Problem Set 3

Raj Thakkar

26 January 2019

Write a document in three sections, giving the relationship of average systolic blood pressure with age, height, and weight respectively. (You can also include an introduction and conclusion if you really want to.) Each section should include approximately TWO graphs (a set of faceted plots counts as one graph) examining the trend and the residuals. Including many more graphs than this may be penalized. In each section, include a brief justification of your modeling choices (type of model, transformations or lack of transformations) and a verbal description of the differences you see between men and women. Some (sensibly rounded) quantitative measures will probably be useful, but you do not (and should not) list every single statistic you can think of.

Section 1: Relationship of average systolic blood pressure with age

We will load the desired libraries and make the desired transformations on the variables

```
# Loading the desired libraries
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.5.2

## -- Attaching packages ---------------------------------------------------------------------------

## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.6
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0

## Warning: package 'ggplot2' was built under R version 3.5.2

## Warning: package 'tidyr' was built under R version 3.5.2

## -- Conflicts ---------------------------------------------------------------------------------- tid
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(ggplot2)
library(NHANES)
```

```
## Warning: package 'NHANES' was built under R version 3.5.2
```

```
# Removing the NAs from BPSysAve and Age variables
nhanes_data_age <- NHANES %>% filter(!is.na(BPSysAve) & !is.na(Age)) %>% dplyr::select(ID,
    Age, Gender, BPSysAve)

# Creating separate dataframes for males and females
nhanes_data_age_male <- nhanes_data_age %>% filter(Gender == "male")
nhanes_data_age_female <- nhanes_data_age %>% filter(Gender == "female")
```

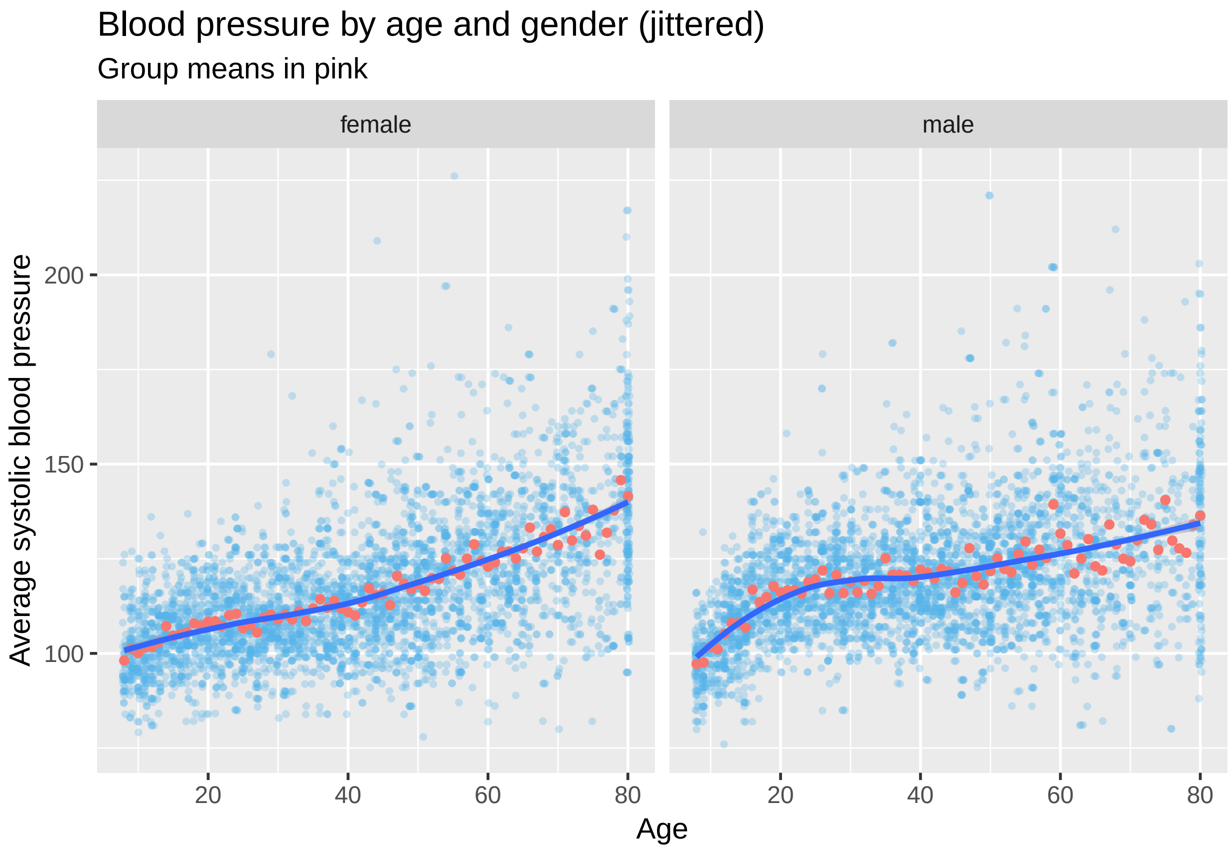
First we will plot the scatter plot of average blood pressure v/s age for men and women separately and fit a loess model on the points.

```
# Using geom_jitter as there are only a few values that the age variable
# takes and thus many observations are on top of each other
```

```
gg <- ggplot(nhanes_data_age, aes(x = Age, y = BPSysAve)) + geom_jitter(alpha = 0.3,
    height = 0.125, width = 0.25, size = 0.7, col = "#56B4E9") + facet_wrap(~Gender,
    ncol = 2) + ylab("Average systolic blood pressure") + ggtitle("Blood pressure by age and gender (ji

age.means <- aggregate(BPSysAve ~ Age + Gender, mean, data = nhanes_data_age)

gg + geom_point(data = age.means, aes(color = "#CC0000", shape = ".")) + theme(legend.position = "none"
    geom_smooth(method = "loess") + labs(subtitle = "Group means in pink")
```



