Assignment3

Apurva Gupta,Shailendra Patil,Surbhi Paithankar

1/27/2018

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.2
```

```
library(NHANES)
Male_data = subset(NHANES, Gender == 'male')
Female_data = subset(NHANES, Gender == 'female')
```

As the data is huge that is 10000 , let us not use loess method for our analysis and do our analysis using linear, polynomial and general additive models

In the given data, number of people above the age 18 is more as compared to number of people less than 18. The reason we mention this because there is generally a relation between age and height , if not the weight. We know that after certain age , the height of the person reaches maximum and thereby doesnt change. So in the data we might not find exact trends as needed because of the majority of data following in some specific range as the data consists mostly adults.
And the weight is variable that is fluctuates and it might not have a strong relation with Age as compared to Height.

And the given problem doesn't require any inferences or predicting and hence we will avoid checking of heteroskedasticity/homoskedasticity

Keeping all these general ideas in mind, let us start th analysis

Age Vs BPSysAve

Checking the trend

```
Age_data=ggplot(NHANES, aes(x = Age, y = BPSysAve)) + geom_point(alpha=0.1,size=1)

Age_data+geom_smooth(method='lm')+
geom_smooth(method='lm',formula = y ~ x + I(x^2),col="red")+geom_smooth(col="yellow")+
facet_wrap(~Gender,ncol=1)
```

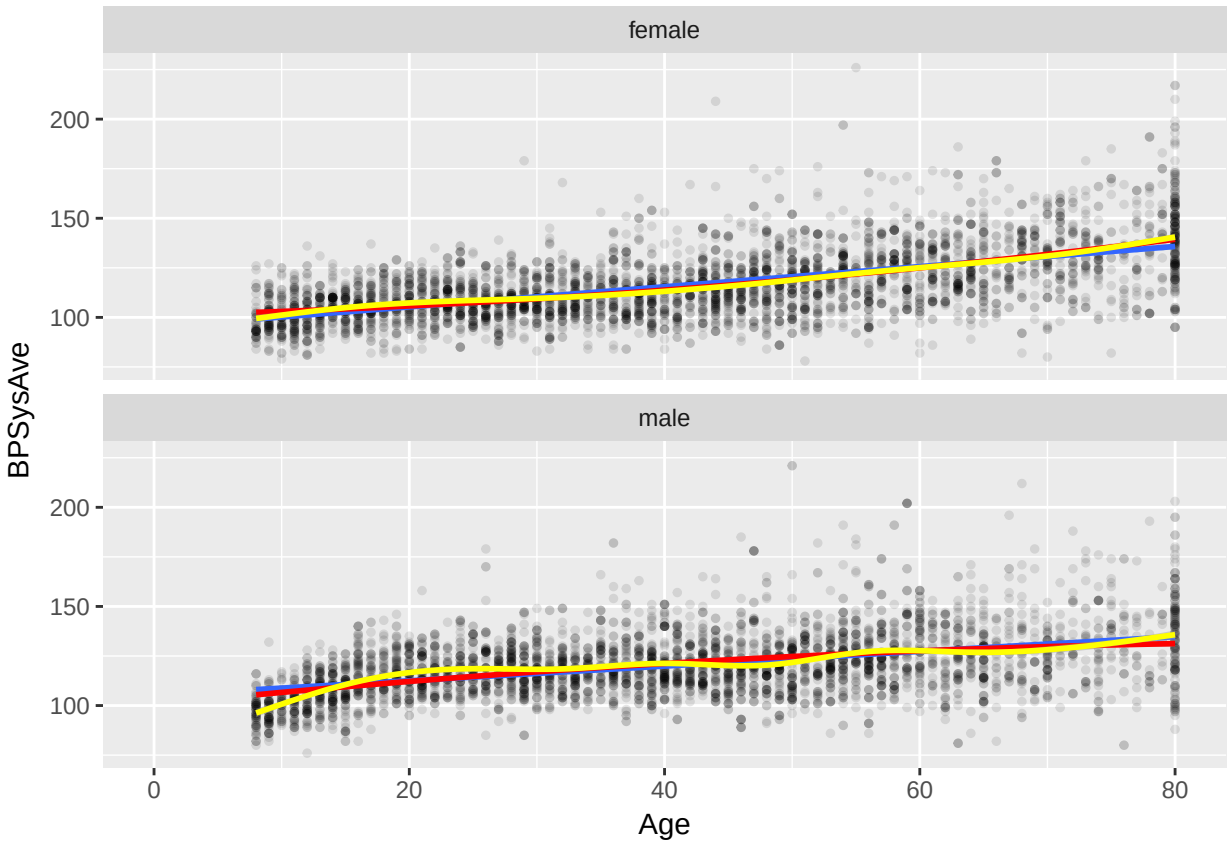
```
## Warning: Removed 1449 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 1449 rows containing non-finite values (stat_smooth).
```

