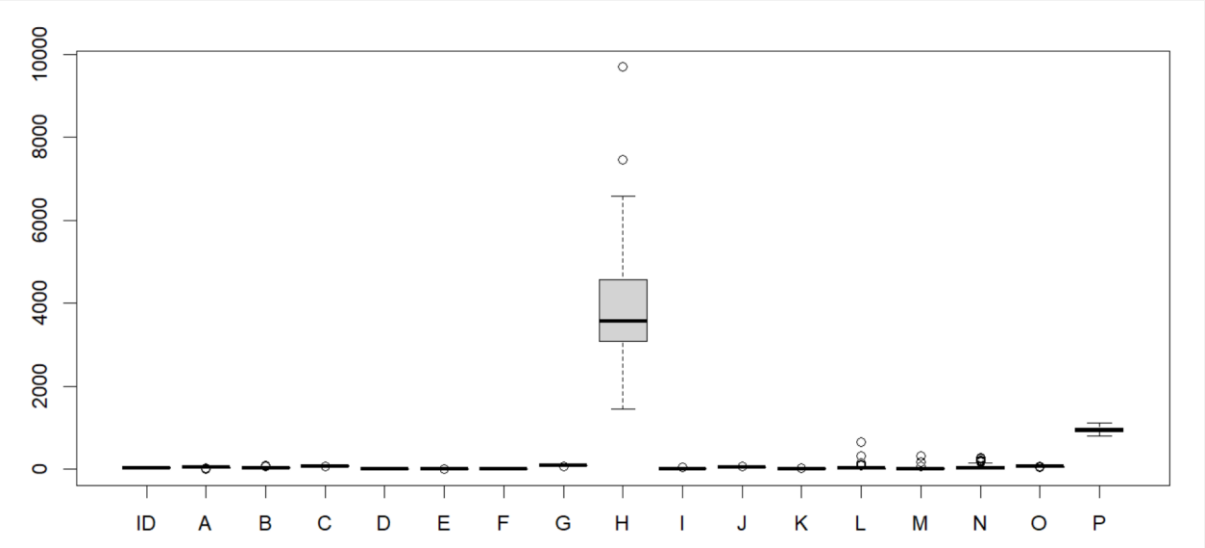
Abhishek Vyas – m2021anlt001

Predicting the death rate in city ABCD

1. **Examine the data as a large part of regression model.**
2. **Formulate a statistical model - select dependent and independent variables of your choice (build an Ordinary Least Squares multiple regression model to predict death rate in the city ABCD).**

- After examining the data it was observed that there are no null values, and some of the variables were observed to be having outliers as seen in the box-plot visualisation. After selecting the relevant variables we will deal with outliers of those particular variables.

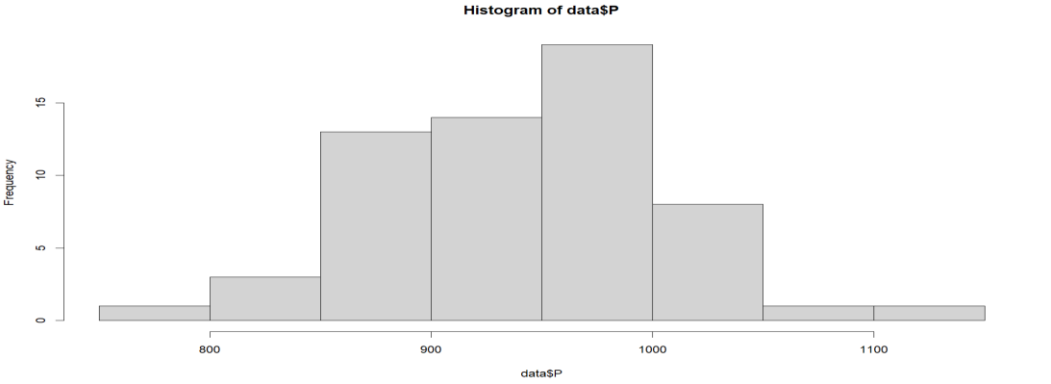


- As per my understanding there are few variables which can be of significant relevance to the dependant variable, i.e., Death rate. Those variables are :

ID	Variable
D	percent of 2015 ABCD population 70 years old or older
I	percent nonwhite population
K	poor families (annual income under $3000)
L	relative pollution potential of hydrocarbons
M	relative pollution potential of oxides of Nitrogen
N	relative pollution of Sulfur Dioxide

- As there can arise difficulties when selecting too many independent variables together such as the variables can be correlated to one another. To avoid this we will select 3-4 variables only from above table. And the dependant variable will be the Death rate that is “P”.

