Air Pollution to Mortality: A Case Study Using Regression

Abir Naha (201257)
Anjan Kumar Kayal (201271)
Arkonil Dhar (201279)
Koyel Pramanick (201333)
Suchismita Roy (201440)

Under Guidance of
Prof. Sharmishtha Mitra,
Dept. of Mathematics and Statistics,
IIT Kanpur

ACKNOWLEDGEMENT

Our journey of accomplishing this project really involves many ones to whom we are highly obliged. We would like to express our deepest appreciation to all those who have provided us the possibility to complete this project. We give a special gratitude to our respected instructor Dr. Sharmishtha Mitra, Department of Mathematics and Statistics, IIT KANPUR, whose contribution in stimulating suggestion, valuable guidance, constructive criticism and encouragement help us to coordinate our project.

We take the privilege to thank the authors and publishers of the various books we have consulted. Also thanks to the Wikipedia and various other free website from which we got help. At last we would like to thank our seniors and batch mates for their co-operation throughout the project. Without their guidance and supervision this project would not have been completed.

Table of Contents

ABSTRACT

In this report, we have fitted a multiple linear regression model on 'air pollution.csv' dataset. The dataset contains 16 variables. Among them we have choosed the variable 'MORT' (indicates total adjusted mortality rate per 100,000) as the response variable as we know from our prior knowledge that mortality rate depends on other variables present in the dataset. We have done some basic exploratory data analysis over the data set and moved to fit multiplle linear regression model. Here after fitting multiple linear regression model on all variables initially, we have checked for various assumptions of it like normality of residuals and non-existance of heteroscedasticity, auto- correlation and multicolinearity among residuals. We have found that though our initially fitted model is satisfying other assuptions but is suffering from multicolinearity. We have taken some necessary steps to remove it and finally got a multiple linear regression model with moderately high adjusted R^2 indicating our choosen response can be satisfactorily explianed by the regressors and model finally chosen.

Introduction:

Mortality rate is a measure of the number of deaths in a population in a certain period. It is usually expressed as the number of deaths in a population of size 1000 in a specific year. As different age groups have different amount of risk of death, a single measure can be defined as the weighted average of mortality rates of all the age groups with population proportions of age groups as the weights, also known as the Age-Adjusted Mortality Rate.

Historical data indicates that average mortality rate has declined steadily since World War II as technology became more and more advanced in the developed countries. Advancement of medical science, improving living standards, many social policies undertaken to improve health standards have contributed to its cause. But the same cannot be said about the under-developed and developing nations. There are different factors affecting mortality rate. Some key factors include environmental pollution and different socio-economic problems like poverty, unemployment etc.

Pollution is the introduction of contaminants or harmful substances (also known as pollutants) into the natural environment. Pollutants can be generated by both natural causes (volcanic ash, forest fire etc.) or human activities (industrial wastes, vehicle emission etc.). Some of the main variants of pollution are air pollution, water pollution, land pollution etc. These different kinds of pollution introduce various life-threatening diseases and are responsible for mass outbreak of several harmful virus and bacteria that affects health standard and ultimately causes rise in mortality rate. During the 19th century, cholera spread across the world causing millions of deaths in all the continents. A major cause of these pandemics is the water pollution in the industrial areas. Air pollution, another major type of pollution, is responsible for various respiratory diseases such as lung cancer. Another major type is land pollution which originates from deforestation, soil erosion, agricultural activities, industrial and nuclear waste. Some effects of land pollution are polluted underground water, polluted soil which leads to loss of fertile lands etc.

Another big factor affecting mortality rate is different socio-economic differences. People from lower income groups faces a difficult time availing necessary commodities and medical treatments which leads to higher mortality rate. Life expectancy also varies from different races. According to 2001 report of the President's Commission to Strengthen Social Security, blacks on an average had both lower income rate and shorter life expectancy.

Problem Statement:

The Goal of this project is to study how various environmental and socio-economic factors affect human mortality and to check which factors affect mortality rate significantly more than the others.

For this, we try to fit a linear model to explain the effects of the factors. We checked the validity of different assumptions and tried to provide a remedy in case an assumption is not satisfied. In this context we have checked the validity for the assumptions: Normality, Homoscedasticity, Independence of Errors, Multicolinearity of the regressors.

Data Description:

The data set that we used comes from McDonald and Schwing (1973), "Instabilities of Regression Estimates Relating Air Pollution to Mortality," Technometrics, 15, 463-481. The data can be found here.

This data consists of 15 independent variables and a measure of mortality on 60 US metropolitan areas in 1959-1961. The independent variables are:

1.	PREC (x_1)	Average Annual Precipitation in Inches
2.	JANT (x_2)	Average Temperature in January in degrees Ferenhite
3.	JULT (x_3)	Average Temperature in July in degrees Ferenhite
4.	OVR65 (x_4)	Percent of 1960 SMSA population that is 65 years of age or over
5.	POPN (x_5)	Average household size
6.	EDUC (x_6)	Median school years completed by those over 22 in 1960 SMSA
7.	HOUS (x_7)	Percent of housing units that are found with facilities
8.	DENS (x_8)	Population per sq. mile in urbanized areas, 1960
9.	NONW (x_9)	Percent of non-white population in urbanized areas, 1960
10.	WWDRK (x_{10})	Percent of population employed in white collar occupations
11.	POOR (x_{11})	Percent of families with income less than 3000 dollars
12.	HC (x_{12})	Relative population potential of hydrocarbons
13.	NOX (x_{13})	Relative pollution potential of oxides of nitrogen
14.	SO2 (x_{14})	Relative pollution potential of sulfur dioxide
15.	HUMID (x_{15})	Percent relative humidity, annual average at 1 PM

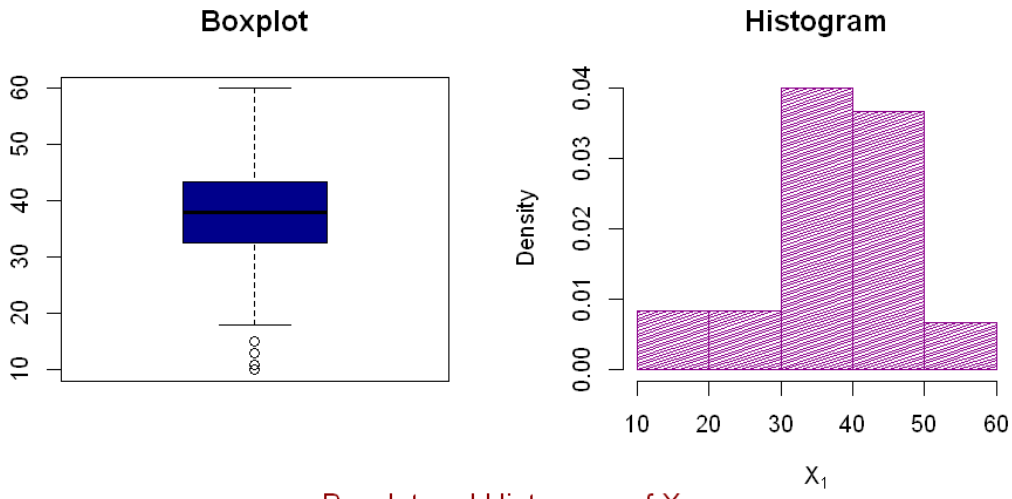
And the dependend variable is the mortality rates in those 60 areas:

MORT (y):	Total age-adjusted mortality rate per 100,000

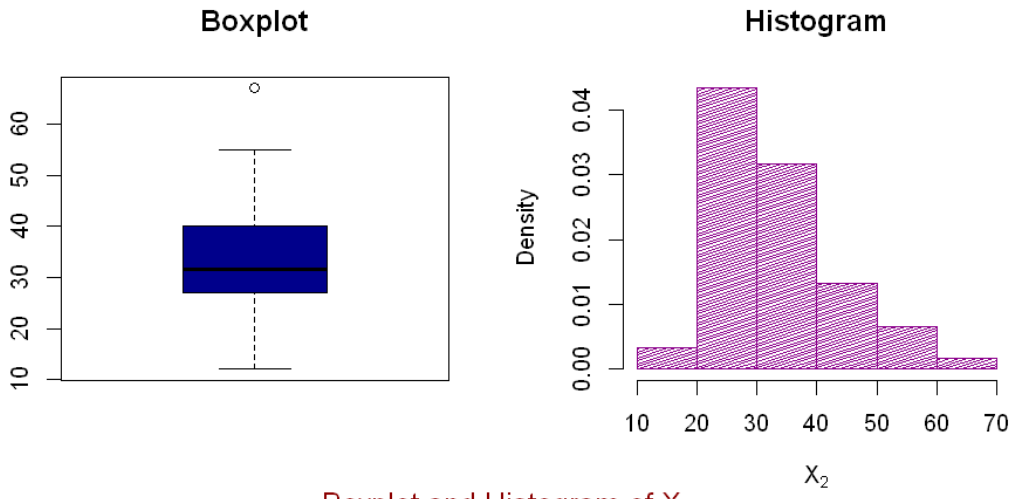
We denote the 15 explanatory variables as $x_1, x_2, \ldots, x_{15}$ and the dependend variable as y.

Exploratory Data Analysis:

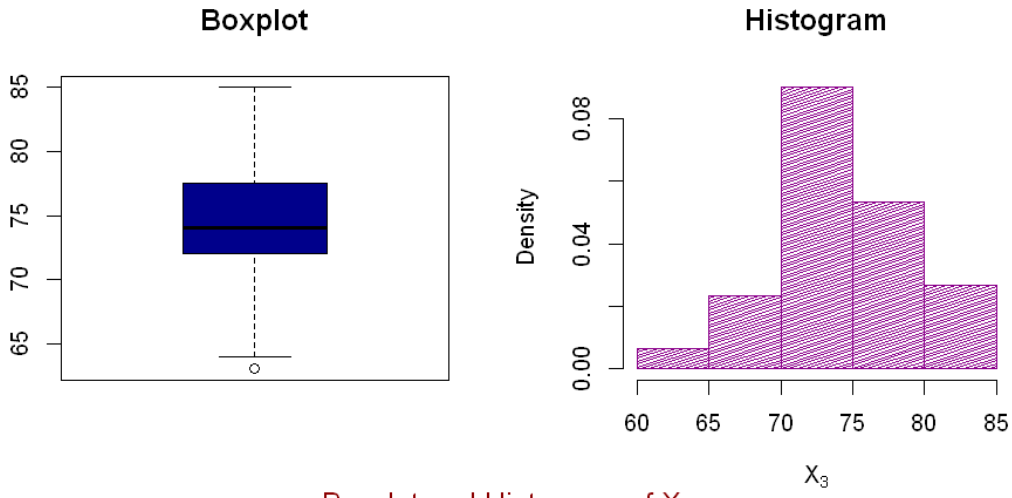
Here We have plotted the boxplot and histogram for the regressors and the regressand.



