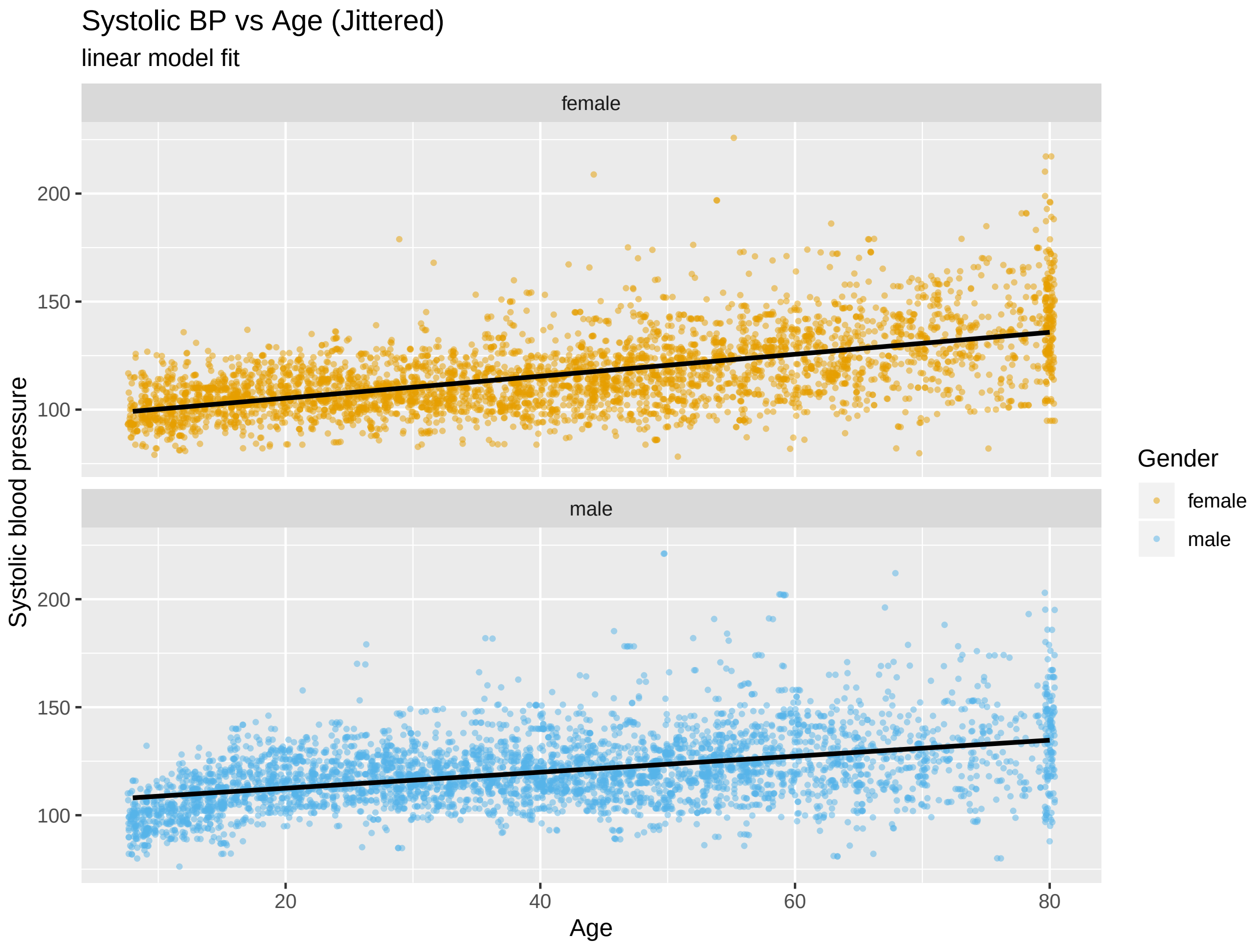
PS3_Answers

vkvats

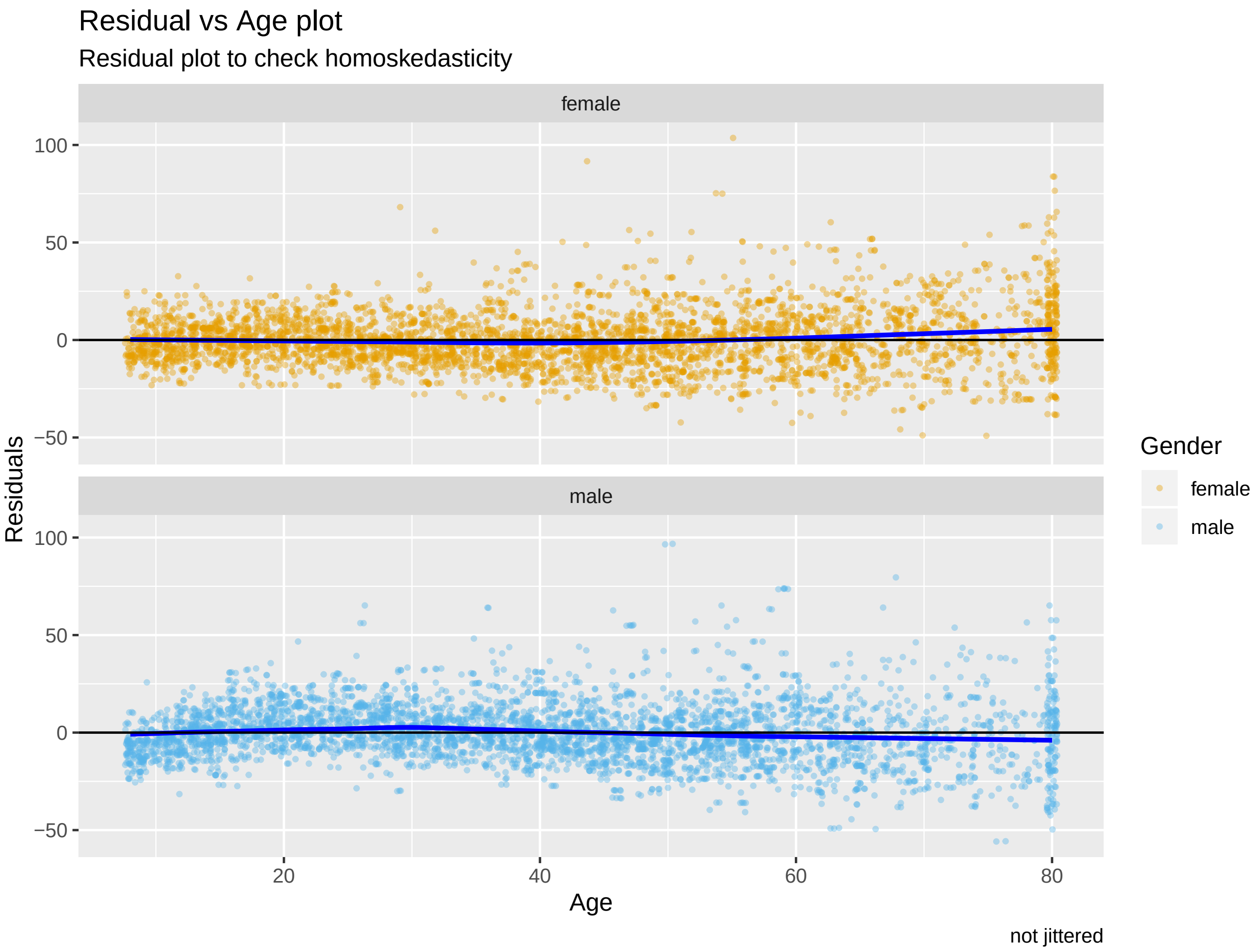
```r
library(ggplot2)
library(NHANES)
library(broom)
library(MASS)
nhanes = NHANES[, c("Age", "Weight", "Height", "Gender", "BPSysAve")]
cb_palette = c("#E69F00", "#56B4E9", "#009E73", "#999999", "#F0E442", "#0072B2", "#D55E00", "#CC79A7")
```