- When we made a histogram of the dependant variable, death rate is almost normally distributed, which pacifies one of the assumptions for OLS. See below:



After trying multiple combinations of independent variables I have decided to go forward with a combination of 3 Independent variable, namely, L+M+I. Where**, L = relative pollution potential of hydrocarbons, H =population per square mile in urbanized areas, I = percent nonwhite population**. Below is the summary of the model:

```
Call:
lm(formula = P ~ L + H + I, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-101.019  -23.265    0.014   29.763   92.053

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 843.835965  18.048550  46.754  < 2e-16 ***
L            -0.132410   0.062791  -2.109   0.0395 *
H             0.012517   0.003971   3.152   0.0026 **
I             4.465371   0.642783   6.947 4.24e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 44.03 on 56 degrees of freedom
Multiple R-squared:  0.5246,    Adjusted R-squared:  0.4991
F-statistic:  20.6 on 3 and 56 DF,  p-value: 4.039e-09
```

➔ **<u>Discussion on above model summary:</u>**

- From the summary above we see that the residual values are pretty decent, we also have a median residual which is very close to 0 & after we calculated the mean of the residuals it turned out to be almost 0 (2.034252e-15) which satisfies one of the assumptions. We can see that all the independent values selected, and the Intercept are significant.
- We also tried removing the outliers from the model, but it led to a drop in the R2 value, so the data still contains the all the rows.
- ******While interpreting the estimate values we saw that L(**relative pollution potential of hydrocarbons**) has a negative effect or we can say that it reduces the death rate. Maybe it signifies that households with more hydrocarbon are probably in cities with more cars etc which

might lead to release of more hydrocarbons and maybe more facilities as compared to villages where although hydrocarbon pollution is less, but the amenities such as advanced hospitals might not be there in villages as compared to cities.*****

- Standard error as seen in the summary are pretty low (almost 0), except for the intercept.
- If we take 0.05 as a benchmark we observe that all the coefficients are significant.
- Adjusted R2 is lower than Multiple R2, and there is not a big difference between them so we can say that we don't have the issue of model overfitting. Although the R2 value is a little lower.
- Lastly, the F statistic shows that model as a whole is significant, since we have a p-value which is 4.039e-09.

3. Estimate the parameters of the model you chose in (2) and

Write out the following

i. Examine and interpret the relationships between variables.

- After trying with multiple combinations of independent variables we saw that variables L,H and I have a very low correlation with each other and all the 3 variables have a certain reasoning to be included in the model. In the Question 4 we will explore some more ways to look at relationship and multicollinearity among the variables.

```
> cor(data3)
             L           H           I
L  1.00000000  0.12028225 -0.02600206
H  0.12028225  1.00000000 -0.00552433
I -0.02600206 -0.00552433  1.00000000
```

ii. What is the value R^2 and what does it tell you? Paste what you read in multicollinearity chapter.

- We have values of R2 and Adjusted R2. Since in multiple regression as we increase the number of independent variables the Multiple R2 tends to increase, so better choice is looking at the Adjusted R2. We have adjusted R2 of 0.4991 which is smaller than the multiple R2, as it should be. This means that almost 50% of the movement can be explained by the independent variables.

iii. Perform t-test (two-sided) (α at 5 percent). What does it convey?

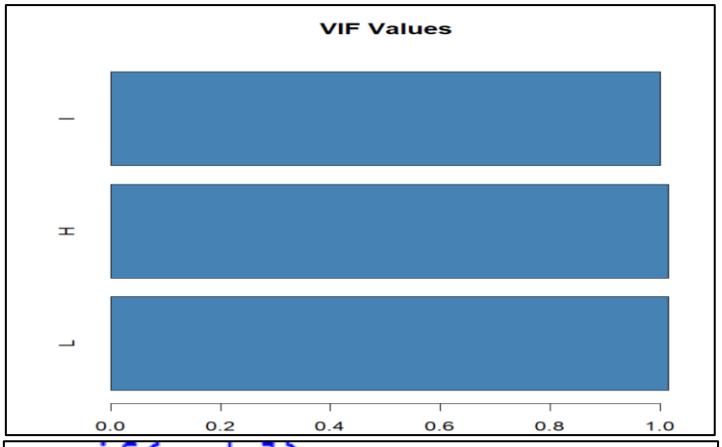
- From the summary table, we can see that even if we take the level of significance to be 5%, we will still have independent variables which are significant. Based on the standard error and t-test in the summary of the model, we get the p-values. These p-values are supposed to tell us whether the estimates of the intercept and the slopes are 0 or not. If they are 0 then we won't have much use of the Independent variables.
- For intercept although we have a significant p-value, but in case if it isn't significant we can still proceed with the, it just means that effect of all omitted variables may not be important.

