Problem Set 3

In [2]:
```
library(ggplot2)
library(NHANES)
df = NHANES
df = df[,c("BPSysAve","Age","Weight","Height","Gender")]
```

In [3]:
```
head(df)
```

BPSysAve	Age	Weight	Height	Gender
113	34	87.4	164.7	male
113	34	87.4	164.7	male
113	34	87.4	164.7	male
NA	4	17.0	105.4	male
112	49	86.7	168.4	female
86	9	29.8	133.1	male

Here we are dropping those rows which does not have the BPSysAve as replacing them with mean gives us spurious results

In [4]:
```
colSums(is.na(df))
df = df[complete.cases(df, df$BPSysAve),]
```

BPSysAve	1449
Age	0
Weight	78
Height	353
Gender	0

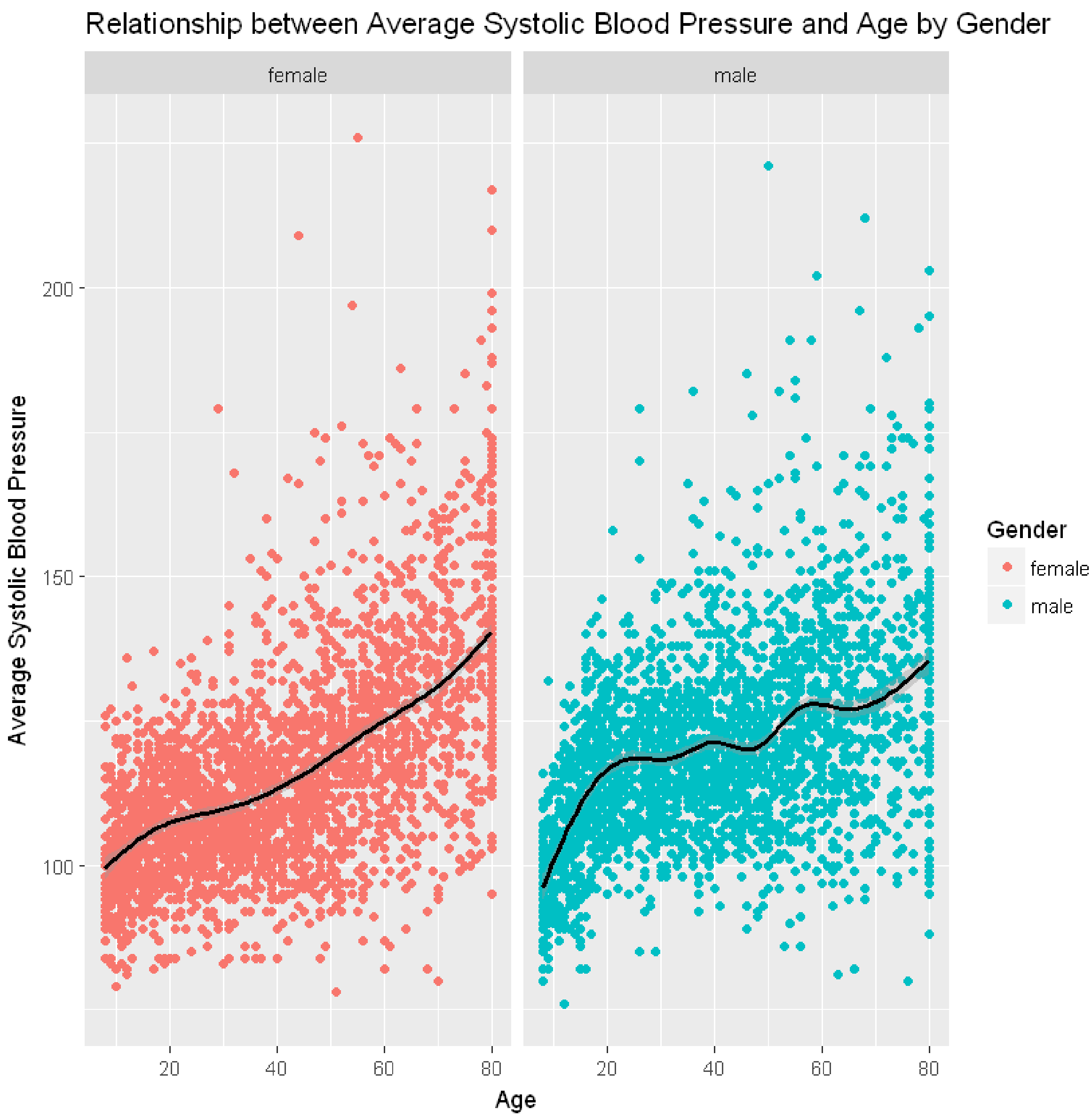
In [36]:
```
colSums(is.na(df))
```

BPSysAve	0
Age	0
Weight	0
Height	0
Gender	0

Section1: Relationship between Average Systolic Blood Pressure and Age

In [39]:
```
ggplot(df,aes(x=Age,y=BPSysAve,color=Gender)) +  geom_point()  +
  geom_smooth( method.args= list(degree=1),col = "black",span=1) + facet_grid(
~Gender) +
ylab("Average Systolic Blood Pressure") + ggtitle("Relationship between Averag
e Systolic Blood Pressure and Age by Gender")
```