We observe that Average systolic blood pressure increases monotonically with Age of the participants in the research. The loess model captures the mean of the Average Systolic Pressure variable (represented by pink dots above) well but misses many of the outlier observations as observed in the graph above. “loess” model with degree 1 as tried to but the fit is better especially for males when degree = 2 (default value) is used.

The trend of Average systolic pressure with age is increases at a a steady for males but it increases quickly in the beginning for females and then the rate of increase is smaller. Average systolic pressure increases monotonically with age for both males and females.

Now plotting the residuals we get based on Loess model

```
bp.lm_male <- loess(formula = BPSysAve ~ Age, data = nhanes_data_age_male)
# summary(bp.lm)
library(broom)
```

```
## Warning: package 'broom' was built under R version 3.5.2
```

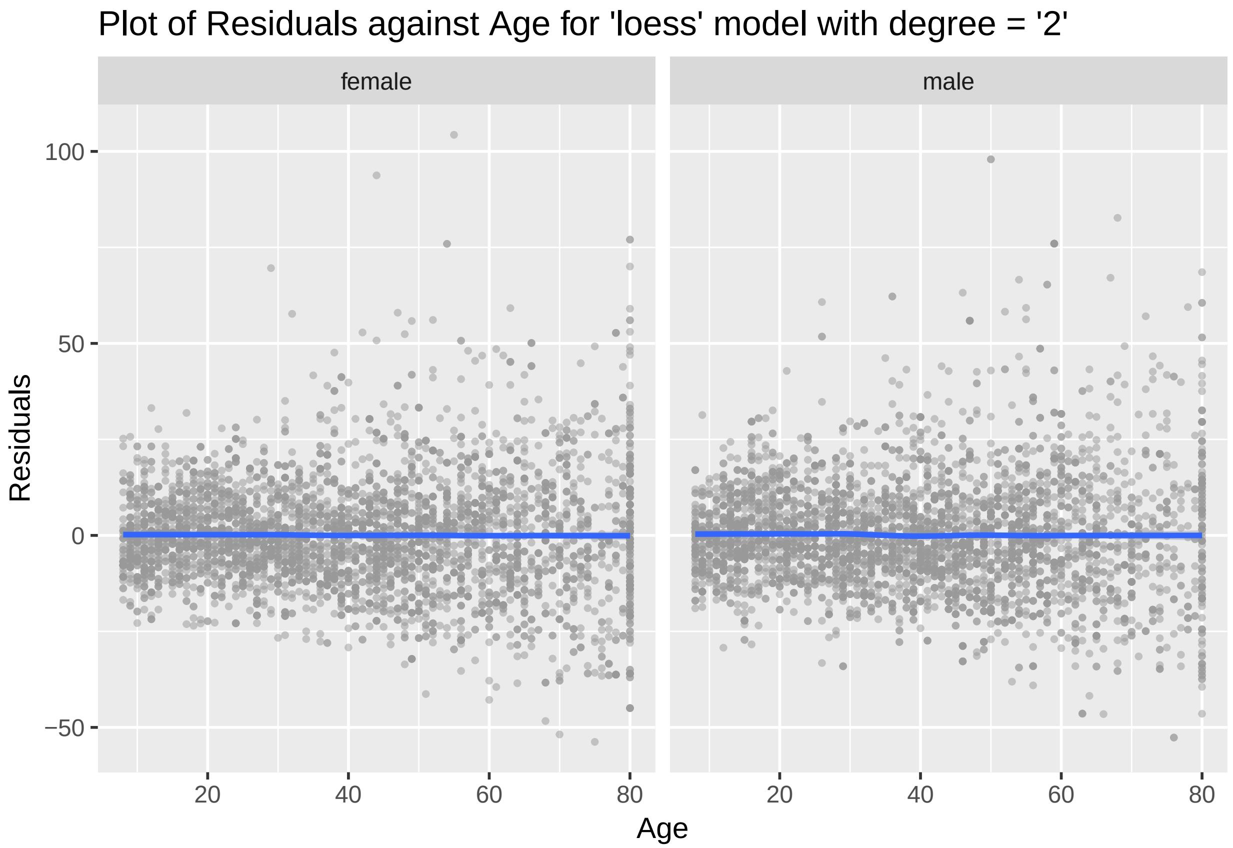
```
bp.lm.df_male <- augment(bp.lm_male)
bp.lm.df_male$Gender <- "male"
```

```
# bp.lm.df

bp.lm_female <- loess(formula = BPSysAve ~ Age, data = nhanes_data_age_female)
# summary(bp.lm) library(broom)
bp.lm.df_female <- augment(bp.lm_female)
bp.lm.df_female$Gender <- "female"

bp.lm_df_overall <- rbind(bp.lm.df_male, bp.lm.df_female)

ggplot(bp.lm_df_overall, aes(x = Age, y = .resid)) + geom_point(alpha = 0.5,
    col = "#999999", size = 0.7) + geom_smooth(method = "loess", method.args = list(degree = 1)) +
    facet_wrap(~Gender, ncol = 2) + ylab("Residuals") + ggtitle("Plot of Residuals against Age for 'loes
```



We observe that there is no trend in the residuals and the fitted line is parallel coincides with the line x = 0 for both males and females when loess model is used. Moreover, the residuals oscillate around 0. Thus we can say that Age can be used to predict Systolic Blood Pressure of individuals who participated in the study.

Linear model was tried too but the plot of residuals is better when we use the loess model. Log transformation on the Age variable was tried but it didn't help a lot in getting a better residual plot when linear model was fitted on the transformed data so finally loess model without any transformation of variables was used.

Section 2: Relationship of average systolic blood pressure with height

First we will plot the scatter plot of average blood pressure v/s height for men and women separately