```
## `geom_smooth()` using method = 'gam'
```

```
## Warning: Removed 1449 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 1449 rows containing missing values (geom_point).
```



In the above plots, blue line is the linear fit, red line is the quadratic fit and yellow line is the General Additive Model (GAM) fit.

For Women, all the 3 fits are almost similar, let us consider linear fit as its easy for interpretation.

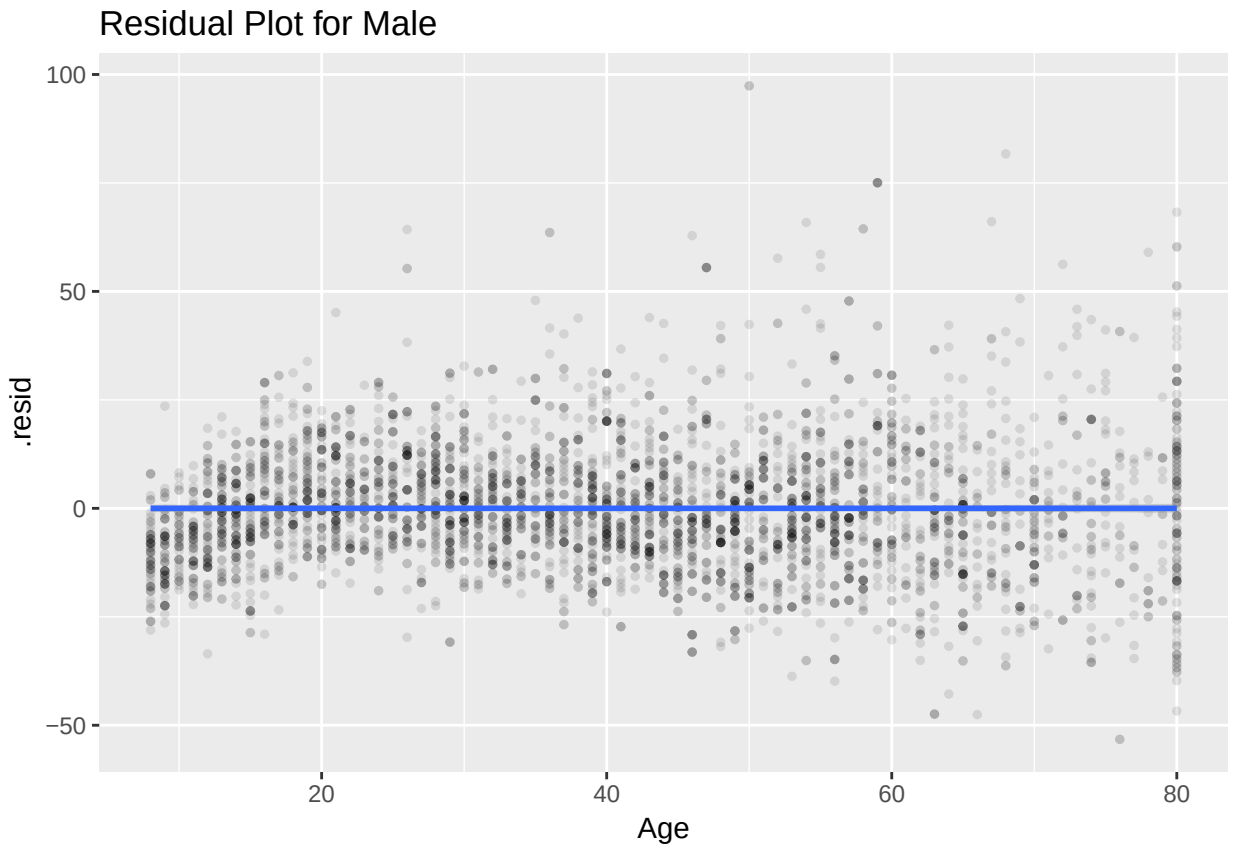
For Men, the GAM fit suggest something which is not normal and has a trend that is varying for few intervals. And this GAM might overfit the data. Hence lets shift the focus to linear and quadratic fit. Both linear and quadratic fits look similar except near the corner points where quadratic fit has a dip and this may be because Age above 80 are considered 80 and the data might be inconistent at the edge because the no of people who are not adults are less in the given data and hence we have less values for analyses. So looking at the trend we can consider either a linear or quadratic trend. Let us consider the linear trend as it is easy to interepret.

Checking the spread

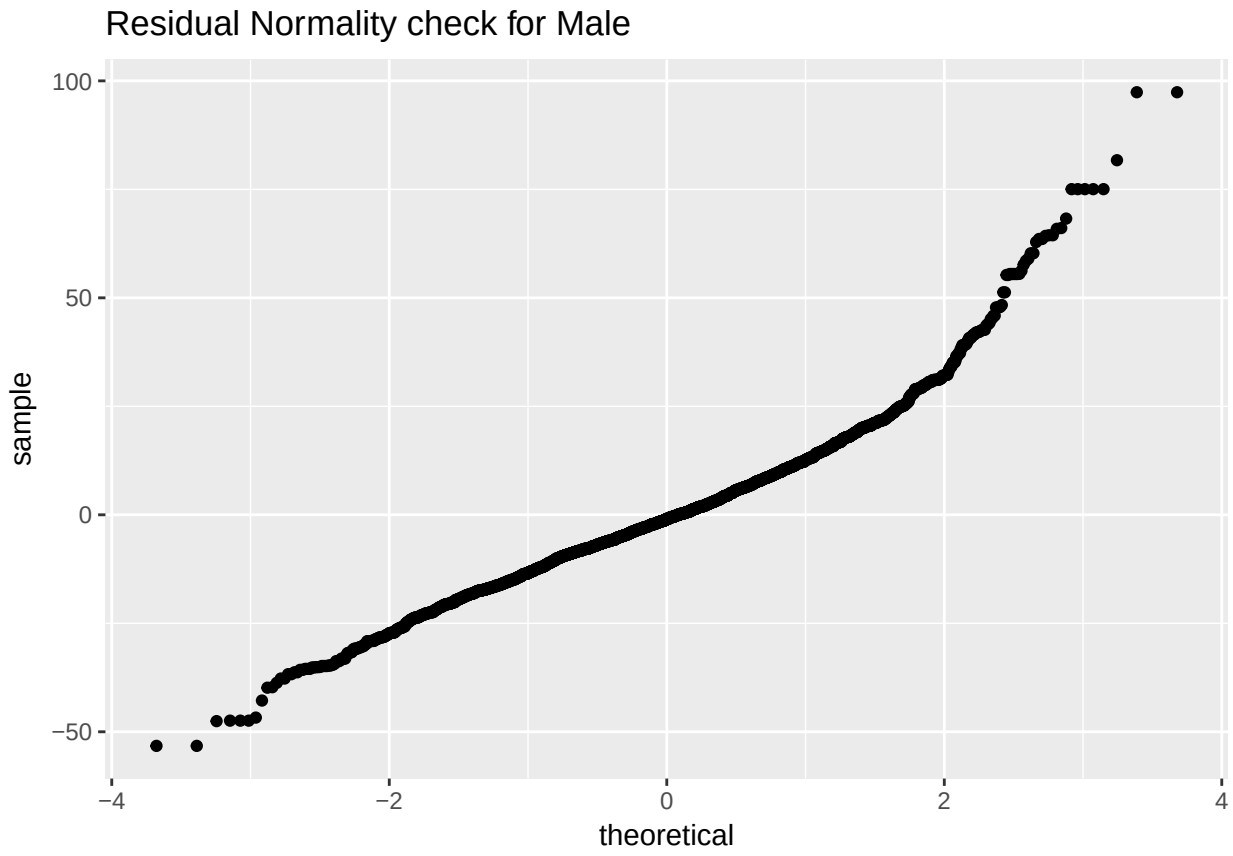
```
library(broom)
```