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

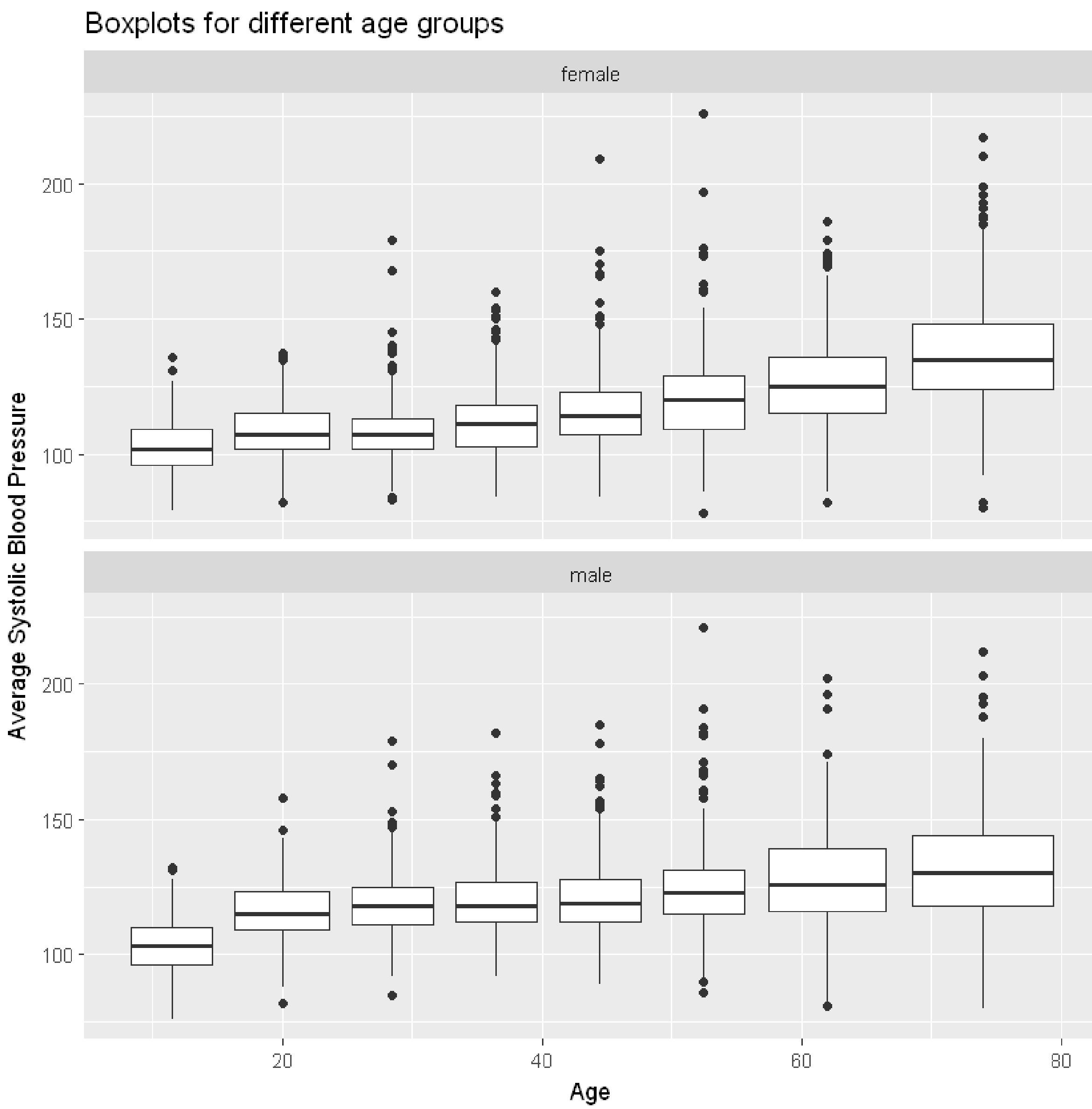


It looks like there is a weak Positive Correlation between Average Systolic Blood Pressure and Age.

The average systolic Blood Pressure increases gradually with age for both genders.

A definitive linear trend is not observed.

In [111]:
```
ggplot(df, aes(x=Age,y=BPSysAve)) +geom_boxplot(aes(group= cut_number(Age, n =
 8))) + facet_wrap(~Gender,ncol=1)+
ylab("Average Systolic Blood Pressure") + ggtitle("Boxplots for different age
 groups")
```