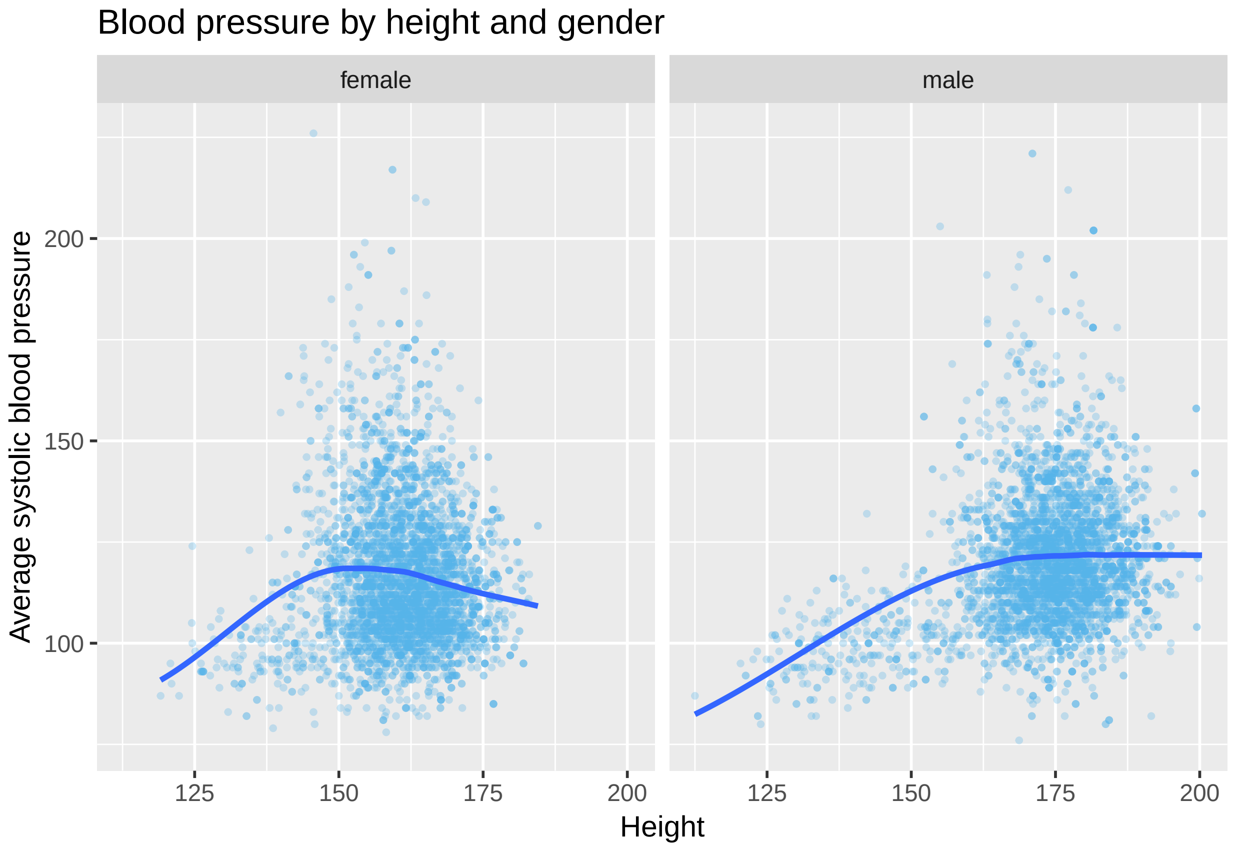
```
# Removing the NAs from BPSysAve variable
nhanes_data_height <- NHANES %>% filter(!is.na(BPSysAve) & !is.na(Height)) %>% dplyr::select(ID, Height
```

```r
# Creating separate dataframes for males and females
nhanes_data_height_male <- nhanes_data_height %>% filter(Gender == "male")
nhanes_data_height_female <- nhanes_data_height %>% filter(Gender == "female")


gg <- ggplot(nhanes_data_height,aes(x = Height, y = BPSysAve)) +
      geom_point(col = "#56B4E9", alpha = 0.3, size = 0.7) +
      facet_wrap(~Gender, ncol = 2) +
      ylab("Average systolic blood pressure") +
      ggtitle("Blood pressure by height and gender")

# height.means <- aggregate(BPSysAve~Height + Gender,mean,data = nhanes_data_height)

gg+# geom_point(data = height.means, aes(color = "pink", shape = ".")) +
  theme(legend.position = 'none') +
  geom_smooth(method = "loess", method.args = list(degree = 1), se = FALSE)
```



We observe that the Average systolic blood pressure doesn't increase monotonically with increase in Height. The loess model with degree = 1 is the best fit. loess model with degree = 2 was tried but the strong fit by degree =2 was not justified by the available data and hence degree = 1 is used above. Various transformations such as log, square root, cube root were tried on the Height variable but the transformations didn't result in a better fit. Thus not transformations have been made.