Section 1: systolic blood pressure with Age

```r
BP.age = na.omit(nhanes[,c("BPSysAve", "Age", "Gender")])
BP.age.male = BP.age[BP.age$Gender == 'male',]
BP.age.female = BP.age[BP.age$Gender == 'female',]

## Model fitting
BP.age.lm = lm(BPSysAve ~ Age + Gender, data = BP.age)
BP.age.lm.df = augment(BP.age.lm)

## Fitted model plot
gg = ggplot(BP.age.lm.df, aes(x = Age, y = BPSysAve, color = Gender)) +
  geom_jitter(height =  0.25, size = 0.7, alpha = 0.5)
gg + geom_smooth( method = "lm", se = F, color = "black") +
  facet_wrap(~Gender, ncol = 1) +
   scale_color_manual(values = cb_palette) +
  labs(title = "Systolic BP vs Age (Jittered)",
       subtitle = "linear model fit") +
  xlab("Age") + ylab("Systolic blood pressure")
```



```
# residual plot
gg = ggplot(BP.age.lm.df, aes(x = Age, y = .resid, color = Gender)) +
  geom_jitter(height =  0.25, size = 0.7, alpha = 0.4)
gg + geom_smooth(method = "loess", method.args=list(degree=1), se = F, col = "blue") +
  geom_abline(slope = 0) +
  facet_wrap(~Gender, ncol=1) +
   scale_color_manual(values = cb_palette) +
  labs(title = "Residual vs Age plot",
       subtitle = "Residual plot to check homoskedasticity",
       caption = "not jittered") +
  xlab("Age") + ylab("Residuals")
```



A general oversavation on the dotplot shows a monotonically increasing tend in the blood pressure with age, The large amount of data enables us to ignore the effect of gross outliers. So, We can model the systolic blood pressure and age realtion using linear model. The distribution of Age is almost symmetric but at the end (age between 60 to 80) there are a little less observations and couting the age of 80 and more as 80 has lead to a peak formation at age 80. The mean of the Age is around 36 years. In the first plot, Systolic BP vs Age (Jittered) plot, it can be seen that choosing to model male and female separately has lead to different linear line that best desribes the model. The value of slope and the intercept are different for male and female as shown in table 1. The model is explained by the line:

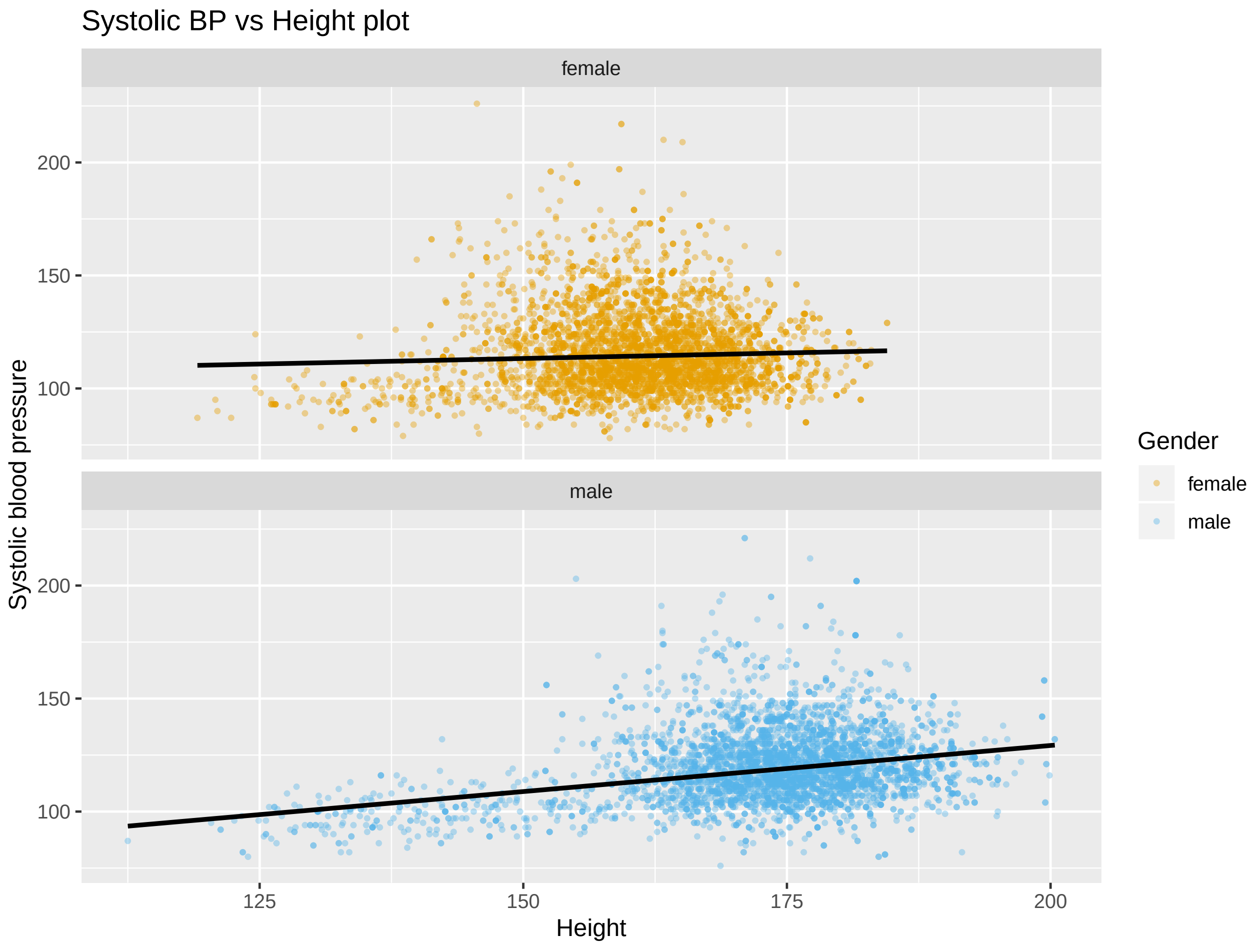
$$BPSysAve_{male} = 0.37 * Age_{male} + 105.09$$

