Problem Set 4 for Methods in Biostatistics 3 [140.653]

Zixuan Yu

Objective

Pulse pressure is the difference between diastolic blood pressure (BP) and systolic blood pressure, and its normal range is 40-60mmHg. Wide pulse pressure can suggest underlying health problems like hypertension and heart diseases (Tang et al., 2020). Our primary objective is to study how the mean pulse pressure of U.S. adults (20 to 80 years old) will differ by age, sex and race. The secondary objective is to study whether height, weight, and body mass index (BMI) are mediators between the relationship of pulse pressure and age, after adjusting for sex and race.

Data

We analyze a subset of data from the National Health and Nutrition Examination Survey. The outcome variable is pulse pressure (calculated by diastolic BP minus systolic BP). Key exposure variables are age (in years), female (indicator variable, 1 for female, 2 for male), race (indicator variable, 1 for Hispanic, 2 for non-Hispanic white, 3 for non-Hispanic black, and 4 for all other race). Height, weight, and BMI are potential mediating variables.

Methods

Data preparation: We remove incomplete records and records with unrealistic values: any records with systolic BP out of the range 80–250 mmHg or with diastolic BP out of the range 30–120 mm are removed.

Descriptive data analysis: We use scatter plots to display the trend between pulse pressure and age, and then plot the same trend but stratified by gender and race using both scatter plots and boxplots.

Model fitting: In ***Model 1***, the pulse pressures are modeled as a non-linear function of age (via a linear spline with a knot at 45 years of age, named as age_sp45) with sex and race as covariates, assuming constant variance in pulse pressure. Then in ***Model 2***, interaction terms between sex and age, sex and age_sp45 are added. In ***Model 3***, we add interaction terms between race and age, race and age_sp45 on the basis of ***Model 2.*** We fit another 3 models using the same covariates as ***Model 1***, ***Model 2***, ***Model 3*** but allow variance in pulse pressure to vary as a function of and name them as ***Model 21***, ***Model 22***, ***Model 23***, respectively.

Model comparison and checking: We compare the AIC values of all aforementioned models and favor those with lower AIC. We check for constant variance assumption on ***Model 3*** by plotting residuals against age and check for the normality of residuals through Q-Q Plots. Likelihood ratio tests (LRT) are conducted between ***Model 21***(null model 1) and ***Model 22*** to see if the pulse pressure change rate will differ between females and males. LRT between ***Model 22*** (null model2) and ***Model 23*** is used to determine if pulse pressure change rate will differ by race. Since the distribution of pulse pressure is positively skewed, we apply the bootstrap method (sampling 11,000 times with replacement, using the percentile method) to calculate the 95% confidence interval (CI) of difference in mean pulse pressure between females and males at the age of 35, 55, 75, accordingly.

Mediation Analysis: After confirming pulse pressure will change significantly as age grows, we regress BMI, height, weight on age to see if the coefficient of age and age_sp45 has a significant p-value after controlling for sex and race. Then we add BMI, height, weight to *model23*, one variable at a time. If the coefficients of age and age_sp change significantly after adding a certain variable, the added variable is a potential mediator.

Results

Data preparation: After data cleaning, we get 4,994 records ready for further analysis out of the original 5,395.

Descriptive data analysis: The distributions of pulse pressure are right-skewed as depicted in Figure 1C. The medians are similar for Hispanics and Non-Hispanics (~ 52 mmHg) but are higher than that of 'Other' race (46 mmHg). As shown in Figure1A, pulse pressure decreases slightly as age increases before 45 years of age and then increases. The trend is consistent in sex groups as well as across race groups as depicted by Figure1B.

Model fitting, comparison and checking: The AIC of ***model 3***, ***Model 23*** is 41294, 40966, respectively. The variance of pulse pressure increases as age increases and the residuals of Model 3 is not exactly normally distributed. The two LRT suggest pulse pressure growth rate will differ by sex (p=1.19e-11) as well as differ by race after adjusting for sex (p=0.015). Based on the AIC values and results of LRT, we favor ***Model 23*** (recap: its covariates are age, age_sp45, interaction terms between age, age_sp45 and sex, race, respectively). The predicted value of ***Model 23*** is displayed in **Figure 1D**, stratified by sex. Based on the bootstrap method, the difference in mean pulse pressure between females and males is -2.92 at age 35 (95% CI: -3.81 to -2.00), 0.82 at age 55 (95% CI: -0.24 to -1.90), 3.81 at age 75 (95% CI: 1.69 to 5.97). Females have lower pulse pressure as compared to males of the same race but have higher pulse pressure after mid-age (age >45-55). The difference, though statistically significant at age 35 and 75, may not be clinically meaningful due to the small magnitude.

Model interpretation: As shown in **Table 1**, for males aged between 20-45, pulse pressure changes by -0.30 (95% CI: -0.45 to -0.15) per year of age, adjusted for race; for females the change is not significant ($p>0.05$). For people aged 45-80, pulse pressure increases significantly in both sexes (both $p<2e-16$). It increases by 0.79 per year of age for males, increase by 0.94 for females, adjusted for race. The estimated difference in pulse pressure change rate between females and males is 0.22 (p=0.0001) before 45 and is 0.15 (p=0.002) after 45. The coefficient of race is -2.33 (p=0.013), suggesting significant differences across race categories at age 20. The change rate difference is not significant across race groups before 45 (p-value = 0.12) or after (p-value = 0.27), after adjusting for sex.

Mediator Analysis: Height is not significantly correlated with age or age_sp after controlling for sex and race (p=0.053 for age, p=0.999 for age_sp), thus cannot be a mediator between the relationship of pulse pressure and age. Weight and BMI both change significantly as age changes ($p<0.05$ for both age and age_sp). The coefficient of age changes from -0.32 to -0.30 when comparing before and after adding weight as a covariate, the coefficient of age changes from -0.33 to -0.30 when comparing before and after adding BMI as a covariate. We have evidence that weight and BMI are potential mediators, but their effects are small in magnitude (<10% change in coefficients).

Conclusion: Male's pulse pressure decreases between 20 to 45 years of age while female's pulse pressure remains constant. After 45 years of age, pulse pressures increase as age increases in all sexes and all races. Females have statistically higher pulse pressure at a younger age as compared to males of the same age and race but will have a lower pulse pressure than males as they age. The difference between races is not significant, adjusting for sex and age. Weight and BMI are potential mediators with weak effects.

Discussion of limitations: Our final model contains 8 covariates, including 3 with non-significant coefficients, and may overfit the data. Race is modeled as a nominal variable, assuming the differences between two adjacent race groups are the same while holding age and sex constant. This assumption is likely violated, thus challenging the validity of the model. Whether weight is a mediator between the relationship of pulse and pressure can not be simply determined by comparing coefficients change. Methods like the Sobel test (Sobel, 1982) could be applied.

Figure and Table

Figure 1

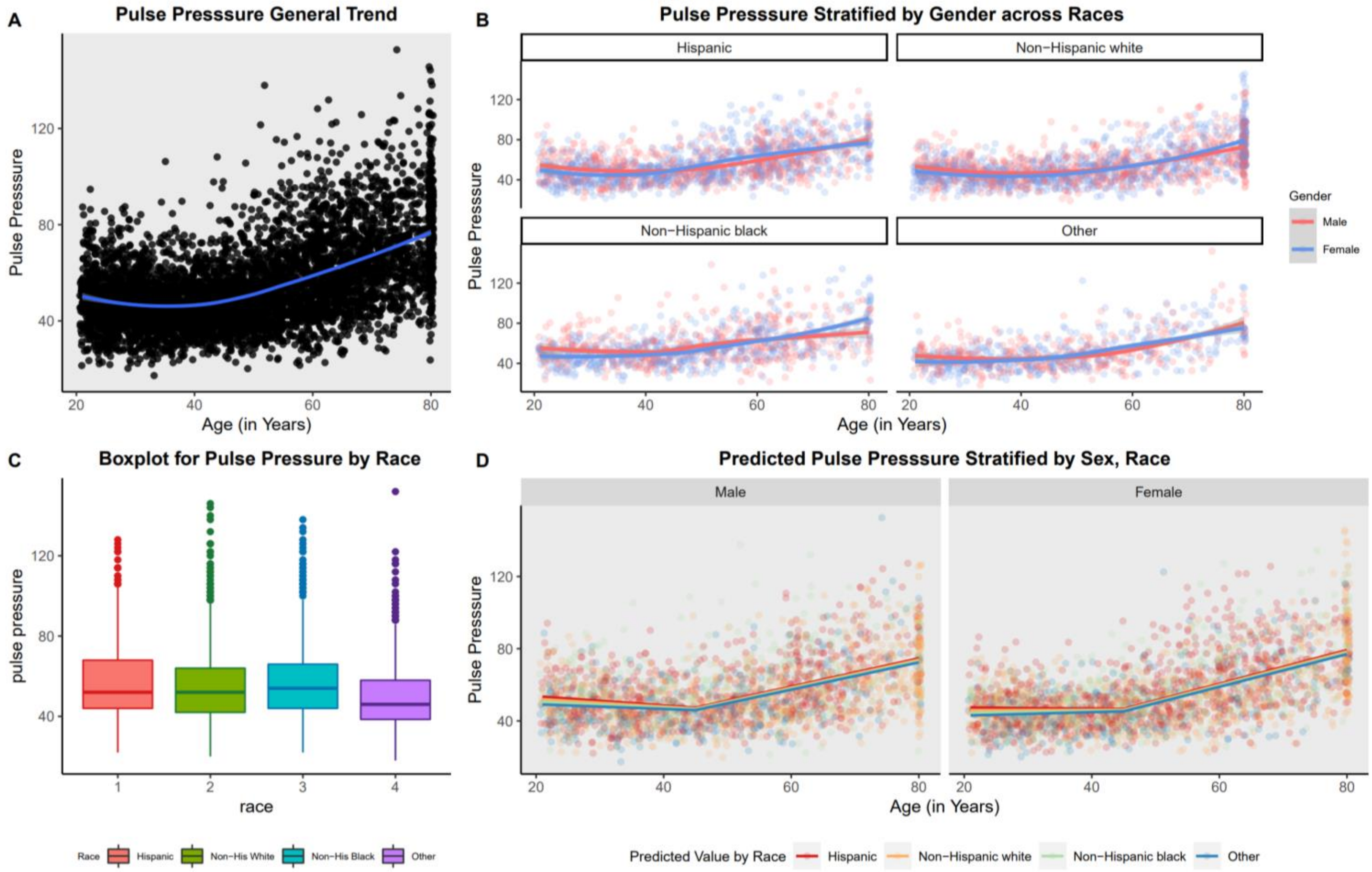


Table 1

variable	Value	Std.Error	p-value	2.5 %	97.5 %	significant
(Intercept)	61.167	2.635	0.000	56.00	66.33	TRUE
age	-0.300	0.076	0.000	-0.45	-0.15	TRUE
age_sp45	1.092	0.127	0.000	0.84	1.34	TRUE
race	-2.335	0.938	0.013	-4.17	-0.50	TRUE
female	-10.764	2.003	0.000	-14.69	-6.84	TRUE
age:female	0.224	0.058	0.000	0.11	0.34	TRUE
age_sp45:female	-0.075	0.095	0.433	-0.26	0.11	FALSE
age:race	0.043	0.027	0.118	-0.01	0.10	FALSE
age_sp45:race	-0.052	0.047	0.271	-0.14	0.04	FALSE

References

Tang, K., Medeiros, E., & Shah, A. (2020). Wide pulse pressure: A clinical review. *The Journal Of Clinical Hypertension*, *22*(11), 1960-1967. doi: 10.1111/jch.14051

http://www.cdc.gov/nhanes

https://data.library.virginia.edu/introduction-to-mediation-analysis/

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 5,* 1173-1182.

