STAT-670-Assignment 3

Prashanth Thirukkurungudi Sekar (UID: 2000376633)

January 30, 2019

Worked with Pravin Sundar and Barathwaaj Parthasarathy

Introduction

It is know that High systolic blood pressure is a strong predictor of heart attacks and strokes. In this analysis, we attempt to verify if a person's age, weight and height changes with average systolic blood pressure. Eventually, we report if the variation between blood pressure and the above mentioned factors vary between male and female to conclude more insights.
NHANES data is used for this experiment and with the help of some plots suitable conclusions are drawn.

Loading the libraries:

```
library(readr)
library(tidyverse)
library(ggplot2)
library(lattice)
library(NHANES)
library(broom)
library(gridExtra)
library(arm)
```

Loading the data with the required columns:

```
req_cols = c("Age","Weight","Height","Gender","BPSysAve")
nhanes = NHANES[,req_cols]
nhanes = nhanes[complete.cases(nhanes),]
```

Splitting the data for Male and Female:

```
male =  nhanes[nhanes$Gender=="male",]
female =  nhanes[nhanes$Gender=="female",]
```

Summarising Male and Female data:

```
summary(male)
```

```
##       Age             Weight           Height         Gender
##  Min.   : 8.00   Min.   : 17.10   Min.   :112.5   female:   0
##  1st Qu.:24.00   1st Qu.: 70.30   1st Qu.:168.9   male  :4217
##  Median :40.00   Median : 82.80   Median :174.8
##  Mean   :40.18   Mean   : 83.81   Mean   :173.1
##  3rd Qu.:55.00   3rd Qu.: 97.30   3rd Qu.:179.9
##  Max.   :80.00   Max.   :223.00   Max.   :200.4
##     BPSysAve
##  Min.   : 76
##  1st Qu.:110
##  Median :118
##  Mean   :120
##  3rd Qu.:128
##  Max.   :221
```

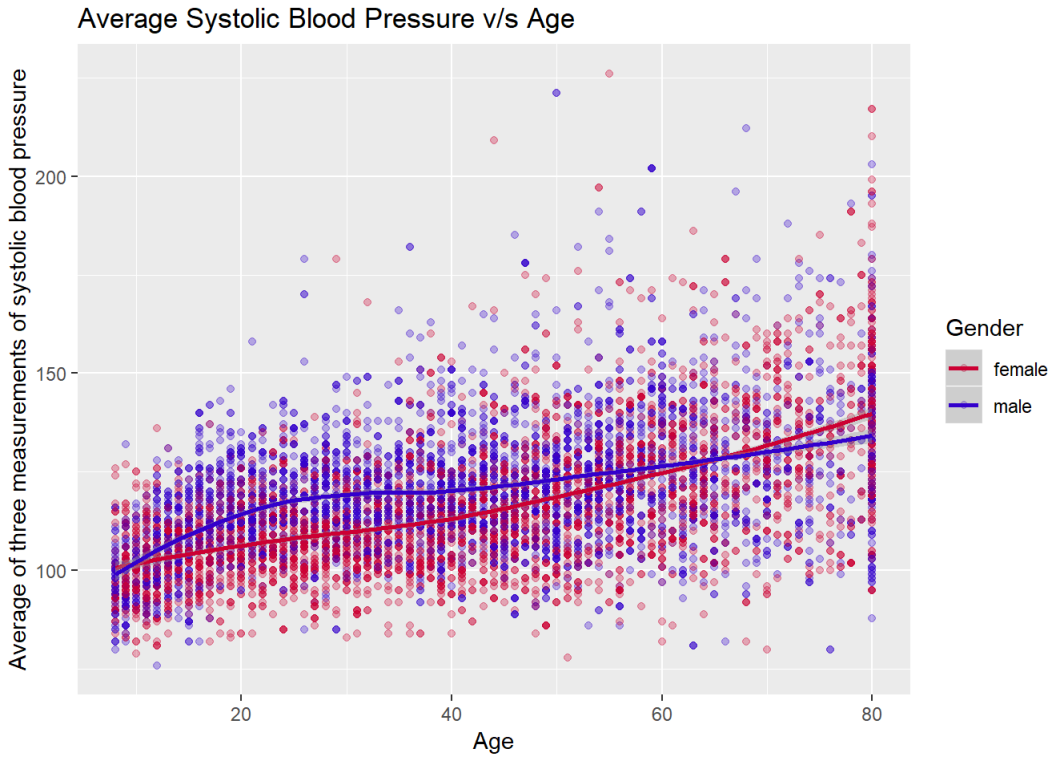
```
summary(female)
```

```
##       Age             Weight           Height         Gender     
##  Min.   : 8.00   Min.   : 21.70   Min.   :119.1   female:4270  
##  1st Qu.:24.00   1st Qu.: 58.00   1st Qu.:156.2   male   :   0  
##  Median :41.00   Median : 68.50   Median :161.3                
##  Mean   :41.68   Mean   : 71.99   Mean   :160.8                
##  3rd Qu.:58.00   3rd Qu.: 83.00   3rd Qu.:166.6                
##  Max.   :80.00   Max.   :230.70   Max.   :184.5                
##     BPSysAve    
##  Min.   : 78.0  
##  1st Qu.:104.0  
##  Median :113.0  
##  Mean   :116.3  
##  3rd Qu.:125.0  
##  Max.   :226.0
```