4. Carry out the necessary diagnostics – (multicollinearity, homoscedasticity and autocorrelation) to convince that the model assumptions are met (or not met) to a reasonable degree.

- After running the gylma() function on our model we saw that all the assumptions were satisfied, where level of significance is 5%.

```
> gvlma(model)

Call:
lm(formula = P ~ L + H + I, data = data)

Coefficients:
(Intercept)            L            H            I
  843.83596     -0.13241      0.01252      4.46537


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = model)

                     Value p-value                Decision
Global Stat        2.14831  0.7085 Assumptions acceptable.
Skewness           0.46455  0.4955 Assumptions acceptable.
Kurtosis           0.02586  0.8722 Assumptions acceptable.
Link Function      0.25807  0.6114 Assumptions acceptable.
Heteroscedasticity 1.39982  0.2368 Assumptions acceptable.
```

Multicollinearity –



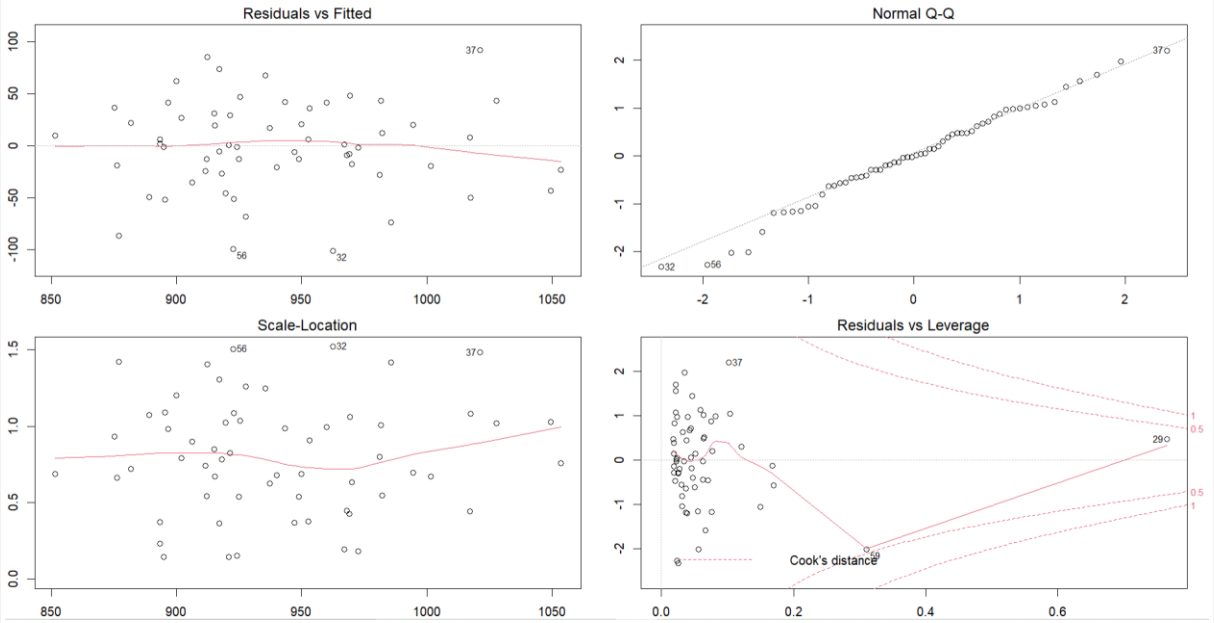

```
> vif(model)
       L        H        I
1.015342 1.014686 1.000682
```

From the above graph and VIF values we can deduce that there is no issue of multicollinearity in our model. As the VIF values are close to 1. A general rule of thumb would be to start questioning the model with VIF values of IV around 5 or greater than 5. Hence, no remedial measures are required here in our model.

Homoscedasticity –

We can use scale-location graphical analysis, White test, BP test for identifying if there is Homoscedasticity or not. In the scale location graph, if the red line is somewhat straight and the residual dots do not have any pattern.

