Introduction :-

What is Blood Pressure?

BP or Blood pressure is the force of blood pushing against the walls of the arteries. It is determined by how much blood your heart pumps and how much resistance to the flow of blood occurs in your arteries. It is through the arteries that oxygenated blood is transported throughout your body. Blood pressure is high if resistance to blood flow is high – this means that blood is not able to easily be transported throughout your body.

How is blood pressure measured?

When blood pressure is measured, two numbers are recorded: one for the systolic blood pressure and one for the diastolic blood pressure.

The systolic pressure measures the pressure in the arteries when the heart beats (the heart muscle contracts). This is when your blood pressure is highest. The diastolic pressure measures the pressure in the arteries between heartbeats (the heart muscle is resting and refilling with blood). This is when your blood pressure is lowest.

Blood pressure is usually expressed as a ratio of systolic to diastolic blood pressure. The units of blood pressure are typically millimeters of mercury (mm Hg).

Symptoms and Health Risks

High Blood Pressure (or High BP) i.e. Hypertension increases the risk for other conditions, specifically heart disease and stroke, two of the leading causes of death in the U.S. Hypertension is known as the "silent killer" due to there often being a lack of signs and symptoms associated with its onset; many people do not even know they have it until they experience other health issues! Thus, it is important to have blood pressure measured regularly! In the U.S., about 1 in 3 adults (67 million people) have high blood pressure, but only around 47% have the condition under control.

Factors That Affect BP

BP of a human being can be affected by - Age, Alcohol intake, Chronic conditions (Diabetes, High cholesterol, Kidney disease, Sleep apnea, Pulse) , Diet-Obesity-Weight (Too much salt, Too little potassium, Too little vitamin D) , Family history, Gender-related risk patterns, Physical activity, Stress, Smoking etc.

Description of the Project -

In this project our main objective is to find the relationship between blood pressure (BP), age, weight,body surface area (BSA), duration of hypertension (Dur), Pulse rate and Stress level of an individual .We also find out the outliers in response variables and explanatory variables and also find the influential observation in the data set. We check wheather the heterosecdasticity , autocorrelation , multicollinearity are present in the data or not . we also check wheather the errors of the linear model follows normality assumption or not .And finally we shall find out the best linear regression model on the data set .

So, here collect a data of 20 people, related to BP and some of its causing factors such as Age, Weight, Body Surface Area (BSA), Duration of hypertension (Dur), Pulse & Stress .

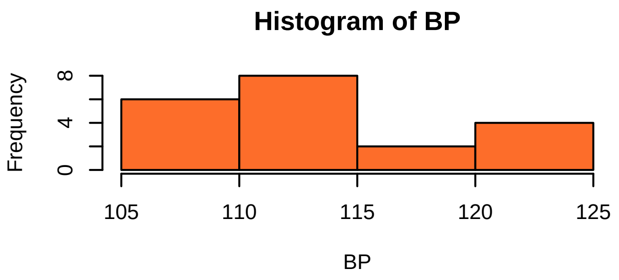
Atfirst, we load the data on R from the .csv file.

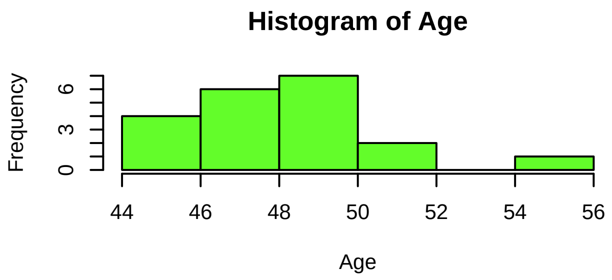
```
   Pt  BP Age Weight  BSA  Dur Pulse Stress
1   1 105  47   85.4 1.75  5.1    63     33
2   2 115  49   94.2 2.10  3.8    70     14
3   3 116  49   95.3 1.98  8.2    72     10
4   4 117  50   94.7 2.01  5.8    73     99
5   5 112  51   89.4 1.89  7.0    72     95
6   6 121  48   99.5 2.25  9.3    71     10
7   7 121  49   99.8 2.25  2.5    69     42
8   8 110  47   90.9 1.90  6.2    66      8
9   9 110  49   89.2 1.83  7.1    69     62
10 10 114  48   92.7 2.07  5.6    64     35
```

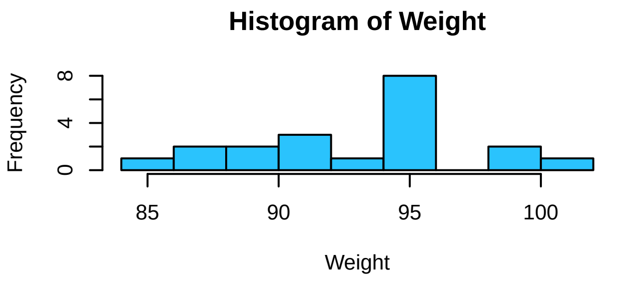
```
11 11 114  47   94.4 2.07  5.3    74     90
12 12 115  49   94.1 1.98  5.6    71     21
13 13 114  50   91.6 2.05 10.2    68     47
14 14 106  45   87.1 1.92  5.6    67     80
15 15 125  52  101.3 2.19 10.0    76     98
16 16 114  46   94.5 1.98  7.4    69     95
17 17 106  46   87.0 1.87  3.6    62     18
18 18 113  46   94.5 1.90  4.3    70     12
19 19 110  48   90.5 1.88  9.0    71     99
20 20 122  56   95.7 2.09  7.0    75     99
       Pt              BP             Age            Weight
 Min.   : 1.00   Min.   :105.0   Min.   :45.00   Min.   : 85.40
 1st Qu.: 5.75   1st Qu.:110.0   1st Qu.:47.00   1st Qu.: 90.22
 Median :10.50   Median :114.0   Median :48.50   Median : 94.15
 Mean   :10.50   Mean   :114.0   Mean   :48.60   Mean   : 93.09
 3rd Qu.:15.25   3rd Qu.:116.2   3rd Qu.:49.25   3rd Qu.: 94.85
 Max.   :20.00   Max.   :125.0   Max.   :56.00   Max.   :101.30
      BSA             Dur            Pulse           Stress
 Min.   :1.750   Min.   : 2.50   Min.   :62.00   Min.   : 8.00
 1st Qu.:1.897   1st Qu.: 5.25   1st Qu.:67.75   1st Qu.:17.00
 Median :1.980   Median : 6.00   Median :70.00   Median :44.50
 Mean   :1.998   Mean   : 6.43   Mean   :69.60   Mean   :53.35
 3rd Qu.:2.075   3rd Qu.: 7.60   3rd Qu.:72.00   3rd Qu.:95.00
 Max.   :2.250   Max.   :10.20   Max.   :76.00   Max.   :99.00
```