```
## Warning: package 'broom' was built under R version 3.3.2
```

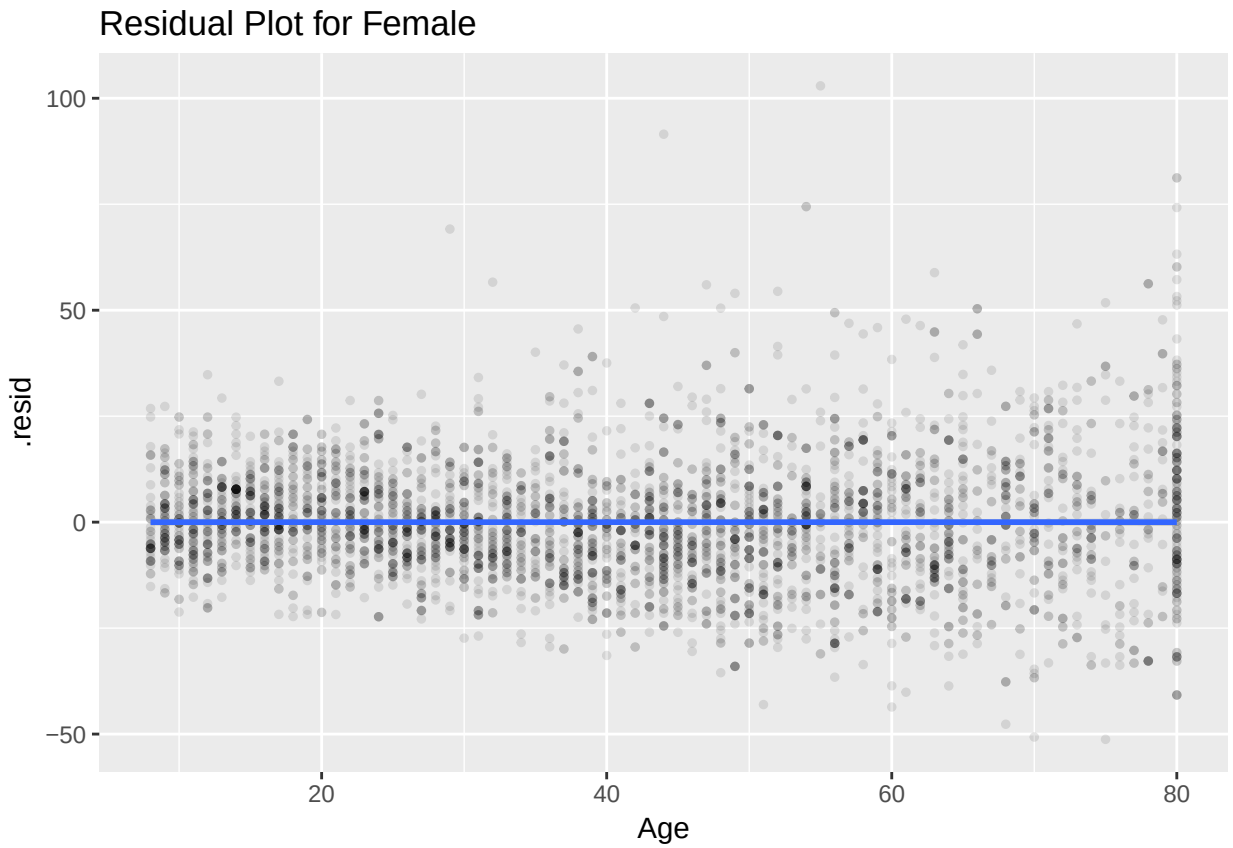
```
Male_age.lo = lm(BPSysAve ~ Age,data = Male_data)
Male_age.lo.df=augment(Male_age.lo)
ggplot(Male_age.lo.df, aes(x = Age, y = .resid)) + geom_point(alpha=0.1,size=1) + geom_smooth(method='lm
```



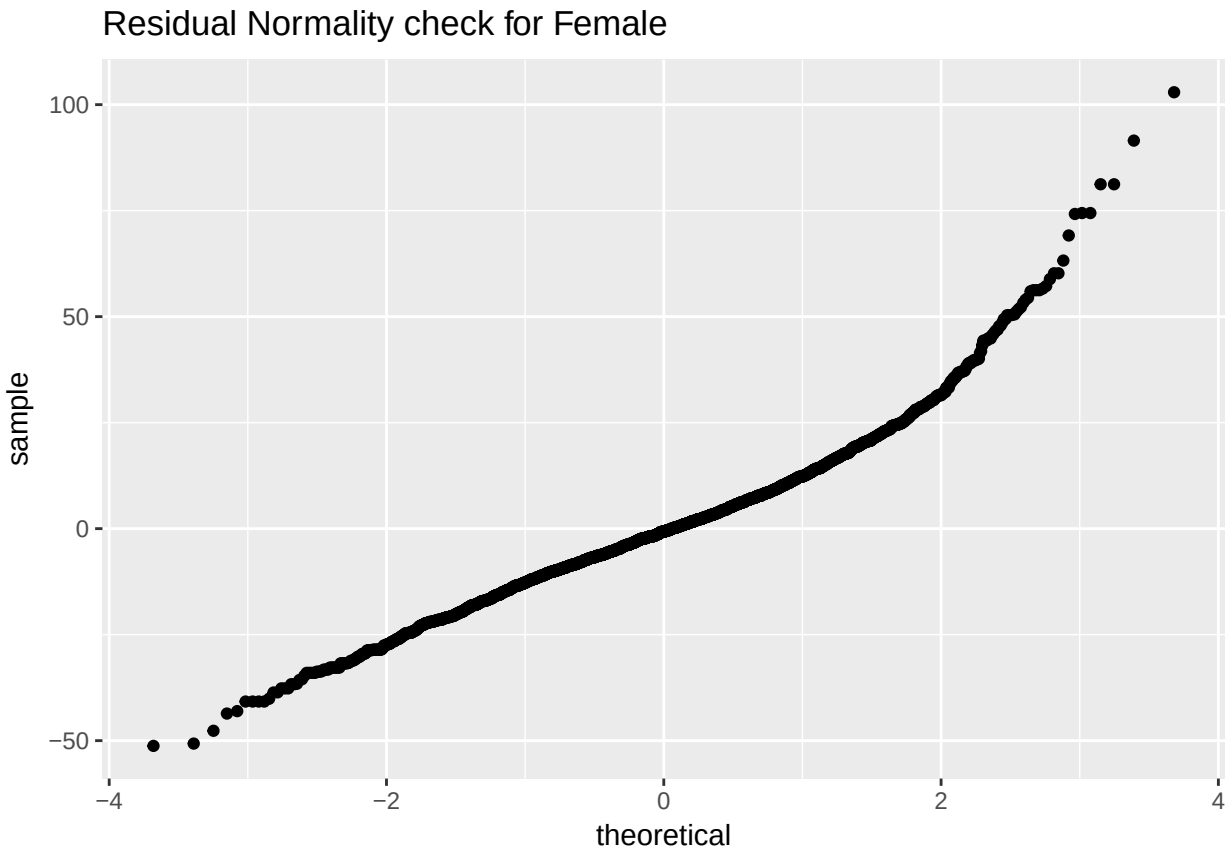

```
ggplot(Male_age.lo.df, aes(sample = .resid)) + stat_qq()+ggtitle("Residual Normality check for Male")
```




```
Female_age.lo = lm(BPSysAve ~ Age,data = Female_data)
Female_age.lo.df=augment(Female_age.lo)
ggplot(Female_age.lo.df, aes(x = Age, y = .resid)) + geom_point(alpha=0.1,size=1) + geom_smooth(method=
```




```
ggplot(Female_age.lo.df, aes(sample = .resid)) + stat_qq()+ggtitle("Residual Normality check for Female
```



Residuals plots for both male and female are almost similar and it doesn't look like there is any trend in the residuals. When we check the normality , the residuals for both male and female are normal except for the edges.

Checking the mean of residuals

```
mean(Male_age.lo.df$.resid)
```

```
## [1] 6.723514e-16
```

```
mean(Female_age.lo.df$.resid)
```

```
## [1] 1.609629e-16
```

We can see that mean of the residuals is almost 0.

BPSysAve vs Weight

Checking the trend

```
library(MASS)
Weight_data=ggplot(NHANES, aes(x = Weight, y = BPSysAve)) + geom_point(alpha=0.1,size=1)