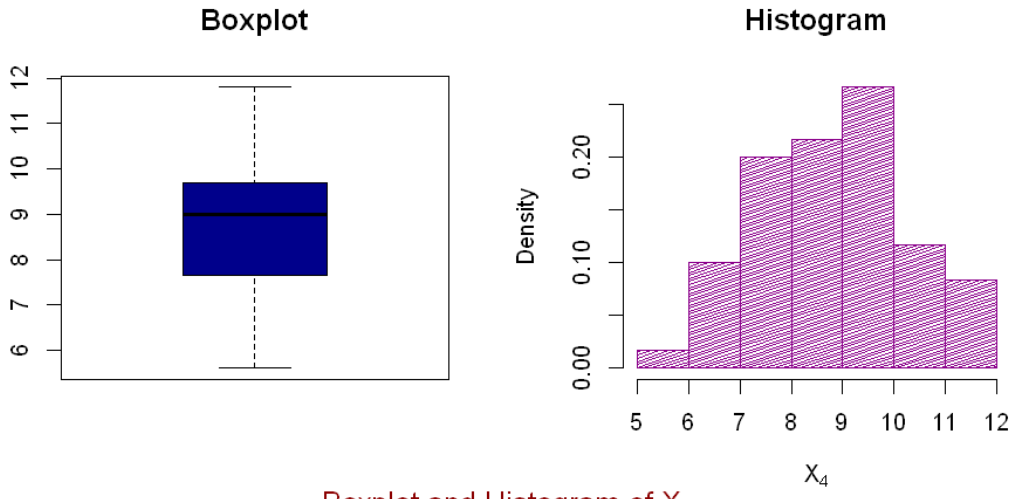
Boxplot and Histogram of X_1



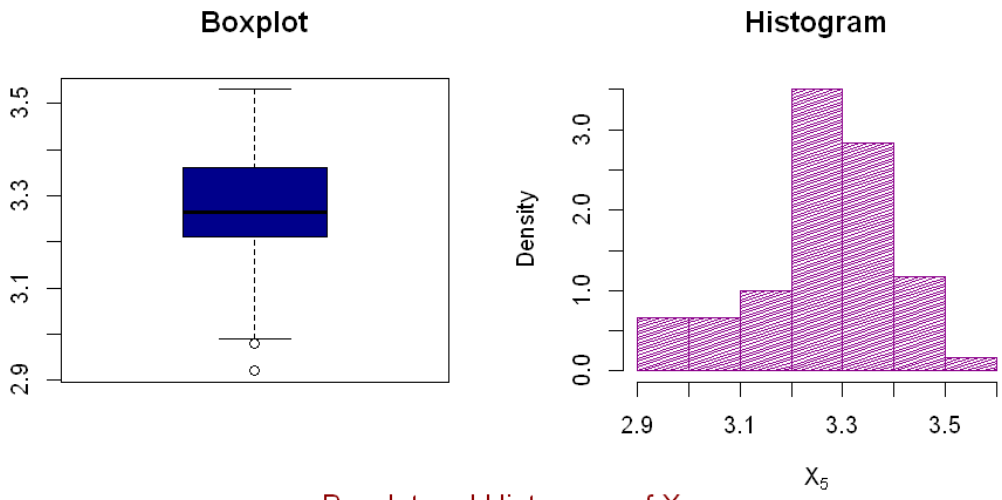
Boxplot and Histogram of X_2



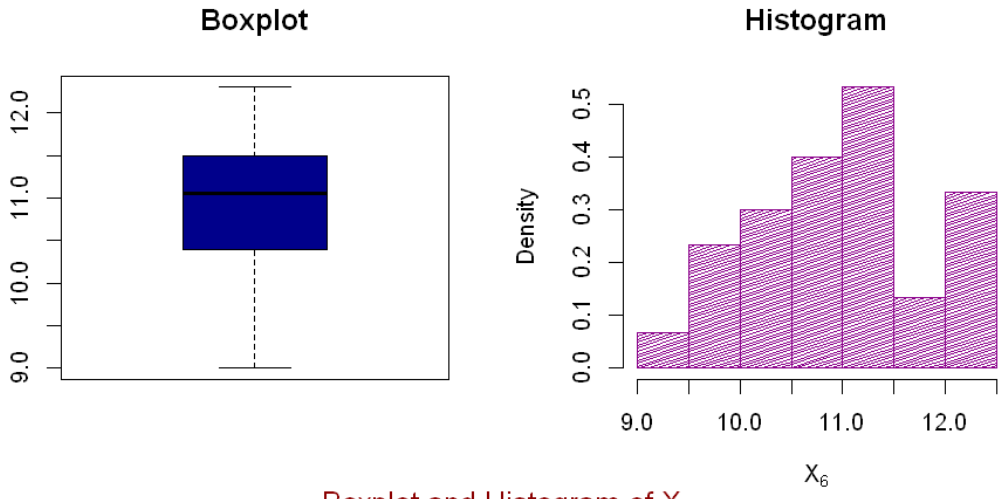
Boxplot and Histogram of X_3



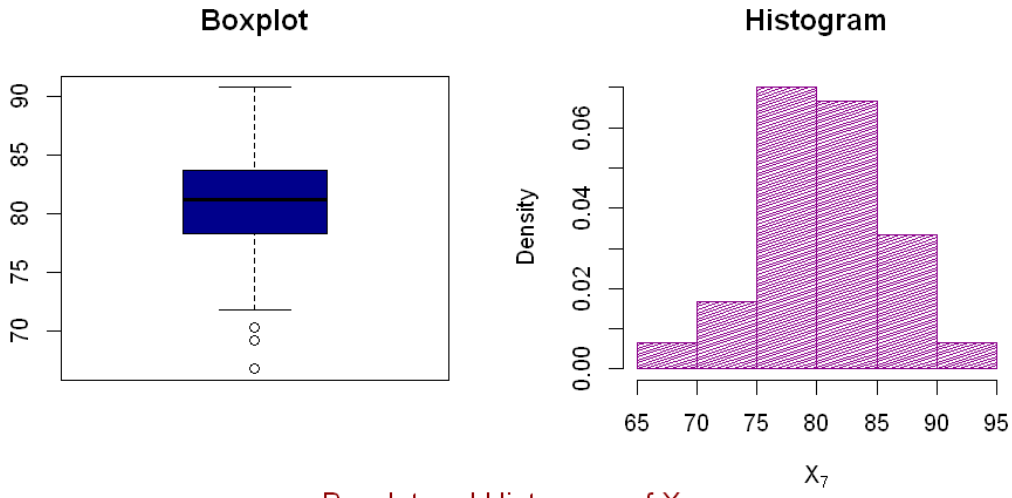
Boxplot and Histogram of X_4



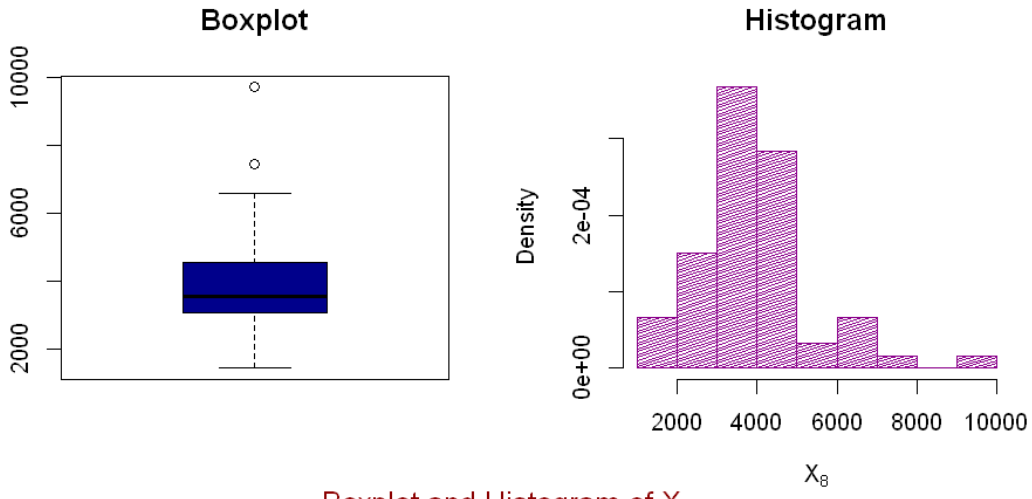
Boxplot and Histogram of X_5



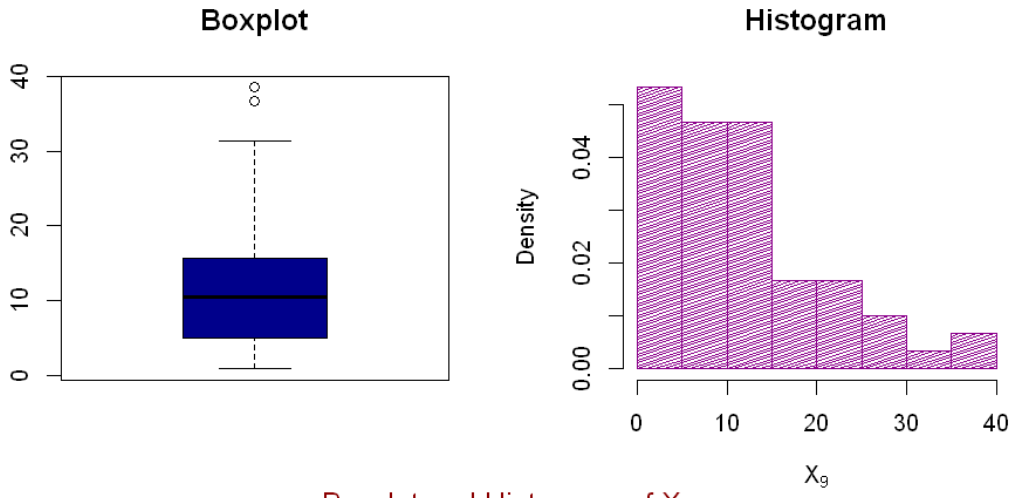
Boxplot and Histogram of X_6



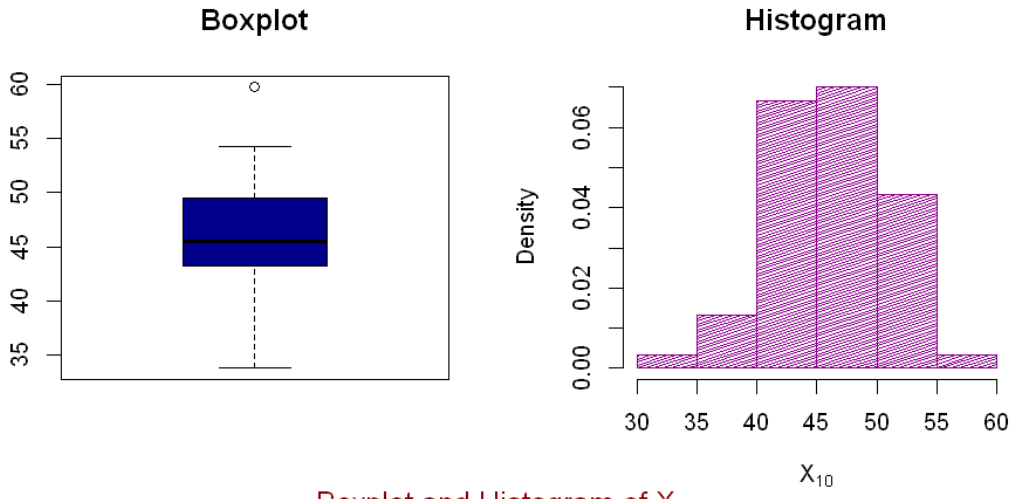
Boxplot and Histogram of X_7



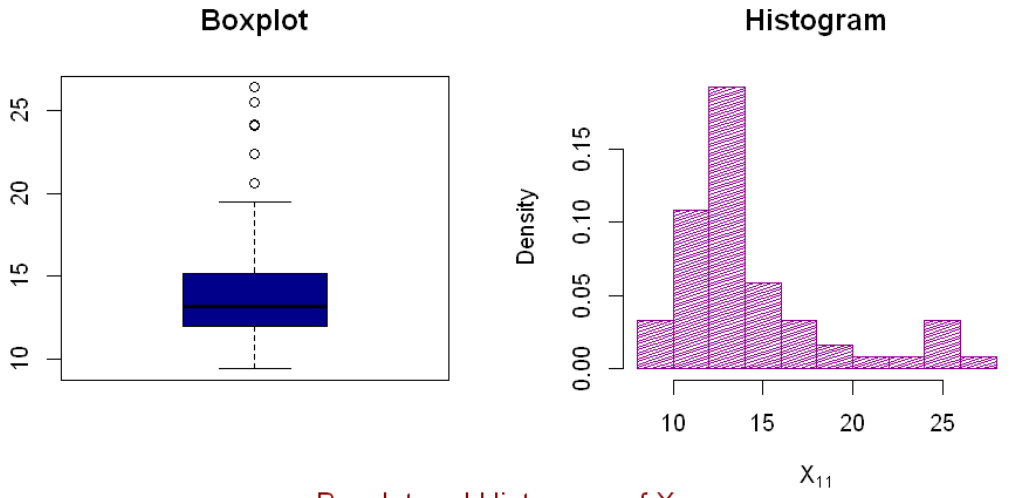
Boxplot and Histogram of X_8



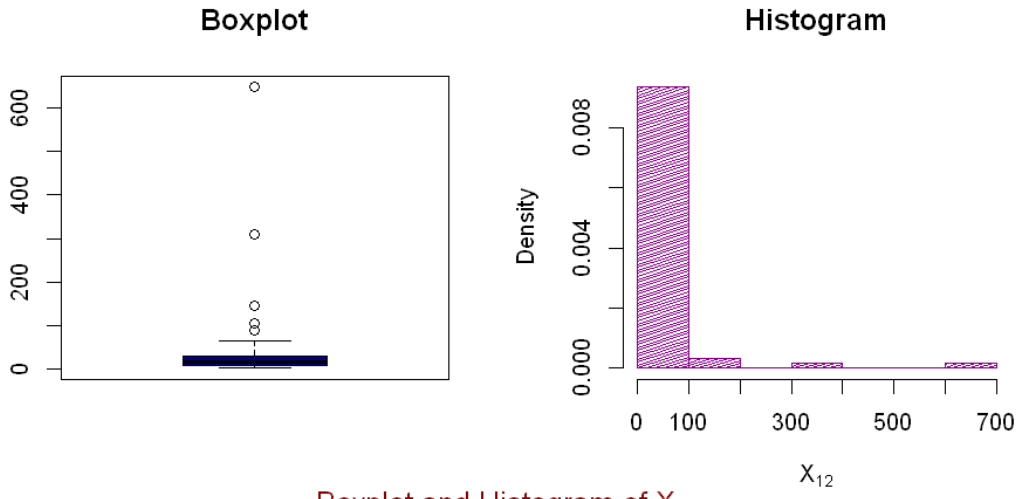
Boxplot and Histogram of X_9



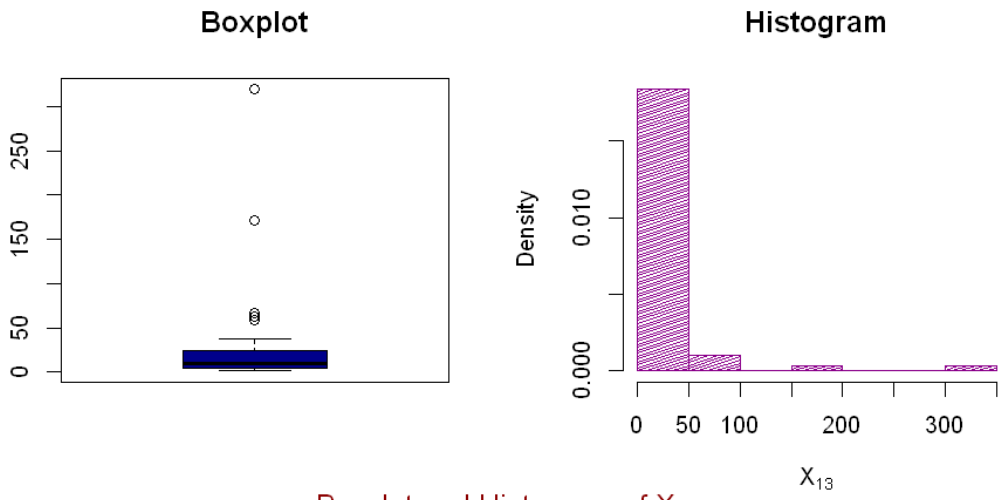
Boxplot and Histogram of X_{10}



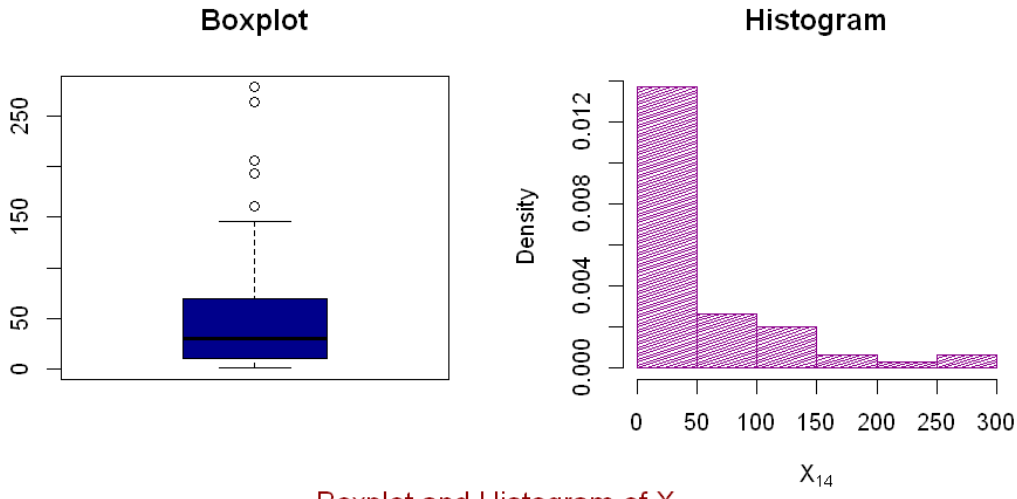
Boxplot and Histogram of X_{11}



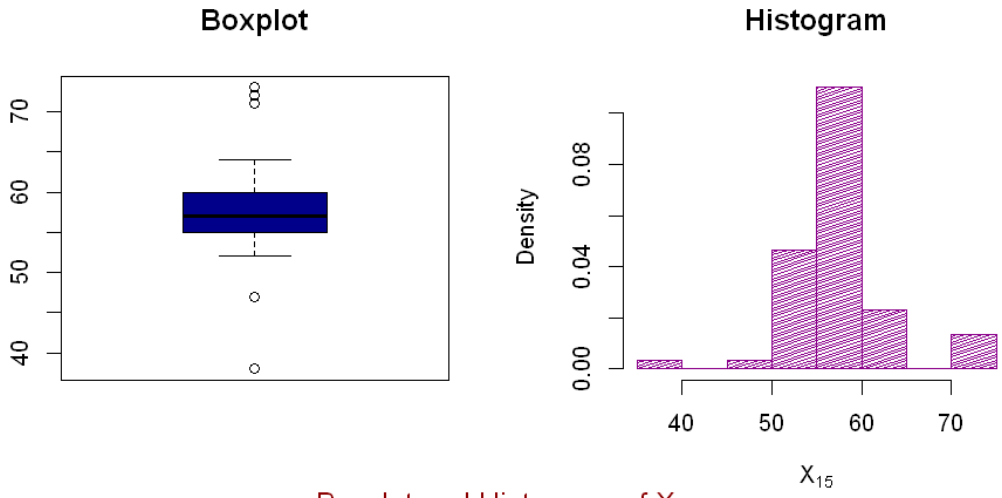
Boxplot and Histogram of X_{12}



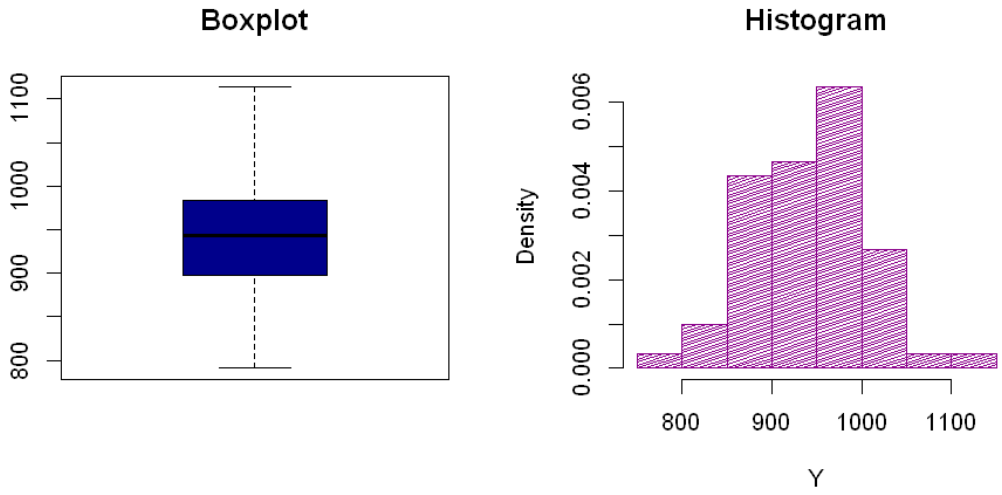
Boxplot and Histogram of X_{13}



Boxplot and Histogram of X_{14}



Boxplot and Histogram of X_{15}



Boxplot and Histogram of Y

Model Assumption:

Our multiple regression model assumption is,

$$y_i = \beta_0 + \sum_{j=1}^{15} \beta_j x_{ij} + \epsilon_i, \forall i = 1(1)n$$

Where,

- y_i = Age-adjusted mortality rate of the i^{th} area,
- x_{ij} = value of the j^{th} variable of the i^{th}area,
- $\beta = (\beta_1, \beta_2, \ldots, \beta_{15})' =$ Vector of unknown parameters,
- $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)' =$ Vector of random errors with usual error assumptions.

First we fit the model with the data and check the validity of the error assumptions.

Model fitting & Summary:

We fit a multiple linear regression model to the data and note the regression coefficients and summary of the model.

[3]:
```
#fitting MLR model
X<-data.matrix(subset(data,select=-c(Y)))
model<-lm(Y ~ .,data=data)
