Assignment-3

Saniya Ambavanekar, Akshay Naik, & Dheeraj Singh

January-31-2018

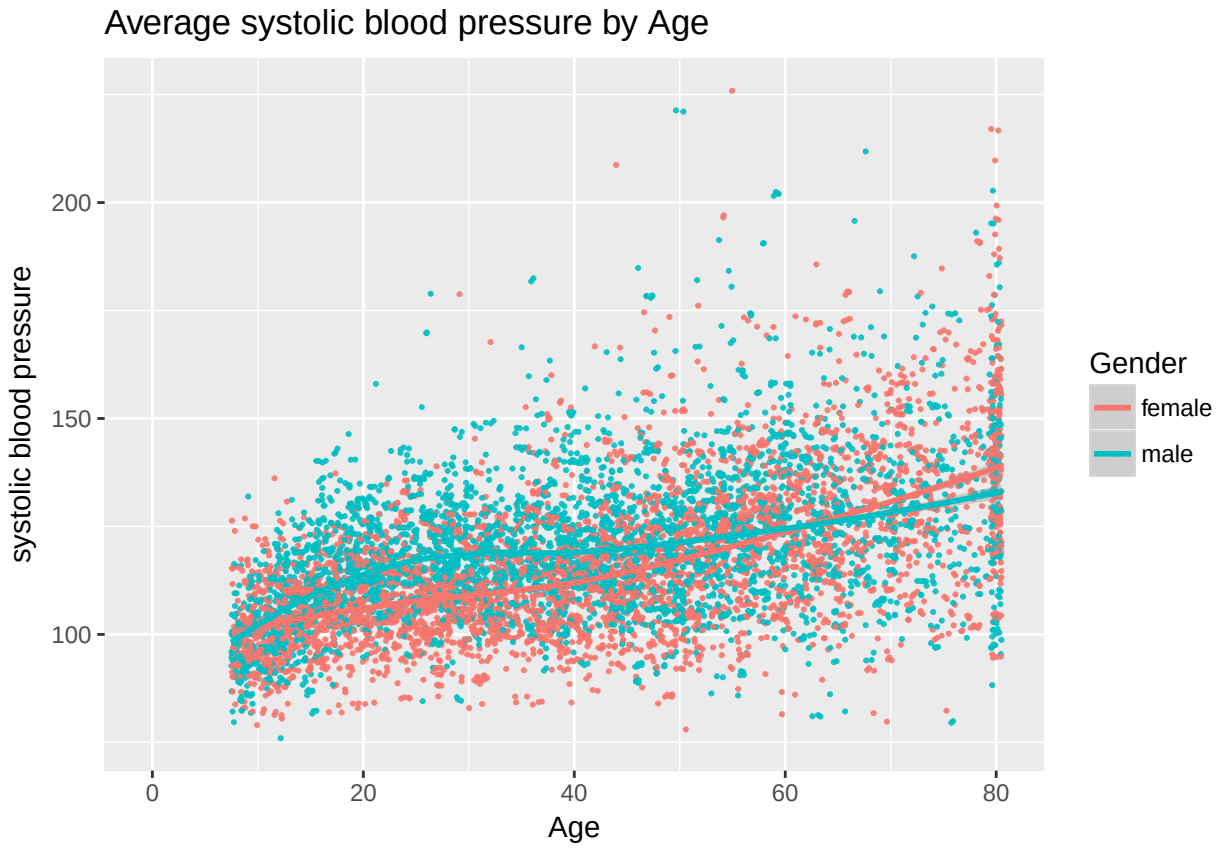
```
library('NHANES')
library('ggplot2')
library('broom')
```

```
data <- NHANES
n <- nrow(data)
male <- subset(data, Gender=='male')
female <- subset(data, Gender=='female')
```

