BST 210 Homework 1

Wenjie Gu

Question 1

#1(a)

```
dat <- read.csv("~/Documents/Class Info/Harvard HSPH/Fall2019/BST210/Homework#1/Data and Programs/SCCS2
head(dat)
```

```
##   X caseID ethnic height weight waist   hip   tc   tg  hdl  ldl diabetes
## 1 1     23      1 177.35   58.0 69.50 92.00 4.65 0.54 1.07 3.34        0
## 2 2     39      1 173.35   64.1 75.55 95.50 6.26 1.20 1.20 4.52        0
## 3 3     42      1 177.10   52.2 63.00 89.00 4.81 0.79 1.64 2.81        0
## 4 4     51      1 167.80   59.7 72.00 92.75 5.06 2.19 1.13 2.94        0
## 5 5     68      1 172.30   65.7 72.50 99.00 4.85 1.35 1.15 3.09        0
## 6 6     86      1 172.00   53.3 64.50 89.00 3.86 1.69 1.42 1.67        0
##   hypertension educ drink smoke gender alcohol      age ihd Dummy2
## 1            1    3     1     1      0       2 35.97262   0      2
## 2            1    2     1     1      0       2 31.66324   0      0
## 3            1    5     0     1      0       1 29.96304   0      2
## 4            1    5     1     1      0       2 30.69678   0      0
## 5            1    3     1     1      0       2 27.97810   0      2
## 6            1    3     0     1      0       1 25.97673   0      0
```

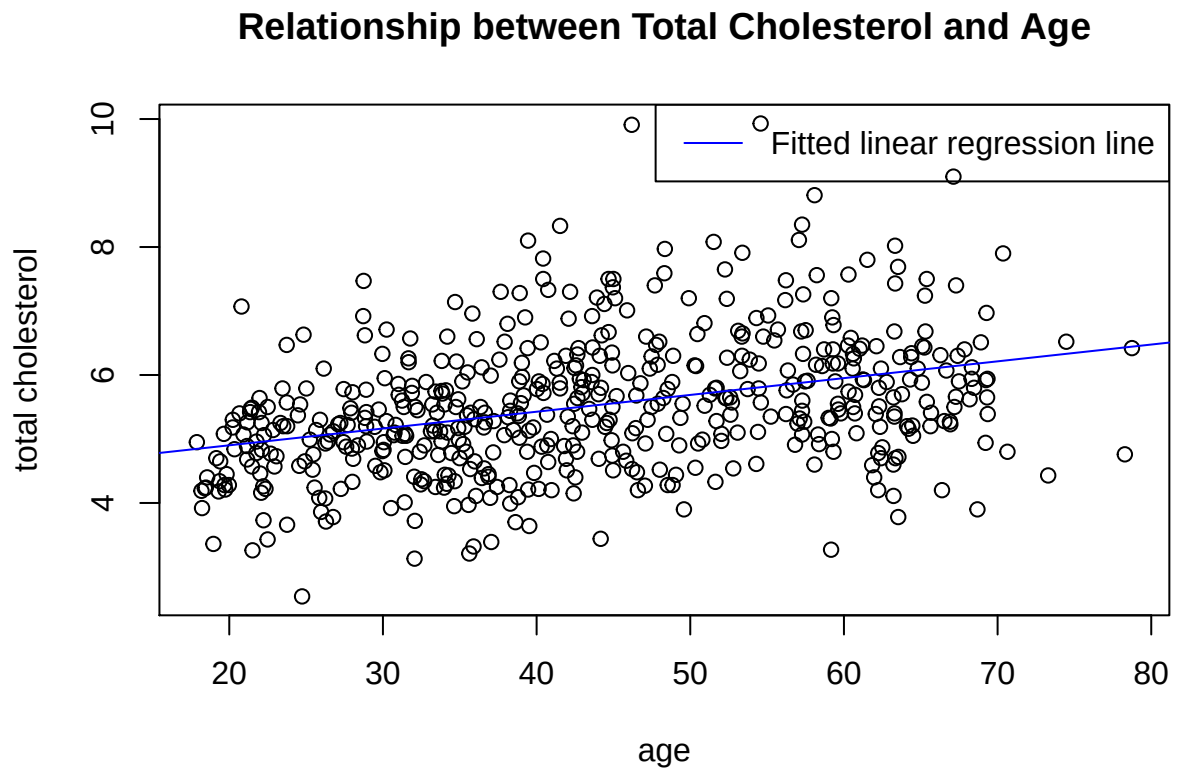
```
age.tc = lm(tc~age, data = dat)
summary(age.tc)
```

```
##
## Call:
## lm(formula = tc ~ age, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6589 -0.6383 -0.0557  0.5009  4.3216
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.376191   0.135848  32.214   <2e-16 ***
## age         0.026243   0.002966   8.849   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9801 on 525 degrees of freedom
## Multiple R-squared:  0.1298, Adjusted R-squared:  0.1281
## F-statistic: 78.31 on 1 and 525 DF,  p-value: < 2.2e-16
```

```
confint(age.tc)
```

```
##                 2.5 %     97.5 %
## (Intercept) 4.1093176 4.64306405
## age         0.0204172 0.03206905
```

```r
par(mfrow = c(1,1))
plot(dat$tc~dat$age, xlab = "age", ylab = "total cholesterol", type = "p", main="Relationship between T
abline(age.tc, col = "blue")
legend("topright", "Fitted linear regression line", col = "blue", lwd = 1)
```




```r
cor.test(dat$age, dat$tc)
```

```
##
##  Pearson's product-moment correlation
##
## data:  dat$age and dat$tc
## t = 8.8491, df = 525, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2835867 0.4323806
## sample estimates:
##      cor
## 0.360273
```