```

Estimates of Regression Coefficients:

Parameter	Estimate	Std. Error	t value	Pr(>|t|)
β_0	1.763981e+03	437.32672826	4.03355489	0.0002153976
β_1	1.905419e+00	0.92373242	2.06273860	0.0450705328
β_2	-1.937620e+00	1.10838287	-1.74815010	0.0874131937
β_3	-3.100426e+00	1.90165071	-1.63038666	0.1101586334
β_4	-9.065397e+00	8.48615229	-1.06825769	0.2912302968
β_5	-1.068257e+02	69.77950184	-1.53090377	0.1329519245
β_6	-1.715718e+01	11.86002521	-1.44663902	0.1550848875
β_7	-6.511172e-01	1.76775201	-0.36833062	0.7143929529
β_8	3.600485e-03	0.00402706	0.89407295	0.3761474785
β_9	4.459600e+00	1.32719663	3.36016532	0.0016182285
β_{10}	-1.870552e-01	1.66167677	-0.11257014	0.9108833755
β_{11}	-1.676426e-01	3.22727440	-0.05194556	0.9588071911
β_{12}	-6.721450e-01	0.49101863	-1.36887873	0.1779853622
β_{13}	1.340097e+00	1.00558577	1.33265273	0.1895062265
β_{14}	8.625178e-02	0.14751807	0.58468621	0.5617450516
β_{15}	1.068042e-01	1.16942177	0.09133080	0.9276442866

Anova Table:

Parameter	Df	Sum Sq	Mean Sq	F value	Pr(>F)
β_1	1	59266	59266	48.579	1.27e-08
β_2	1	1365	1365	1.119	0.295897
β_3	1	670	670	0.549	0.462506
β_4	1	19993	19993	16.388	0.000206
β_5	1	9	9	0.007	0.931723
β_6	1	28840	28840	23.639	1.52e-05
β_7	1	1748	1748	1.433	0.237668
β_8	1	10927	10927	8.957	0.004521
β_9	1	42072	42072	34.485	5.20e-07
β_{10}	1	56	56	0.046	0.831954
β_{11}	1	58	58	0.048	0.828466
β_{12}	1	45	45	0.037	0.848862
β_{13}	1	9133	9133	7.486	0.008931
β_{14}	1	434	434	0.356	0.554039
β_{15}	1	10	10	0.008	0.927644
Residuals	44	53680	1220	-	-

Model Summary:

- Residual Standard Error: 34.93 (with degrees of freedom 44)
- Multiple R squared: 0.7649
- Adjusted R squared: 0.6847
- F-statistic: 9.542 (with degrees of freedom 15, 44)
- p-value: 0.002193
- $F_{0.95;15,44}$: 1.900236 (upper 5% point of F-distribution with df 15, 44)

Test of the Significance of Regression Coefficients:

To test the significance of regression coefficients we consider the null hypothesis:

$$H_0 : \beta_j = 0, \forall j \neq 0$$

against

$$H_A : H_0 \text{ is not true}$$

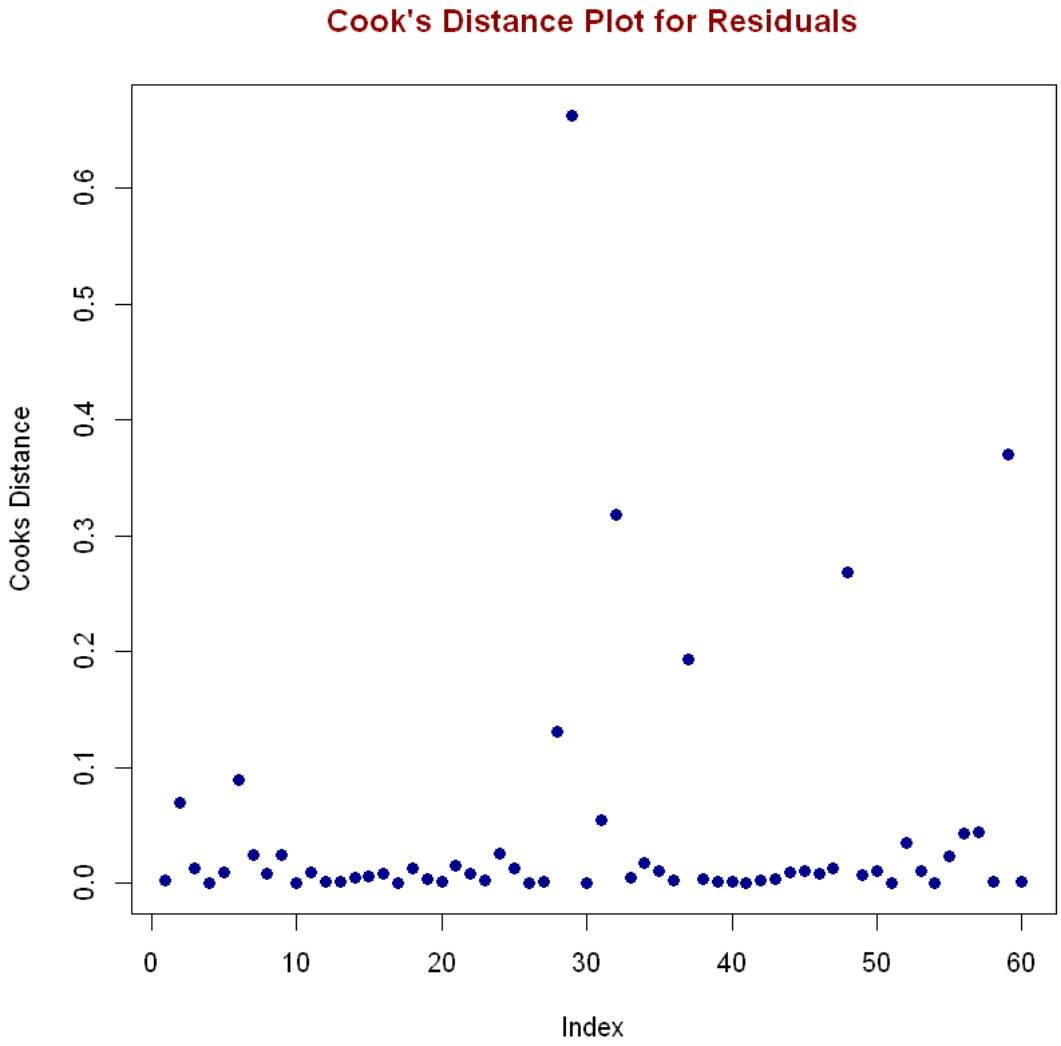
From the above summary we get the value of observed F–statistic as 9.542 and the observed p–value to be 2.193×10^{-09}. As the p–value of the test is very small we reject H_0 at level $\alpha = 0.05$ and conclude that at least one of the β_j's are significant. Hence, we proceed to fit a linear model.

Outlier Detection:

Cook's distance or Cook's D, named after the American statistician R. Dennis Cook, is a commonly used estimate of the influence of a data point when performing a least-squares regression analysis. In a practical ordinary least squares analysis, Cook's distance can be used in several ways: to indicate influential data points that are particularly worth checking for validity. The Cook's distance D of observation i (for $i = 1(1)n$) is defined as,

$$D_i = \frac{y_i - \hat{y}_i}{p \times MSE} \left(\frac{h_{ii}}{(1 - h_{ii})^2} \right)$$

Where, $\hat{y}_i$ denoted the predicted value of the i^{th} observation and h_{ii} denotes the i^{th} diagonal element of the hat matrix $H = X(X'X)^{-1}X'$ and p be the total no. of regressors. We say that the i^{th} observation is an outliar if $D_i > F_{\alpha;n,n-p-1}$. The plot of the corresponding Cook's distance for the model is given below:



Conclusion:

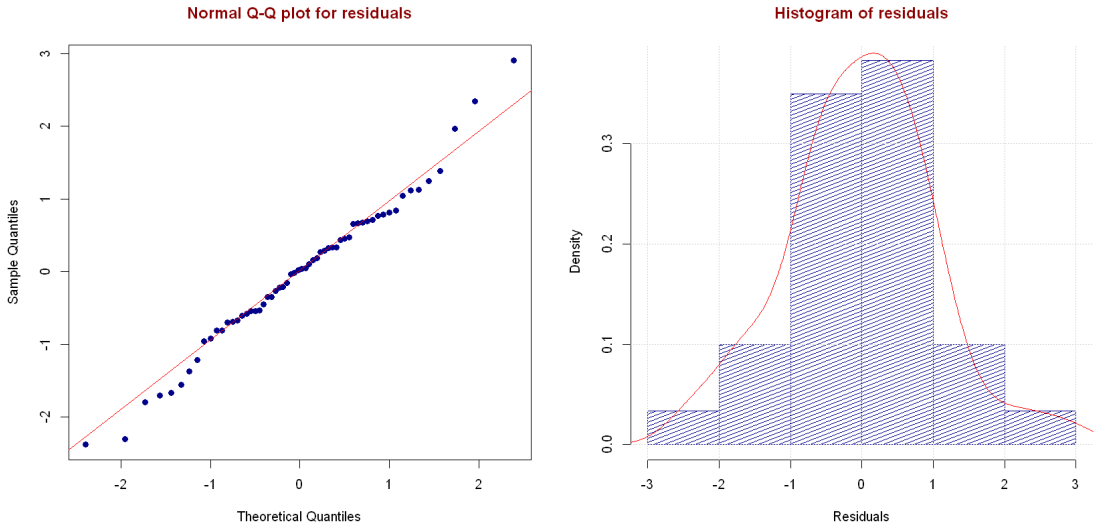
From the above plot we can observe that the maximum value of the observed Cook's distance is about 0.66. As the maximum value is less than $F_{0.95;15,44} = 1.9$, we can conclude that the model is not affected by any outliers.

Checking Error Assumptions of MLR Model:

Now we check for the validity of normality and homoscedasticity of error. As the data is not a time series data or a spatial data we omit the check for autocorrelation among the errors.

Assumption 1: Errors are Normally Distributed

We plot QQ-plot and histogram of the errors to check if the errors comes from a normal distribution:



Conclusion:

1. In the QQ-plot we observe that almost all the points fall on the 45 degree straight line.
2. Combined with the QQ-plot the histogram shows the normality assumptions of the errors are quite appropiate.

Now we perform Shapiro-Wilk test and Anderson-Darling normality test to check if the errors follows a normal distribution:

Shapiro-Wilk Test:

Test statistic:

$$W = \frac{\left(\sum_{i=1}^{n} a_i \hat{\epsilon}_{(i)}\right)^2}{\sum_{i=1}^{n} \left(\hat{\epsilon}_i - \bar{\hat{\epsilon}}\right)^2}$$

Here, $a' = (a_1, a_2, \ldots, a_n) = \frac{m^T V^{-1}}{\sqrt{m^T V^{-1} V^{-1} m}}$ where, $m = E(\hat{\epsilon}_{ordered})$ and $V = Cov(\hat{\epsilon}_{ordered})$

Here the hypotheses under consideration are,

$$H_0 : \text{The residuals are normally distributed vs } H_1 : \text{not } H_0$$

Test Criterion: We reject the null hypothesis if the observed p-value is less than α.

Anderson-Darling Test:

Test statistic:

$$A = -n - \sum_{i=1}^{n} \frac{2i-1}{n} \left[\ln\left(\Phi\left(\hat{\epsilon}_i\right)\right) + \ln\left(1 - \Phi\left(\hat{\epsilon}_{n-i+1}\right)\right) \right]$$

Here the hypotheses under consideration are,

$$H_0 : \text{The residuals are normally distributed vs } H_1 : \text{not } H_0$$

Test Criterion: We reject the null hypothesis if the observed p-value is less than α.

[9]:
```
#Shapiro-Wilk test for normality
shapiro.test(resid)

#Anderson-Darling normality test
library(nortest)
ad.test(resid)
```

```
        Shapiro-Wilk normality test

data:  resid
W = 0.98494, p-value = 0.6669
```

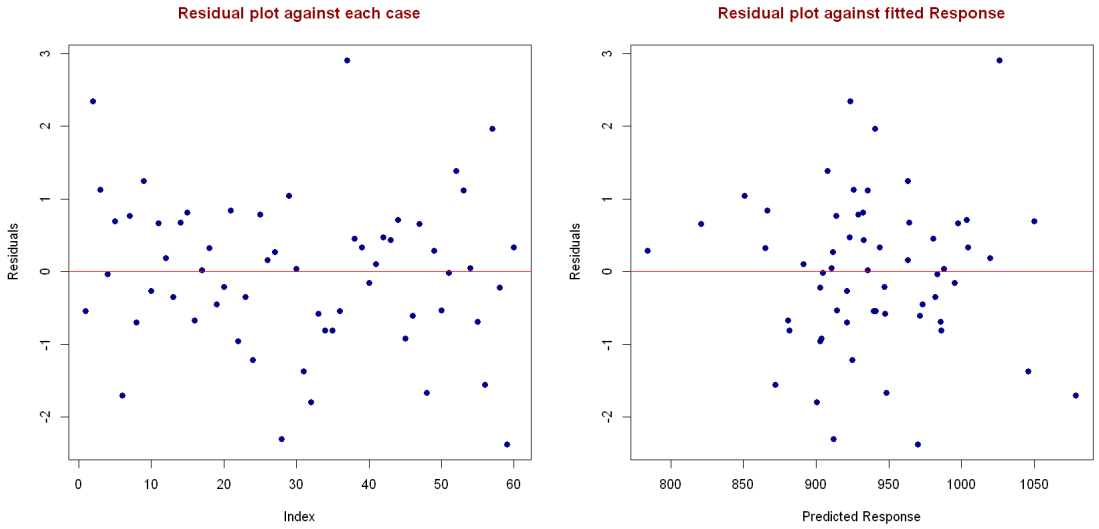
```
        Anderson-Darling normality test

data:  resid
A = 0.28976, p-value = 0.6015
```

Conclusion:

Higher p-value (> 0.05) for both the tests supports the null hypothesis that errors are normally distributed.

Assumption 2: Errors are Homoscedastic

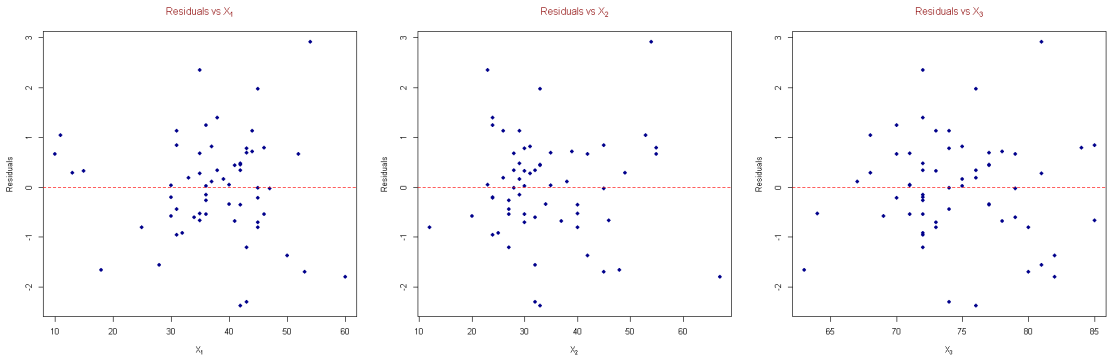
Now we check the homoscedasticity of errors. For this we make a scatterplot of the residuals and then plot the residuals against the fitted response.