$$BPSysAve_{female} = 0.5 * Age_{female} + 95.11$$

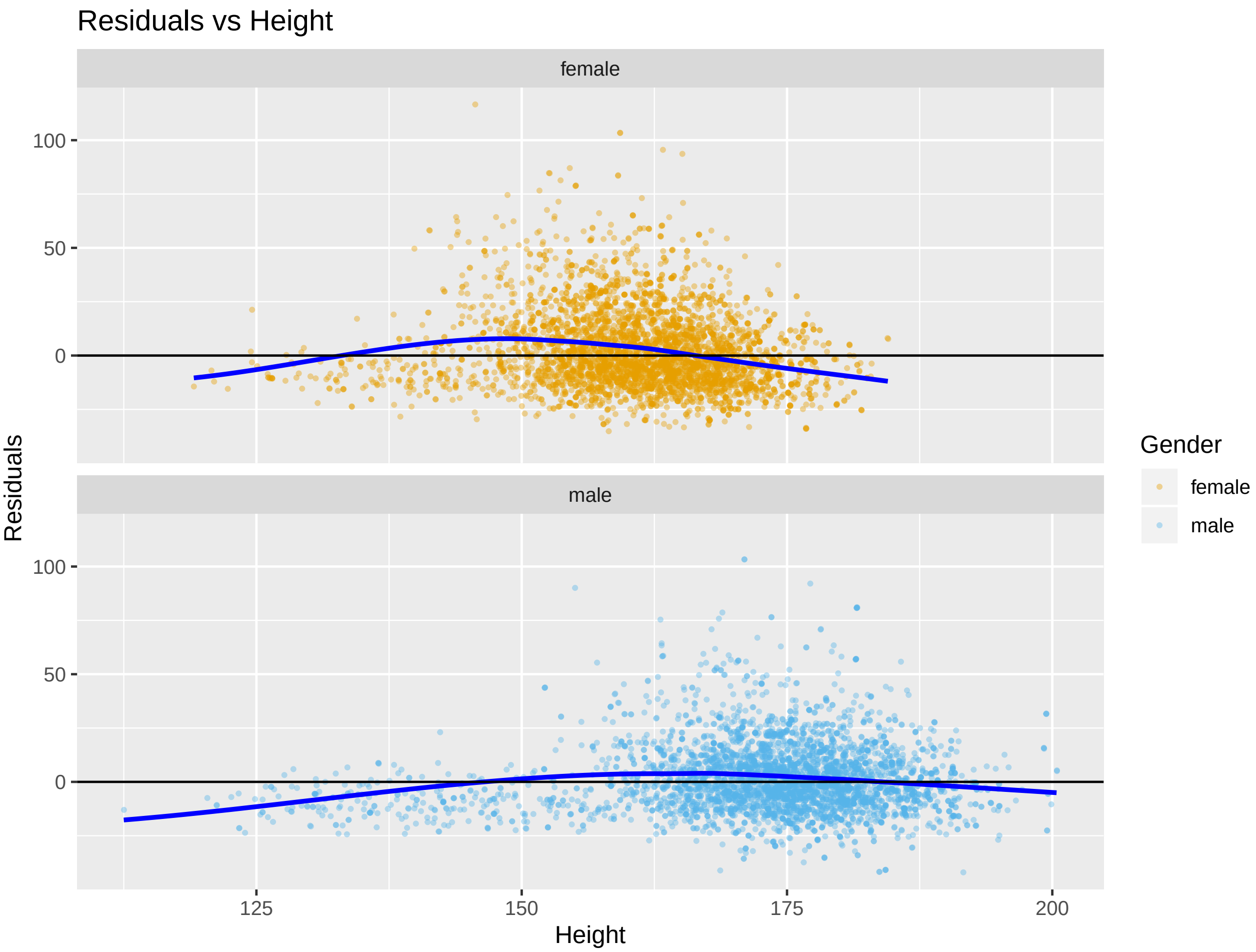
There is a general monotonously increasing trend between blood pressure and age, for both male and female, as shown by the equation above, so, we didn't need to use any transformation. The rate of inrease of blood pressure for females are greater than the rate for increase in males. It can be noted from the plot that the male blood pressure is not adequatley explained at initial values of age, the blood pressure values are smaller for small age, and at that small region this model will not properly explain the relation. The same trend is followed in residual plot as well for male, otherwise the residuals don't show any pattern. The residuals for both male and female is fairly distributed across line $y = 0$, indicating homoskedasticity. The fitted model for male explains approximately 20 percent of variance and the fitted model for female explains approximately 35 percent of the variance. This models have some significant explainatory value, and this will be best suited among Height, weight and Age, if needed for predition purposes. Still, we should never try to do predition beyong the range of data avaialable to us.

Section 2: Systolic blood pressure with Height

```r
BP.height = na.omit(nhanes[,c("BPSysAve", "Height", "Gender")])
BP.height.male = BP.height[BP.height$Gender == "male",]