Weight_data+geom_smooth(method='lm')+
```

```
geom_smooth(method='lm',formula = y ~ x + I(x^2),col="red")+geom_smooth(col="yellow")+
geom_smooth(method = "rlm", se = FALSE,
col = "orange",method.args = list(psi = psi.bisquare))+facet_wrap(~Gender,ncol=1)
```

```
## Warning: Removed 1507 rows containing non-finite values (stat_smooth).
```

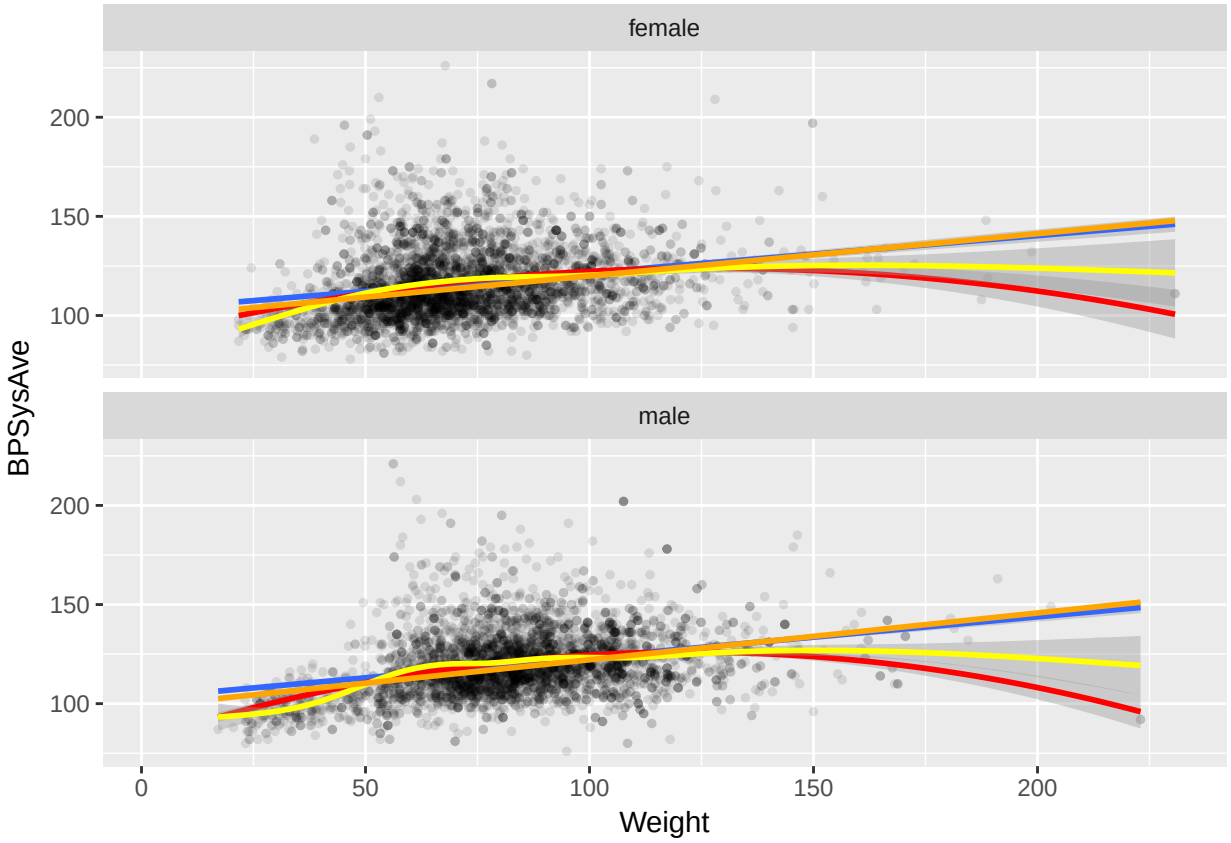
```
## Warning: Removed 1507 rows containing non-finite values (stat_smooth).
```