Problem Set 4

Zixuan Yu

Objective

Pulse pressure is the difference between diastolic blood pressure (BP) and systolic blood pressure, and its normal range is 40-60mmHg. Wide pulse pressure can suggest underlying health problems like hypertension and heart diseases (Tang et al., 2020). Our primary objective is to study how the mean pulse pressure of U.S. adults (20 to 80 years old) will differ by age, sex and race. The secondary objective is to study whether height, weight, and body mass index (BMI) are mediators between the relationship of pulse pressure and age, after adjusting for sex and race.

Data

We analyze a subset of data from the National Health and Nutrition Examination Survey. The outcome variable is pulse pressure (calculated by diastolic BP minus systolic BP). Key exposure variables are age (in years), female (indicator variable, 1 for female, 2 for male), race (indicator variable, 1 for Hispanic, 2 for non-Hispanic white, 3 for non-Hispanic black, and 4 for all other race). Height, weight, and BMI are potential mediating variables.

```
setwd("C:/Users/19092/OneDrive - Johns Hopkins/Term3/Biostat653")
library(readr)
dat <- read_csv("./Data/hw4.2018.nhanes.csv")
```

QUESTION: How does mean pulse pressure differ by age, sex and race?

```
## New names:
## * `` -> ...1

## Rows: 5395 Columns: 10
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (10): ...1, id, female, age, race, bp.sys, bp.di, weight, height, bmi
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(dat)
```

```
## # A tibble: 6 x 10
##     ...1    id female   age  race bp.sys bp.di weight height   bmi
##    <dbl> <dbl>  <dbl> <dbl> <dbl>  <dbl> <dbl>  <dbl>  <dbl> <dbl>
## 1      1 83732      0    62     2    128    70   94.8   184.  27.8
## 2      2 83733      0    53     2    146    88   90.4   171.  30.8
## 3      3 83734      0    78     2    138    46   83.4   170.  28.8
## 4      4 83735      1    56     2    132    72  110.    161.  42.4
## 5      5 83736      1    42     3    100    70   55.2   165.  20.3
## 6      6 83737      1    72     1    116    58   64.4   150   28.6
```

Methods

Data preparation: We remove incomplete records and records with unrealistic values: any records with systolic BP out of the range 80–250 mmHg or with diastolic BP out of the range 30–120 mm are removed.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
d <- dat %>% select(id, age, female, race, bp.sys, bp.di, height, weight, bmi)
table(complete.cases(d))
```

```
##
## FALSE  TRUE
##   369  5026
```

```
d.cc <- d[complete.cases(d),]
length(unique(d.cc$id));
```

```
## [1] 5026
```

```
##No duplicate id -> No duplicate record
str(d.cc)
```

```
## tibble [5,026 x 9] (S3: tbl_df/tbl/data.frame)
##  $ id    : num [1:5026] 83732 83733 83734 83735 83736 ...
##  $ age   : num [1:5026] 62 53 78 56 42 72 22 32 56 46 ...
##  $ female: num [1:5026] 0 0 0 1 1 1 0 1 0 0 ...
##  $ race  : num [1:5026] 2 2 2 2 3 1 3 1 3 2 ...
##  $ bp.sys: num [1:5026] 128 146 138 132 100 116 110 120 178 144 ...
##  $ bp.di : num [1:5026] 70 88 46 72 70 58 70 70 116 94 ...
##  $ height: num [1:5026] 184 171 170 161 165 ...
##  $ weight: num [1:5026] 94.8 90.4 83.4 109.8 55.2 ...
##  $ bmi   : num [1:5026] 27.8 30.8 28.8 42.4 20.3 ...
```

```
sum(d.cc$bp.di==0)
```

```
## [1] 24
```

```
sum(d.cc$bp.sys==0)
```

```
## [1] 0
```

Though there is no missing values, some of the blood pressure is 0, which is invalid.
I will remove these invalid records.

```
d2 <- d.cc[d.cc$bp.di!=0,]
table(d2$bp.sys)
```

```
##
##  82  84  86  88  90  92  94  96  98 100 102 104 106 108 110 112 114 116 118 120
##   1   6   4   4  11  16  31  49  64  74  92 124 196 173 185 205 260 285 227 223
```

```
## 122 124 126 128 130 132 134 136 138 140 142 144 146 148 150 152 154 156 158 160
## 230 253 265 215 186 173 171 157 127 100  95 103  91  85  52  45  55  66  40  26
## 162 164 166 168 170 172 174 176 178 180 182 184 186 188 190 192 194 196 198 200
##  22  24  37  28  15   9  11  15  14   7   5   4   6   7   4   2   5   5   3   1
## 202 204 206 208 210 218 230 236
##   1   4   3   1   1   1   1   1
```

```
table(d2$bp.di)
```

```
##
##  14  22  26  28  30  32  34  36  38  40  42  44  46  48  50  52  54  56  58  60
##   1   2   1   2   2   1   2   4   8   6  19  19  35  52  77 107 135 185 169 198
##  62  64  66  68  70  72  74  76  78  80  82  84  86  88  90  92  94  96  98 100
## 263 316 363 355 314 339 338 322 272 244 163 171 135 107  62  55  52  24  25  17
## 102 104 106 108 110 112 114 116 120
##  14   8   4   6   2   2   1   2   1
```

Then I filter out unreasonable values.
Threshold for systolic bp: 80–250 mm
for diastolic bp: 30–120 mm

```
d2 <- d2 %>% filter(bp.di>30)
```

Calculated the outcome variable pulse pressure by diastolic blood pressure minus systolic blood pressure.

```
d2 <- d2 %>% mutate(pp=bp.sys-bp.di)
str(d2)
```

```
## tibble [4,994 x 10] (S3: tbl_df/tbl/data.frame)
##  $ id    : num [1:4994] 83732 83733 83734 83735 83736 ...
##  $ age   : num [1:4994] 62 53 78 56 42 72 22 32 56 46 ...
##  $ female: num [1:4994] 0 0 0 1 1 1 0 1 0 0 ...
##  $ race  : num [1:4994] 2 2 2 2 3 1 3 1 3 2 ...
##  $ bp.sys: num [1:4994] 128 146 138 132 100 116 110 120 178 144 ...
##  $ bp.di : num [1:4994] 70 88 46 72 70 58 70 70 116 94 ...
##  $ height: num [1:4994] 184 171 170 161 165 ...
##  $ weight: num [1:4994] 94.8 90.4 83.4 109.8 55.2 ...
##  $ bmi   : num [1:4994] 27.8 30.8 28.8 42.4 20.3 ...
##  $ pp    : num [1:4994] 58 58 92 60 30 58 40 50 62 50 ...
```

After data cleaning, we get 4,994 records ready for further analysis out of the original 5,395 records.

See the distribution of pulse pressure through summary table.

```
table(d2$race);summary(d2$pp)
```

```
##
##    1    2    3    4
## 1549 1658 1029  758
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   42.00   52.00   55.34   64.00  152.00
```

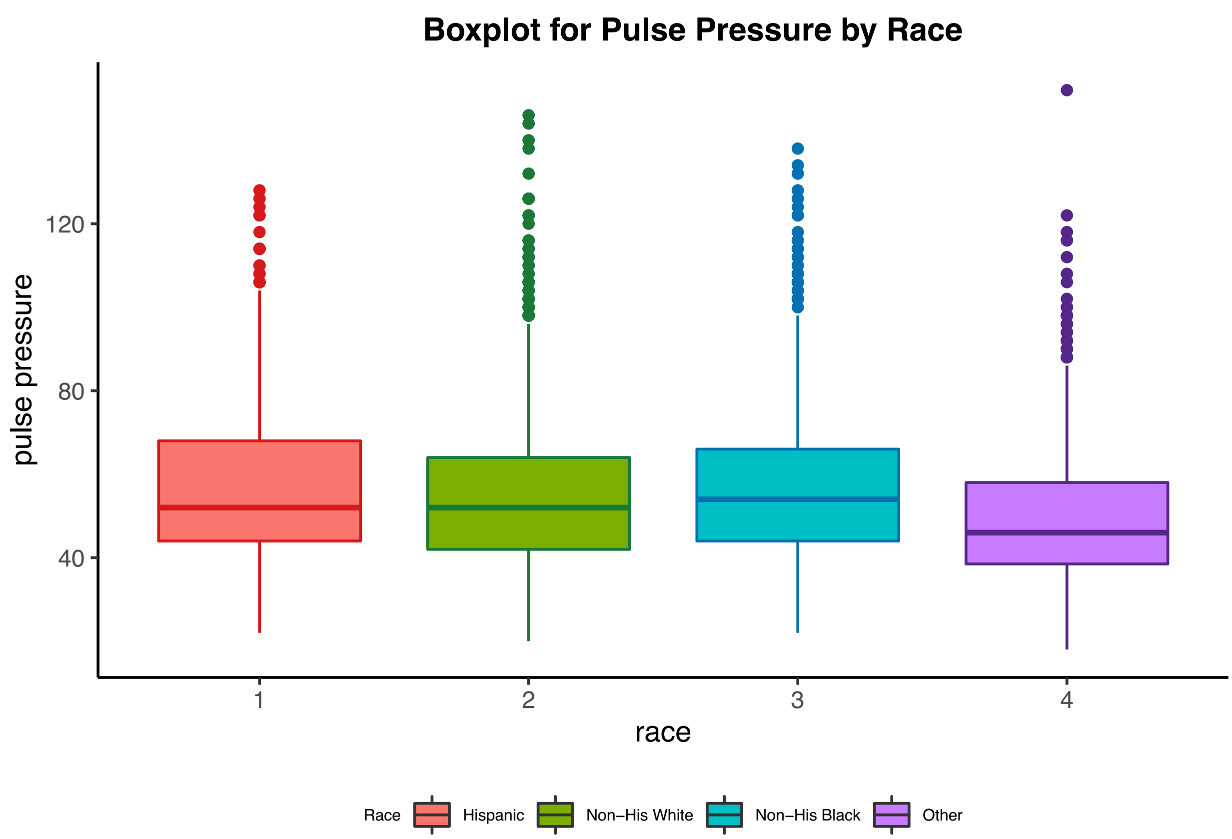
Examine the distribution of pulse pressure across race groups via box plot.

```
library(ggplot2)
p4 <- d2 %>% ggplot()+
    geom_boxplot( aes(x = factor(race), y = pp, group = race, fill = factor(race), color =factor(race))
    theme_bw()+
    scale_fill_discrete(name = "Race",labels=c("Hispanic","Non-His White","Non-His Black", "Other"))+
```

```r
    scale_color_manual(name = "Race",labels = "",values = c("#d7191c","#1b7837","#0571b0","#542788"))+
    ggtitle('Boxplot for Pulse Pressure by Race')+
    ylab('pulse pressure')+
    xlab('race')+
    guides(color=F)+
    theme_bw()+
    theme(plot.title = element_text(hjust = 0.55))+
    theme(legend.title=element_text(size=6),
                legend.text=element_text(size=6),
                legend.position = "bottom",
      panel.border = element_blank(),
            panel.grid.major = element_blank(),
            panel.grid.minor = element_blank(),
            axis.line = element_line(colour = "black"),
            plot.title = element_text(size = 12, face = "bold"))
```

```
## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =
## "none")` instead.
```

```r
p4
```



Find the median pulse pressure in each race group.

```r
d2%>%group_by(race) %>% summarise(mean_pp = median(pp))
```

```
## # A tibble: 4 x 2
##    race mean_pp
##   <dbl>   <dbl>
```

```
## 1      1       52
## 2      2       52
## 3      3       54
## 4      4       46
```

Find the median pulse pressure in Hispanics and Non-Hispanics.