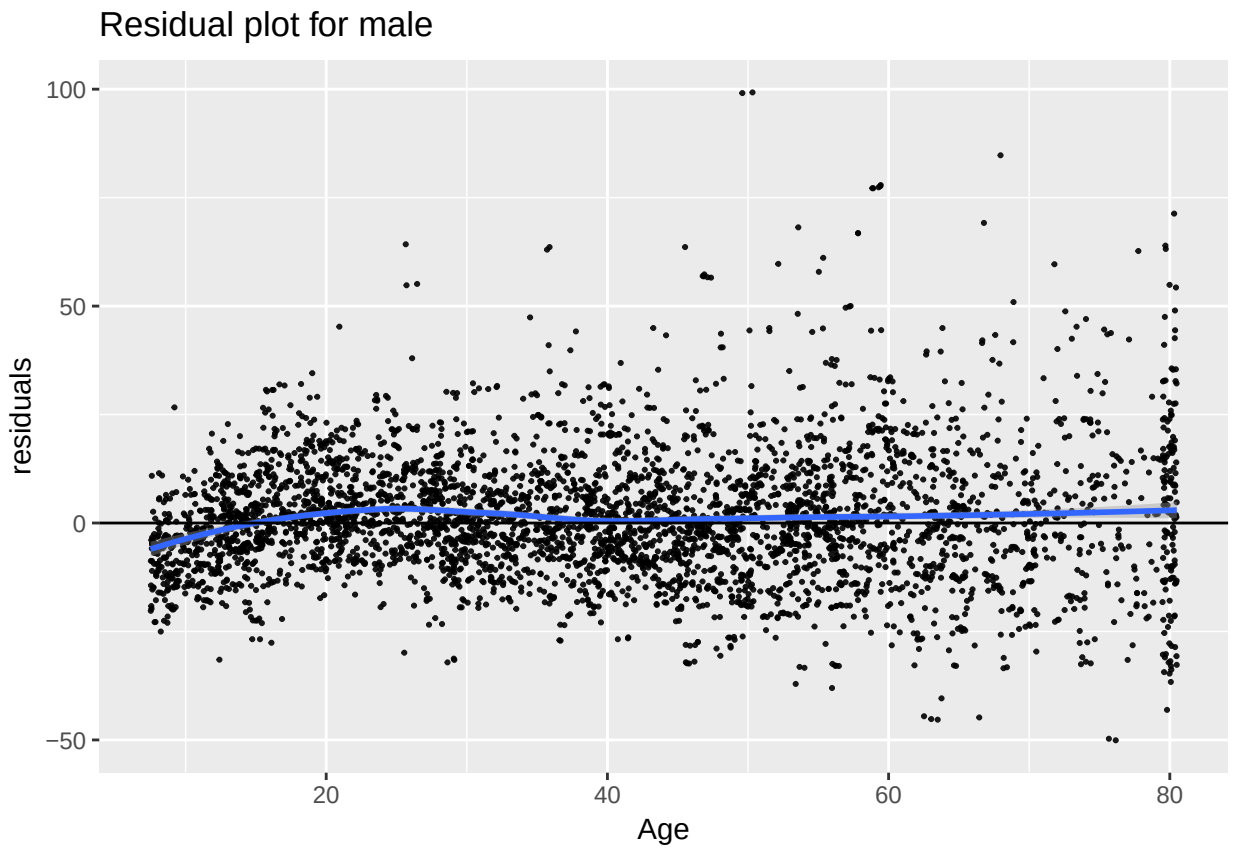
Section 01

Relationship of average systolic blood pressure with age

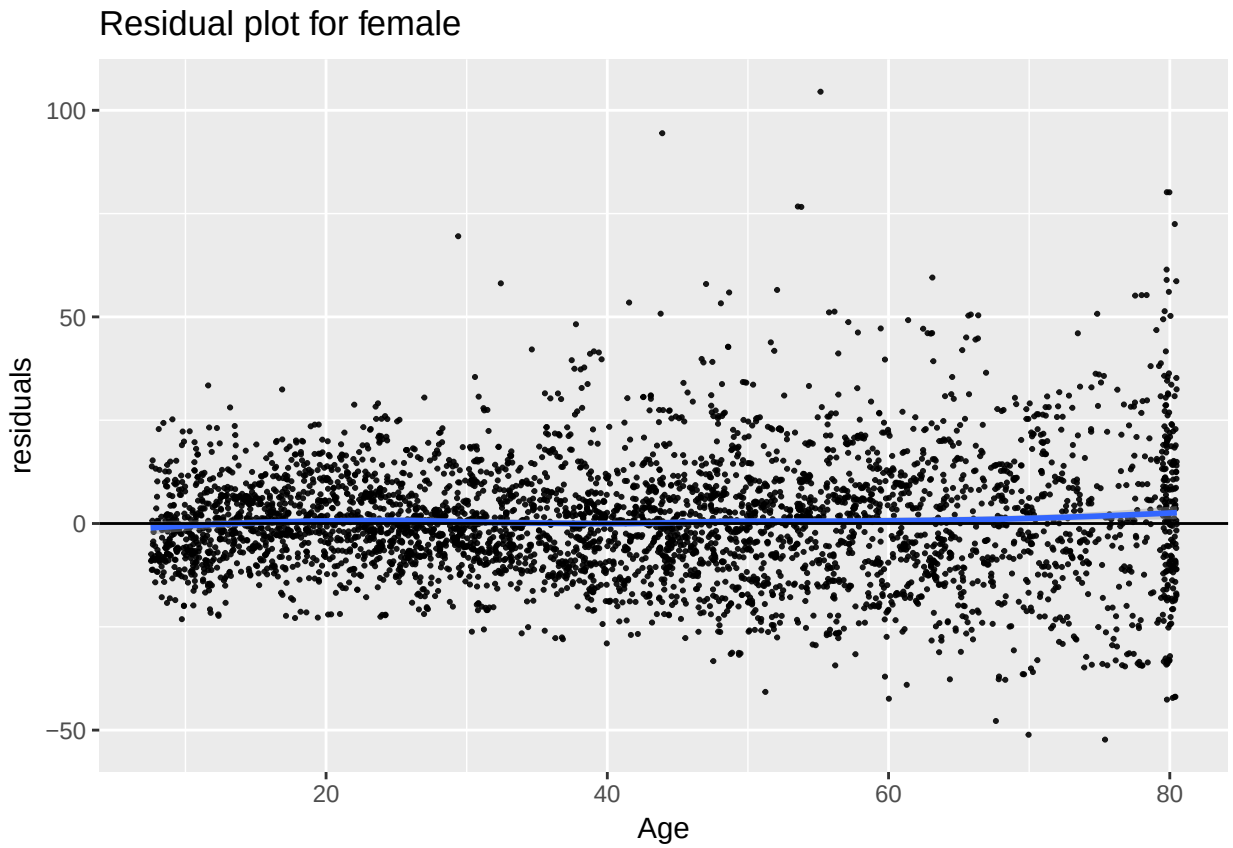
```
ggplot(data, aes(x=Age+runif(n, -0.5, 0.5), y=BPSysAve+runif(n, -0.5, 0.5), color=Gender)) +
  geom_point(size=0.4, alpha=0.9) +
  geom_smooth(method="loess", method.args=list(degree=2, family="symmetric")) +
  labs(x='Age', y='systolic blood pressure',
       title='Average systolic blood pressure by Age')
```



```
male_age.lo <- loess(male$BPSysAve ~ male$Age, degree=1, family='symmetric')
male_age.lo.df <- augment(male_age.lo)
n_male <- nrow(male_age.lo.df)
ggplot(male_age.lo.df, aes(x=male.Age+runif(n_male, -0.5, 0.5),
                           y=.resid+runif(n_male, -0.5, 0.5)))+
  geom_point(size=0.4,alpha=0.9) + geom_smooth(method='loess',
  method.args=list(degree=2)) + geom_abline(slope=0) +
  labs(x="Age", y="residuals", title='Residual plot for male')
```




```
female_age.lo <- loess(female$BPSysAve ~ female$Age, degree=1, family='symmetric')
female_age.lo.df <- augment(female_age.lo)
n_female <- nrow(female_age.lo.df)
ggplot(female_age.lo.df, aes(x=female.Age+runif(n_female, -0.5, 0.5),
                             y=.resid+runif(n_female, -0.5, 0.5))) +
      geom_point(size=0.4,alpha=0.9) + geom_smooth(method='loess',
      method.args=list(degree=2)) + geom_abline(slope=0) +
      labs(x="Age", y="residuals", title='Residual plot for female')
```




```
var(male_age.lo.df$.fitted)/var(male_age.lo.df$male.BPSysAve)
```

```
## [1] 0.1696802
```

```
var(female_age.lo.df$.fitted)/var(female_age.lo.df$female.BPSysAve)
```

```
## [1] 0.3273877
```