BP.height.female = BP.height[BP.height$Gender == "female",]

# Model fitting
BP.height.rlm = rlm(BPSysAve ~ Height + Gender, data = BP.height, psi = psi.bisquare)
BP.height.rlm.df = augment(BP.height.rlm)

#BP.male.rlm = rlm(BPSysAve ~ Height, data = BP.height.male, psi = psi.bisquare)
#male.rlm = augment(BP.male.rlm)
#var(male.rlm$.fitted)/ var(male.rlm$BPSysAve)
#summary(BP.male.rlm)
#BP.female.rlm = rlm(BPSysAve ~ Height, data = BP.height.female, psi = psi.bisquare)
#female.rlm = augment(BP.female.rlm)
#var(female.rlm$.fitted)/ var(female.rlm$BPSysAve)
#summary(BP.female.rlm)

# plot
gg = ggplot(BP.height.rlm.df, aes(x = Height, y = BPSysAve, color = Gender)) +
  geom_point(size = 0.7, alpha = 0.4)+
  geom_smooth(method = "rlm", se = F, method.args = list(psi = psi.bisquare), color = "black") +
  facet_wrap(~Gender, ncol = 1) +
  scale_color_manual(values = cb_palette) +
  xlab("Height") + ylab("Systolic blood pressure") +
  labs(title = "Systolic BP vs Height plot")
gg
```



```
gg = ggplot(BP.height.rlm.df, aes(x = Height, y = .resid, color = Gender)) +
  geom_jitter(height =  0.25, size = 0.6, alpha = 0.4) +
  facet_wrap(~Gender, ncol = 1) +
  xlab("Height") + ylab("Residuals") +
  labs(title = "Residuals vs Height") +
  geom_smooth(method = "loess", method.args=list(degree=1), se = F, col = "blue") +
  geom_abline(slope = 0) +
   scale_color_manual(values = cb_palette)
gg
```