```
d2 %>% filter(race %in% c(1,2,3)) %>% summarise(median_pp = median(pp))
```

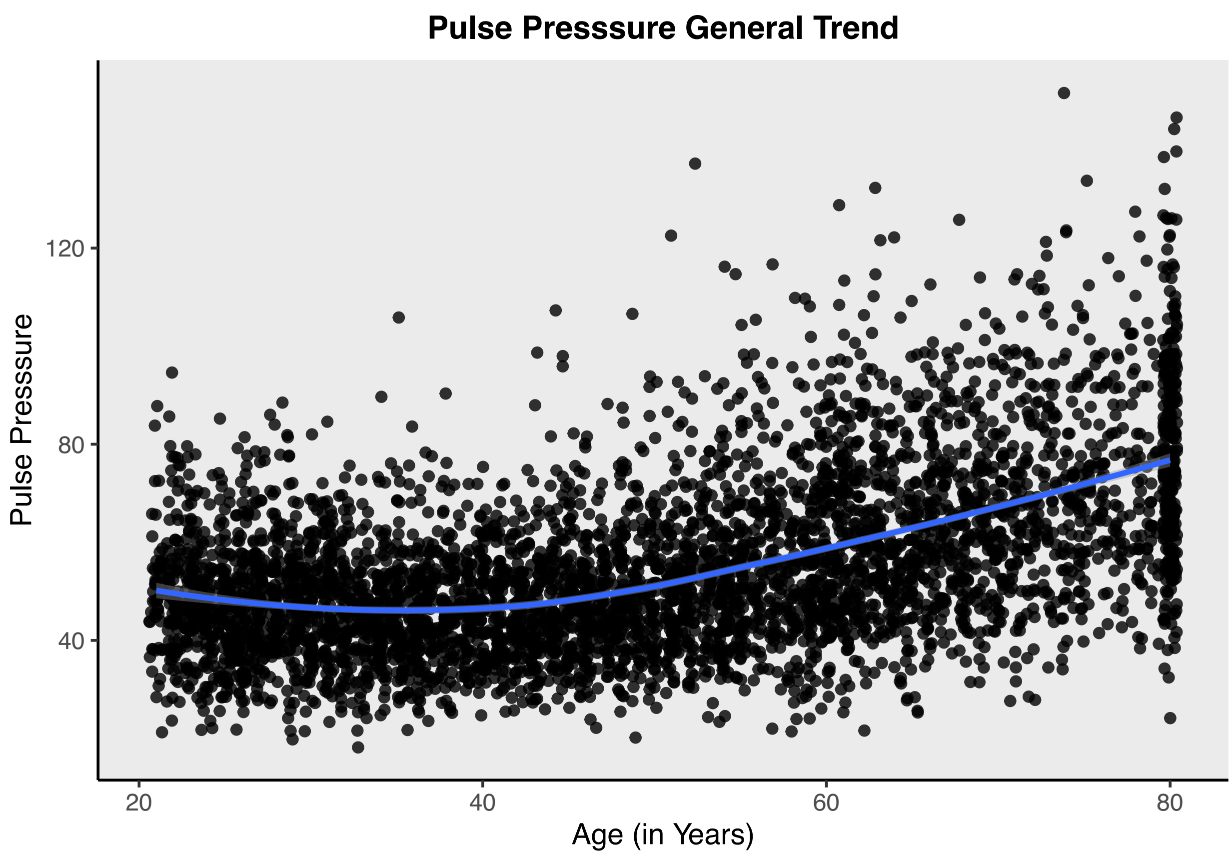
```
## # A tibble: 1 x 1
##   median_pp
##       <dbl>
## 1        52
```

From the box plot, we could see the distribution of pulse pressure are right-skewed. The mean is similar for Hispanics and Non-Hispanics (around 56.25), but higher than the mean of people with 'Other' race (50.22).

See how will pulse pressure change as a function of age.

```
#Pulse Pressure General Trend
library(ggplot2)
p1 <- d2 %>% ggplot(aes(x=age, y=pp))+
    geom_jitter(alpha=0.8)+
    geom_smooth(method = 'loess')+
    ggtitle("Pulse Presssure General Trend")+
    labs(y="Pulse Presssure",x="Age (in Years)") +
    theme(plot.title = element_text(hjust = 0.5))+
      theme(panel.border = element_blank(),
            panel.grid.major = element_blank(),
            panel.grid.minor = element_blank(),
            axis.line = element_line(colour = "black"),
            plot.title = element_text(size = 12, face = "bold"))
p1
```

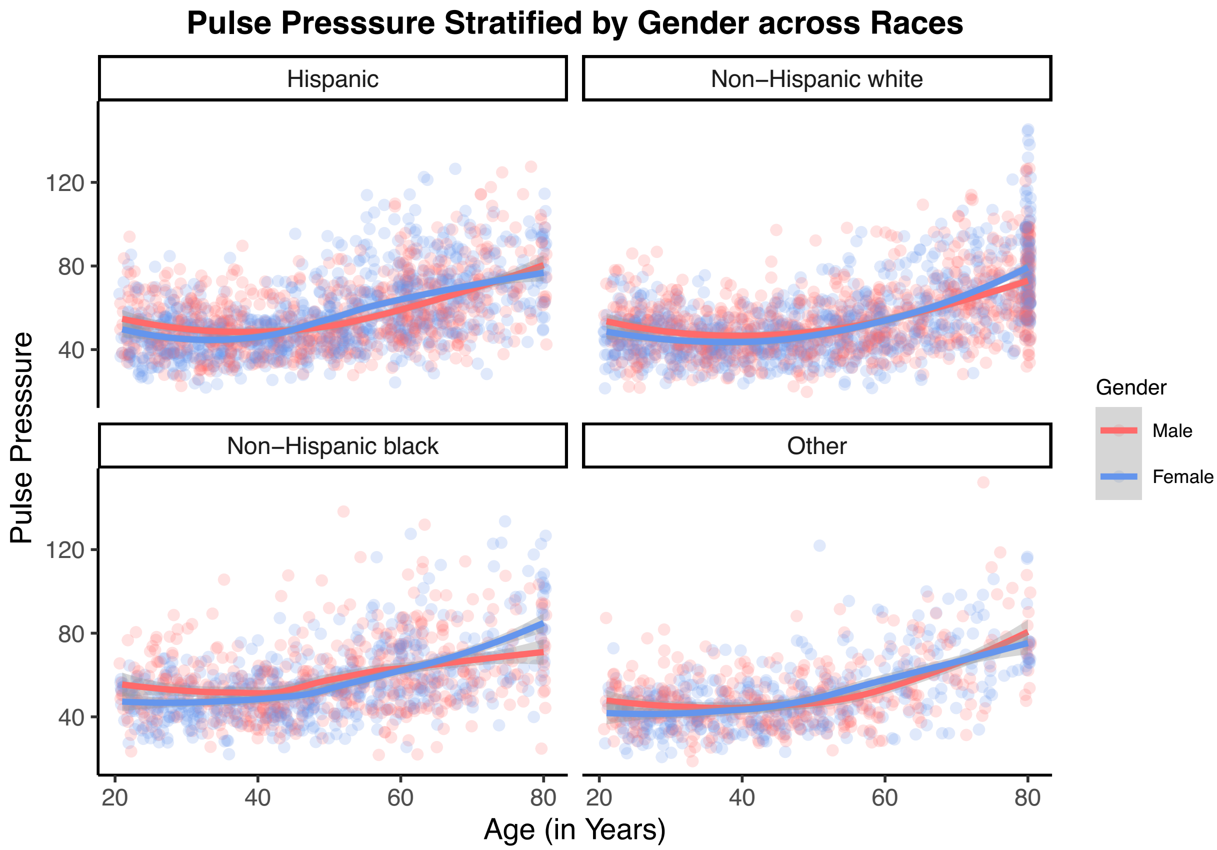
```
## `geom_smooth()` using formula 'y ~ x'
```



Pulse pressure decreases slightly as age increases before 45 years of age and then increases as age increases. The variance of pulse pressure increases as age increases according to Figure1A.

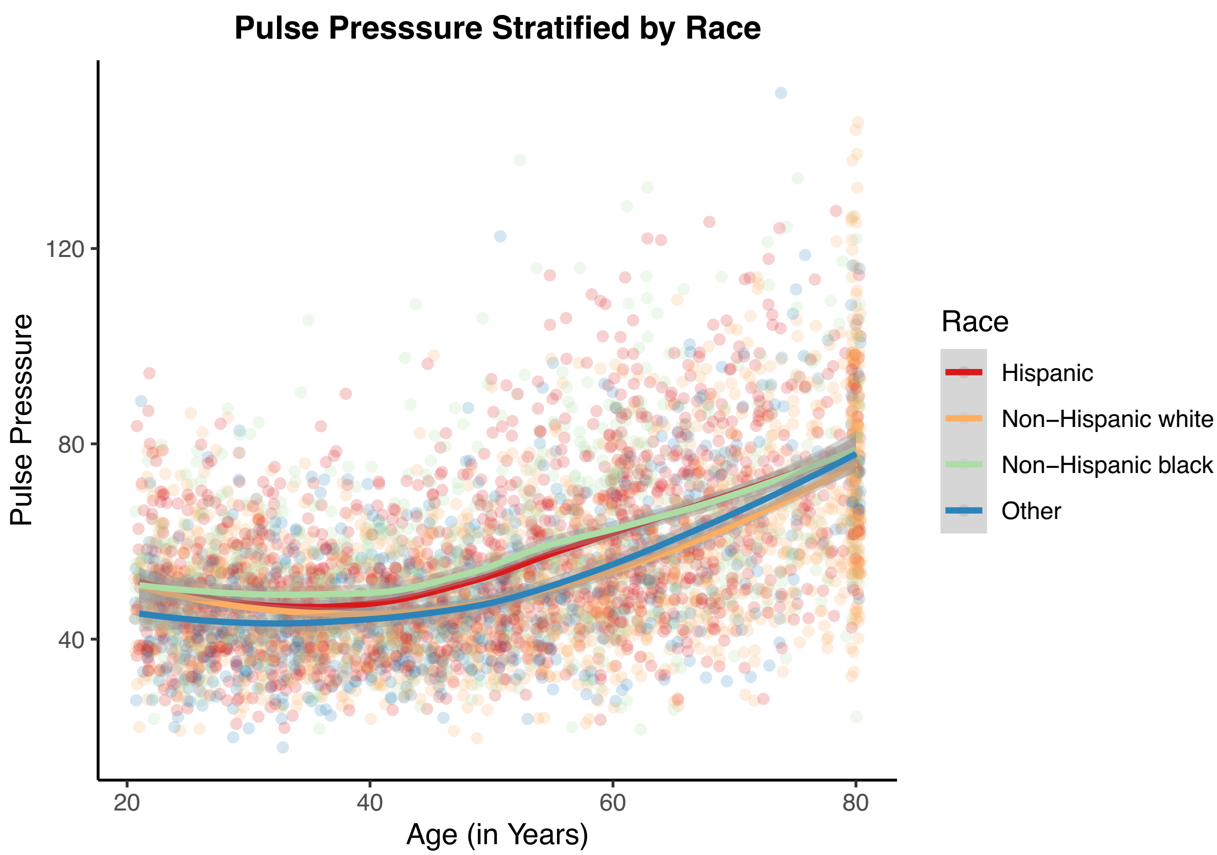
```
#Pulse Pressure Stratified by Gender across Races
race_labs <- c( '1' = "Hispanic",
                 '2' = "Non-Hispanic white",
                 '3' = "Non-Hispanic black",
                 '4' = "Other")
p2 <- d2 %>% ggplot(aes(x=age, y=pp, color = factor(female)))+
    geom_jitter(alpha=0.2)+
    geom_smooth(method = 'loess', size = 1.1,aes(x=age, y=pp, color = factor(female)))+
    scale_color_manual(name = "Gender",labels = c("Male", "Female"),values = c("Indianred1", "cornflowe
    facet_wrap(~race, nrow = 2,
               labeller = as_labeller(race_labs))+
    ggtitle("Pulse Presssure Stratified by Gender across Races")+
    labs(y="Pulse Presssure",x="Age (in Years)") +
    theme_classic()+
    theme(plot.title = element_text(hjust = 0.5, size = 12, face = "bold"))+
    theme(legend.title=element_text(size=8),
              legend.text=element_text(size=7))

p2
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
#Pulse Pressure Stratified by Gender across Races
p3 <-d2 %>%
    ggplot(aes(x=age, y=pp, color = as.factor(race)))+
    geom_jitter(alpha=0.2)+
    geom_smooth(method = 'loess', size = 1.1,aes(x=age, y=pp, color = as.factor(race)))+
    scale_color_manual(name = "Race",labels = c("Hispanic","Non-Hispanic white","Non-Hispanic black","O
    ggtitle("Pulse Presssure Stratified by Race")+
    labs(y="Pulse Presssure",x="Age (in Years)") +
    theme_bw()+
    theme(plot.title = element_text(hjust = 0.5))+
          theme(panel.border = element_blank(),
            panel.grid.major = element_blank(),
            panel.grid.minor = element_blank(),
            axis.line = element_line(colour = "black"),
            plot.title = element_text(size = 12, face = "bold"))
p3
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Pulse pressure decreases slightly as age increases before 45 years of age and then increases as age increases. The trend is consistent in sex groups as well as across race groups.