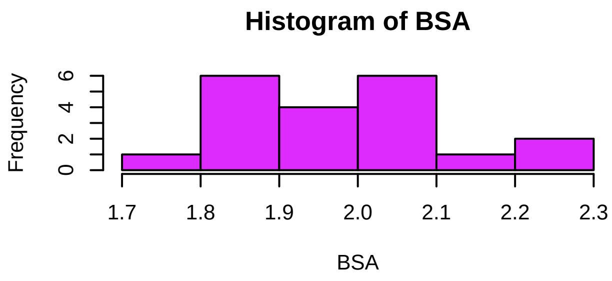
Univariate Data Analysis :-

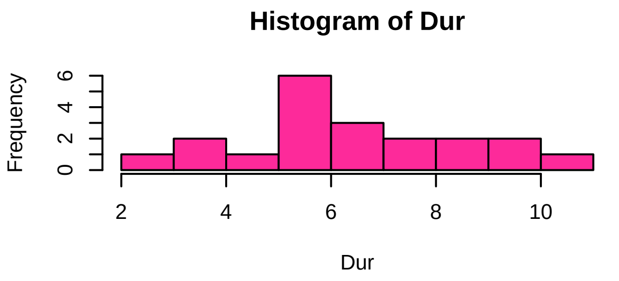
Here, we draw the histograms corresponding to all the variables. And to understand the nature of the distribution nature of the each variables which will helps us to construct the linear model with the help of those variables.

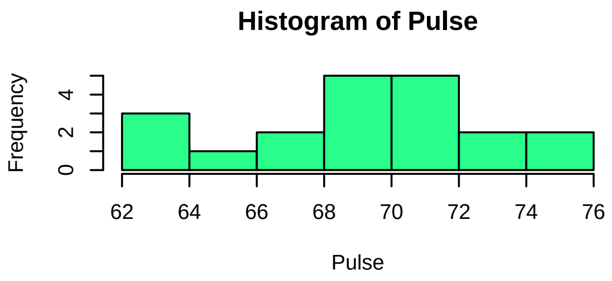


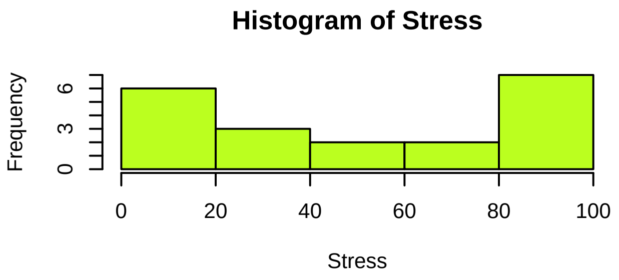






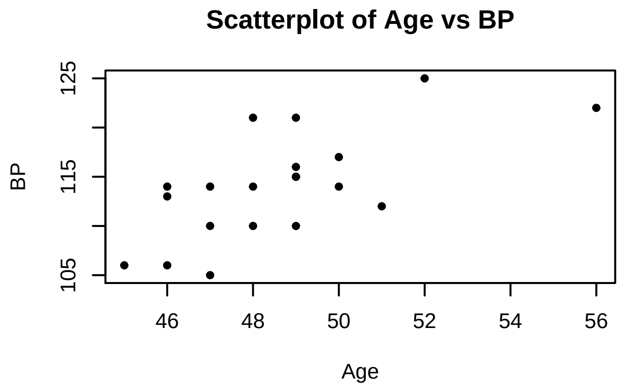


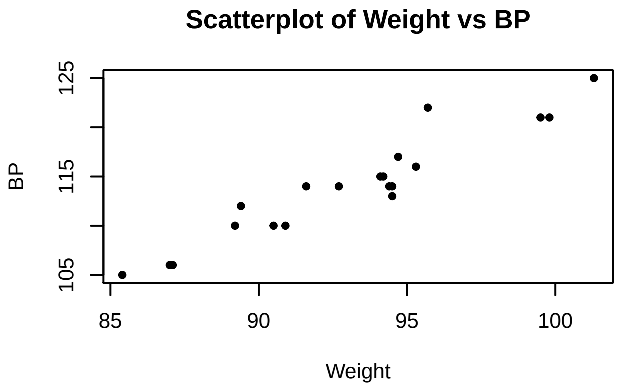


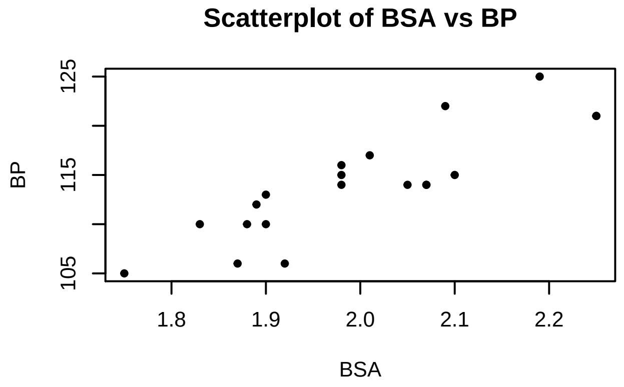


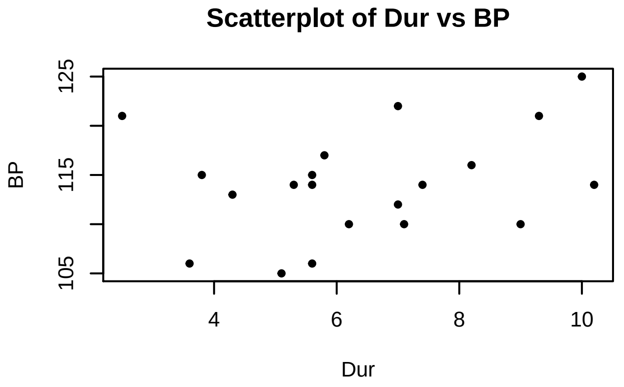
Bivariate Data Analysis :-

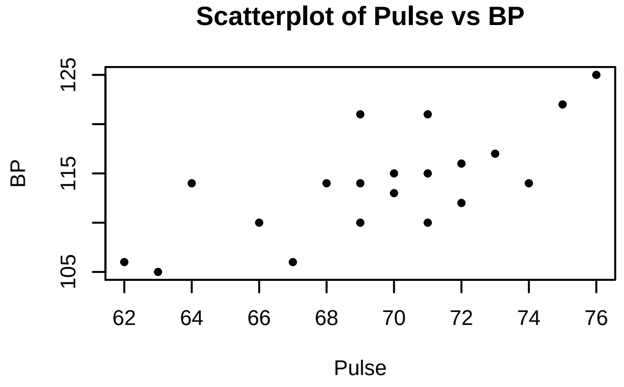
Here, we make a scatterplot between BP and rest of the variables.