Age:

Trend:

```
cb_palette = c("#CC0033", "#3300CC")
ggplot(nhanes, aes(x = Age, y = BPSysAve, color = Gender)) + geom_point(alpha = 0.3)+ scale_color_manual(val
ues = cb_palette) + geom_smooth(method=loess) + labs(x="Age",y="Average of three measurements of systolic bl
ood pressure") + ggtitle("Average Systolic Blood Pressure v/s Age")
```



Building model:

```
age_male.lm = loess(BPSysAve~Age, data = male,method.args =list(degree=1),span=.8)
age_male.lm.df = augment(age_male.lm)
hat <- predict(age_male.lm)
r_sq_loess_male <- cor(male$BPSysAve, hat)^2
print("The R Squared value for the loess model for male:")
```

```
## [1] "The R Squared value for the loess model for male:"
```

```
round(r_sq_loess_male,2)
```

```
## [1] 0.23
```

```
age_female.lm = loess(BPSysAve~Age, data = female,method.args =list(degree=1))
age_female.lm.df = augment(age_female.lm)
hat <- predict(age_female.lm)
r_sq_loess_female <- cor(female$BPSysAve, hat)^2
print("The R Squared value for the loess model for female:")
```

```
## [1] "The R Squared value for the loess model for female:"
```