It is obvious that there is non-linear relationship between age and pulse pressure, so I decided to add a spline term at age=45. I define age_sp45 as $= (age - 45)^{+} = age - 45$ if age>45, 0 if not.

```
d2 <- d2 %>% mutate(age_sp45=ifelse(age-45>0,age-45,0))
```

It is intuitive to fit a multiple linear regression model to the data.

```
model1 <- lm(pp~age+age_sp45+female+race,data = d2)
summary(model1)
```

```
##
## Call:
## lm(formula = pp ~ age + age_sp45 + female + race, data = d2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.604 -10.187  -1.689   8.128  85.985
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 51.45914    1.39493  36.890  < 2e-16 ***
## age         -0.07014    0.03498  -2.005 0.044993 *
## age_sp45     0.91260    0.05142  17.747  < 2e-16 ***
## female      -0.34007    0.42984  -0.791 0.428890
## race        -0.72832    0.20675  -3.523 0.000431 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.17 on 4989 degrees of freedom
## Multiple R-squared:  0.2924, Adjusted R-squared:  0.2918
## F-statistic: 515.3 on 4 and 4989 DF,  p-value: < 2.2e-16
```

```
model2 <- lm(pp~age + age_sp45 + race +female:age +female:age_sp45,data = d2)
summary(model2)
```

```
##
## Call:
## lm(formula = pp ~ age + age_sp45 + race + female:age + female:age_sp45,
##     data = d2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.634 -10.190  -1.616   8.104  85.575
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     51.25210    1.36949  37.424  < 2e-16 ***
## age             -0.03731    0.03582  -1.041 0.297710
## age_sp45         0.75906    0.05815  13.053  < 2e-16 ***
## race            -0.73362    0.20608  -3.560 0.000375 ***
## age:female      -0.05967    0.01524  -3.916 9.13e-05 ***
## age_sp45:female  0.29189    0.05182   5.633 1.87e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.12 on 4988 degrees of freedom
## Multiple R-squared:  0.2971, Adjusted R-squared:  0.2964
## F-statistic: 421.7 on 5 and 4988 DF,  p-value: < 2.2e-16
```

```
AIC(model2)
```

```
## [1] 41307.18
```

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
lrtest(model1, model2)
```

```
## Likelihood ratio test
##
## Model 1: pp ~ age + age_sp45 + female + race
## Model 2: pp ~ age + age_sp45 + race + female:age + female:age_sp45
##   #Df LogLik Df  Chisq Pr(>Chisq)
## 1   6 -20663
## 2   7 -20647  1 33.582  6.832e-09 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model3 <- lm(pp~age+age_sp45+race + female + female:age +female:age_sp45 +race:age +race:age_sp45,data =
summary(model3)
```

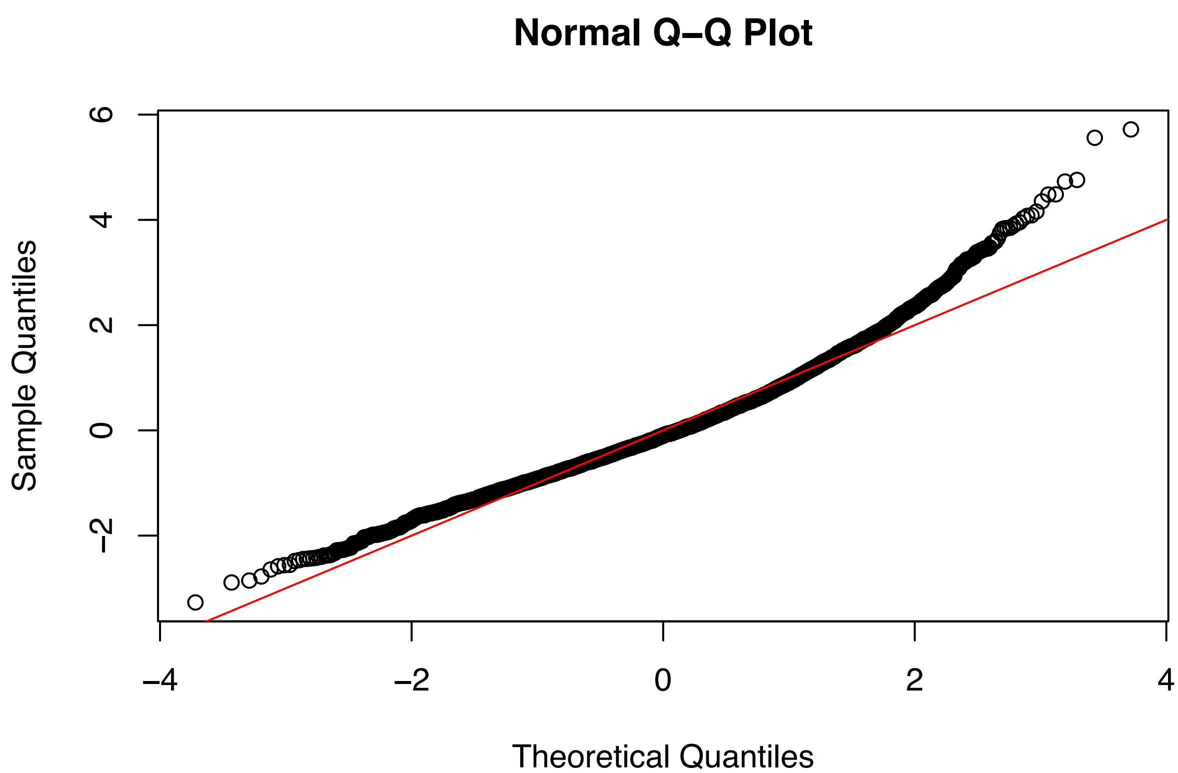
```
##
## Call:
## lm(formula = pp ~ age + age_sp45 + race + female + female:age +
##     female:age_sp45 + race:age + race:age_sp45, data = d2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.323 -10.087  -1.510   8.119  86.252
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      59.76487    3.37823  17.691  < 2e-16 ***
## age              -0.25952    0.09187  -2.825 0.004750 **
## age_sp45          1.02841    0.13731   7.490 8.10e-14 ***
## race             -1.87867    1.20484  -1.559 0.118995
## female          -11.02347    2.58005  -4.273 1.97e-05 ***
## age:female        0.23195    0.06993   3.317 0.000917 ***
## age_sp45:female  -0.08711    0.10273  -0.848 0.396524
## age:race          0.02892    0.03287   0.880 0.378955
## age_sp45:race    -0.02837    0.05025  -0.565 0.572374
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.09 on 4985 degrees of freedom
## Multiple R-squared:  0.2998, Adjusted R-squared:  0.2987
## F-statistic: 266.8 on 8 and 4985 DF,  p-value: < 2.2e-16
```

```
AIC(model3)
```

```
## [1] 41294.15
```

Model Checking for model 3

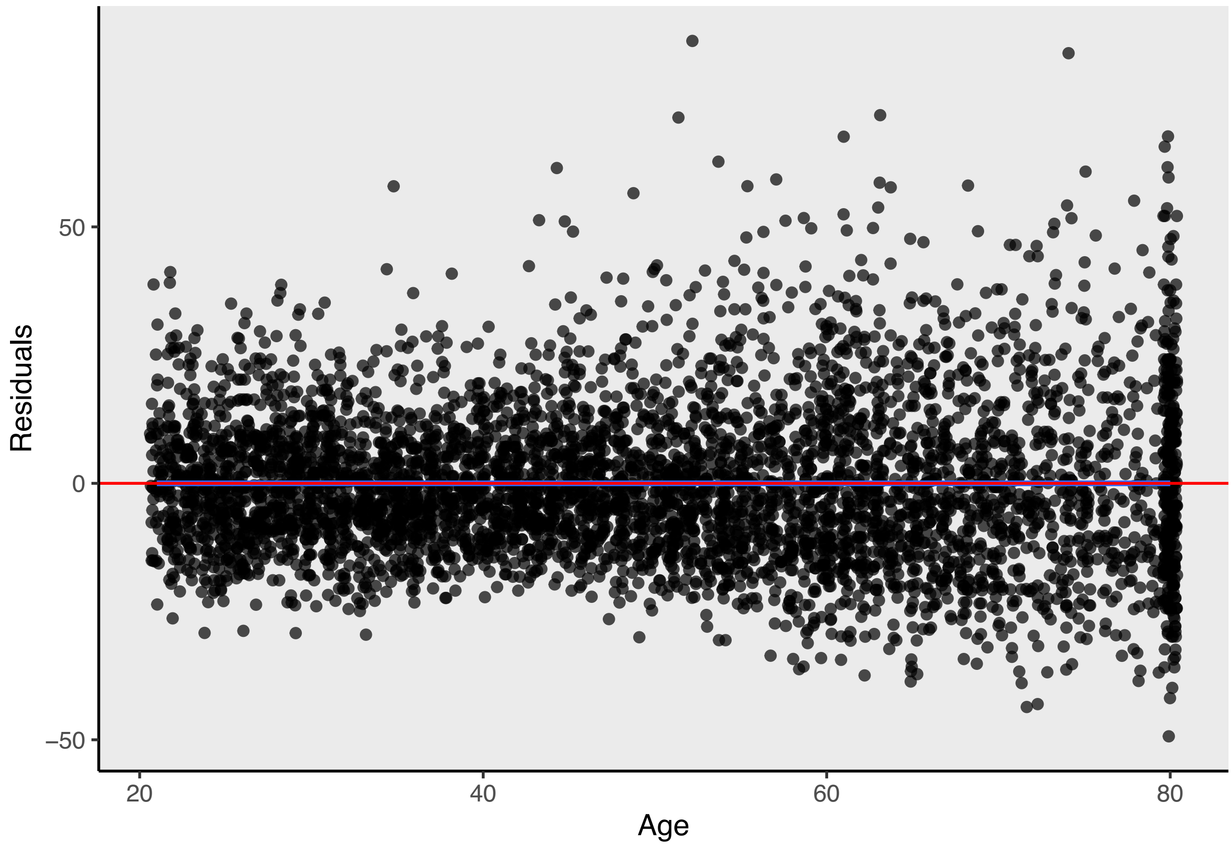
```
d2$res_m3 = residuals(model3)
qqnorm(scale(d2$res_m3))
abline(0,1,col='red')
```



The residuals of Model 3 is not exactly normally distributed, with more deviations at larger age.