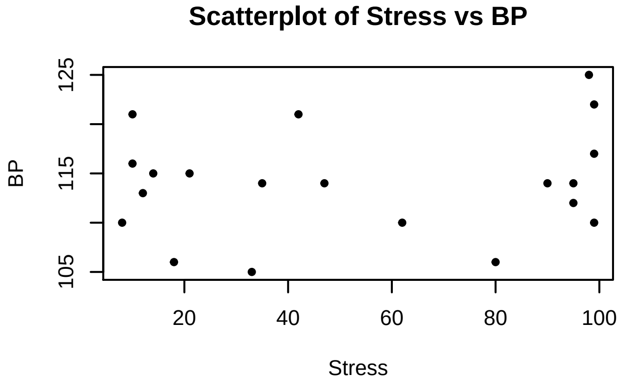

Scatterplot of Age vs BP


Scatterplot of Weight vs BP


Scatterplot of BSA vs BP


Scatterplot of Dur vs BP


Scatterplot of Pulse vs BP


Scatterplot of Stress vs BP

Fitting a Linear Model :-

Here the response variable is BP and explanatory variables are Age, Weight, BSA, Dur, Pulse & Stress.We shall fit a linear model to the data set and we will find the regression coefficents by the help of the Least Square method. We perform this in "R" software.

Here our linear model is of the form -

$y_i = \beta_0+\sum_{j=1}^{6}\beta_j x_{ij}+\varepsilon_i, \forall i = 1(1)20$, where x_{i1},x_{i2},x_{i3},x_{i4},x_{i5}&x_{i6}represents all the covariates "Age", "Weight", "Body Surface Area (BSA)", "Duration of hypertension (Dur)", "Pulse" & "Stress" respectively .

```
Call:
lm(formula = BP ~ Age + Weight + BSA + Dur + Pulse + Stress,
    data = data)
```

```
Coefficients:
(Intercept)          Age       Weight          BSA          Dur        Pulse
 -12.870476     0.703259     0.969920     3.776491     0.068383    -0.084485
     Stress
   0.005572
```

From above, we get the Least-Square(LS) estimates of the model parameters $\beta_i's$.

Hence our linear model is -

$y_i = -12.870476 + 0.703259x_{i1} + 0.969920x_{i2} + 3.776491x_{i3} + 0.068383x_{i4} - 0.084485x_{i5} + 0.005572x_{i6} + \varepsilon_i, \forall i = 1(1)20$.

Detecting outliers & influential observations :-

Now our objective is to find the outliers in the dataset . we shall find the outliers in response variable and explanatoray variables seperatley . outliers in explanatary variables- we shall calculate the hat values .

To find out the outliers in the explanatory variables we shall compute the hatvalues from the data set. we know that here the number of covariates (p) is 6 and number of observations (n)is 20 so $(2 * p)/n = 0.6$.

we know that if the hatvalues of the observations are more than $(2 * p)/n$ then observations corresponding to the hatvalues are X- outliers .

```
        1         2         3         4         5         6         7         8 
0.3130173 0.3385942 0.3213278 0.1860538 0.2566490 0.5022887 0.5404622 0.1727613 
        9        10        11        12        13        14        15        16 
0.1605077 0.3207031 0.4533881 0.1760918 0.4179393 0.4472866 0.4037528 0.4544102 
       17        18        19        20 
0.2854594 0.4338787 0.2478883 0.5675397 
named integer(0)
```

Here we can see that there is not any observation whose hat values are more than 0.6 so we can say that there is no outliers in the explanatory variables.

To find out the outliers in the response variables we shall compute the studentized residulas (rstudent) values from the data set.

we know that if the rstudents of the observations are more than 2 then observations corresponding to the rstudent values are Y- outliers .

```
          1           2           3           4           5           6 
 0.48048277 -0.93199923 -0.09626208  1.36144665  0.64225896  1.52184747 
          7           8           9          10          11          12 
-1.80456478 -1.13654358 -0.04781211  0.62298481 -0.55747445  0.42509089 
         13          14          15          16          17          18 
-0.28971726  1.07940342 -0.09598398  0.58232001 -0.09194255  0.77371379 
         19          20 
-3.72213371  0.28746645 
named integer(0)
```

Here we can see that there is not any observation whose rstudent values are more than 2 so we can say that there is no outliers in the response variables.

Finally we want to check if there exists any influentitional observation in the data set or not. To check this we calculate the cook distance from the data set .

```
Influence measures of
 lm(formula = BP ~ Age + Weight + BSA + Dur + Pulse + Stress,      data = data) :

     dfb.1_  dfb.Age dfb.Wght   dfb.BSA  dfb.Dur  dfb.Puls  dfb.Strs   dffit
1   0.13449  0.10322  0.05530 -0.128531 -0.01325 -0.139124  2.15e-02  0.3243
```