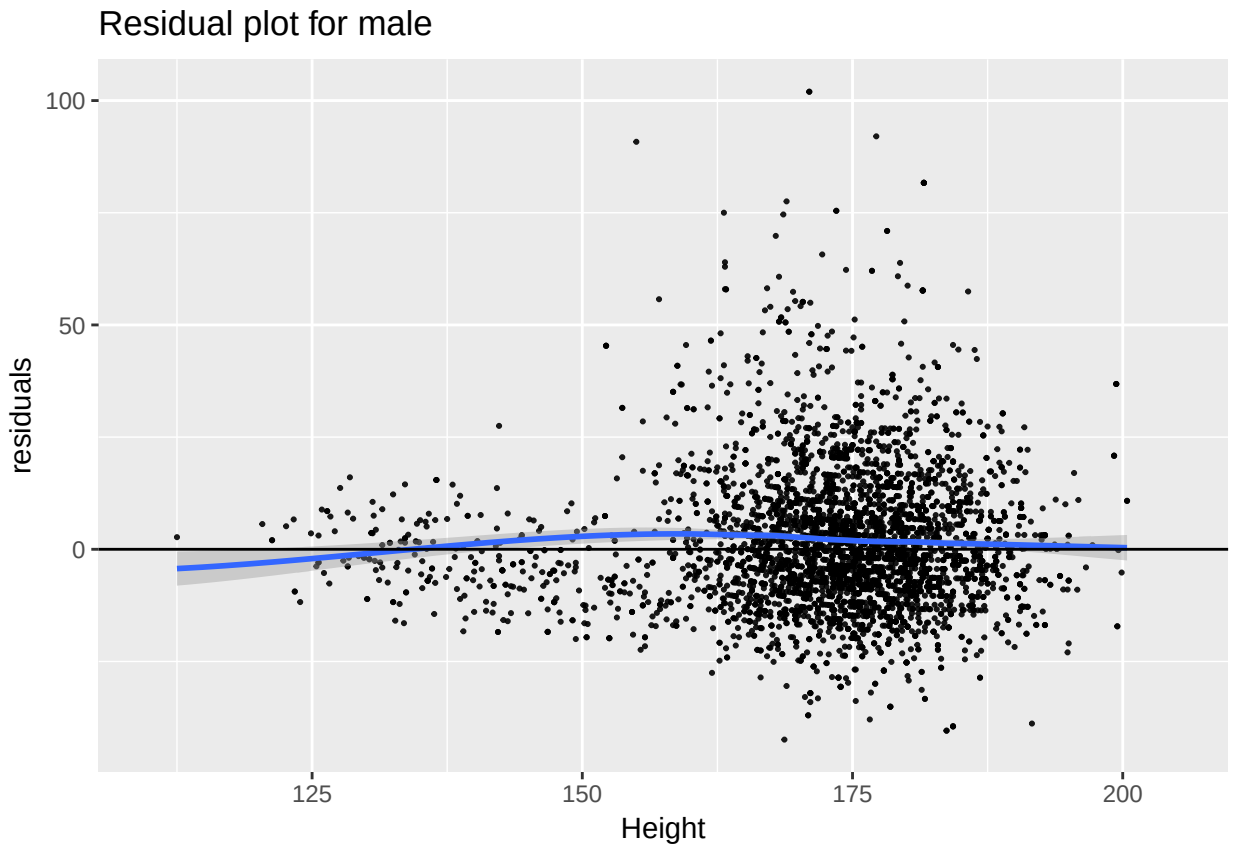
Since most of the data points were stacked on each other (this might be because of rounding of both age as well as average blood pressure), we jittered both using uniform random. As linear function was not effectively capturing the overall trend, used Loess to fit the distribution. Because there are few outliers in the distribution, used symmetric family to counter the effect. Comparing trend lines for male and female, we can say that the average systolic blood pressure is higher for male till the age of 60. Post this females have relatively higher average systolic blood pressure. Residuals plots also show that there is no systematic trend in residuals. Looking at R^2, age has more explanatory value for female (~33%) in comparison to male (~17%).