```
pres_lm <- ggplot(d2, aes(x = age, y = res_m3))+
geom_jitter(alpha = 0.7)+
geom_smooth()+
labs(y = "Residuals", x = 'Age')+
geom_hline(yintercept = 0, color = "red") +
    theme(plot.title = element_text(hjust = 0.5))+
          theme(legend.title=element_text(size=9),
                legend.text=element_text(size=8),
              panel.border = element_blank(),
            panel.grid.major = element_blank(),
            panel.grid.minor = element_blank(),
            axis.line = element_line(colour = "black"),
            plot.title = element_text(size = 12, face = "bold"))
pres_lm
```

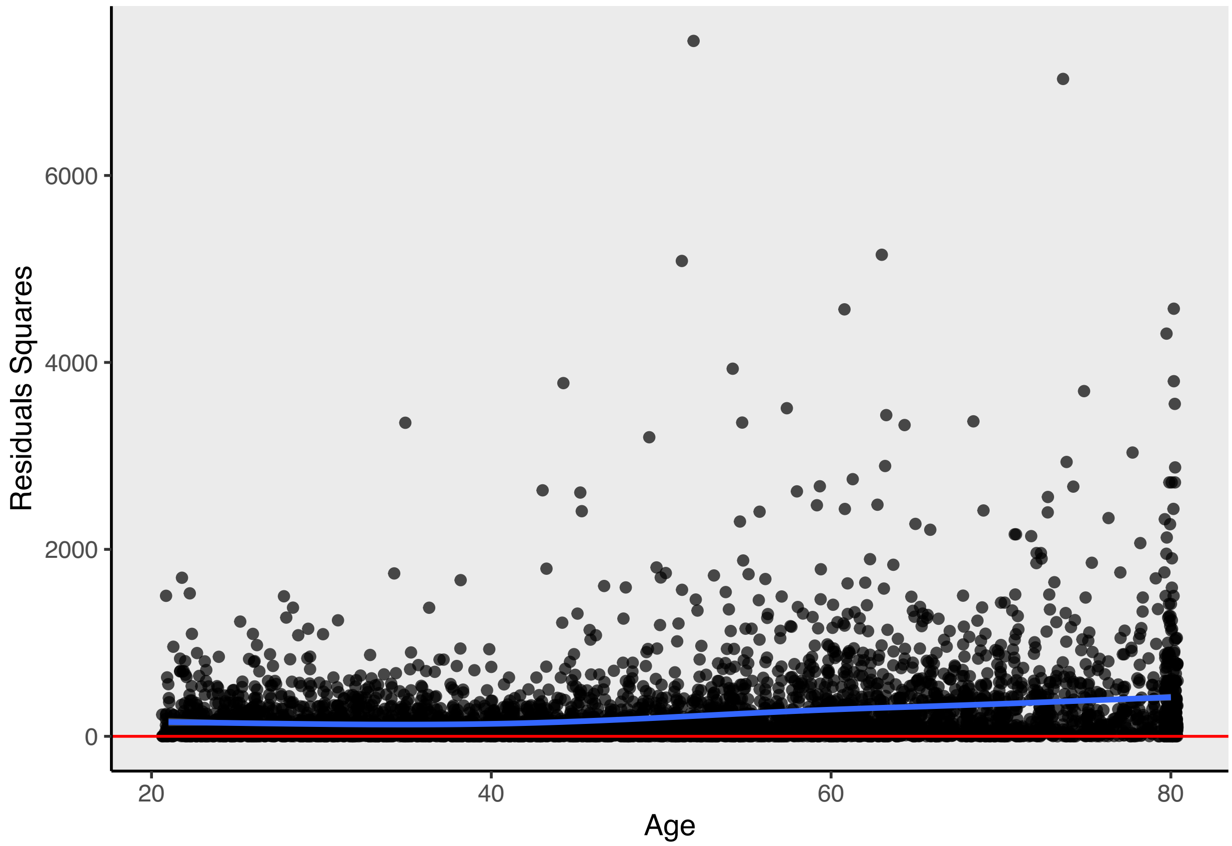
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



Residuals Squares

```
pressq_lm <- ggplot(d2, aes(x = age, y = res_m3^2))+
geom_jitter(alpha = 0.7)+
geom_smooth()+
labs(y = "Residuals Squares", x = 'Age')+
geom_hline(yintercept = 0, color = "red") +
    theme(plot.title = element_text(hjust = 0.5))+
          theme(legend.title=element_text(size=9),
                legend.text=element_text(size=8),
              panel.border = element_blank(),
            panel.grid.major = element_blank(),
            panel.grid.minor = element_blank(),
            axis.line = element_line(colour = "black"),
            plot.title = element_text(size = 12, face = "bold"))
pressq_lm
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



The Residuals and squared reisduals of Model 3 increases as age increases.

Using the `gls` command to allow variance to change as a function of age:

```
library(nlme)
```

```
##
## Attaching package: 'nlme'

## The following object is masked from 'package:dplyr':
##
##     collapse
```

```
model21 = gls(pp ~ age + age_sp45 + race + female,
              weights = varFunc(~age),data = d2)
summary(model21)
```

```
## Generalized least squares fit by REML
##   Model: pp ~ age + age_sp45 + race + female
##   Data: d2
##          AIC      BIC   logLik
##   40999.61 41038.7 -20493.8
##
## Variance function:
##  Structure: fixed weights
##  Formula: ~age
##
## Coefficients:
##                  Value Std.Error  t-value p-value
```

```
## (Intercept) 52.77560 1.1126181 47.43371  0.0000
## age         -0.08309 0.0289439 -2.87074  0.0041
## age_sp45     0.92687 0.0476545 19.44981  0.0000
## race        -0.81854 0.1890080 -4.33071  0.0000
## female      -1.52835 0.4001271 -3.81966  0.0001
##
##  Correlation:
##          (Intr) age    ag_s45 race
## age      -0.875
## age_sp45  0.740 -0.911
## race     -0.389 -0.001  0.027
## female   -0.204  0.002  0.009  0.029
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -2.7156357 -0.6935188 -0.1143659  0.5660296  5.5270973
##
## Residual standard error: 2.145515
## Degrees of freedom: 4994 total; 4989 residual
```

```
AIC(model21)
```

```
## [1] 40999.61
```

```
model22 = gls(pp ~ age + age_sp45 + race + female +
                  female:age + female:age_sp45,
              weights = varFunc(~age),data = d2)
summary(model22)
```

```
## Generalized least squares fit by REML
##   Model: pp ~ age + age_sp45 + race + female + female:age + female:age_sp45
##   Data: d2
##       AIC      BIC    logLik
##   40953.3 41005.41 -20468.65
##
## Variance function:
##  Structure: fixed weights
##  Formula: ~age
##
## Coefficients:
##                     Value Std.Error  t-value p-value
## (Intercept)      57.65343 1.5322886 37.62570  0.0000
## age              -0.20057 0.0423916 -4.73131  0.0000
## age_sp45          0.96941 0.0694916 13.94998  0.0000
## race             -0.83230 0.1879454 -4.42840  0.0000
## female          -10.55278 1.9991757 -5.27856  0.0000
## age:female        0.21834 0.0577381  3.78156  0.0002
## age_sp45:female  -0.06695 0.0949864 -0.70484  0.4809
##
##  Correlation:
##                 (Intr) age    ag_s45 race   female ag:fml
## age             -0.933
## age_sp45         0.800 -0.916
## race            -0.288  0.006  0.014
## female          -0.707  0.714 -0.616  0.015
```

```
## age:female       0.686 -0.734  0.672 -0.009 -0.972
## age_sp45:female -0.590  0.670 -0.731  0.006  0.835 -0.912
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -2.79158204 -0.70029379 -0.09819766  0.56598442  5.62201438
##
## Residual standard error: 2.13333
## Degrees of freedom: 4994 total; 4987 residual
```

```r
AIC(model22)
```

```
## [1] 40953.3
```

```r
model23 = gls(pp ~ age + age_sp45 + race + female +
                  female:age + female:age_sp45 +
                  race:age + race:age_sp45,
              weights = varFunc(~age),data = d2)
summary(model23)
```

```
## Generalized least squares fit by REML
##   Model: pp ~ age + age_sp45 + race + female + female:age + female:age_sp45 +      race:age + race:a
##   Data: d2
##        AIC      BIC    logLik
##   40965.67 41030.81 -20472.84
##
## Variance function:
##  Structure: fixed weights
##  Formula: ~age
##
## Coefficients:
##                     Value Std.Error   t-value p-value
## (Intercept)      61.16679 2.6345504 23.217163  0.0000
## age              -0.30000 0.0761587 -3.939092  0.0001
## age_sp45          1.09151 0.1271624  8.583627  0.0000
## race             -2.33484 0.9375112 -2.490471  0.0128
## female          -10.76447 2.0028881 -5.374473  0.0000
## age:female        0.22421 0.0578369  3.876557  0.0001
## age_sp45:female  -0.07459 0.0950876 -0.784419  0.4328
## age:race          0.04255 0.0271942  1.564744  0.1177
## age_sp45:race    -0.05158 0.0468168 -1.101663  0.2707
##
##  Correlation:
##                 (Intr) age    ag_s45 race   female ag:fml ag_sp45:f age:rc
## age             -0.973
## age_sp45         0.836 -0.911
## race            -0.831  0.810 -0.702
## female          -0.461  0.446 -0.376  0.063
## age:female       0.447 -0.456  0.406 -0.060 -0.972
## age_sp45:female -0.382  0.412 -0.432  0.048  0.836 -0.912
## age:race         0.807 -0.831  0.765 -0.973 -0.060  0.058 -0.047
## age_sp45:race   -0.680  0.743 -0.837  0.822  0.046 -0.045  0.038    -0.898
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
```

```
## -2.8075797 -0.6979236 -0.1055712  0.5746900  5.6065954
##
## Residual standard error: 2.133132
## Degrees of freedom: 4994 total; 4985 residual
```

```
AIC(model23)
```

```
## [1] 40965.67
```

Check for AIC to select best predicting models:

```
AIC(model1)
```

```
## [1] 41338.77
```

```
AIC(model2)
```

```
## [1] 41307.18
```

```
AIC(model3)
```

```
## [1] 41294.15
```

```
AIC(model21)
```

```
## [1] 40999.61
```

```
AIC(model22)
```

```
## [1] 40953.3
```

```
AIC(model23)
```

```
## [1] 40965.67
```

The AIC of model3, Model23 is 41294, 40966, respectively.

Models allow variance to change as a function of age are better.

```
lrtest(model21, model22)
```

```
## Likelihood ratio test
##
## Model 1: pp ~ age + age_sp45 + race + female
## Model 2: pp ~ age + age_sp45 + race + female + female:age + female:age_sp45
##   #Df LogLik Df  Chisq Pr(>Chisq)
## 1   6 -20494
## 2   8 -20469  2 50.309   1.19e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Likelihood ratio tests (LRT) are conducted between Model 21(null model 1) and Model22 to see if pulse pressure change rate will differ between females and male. With a p-value of 1.19e-11, we reject the null that pulse pressure change rate is same in females and males.

```
lrtest(model21, model23)
```

```
## Likelihood ratio test
##
## Model 1: pp ~ age + age_sp45 + race + female
## Model 2: pp ~ age + age_sp45 + race + female + female:age + female:age_sp45 +
##     race:age + race:age_sp45
##   #Df LogLik Df  Chisq Pr(>Chisq)
```

```
## 1   6 -20494
## 2  10 -20473  4 41.935  1.721e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