It is clear from the graph above that Average systolic pressure changes in a different manner for males and females with increase in Height. For females, the systolic pressure first increases and then it starts decreasing while for males, the systolic pressure increases almost linearly with height but then it remains constant. For females, the data is concentrated between 150 and 175 cm while for males data is concentrated between 160 and 180 cm. This can be attributed to the fact that males on an average are taller than females.

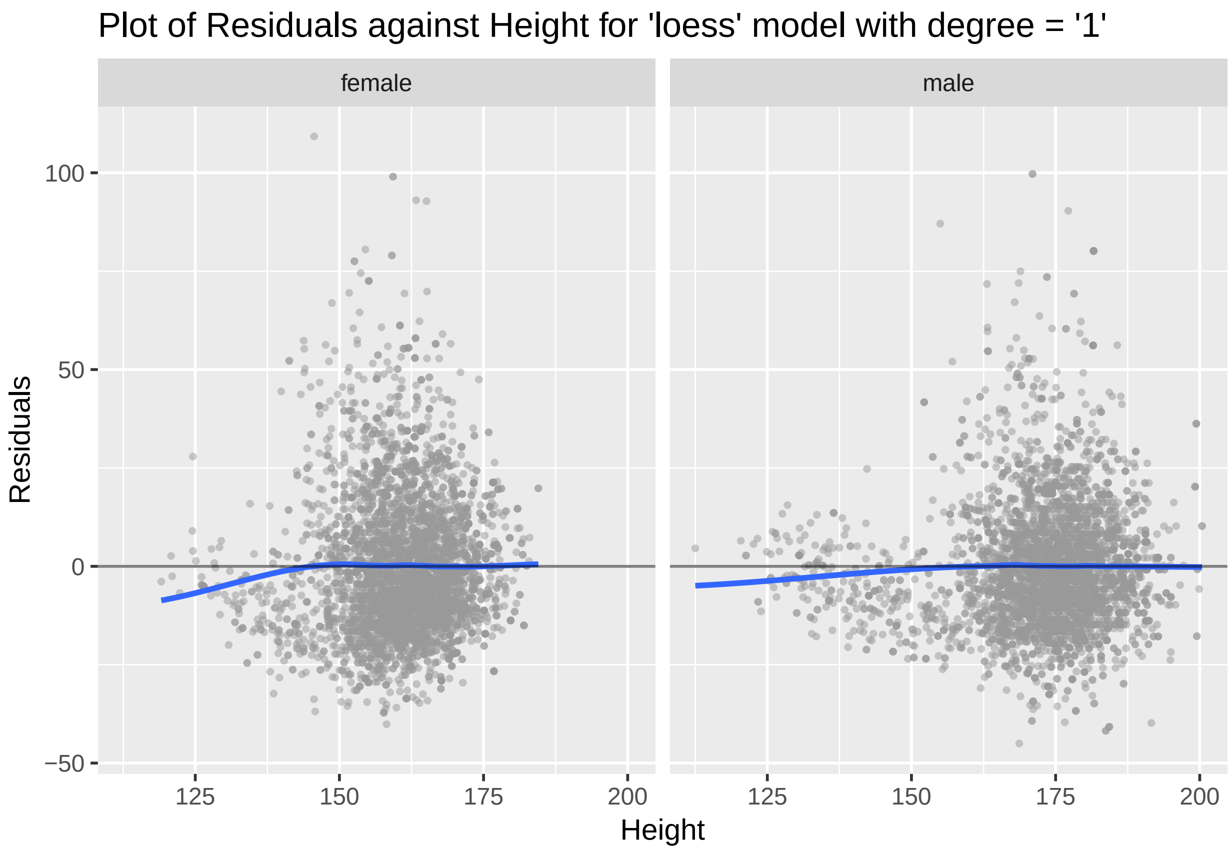
Now plotting the residuals we get based on Loess model with degree = 1

```
bp.lm_male <- loess(formula = BPSysAve ~ Height, data = nhanes_data_height_male,
    degree = 1)
# summary(bp.lm)
library(broom)
bp.lm.df_male <- augment(bp.lm_male)
bp.lm.df_male$Gender <- "male"
# bp.lm.df

bp.lm_female <- loess(formula = BPSysAve ~ Height, data = nhanes_data_height_female,
    degree = 1)
# summary(bp.lm) library(broom)
bp.lm.df_female <- augment(bp.lm_female)
bp.lm.df_female$Gender <- "female"

bp.lm_df_overall <- rbind(bp.lm.df_male, bp.lm.df_female)

ggplot(bp.lm_df_overall, aes(x = Height, y = .resid)) + geom_point(alpha = 0.5,
    col = "#999999", size = 0.7) + geom_smooth(method = "loess", method.args = list(degree = 1),
    se = F) + facet_wrap(~Gender, ncol = 2) + geom_abline(slope = 0, intercept = 0,
    alpha = 0.5) + ylab("Residuals") + ggtitle("Plot of Residuals against Height for 'loess' model with
```



```
cor(nhanes_data_height_male$Height, nhanes_data_height_male$BPSysAve)
```

```
## [1] 0.265769
```

```
cor(nhanes_data_height_female$Height, nhanes_data_height_female$BPSysAve)
```

```
## [1] 0.009856986
```

