STAT363 Homework 3

Berfin AKDEMİR, Zeynep KANAR

2024-08-26

INTRODUCTION

Weight is an important indicator of health and plays a significant role in managing obesity and other weight-related health issues. Understanding the factors affecting weight is crucial for developing strategies for healthy living and weight management. This study aims to investigate the effects of variables such as blood pressure, age, body surface area (BSA), duration of hypertension, pulse, and stress on weight. Identifying the impact of these variables on weight management can aid in developing data-driven strategies to improve individuals' health.

This research focuses on identifying the factors affecting weight. Specifically, it analyzes the effects of variables such as blood pressure, age, body surface area, duration of hypertension, pulse, and stress on weight.

Research Question: How do factors such as blood pressure, age, body surface area, duration of hypertension, pulse, and stress affect weight? Which variables have a significant relationship with weight, and how can this information be used in weight management strategies?

Variable information as follows: Age: Age can influence metabolism rate and its effects on weight. Body Surface Area (BSA): Body surface area is related to energy requirements and may impact weight. Duration of Hypertension: The duration of hypertension can indirectly affect weight management. Pulse: Pulse rate can influence metabolism rate, which in turn can affect weight. Stress: Stress can have a direct impact on weight management. Gender: Gender may be associated with biological and genetic factors that affect weight. Blood Pressure: Blood pressure can influence metabolic processes related to weight.

ANALYSIS & TESTING

We started by reading and observing the data

```r
blood <- read.table("/Users/berfinakdemir/Desktop/STAT363/Homework/HW3/bloodpress.txt", header = T)
```

```r
head(blood)
```

```
##   Pt  BP Age Weight  BSA Dur Pulse Stress Gender
## 1  1 105  47   85.4 1.75 5.1    63     33   Male
## 2  2 115  49   94.2 2.10 3.8    70     14   Male
## 3  3 116  49   95.3 1.98 8.2    72     10 Female
## 4  4 117  50   94.7 2.01 5.8    73     99   Male
## 5  5 112  51   89.4 1.89 7.0    72     95 Female
## 6  6 121  48   99.5 2.25 9.3    71     10 Female
```

```r
str(blood)
```

```
## 'data.frame':    20 obs. of  9 variables:
##  $ Pt    : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ BP    : int  105 115 116 117 112 121 121 110 110 114 ...
##  $ Age   : int  47 49 49 50 51 48 49 47 49 48 ...
##  $ Weight: num  85.4 94.2 95.3 94.7 89.4 99.5 99.8 90.9 89.2 92.7 ...
##  $ BSA   : num  1.75 2.1 1.98 2.01 1.89 2.25 2.25 1.9 1.83 2.07 ...
##  $ Dur   : num  5.1 3.8 8.2 5.8 7 9.3 2.5 6.2 7.1 5.6 ...
##  $ Pulse : int  63 70 72 73 72 71 69 66 69 64 ...
##  $ Stress: int  33 14 10 99 95 10 42 8 62 35 ...
##  $ Gender: chr  "Male" "Male" "Female" "Male" ...
```

Then prepared our dummy variables

```
blood$Gender <- as.factor(blood$Gender)
blood$Gender <- relevel(blood$Gender, ref = "Male")
```

Checking Multiple Linear Regression Assumptions

Linearity of the data: The relationship between the predictor (x) and the outcome (y) is assumed to be linear.

Simplest way to see if there is linear relationship is with scatter plots. We have also conducted correlation tests and tables to choose which of the variables we are going to use.

We observed the data thoroughly with tests and applied necessary transformations as we go on the assumptions throughout this study.