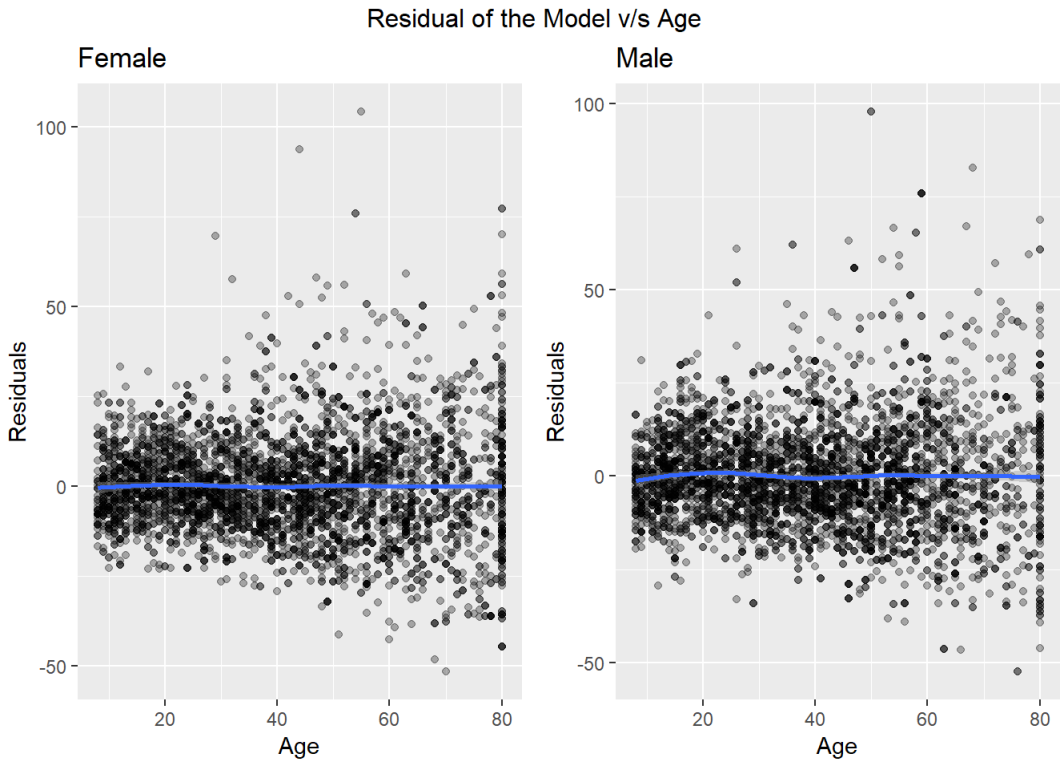
```
round(r_sq_loess_female,2)
```

```
## [1] 0.36
```

Plot residuals to check the model:

```
age_male_resid = ggplot(age_male.lm.df, aes(x = Age, y = .resid)) + geom_point(alpha=0.3)+ geom_smooth(metho
d=loess) + ggtitle("Male") + labs(y="Residuals")
age_female_resid = ggplot(age_female.lm.df, aes(x = Age, y = .resid)) + geom_point(alpha=0.3)  + geom_smooth
(method=loess) + ggtitle("Female") + labs(y="Residuals")
```

```
grid.arrange(age_female_resid, age_male_resid, top = "Residual of the Model v/s Age",ncol = 2)
```



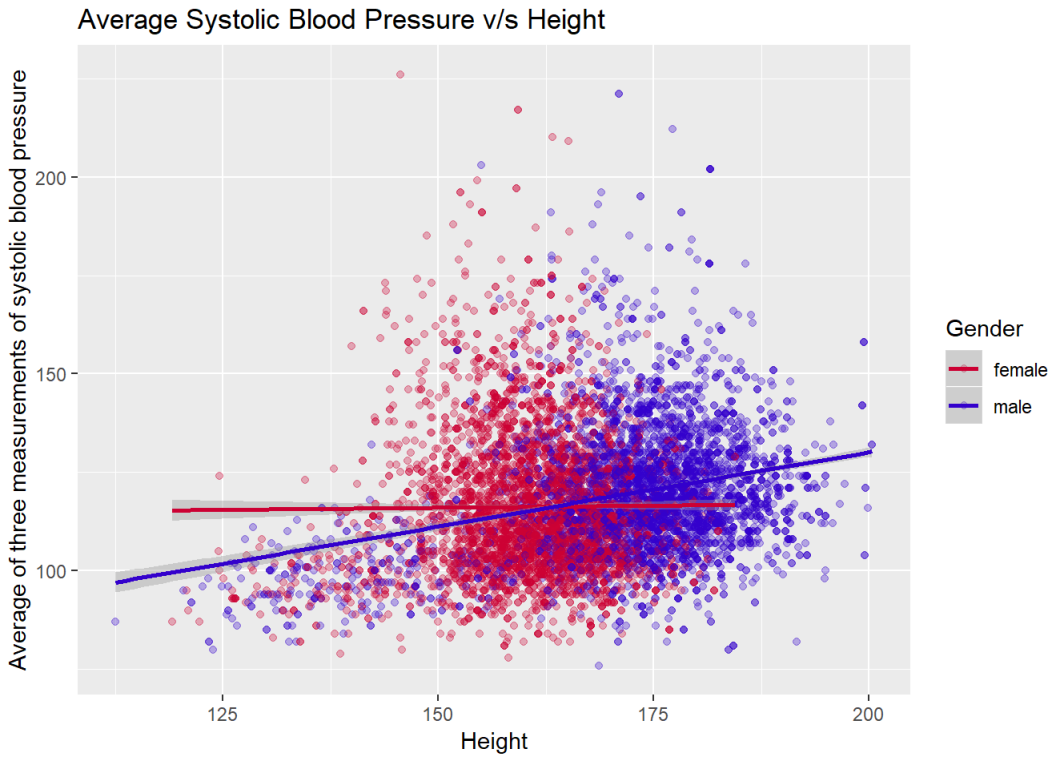
Description :