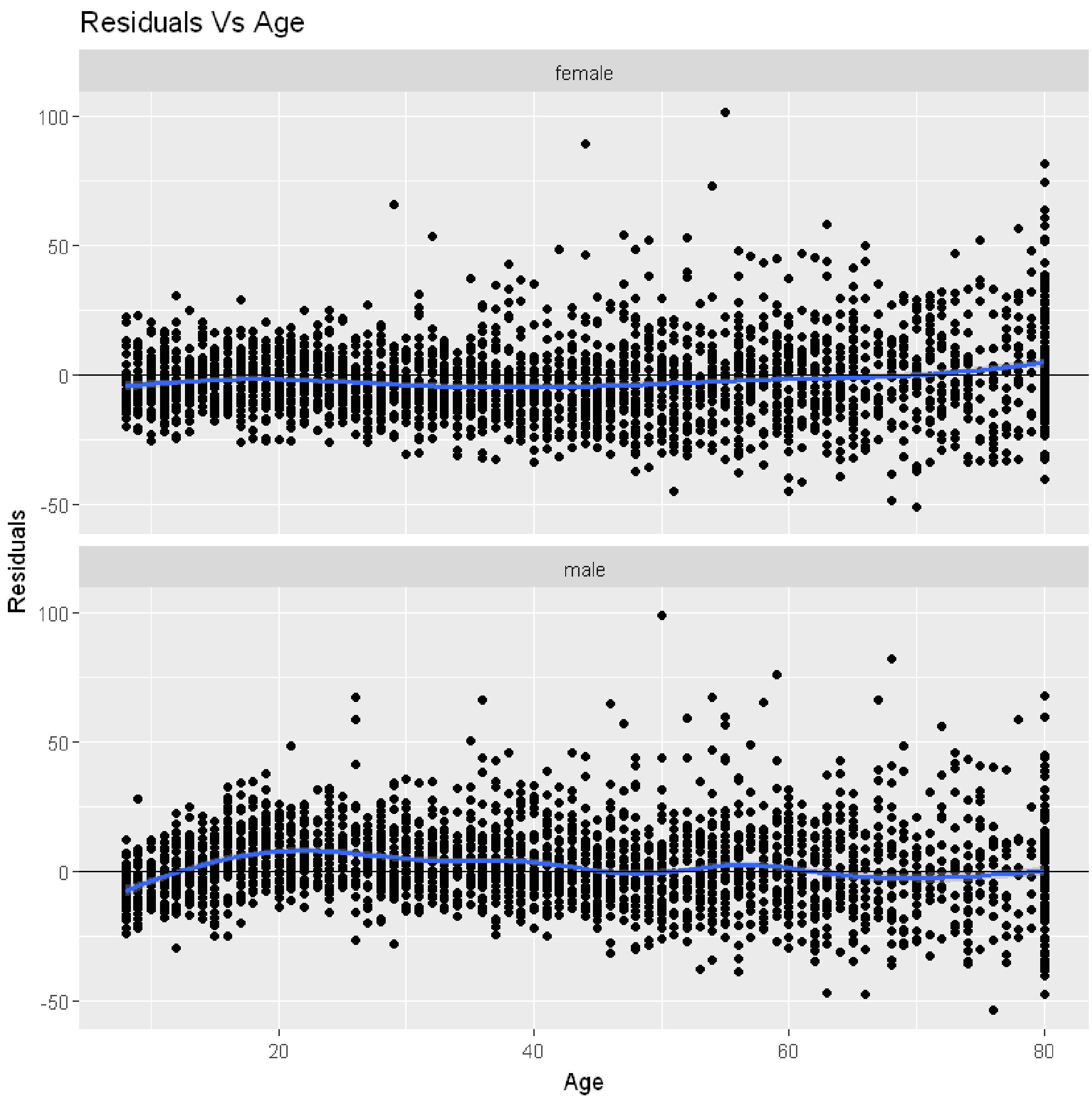


For each of the 8 age groups, it can be observed that the spread of average systolic blood pressure is similar with age-group for both genders.

In [113]:
```
library(broom)
df.lm= lm(BPSysAve ~ Age, data=df)
df.lm.au = augment(df.lm)
df.lm.au$Gender = df$Gender
```

In [114]:
```
ggplot(df.lm.au, aes(x=Age, y= .resid)) + geom_point() + geom_smooth() +
  geom_abline(slope = 0, intercept = 0)+ facet_wrap(~Gender,ncol=1) + ylab("Re
siduals") + ggtitle("Residuals Vs Age")
```