But lets do a




```
> white_lm(model)
# A tibble: 1 x 5
  statistic p.value parameter method       alternative
      <dbl>   <dbl>     <dbl> <chr>        <chr>
1      2.23   0.898         6 White's Test greater
```

- By running the white_lm() function on our model we observed that p-value is very high and hence we fail to reject the null hypothesis. Null hypothesis being Homoscedasticity is present.

Autocorrelation –

```
> # autocorrelation test - DW test
> dwtest(model)

        Durbin-Watson test

data:  model
DW = 1.7046, p-value = 0.1221
alternative hypothesis: true autocorrelation is greater than 0

> # Durbin's M test(as p=1)
> bgtest(P ~ H + I + L, order=1, data=data)

        Breusch-Godfrey test for serial correlation of order up to 1

data:  P ~ H + I + L
LM test = 1.3682, df = 1, p-value = 0.2421

>
> bgtest(P ~ H + I + L, order=35, data=data)

        Breusch-Godfrey test for serial correlation of order up to 35

data:  P ~ H + I + L
LM test = 45.011, df = 35, p-value = 0.1197
```

- After observing the results from BG-test and DW-test we cannot reject the null hypothesis [H0 (null hypothesis): There is no autocorrelation], which means we do have autocorrelation in our model.
- As we know that a drawback of the BG test is that the value of p, the length of the lag, cannot be specified a priori. Some experimentation with the p value is inevitable.
-
- After doing multiple tests for autocorrelation which is recommended, we conclude that true autocorrelation is present which is greater than 0. Although the p-value is very close to 10%. There are ways to fix the autocorrelation issue if the issue is too concerning. Based on whether we know the value of ρ or not we can conduct required tests to address the issue.

Finally, we can conclude that apart from issue of autocorrelation and low R2 value our model satisfies the other assumptions.

Abhishek Vyas – m2021anlt001

- **Code from R library**

```
#reading the excel file


install.packages("readxl")

library("readxl")

library("dplyr")

library("magrittr")

data <- read_xlsx("C:/Users/dell/Desktop/Sem 2/Multivariate Predictive Analysis I/assignment/xyz.xlsx",sheet=1)

View(data)

summary(data)


# checking for assumptions

# no null values detected

which(is.na(data))


#checking for outliers

boxplot(data)

#we can see some columns have the outliers which are to be taken care of once selected for the model


#checking for normality of dependant variables

hist(data$P)

#it can be said that there is a certain normality in the depandant variable


#Forming new table with relevant columns

data1 <- data %>% select(P,H,I,L)




#Create a boxplot that labels the outliers

# install the package
```

Abhishek Vyas – m2021anlt001

```
#install.packages("ggstatsplot")
# Load the package
#library(ggstatsplot)
#ggbetweenstats(warpbreaks, wool, breaks, outlier.tagging = TRUE)
#?ggbetweenstats

#treating outliers for independant variables 'L' & 'N'
Q <- quantile(data1$L, probs=c(.25, .75), na.rm = FALSE)
iqr <- IQR(data1$L)
up <-  Q[2]+1.5*iqr # Upper Range
low<- Q[1]-1.5*iqr # Lower Range
data2<- subset(data1, data1$L > (Q[1] - 1.5*iqr) & data1$L < (Q[2]+1.5*iqr))
boxplot(data2)
View(data2)

Q <- quantile(data2$H, probs=c(.25, .75), na.rm = FALSE)
iqr <- IQR(data2$H)
up <-  Q[2]+1.5*iqr # Upper Range
low<- Q[1]-1.5*iqr # Lower Range
data2<- subset(data2, data2$H > (Q[1] - 1.5*iqr) & data2$H < (Q[2]+1.5*iqr))
boxplot(data2)
View(data2)



#fit the regression model
model <- lm(P ~ L + H  + I, data = data)

#view the output of the regression model
summary(model)
mean(model$residuals)
#define the variables we want to include in the correlation matrix
```

Abhishek Vyas – m2021anlt001

```
data3 <- data[ , c( "L", "H", "I","P")]
boxplot(data3)
#create correlation matrix
cor(data3)



#test for multicollinearity

library(car)
vif(model)
vif_values <- vif(model)
barplot(vif_values, main = "VIF Values", horiz = TRUE, col = "steelblue")

#Homoscedasticity test

plot(model)

install.packages("gvlma")
library(gvlma)
gvlma(model)

white_lm(model)



# autocorrelation test - DW test
dwtest(model)
# Durbin’s M test(as p=1)
bgtest(P ~ H + I + L, order=1, data=data)

bgtest(P ~ H + I + L, order=35, data=data)
```