Simple linear regression model:

intercept: 4.376, p-value < 2e-16, 95% conf interval: (4.109, 4.643)

slope: 0.026, p-value < 2e-16, 95% conf interval: (0.020, 0.032)

Interpretation: For a change of an age decade, the mean total cholesterol level increases by 2.6 units. The p-value is smaller than 0.05, indicating the statistical significance of the slope value.

Pearson's correlation test:

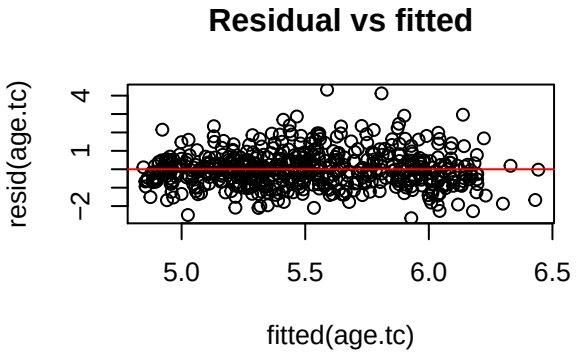
p-value < 2.2e-16, 95% confidence interval: (0.2836,0.4324), sample estimates: 0.36, the result from pearson's correlation test indicates that there is a weak positive correlation between age and total cholesterol(r = 0.36)

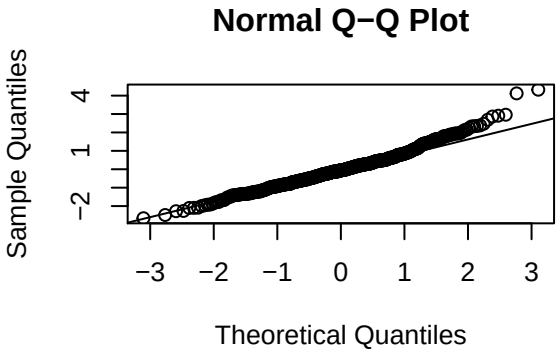
Comparing simple linear regression model with Pearson's test:

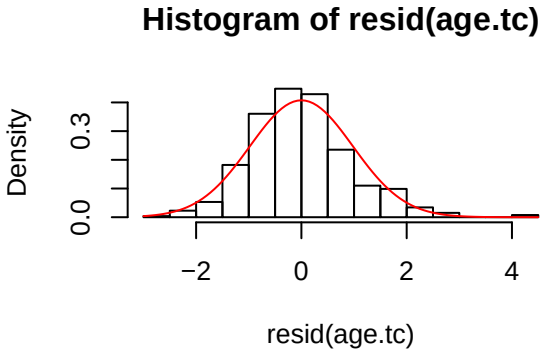
Multiple R-squared value from linear regression model is 0.1298, which means that only 13% of the observed total cholesterol data can be explained by the linear regression of total cholesterol on age, indicating that the linear model is not a strong fit for the date. The Pearson's correlation coefficient corroborates this suggestion, suggesting the linear association exists but not strong.