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

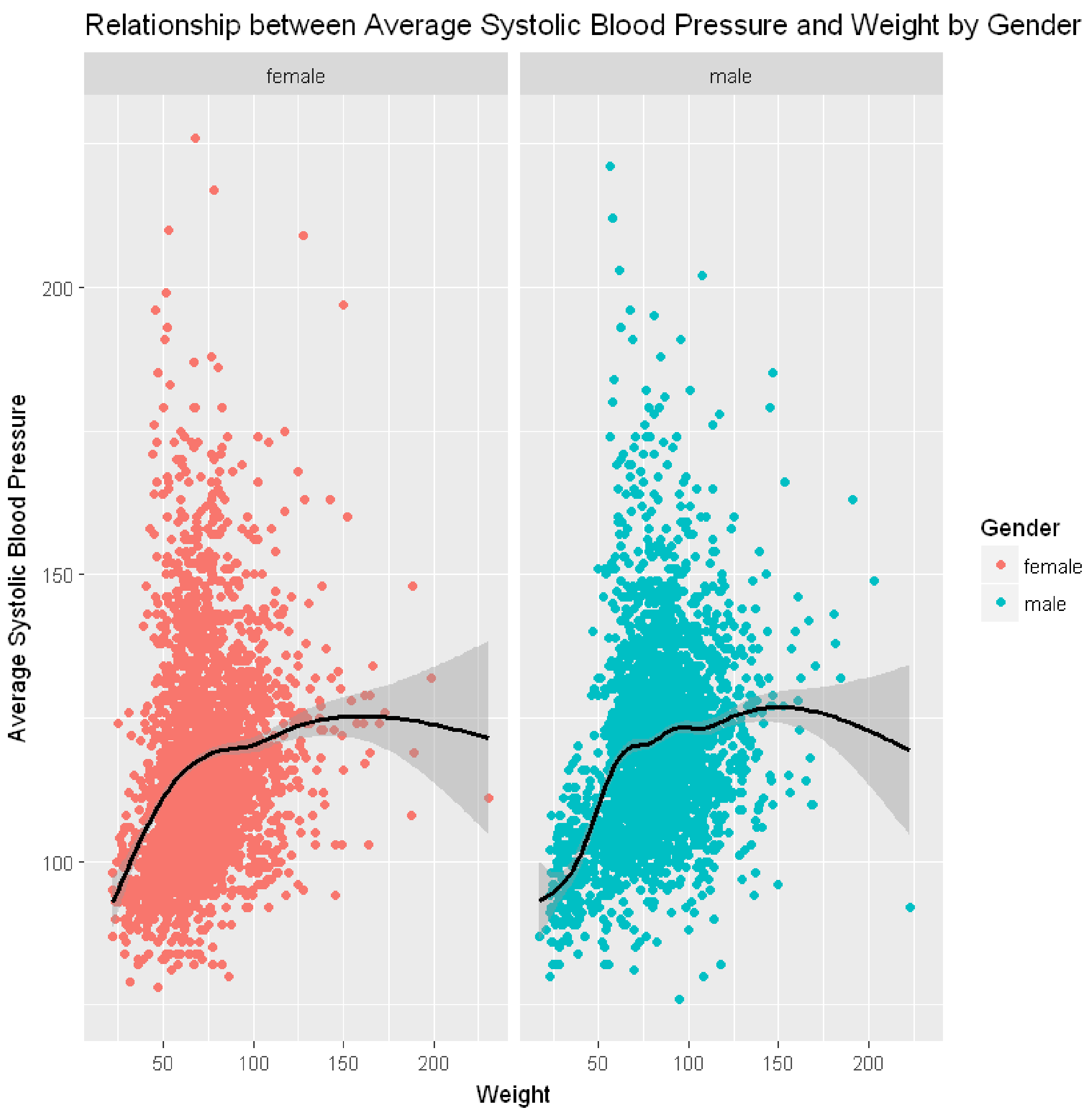


It looks like the model can be fitted best with a linear model. But the confidence level does not contain all the points, it means there is a less correlation between Average Systolic Blood Pressure and Age.

Section2: Relationship between Average Systolic Blood Pressure and Weight

In [104]:
```
ggplot(df,aes(x=Weight,y=BPSysAve,color=Gender)) +  geom_point()  +
  geom_smooth( method.args= list(degree=1),col = "black",span=1) + facet_grid(
~Gender) +
ylab("Average Systolic Blood Pressure") + ggtitle("Relationship between Averag
e Systolic Blood Pressure and Weight by Gender")
```

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



Eyeballing at the above plot, it can be said that Linear approximation model is not appropriate for this data for both the genders.