Section 02

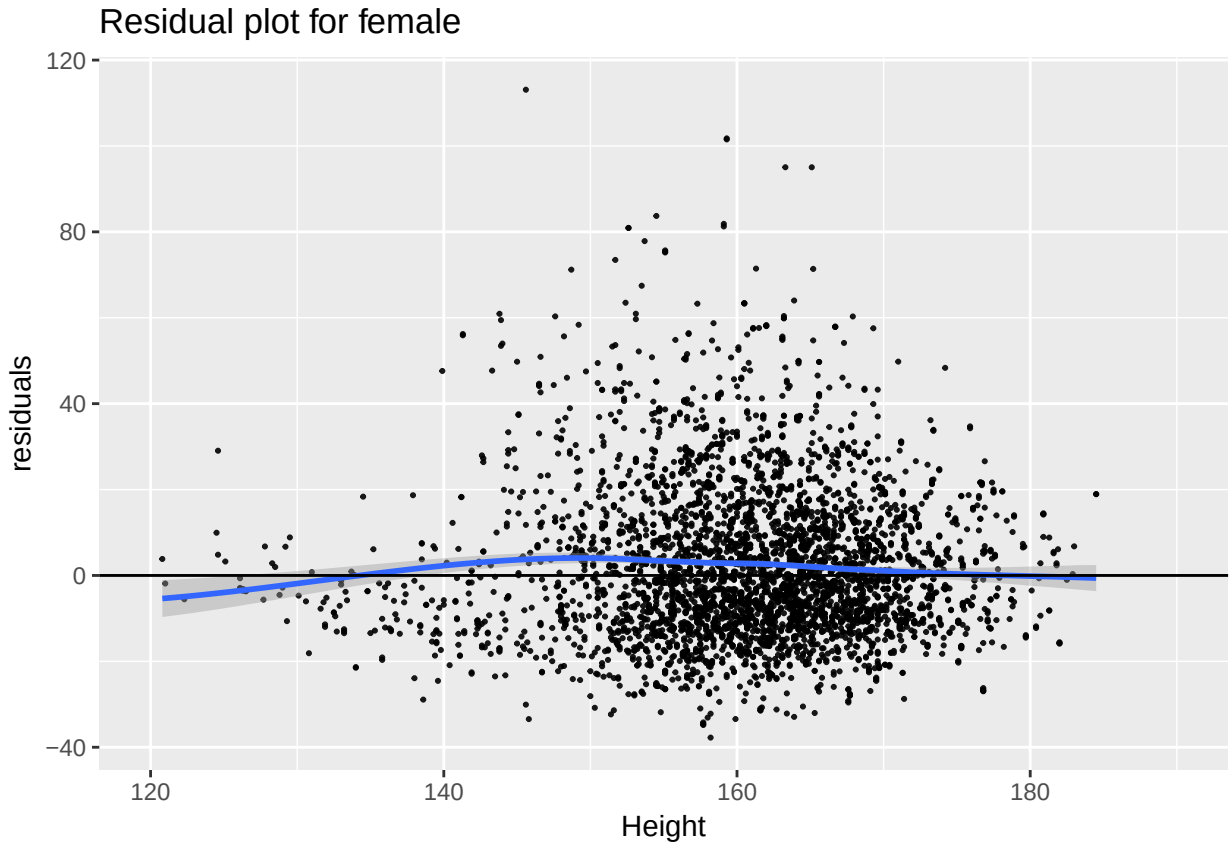
Relationship of average systolic blood pressure with Height

```
ggplot(data, aes(x=Height, y=BPSysAve+runif(n, -0.5, 0.5), color=Gender)) +
  geom_point(size=0.4, alpha=0.9) +
  geom_smooth(method="loess", method.args=list(degree=1, family="symmetric")) + xlim(110, 205) +
  labs(x='Height', y='systolic blood pressure',
       title='Average systolic blood pressure by Height')
```


```
male_height.lo <- loess(male$BPSysAve ~ male$Height, degree=1, family='symmetric')
male_height.lo.df <- augment(male_height.lo)
n_male <- nrow(male_height.lo.df)
ggplot(male_height.lo.df, aes(x=male.Height, y=.resid))+ xlim(110, 205) +
  geom_point(size=0.4,alpha=0.9) + geom_smooth(method='loess',
  method.args=list(degree=1)) + geom_abline(slope=0) +
  labs(x="Height", y="residuals", title='Residual plot for male')
```




```
female_height.lo <- loess(female$BPSysAve ~ female$Height, degree=1, family='symmetric')
female_height.lo.df <- augment(female_height.lo)
n_female <- nrow(female_height.lo.df)
ggplot(female_height.lo.df, aes(x=female.Height,
        y=.resid+runif(n_female, -0.5, 0.5))) + xlim(120, 190) +
      geom_point(size=0.4,alpha=0.9) + geom_smooth(method='loess',
      method.args=list(degree=1)) + geom_abline(slope=0) +
      labs(x="Height", y="residuals", title='Residual plot for female')
```




```
var(male_height.lo.df$.fitted)/var(male_height.lo.df$male.BPSysAve)
```

```
## [1] 0.07443659
```

```
var(female_height.lo.df$.fitted)/var(female_height.lo.df$female.BPSysAve)
```

```
## [1] 0.01572679
```