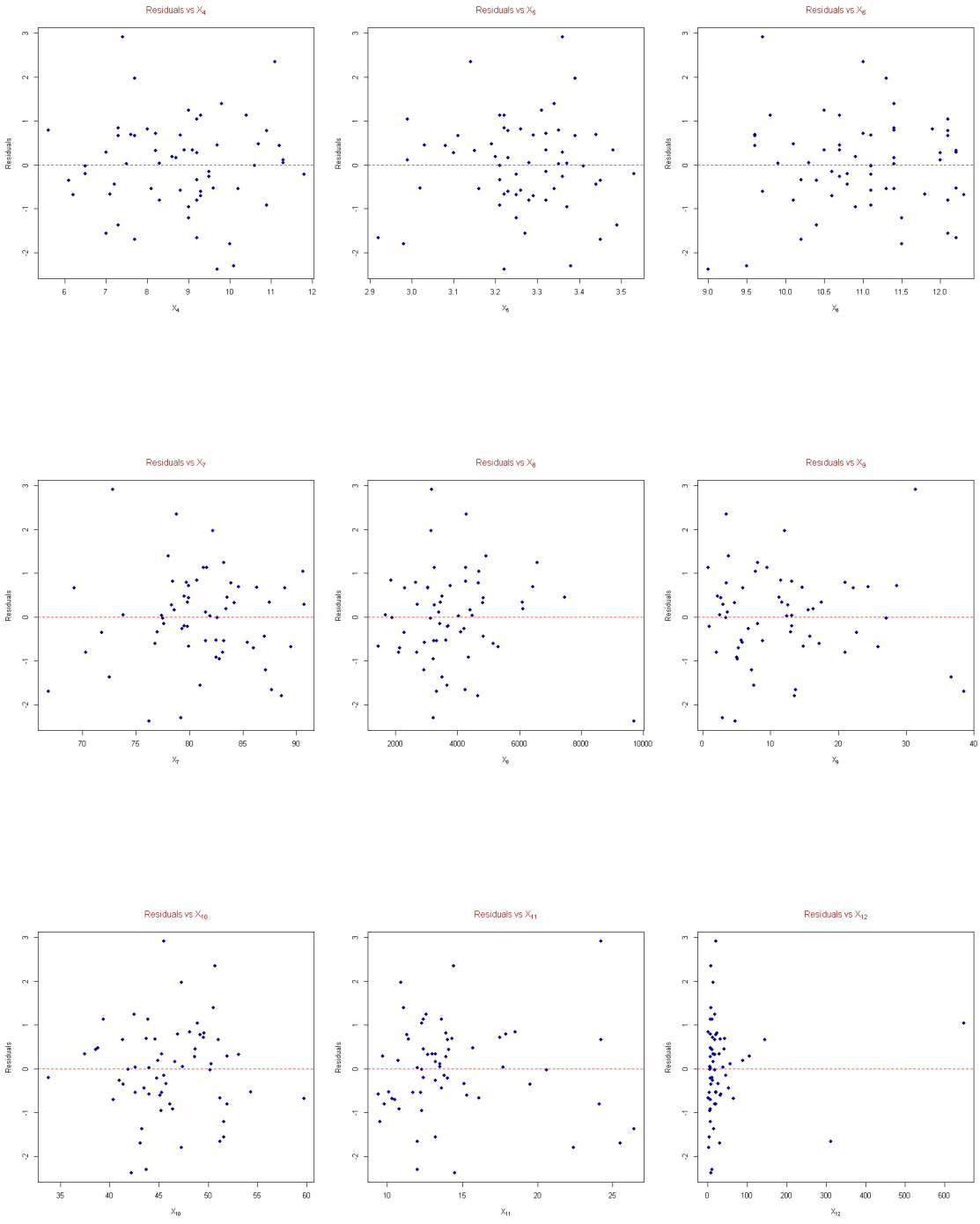


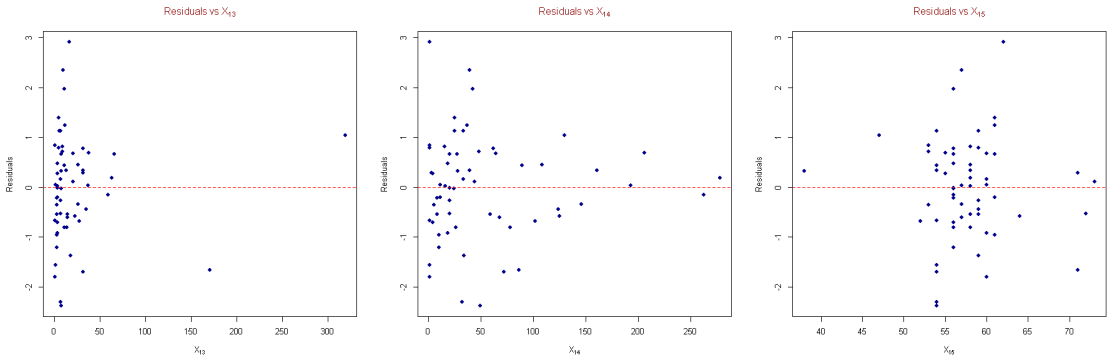
Conclusion:

There is no specific pattern among the residuals in the first plot indicating that there is no non-constant variance and the residuals are not correlated over different cases. Again no pattern detected in the plot against residuals against fitted response which indicates variance of residuals does not increase or decrease with predicted responses.

Now we plot the residuals against each regressor to check if the residuals are correlated to any of the explanatory variables:




Residuals vs X_4
Residuals vs X_5
Residuals vs X_6
Residuals vs X_7
Residuals vs X_8
Residuals vs X_9
Residuals vs X_{10}
Residuals vs X_{11}
Residuals vs X_{12}
Residuals



Conclusion:

From the plots we can't find any patterns of residuals for the regressors except for X_{14}. Hence the residuals can be assumed to be independently distributed with the regressors. Again, residuals are plotted randomly against each regressors, i.e. linear term in regressors are enough in the model. Hence no higher order term of the regressors is necessary.

In case of X_{14} there seems to be an inward funnel in the plot. So we are going for Glejser's test and Golffeld-Quandt test for confirmation.

Glejser's Test and Goldfeld-Quandt Test:

To check if the residuals have any functional relationship with any regressors we fit the following model for each of the regressors:

$$|\epsilon_j| = \delta_0 + \delta_1 x_{ij}^{\delta_2} + v_j$$

for each $i = 1(1)15$. To detect the form of heteroscedasticity, the values of δ_2 are taken as $-1, -\frac{1}{2}, \frac{1}{2}, 1$. The R^2 values for each of the models are shown below:

δ_2	-1	-0.5	0.5	1
X_1	0.0313	0.0386	0.0577	0.0681
X_2	0.0038	0.0064	0.0131	0.0166
X_3	0.0378	0.0385	0.0398	0.0403
X_4	0.0008	0.0006	0.0002	0.0001
X_5	0.0037	0.0037	0.0037	0.0037
X_6	0.0587	0.0565	0.0524	0.0504
X_7	0.0332	0.0326	0.0313	0.0307
X_8	0.0112	0.0101	0.0097	0.0108
X_9	0.0000	0.0000	0.0050	0.0142
X_{10}	0.0087	0.0077	0.0058	0.0050
X_{11}	0.0198	0.0252	0.0381	0.0450
X_{12}	0.0033	0.0138	0.0215	0.0097
X_{13}	0.0016	0.0027	0.0106	0.0080
X_{14}	0.0712	0.0551	0.0479	0.0602
X_{15}	0.0014	0.0007	0.0001	0.0000

Here the highest value of R^2 corresponds to x_{14} with $\delta_2 = -1$. So we perform Goldfeld-Quandt test for decreasing oreder of x_{14}.

Here we consider the hypotheses,

$$H_0 : \text{The residuals are Homoscedastic}$$

against,

$$H_1 : \text{The residuals are Heteroscedastic}$$

For, Goldfeld-Quant test we arrange the data in descending order of x_{14} (as $\delta_2 = -1$ for x_{14}).

Then we divide the dataset into three parts with number of observations $\frac{n-c}{2}$, c and $\frac{n-c}{2}$ respectively. Then we discard the 2nd part and calculate the RSS for the remaining two parts. Let us denote these as RSS_1 and RSS_3 respectively. Then we have the test statistic of Goldfeld-Quandt test as,

$$F = \frac{\text{RSS}_3}{\text{RSS}_1}$$

Test Criterion: We reject H_0 if $F > F_{\alpha;\frac{n-c}{2}-p-1,\frac{n-c}{2}-p-1}$ and accept H_0 otherwise.

[13]:
```
#GQ-test for heteroscedasticity
library(lmtest)
gqtest(Y ~ ., fraction=20, order.by=~-X14, data=data)
qf(0.95, 4, 4)
```

```
        Goldfeld-Quandt test

data:  Y ~ .
GQ = 1.3446, df1 = 4, df2 = 4, p-value = 0.3906
alternative hypothesis: variance increases from segment 1 to 2
```

6.38823290869587

Conclusion:

As the observed value of GQ statistic is less than $F_{0.95;4,4} = 6.388$ we cannot reject the null hypothesis at 5% level of significance and conclude that the errors are homoscedastic.

Checking for Multicolinearity in Model:

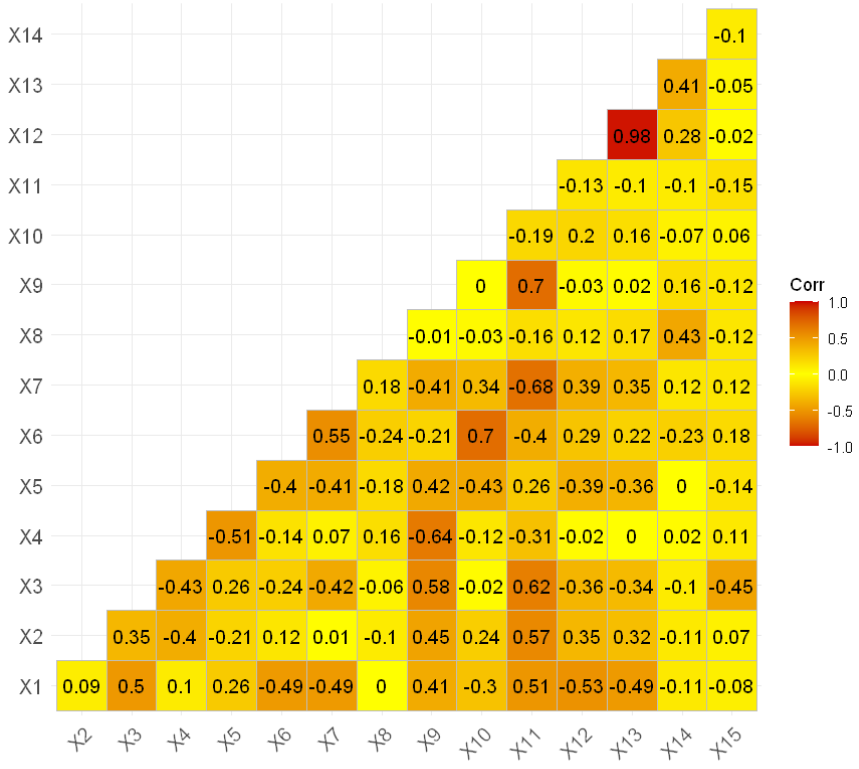
Now we check for existance of multicolinearity in the regressors. For a model with p regressors, it may be a case such that $\sum_{j=1}^{k} a_j X_{i_j} = 0$, for some $a_j \neq 0, k \leq p, j = 1(1)k$, implying perfect linear relationship or multicolinearity among regressors. Im this case $|X'X| = 0$ or $(X'X)^{-1}$ does not exists and we can't estimate the model parameters. But in practice, X_i's are realisations of some random variables and such perfect linear relationship is generally not possible. But when p is large, it may be that $\sum_{j=1}^{k} a_j X_{i_j} \approx 0$ i.e. there may exists near linear relationship between the regressors. As a result $|X'X|$ becomes small and the estimates of regresssion parameters becomes unstable as $Cov(\hat{\beta}) = \sigma^2 (X'X)^{-1}$.

Another way of formulating this is by the condition number of $X'X$ $\left(= \frac{\text{maximum eigenvalue of } X'X}{\text{minimum eigenvalue of } X'X}\right)$ which becomes very large as atleast one eigenvalue becomes small. In this case $X'X$ is said to be ill-conditioned.

If this situation occures, we say that the model is suffering from multicolinearity issue.

If our model suffers from multicolinearity issue, the parameter estimates have exceptionally high value (in magnitude), some parameter estimates may have large variance, some important regressors may become insignificant, even some regressors may turn up with 'wrong sign'. That's why, if our model suffers from multicolinearity issue, we have to take necessary steps to remove this issue.

Now we plot the correlogram of X matrix and calculate the condition number of $X'X$.



[18]:
```
# Calculating Condition Number
eigen.values <- eigen(t(X) %*% X)$values
cond.num <- max(eigen.values)/min(eigen.values)
cond.num
```

1253619450.55388

In our model the condition number of $X'X$ is 1253619451 which is very high i.e. $X'X$ is ill-conditioned. This clearly indicates that there is multicolinearity in the regressors in the model.

Now, here is no such condition or criteria that prevent us from excluding any variable for better model. So we will remove the variables responsible for multicolinearity issue.

To do that, first we applied VIF (Variance Inflation Factor) to detect multicolinearity.

If we fit linear regression of X_j on $X_1, X_2, \ldots, X_{15}$ without X_j, $j = 1(1)15$, then let $R^2_{(j)}$ be the coefficient of determination for this model and in this case VIF value for variable X_j is given by,

$$\text{VIF}_{(j)} = \frac{1}{1 - R^2_{(j)}}, j = 1(1)15$$

i.e. $\text{VIF}_{(j)}$ measures combined effect of dependence between X_j and any subset of the remaining $p - 1 = 14$ regressors.