From the trend plot it can be observed that the average systolic blood pressure increases with age and this trend seems to hold good for both male and female. Also, it can be observed that there is not any evident/ significant difference in the distribution of age v/s systolic blood pressure with respect to gender. Also, since 80 and older age groups have been coded as 80, it can be observed that a variety of values are present for the age group 80.Given the slight upward bump in the distirubtion for male, a “loess” model (non-parametric smoother) was selected to fit the data since it captures more variance present in the data. Also, since the data already looks fairly linear (for both male and female), no transformation on the age variable was warranted for this model. Looking at the residual plot, it is evident that the error terms have a constant variance and no evident pattern is found. Hence we can conclude that there is no presence of heteroskedasticity in this case. Finally, we can conclude that for male and female the trend with respect to age and average systolic blood pressure is positive and it can be seen that the observations are fairly close to the trend however, with the models explaining only 23% of the variance for male and 36% of the variance for females suggesting that the variations/trend is not strictly linear in nature.

Height:

Trend:

```
cb_palette = c("#CC0033", "#3300CC")
ggplot(nhanes, aes(x = Height, y = BPSysAve, color = Gender)) + geom_point(alpha = 0.3)+ scale_color_manual(
values = cb_palette) + geom_smooth(method=lm)+ labs(x="Height",y="Average of three measurements of systolic
blood pressure") + ggtitle("Average Systolic Blood Pressure v/s Height")
```



Buidling model :

```
height_male.lm = lm((BPSysAve)~(Height), data = male)
height_male.lm.df = augment(height_male.lm)
display((height_male.lm))
```

```
## lm(formula = (BPSysAve) ~ (Height), data = male)
##             coef.est coef.se
## (Intercept) 54.67     3.67
## Height       0.38     0.02
## ---
## n = 4217, k = 2
## residual sd = 15.82, R-Squared = 0.07
```