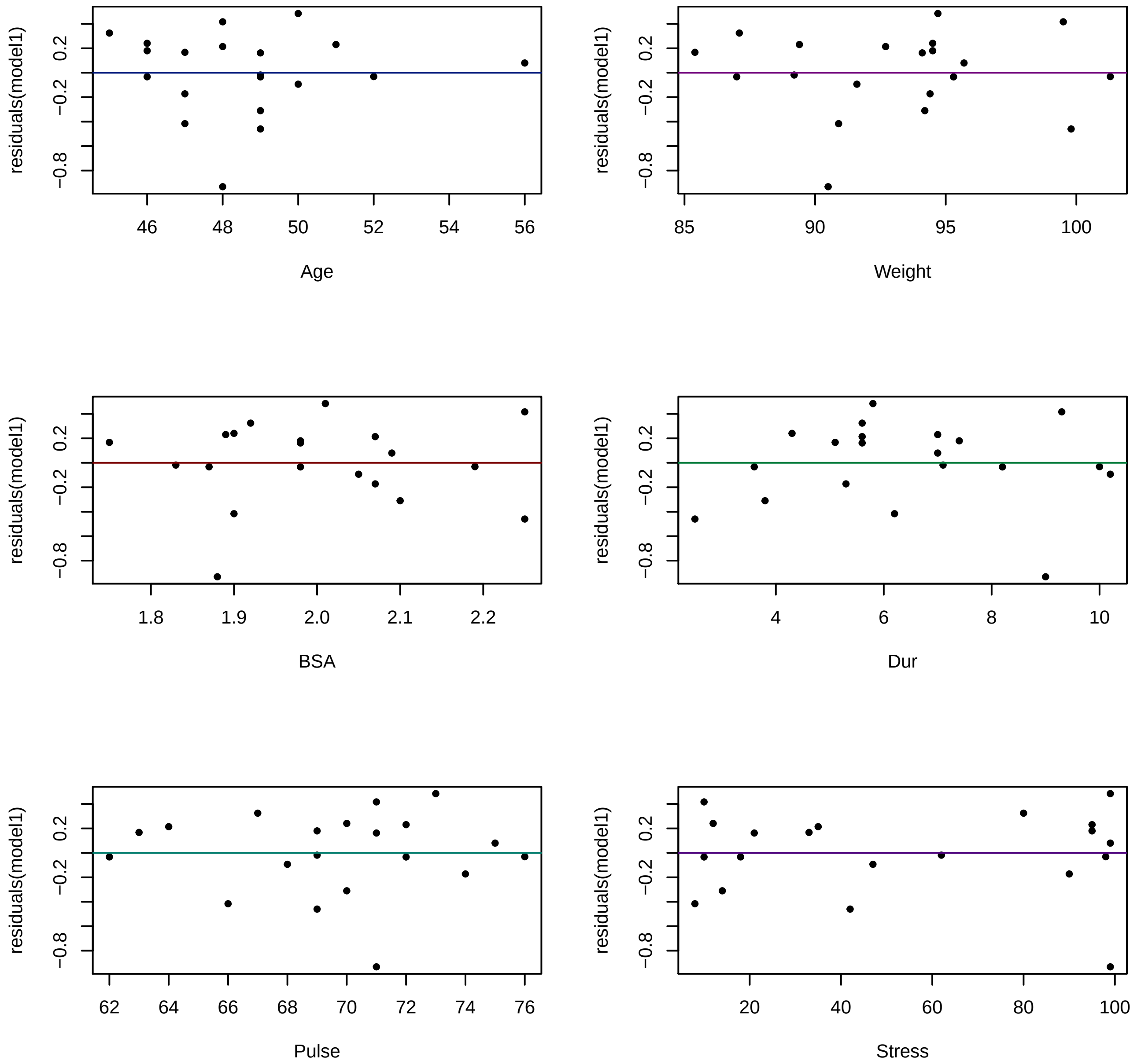
```
2   0.03111 -0.01252  0.37691 -0.369174  0.32080 -0.360029  3.95e-01 -0.6668
3   0.02002  0.00164 -0.00548  0.017939 -0.02255 -0.023215  4.64e-02 -0.0662
4  -0.21113  0.05263  0.16710 -0.170592 -0.29606  0.007650  3.58e-01  0.6509
5   0.03263  0.11262 -0.16334  0.039514 -0.04197  0.140866  1.19e-02  0.3774
6  -0.06035 -0.50836 -0.29341  0.684205  0.83769  0.234797 -6.42e-01  1.5288
7   0.68360 -0.25669 -0.60947 -0.128149  1.14296  0.709869 -5.70e-01 -1.9570
8  -0.08570 -0.03542 -0.12485  0.191312 -0.11116  0.112473  2.15e-01 -0.5194
9  -0.00267 -0.00597 -0.00238  0.010130 -0.00309  0.000863  6.21e-05 -0.0209
10  0.07577  0.11244  0.08073  0.057011  0.02629 -0.316116  1.30e-01  0.4281
11 -0.10482  0.32967  0.24130 -0.239287  0.15749 -0.347769 -2.26e-02 -0.5077
12 -0.06551  0.01920 -0.00465 -0.038057 -0.04973  0.088421 -1.31e-01  0.1965
13 -0.06667 -0.03929  0.09933 -0.113807 -0.17835  0.022187  4.97e-02 -0.2455
14  0.77552 -0.57372 -0.62319  0.583101 -0.02904  0.342979  1.49e-01  0.9710
15  0.04756 -0.01000 -0.03882  0.020508 -0.02918  0.021625 -2.90e-02 -0.0790
16  0.05298 -0.19747  0.33030 -0.227090  0.09962 -0.269746  3.90e-01  0.5314
17 -0.03191 -0.00713  0.00462 -0.003140  0.01533  0.020779 -3.48e-04 -0.0581
18 -0.15190 -0.17031  0.31517 -0.410564 -0.20623  0.100079 -1.69e-01  0.6773
19 -0.69037  0.66139 -0.06495  0.413965 -1.02827 -0.104963 -7.96e-01 -2.1369
20 -0.18064  0.26180 -0.01742  0.000909 -0.07576 -0.011748  2.65e-02  0.3293
    cov.r   cook.d   hat inf
1  2.2308 1.60e-02 0.313
2  1.6233 6.42e-02 0.339
3  2.5664 6.78e-04 0.321
4  0.7872 5.68e-02 0.186
5  1.8595 2.13e-02 0.257
6  1.0230 3.03e-01 0.502
7  0.7098 4.66e-01 0.540   *
8  1.0349 3.77e-02 0.173
9  2.0832 6.76e-05 0.161
10 2.0630 2.75e-02 0.321
11 2.6787 3.89e-02 0.453   *
12 1.9143 5.89e-03 0.176
13 2.8653 9.26e-03 0.418   *
14 1.6563 1.33e-01 0.447
15 2.9213 9.65e-04 0.404   *
16 2.6409 4.25e-02 0.454   *
17 2.4387 5.22e-04 0.285
18 2.2000 6.76e-02 0.434
19 0.0108 3.28e-01 0.248   *
20 3.8594 1.67e-02 0.568   *
```

Since the Cook's distance corresponding to 7th,11th,13th,15th,16th,19th,20th obersvations are large enough. So, we can conclude that 7th,11th,13th,15th,16th,19th,20th paitents' BP are influentitonal .

Detecting Heteroscedasticity :-

To detect whether heteroscedasticity is present or not in the dataset, atfirst we observe the diagnostic plots (each variable vs residuals of the model).

Also, we check whether the average value of residuals is close to zero or not. So, in each plot we also draw a horizontal line with height zero.



From the diagnostic plots, we can not get a clear idea about the on an average value of the residuals (i.e. whether its zero or not). So, here we can not conclude anything about heteroscedasticity.

To confirm whether heteroscedasticity is present or not, here we perform Goldfeld-Quandt test.

Suppose the plot indicate that the error variability (σ_i^2) increases with the j^{th}variable (x_{ji}) .

Step I : Arrange the observation in increasing order.

Step II : Delete c(constant) central observations so that there are now two sets of $\frac{20-c}{2}$observations. (Set 1 comprising of the smaller x_{ji}, and Set 2 comprising of the larger x_{ji}.)

Step III : Assuming constant variance in each set, fit two seperate regressions to the sets and calculate the corresponding residual sum of squares, $RSS_1$and RSS_2, respectively.

Step IV : Since each RSS has the same d.f. $\frac{20-c}{2} - p$, our test statistic $F = \frac{RSS_2}{RSS_1} \sim F_{\frac{20-c}{2}-6, \frac{20-c}{2}-6;\alpha}$

We will reject our null hypothesis i.e. assumption of homoscedasticity if $F_{obs} > F_{\frac{20-c}{2}-6, \frac{20-c}{2}-6;\alpha}$or the corresponding p-value of each test is less than 0.05 .