Now, $R^2_{(j)} \approx 1$, implies $\text{VIF}_{(j)} \to \infty$. But if coefficient of determination is greater than 0.8 for some linear model, it indicates quite high linearity between regressand and regressors. So we are fixing threshold value of $R^2_{(j)}$ as 0.8, i.e. threshold value of $\text{VIF}_{(j)}$ as 5, $j = 1(1)15$. We will doubt for multicolinearity issue if one or more variable has VIF value greater than 5.

The VIF's of the regressors are as follows:

Regressors:	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
VIF:	4.11389	6.14355	3.96777	7.47005	4.30762	4.86054	3.99478	1.65828
Regressors:	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	
VIF:	6.77960	2.84158	8.71707	98.6399	104.982	4.22893	1.90709	

Here in this model we get that $X_2, X_4, X_9, X_{11}, X_{12}, X_{13}$ have VIF-value greater than 5. This indicates that these variables may include in multicolinearity. Now to get more evidence against subsets of variables creating multicolinarity issue, we will perform Variance Decomposition method.

To apply variance decomposition method we take the standardize version of X matrix. Then we calculate,

$$\pi_{kj} = \frac{v^2_{kj} \Big/ l_k}{\text{VIF}_j}$$

where, l_k is the k^{th} eigen value of $X'X$ and v_{kj} is the j^{th} component of the orthonormal eigen vector corresponding to l_k. Also $\sum_k \pi_{kj} = 1$.

Here if high proportion of π_{kj} (generally > 0.5) are found for more than one $\hat{\beta}_j, j = 1(1)15$ then multicolinearity is indicated in the corresponding subset.

[20]:
```
eigprop(model.std, Inter=F)
```

```
Call:
eigprop(mod = model.std, Inter = F)

   Eigenvalues      CI     X1     X2     X3     X4     X5     X6     X7     X8
1       4.5284  1.0000 0.0064 0.0002 0.0066 0.0008 0.0045 0.0037 0.0072 0.0007
2       2.7548  1.2821 0.0009 0.0137 0.0035 0.0065 0.0004 0.0022 0.0002 0.0001
3       2.0545  1.4846 0.0001 0.0009 0.0007 0.0016 0.0001 0.0185 0.0004 0.0569
4       1.3484  1.8326 0.0200 0.0131 0.0001 0.0269 0.0539 0.0019 0.0041 0.0014
5       1.2232  1.9241 0.0030 0.0010 0.0327 0.0001 0.0101 0.0031 0.0076 0.0860
6       0.9604  2.1714 0.0085 0.0010 0.0035 0.0027 0.0002 0.0002 0.0069 0.0947
7       0.6127  2.7185 0.0001 0.0348 0.0035 0.0096 0.0005 0.0077 0.1002 0.0968
8       0.4720  3.0974 0.1218 0.0016 0.0086 0.0044 0.0002 0.0026 0.1252 0.3565
9       0.3709  3.4944 0.1715 0.0239 0.0542 0.0059 0.0514 0.0221 0.0002 0.0694
10      0.2164  4.5746 0.0021 0.1134 0.1846 0.0028 0.1414 0.1659 0.0208 0.0125
11      0.1664  5.2175 0.0198 0.0037 0.4377 0.0015 0.0318 0.1519 0.0338 0.0066
12      0.1270  5.9712 0.0593 0.0744 0.0198 0.0032 0.2621 0.4750 0.1165 0.2102
13      0.1140  6.3030 0.1980 0.0596 0.2173 0.2653 0.2502 0.0004 0.3507 0.0080
14      0.0460  9.9176 0.3837 0.6572 0.0259 0.6140 0.1715 0.0467 0.2031 0.0000
15      0.0049 30.5051 0.0050 0.0015 0.0014 0.0547 0.0217 0.0981 0.0231 0.0003
       X9    X10    X11    X12    X13    X14    X15
1  0.0030 0.0030 0.0032 0.0002 0.0001 0.0002 0.0015
2  0.0073 0.0074 0.0030 0.0005 0.0005 0.0008 0.0013
3  0.0003 0.0190 0.0000 0.0003 0.0004 0.0311 0.0093
4  0.0011 0.0005 0.0066 0.0000 0.0000 0.0055 0.0126
5  0.0000 0.0424 0.0017 0.0004 0.0003 0.0020 0.1162
6  0.0114 0.0133 0.0006 0.0005 0.0004 0.0223 0.2494
7  0.0060 0.1265 0.0017 0.0000 0.0001 0.0729 0.0000
8  0.0006 0.0217 0.0063 0.0000 0.0001 0.0714 0.0031
9  0.0395 0.0133 0.0033 0.0017 0.0012 0.1135 0.0195
10 0.0396 0.3802 0.0029 0.0002 0.0004 0.0028 0.1007
11 0.0512 0.1533 0.1221 0.0043 0.0023 0.0188 0.2852
12 0.3236 0.0425 0.0121 0.0000 0.0005 0.0781 0.0069
13 0.0089 0.0259 0.2071 0.0001 0.0002 0.0104 0.1291
14 0.5002 0.0971 0.6264 0.0032 0.0008 0.0000 0.0587
15 0.0074 0.0537 0.0029 0.9886 0.9927 0.5701 0.0065

================================
Row 14==> X2, proportion 0.657171 >= 0.50
Row 14==> X4, proportion 0.614033 >= 0.50
Row 14==> X9, proportion 0.500171 >= 0.50
Row 14==> X11, proportion 0.626417 >= 0.50
Row 15==> X12, proportion 0.988630 >= 0.50
Row 15==> X13, proportion 0.992728 >= 0.50
Row 15==> X14, proportion 0.570060 >= 0.50
```

In row 14, the proportion corresponding to X_2, X_4, X_9, X_{11} are greater than 0.5. Hence we conclude that these variables have multicolinearity among them. As the VIF corresponding to X_{11} is highest among them we fit a linear regression model with regressand X_{11} and the rest as regressors.

[21]:
```
summary(lm(X11 ~ X2 + X4 + X9, data=data.std))
```

```
Call:
lm(formula = X11 ~ X2 + X4 + X9, data = data.std)

Residuals:
     Min       1Q   Median       3Q      Max
-1.71051 -0.40200  0.03159  0.37045  1.47840

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.994e-17  8.117e-02   0.000 1.000000
X2          3.501e-01  9.303e-02   3.764 0.000403 ***
X4          2.999e-01  1.076e-01   2.786 0.007269 **
X9          7.373e-01  1.108e-01   6.654 1.29e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6287 on 56 degrees of freedom
Multiple R-squared:  0.6248,        Adjusted R-squared:  0.6047
F-statistic: 31.08 on 3 and 56 DF,  p-value: 5.81e-12
```

As the value of adjusted R^2 is 0.6047 we conclude that X_{11} can be determined effectively by using variables X_2, X_4, X_9. So we can remove X_{11} from regressors' list.

Similarly, in row 15 the proportion corresponding to X_{12}, X_{13}, X_{14} are greater than 0.5. Hence we conclude that these variables have multicolinearity among them. As the VIF corresponding to X_{13} is the highest among them we fit a linear regression model with regressand X_{13} and the rest as regressors.

[22]:
```
summary(lm(X13 ~ X12 + X14, data=data.std))
```

```
Call:
lm(formula = X13 ~ X12 + X14, data = data.std)

Residuals:
     Min       1Q   Median       3Q      Max
-0.37242 -0.03887 -0.00487  0.02202  0.35606

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) 6.697e-17  1.510e-02    0.00        1
X12         9.435e-01  1.588e-02   59.42  < 2e-16 ***
X14         1.431e-01  1.588e-02    9.01 1.48e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.117 on 57 degrees of freedom
Multiple R-squared:  0.9868,        Adjusted R-squared:  0.9863
F-statistic:  2127 on 2 and 57 DF,  p-value: < 2.2e-16
```

As the value of adjusted R^2 is 0.9863 we conclude that X_{13} can be determined effectively using variables X_{12} and X_{14}. So we can remove X_{13} from regressor's list.

Now we fit the model again without the regressors X_{11} and X_{13}.

The VIF of the regressors in the modified model is:

Regressors:	X_1	X_2	X_3	X_4	X_5	X_6	X_7
VIF:	3.893751	3.353352	3.883299	6.061823	4.179623	4.395038	2.455549
Regressors:	X_8	X_9	X_{10}	X_{12}	X_{14}	X_{15}	
VIF:	1.635778	4.85897	2.557248	3.303399	1.769514	1.820013	

Here only VIF corresponding to X_4 is greater than 5. So we calculate the variance decomposition proportions to get better evidence for multicolinearity.

[24]:
```
eigprop(model.std1, Inter=F)
```

```
Call:
eigprop(mod = model.std1, Inter = F)

   Eigenvalues     CI     X1     X2     X3     X4     X5     X6     X7     X8
1       3.6992 1.0000 0.0097 0.0001 0.0095 0.0015 0.0083 0.0074 0.0167 0.0004
2       2.3871 1.2448 0.0010 0.0288 0.0076 0.0156 0.0000 0.0102 0.0012 0.0065
3       1.7632 1.4484 0.0010 0.0012 0.0001 0.0002 0.0004 0.0078 0.0026 0.1013
4       1.2572 1.7153 0.0368 0.0213 0.0167 0.0241 0.0620 0.0004 0.0018 0.0211
5       1.0978 1.8356 0.0044 0.0406 0.0196 0.0005 0.0009 0.0083 0.0068 0.0031
6       0.8538 2.0815 0.0116 0.0205 0.0000 0.0029 0.0031 0.0093 0.0147 0.1080
7       0.5944 2.4947 0.0001 0.0301 0.0027 0.0145 0.0030 0.0030 0.1941 0.1767
8       0.4574 2.8438 0.0864 0.0020 0.0068 0.0036 0.0002 0.0042 0.3072 0.3292
9       0.3414 3.2918 0.1730 0.0285 0.1237 0.0133 0.0540 0.0276 0.0102 0.0283
10      0.2130 4.1671 0.0022 0.2091 0.0969 0.0015 0.1560 0.2359 0.0177 0.0134
11      0.1342 5.2495 0.0026 0.1934 0.6371 0.0451 0.1069 0.1078 0.0851 0.0031
12      0.1247 5.4468 0.1030 0.1346 0.0452 0.0006 0.2463 0.4592 0.3072 0.1875
13      0.0765 6.9537 0.5682 0.2898 0.0340 0.8765 0.3588 0.1187 0.0346 0.0213
       X9    X10    X12    X14    X15
1  0.0063 0.0068 0.0067 0.0001 0.0035
2  0.0115 0.0222 0.0057 0.0016 0.0004
3  0.0046 0.0017 0.0224 0.1196 0.0202
```

```
4  0.0002 0.0150 0.0016 0.0101 0.0041
5  0.0071 0.0191 0.0133 0.0021 0.2665
6  0.0074 0.0684 0.0842 0.0345 0.1224
7  0.0101 0.1012 0.0061 0.2037 0.0012
8  0.0003 0.0342 0.0026 0.2138 0.0040
9  0.0687 0.0095 0.1267 0.1574 0.0688
10 0.0898 0.4902 0.0179 0.0035 0.0492
11 0.0115 0.1497 0.4686 0.0733 0.4354
12 0.4614 0.0393 0.0562 0.1804 0.0178
13 0.3210 0.0427 0.1878 0.0000 0.0064

==============================
Row 13==> X1, proportion 0.568162 >= 0.50
Row 11==> X3, proportion 0.637139 >= 0.50
Row 13==> X4, proportion 0.876512 >= 0.50
```

Here in row 13, the proportions corresponding to X_1 and X_4 are greater than 0.5. So we fit linear regression model with X_4 as regressand and X_1 as regressor.

[25]:
```
summary(lm(X4 ~ X1, data = data.std))
```

```
Call:
lm(formula = X4 ~ X1, data = data.std)

Residuals:
    Min      1Q  Median      3Q     Max
-2.2713 -0.8733  0.1161  0.5694  1.9722

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.164e-16  1.295e-01   0.000    1.000
X1           1.011e-01  1.306e-01   0.774    0.442

Residual standard error: 1.003 on 58 degrees of freedom
Multiple R-squared:  0.01022,        Adjusted R-squared:  -0.006841
F-statistic: 0.5991 on 1 and 58 DF,  p-value: 0.4421
```

The value of multiple R^2 (0.01022) is very close to 0 and observed p-value (0.4421) is greater than the level of significance (0.05). So we can see that there doesnot exist any significant linear relationship between X_1 and X_4.Moreover, condition number is approximately 7 which is less than 15. Hence we conclude that the remaining regressors doesnot have multicolinearity among them.