We observe that there is a trend in the residual and it increases slightly in the beginning with increase in height and then fitted line overlaps the line x = 0 (x- axis). The residuals don't oscillate around 0 with respect to height .This trend can be attributed to the fact that we cannot fit a good model for the data available as the correlation between average systolic pressure and height of the participants is very less. The correlation between average systolic pressure and height is 0.266 approximately for males while it is 0.00986 approximately for females. From the correlation values, we can say that Height is not a good predictor of Average Systolic pressure for the participants.

Section 3: Relationship of average systolic blood pressure with weight

First we will plot the scatter plot of average blood pressure v/s weight for men and women separately

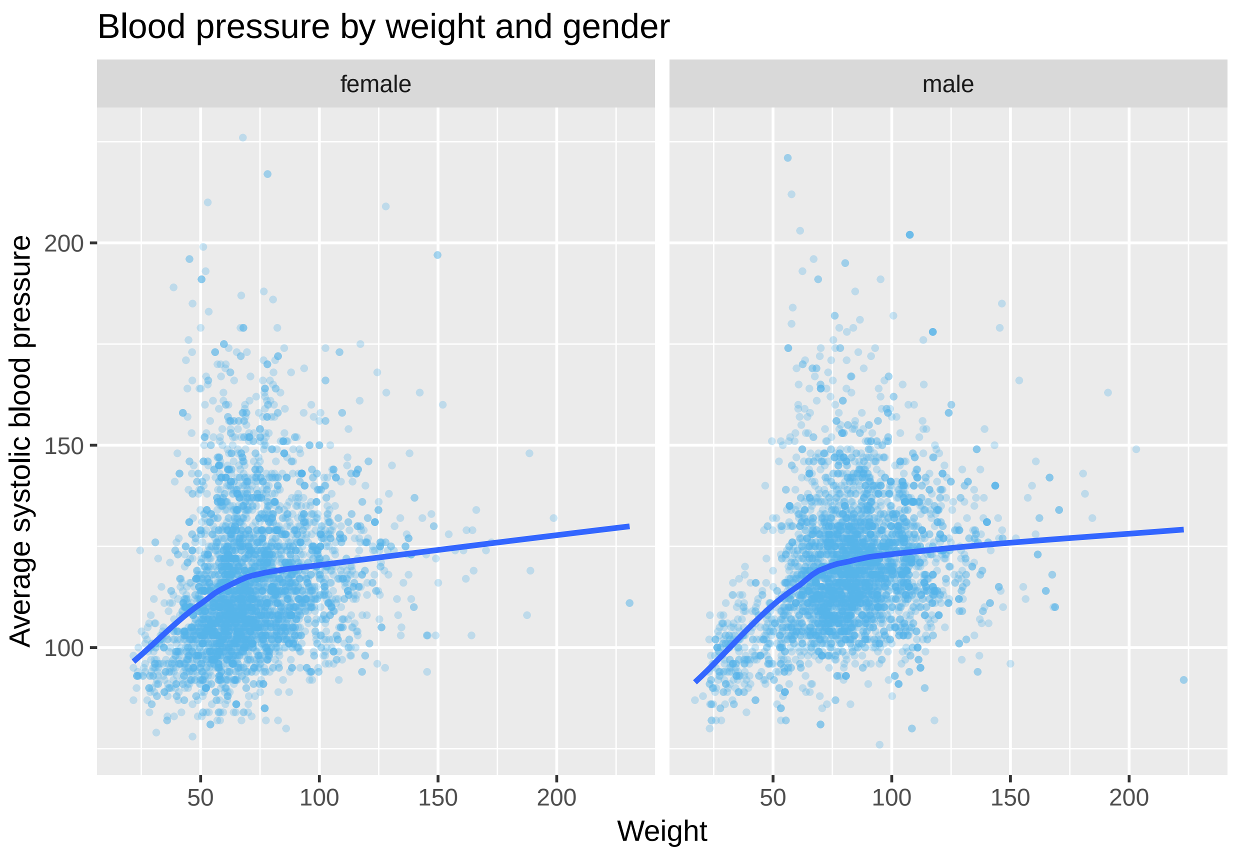
```
# Removing the NAs from BPSysAve variable
nhanes_data_weight <- NHANES %>% filter(!is.na(BPSysAve) & !is.na(Weight)) %>%
    dplyr::select(ID, Weight, Gender, BPSysAve)

# Creating separate dataframes for males and females
nhanes_data_weight_male <- nhanes_data_weight %>% filter(Gender == "male")
nhanes_data_weight_female <- nhanes_data_weight %>% filter(Gender == "female")

gg <- ggplot(nhanes_data_weight, aes(x = Weight, y = BPSysAve)) + geom_point(col = "#56B4E9",
    alpha = 0.3, size = 0.7) + facet_wrap(~Gender, ncol = 2) + ylab("Average systolic blood pressure") +
    ggtitle("Blood pressure by weight and gender")

# age.means <- aggregate(BPSysAve~Age + Gender,mean,data = nhanes_data_age)

gg + theme(legend.position = "none") + geom_smooth(method = "loess", method.args = list(degree = 1),
    se = FALSE)
```