[Here, we take the constant $c = 6$]

```
Warning:  package 'lmtest' was built under R version 4.0.5
Warning:  package 'zoo' was built under R version 4.0.5


Goldfeld-Quandt test

data:  model1
GQ = 0.26274, df1 = 1, df2 = 1, p-value = 0.6985
alternative hypothesis: variance increases from segment 1 to 2

Goldfeld-Quandt test

data:  model1
GQ = 3.6454, df1 = 1, df2 = 1, p-value = 0.3071
alternative hypothesis: variance increases from segment 1 to 2

Goldfeld-Quandt test

data:  model1
GQ = 60.923, df1 = 1, df2 = 1, p-value = 0.08112
alternative hypothesis: variance increases from segment 1 to 2

Goldfeld-Quandt test

data:  model1
GQ = 53.773, df1 = 1, df2 = 1, p-value = 0.08628
alternative hypothesis: variance increases from segment 1 to 2

Goldfeld-Quandt test

data:  model1
GQ = 233.41, df1 = 1, df2 = 1, p-value = 0.04161
alternative hypothesis: variance increases from segment 1 to 2

Goldfeld-Quandt test

data:  model1
GQ = 2.3491, df1 = 1, df2 = 1, p-value = 0.368
alternative hypothesis: variance increases from segment 1 to 2
```

Here, we perform the test $\forall j = 1, 2, ..., 6$.

From, the p-values of all the tests, we observe that only for the variable “Pulse”(i.e. $j = 5$) the corresponding p-value is less than 0.05.

So, heteroscedasticity arises only when we perform the test based on the “Pulse” variable .

Detecting Multicollinearity :-

We need to check wheather the multicolinearity is present or not in the data set.To check this we have to calculate the condition number for the data set.

Condition Number (CN) $=\sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$, where $\lambda_{max}$and λ_{min} are the maximum and minimum the eigen values of $X^T X$, where X is the design matrix (or model matrix).

```
[1] 4016.581
```

Since CN $>$30 .Hence we can say that there is definite multicolinearity in the data set.

Now we are interst to find out that which variables are responsible for the multicolinearity .To find out this we need to calculate the variance inflation factor (VIF) for the data set.

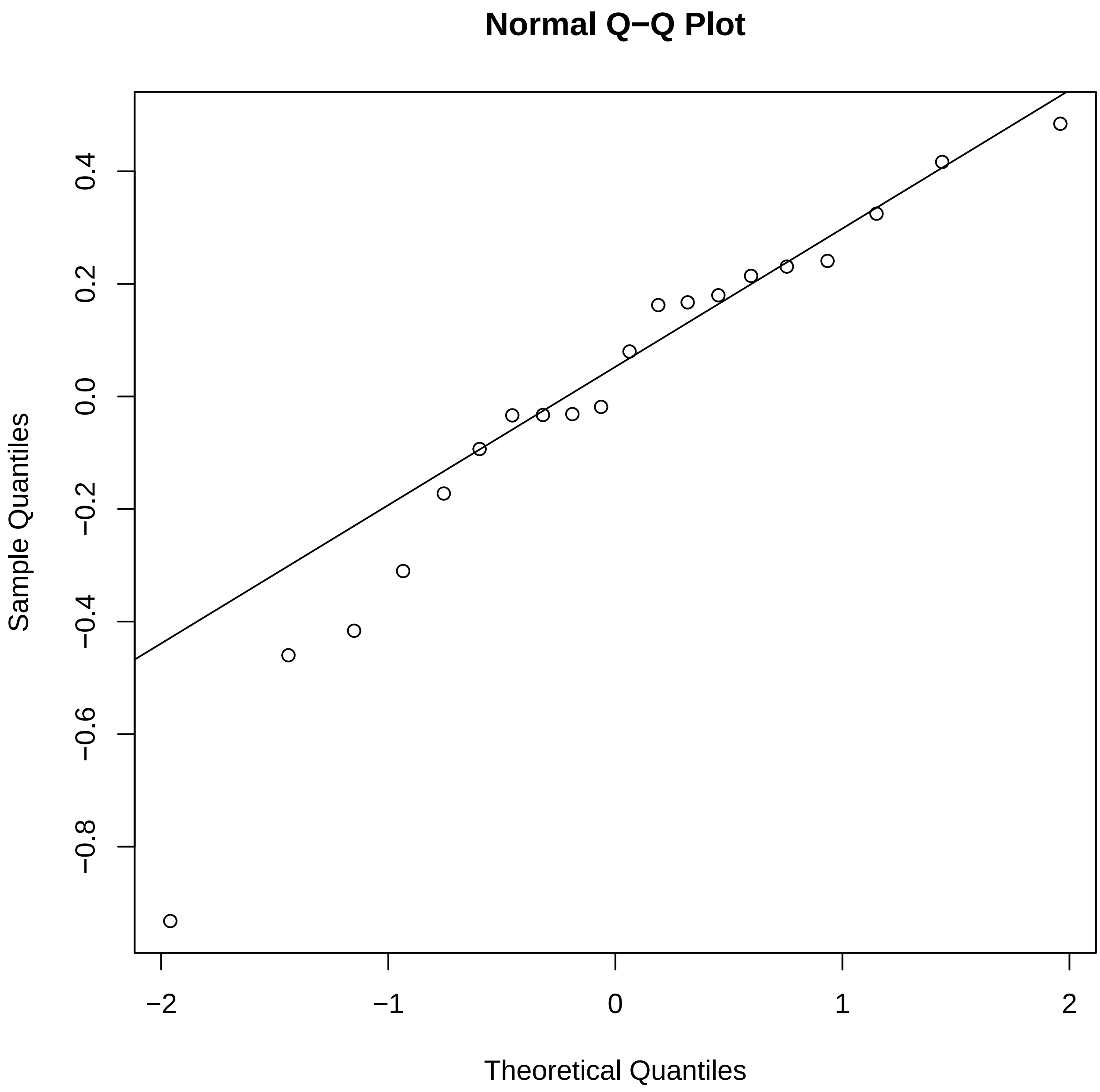
$VIF = \frac{1}{1-R_j^2}$,where $R_j$is the multiple correlation coefficient of $x_j$with $x_1, x_2, ...x_{j-1}, x_{j+1}, ...x_{20}$.

```
[1] 1.762807 8.417035 5.328751 1.237309 4.330443 1.834845
```

Values of VIF corresponding to the covariate weight is higher than the others , Hence we can aviod weight variable to aviod multicolinearity in the data set.

Detecting normality in errors :-

Now we need to check whether the assumption of normality of errors is valid in this case for this we do a Q-Q plot of the residuals of the model:



The normality of errors assumption is vaild in this case suggested by the plot since the sample quantiles are lies on the line of the therotical quantiles.

Further to justify the above mentioned fact Shapiro-Wilk's test is performed to check the normality of errors.

Here, our null hypothesis is that the data came from a normal distribution against the alternative that it doesn't.

```
Shapiro-Wilk normality test

data:  residuals(model1)
W = 0.92408, p-value = 0.1187
```

The p-value of the test is 0.19(>0.05). Hence, we can not reject the null hypothesis. So, we can conclude that the data follows a normal distribution i.e. our assumption of normality of the errors is vaild here.

Detecting Autocorrelation :-

We need to check wheather the Autocorrelation is present or not in the data set.To check this we have to perform Durbin Watson test for Autocorrelation.

Here our null hypothesis is there is no autocorrelation in the data set aganist the alternative that there exists autocorrelation .