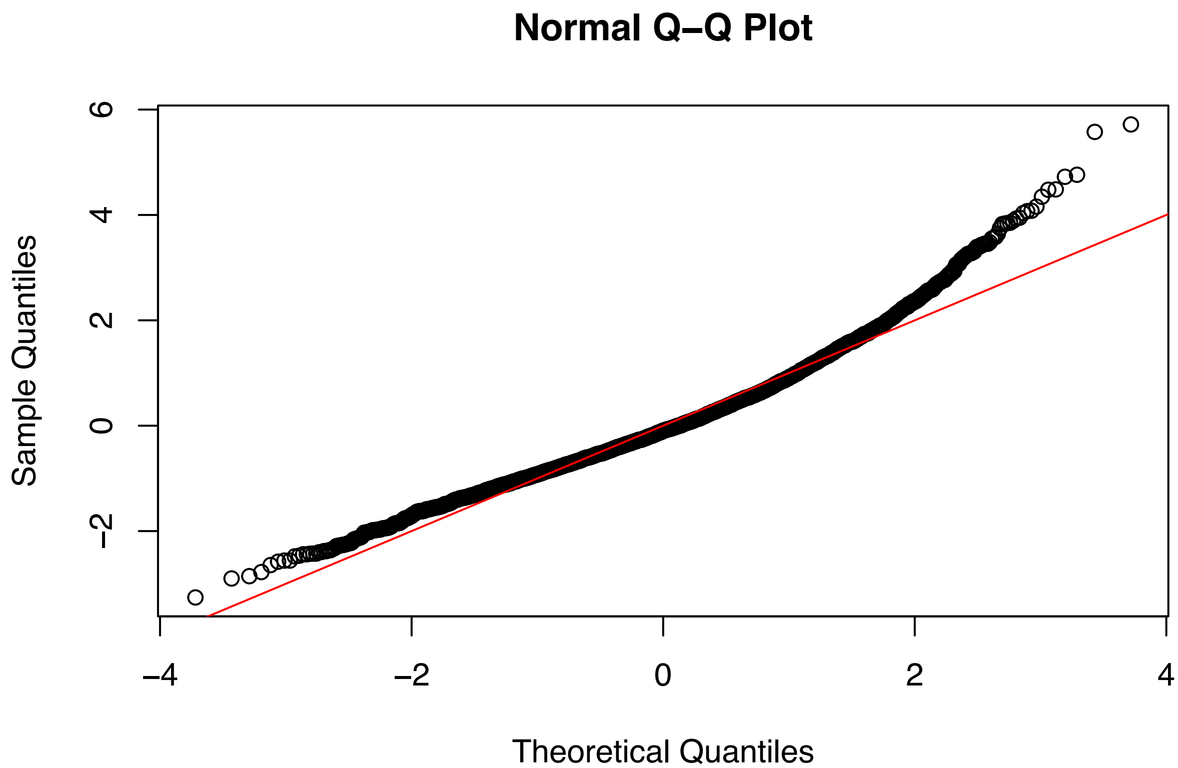
```r
lrtest(model22, model23)
```

```
## Likelihood ratio test
##
## Model 1: pp ~ age + age_sp45 + race + female + female:age + female:age_sp45
## Model 2: pp ~ age + age_sp45 + race + female + female:age + female:age_sp45 +
##     race:age + race:age_sp45
##   #Df LogLik Df  Chisq Pr(>Chisq)
## 1   8 -20469
## 2  10 -20473  2 8.3747    0.01519 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

LRT are conducted between Model 22(null model 2) and Model23 to determine if mean pulse pressure change rate will change by race. With a p-value of 0.016, we reject the null that pulse pressure change rate is the same across races.

Based on the AIC values and results of LRT, we favor Model 23 (recap: its covariate are age, age_sp45, interaction terms between age, age_sp45 and sex, race, respectively).
Model Checking for model 23

```r
d2$fit_m23 = predict(model23)
d2$res_m23 = residuals(model23)
qqnorm(scale(d2$res_m23))
abline(0,1,col='red')
```

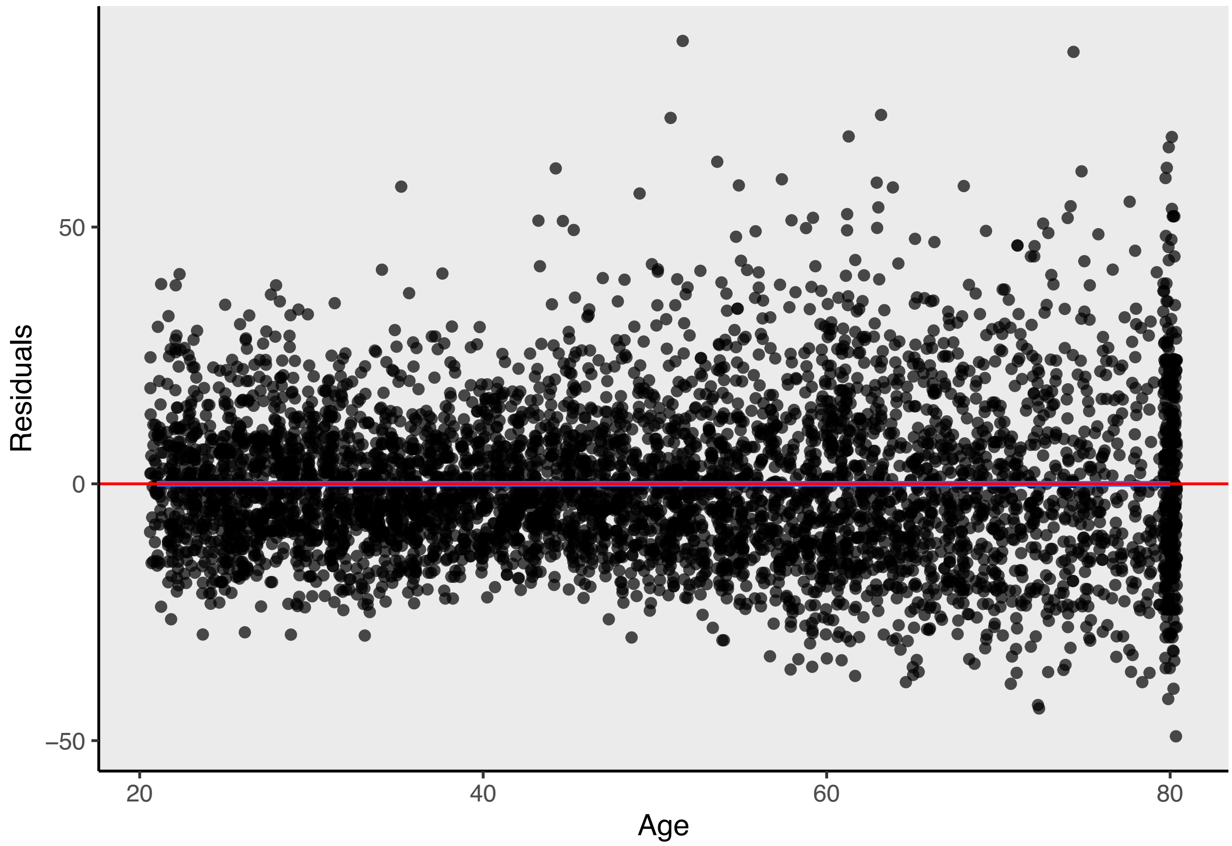


The residuals of Model 23 is not exactly normally distributed.

Residuals

```
pres <- ggplot(d2, aes(x = age, y = res_m23))+
geom_jitter(alpha = 0.7)+
geom_smooth()+
labs(y = "Residuals", x = 'Age')+
geom_hline(yintercept = 0, color = "red") +
    theme(plot.title = element_text(hjust = 0.5))+
          theme(legend.title=element_text(size=9),
                legend.text=element_text(size=8),
              panel.border = element_blank(),
            panel.grid.major = element_blank(),
            panel.grid.minor = element_blank(),
            axis.line = element_line(colour = "black"),
            plot.title = element_text(size = 12, face = "bold"))
pres
```

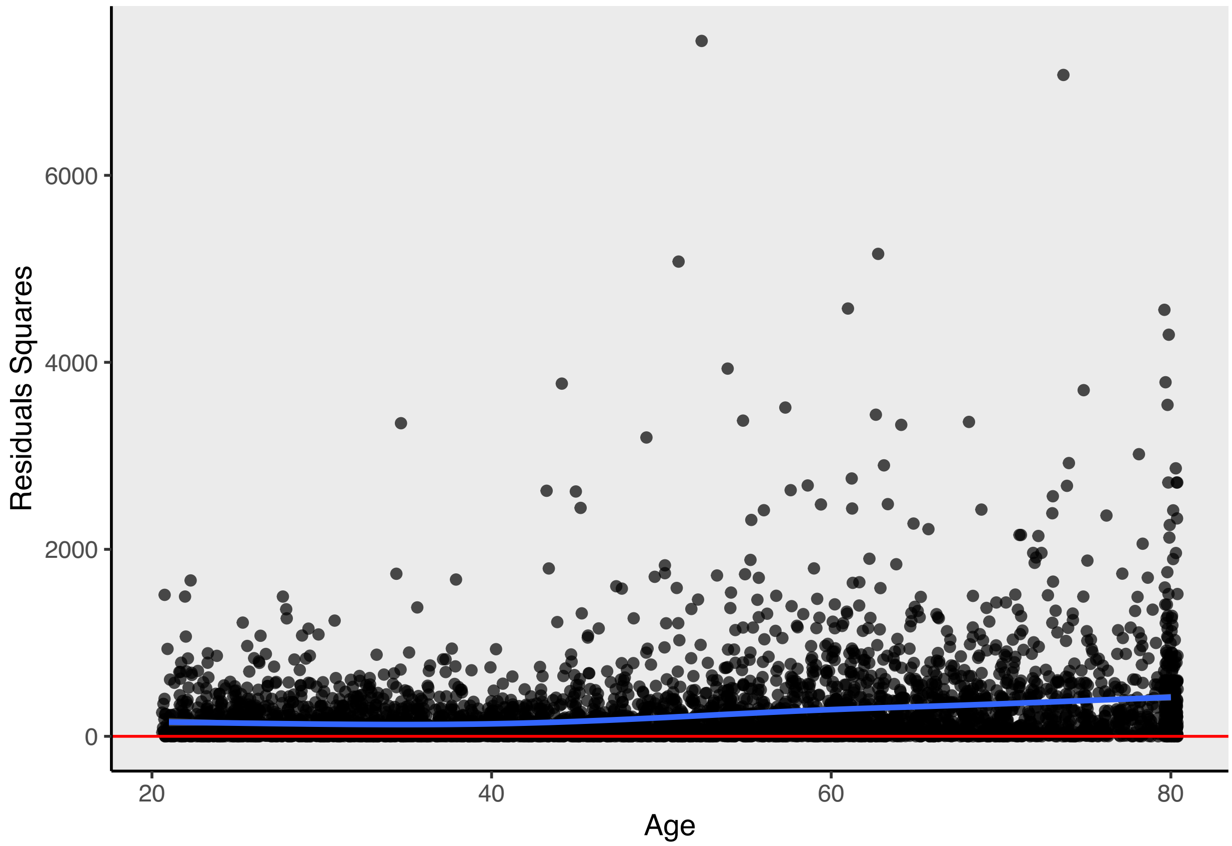
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



The variance of pulse pressure increases as age increases.
Residuals Squares

```
pressq <- ggplot(d2, aes(x = age, y = res_m23^2))+
geom_jitter(alpha = 0.7)+
geom_smooth()+
labs(y = "Residuals Squares", x = 'Age')+
geom_hline(yintercept = 0, color = "red") +
    theme(plot.title = element_text(hjust = 0.5))+
          theme(legend.title=element_text(size=9),
                legend.text=element_text(size=8),
              panel.border = element_blank(),
            panel.grid.major = element_blank(),
            panel.grid.minor = element_blank(),
            axis.line = element_line(colour = "black"),
            plot.title = element_text(size = 12, face = "bold"))
pressq
```

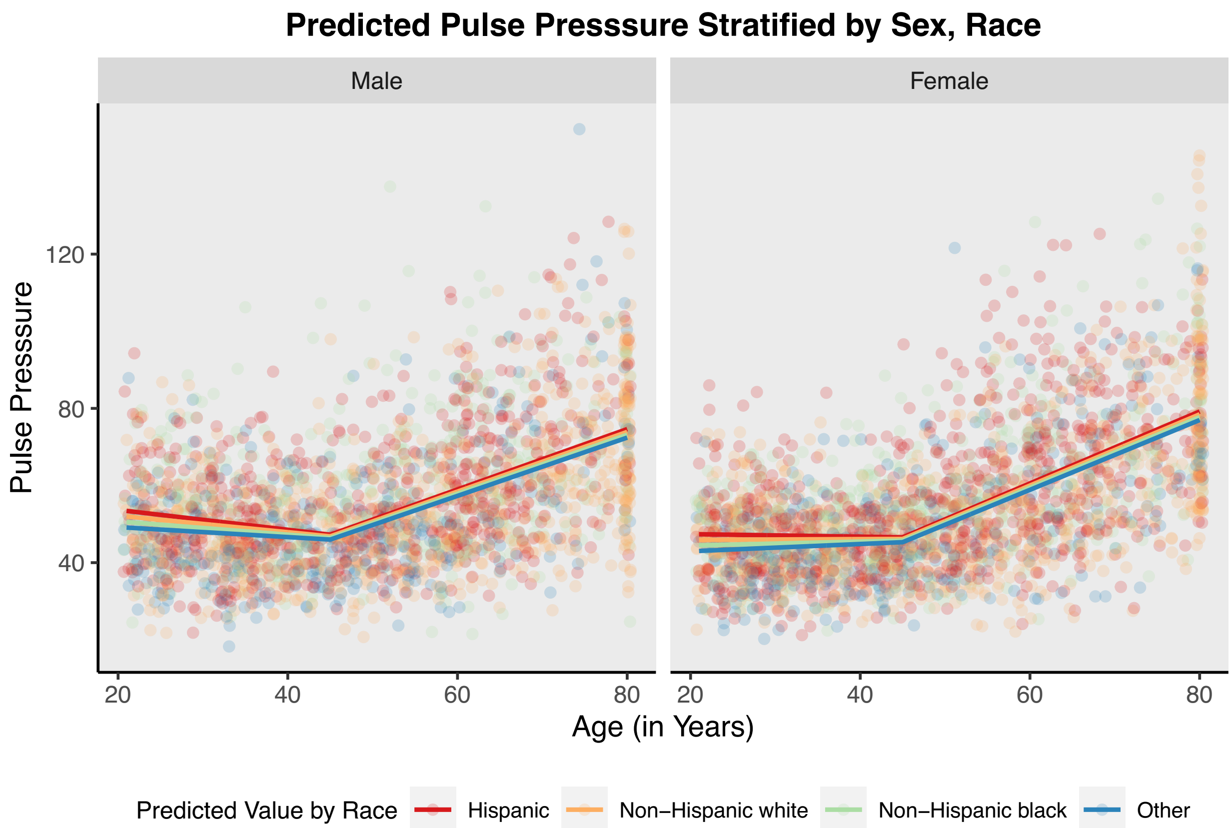
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



The residuals and residual squares gradually increases as age increases.

Plot the predicted value based on Model23.