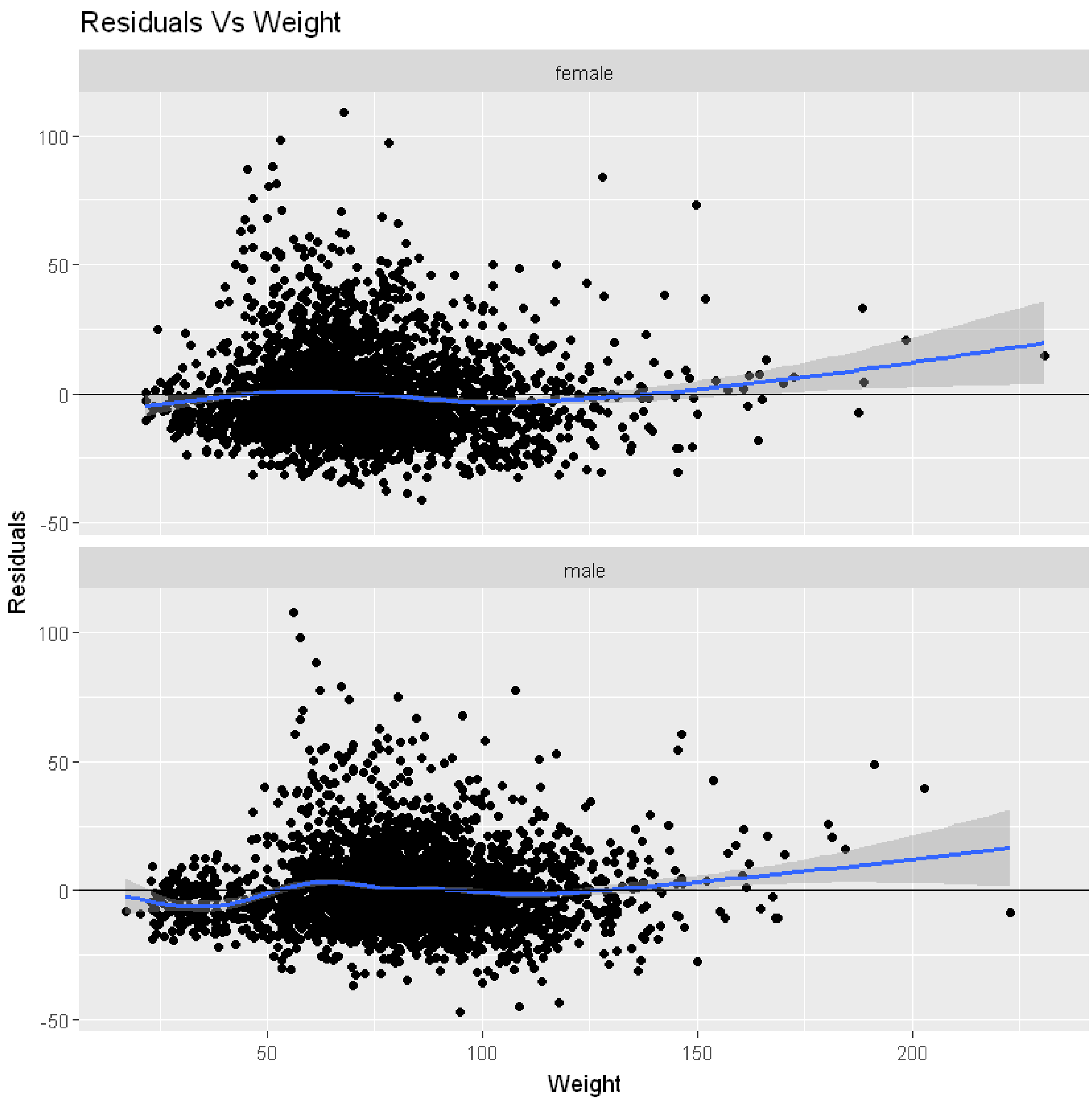
For weights about 150 and above, the Blood Pressure values deviate from the existing trend.

In [100]:
```
library(broom)
df.lm.wt = lm(BPSysAve ~ Weight + I(Weight^2), data=df)
df.lm.wt.au = augment(df.lm.wt)
df.lm.wt.au$Gender = df$Gender
```

In [115]:
```
ggplot(df.lm.wt.au, aes(x=Weight, y= .resid)) + geom_point() + geom_smooth() +

  geom_abline(slope = 0, intercept = 0)+ facet_wrap(~Gender,ncol=1) + ggtitle(
"Residuals Vs Weight") +
ylab("Residuals")
```

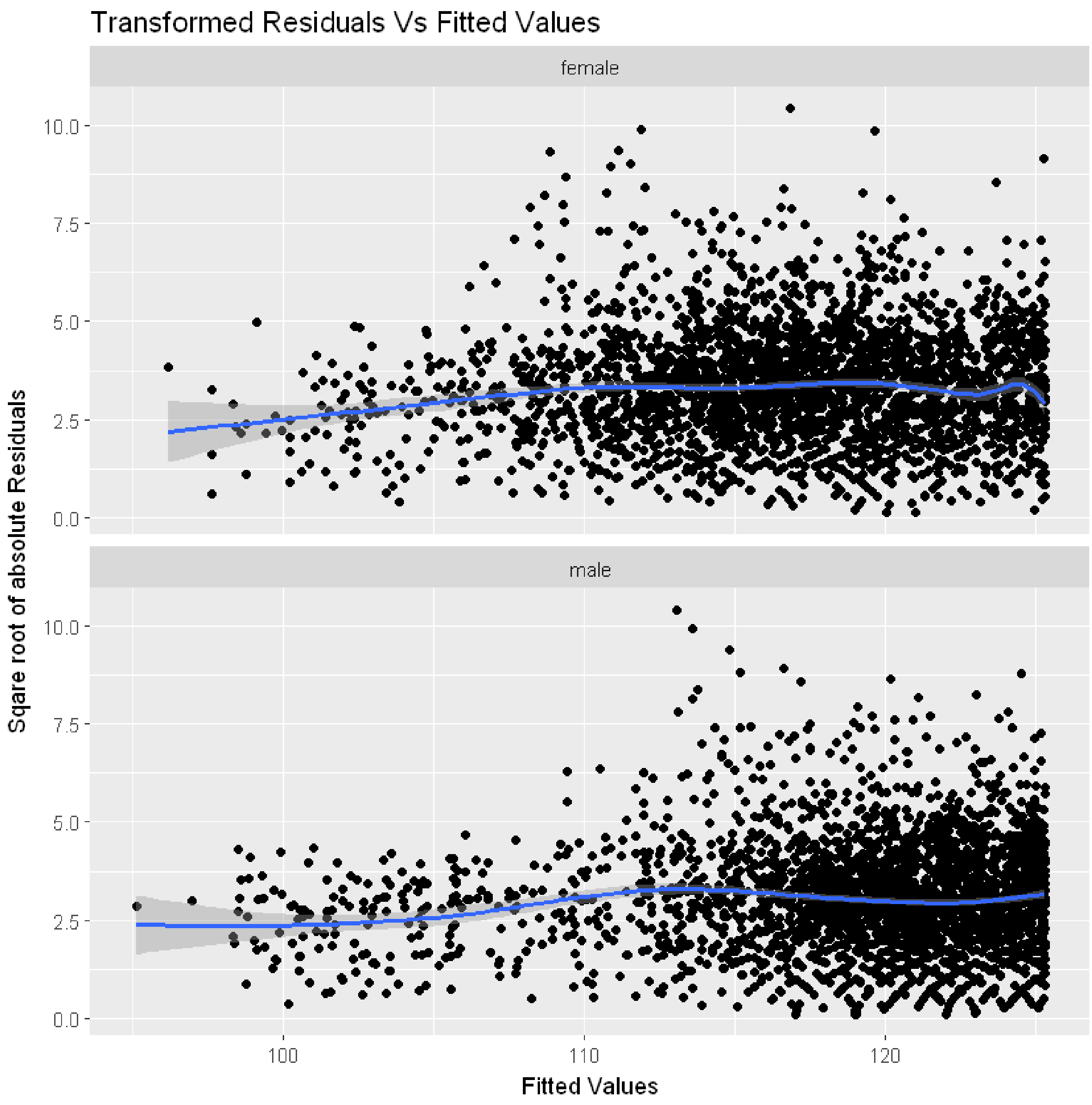
```
`geom_smooth()` using method = 'gam'
```