Here our test statistics is-
$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=2}^{n} e_{t-1}^2}$, $0 \leq d \leq 4$.
For, postive autocorrelation $d \simeq 0$,
no autocorrelation $d \simeq 2$,
& negative autocorrelation $d \simeq 4$.

```
Durbin-Watson test

data:  model1
DW = 2.2486, p-value = 0.5347
alternative hypothesis: true autocorrelation is not 0
```

Here the value of Durbin Watson test statistics is 2.2486 which is very close to 2 . Hence we can say that there is no autocorrelation.

Also, the p-value of the test is 0.53(>0.05) .Hence we can not reject the null hypothesis. So there does not exist in any autocorrelation in the data set .

Variable Selection Method :-

Now our main objective is to find the best regression model for the data given data set. for that we will use variable selection method to the best regression model.

Using Stepwise regression Method

Atfirst, we fit linear model of BP on every other covariates one by one and observe the values of all the F statistic of the tests.

```
Call:
lm(formula = BP ~ Age, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.7104 -2.9217  0.4276  2.3973  7.8586 
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  44.4545    18.7277   2.374  0.02894 *
Age           1.4310     0.3849   3.718  0.00157 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.195 on 18 degrees of freedom
Multiple R-squared:  0.4344,Adjusted R-squared:  0.403
F-statistic: 13.82 on 1 and 18 DF,  p-value: 0.001574

Call:
lm(formula = BP ~ Weight, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6933 -0.9318 -0.4935  0.7703  4.8656

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.20531    8.66333   0.255    0.802
Weight       1.20093    0.09297  12.917 1.53e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.74 on 18 degrees of freedom
Multiple R-squared:  0.9026,Adjusted R-squared:  0.8972
F-statistic: 166.9 on 1 and 18 DF,  p-value: 1.528e-10

Call:
lm(formula = BP ~ BSA, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-5.314 -1.963 -0.197  1.934  4.831

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   45.183      9.392   4.811  0.00014 ***
BSA           34.443      4.690   7.343 8.11e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.79 on 18 degrees of freedom
Multiple R-squared:  0.7497,Adjusted R-squared:  0.7358
F-statistic: 53.93 on 1 and 18 DF,  p-value: 8.114e-07

Call:
lm(formula = BP ~ Dur, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-8.0144 -3.9963  0.5968  3.0785  9.9124

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

```
(Intercept) 109.2350     3.8563  28.327   <2e-16 ***
Dur           0.7411     0.5703   1.299     0.21
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.333 on 18 degrees of freedom
Multiple R-squared:  0.08575,Adjusted R-squared:  0.03496
F-statistic: 1.688 on 1 and 18 DF,  p-value: 0.2102

Call:
lm(formula = BP ~ Pulse, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.4418 -2.4978 -0.3672  1.8455  7.6179

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   42.323     16.240   2.606 0.017871 *
Pulse          1.030      0.233   4.420 0.000331 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.863 on 18 degrees of freedom
Multiple R-squared:  0.5204,Adjusted R-squared:  0.4938
F-statistic: 19.53 on 1 and 18 DF,  p-value: 0.0003307

Call:
lm(formula = BP ~ Stress, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-8.6394 -3.3014  0.0722  2.2181  9.9287

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 112.71997    2.19345  51.389   <2e-16 ***
Stress        0.02399    0.03404   0.705     0.49
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.502 on 18 degrees of freedom
Multiple R-squared:  0.02686,Adjusted R-squared:  -0.0272
F-statistic: 0.4969 on 1 and 18 DF,  p-value: 0.4899
```

Covariate Weight is significant since the value of F statsistic corresponding to Weight is large than others.

```
Call:
lm(formula = BP ~ Weight + Age, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.89968 -0.35242  0.06979  0.35528  0.82781

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

```
(Intercept) -16.57937    3.00746  -5.513 3.80e-05 ***
Weight        1.03296    0.03116  33.154  < 2e-16 ***
Age           0.70825    0.05351  13.235 2.22e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5327 on 17 degrees of freedom
Multiple R-squared:  0.9914,Adjusted R-squared:  0.9904
F-statistic: 978.2 on 2 and 17 DF,  p-value: < 2.2e-16

Call:
lm(formula = BP ~ Weight + BSA, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8932 -1.1961 -0.4061  1.0764  4.7524

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.6534     9.3925   0.602    0.555
Weight        1.0387     0.1927   5.392 4.87e-05 ***
BSA           5.8313     6.0627   0.962    0.350
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.744 on 17 degrees of freedom
Multiple R-squared:  0.9077,Adjusted R-squared:  0.8968
F-statistic: 83.54 on 2 and 17 DF,  p-value: 1.607e-09

Call:
lm(formula = BP ~ Weight + Dur, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.0812 -1.2496  0.1100  0.4747  4.7825

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.98677    8.41850   0.355    0.727
Weight       1.17392    0.09203  12.756 3.93e-10 ***
Dur          0.26949    0.18425   1.463    0.162
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.688 on 17 degrees of freedom
Multiple R-squared:  0.9135,Adjusted R-squared:  0.9033
F-statistic: 89.78 on 2 and 17 DF,  p-value: 9.207e-10

Call:
lm(formula = BP ~ Weight + Pulse, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5918 -0.9456 -0.0553  0.7448  3.9357

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.4535     8.3937  -0.173   0.8646
Weight        1.0609     0.1163   9.118 5.88e-08 ***
Pulse         0.2399     0.1314   1.826   0.0855 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.638 on 17 degrees of freedom
Multiple R-squared:  0.9186,Adjusted R-squared:  0.909
F-statistic: 95.91 on 2 and 17 DF,  p-value: 5.503e-10

Call:
lm(formula = BP ~ Weight + Stress, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4865 -0.9395  0.1950  0.5080  4.0023

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.71023    8.09054   0.211   0.8351
Weight       1.19522    0.08683  13.765  1.2e-10 ***
Stress       0.01924    0.01006   1.913   0.0727 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.625 on 17 degrees of freedom
Multiple R-squared:  0.9199,Adjusted R-squared:  0.9105
F-statistic: 97.59 on 2 and 17 DF,  p-value: 4.807e-10
```