#1(b)

```
par(mfrow = c(2,2))
plot(resid(age.tc)~fitted(age.tc), main = "Residual vs fitted")
abline(h = 0, col = "red")
qqnorm(resid(age.tc))
qqline(resid(age.tc))
hist(resid(age.tc),prob = TRUE)
m = mean(resid(age.tc))
std = sqrt(var(resid(age.tc)))
curve(dnorm(x,mean = m, sd = std),add = TRUE, col = "red")
```



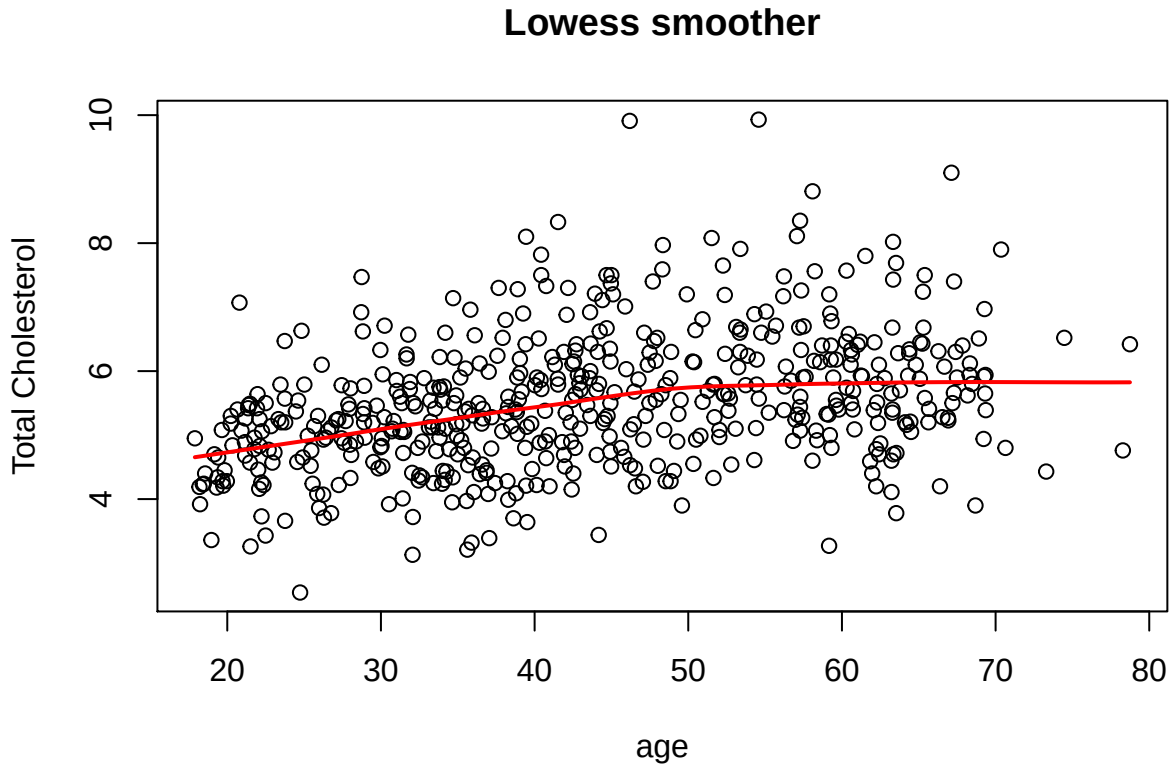




No, the residuals are not normally distributed according to the normal Q-Q plot. The right tail of the sample quantile deviates from the reference line. Also, according to the histogram of the residual, the residual data is slightly skewed to the right. There are a few outliers in the residual plot that might contribute to the skewed distribution.