```
#Pulse Pressure Stratified by Gender across Races
#Predict value by model23 vs. true values
gender_labs  <- c( '1' = "Female",
                   '0' = "Male")

p33 <-d2 %>%
    ggplot(aes(x=age, y=pp, color = as.factor(race)))+
    geom_jitter(alpha=0.2)+
    geom_line(size = 0.8,aes(x=age, y=fit_m23, color = as.factor(race)))+
    scale_color_manual(name = "Predicted Value by Race",labels = c("Hispanic","Non-Hispanic white","Non-
      facet_wrap(~female, nrow = 1,
               labeller = as_labeller(gender_labs))+
    ggtitle("Predicted Pulse Presssure Stratified by Sex, Race")+
    labs(y="Pulse Presssure",x="Age (in Years)") +
    theme(plot.title = element_text(hjust = 0.5))+
          theme(legend.title=element_text(size=9),
                legend.text=element_text(size=7.5),
                legend.position = "bottom",
              panel.border = element_blank(),
            panel.grid.major = element_blank(),
            panel.grid.minor = element_blank(),
            axis.line = element_line(colour = "black"),
            plot.title = element_text(size = 12, face = "bold"))
p33
```



Bootstrap Standard Error and 95% Confidence Interval between female and male

Since the distribution of pulse pressure is positively skewed, we use the bootstrap method (sampling 11,000 times with replacement, using the percentile method) to calculate the 95% confidence interval (CI) of difference in mean pulse pressure between females and males at the age of 35, 55, 75, accordingly.

```
library(boot)
pp.est <- function(data, id){
 dt <- data[id,] # allows boot to select sample
 fit1 <- gls(pp ~ age + age_sp45 + race + female +
                  female:age + female:age_sp45 +
                  race:age + race:age_sp45,
              weights = varFunc(~age),data = dt)
 diff35 <- coef(fit1)[5] + 35*coef(fit1)[6]
 diff55 <- coef(fit1)[5] + 55*coef(fit1)[6] + 10*coef(fit1)[7]
 diff75 <- coef(fit1)[5] + 75*coef(fit1)[6] + 30*coef(fit1)[7]
 c(diff35,diff55,diff75)
}
set.seed(317)
pp.result <- boot(d2, pp.est, R=11000)
pp.result
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
```

```
##
## Call:
## boot(data = d2, statistic = pp.est, R = 11000)
##
##
## Bootstrap Statistics :
##       original       bias    std. error
## t1* -2.9171821 -0.006508211   0.4647761
## t2*  0.8210958 -0.007859852   0.5400343
## t3*  3.8134885  0.008395939   1.0836997
```

```
#The element $t contains a total number of R values (i.e. number of bootstrap samples) for each statist
head(pp.result$t)
```

```
##           [,1]        [,2]     [,3]
## [1,] -4.190090 -0.41887722 2.745690
## [2,] -3.277647  1.48451666 4.948640
## [3,] -2.886391  0.57448735 1.809999
## [4,] -3.922656 -0.08183851 1.954853
## [5,] -2.244474  0.45631817 2.039469
## [6,] -2.393784  0.94266003 2.834705
```

```
pp.bootci <- sapply(1:3,function(x) boot.ci(pp.result,index = x,type = "basic")$basic[4:5])
pp.boot.result <- data.frame(rbind(pp.result$t0,pp.bootci))
rownames(pp.boot.result) <- c("Est","Lower","Upper")
colnames(pp.boot.result) <- c("diff35","diff55","diff75")
round(pp.boot.result, 3)
```

```
##       diff35 diff55 diff75
## Est   -2.917  0.821  3.813
## Lower -3.805 -0.239  1.685
## Upper -2.008  1.895  5.968
```

The average difference between female and male at age 35 is -2.92 (95% CI: -3.81 to -2.00, at age 55 is 0.82 (95% CI: -0.24 to -1.90), at age 75 is 3.81 (95% CI: 1.69 to 5.97). The change rate difference is not significant across race groups before 45 (p-value = 0.12) or after (p-value = 0.27), after adjusting for sex.

```
library(ggpubr)
library(cowplot)
```

Figure1 for the report

```
##
## Attaching package: 'cowplot'

## The following object is masked from 'package:ggpubr':
##
##     get_legend
```

```
library(grid)
library(gridExtra)
```

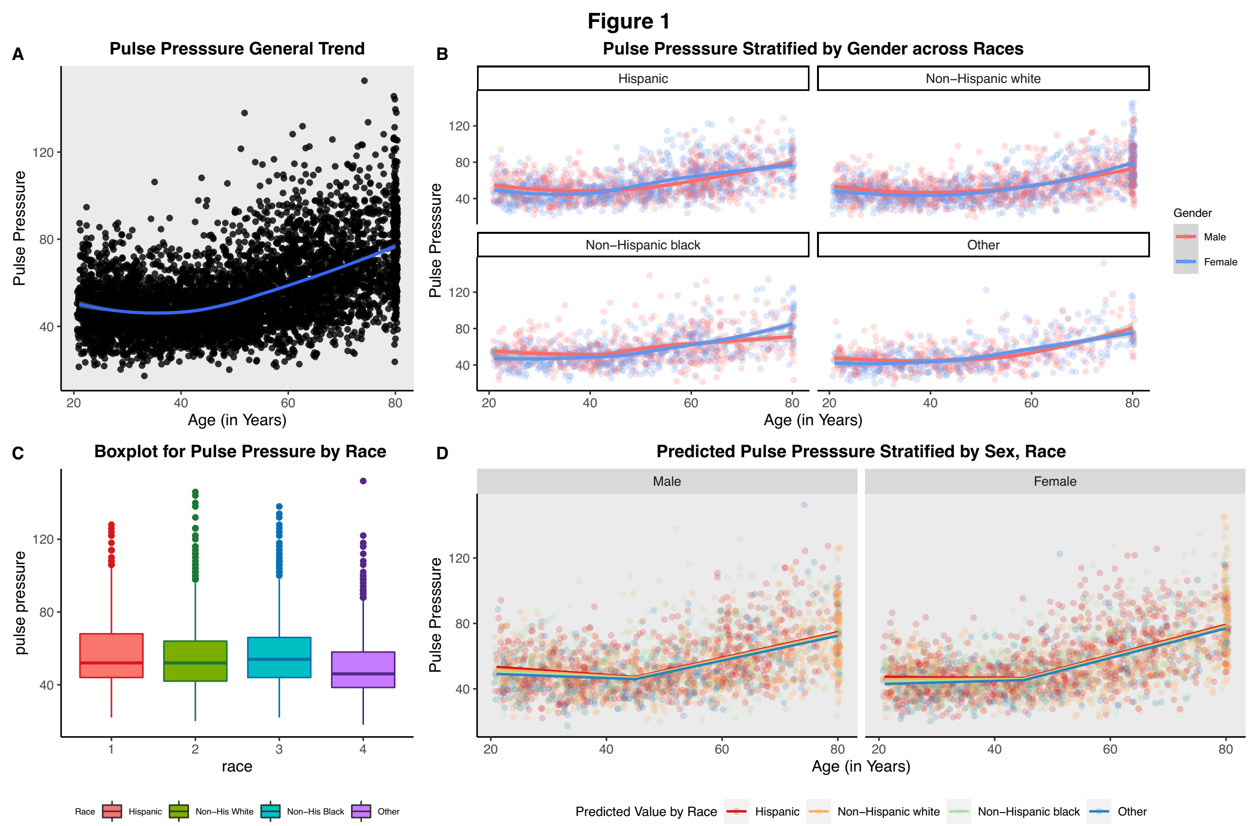
```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
```

```
##     combine
```

```r
gt <- arrangeGrob(p1,p2,
                  p4, p33,
                  ncol = 2, nrow = 2,
                  top=textGrob("Figure 1", gp=gpar(fontsize=15,font=2)),
                  layout_matrix = rbind(c(1,2,2), c(3,4,4)))
```

```
## `geom_smooth()` using formula 'y ~ x'

## `geom_smooth()` using formula 'y ~ x'
```

```r
# Add labels to the arranged plots
figure1 <- as_ggplot(gt) +
  draw_plot_label( label = c("A", "B", "C", "D"),size = 12.5,
                  x = c(0, 0.34, 0,0.34), y = c(0.96, 0.96, 0.48,0.48))
figure1
```




```r
tTable<-as.data.frame(summary(model23)$tTable[,c('Value', 'Std.Error', 'p-value')])
tTable$variable <- rownames(tTable)
ci<-as.data.frame(round(confint(model23, level = 0.95), 2))
ci$variable <- rownames(ci)
coef_table <- left_join(tTable, ci, by = 'variable')
table1 <- coef_table[,c("variable", "Value", "Std.Error", 'p-value', '2.5 %', '97.5 %')]
table1$significant = table1$`p-value`<0.05
table1 <- table1 %>% mutate(across(is.numeric, round, digits=3))
```

Figure1 for the report

```
## Warning: Predicate functions must be wrapped in `where()`.
##
##   # Bad
##   data %>% select(is.numeric)
##
##   # Good
##   data %>% select(where(is.numeric))
##
## i Please update your code.
## This message is displayed once per session.
```

```
library(knitr)
knitr::kable(table1)
```

variable	Value	Std.Error	p-value	2.5 %	97.5 %	significant
(Intercept)	61.167	2.635	0.000	56.00	66.33	TRUE
age	-0.300	0.076	0.000	-0.45	-0.15	TRUE
age_sp45	1.092	0.127	0.000	0.84	1.34	TRUE
race	-2.335	0.938	0.013	-4.17	-0.50	TRUE
female	-10.764	2.003	0.000	-14.69	-6.84	TRUE
age:female	0.224	0.058	0.000	0.11	0.34	TRUE
age_sp45:female	-0.075	0.095	0.433	-0.26	0.11	FALSE
age:race	0.043	0.027	0.118	-0.01	0.10	FALSE
age_sp45:race	-0.052	0.047	0.271	-0.14	0.04	FALSE