Covariates Weight and Age is significant since the value of F statsistics corresponding to Weight and Age is larger than others.

```
Call:
lm(formula = BP ~ Weight + Age + BSA, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.75810 -0.24872  0.01925  0.29509  0.63030

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -13.66725    2.64664  -5.164 9.42e-05 ***
Weight        0.90582    0.04899  18.490 3.20e-12 ***
Age           0.70162    0.04396  15.961 3.00e-11 ***
BSA           4.62739    1.52107   3.042  0.00776 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.437 on 16 degrees of freedom
Multiple R-squared:  0.9945,Adjusted R-squared:  0.9935
F-statistic: 971.9 on 3 and 16 DF,  p-value: < 2.2e-16

Call:
lm(formula = BP ~ Weight + Age + Dur, data = data)
```

```
Residuals:
     Min       1Q   Median       3Q      Max 
-1.03592 -0.29671  0.05216  0.32551  0.85934 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -16.09486    3.10435  -5.185 9.04e-05 ***
Weight        1.03121    0.03159  32.639 4.54e-16 ***
Age           0.69526    0.05661  12.280 1.47e-09 ***
Dur           0.04821    0.06152   0.784    0.445    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5388 on 16 degrees of freedom
Multiple R-squared:  0.9917,Adjusted R-squared:  0.9901 
F-statistic: 637.6 on 3 and 16 DF,  p-value: < 2.2e-16

Call:
lm(formula = BP ~ Weight + Age + Pulse, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.71174 -0.45422 -0.01909  0.41745  0.88743 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -16.69000    2.93761  -5.681 3.40e-05 ***
Weight        1.06135    0.03695  28.722 3.40e-15 ***
Age           0.75018    0.06074  12.350 1.36e-09 ***
Pulse        -0.06566    0.04852  -1.353    0.195    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5201 on 16 degrees of freedom
Multiple R-squared:  0.9923,Adjusted R-squared:  0.9908 
F-statistic: 684.7 on 3 and 16 DF,  p-value: < 2.2e-16

Call:
lm(formula = BP ~ Weight + Age + Stress, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.0252 -0.3277  0.0368  0.2274  0.8901 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -16.196316   3.090002  -5.242 8.07e-05 ***
Weight        1.036206   0.031865  32.518 4.82e-16 ***
Age           0.691179   0.058833  11.748 2.80e-09 ***
Stress        0.002710   0.003625   0.748    0.465    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5397 on 16 degrees of freedom
Multiple R-squared:  0.9917,Adjusted R-squared:  0.9901 
F-statistic: 635.4 on 3 and 16 DF,  p-value: < 2.2e-16
```

Covariates Weight, Age, BSA is significant since the value of F statsistics corresponding to Weight, Age and BSA is larger than others.

```
Call:
lm(formula = BP ~ Weight + Age + BSA + Dur, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.86420 -0.26320  0.08341  0.25020  0.58272 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -12.85206    2.64804  -4.853 0.000211 ***
Weight        0.89701    0.04818  18.618 8.88e-12 ***
Age           0.68335    0.04490  15.220 1.58e-10 ***
BSA           4.86037    1.49220   3.257 0.005305 ** 
Dur           0.06653    0.04895   1.359 0.194184    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4259 on 15 degrees of freedom
Multiple R-squared:  0.9951,Adjusted R-squared:  0.9938 
F-statistic:   768 on 4 and 15 DF,  p-value: < 2.2e-16

Call:
lm(formula = BP ~ Weight + Age + BSA + Pulse, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.74386 -0.22036  0.01533  0.30937  0.62375 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -13.88934    2.75905  -5.034 0.000148 ***
Weight        0.92288    0.06284  14.686 2.62e-10 ***
Age           0.71516    0.05412  13.213 1.15e-09 ***
BSA           4.32945    1.69374   2.556 0.021929 *  
Pulse        -0.02053    0.04539  -0.452 0.657579    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4483 on 15 degrees of freedom
Multiple R-squared:  0.9946,Adjusted R-squared:  0.9932 
F-statistic: 692.8 on 4 and 15 DF,  p-value: < 2.2e-16

Call:
lm(formula = BP ~ Weight + Age + BSA + Stress, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.8287 -0.1228  0.0521  0.2163  0.5427 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -13.175722   2.673818  -4.928 0.000182 ***
Weight        0.907477   0.048785  18.602 8.99e-12 ***
```

```
Age           0.681748   0.047514  14.348 3.63e-10 ***
BSA           4.703864   1.515632   3.104 0.007264 **
Stress        0.003137   0.002925   1.073 0.300366
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.435 on 15 degrees of freedom
Multiple R-squared:  0.9949,Adjusted R-squared:  0.9936
F-statistic: 736.1 on 4 and 15 DF,  p-value: < 2.2e-16
```

Covariates Weight, Age, BSA and Dur is significant since the value of F statsistics corresponding to Weight, Age, BSA and Dur is larger than others.

```
Call:
lm(formula = BP ~ Weight + Age + BSA + Dur + Pulse, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.80879 -0.20964  0.05041  0.22437  0.64247

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -13.12652    2.69940  -4.863 0.000251 ***
Weight        0.92609    0.06042  15.328 3.82e-10 ***
Age           0.70475    0.05247  13.433 2.17e-09 ***
BSA           4.36419    1.62758   2.681 0.017897 *
Dur           0.07642    0.05098   1.499 0.156036
Pulse        -0.03657    0.04491  -0.814 0.429059
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4308 on 14 degrees of freedom
Multiple R-squared:  0.9954,Adjusted R-squared:  0.9937
F-statistic: 600.7 on 5 and 14 DF,  p-value: 8.185e-16

Call:
lm(formula = BP ~ Weight + Age + BSA + Dur + Stress, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9484 -0.1951  0.0895  0.2137  0.4930

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -12.587398   2.699492  -4.663 0.000366 ***
Weight        0.899467   0.048847  18.414 3.29e-11 ***
Age           0.670659   0.048081  13.948 1.33e-09 ***
BSA           4.887143   1.510274   3.236 0.005978 **
Dur           0.057500   0.050780   1.132 0.276520
Stress        0.002396   0.002971   0.806 0.433510
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.431 on 14 degrees of freedom
Multiple R-squared:  0.9954,Adjusted R-squared:  0.9937
F-statistic: 600.2 on 5 and 14 DF,  p-value: 8.236e-16
```

Now the Pulse and Stress are insignificant since the value of F statsistics corresponding to Pulse and Stress is lesser than others.

The best model according to Stepwise regression method is $y = \alpha + \beta * Weight + \gamma * Age + \delta * BSA + \eta * Dur + \epsilon$ where α, β, γ,$\delta, \eta$are the regression parameters and $\epsilon$is the random error in the model.

Now we have to estimate the values of the unknown parameters by the help method of least squares

```
Call:
lm(formula = BP ~ Weight + Age + BSA + Dur, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.86420 -0.26320  0.08341  0.25020  0.58272

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -12.85206    2.64804  -4.853 0.000211 ***
Weight        0.89701    0.04818  18.618 8.88e-12 ***
Age           0.68335    0.04490  15.220 1.58e-10 ***
BSA           4.86037    1.49220   3.257 0.005305 **
Dur           0.06653    0.04895   1.359 0.194184
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4259 on 15 degrees of freedom
Multiple R-squared:  0.9951,Adjusted R-squared:  0.9938
F-statistic:   768 on 4 and 15 DF,  p-value: < 2.2e-16
```

Hence our final model is -
$y = -12.85206 + 0.89701 * weight$+$0.68335 * Age + 4.86037 * BSA$+$0.06653 * Dur$+$\epsilon$.