The curve moves just around zero. Thus the Quadratic model seems apt for this data.

In [116]:
```
ggplot(df.lm.wt.au, aes(x=.fitted, y= sqrt(abs(.resid)))) + geom_point() + geo
m_smooth() +
  facet_wrap(~Gender,ncol=1) + xlab("Fitted Values") + ylab("Sqare root of abs
olute Residuals") +
ggtitle("Transformed Residuals Vs Fitted Values")
```

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

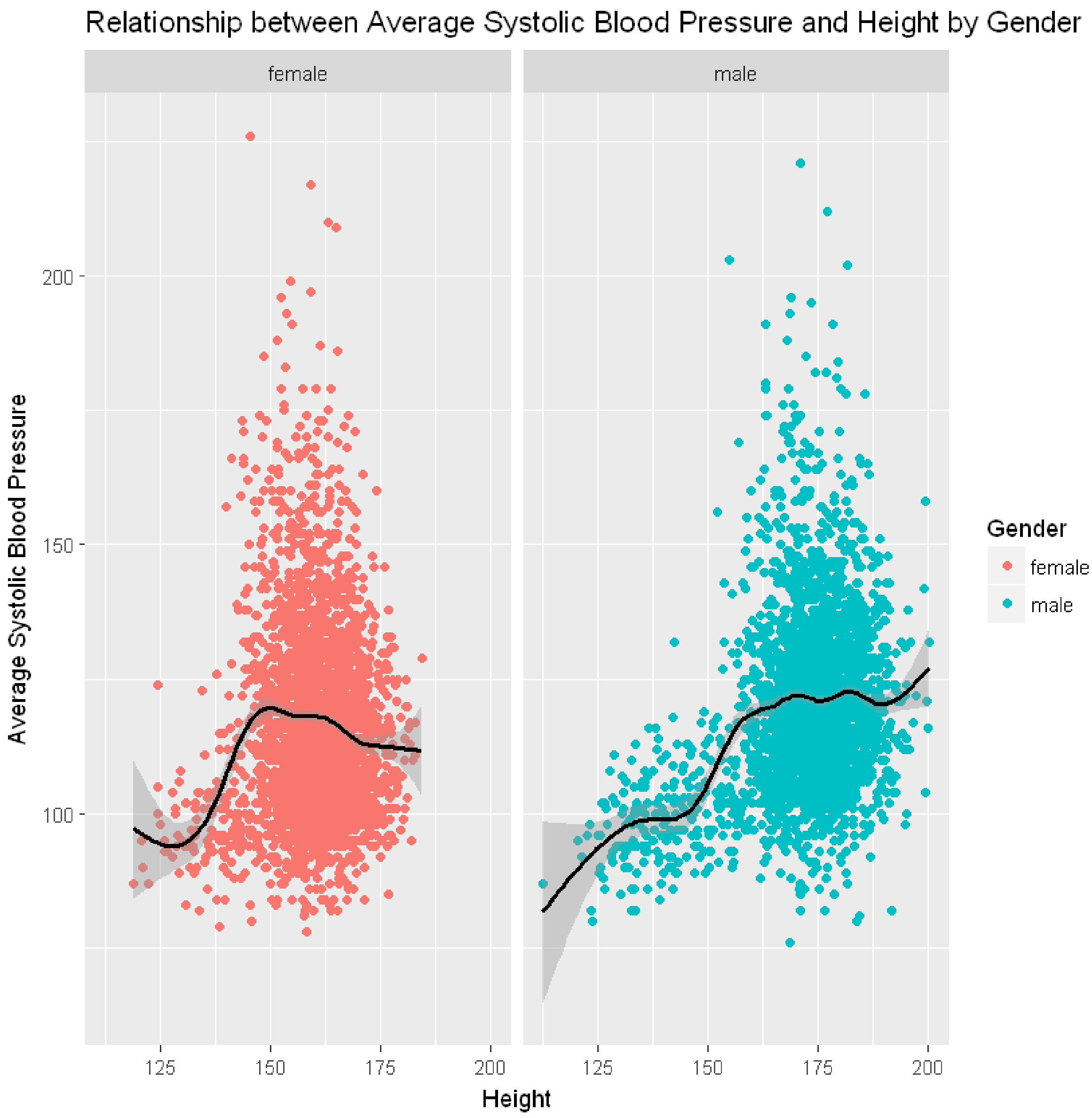


The curve clearly appears to be a horizontal line which proves Homoscedasticity in the data. But the confidence band does not cover majority of data points.