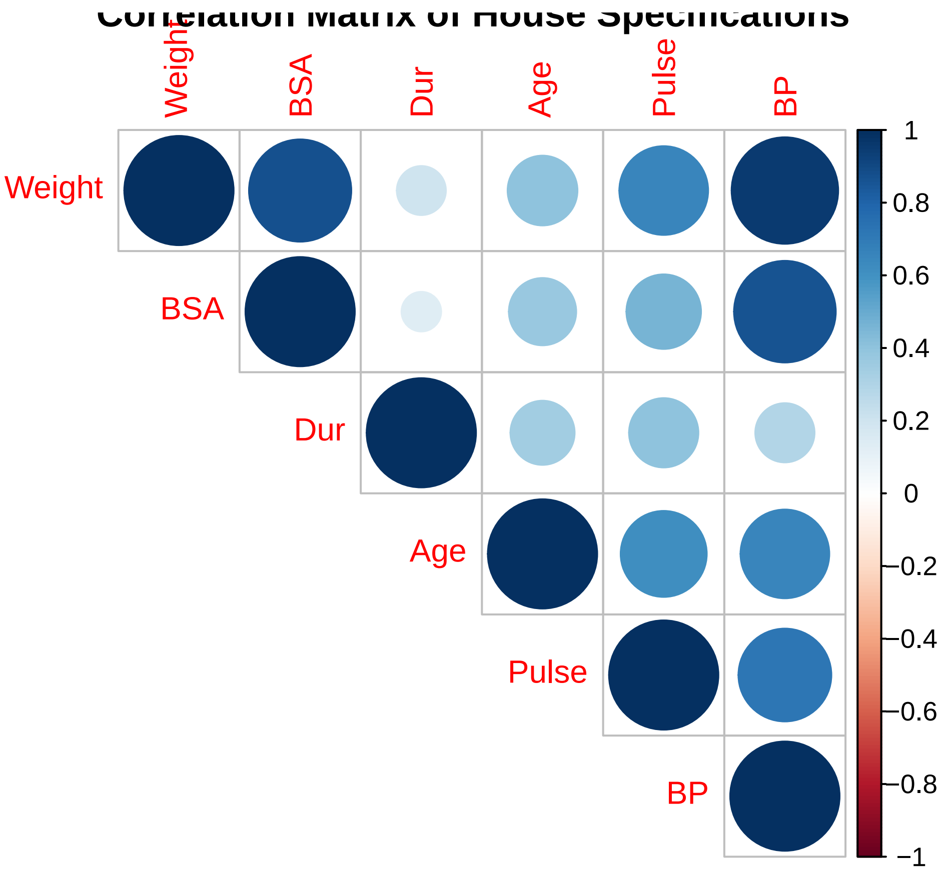
```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

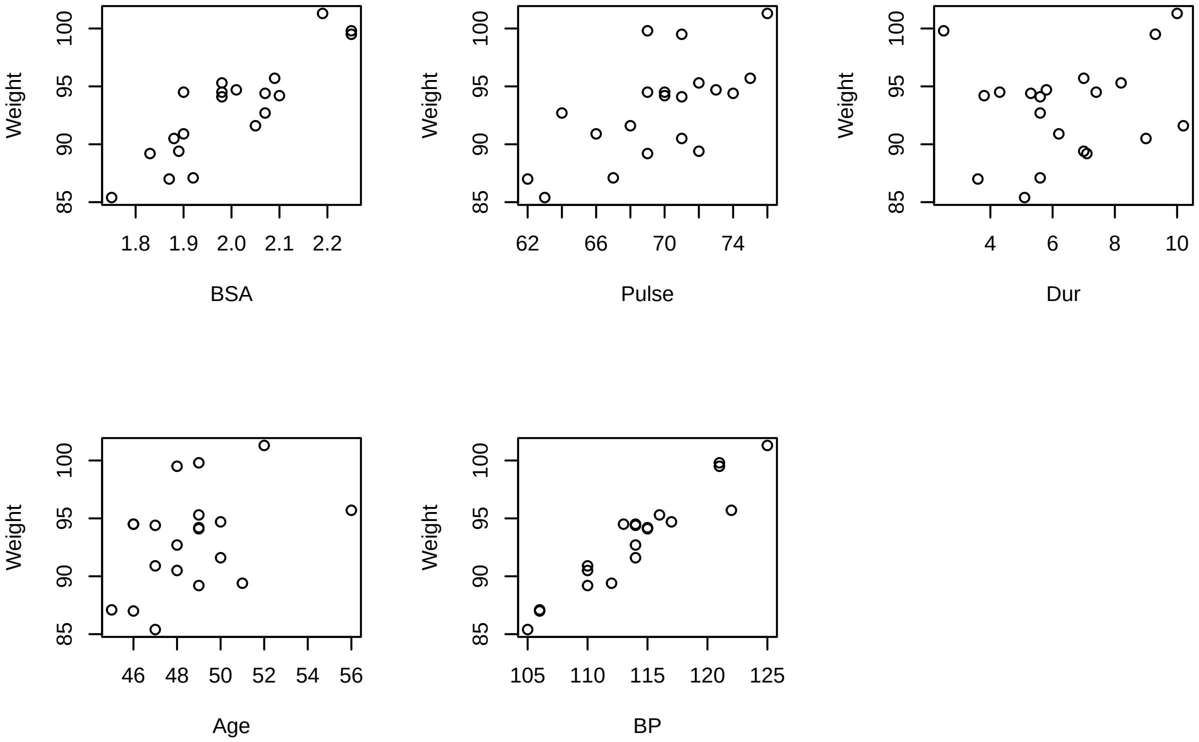
```
numer <- blood[, c("Weight", "BSA", "Dur", "Age", "Pulse", "BP")]
cor_matrix <- cor(numer)
print(cor_matrix)
```

```
##           Weight       BSA       Dur       Age     Pulse        BP
## Weight 1.0000000 0.8753048 0.2006496 0.4073493 0.6593399 0.9500677
## BSA    0.8753048 1.0000000 0.1305400 0.3784546 0.4648188 0.8658789
## Dur    0.2006496 0.1305400 1.0000000 0.3437921 0.4015144 0.2928336
## Age    0.4073493 0.3784546 0.3437921 1.0000000 0.6187643 0.6590930
## Pulse  0.6593399 0.4648188 0.4015144 0.6187643 1.0000000 0.7214132
## BP     0.9500677 0.8658789 0.2928336 0.6590930 0.7214132 1.0000000
```

```
par(mfrow=c(1,1))
corrplot(cor_matrix, , type = "upper",
         title = "Correlation Matrix of House Specifications")
```



