.2
-1.0
-0.8
-0.6
-1





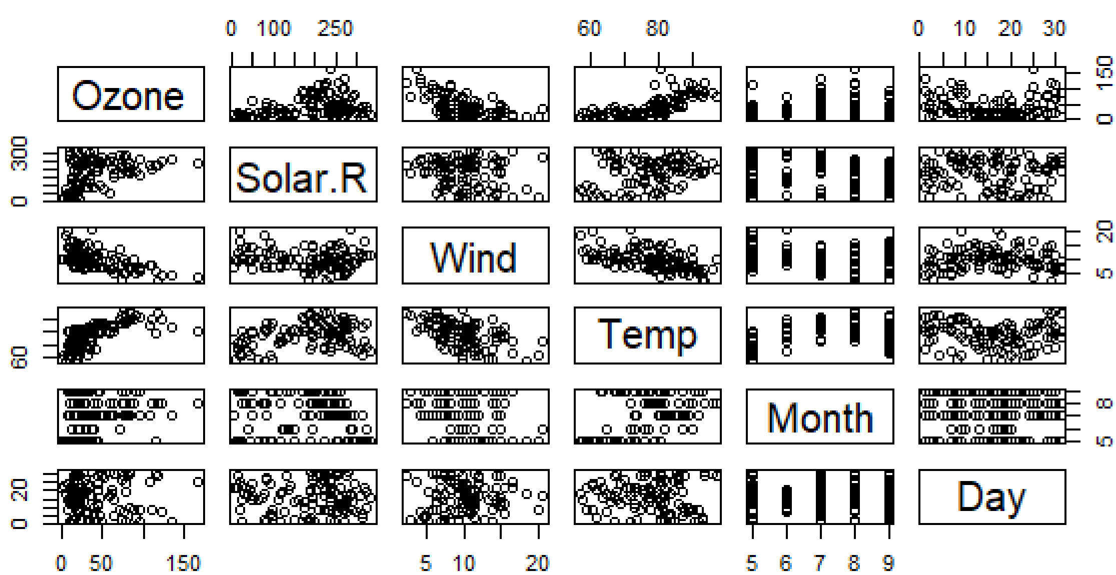
C. Correlation

```
#we find correlation between the variables
correlations <- cor(na[,1:5])
corrplot(correlations, method = "circle")
```



D. Overall plot

```
#overall plot
plot(na)
```



Models

a. **Logistic Regression model**

1. **Model with 2 predictors are Temperature and wind speed**

```
#we will be building some plots using multinomial logistic Regression,
#linear Discriminant Analysis and K-nearest Neighbor
#Losgistic regression
#We use logistic regression with two predictor
#1.first predictors are Wind and Temp according to month
log_fit1=multinom(Month~Temp+Wind, data=na)
print(log_fit1)
```

```
Call:
multinom(formula = Month ~ Temp + Wind, data = na)

Coefficients:
  (Intercept)      Temp       Wind
6   -24.45673 0.2825699 0.26055271
7   -26.49065 0.3508314 0.02329945
8   -27.22940 0.3545852 0.05800654
9   -15.50192 0.2137201 0.04011390

Residual Deviance: 269.7798
AIC: 293.7798
```

2. **Model with 2 predictors are Ozone layer and Solar radiation**

```
#2.Second predictor are Ozone and Solar radiation according to month
log_fit2=multinom(Month~Ozone+Solar.R, data = na)
print(log_fit2)
```

```
Call:
multinom(formula = Month ~ Ozone + Solar.R, data = na)

Coefficients:
  (Intercept)      Ozone       Solar.R
6 -1.28082530 0.01678858 -0.0007970226
7 -1.96226715 0.05233132  0.0004868905
8 -0.86248972 0.06045527 -0.0082756481
9  0.07253269 0.02442115 -0.0031563279

Residual Deviance: 309.1981
AIC: 333.1981
```

3. Model With all predictors

```
#3.Next we use model logit with all predictors
log_fit_all=multinom(Month~Ozone+Solar.R+Wind+Temp, data = na)
print(log_fit_all)
```

```
Call:
multinom(formula = Month ~ Ozone + Solar.R + Wind + Temp, data = na)

Coefficients:
  (Intercept)        Ozone      Solar.R        Wind      Temp
6   -29.31985 -0.042873518 -0.006417246  0.20878475 0.3892660
7   -28.16260 -0.021880093 -0.003396519 -0.01495137 0.3984670
8   -28.83435 -0.004756026 -0.014336353  0.10560787 0.4058014
9   -19.02265 -0.036004438 -0.006715329 -0.02898424 0.3036300

Residual Deviance: 253.1663
AIC: 293.1663
```

Observation

In prediction model 3 with all predictor has highest accuracy and lowest error

b. Linear regression

```
Modelm <- lm(Month~Ozone+ Solar.R+ Temp+ Wind, data = na)
print(Modelm)
```

```
Call:
lm(formula = Month ~ Ozone + Solar.R + Temp + Wind, data = na)

Coefficients:
(Intercept)        Ozone      Solar.R         Temp         Wind
   1.339740    -0.011846    -0.002709     0.092702    -0.033780
```

Observation

Every month there are variable changes in ozone layer, solar radiation, temperature, and wind.

c. Anova

```
> Anova(Modelm)
Anova Table (Type II tests)

Response: Month
           Sum Sq  Df F value    Pr(>F)
Ozone       6.736   1  3.8946   0.05104 .
Solar.R     5.765   1  3.3327   0.07073 .
Temp       42.939   1 24.8252 2.455e-06 ***
Wind        0.962   1  0.5562   0.45744
Residuals 183.344 106
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Conclusion

We modeled logistic regression and linear regression it is observed that logistic regression is best fit for analysing multivariate data due to which