```
## `geom_smooth()` using method = 'gam'
```

```
## Warning: Removed 1507 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 1507 rows containing non-finite values (stat_smooth).
```

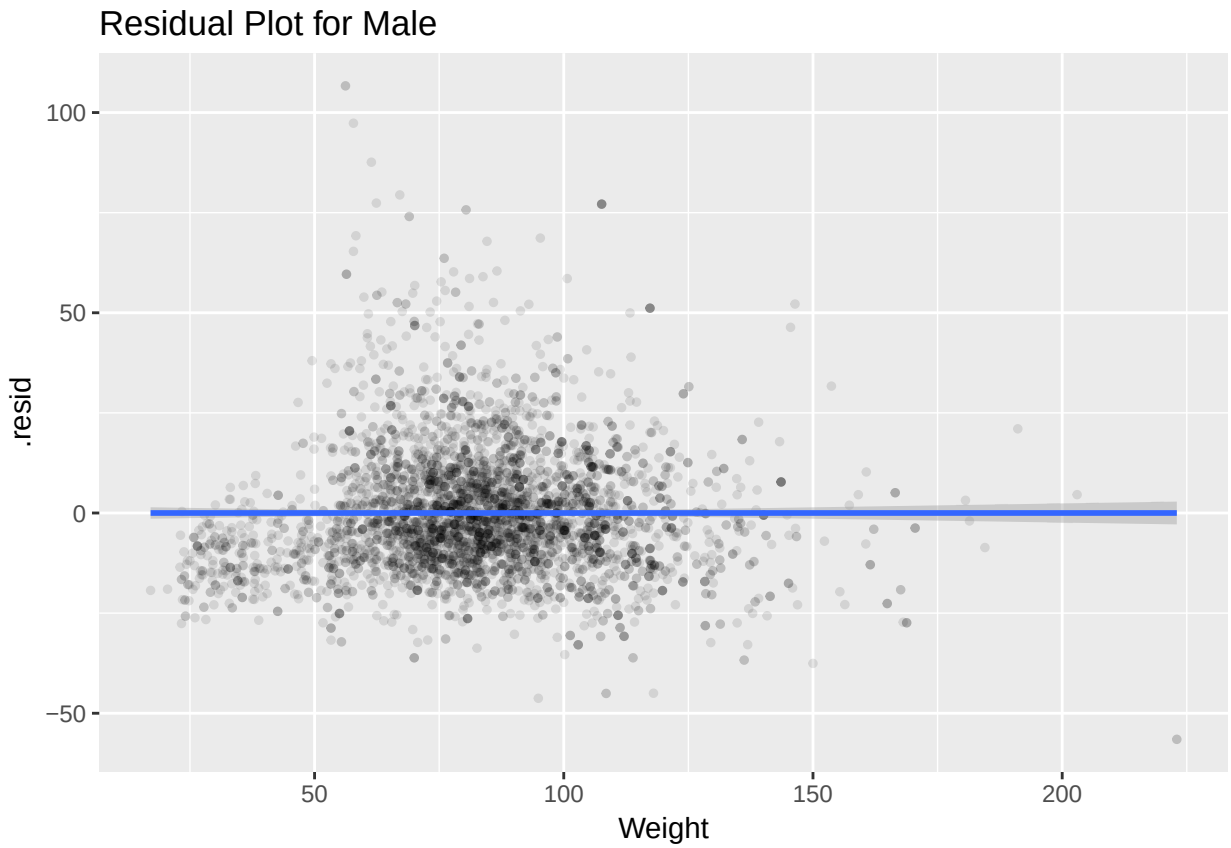
```
## Warning: Removed 1507 rows containing missing values (geom_point).
```



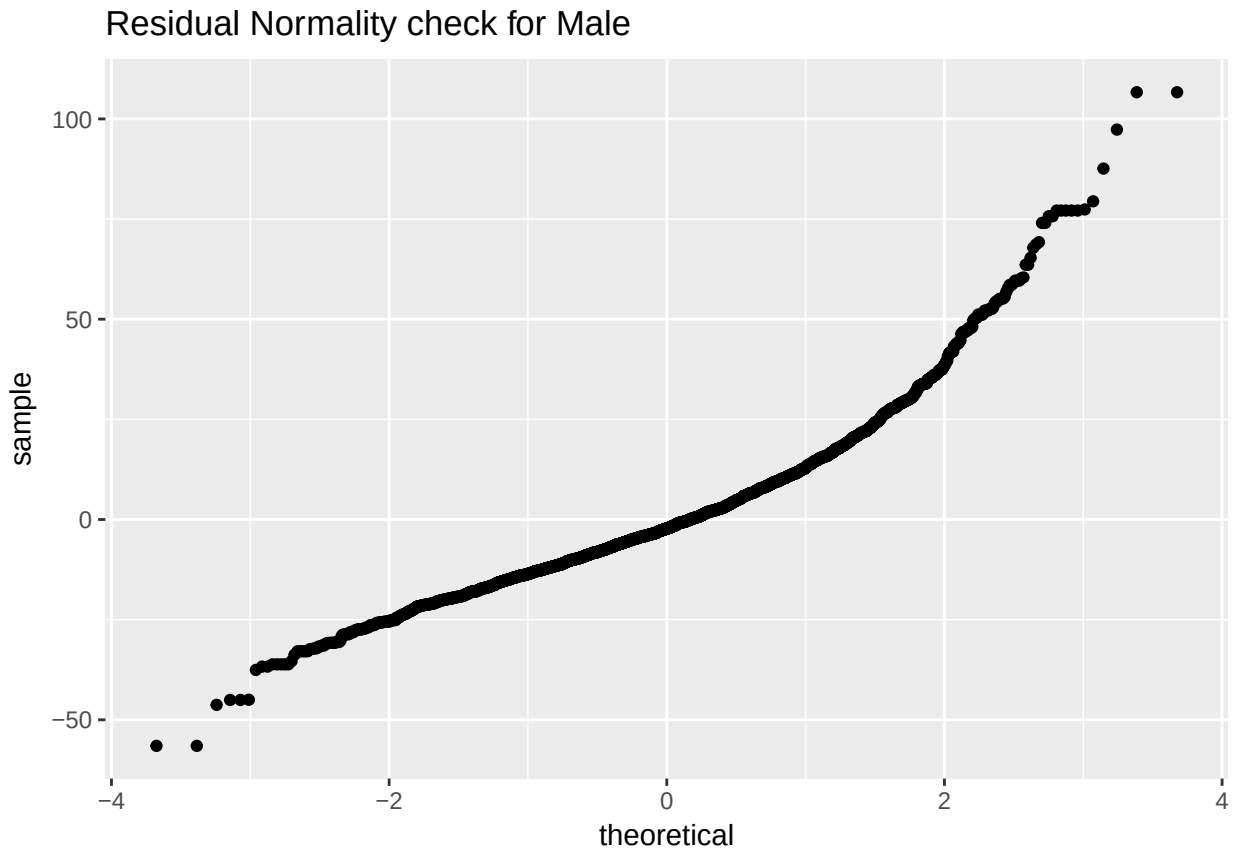
The orange color fit is using the Tukey's bisquare method. Both linear and bisquare fit are almost similar. The quadratic and GAM trend show a bend at the end and as the number of data points is less towards end, we cannot assume the curve to be a good fit at th end. Hence we can consider the linear fit for this case.

Checking the spread

```
Male_weight.lo = lm(BPSysAve ~ Weight,data = Male_data)
Male_weight.lo.df=augment(Male_weight.lo)
ggplot(Male_weight.lo.df, aes(x = Weight, y = .resid)) + geom_point(alpha=0.1,size=1) + geom_smooth(meth
```

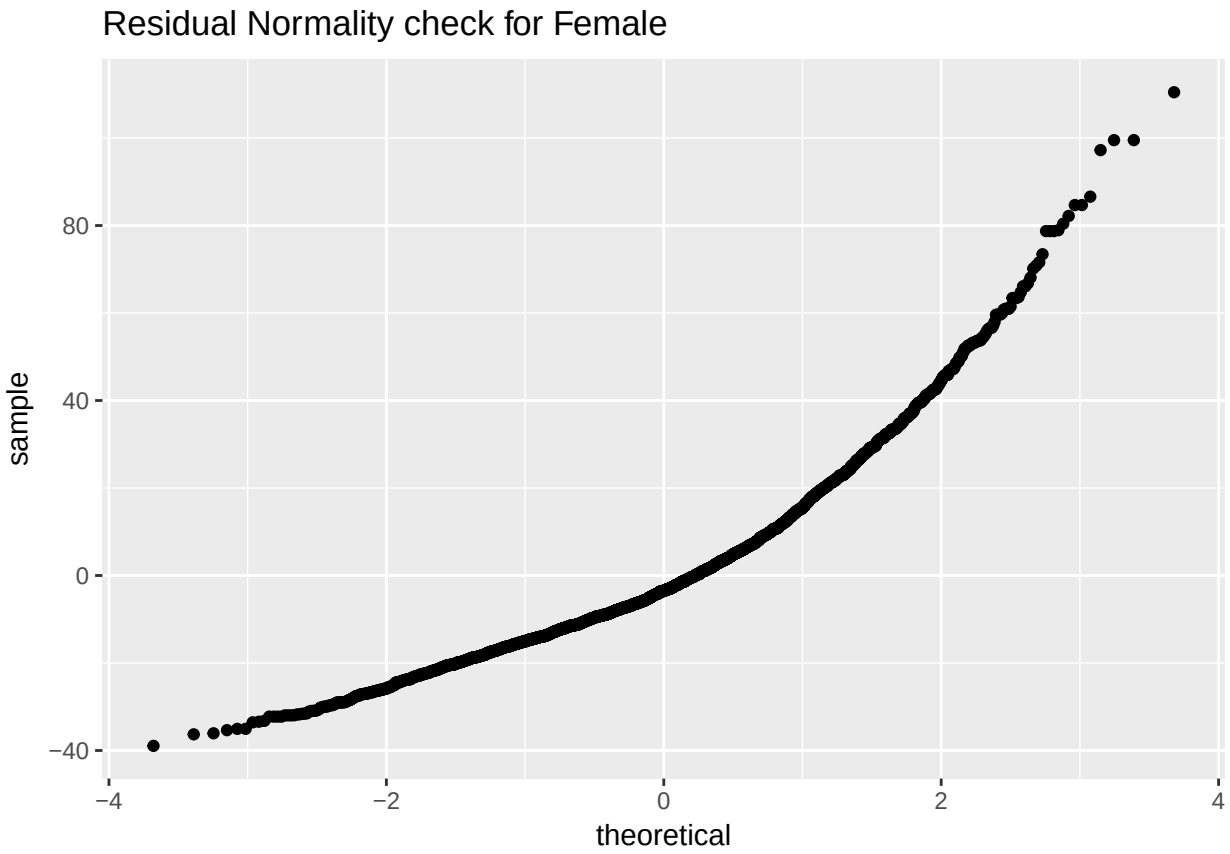