```
par(mfrow = c(2,3))
plot(Weight ~ BSA, data = blood)
plot(Weight ~ Pulse, data = blood)
plot(Weight ~ Dur, data = blood)
plot(Weight ~ Age, data = blood)
plot(Weight ~ BP, data = blood)
```



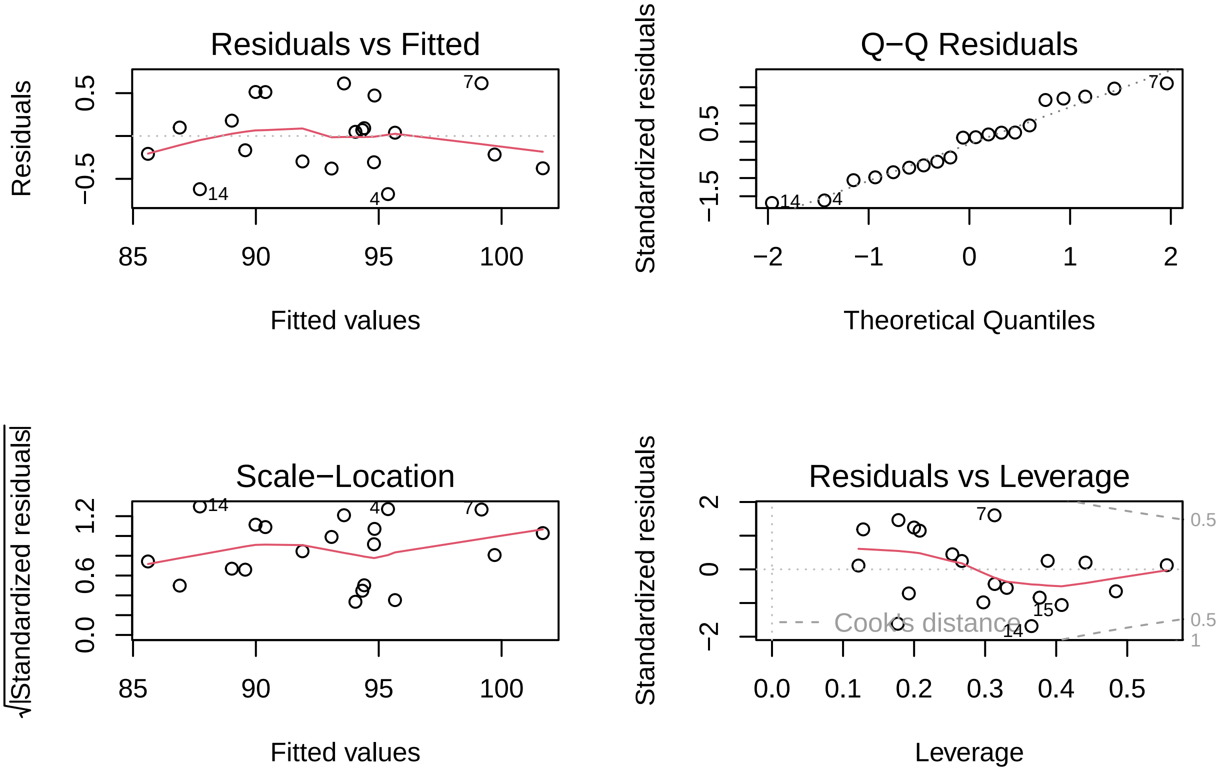
There seems a small disturbance in linearity. To solve this as practically as possible, we will reduce the number of regressors (kind of a backward elimination), apply Box-cox transformation and the necessary log() or exp() transformation.(not necessarily in that order)

Since our response variable is weight, observing the correlation plot we eliminated “Dur” variable then proceeded.

```
model1 <- lm(Weight ~ BSA + Pulse + Age + BP + Gender, data = blood)
summary(model1)
```

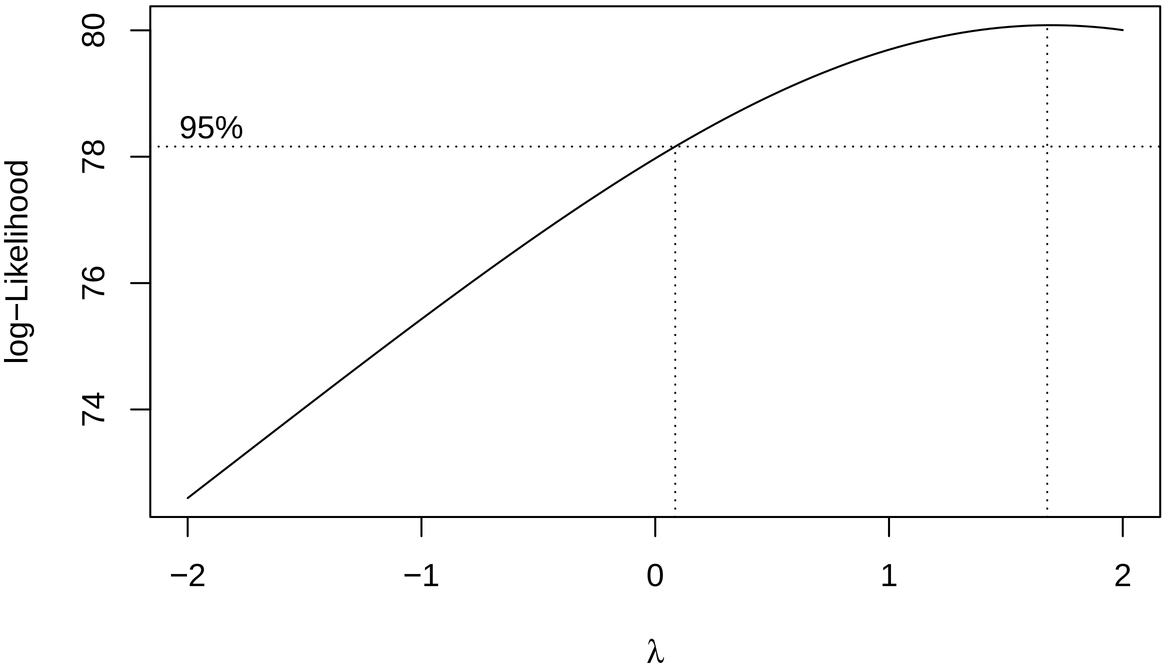
```
##
## Call:
## lm(formula = Weight ~ BSA + Pulse + Age + BP + Gender, data = blood)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6778 -0.2993  0.0434  0.2522  0.6146
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  14.74804    2.53601   5.815 4.48e-05 ***
## BSA          -3.43531    2.01366  -1.706    0.110
## Pulse         0.05212    0.04552   1.145    0.271
## Age          -0.73004    0.06592 -11.074 2.60e-08 ***
## BP            1.02762    0.06892  14.910 5.51e-10 ***
## GenderFemale -0.30326    0.24848  -1.220    0.242
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4623 on 14 degrees of freedom
## Multiple R-squared:  0.9915, Adjusted R-squared:  0.9884
## F-statistic: 325.2 on 5 and 14 DF,  p-value: 5.811e-14
```