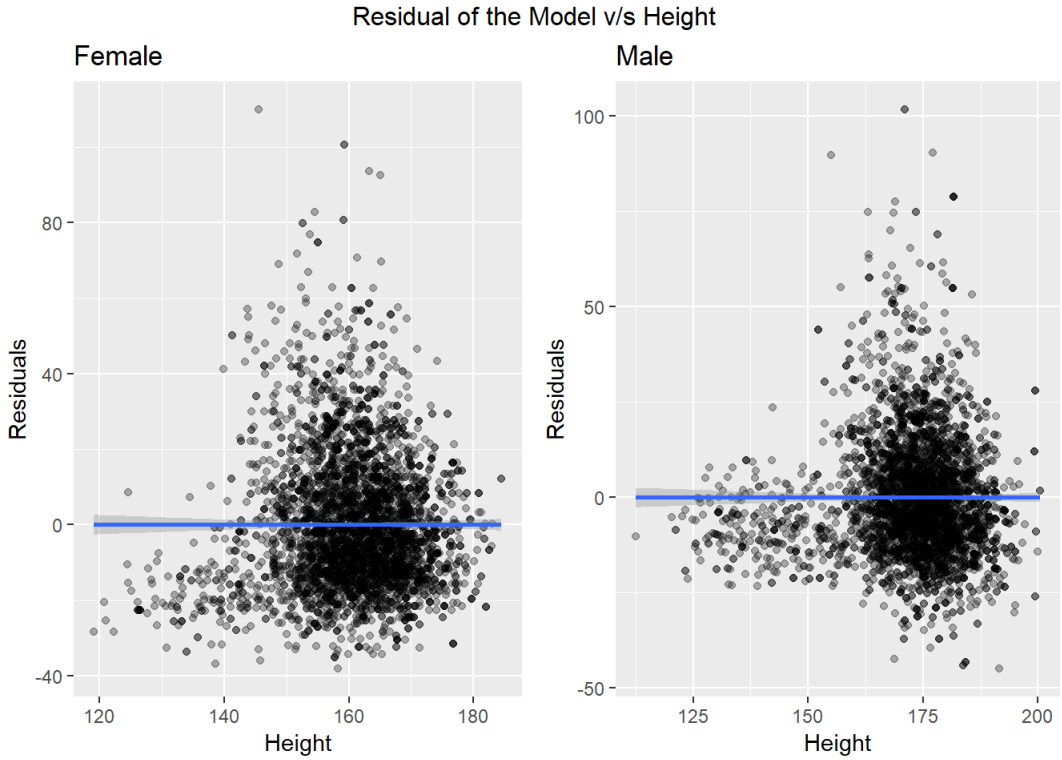
```
height_female.lm = lm((BPSysAve)~(Height), data = female)
height_female.lm.df = augment(height_female.lm)
display((height_female.lm))
```

```
## lm(formula = (BPSysAve) ~ (Height), data = female)
##             coef.est coef.se
## (Intercept) 112.87     5.02
## Height        0.02     0.03
## ---
## n = 4270, k = 2
## residual sd = 17.77, R-Squared = 0.00
```

Residual Plot:

```
height_male_resid = ggplot(height_male.lm.df, aes(x =Height, y = .resid)) + geom_point(alpha=0.3)+ geom_smoo
th(method="lm") + ggtitle("Male") + labs(y="Residuals")
height_female_resid = ggplot(height_female.lm.df, aes(x = Height, y = .resid)) + geom_point(alpha=0.3)  + ge
om_smooth(method="lm") + ggtitle("Female") +labs(y="Residuals")
```

```
grid.arrange(height_female_resid, height_male_resid, top = "Residual of the Model v/s Height",ncol = 2)
```



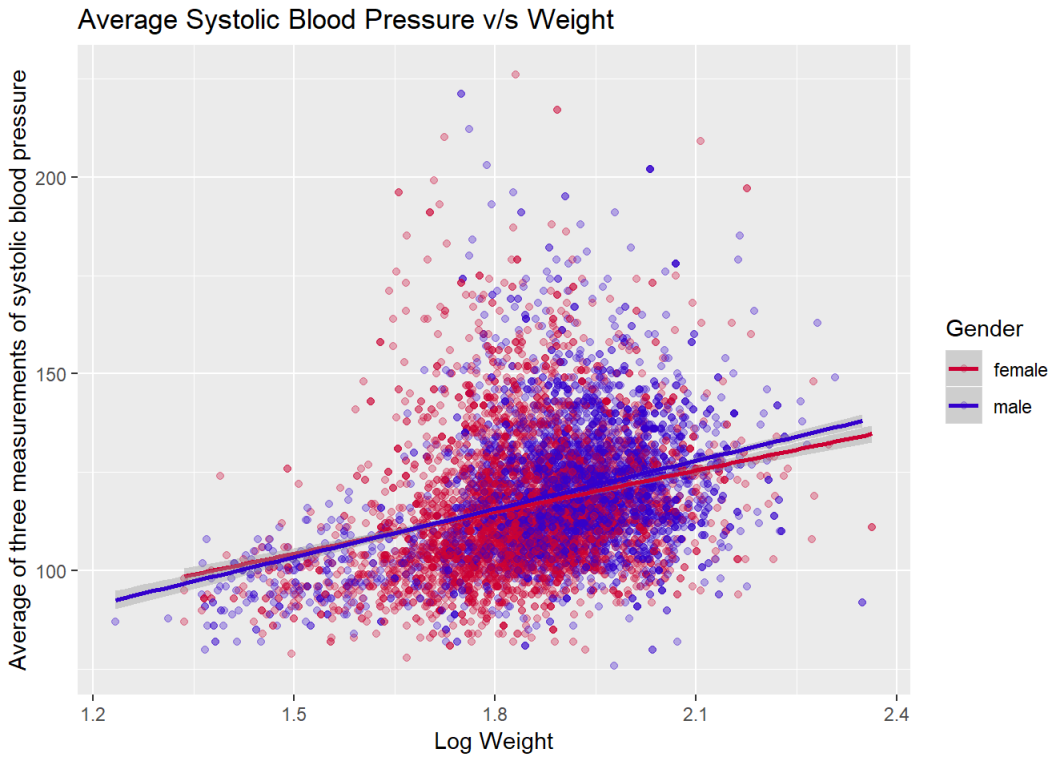
Description :