Section3: Relationship between Average Systolic Blood Pressure and Height

In [117]:
```
ggplot(df,aes(x=Height,y=BPSysAve,color=Gender)) +  geom_point()  +
  geom_smooth(col = "black", span = 1) + facet_grid(~Gender) +
ylab("Average Systolic Blood Pressure") + ggtitle("Relationship between Averag
e Systolic Blood Pressure and Height by Gender")
```

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



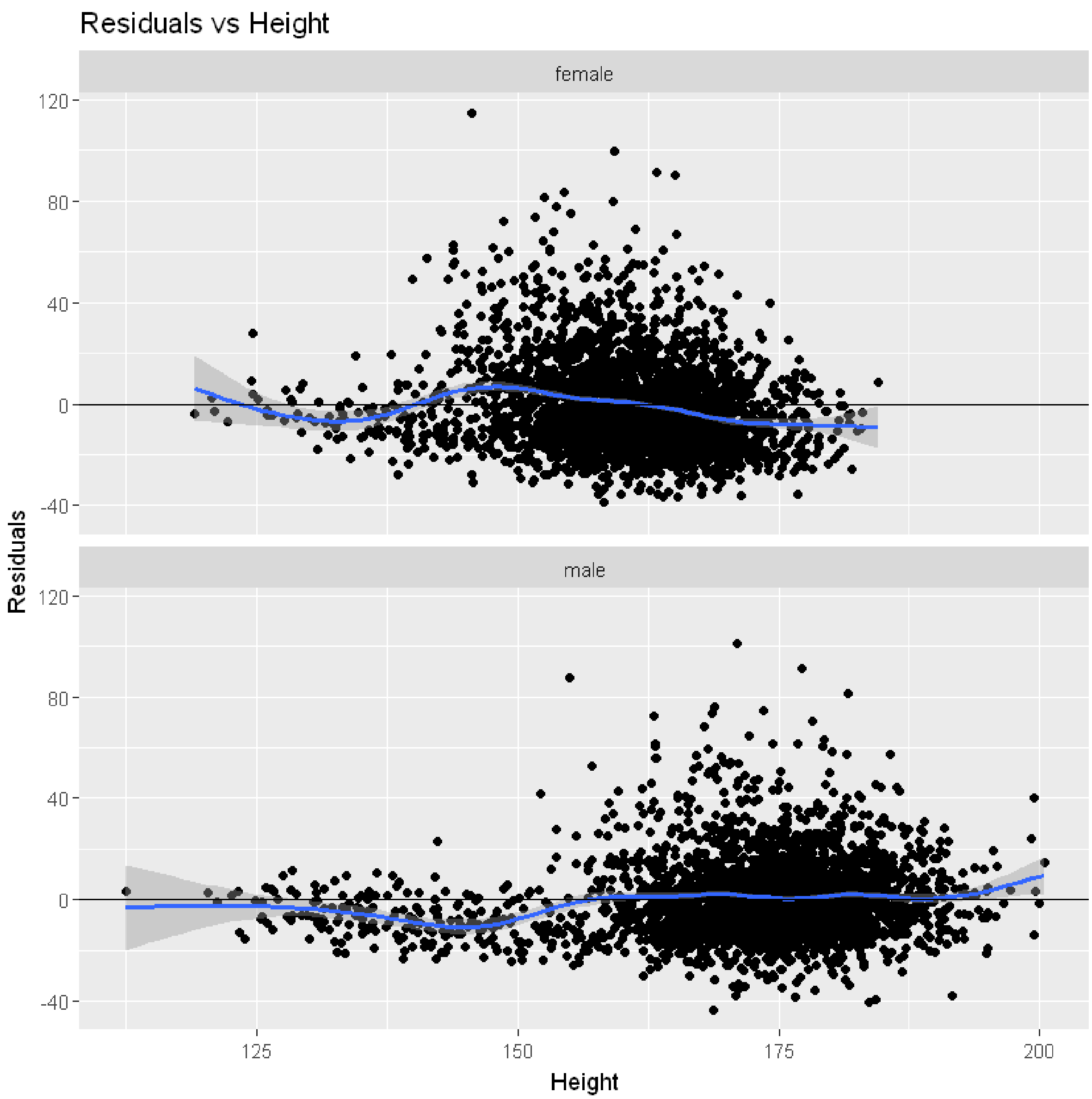
By looking at the above plot we can say that Linear model is not a best fit for the given data