```
par(mfrow = c(2,2))
plot(model1)
```



Applying Box-cox transformation on response variable

```
#box-cox transformation with built-in function
library(MASS)
bcx <- boxcox(model1)
```



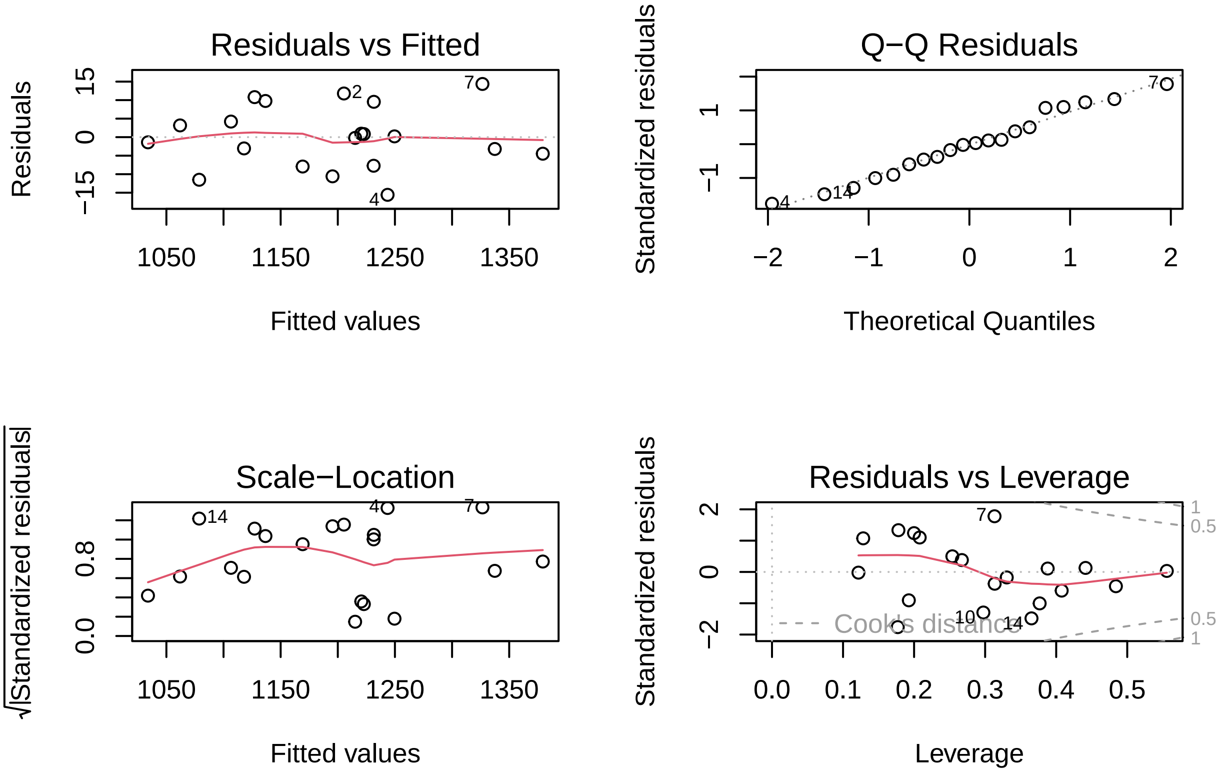
```r
# Best lambda value
lambda <- bcx$x[which.max(bcx$y)]

# Apply transformation with optimal lambda
blood$Weight <- (blood$Weight ^ lambda - 1) / lambda
```

Fitting the model again

```r
model2 <- lm(Weight ~ BSA + Pulse + Age + BP + Gender, data = blood)
summary(model2)
```