For male, the average systolic blood pressure tends to increases with height while for female, it almost tends to remain cosntant throughout and this can be confirmed through the very small coefficient value obtained through the model (0.02) female while the model for male having a positive coefficient of 0.38. Also, for female most of the points remains highly concentrated within the range of 150 to 175. Looking at the trend plot, it is evident that male and female have 2 different pattern or trend for systolic blood pressure with height. Also, the mean heights differ with the mean female height being 161 cm and the mean male height being 173 cm. Given the fairly linear nature present in the trend plot, a standard linear model (lm) was chosen. Also, since the data already looks fairly linear (for both male and female), no transformation on the height variable was warranted for this model. Looking at the residual plot, it is evident that the error terms have a constant variance and no evident pattern is found. Hence we can conclude that there is no presence of heteroskedasticity in this case. Finally, we can conclude that for male there is a slight increasing trend of average systolic blood pressure with height while for female it tends to remain a constant and it can be seen that the observations are not very close to the trend given the large number of outlier observations and this is supported by the models explaining only 7% of the variance for male and 0% of the variance for females (no trend - constant trend).

Weight:

Trend:

```
cb_palette = c("#CC0033", "#3300CC")
ggplot(nhanes, aes(x = log10(Weight), y = BPSysAve, color = Gender)) + geom_point(alpha = 0.3)+ scale_color_
manual(values = cb_palette) + geom_smooth(method=lm) + labs(x="Log Weight",y="Average of three measurements
of systolic blood pressure") + ggtitle("Average Systolic Blood Pressure v/s Weight")
```



Buidling model :

```
male$Weight = log(male$Weight)
female$Weight = log(female$Weight)

weight_male.lm = lm((BPSysAve)~(Weight), data = male)
weight_male.lm.df = augment(weight_male.lm)
display(weight_male.lm)
```

```
## lm(formula = (BPSysAve) ~ (Weight), data = male)
##             coef.est coef.se
## (Intercept) 42.30     3.34
## Weight      17.72     0.76
## ---
## n = 4217, k = 2
## residual sd = 15.45, R-Squared = 0.11
```