Conclusion:-

About the data set-

After implementing the Regression Diagnostics on the data set, we observe that the data set doesn't contain any kind of outliers in explanatory variables as well as response variables. But, we get some influential observation from the data set. Not only that, there doesn't exist in any autocorrelation in the data set . But, heteroscedasticity is present in the data set when we worked on the basis of the factor "Pulse". Also, multicollinearity is also present in the data set due to the factor "Weight". Finally, we observe that the errors are normally distributed. Atlast, we performed the variable selection method which leads us to detect the four most important covariates from six covariates viz "Age", "Weight", "BSA" & "Dur".

On the aspect of the project

From the project, we conclude that all the six factors are responsible for high BP in human body. However, in the end we provide the four most important reasons among the six factors which are age, weight, body surface area and duration of hypertension of an individual.

Appendix :-

(The used R codes)

```r
#... load and attach the dataset.
data=read.csv("F:\\Downloads\\data_set.csv",header=T,sep=",")
data
summary(data)
attach(data)

#...univariate data analysis.
par(mfrow=c(4,2))
hist(BP,col="#FD6D2A")
hist(Age,col="#63FD2A")
hist(Weight,col="#2AC3FD")
hist(BSA,col="#DD2AFD")
hist(Dur,col="#FD2A9A")
hist(Pulse,col="#2AFD8A")
hist(Stress,col="#BBFF1F")
par(mfrow=c(1,1))

#...bivariate data analysis
par(mfrow=c(3,2))
plot(Age,BP,main="Scatterplot of Age vs BP",pch=20)
plot(Weight,BP,main="Scatterplot of Weight vs BP",pch=20)
plot(BSA,BP,main="Scatterplot of BSA vs BP",pch=20)
plot(Dur,BP,main="Scatterplot of Dur vs BP",pch=20)
plot(Pulse,BP,main="Scatterplot of Pulse vs BP",pch=20)
plot(Stress,BP,main="Scatterplot of Stress vs BP",pch=20)


#...linear model.
model1=lm(BP~Age+Weight+BSA+Dur+Pulse+Stress,data=data)
model1


#...hatvalues
hat=hatvalues(model1)
which(hat>0.6)


#...studentized residuals.
rstudent=rstudent(model1)
rstudent
which(rstudent>2)

#....influentational observations.
influence.measures(model1)

#....residuals vs covariates plot.
par(mfrow=c(3,2))
plot(Age,residuals(model1),pch=20)
abline(h=0,col="#01197C")
plot(Weight,residuals(model1),pch=20)
abline(h=0,col="#71017C")
plot(BSA,residuals(model1),pch=20)
```

```
abline(h=0,col="#7C0101")
plot(Dur,residuals(model1),pch=20)
abline(h=0,col="#017C3B")
plot(Pulse,residuals(model1),pch=20)
abline(h=0,col="#017C6D")
plot(Stress,residuals(model1),pch=20)
abline(h=0,col="#4C017C")

#...test for heteroscedasticity
library(lmtest)
gqtest(model1,order.by=Age,data=data,fraction=4)
gqtest(model1,order.by=Weight,data=data,fraction=4)
gqtest(model1,order.by=BSA,data=data,fraction=4)
gqtest(model1,order.by=Dur,data=data,fraction=4)
gqtest(model1,order.by=Pulse,data=data,fraction=4)
gqtest(model1,order.by=Weight,data=Stress,fraction=4)

#....condition number for multicolinearity .
X=model.matrix(model1)
M=t(X)%*%X
lambda=eigen(M)
M.lambda=max(lambda$values)
m.lambda=min(lambda$values)
CN=sqrt(M.lambda/m.lambda)
CN

#...variance inflation factors for multicolinearity .
lm1=lm( Age~Weight+BSA+Dur+Pulse+Stress,data=data)
R1=cor(data[,3],fitted(lm1))
lm2=lm(Weight~Age+BSA+Dur+Pulse+Stress,data=data)
R2=cor(data[,4],fitted(lm2))
lm3=lm(BSA~Weight+Age+Dur+Pulse+Stress,data=data)
R3=cor(data[,5],fitted(lm3))
lm4=lm(Dur~Age+Weight+BSA+Pulse+Stress,data=data)
R4=cor(data[,6],fitted(lm4))
lm5=lm(Pulse~Age+Weight+BSA+Age+Stress,data=data)
R5=cor(data[,7],fitted(lm5))
lm6=lm(Stress~Weight+BSA+Dur+Pulse+Age,data=data)
R6=cor(data[,8],fitted(lm6))
R=c(R1,R2,R3,R4,R5,R6)
VIF=1/(1-(R)^2)
VIF

#...Q-Q plot for normality
qqnorm(residuals(model1))
qqline(residuals(model1))

#..test for normality
shapiro.test(residuals(model1))

#...test for autocorrelation
library(lmtest)
dwtest(model1,alternative="two.sided")

#....stepwise regression method.
```

```r
lm1=lm(BP~Age,data=data)
summary(lm1)
lm2=lm(BP~Weight,data=data)
summary(lm2)
lm3=lm(BP~BSA,data=data)
summary(lm3)
lm4=lm(BP~Dur,data=data)
summary(lm4)
lm5=lm(BP~Pulse,data=data)
summary(lm5)
lm6=lm(BP~Stress,data=data)
summary(lm6)

lm1=lm(BP~Weight+Age,data=data)
summary(lm1)
lm2=lm(BP~Weight+BSA,data=data)
summary(lm2)
lm3=lm(BP~Weight+Dur,data=data)
summary(lm3)
lm4=lm(BP~Weight+Pulse,data=data)
summary(lm4)
lm5=lm(BP~Weight+Stress,data=data)
summary(lm5)

lm1=lm(BP~Weight+Age+BSA,data=data)
summary(lm1)
lm2=lm(BP~Weight+Age+Dur,data=data)
summary(lm2)
lm3=lm(BP~Weight+Age+Pulse,data=data)
summary(lm3)
lm4=lm(BP~Weight+Age+Stress,data=data)
summary(lm4)

lm1=lm(BP~Weight+Age+BSA+Dur,data=data)
summary(lm1)
lm2=lm(BP~Weight+Age+BSA+Pulse,data=data)
summary(lm2)
lm3=lm(BP~Weight+Age+BSA+Stress,data=data)
summary(lm3)

lm1=lm(BP~Weight+Age+BSA+Dur+Pulse,data=data)
summary(lm1)
lm2=lm(BP~Weight+Age+BSA+Dur+Stress,data=data)
summary(lm2)

lm1=lm(BP~Weight+Age+BSA+Dur,data=data)
summary(lm1)
```

Acknowledgement:-

I would want to convey my heartfelt gratitude to Prof. ...., my mentor, for his invaluable advice and assistance in completing my project. He was there to assist me every step of the way, and his motivation is what enabled me to accomplish my task effectively. I would also like to thank all of the other supporting personnel who assisted me by supplying the equipment that was essential and vital, without which I would not have been able to perform efficiently on this project.

I'd also like to thank my friends and parents for their support and encouragement as I worked on this assignment.