```
##
## Call:
## lm(formula = Weight ~ BSA + Pulse + Age + BP + Gender, data = blood)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.5750  -5.2840   0.0051   5.5395  14.3600
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -490.7770    53.4463  -9.183 2.66e-07 ***
## BSA           -73.9459    42.4377  -1.742    0.103
## Pulse           0.9464     0.9593   0.986    0.341
## Age           -15.6999     1.3893 -11.301 2.01e-08 ***
## BP             22.2127     1.4525  15.293 3.94e-10 ***
## GenderFemale   -6.7947     5.2366  -1.298    0.215
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.742 on 14 degrees of freedom
## Multiple R-squared:  0.9918, Adjusted R-squared:  0.9889
## F-statistic: 339.6 on 5 and 14 DF,  p-value: 4.31e-14
```

```r
par(mfrow = c(2,2))
plot(model2)
```

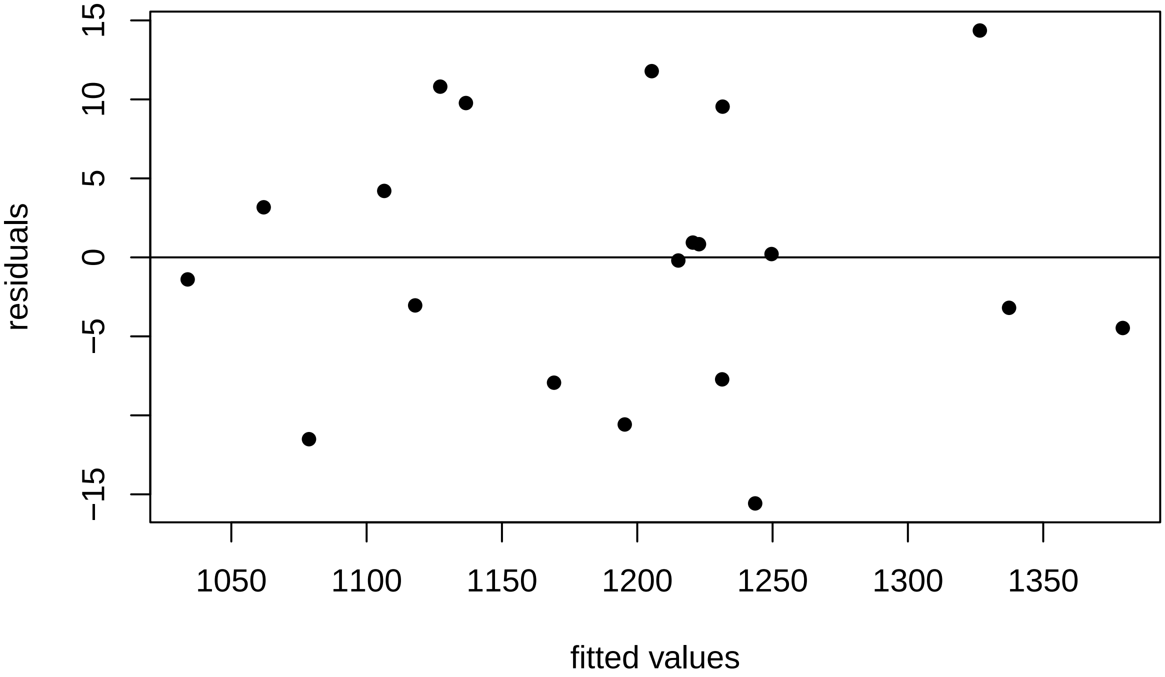


We observed the desired change on residuals vs fitted and reesiduals vs leverage plot.

Homogeneity of residuals variance: The residuals are assumed to have a constant variance (homoscedasticity)

After applying Box-cox transformation we also helped solve the problem for this assumption also

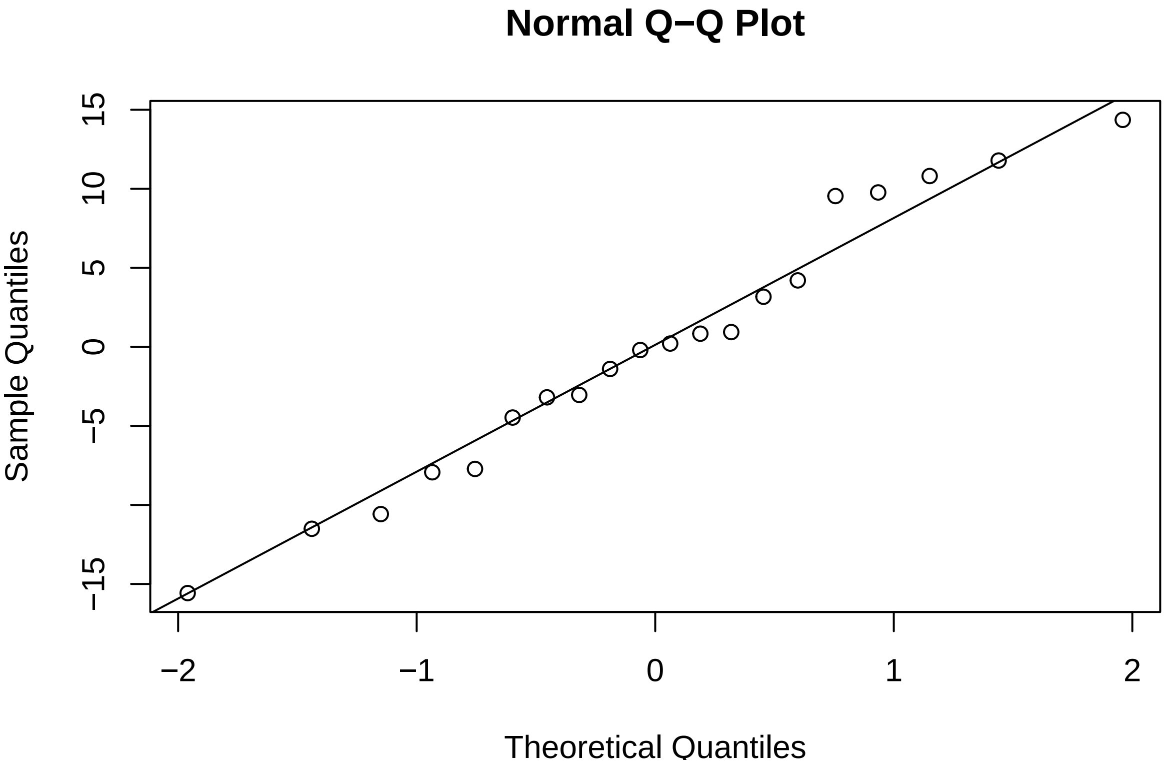
```r
yhat=predict(model2)  ### obtaining yhat by predict()
resid=model2$residuals
plot(yhat,resid,pch=16,xlab="fitted values",ylab="residuals")
abline(h=0)
```



The values bounce randomly around the 0 line, indicating homogeneous spread of variance.

Normality of residuals: The residual errors are assumed to be normally distributed.