It can be seen from dotplot of the male and female systolic blood pressure vs height that both the data are left skewed to some extent. For initial heights especially before 150 cm, there are not much data points so it would be best to get more such data if we really want this model to be perfect. Apart from that, there are some outliers which goes beyond 175 mmg of blood pressure. It is hard to capture and explain such data with this model. To nullify the effect of outliers, robust linear model is fitted for both male and female. The model is represented by the following equation

$$BPSysAve_{male} = 0.41 * Height_{male} + 47.68$$

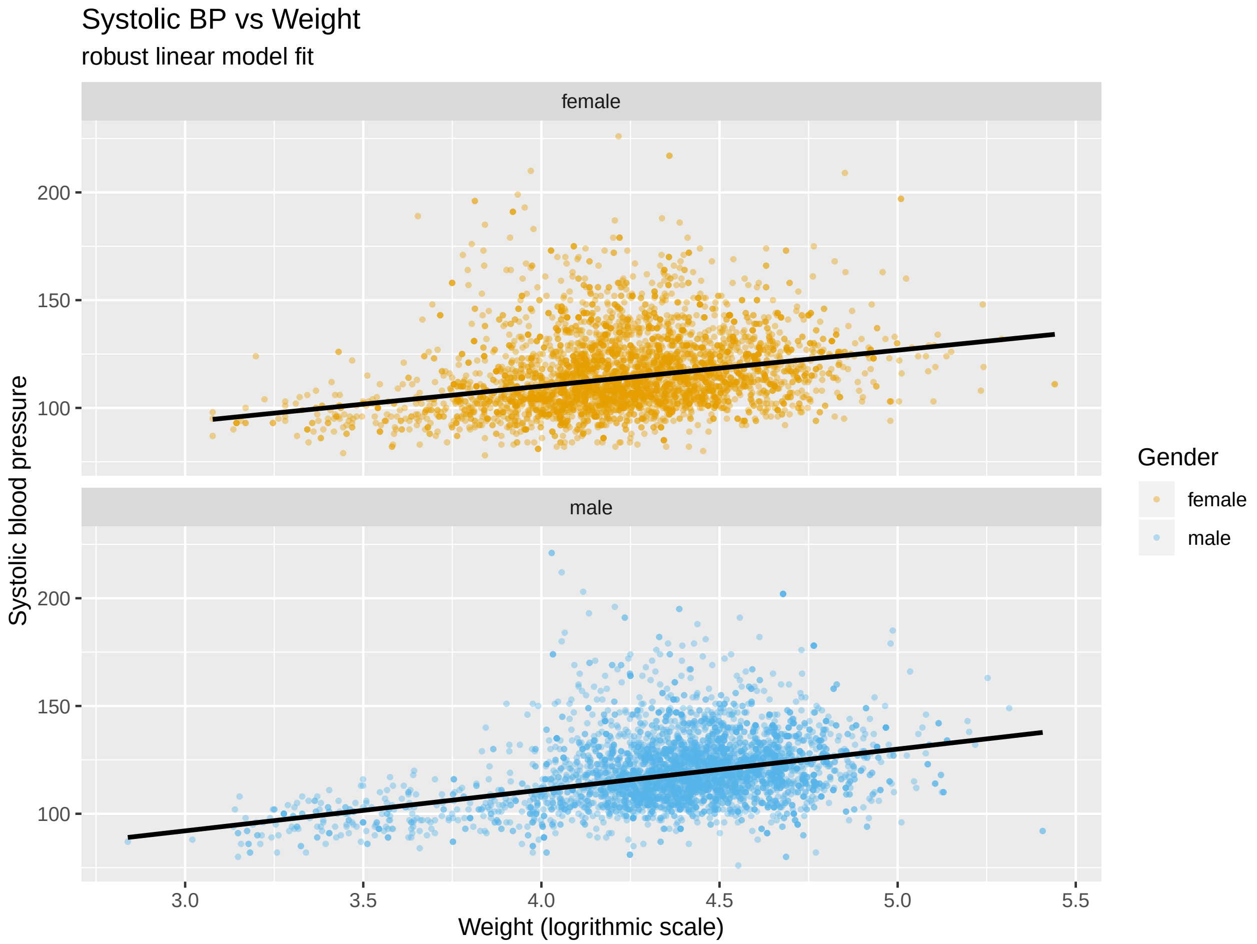
$$BPSysAve_{female} = 0.098 * Height_{female} + 98.48$$