```
ggplot(Male_weight.lo.df, aes(sample = .resid)) + stat_qq()+ggtitle("Residual Normality check for Male")
```




```
Female_weight.lo = lm(BPSysAve ~ Weight,data = Female_data)
Female_weight.lo.df=augment(Female_weight.lo)
ggplot(Female_weight.lo.df, aes(x = Weight, y = .resid)) + geom_point(alpha=0.1,size=1) + geom_smooth(m
ggtitle("Residual Plot for Female")
```


```
ggplot(Female_weight.lo.df, aes(sample = .resid)) + stat_qq()+ggtitle("Residual Normality check for Fema
```



Residuals plots for both male and female are almost similar and it doesn't look like there is any trend in the residuals. When we check the normality , the residuals for both male and female is close to normal but not perfectly normal except for the edges.

Checking the mean of residuals

```
mean(Male_age.lo.df$.resid)
```

```
## [1] 6.723514e-16
```

```
mean(Female_age.lo.df$.resid)
```

```
## [1] 1.609629e-16
```

We can see that mean of the residuals is almost 0

Height Vs BPSysAve

Checking the trend

```
library(MASS)
Height_data=ggplot(NHANES, aes(x = Height, y = BPSysAve)) + geom_point(alpha=0.1,size=1)

Height_data+geom_smooth(method='lm')+
```

```
geom_smooth(method='lm',formula = y ~ x + I(x^2),col="red")+geom_smooth(col="yellow")+
geom_smooth(method = "rlm", se = FALSE,
col = "orange",method.args = list(psi = psi.bisquare))+facet_wrap(~Gender,ncol=1)
```

```
## Warning: Removed 1501 rows containing non-finite values (stat_smooth).
```

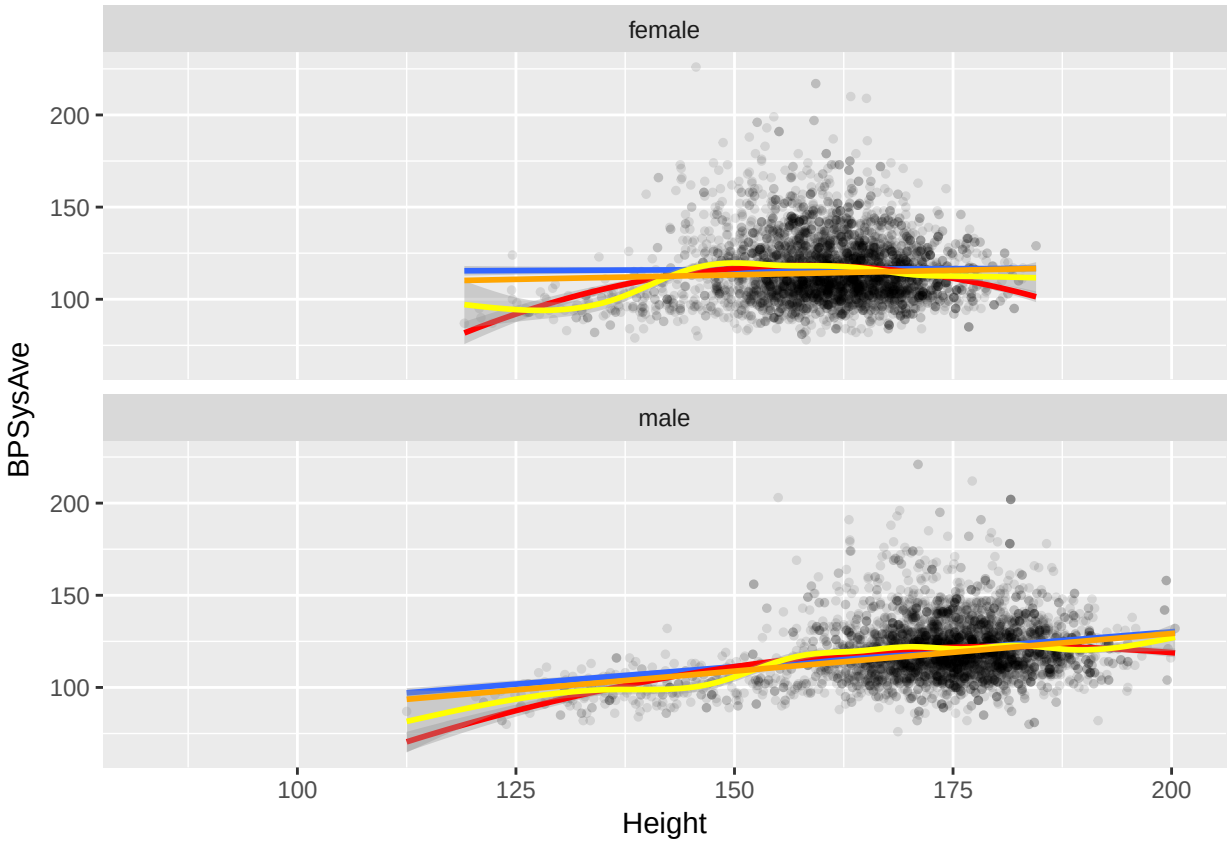
```
## Warning: Removed 1501 rows containing non-finite values (stat_smooth).
```