The model after removing multicolinearity is as follows:

$$y_i = \beta_0 + \sum_{j=1}^{10} \beta_j x_{ij} + \beta_{12} x_{12} + \beta_{14} x_{14} + \beta_{15} x_{15} + \epsilon_i, \forall i = 1(1)n$$

Variable Selection:

After removing multicolinearity, we perform variable selection method to discard insignificant regressors (if any) from the model. In order to do this we apply stepwise selection method. We start with forward selection, but at each step all regressors previously entered are rechecked for possible deletion by backward elimination since the regressor added in earlier step may now become unnecessary because of the presence of new regressors.

In forward selection we start with only intercept model,

$$y_i = \beta_0 + \epsilon_i, \forall i = 1(1)n$$

Now $\forall j = 1(1)p$ (p = number of remaining variables in the model), we add X_j in the model as,

$$y_i = \beta_0 + \beta_j x_{ij} + \epsilon_i, \forall i = 1(1)n$$

and test for the significance of β_j. That is we test the hypotheses:

$$H_{10} : \beta_j = 0 \text{ vs } H_{11} : \beta_j \neq 0$$

The test statistic is the partial F statistic which, under H_0,

$$F_j = \frac{R_j^2 - R_0^2}{R_0^2/(n-2)} \sim F_{1,n-2}$$

Here R_j^2 is restricted residual Sum of Square under null hypothesis and R_0^2 is unrestricted residual Sum of Square. Now we say β_j is significant if $F_j > F_{1-\alpha;1,n-2}$. We choose the regressor for which the regression coeffecient is significant and the test statistic is the highest.

If a regressor X_j is chosen then our model becomes

$$y_i = \beta_0 + \beta_j x_{ij} + \epsilon_i, \forall i = 1(1)n$$

We then apply forward selection again. We consider the models for $l(\neq j) = 1(1)p$,

$$y_i = \beta_0 + \beta_j x_{ij} + \beta_l x_{il} + \epsilon_i, \forall i = 1(1)n$$

and test for the significance of β_l. That is we test the hypotheses:

$$H_{20} : \beta_l = 0 \text{ vs } H_{21} : \beta_l \neq 0$$

The test statistic in this case is the partial F statistic which, under H_0,

$$F_{j,l} = \frac{\text{SSRes}(\beta_l|\beta_j)}{\text{MSRes}(\beta_j, \beta_l)} \sim F_{1,n-3}$$

where, SSRes$(\beta_l|\beta_j)$ = Extra sum of square due to β_l when β_0 and β_j are already in the model and MSRes(β_j, β_l) is the mean square residual of the model having β_0, β_j and β_l.

We take the regressor (X_l) with highest value of $F_{j,l}$ in the model.

Let after the 2nd step our model is,

$$y_i = \beta_0 + \beta_j x_{ij} + \beta_l x_{il} + \epsilon_i, \forall i = 1(1)n$$

After we have added a new regressor in the model, we check if the regressor added in the 1st step is significant or not. For this we test the hypothesis: H_{10} vs H_{11} on the model. In this case the test statistic is,

$$F_l = \frac{\text{SSRes}(\beta_j|\beta_l)}{\text{MSRes}(\beta_l, \beta_j)} \sim F_{1,n-3}, \text{ under } H_0$$

If observed $F_l > F_{1-\alpha;1,n-3}$ we will reject H_{10}.

Similarly, at the s^{th} step, we test for all the $s-1$ regressors previously added. Proceeding in this way, that is selecting variable using Forward selection and then removing using Backward selection, we carry on stepwise selection.

[26]:
```
library(olsrr)
ols_step_both_p(model.std1)
```

Stepwise Selection Summary

Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	X_9	addition	0.414	0.404	54.1220	638.8107	48.0112
2	X_6	addition	0.563	0.547	28.2370	623.2920	41.8521
3	X_2	addition	0.639	0.620	15.8620	613.7640	38.3565
4	X_{14}	addition	0.684	0.661	9.5110	607.8855	36.2442
5	X_1	addition	0.717	0.691	5.2380	603.2015	34.5966
6	X_3	addition	0.735	0.705	3.8650	601.2743	33.7971

Conclusion:

Hence, after performing stepwise variable selection method, we are left with 6 regressors viz X_1, X_2, X_3, X_6, X_9 and X_{14}. So we drop the remaining redundant regressors and continue analysis with these 6 regressors.

Final Model:

Thus our final model is:

$$y_i = \beta_0 + \beta_1 x_{1,j} + \beta_2 x_{2,j} + \beta_3 x_{3,j} + \beta_6 x_{6,j} + \beta_9 x_{9,j} + \beta_{14} x_{14,j} + \epsilon_i, \forall i = 1(1)n$$

where $\epsilon_i \sim N(0, \sigma^2), \forall i = 1(1)60$ under ususal error assumptions.

Final Model Fitting and Summary:

[27]:
```
final.model <- lm(Y ~ X1 + X2 + X3 + X6 + X9 + X14, data=data)
```

Estimates of Regression Coefficients:

Parameters	Estimate	Std. Error	t value	Pr(>|t|)
β_0	1180.3564660	120.45217409	9.799379	1.700813e-13
β_1	1.7969773	0.59896219	3.000151	4.105855e-03
β_2	-1.4835869	0.51363486	-2.888408	5.595751e-03
β_3	-2.3553238	1.24394410	-1.893432	6.376507e-02
β_6	-13.6190403	6.43196019	-2.117401	3.893543e-02
β_9	4.5853483	0.69629249	6.585377	2.094800e-08
β_{14}	0.2596076	0.07837353	3.312440	1.670672e-03

Anova Table:

Parameter	Df	Sum Sq	Mean Sq	F value	Pr(>F)
β_1	1	59266	59266	51.886	2.12e-09
β_2	1	1365	1365	1.195	0.279222
β_3	1	670	670	0.587	0.447063
β_6	1	18887	18887	16.535	0.000159
β_9	1	75047	75047	65.702	7.59e-11
β_{14}	1	12533	12533	10.972	0.001671
Residuals	53	60539	1142	-	-

Model Summary:

- Residual Standard Error: 33.8 (with degrees of freedom 53)
- Multiple R squared: 0.7348
- Adjusted R squared: 0.7048
- F-statistic: 24.48 (with degrees of freedom 6, 53)
- p-value: 1.148×10^{-13}
- $F_{0.95;6,53}$: 2.275 (upper 5% point of F-distribution with df 6, 53)

Test of Regression coefficients:

Now we want to test which variables are significant in this model and which are not. In order to do this we want to test, for each $j = 1(1)p$ (p is the number of regressors in the final model),

$$H_{0j} : \beta_j = 0 \text{ vs } H_1 j : \beta_j \neq 0$$

Let $C = (X'X)^{-1} = ((c_{ij}))_{p\times p}$ and $P_X = X(X'X)^{-1}X'$. Then the test statistic is given by,

$$T_j = \frac{\hat{\beta}_j}{\sqrt{\frac{c_{jj}y'(I-P_X)y}{n-p-1}}} \sim t_{n-p-1}, \text{under } H_0$$

We reject H_0 against H_1 at α% level if and only if the absolute observed value of T_j, i.e., $|T_j| > t_{1-\frac{\alpha}{2},n-p-1}$

Test of linear model:

Now we want to test $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$ against $H_1 :$ not H_0
Let R^2 denotes the multiple R-square of the model.
The test statistic is given by,

$$F = \frac{R^2(n-p-1)}{(1-R^2)p} \sim F_{p,n-p-1}, \text{under } H_0$$

We reject H_0 against H_1 at α% level if and only if observed $F > F_{1-\alpha;p,n-p-1}$.

Conclusion:

From the above summary we get, observed $|T_j| > t_{0.975;53}$ for all j except $j = 3$. Hence we accept H_{03} and conclude that at 5% level, the regressor X_3 is insignificant.

Again, observed $F > F_{0.95;6,53}$. Hence we reject H_0 and conclude that the regression model is significant.

Here the value of adjusted R^2 is 0.7048 implying that more than 70% of the response can be explained by this linear model.

Checking Error Assumption of the Final MLR Model:

Now we check again for normality and homoscedasticity of the final model.

Shapiro-Wilk Test for Normality:

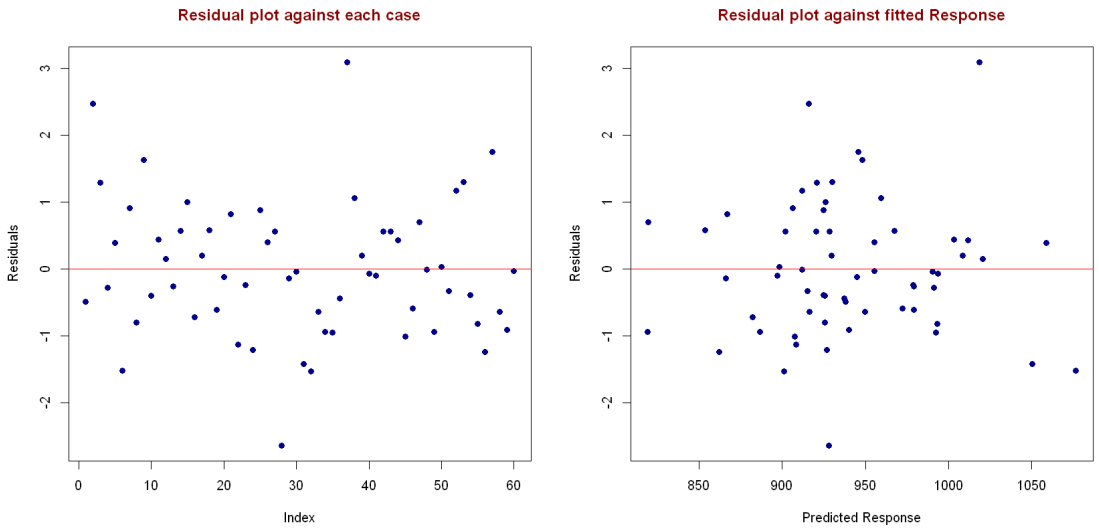
[30]:
```
resid <- rstandard(final.model)
shapiro.test(resid)
```

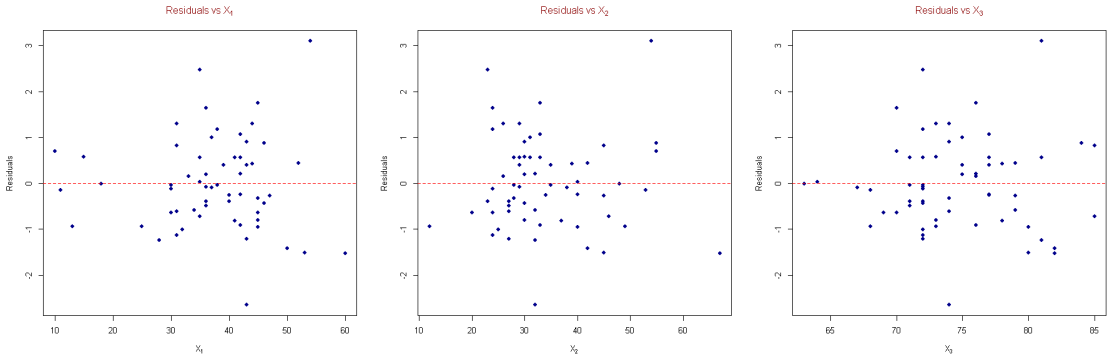
```
        Shapiro-Wilk normality test

data:  resid
W = 0.97987, p-value = 0.423
```

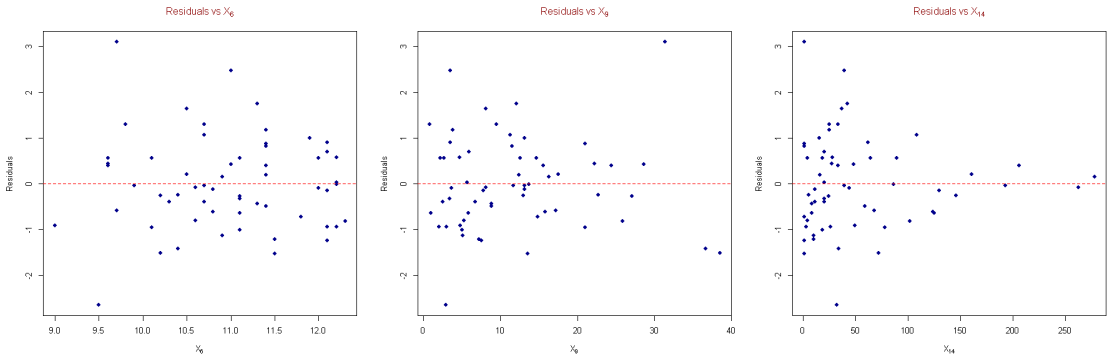
Here the observed p-value (0.423) is greater than the level of significance (0.05). Hence we conclude that the residuals follows a normal distribution.