```
library(multcomp)
```

pp change rate of woman aged 20-45

```
## Loading required package: mvtnorm

## Loading required package: survival

##
## Attaching package: 'survival'

## The following object is masked from 'package:boot':
##
##     aml

## Loading required package: TH.data

## Loading required package: MASS

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

##
## Attaching package: 'TH.data'

## The following object is masked from 'package:MASS':
##
##     geyser
```

```r
summary(glht(model23, linfct = "`age` + `age:female`=0"))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: gls(model = pp ~ age + age_sp45 + race + female + female:age +
##     female:age_sp45 + race:age + race:age_sp45, data = d2, weights = varFunc(~age))
##
## Linear Hypotheses:
##                          Estimate Std. Error z value Pr(>|z|)
## age + `age:female` == 0 -0.07579    0.07160  -1.058     0.29
## (Adjusted p values reported -- single-step method)
```

```r
se_female1 = 0.07160
-0.07579 + c(-1,1)*qnorm(0.975)*se_female1
```

95% CI for pp change rate of woman aged 20-45

```
## [1] -0.21612342  0.06454342
```

```r
library(multcomp)
summary(glht(model23, linfct = "`age` + `age_sp45`=0"))
```

pp change rate of male aged 45-80

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: gls(model = pp ~ age + age_sp45 + race + female + female:age +
##     female:age_sp45 + race:age + race:age_sp45, data = d2, weights = varFunc(~age))
##
## Linear Hypotheses:
##                      Estimate Std. Error z value Pr(>|z|)
## age + age_sp45 == 0  0.79152    0.06578   12.03   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

```r
se_man45 =  0.06578
0.79152 + c(-1,1)*qnorm(0.975)*se_man45
```

95% CI for pp change rate of male aged 45-80

```
## [1] 0.6625936 0.9204464
```

```r
library(multcomp)
summary(glht(model23, linfct = "`age`+`age_sp45`+`age:female` + `age_sp45:female`=0"))
```

pp change rate of female aged 45-80

```
##
##   Simultaneous Tests for General Linear Hypotheses
```

```
##
## Fit: gls(model = pp ~ age + age_sp45 + race + female + female:age +
##     female:age_sp45 + race:age + race:age_sp45, data = d2, weights = varFunc(~age))
##
## Linear Hypotheses:
##                                                         Estimate Std. Error
## age + age_sp45 + `age:female` + `age_sp45:female` == 0  0.94114    0.06437
##                                                         z value Pr(>|z|)
## age + age_sp45 + `age:female` + `age_sp45:female` == 0   14.62   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

```
se_woman45 =  0.06437
0.94114 + c(-1,1)*qnorm(0.975)*se_woman45
```

95% CI for pp change rate of female aged 45-80

```
## [1] 0.8149771 1.0673029
```

```
library(multcomp)
summary(glht(model23, linfct = "`age:female` + `age_sp45:female`=0"))
```

pp difference in change rate , female vs male

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: gls(model = pp ~ age + age_sp45 + race + female + female:age +
##     female:age_sp45 + race:age + race:age_sp45, data = d2, weights = varFunc(~age))
##
## Linear Hypotheses:
##                                        Estimate Std. Error z value Pr(>|z|)
## `age:female` + `age_sp45:female` == 0  0.14962    0.04849   3.085  0.00203 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

```
se_diff45 =  0.04849
0.14962 + c(-1,1)*qnorm(0.975)*se_diff45
```

95% CI for pp difference in change rate , female vs male

```
## [1] 0.05458135 0.24465865
```

Second Object:

Potential mediation effect of Height, BMI, weight

```
model23 = gls(pp ~ age + age_sp45 + race + female +
                  female:age + female:age_sp45 +
                  race:age + race:age_sp45,
```

```r
             weights = varFunc(~age),data = d2)
## Regress Height on Age
model23_h = gls(height ~ age + age_sp45 + race + female +
                  female:age + female:age_sp45 +
                  race:age + race:age_sp45,
             weights = varFunc(~age),data = d2)

summary(model23_h)
```

Check for Height

```
## Generalized least squares fit by REML
##   Model: height ~ age + age_sp45 + race + female + female:age + female:age_sp45 +      race:age + rac
##   Data: d2
##        AIC      BIC    logLik
##   34469.26 34534.4 -17224.63
##
## Variance function:
##  Structure: fixed weights
##  Formula: ~age
##
## Coefficients:
##                     Value Std.Error   t-value p-value
## (Intercept)     175.06598 1.3731623 127.49111  0.0000
## age              -0.07668 0.0396949  -1.93184  0.0534
## age_sp45          0.00011 0.0662787   0.00164  0.9987
## race              0.33250 0.4886432   0.68046  0.4962
## female          -13.64070 1.0439316 -13.06666  0.0000
## age:female        0.00793 0.0301454   0.26310  0.7925
## age_sp45:female  -0.02819 0.0495609  -0.56886  0.5695
## age:race          0.01597 0.0141740   1.12663  0.2600
## age_sp45:race    -0.02898 0.0244015  -1.18756  0.2351
##
##  Correlation:
##                 (Intr) age    ag_s45 race   female ag:fml ag_sp45:f age:rc
## age             -0.973
## age_sp45         0.836 -0.911
## race            -0.831  0.810 -0.702
## female          -0.461  0.446 -0.376  0.063
## age:female       0.447 -0.456  0.406 -0.060 -0.972
## age_sp45:female -0.382  0.412 -0.432  0.048  0.836 -0.912
## age:race         0.807 -0.831  0.765 -0.973 -0.060  0.058 -0.047
## age_sp45:race   -0.680  0.743 -0.837  0.822  0.046 -0.045  0.038    -0.898
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -7.11911890 -0.61710126  0.01701747  0.65048695  4.04596629
##
## Residual standard error: 1.111816
## Degrees of freedom: 4994 total; 4985 residual
```

Height does not change significant as age change, thus can not be a medicator. (p=0.0534 for age, 0.9987 for age__sp)

```r
## Regress Weight on Age
model23_w = gls(weight ~ age + age_sp45 + race + female +
                  female:age + female:age_sp45 +
                  race:age + race:age_sp45,
              weights = varFunc(~age),data = d2)

summary(model23_w)
```

Check for Height

```
## Generalized least squares fit by REML
##   Model: weight ~ age + age_sp45 + race + female + female:age + female:age_sp45 +      race:age + rac
##   Data: d2
##        AIC      BIC    logLik
##   45020.37 45085.51 -22500.19
##
## Variance function:
##  Structure: fixed weights
##  Formula: ~age
##
## Coefficients:
##                     Value Std.Error   t-value p-value
## (Intercept)      82.29053  3.956669 20.797934  0.0000
## age               0.30641  0.114378  2.678895  0.0074
## age_sp45         -0.62532  0.190977 -3.274297  0.0011
## race             -2.79633  1.407990 -1.986045  0.0471
## female          -11.44597  3.008014 -3.805159  0.0001
## age:female        0.02299  0.086862  0.264707  0.7912
## age_sp45:female  -0.01033  0.142806 -0.072349  0.9423
## age:race          0.02469  0.040841  0.604608  0.5455
## age_sp45:race    -0.03173  0.070311 -0.451230  0.6518
##
##  Correlation:
##                 (Intr) age    ag_s45 race   female ag:fml ag_sp45:f age:rc
## age             -0.973
## age_sp45         0.836 -0.911
## race            -0.831  0.810 -0.702
## female          -0.461  0.446 -0.376  0.063
## age:female       0.447 -0.456  0.406 -0.060 -0.972
## age_sp45:female -0.382  0.412 -0.432  0.048  0.836 -0.912
## age:race         0.807 -0.831  0.765 -0.973 -0.060  0.058 -0.047
## age_sp45:race   -0.680  0.743 -0.837  0.822  0.046 -0.045  0.038    -0.898
##
## Standardized residuals:
##        Min         Q1        Med          Q3         Max
## -2.6241988 -0.6590043 -0.1690055  0.4860081  6.9362646
##
## Residual standard error: 3.203619
## Degrees of freedom: 4994 total; 4985 residual
```

```r
## Add Weight to the model
model23_medw = gls(pp~ age + age_sp45 + weight + race + female +
```

```r
                  female:age + female:age_sp45 +
                  race:age + race:age_sp45,
              weights = varFunc(~age),data = d2)

coef(model23_medw)[c('age','age_sp45')]
```

Weight changes significantly as age changes. -> potential mediator

```
##        age   age_sp45
## -0.3217654  1.1359416
```

```r
coef(model23)[c('age','age_sp45')]
```

```
##        age   age_sp45
## -0.2999961  1.0915145
```

The coefficient of age changes from -0.32 to -0.30 when comparing before and after adding weight as a covariate. Weight can be a potential mediator with weak effects.
Check for BMI

```r
## Regress Weight on Age
model23_b = gls(bmi ~ age + age_sp45 + race + female +
                  female:age + female:age_sp45 +
                  race:age + race:age_sp45,
              weights = varFunc(~age),data = d2)

summary(model23_b)
```

```
## Generalized least squares fit by REML
##   Model: bmi ~ age + age_sp45 + race + female + female:age + female:age_sp45 +      race:age + race:a
##   Data: d2
##        AIC      BIC    logLik
##   33980.48 34045.62 -16980.24
##
## Variance function:
##  Structure: fixed weights
##  Formula: ~age
##
## Coefficients:
##                     Value Std.Error   t-value p-value
## (Intercept)     26.962529 1.3074666 20.621964  0.0000
## age              0.128748 0.0377958  3.406411  0.0007
## age_sp45        -0.209435 0.0631078 -3.318691  0.0009
## race            -1.063680 0.4652652 -2.286180  0.0223
## female          -0.165125 0.9939872 -0.166124  0.8681
## age:female       0.032250 0.0287031  1.123553  0.2613
## age_sp45:female -0.034782 0.0471898 -0.737061  0.4611
## age:race        -0.000047 0.0134958 -0.003493  0.9972
## age_sp45:race    0.002140 0.0232341  0.092118  0.9266
##
##  Correlation:
##                 (Intr) age    ag_s45 race   female ag:fml ag_sp45:f age:rc
## age             -0.973
## age_sp45         0.836 -0.911
## race            -0.831  0.810 -0.702
## female          -0.461  0.446 -0.376  0.063
```

```
## age:female       0.447 -0.456  0.406 -0.060 -0.972
## age_sp45:female -0.382  0.412 -0.432  0.048  0.836 -0.912
## age:race         0.807 -0.831  0.765 -0.973 -0.060  0.058 -0.047
## age_sp45:race   -0.680  0.743 -0.837  0.822  0.046 -0.045  0.038    -0.898
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -2.4320238 -0.6408495 -0.1366487  0.4825545  7.1581511
##
## Residual standard error: 1.058624
## Degrees of freedom: 4994 total; 4985 residual
```

```
## Add Weight to the model
model23_medb = gls(pp~ age + age_sp45 + bmi + race + female +
                  female:age + female:age_sp45 +
                  race:age + race:age_sp45,
              weights = varFunc(~age),data = d2)

coef(model23_medb)[c('age','age_sp45')]
```

BMI changes significantly as age changes -> potential mediator

```
##        age   age_sp45
## -0.3310156  1.1419741
```

```
coef(model23)[c('age','age_sp45')]
```

```
##        age   age_sp45
## -0.2999961  1.0915145
```

The coefficient of age changes from -0.33 to -0.30 when comparing before and after adding BMI as a covariate. BMI are potential mediators with weak effects.

Conclusion

Male's pulse pressure decreases between 20 to 45 years of age while female's pulse pressure remains constant. After 45 years of age, pulse pressures increase as age increases in all sexes and all races. Females have statistically higher pulse pressure at a younger age as compared to males of the same age and race but will have a lower pulse pressure than males as they age. The difference between races is not significant, adjusting for sex and age. Weight and BMI are mediators with weak effects.

Discussion of limitations

Our final model contains 9 covariates, including two with non-significant coefficients, and may overfit the data. Race is modeled as a nominal variable, assuming the differences between two adjacent race groups are the same while holding age and sex constant. This assumption is likely violated, thus challenging the validity of the model. Whether weight is a mediator between the relationship of pulse and pressure can not be simply determined by comparing coefficients change. The Sobel test (Sobel, 1982) could be applied.