```
## `geom_smooth()` using method = 'gam'
```

```
## Warning: Removed 1501 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 1501 rows containing non-finite values (stat_smooth).
```

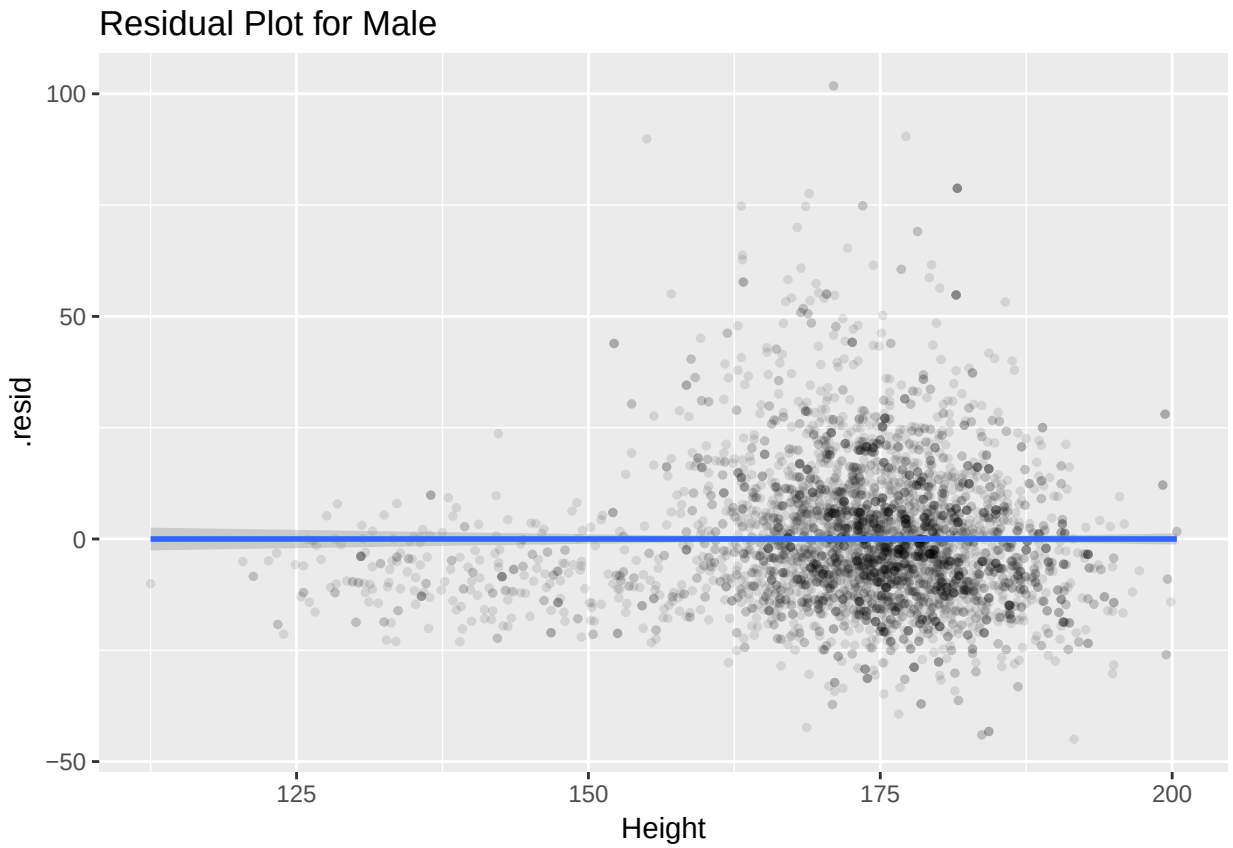
```
## Warning: Removed 1501 rows containing missing values (geom_point).
```



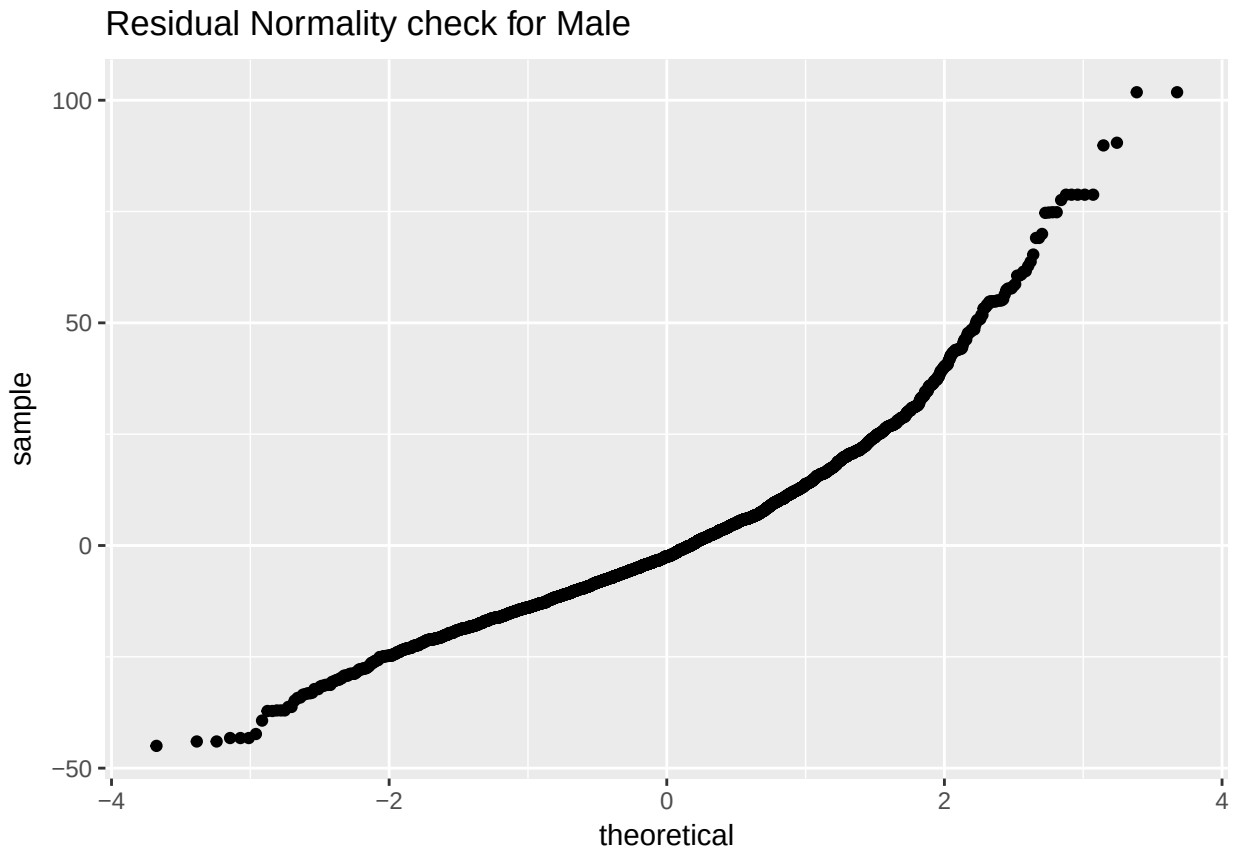
All the trends at the start are different and this might be because as explained earlier height and age are related . And as we have less data for non adults at start there is no valid explanation for the trends. But after a point almost all the trends are similar to each other. If we consider the trend after a specific point we can assume it to be linear.

Checking the spread

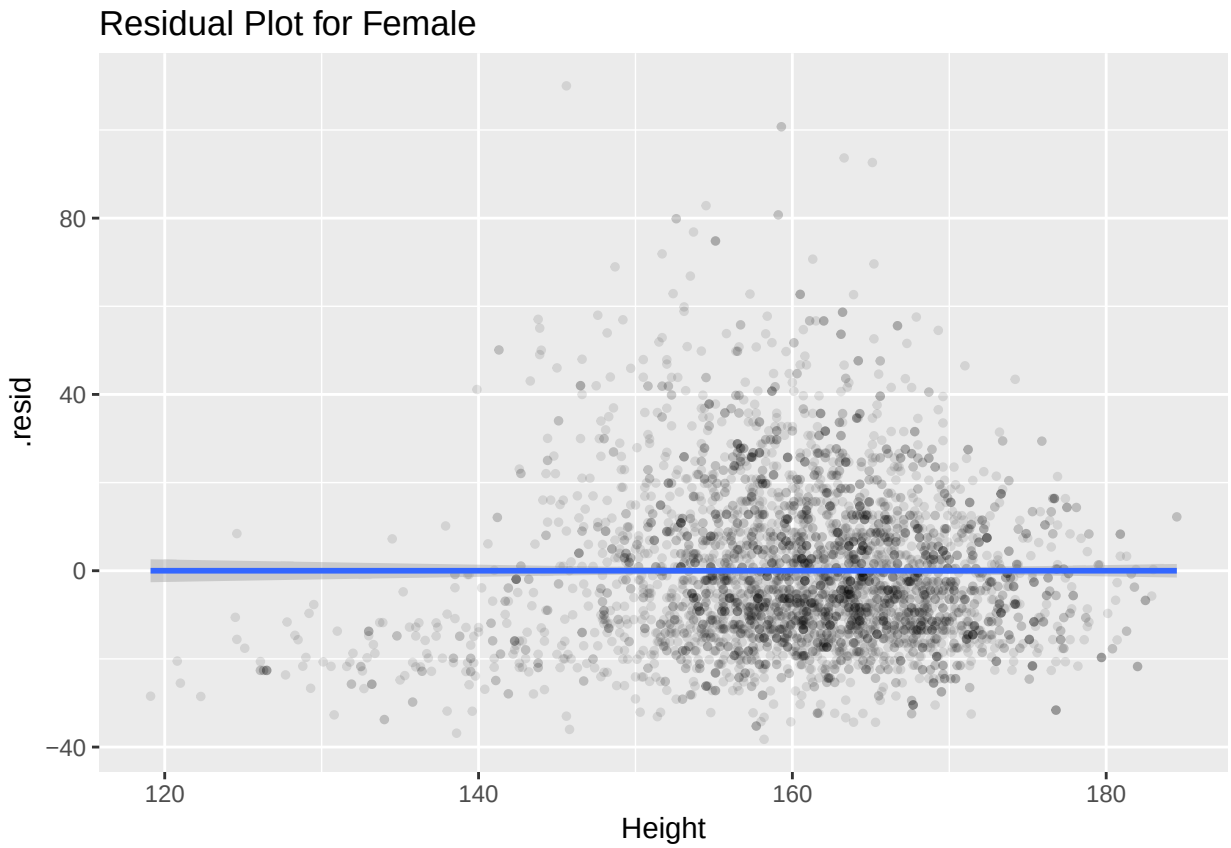
```
Male_height.lo = lm(BPSysAve ~ Height,data = Male_data)