the relation between male blood pressure and height shows a upward trend with significant slope but same is not true for female model, the slope value of fitted line is very low as good as almost flat. The model for female exaplains 0.23 percent of the variance and for males, it explains 8.2 percent of the variance. These values are so small and indicates that the model don't have significant explainatory value. It is best not to use this model for prediction purposes.

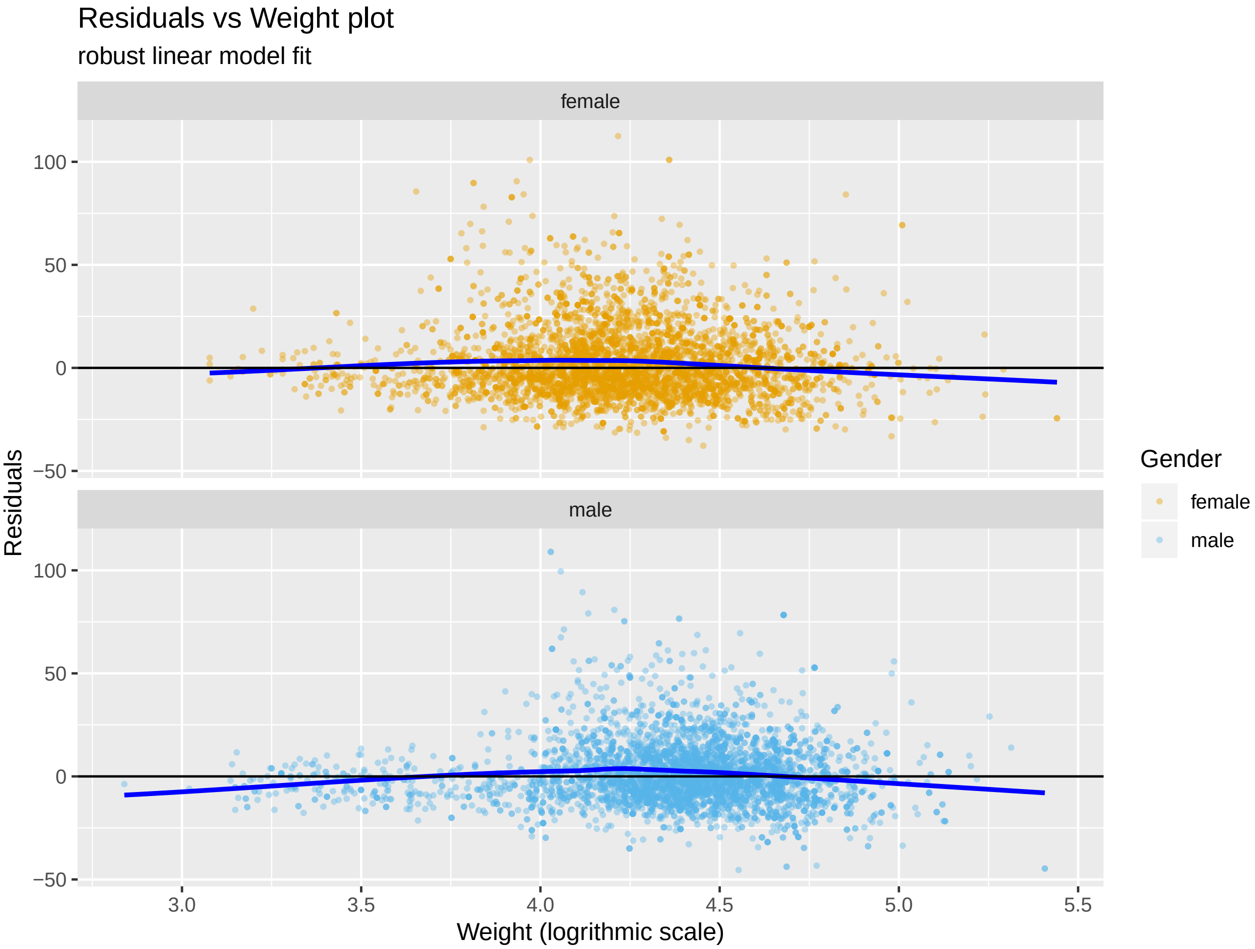
The residual plot for both male and female shows some deviation along $y = 0$ line, especially where data points are sparse and also the "loess" fit line is pulled upwards by the outliers of the range in between the height of 150 cm and 75 cm, but that sould be okay as, we have fitted robust model in this data.