In [120]:
```
library(broom)
df.lm.ht = lm(BPSysAve ~ Height + I(Height^2), data=df)
df.lm.ht.au = augment(df.lm.ht)
df.lm.ht.au$Gender = df$Gender
```

In [123]:
```
ggplot(df.lm.ht.au, aes(x=Height, y= .resid)) + geom_point() + geom_smooth() +

  geom_abline(slope = 0, intercept = 0)+ facet_wrap(~Gender,ncol=1) + ylab("Re
siduals") +
ggtitle("Residuals vs Height")
```

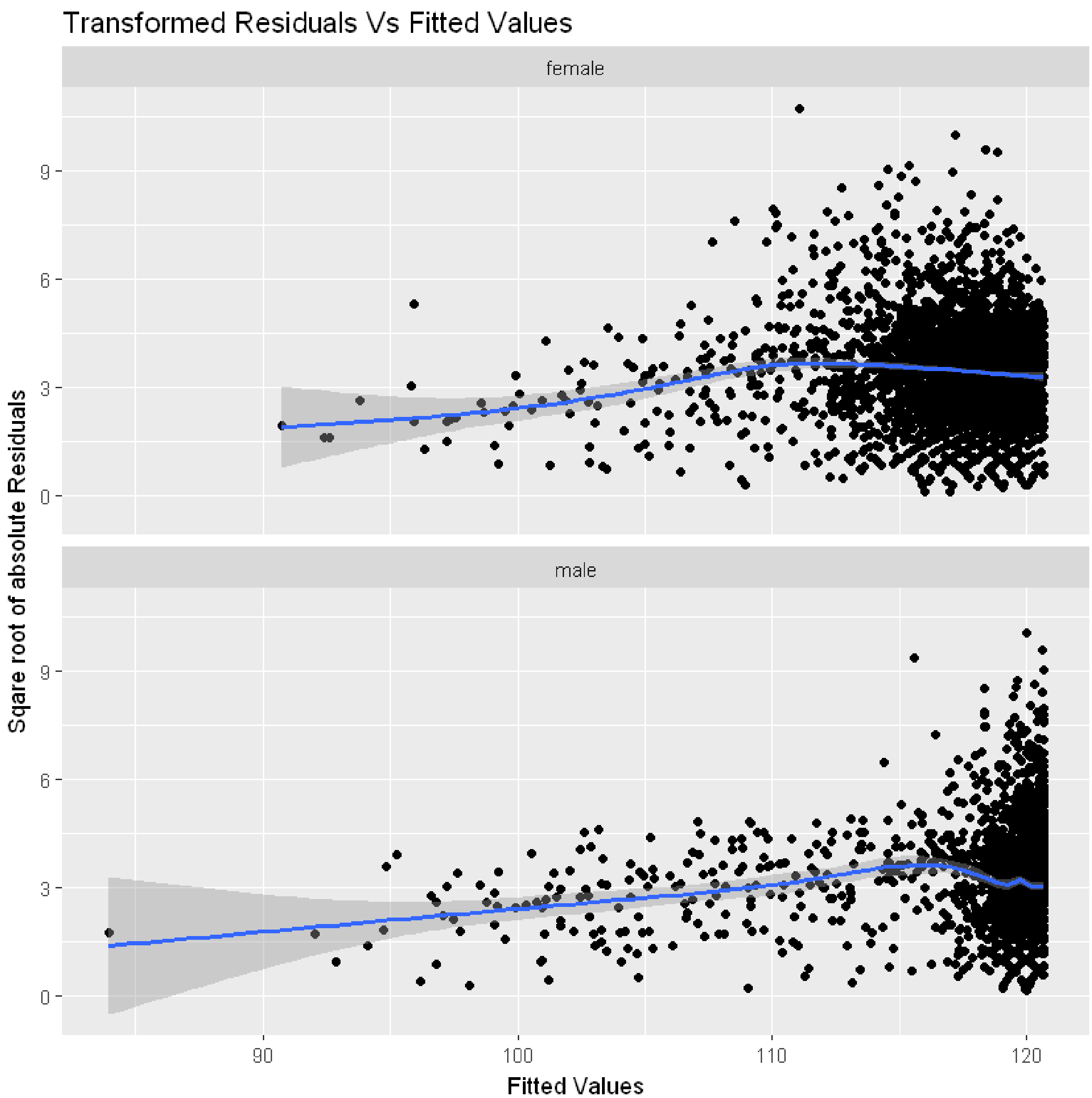
```
`geom_smooth()` using method = 'gam'
```



The residuals move just around zero. It implies that the quadratic model is appropriately fitted to the data

In [124]:
```
ggplot(df.lm.ht.au, aes(x=.fitted, y= sqrt(abs(.resid)))) + geom_point() +
 geom_smooth() +
  facet_wrap(~Gender,ncol=1) + xlab("Fitted Values") + ylab("Sqare root of
 absolute Residuals") +
ggtitle("Transformed Residuals Vs Fitted Values")
```

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



It can clearly be observed that the confidence band does not include majority of the data points. The line is nearly horizontal for most portion of the curve.