Male_height.lo.df=augment(Male_height.lo)
ggplot(Male_height.lo.df, aes(x = Height, y = .resid)) + geom_point(alpha=0.1,size=1) + geom_smooth(meth
```



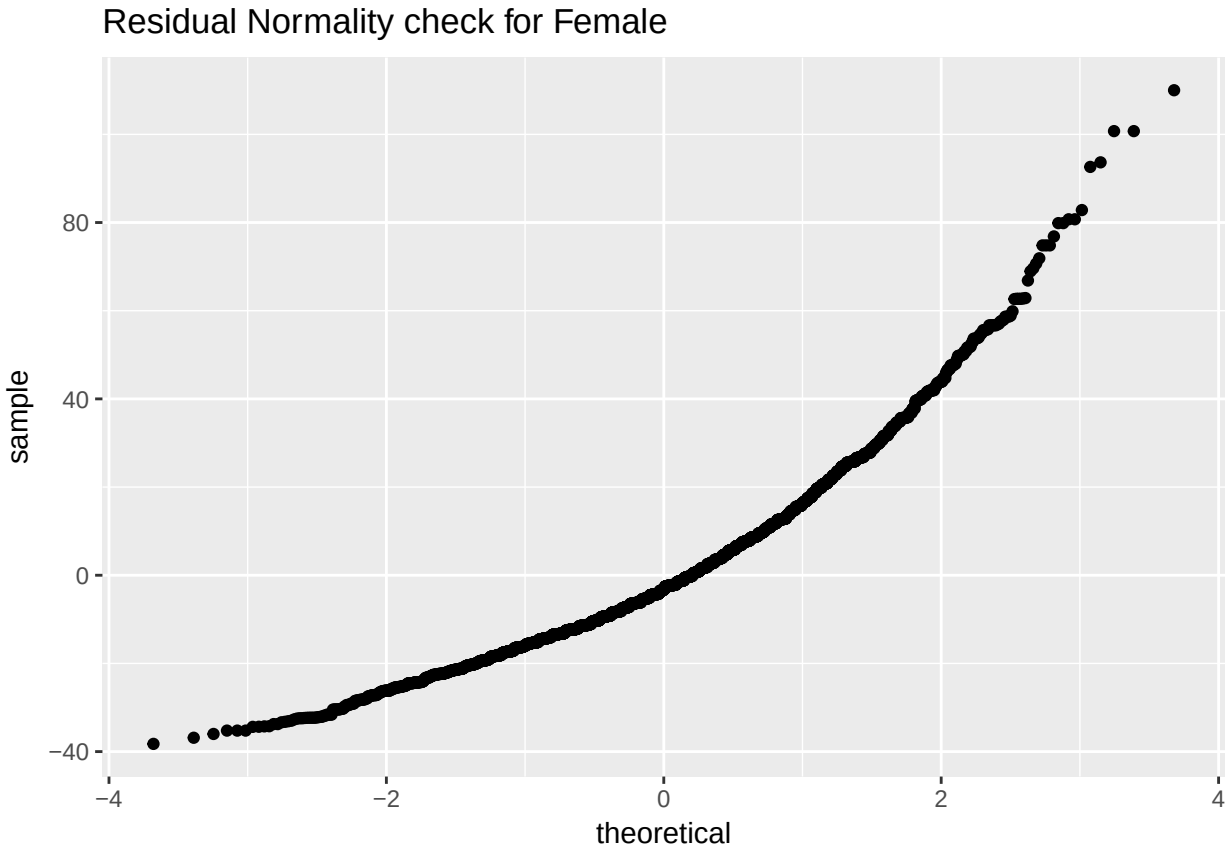

```
ggplot(Male_height.lo.df, aes(sample = .resid)) + stat_qq()+ggtitle("Residual Normality check for Male")
```




```
Female_height.lo = lm(BPSysAve ~ Height,data = Female_data)
Female_height.lo.df=augment(Female_height.lo)
ggplot(Female_height.lo.df, aes(x = Height, y = .resid)) + geom_point(alpha=0.1,size=1) + geom_smooth(m
ggtitle("Residual Plot for Female")
```




```
ggplot(Female_height.lo.df, aes(sample = .resid)) + stat_qq()+ggtitle("Residual Normality check for Fema
```



The residual plot is not exactly normal because the lower values of height are not explained by the fit well and the number of values are less for that range of height.