We observe that the Average systolic blood pressure increases monotonically with increase in Weight. The loess model with degree = 1 is the best fit. loess model with degree = 2 was tried but the strong fit by degree = 2 was not justified by the available data and hence degree = 1 is used above. Various transformations such as log, square root, cube root were tried on the Weight variable but the transformations didn't result in a better fit. Thus no transformations have been made.

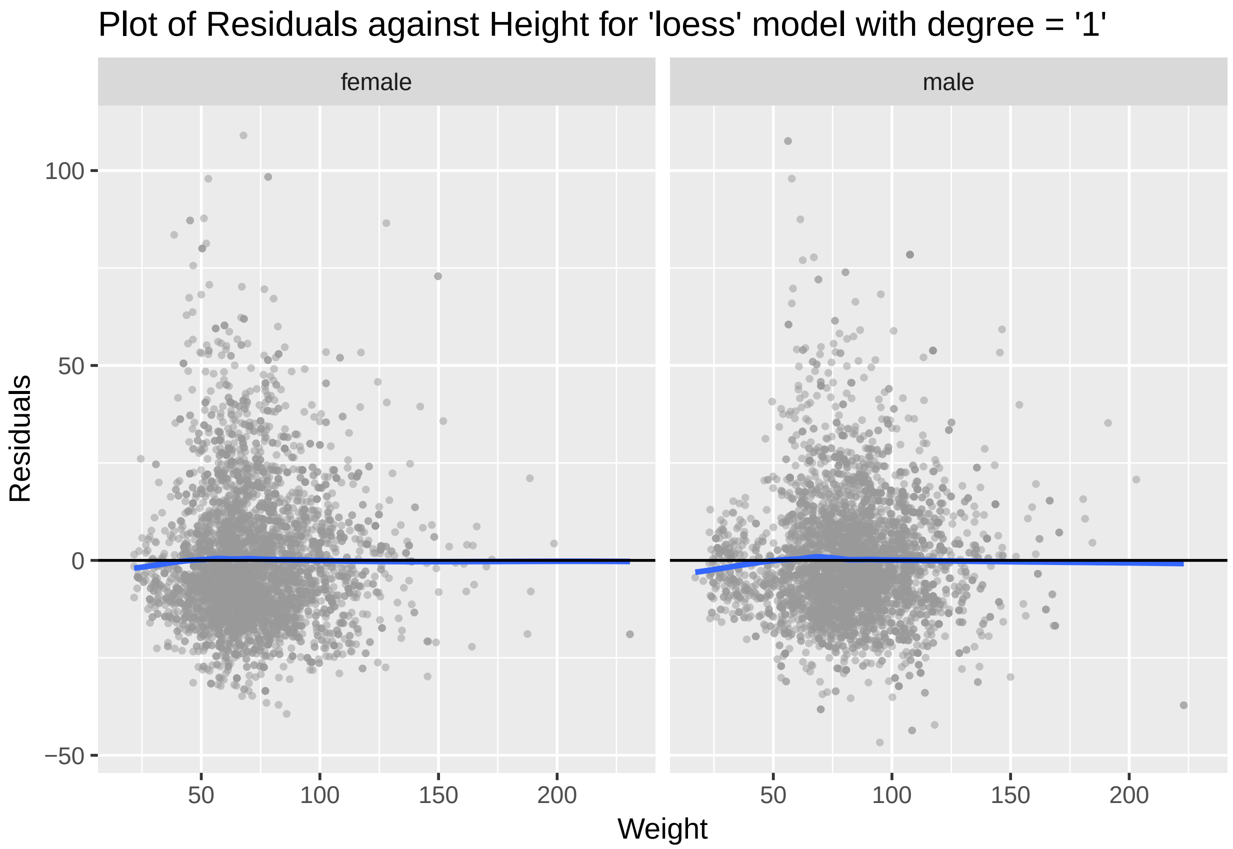
It is clear from the graph above that Average systolic pressure changes in the same way with weight for both males and females. Thus, trend observed by fitting a "loess" model of degree = 1 is almost the same for both males and females.

```r
bp.lm_male <- loess(formula = BPSysAve ~ Weight, data = nhanes_data_weight_male,
    degree = 1)
# summary(bp.lm)
library(broom)
bp.lm.df_male <- augment(bp.lm_male)
bp.lm.df_male$Gender <- "male"
# bp.lm.df

bp.lm_female <- loess(formula = BPSysAve ~ Weight, data = nhanes_data_weight_female,
    degree = 1)
# summary(bp.lm) library(broom)
bp.lm.df_female <- augment(bp.lm_female)
bp.lm.df_female$Gender <- "female"

bp.lm_df_overall <- rbind(bp.lm.df_male, bp.lm.df_female)

ggplot(bp.lm_df_overall, aes(x = Weight, y = .resid)) + geom_point(alpha = 0.5,
    col = "#999999", size = 0.7) + geom_smooth(method = "loess", method.args = list(degree = 1),
    se = F) + facet_wrap(~Gender, ncol = 2) + geom_abline(slope = 0, intercept = 0) +
    ylab("Residuals") + ggtitle("Plot of Residuals against Height for 'loess' model with degree = '1'")
```



```
cor(nhanes_data_weight_male$Weight, nhanes_data_weight_male$BPSysAve)
```

```
## [1] 0.29313
```

```
cor(nhanes_data_weight_female$Weight, nhanes_data_weight_female$BPSysAve)
```

```
## [1] 0.2284306
```

We observe that there is a trend in the residuals and the residuals increase slightly in the beginning with increase in weight and then the fitted line overlaps the line x = 0 (x-axis). But the residuals don't oscillate around 0 randomly with respect to weight (The trend is less obvious as compared to the trend for height) This trend can be attributed to the fact that we cannot fit a good model for the data available as the correlation between average systolic pressure and weight of the participants is very less. The correlation between average systolic pressure and weight is 0.293 approximately for males while it is 0.228 approximately for females. From the correlation values, we can say that Weight is not a good predictor of Average Systolic pressure for the participants.