section 3: systolic blood pressure with Weight

```
BP.weight = na.omit(nhanes[,c("BPSysAve", "Weight", "Gender")])
BP.weight.male = BP.weight[BP.weight$Gender == "male",]
BP.weight.female = BP.weight[BP.weight$Gender == "female",]

# model fitting
BP.weight.rlm = rlm(BPSysAve ~  log(Weight) + Gender, data = BP.weight, psi = psi.bisquare)
BP.weight.rlm.df = augment(BP.weight.rlm)

# model plot
gg = ggplot(BP.weight.rlm.df, aes(x = log.Weight., y = BPSysAve, color = Gender)) +
  geom_point(size = 0.7, alpha = 0.4)+
  geom_smooth(method = "rlm", se = F, method.args = list(psi = psi.bisquare), color = "black") +
  facet_wrap(~Gender, ncol = 1) +
  scale_color_manual(values = cb_palette) +
  xlab("Weight (logrithmic scale)") + ylab("Systolic blood pressure") +
  labs(title = "Systolic BP vs Weight",
       subtitle = "robust linear model fit")
gg
```



```
# Residual plot
gg = ggplot(BP.weight.rlm.df, aes(x =  log.Weight., y = .resid, color = Gender)) +
  geom_point(size = 0.7, alpha= 0.4) +
  geom_smooth(method = "loess", se = F, method.args = list(degree = 1), col = "blue") +
  geom_abline(slope = 0)+
  facet_wrap(~Gender, ncol = 1) +
  scale_color_manual(values = cb_palette) +
  xlab("Weight (logrithmic scale)") + ylab("Residuals") +
  labs(title = "Residuals vs Weight plot",
       subtitle = "robust linear model fit")
gg
```



The dotplot of systolic blood pressure with weight is approximately a football shaped curve, with mean of 71 kg, but the data is condensed only at the center and then it spreads all around with some large outliers, the data is right skewed and it will be better to use logrithmic transformation to better explain the model. On logrithmic scale, we can still see there are some outliers in between 4 to 4.5, which can affect the model. It is hard to capture those outliers in our model, so our model will not be able to explain those outliers. Keepign this in mind, perhaps a robust linear model will better describe our data by making our model unaffect by those outliers. In the first plot, Systolic BP vs Weight, it can be seen that the model appropriately describes that data for both male and female, as written by the quation below

$$BPSysAve_{male} = 19 * log.Weight_{male} + 35.58$$

$$BPSysAve_{female} = 16.69 * log.Weight_{female} + 43.33$$

This model expalin 9.03 of the female model and almost 12 percent variance of the male model. This model doesn't have much explanatory value, it only explains a faction of variance for both the model, so, probabily, it is not a good idea to use this model for prediction. The residual plot shows that the residuals of the model is fits across $y = 0$ line approximately with some deviation from $y = 0$ line as the data points get sparse, but this is a robust model fit, so it will not explain the outliers. The residual plot doesn't show any pattern as such that can act as dealbreaker. It is worth mentioning that the male fitted model won't be the best fit for prediction task as it shows curve in QQ plot but the female fitted model will be good to use with some cautions for prediction as the QQ plot breaks at the end points.So, one should be cautious of using these models for prediction purposes.