```
resid=model2$residuals
qqnorm(resid)
qqline(resid)
```



Values follow the line, also satisfying our assumption. Let’s check again with Shapiro-Test

```r
shapiro.test(residuals(model2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model2)
## W = 0.97108, p-value = 0.7775
```

P-value is clearly larger than 0.05

Independence of residuals error terms. (Multicollinearity)

Variance Inflation Factor and Durbin - Watson test is used to observe the relationship between residuals error terms.

```r
library(car)
```

```
## Loading required package: carData
```

```r
vif(model2)
```

```
##       BSA     Pulse       Age        BP    Gender
##  6.716099  2.664661  2.416129 12.448935  1.213563
```

Here, we observe concerning vif values that are over 5 & 10. After observing the summary we dropped one or more of the regressors.

```r
summary(model2)
```

```
##
## Call:
## lm(formula = Weight ~ BSA + Pulse + Age + BP + Gender, data = blood)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.5750  -5.2840   0.0051   5.5395  14.3600
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -490.7770    53.4463  -9.183 2.66e-07 ***
## BSA          -73.9459    42.4377  -1.742    0.103
## Pulse          0.9464     0.9593   0.986    0.341
## Age          -15.6999     1.3893 -11.301 2.01e-08 ***
## BP            22.2127     1.4525  15.293 3.94e-10 ***
## GenderFemale  -6.7947     5.2366  -1.298    0.215
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.742 on 14 degrees of freedom
## Multiple R-squared:  0.9918, Adjusted R-squared:  0.9889
## F-statistic: 339.6 on 5 and 14 DF,  p-value: 4.31e-14
```

```r
model3 <- lm(Weight ~ BSA + Pulse  + Gender, data = blood)
summary(model3)
```

```
##
## Call:
## lm(formula = Weight ~ BSA + Pulse + Gender, data = blood)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -68.121 -30.882   1.969  27.349  75.558
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -323.043    181.255  -1.782   0.0937 .
## BSA           495.697     75.495   6.566  6.5e-06 ***
## Pulse           7.562      2.845   2.658   0.0172 *
## GenderFemale    2.999     20.382   0.147   0.8849
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39.39 on 16 degrees of freedom
## Multiple R-squared:  0.8472, Adjusted R-squared:  0.8186
## F-statistic: 29.57 on 3 and 16 DF,  p-value: 9.213e-07
```

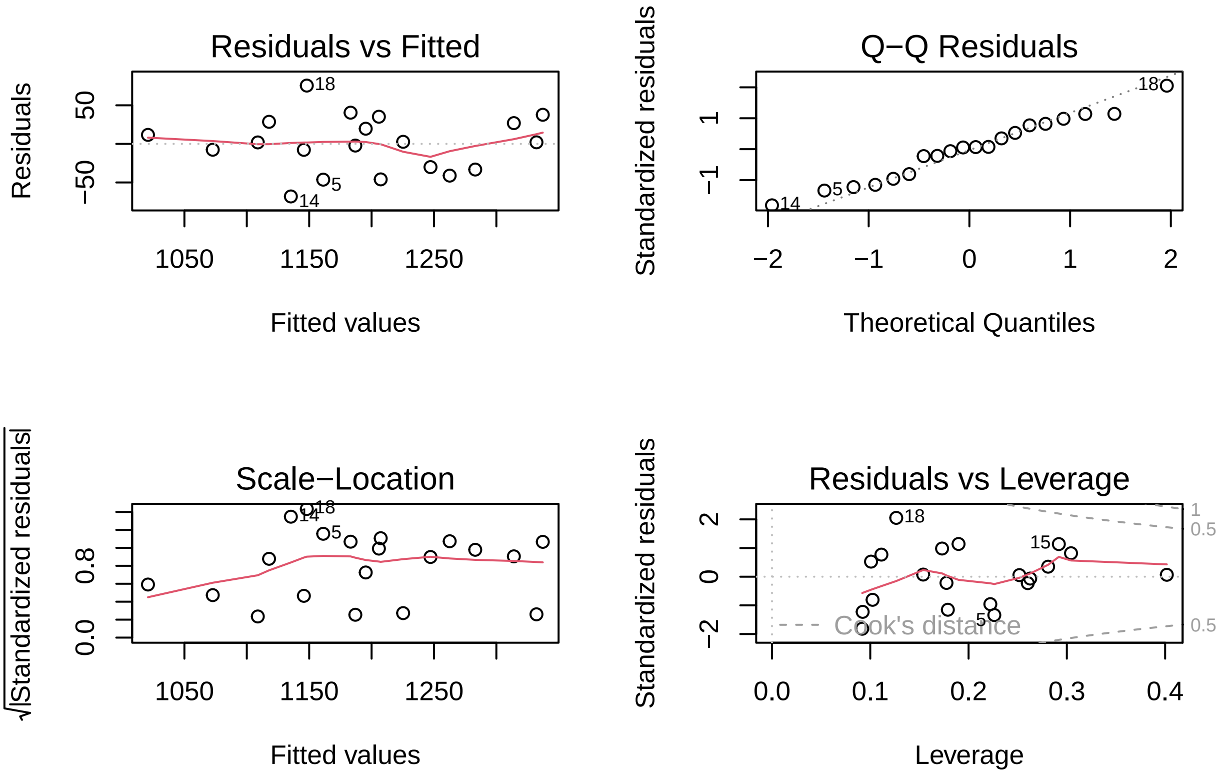
Let's check the vif again