#1(c)

```r
par(mfrow = c(1,1))
plot(dat$age,dat$tc, xlab = "age", ylab = "Total Cholesterol", main = "Lowess smoother")
lines(loess.smooth(dat$age,dat$tc), col = 'red',lty = 1, lwd = 2)
```



The Lowess smoothed curve suggests there might be a nonlinear effect of age on the prediction of total cholesterol.

```r
age_2.tc = lm(tc~ age + I(age^2),data = dat)
summary(age_2.tc)
```

```
##
## Call:
## lm(formula = tc ~ age + I(age^2), data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6542 -0.6410 -0.0461  0.5151  4.1698
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.0658041  0.3907005   7.847 2.41e-14 ***
## age          0.0920305  0.0186508   4.934 1.08e-06 ***
## I(age^2)    -0.0007389  0.0002069  -3.572 0.000387 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9693 on 524 degrees of freedom
## Multiple R-squared:  0.1505, Adjusted R-squared:  0.1472
## F-statistic: 46.41 on 2 and 524 DF,  p-value: < 2.2e-16
```

Since the p-value for age^2 coefficient term is 0.000387, which <005, there is a statistically significant evidence suggesting a nonlinear effect of age to predict total cholesterol.

Question 2

#2(a)

```
t.test(dat$tc ~ dat$gender, var.equal = T)
```

```
##
##  Two Sample t-test
##
## data:  dat$tc by dat$gender
## t = -0.63874, df = 525, p-value = 0.5233
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2392340  0.1218354
## sample estimates:
## mean in group 0 mean in group 1
##        5.490799        5.549498
```

The two sample t-test t value is -0.63874, which gives a p-value of 0.523 => We fail to reject null hypothesis that there's no difference in mean total cholesterol between male and female.

i.e. There's no evidence suggesting a statistically significant difference in mean tc between males and females

```
gender.tc = lm(tc~gender, data= dat)
summary(gender.tc)
```

```
##
## Call:
## lm(formula = tc ~ gender, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9508 -0.6908 -0.0908  0.6242  4.4392
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.49080    0.06189  88.722   <2e-16 ***
## gender       0.05870    0.09190   0.639    0.523
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.05 on 525 degrees of freedom
## Multiple R-squared:  0.0007765,  Adjusted R-squared:  -0.001127
## F-statistic: 0.408 on 1 and 525 DF,  p-value: 0.5233
```

from the linear regression model, p-value of gender coefficient is 0.523, equivalent to p-value in the t-test

from the linear regression model, the coefficients for gender is 0.05870, equivalent to the mean difference of total cholesterol between female and male

y intercept value in the linear regression model is 5.49, equivalent to tc mean of females

#2(b)

```
age_2_gender.tc = lm(tc~age + I(age^2) + gender, data = dat)
summary(age_2_gender.tc)
```

```
##
```

```
## Call:
## lm(formula = tc ~ age + I(age^2) + gender, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6152 -0.6114 -0.0555  0.5250  4.1205
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.0390448  0.3913591   7.765 4.32e-14 ***
## age          0.0909455  0.0186723   4.871 1.48e-06 ***
## I(age^2)    -0.0007241  0.0002073  -3.494 0.000517 ***
## gender       0.0944912  0.0851936   1.109 0.267882
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9691 on 523 degrees of freedom
## Multiple R-squared:  0.1525, Adjusted R-squared:  0.1476
## F-statistic: 31.36 on 3 and 523 DF,  p-value: < 2.2e-16
```

The coefficients of age and age-square change from 0.09203 to 0.0909455, and from -0.0007389 to -0.0007241, respectively. The adjusted analysis does not give a significant different result than the regression model using both linear and quadratic age. Therefore, we do not consider gender as a confounder of the effect of linear and quadratic age on tc.