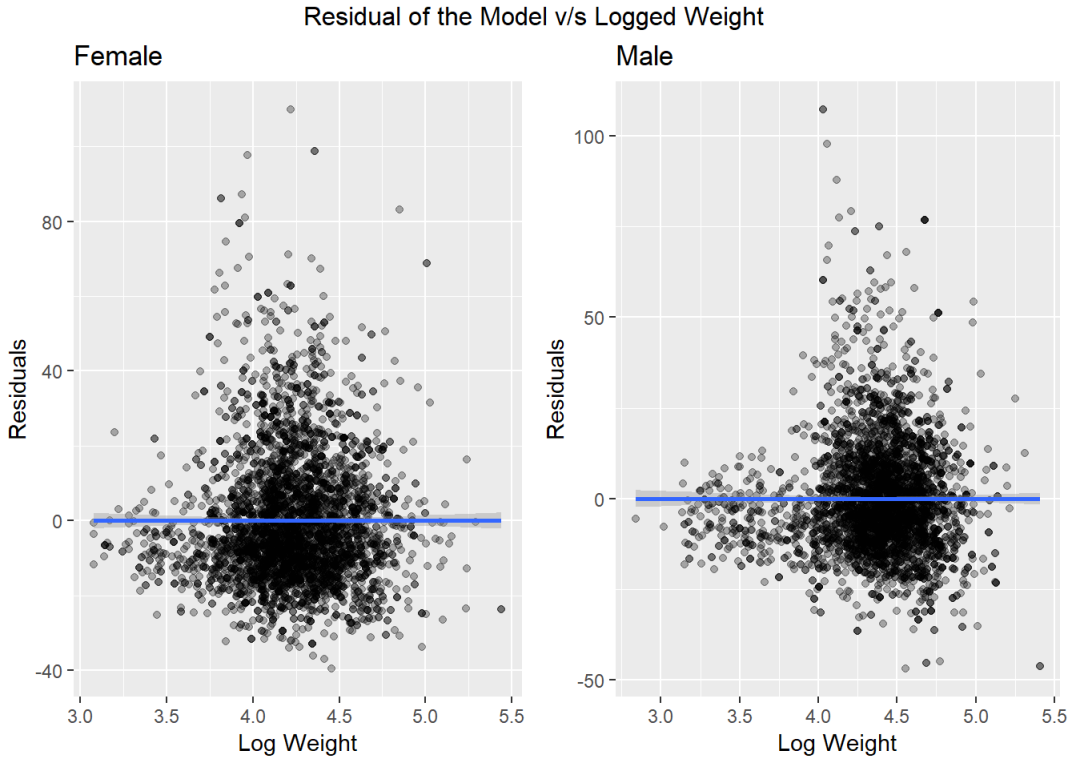
```
weight_female.lm = lm((BPSysAve)~(Weight), data = female)
weight_female.lm.df = augment(weight_female.lm)
display(weight_female.lm)
```

```
## lm(formula = (BPSysAve) ~ (Weight), data = female)
##             coef.est coef.se
## (Intercept) 51.61     3.73
## Weight      15.28     0.88
## ---
## n = 4270, k = 2
## residual sd = 17.17, R-Squared = 0.07
```

Residual Plot:

```
weight_male_resid = ggplot(weight_male.lm.df, aes(x = Weight, y = .resid)) + labs(x="Log Weight",y="Residuals") + geom_point(alpha=0.3)+ geom_smooth(method=lm) + ggtitle("Male")
weight_female_resid = ggplot(weight_female.lm.df, aes(x = Weight, y = .resid)) + labs(x="Log Weight",y="Residuals") + geom_point(alpha=0.3)  + geom_smooth(method=lm) + ggtitle("Female")
```

```
grid.arrange(weight_female_resid, weight_male_resid, top = "Residual of the Model v/s Logged Weight",ncol = 2)
```



Description :

For male and female, the average systolic blood pressure tends to increases with weight and this can be confirmed through the positive coefficient values from the respective models. With almost similar mean weights for both male and female, looking at the trend plot, it is evident that male and female almost have the same pattern. Given the fairly linear nature present in the trend plot, a standard linear model (lm) was chosen. In order to fit a linear model, logged transformation of the weight variable seemed appropriate. Looking at the residual plot, it is evident that the error terms have a constant variance and no evident pattern is found. Hence we can conclude that there is no presence of heteroskedasticity in this case. Finally, we can conclude that for both male and female there is an increasing trend of average systolic blood pressure with respect to weight and it can be seen that the observations are not very close to the trend given the large number of outlier observations and this is supported by the models explaining only 11% of the variance for male and 7% of the variance for females.

Conclusion :

After the above analysis, we can conclude that for males, the average systolic blood pressure has a strictly positive variation/ trend with age, height and weight. While for females, the average systolic blood pressure varies with the age and weight and there is no change or constant trend with increase in height.