From the scatter plot, we can see that most the data points are cluseterd in a local region, so it would be difficult for a line to capture overall trend. Fitting loess trend function with degree 1, we observe that till about height of 160 cm, female have higher average blood pressure and for heights greater than 160 cm, male have higher blood pressure. Residual plots are slightly bent downwards for heights below 125 cm but that can be ignored as those are very less data points. Residual plots are slightly above the x-axis because of presence outliers having relatively higher values of average blood pressure. From R^2, we see that height is not a very good explanatory variable for average blood pressure for both male and female. Though it is relatively higher for male (~7%) than female (~2%).

Section 03

Relationship of average systolic blood pressure with weight

```
ggplot(data, aes(x=Weight, y=BPSysAve+runif(n, -0.5, 0.5), color=Gender)) +
  geom_point(size=0.4, alpha=0.9) +
  geom_smooth(method="loess", method.args=list(degree=1, family="symmetric")) +
  labs(x='Weight', y='systolic blood pressure',
       title='Average systolic blood pressure by Weight')
```


```
male_weight.lo <- loess(male$BPSysAve ~ male$Weight, degree=1, family='symmetric')
male_weight.lo.df <- augment(male_weight.lo)
n_male <- nrow(male_weight.lo.df)
ggplot(male_weight.lo.df, aes(x=male.Weight, y=.resid))+
  geom_point(size=0.4,alpha=0.9) + geom_smooth(method='loess',
  method.args=list(degree=1)) + geom_abline(slope=0) +
  labs(x="Weight", y="residuals", title='Residual plot for male')
```


```
female_weight.lo <- loess(female$BPSysAve ~ female$Weight, degree=1, family='symmetric')
female_weight.lo.df <- augment(female_weight.lo)
n_female <- nrow(female_weight.lo.df)
ggplot(female_weight.lo.df, aes(x=female.Weight, y=.resid)) +
      geom_point(size=0.4,alpha=0.9) + geom_smooth(method='loess',
      method.args=list(degree=1)) + geom_abline(slope=0) +
      labs(x="Weight", y="residuals", title='Residual plot for female')
```


```
var(male_weight.lo.df$.fitted)/var(male_weight.lo.df$male.BPSysAve)
```

```
## [1] 0.1277856
```

```
var(female_weight.lo.df$.fitted)/var(female_weight.lo.df$female.BPSysAve)
```

```
## [1] 0.08008468
```

From the loess trend plots, we can observe that on an average blood pressure is almost similar for male and female upto the weight of 75 kg, after that average blood pressure of male is relatively higher than female and the difference keeps on increasing slightly on increasing weight further. Residual plots are slightly above the x-axis due to presence of few data points with higher blood pressure value and bends downwards in the end but since data points are very less at that end, it can be ignored. R^2 suggests that weight is able to explain almost similar percentage of variance in the blood pressure for both male (~8%) and female (~5%).