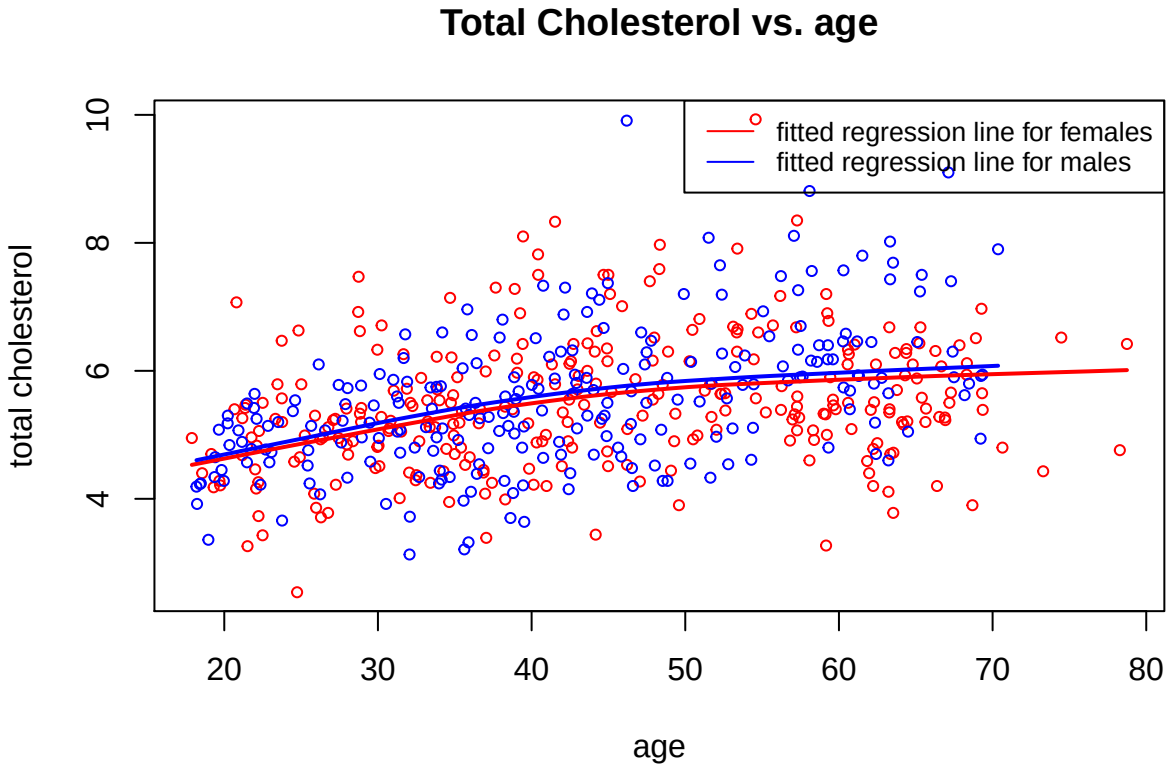
Gender is not an independent predictor as the p-value for gender coefficients > 0.05.

#2(c)

```r
par(mfrow = c(1,1))

color <- function(gender){
  color = c(0, length(gender))
  for (i in 1:length(gender)){
    if (gender[i] == 0) {
      color[i] = "red"
      }
    else {
      color[i] = "blue"
    }
  }
  return (color)
}

plot(dat$age, dat$tc, xlab = "age", ylab = "total cholesterol", col = color(dat$gender), main = "Total
lines(loess.smooth(dat$age[which(dat$gender == 0)], fitted(age_2_gender.tc)[which(dat$gender == 0)]), c
lines(loess.smooth(dat$age[which(dat$gender == 1)], fitted(age_2_gender.tc)[which(dat$gender == 1)]), c
legend("topright",c("fitted regression line for females", "fitted regression line for males"),col = c("
```



The two fitted regression lines are parallel to each other.

#2(d)