```
vif(model3)
```

```
##      BSA    Pulse   Gender
## 1.300248 1.433710 1.124722
```

Vif values indicates the independence of regressors. But we have also seen on the summary that standard error for BSA is considerably high, which might indicate influential points. We continue to observe with cook's distance.

```
par(mfrow = c(2,2))
plot(model3)
```



```
cooks1 <- cooks.distance(model3)
cooks1
```

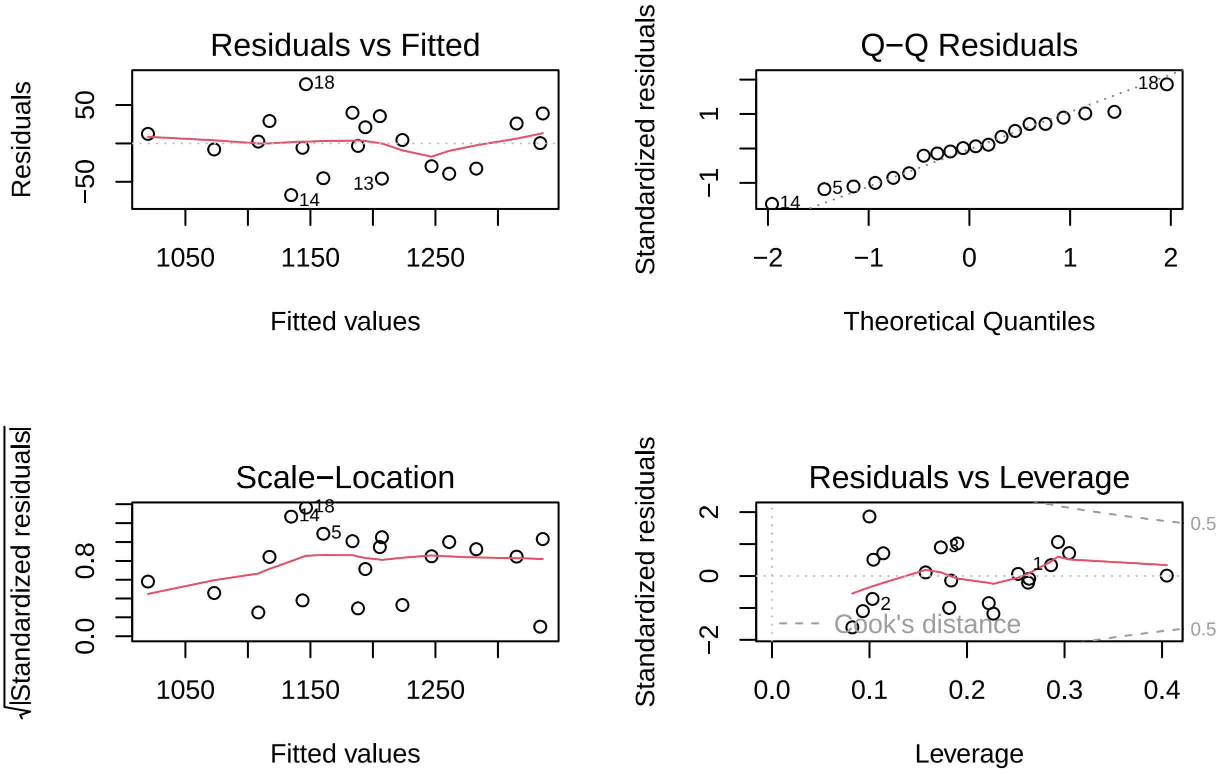
```
##            1            2            3            4            5            6
## 0.0119442644 0.0185362051 0.0505496688 0.0002513610 0.1308375801 0.0007592894
##            7            8            9           10           11           12
## 0.0734402258 0.0186018587 0.0002583194 0.0003733669 0.0723417772 0.0077650909
##           13           14           15           16           17           18
## 0.0382062353 0.0832231754 0.1334295883 0.0763138903 0.0044071804 0.1524308328
##           19           20
## 0.0025266086 0.0654326593
```

Points 5, 15 & 18 have cook's D > 1. Since they are not considerably larger we solve this by using robust linear regression which decreases the effect of influential points on model.