Checking for Heteroscedasticity:

Now we plot the residuals for each sample, against the fitted response and each of the regressors and perform Goldfeld-Quandt test.







In case of X_{14} there seems to be an inward funnel in the plot. So we are going for Goldfeld-Quandt test for confirmation.

[32]:
```
# Goldfeld-Quandt test
gqtest(Y ~ X1 + X2 + X3 + X6 + X9 + X14, fraction=20, order.by=~-X14, data=data)
```

```
        Goldfeld-Quandt test

data:  Y ~ X1 + X2 + X3 + X6 + X9 + X14
GQ = 1.0865, df1 = 13, df2 = 13, p-value = 0.4417
alternative hypothesis: variance increases from segment 1 to 2
```

Conclusion:

Here, the obsered p-value (0.4417) of the Goldfeld-Quandt test is greater than the level of significance (0.05). Again we can observe that the residual vs regressor plot for each regressor exhibit random behaviour in this new revised model as well. So we conclude that the residuals in the final model are homoscedastic.

Hence our final model satisfies usual error assumptions and does not have multicoliearity among the regressors.

Final Fitted Model:

Our finally fitted Model is :

$$\begin{aligned}\text{MORT} = 1180.356 + 1.797 \cdot \text{PREC} - 1.484 \cdot \text{JANT} - 2.355 \cdot \text{JULT} \\ -13.619 \cdot \text{EDUC} + 4.585 \cdot \text{NONW} + 0.260 \cdot \text{SO2}\end{aligned}$$

Evaluation of the Final Model: The PRESS Statistic

The predicted residual error sum of squares (PRESS) statistic is a form of cross-validation used in regression analysis to provide a summary measure of the fit of a model to a sample of observations that were not themselves used to estimate the model. It is calculated as the sums of squares of the prediction residuals for those observations.

The PRESS Statistic is given by,

$$\text{PRESS} = \sum_{i=1}^{n} \left(\frac{\hat{\epsilon}_i}{1 - h_{ii}} \right)^2$$

where h_{ii} is the i^{th} diagonal element of the matrix $H = X(X'X)^{-1}X'$.

Then the R^2 for prediction based on the PRESS statistic is given by $R^2_{PRED} = 1 - \frac{\text{PRESS}}{\text{SST}}$. So a model with small PRESS statistic value is desired to maximize R^2_{PRED}.

In our model the value of the PRESS statistic is 72051.9 and the value of R^2_{PRED} is 0.6844. Therefore, we can expect this model to explain about 68% of the variability in predicting new observations, as compared to the approximately 67% of the variability in the original data explained by the least- squares fit.

Conclusion:

Based on the final model we can conclude that,

1. Average annual precipitation
2. Average temparature in January and July
3. Median School years completed
4. Percentage of non-white population in urbanised area
5. Relative pollution potential of SO_2

has significant linear effect in prediting the mortality rate of the persons living in that area.

Moreover it is quite natural that this socio economic factors will surely affect the life of a person in a linear fashion. For example if annual precipitation is on the higher side then the area will be more homogenous place to live in but excess or lower rainfall will make lifestyle harder thus affecting lives. Same is true for the temparature or the pollution level in the air. So based on our study we could predict the age adjusted mortality rate given the factors with an accuracy of nearly 68%.

Areas of Improvement:

1. Our dataset contains only 60 values. The more number of data will surely boost our model accuracy but we were unable to collect more data in this aspect.
2. Our model is only able to explian 67% of variablity in the dataset. We can improve the adjusted R-squared value by choosing a more promising set of regressors.

Verdict:

This model is an average model to predict the age adjusted mortality rate of a group of person based on their socio-economic data. We chose best set of varaibles after removing multicolinearity and variable selection but still can't get a accuracy greater than 70%.

Reference:

1. Introduction to linear regression analysis by Douglas c. Montgomery, Elizbeth A . peck,G. Geoffrey vining.
2. Class Notes by Prof. Sharmishta Mitra at IIT Kanpur, Dept. of Mathematics and Statistics
3. Wikipidia
4. Numerous blogposts realted to statistics. Eg. - R-bloggers,STHDA etc.

Appendix: R Codes

```
library(latex2exp)

# retriving data
data<-read.csv('project_data.csv')

# changing columns of the data to suitable format
var_names <- names(data)
colnames(data)<-c('X1', 'X2', 'X3', 'X4', 'X5', 'X6',
                  'X7', 'X8', 'X9', 'X10', 'X11',
                  'X12', 'X13', 'X14', 'X15', 'Y')

# Elementary Data Analysis
for(i in 1:15)
{
    par(mfrow=c(1,2), cex.main=1.25)
    boxplot(data[,i],col="darkblue",freq=TRUE,main='Boxplot')
    hist(data[,i],col="darkmagenta",freq=FALSE,density=30,main='Histogram',
         xlab=TeX(paste0('$X_{', i, '}$')))
    mtext(TeX(paste0("Boxplot and Histogram of $X_{", i, "}$")), line=-20,␣
 ↪outer=TRUE, cex=1.25, col='darkred')
}
par(mfrow=c(1,2))
boxplot(data[,16],col="darkblue",freq=TRUE,main='Boxplot')
hist(data[,16],col="darkmagenta",freq=FALSE,density=30,main='Histogram',
     xlab=TeX(paste0('$Y$')))
mtext("Boxplot and Histogram of Y", line=-20, outer=TRUE, cex=1.25,␣
 ↪col='darkred')

#fitting MLR model
X<-data.matrix(subset(data,select=-c(Y)))
model<-lm(Y ~ .,data=data)

#cooks distance plot
plot(cooks.distance(model),
     pch=16,
     col='darkblue',
     main="Cook's Distance Plot for Residuals",
     col.main='darkred',
     ylab="Cooks Distance",
     xlab="Index")

max(cooks.distance(model))

#qqplot of residuals
resid<-rstandard(model)
```

```r
par(mfrow=c(1,2))
qqnorm(resid, pch=16, col='darkblue',
       main = 'Normal Q-Q plot for residuals',
       col.main='darkred')
qqline(resid, lwd=1.5, col='red')

hist(resid, prob = TRUE, col='darkblue',
     density=30, main = 'Histogram of residuals',
     col.main='darkred', xlab = 'Residuals',)
grid()
lines(density(resid), col='red')

#Shapiro-Wilk test for normality
shapiro.test(resid)

#Anderson-Darling normality test
library(nortest)
ad.test(resid)

# Fitted values
y_hat<-predict(model,subset(data,select = -c(Y)))

par(mfrow=c(1,2))
# residual plot against each case
plot(resid,
     col='darkblue',
     pch=16,
     main='Residual plot against each case',
     col.main='darkred',
     xlab='Index',
     ylab='Residuals')
abline(h=0,col='red')

# residual plot against fitted response
plot(y_hat,resid,
     col='darkblue',
     pch=16,
     main='Residual plot against fitted Response',
     col.main='darkred',
     xlab='Predicted Response',
     ylab='Residuals')
abline(h=0,col='red')

# residual plot against each regressor
for(i in 0:4){
    par(mfrow=c(1,3))
```

```
    for(j in 0:2)
    {
        c = 3*i+j+1
        plot(data[,c],resid,
             col='darkblue',
             pch=16,
             col.main='darkred',
             xlab=TeX(paste0('$X_{', c, '}$')),
             ylab='Residuals',
             main=TeX(paste0('Residuals vs $X_{', c, '}$')))
        abline(h=0,col='red',lty=2)
    }
}

M = matrix(0, nrow=15, ncol=4)
rownames(M) <- colnames(data)[-16]
colnames(M) <- c('-1', '-0.5', '0.5', '1')

absrss <- abs(rstandard(model))
for(i in 1:15){
    M[i, 4] = round(summary(lm(absrss ~ data[, i]))$r.squared, digits=4)
    z = sign(data[, i]) * sqrt(abs(data[, i]))
    M[i, 3] = round(summary(lm(absrss ~ z))$r.squared, digits=4)
    z = sign(data[, i]) / sqrt(abs(data[, i]))
    M[i, 2] = round(summary(lm(absrss ~ z))$r.squared, digits=4)
    z = 1 / data[, i]
    M[i, 1] = round(summary(lm(absrss ~ z))$r.squared, digits=4)
}

#GQ-test for heteroscedasticity
suppressPackageStartupMessages(library(lmtest))
gqtest(Y ~ ., fraction=20, order.by=~-X14, data=data)
qf(0.95, 4, 4)

#plotiing residuals vs lagged residuals graph
plot(resid[-length(resid)],
     resid[-1],
     main = 'Residuals vs Lagged Residuals Plot',
     xlab = "Residuals",
     ylab = "Residuals with lag 1",
     pch=16,
     col='darkblue',
     col.main='darkred')
abline(v=0, lty=3, col='red')
abline(h=0, lty=3, col='red')
sprintf('Correlation coefficient between origand and lagged residuals :␣
 ↪%f',cor(resid[-length(resid)],resid[-1]))
```

```r
X <- data.matrix(subset(data, select=-c(Y)))
Y <- data$Y

# Correlogram of X matrix
suppressPackageStartupMessages(library(ggcorrplot))
cormat<-cor(X)
options(repr.plot.width=7, repr.plot.height=7)
ggcorrplot(cormat,lab=TRUE,type = 'lower',colors = c('red3','yellow','red3'))

# Calculating Condition Number
eigen.values <- eigen(t(X) %*% X)$values
cond.num <- max(eigen.values)/min(eigen.values)
cond.num

library(carData)
library(car)
library(mctest)

data.std <- as.data.frame(sapply(data, scale))
data.std$Y <- data$Y

# Multicolinearity Checking by VIF and Variance decomposition
model.std <- lm(Y~., data=data.std)
t(as.matrix(vif(model.std)))
eigprop(model.std, Inter=F)
summary(lm(X11 ~ X2 + X4 + X9, data=data.std))
summary(lm(X13 ~ X12 + X14, data=data.std))
model.std1 <- lm(Y ~ . - X11 - X13, data=data.std)
t(as.matrix(vif(model.std1)))
eigprop(model.std1, Inter=F)
summary(lm(X4 ~ X1, data = data.std))

# Variable Selection
suppressPackageStartupMessages(library(olsrr))
ols_step_both_p(model.std1)

# Final Model
final.model <- lm(Y ~ X1 + X2 + X3 + X6 + X9 + X14, data=data)
summary(final.model)

# Anova Table
summary(aov(final.model))

# Test of Normality
resid <- rstandard(final.model)
shapiro.test(resid)
```

```r
#fitted values
y_hat<-predict(final.model)

par(mfrow=c(1,2))
# residual plot against each case
plot(resid,
     col='darkblue',
     pch=16,
     main='Residual plot against each case',
     col.main='darkred',
     xlab='Index',
     ylab='Residuals')
abline(h=0,col='red')

# residual plot against fitted response
plot(y_hat,resid,
     col='darkblue',
     pch=16,
     main='Residual plot against fitted Response',
     col.main='darkred',
     xlab='Predicted Response',
     ylab='Residuals')
abline(h=0,col='red')

# residual plot against each regressor
par(mfrow=c(1,3))
for(i in 1:3){
    plot(data[,i],resid,
         col='darkblue',
         pch=16,
         col.main='darkred',
         xlab=TeX(paste0('$X_{', i, '}$')),
         ylab='Residuals',
         main=TeX(paste0('Residuals vs $X_{', i, '}$')))
    abline(h=0,col='red',lty=2)
}

par(mfrow=c(1,3))
for(i in c(6,9,14)){
    plot(data[,i],resid,
         col='darkblue',
         pch=16,
         col.main='darkred',
         xlab=TeX(paste0('$X_{', i, '}$')),
         ylab='Residuals',
         main=TeX(paste0('Residuals vs $X_{', i, '}$')))
```

```
    abline(h=0,col='red',lty=2)
}

# Test of Homoscedasticity
gqtest(final.model)
```