```
model4 <-  rlm(Weight ~ BSA + Pulse  + Gender, data = blood)
summary(model4)
```

```
##
## Call: rlm(formula = Weight ~ BSA + Pulse + Gender, data = blood)
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.426 -30.538   1.373  26.793  77.185
##
## Coefficients:
##              Value     Std. Error t value
## (Intercept)  -318.3752  190.6971    -1.6695
## BSA           502.2303   79.4278     6.3231
## Pulse           7.2943    2.9932     2.4370
## GenderFemale    4.3078   21.4442     0.2009
##
## Residual standard error: 43.69 on 16 degrees of freedom
```

```
par(mfrow = c(2,2))
plot(model4)
```



In the last model's summary we did not observe the standard error decrease, but last plots show that influential points have less effect on our model. Last plotting of the model shows the regression assumptions satisfied since: Residuals vs fitted plot does not show any pattern. QQ-plot values follows the line. Scale-Location values bounce randomly around the line. Residuals vs leverage plot shows the cook's distance of the values are at desired spread.

```r
anova(model3)
```

```
## Analysis of Variance Table
##
## Response: Weight
##           Df Sum Sq Mean Sq F value    Pr(>F)
## BSA        1 124836  124836 80.4681 1.218e-07 ***
## Pulse      1  12775   12775  8.2343   0.01112 *
## Gender     1     34      34  0.0217   0.88486
## Residuals 16  24822    1551
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

CONCLUSION

Throughout this study, we aimed to examine the effects of body surface area (BSA), pulse, age, blood pressure, and gender on weight, using multiple linear regression. We thoroughly evaluated the data and made necessary transformations to ensure the regression assumptions were met.

Linearity of the data: Initially, scatter plots indicated potential non-linear relationships between the predictors and the outcome (weight), particularly for BSA and age. To address these issues, we applied a Box-Cox transformation to the response variable. This improved the linearity of the model, as observed in the residuals vs. fitted plot, where the residuals became more randomly scattered around zero.

Homogeneity of residual variance: We tested the assumption of homoscedasticity (constant variance of residuals) by examining the residuals vs. fitted values plot. After the transformation, the spread of residuals was more consistent across fitted values, indicating that this assumption was satisfied.

Normality of residuals: The normality assumption was evaluated using the Q-Q plot and further verified with the Shapiro-Wilk test. After transforming the data, the residuals closely followed the theoretical quantile line in the Q-Q plot. The Shapiro-Wilk test supported this, returning a p-value significantly larger than 0.05, confirming that the residuals were normally distributed.

Independence of residual error terms (Multicollinearity): To ensure the independence of the predictors, we computed the Variance Inflation Factor (VIF) and performed the Durbin-Watson test. The VIF values were below the critical threshold of 5, and the Durbin-Watson test yielded a p-value greater than 0.05, suggesting no significant multicollinearity and that the residuals were independent.

Once all assumptions were confirmed, we proceeded with our regression analysis. The results indicated that age and blood pressure were statistically significant predictors of weight, demonstrating strong effects in the model. However, gender and pulse were not significant in this analysis. Overall, the model explained a substantial portion of the variability in weight, as indicated by the high adjusted R-squared value.

FUTURE DIRECTIONS

Looking forward, if we had access to unlimited resources, we would enhance this study by expanding the dataset and incorporating additional variables to provide a more comprehensive analysis of the factors affecting weight.

Firstly, we would consider including variables such as genetic factors, physical activity levels, detailed dietary habits, sleep patterns, and stress levels, which likely have a significant impact on weight. This would allow us to build a more holistic model, addressing both biological and lifestyle-related factors.

Moreover, we would collect data from a significantly larger and more diverse population, including individuals from various demographic backgrounds, to ensure the generalizability of our findings. Ideally, this data would be collected longitudinally, allowing us to capture changes over time and better understand the dynamic nature of weight fluctuations.

We would also consider employing advanced data collection techniques, such as wearable health monitors that track real-time data on activity levels, caloric intake, and physiological responses. This could enable the development of predictive models that account for real-time variations in weight-related behaviors, leading to more personalized interventions.

Additionally, the use of machine learning and non-linear modeling techniques could further improve the accuracy and predictive power of our models. By capturing more complex interactions between variables, these models could uncover hidden patterns in the data, allowing for a deeper understanding of the factors that influence weight.

Finally, integrating our findings into practical health applications, such as personalized health monitoring systems, would allow for early detection of weight-related health risks and offer tailored recommendations for individuals aiming to manage their weight effectively.

By pursuing these directions, future studies could significantly advance the understanding of weight management, providing actionable insights that improve health outcomes at both the individual and population levels.