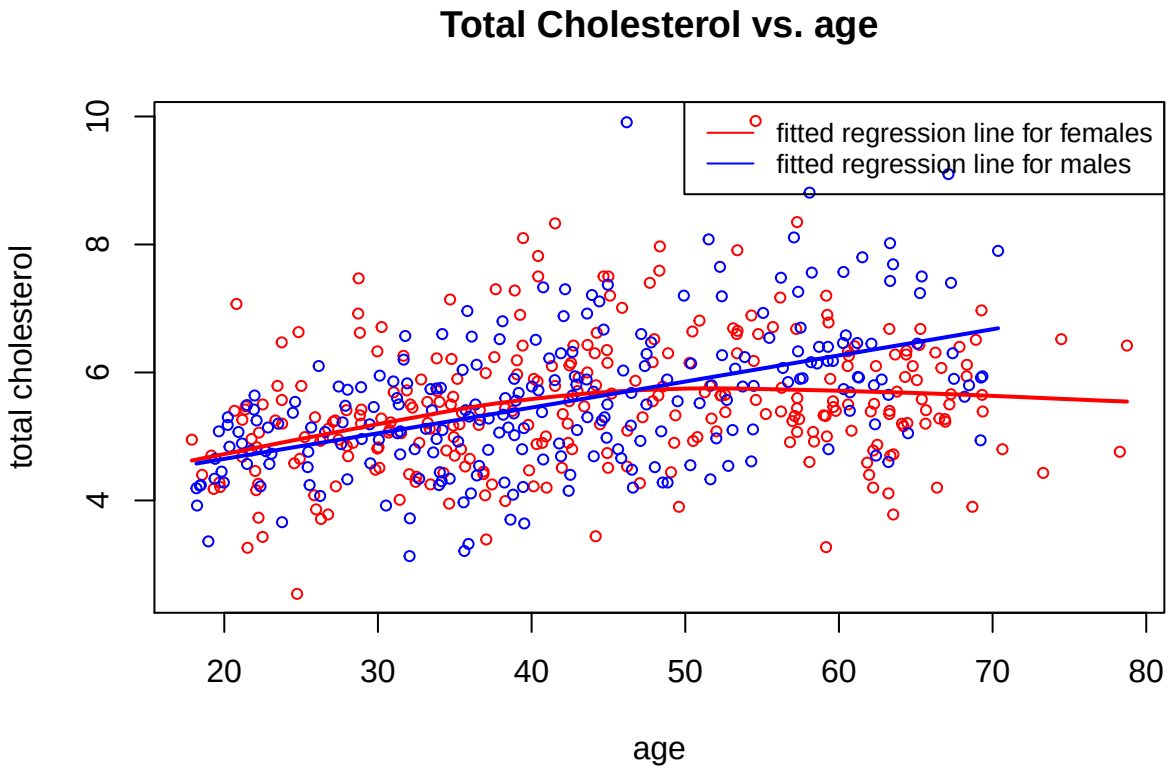
```
full_int = lm(tc ~ age + I(age^2) + gender + age*gender + I(age^2)*gender, data = dat)
summary(full_int)
```

```
##
## Call:
## lm(formula = tc ~ age + I(age^2) + gender + age * gender + I(age^2) *
##     gender, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4971 -0.6287 -0.0584  0.5450  4.2082
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.7876252  0.5259996   5.300 1.72e-07 ***
## age               0.1137966  0.0247662   4.595 5.44e-06 ***
## I(age^2)         -0.0010722  0.0002699  -3.973 8.10e-05 ***
## gender            1.0753560  0.7772306   1.384  0.16708
## age:gender       -0.0747839  0.0373402  -2.003  0.04572 *
## I(age^2):gender   0.0010894  0.0004176   2.609  0.00935 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.953 on 521 degrees of freedom
## Multiple R-squared:  0.1836, Adjusted R-squared:  0.1757
## F-statistic: 23.43 on 5 and 521 DF,  p-value: < 2.2e-16
```

P-value for the coefficient of interaction between gender and age is 0.04572, and for interaction between gender and quadratic age is 0.00935, both of which < 0.05, indicating that both interactions are statistically significant.

Thus gender is an effect modifier of both the effect of linear and quadratic age on total cholesterol.

```
plot(dat$age, dat$tc, xlab = "age", ylab = "total cholesterol", col = color(dat$gender), main = "Total C
lines(loess.smooth(dat$age[which(dat$gender == 0)], fitted(full_int)[which(dat$gender == 0)]), col = "re
lines(loess.smooth(dat$age[which(dat$gender == 1)], fitted(full_int)[which(dat$gender == 1)]), col = "bl
legend("topright",c("fitted regression line for females", "fitted regression line for males"),col = c("r
```



The curves for females and males are no longer parallel. After the age of 50, the difference of predicted total cholesterol between females and males enlarges as age increases.

#2(e)

```
r.square = c( full = summary(full_int)$r.squared,
              age = summary(age.tc)$r.squared,
              age_2 = summary(age_2.tc)$r.squared,
              age.gender = summary(age_2_gender.tc)$r.squared,
              gender = summary(gender.tc)$r.squared)
r.square
```

```
##         full          age        age_2   age.gender       gender
## 0.1835561450 0.1297966337 0.1504793911 0.1524729114 0.0007765133
```

Full interaction model has the highest r-squared value.

```
adj.r.square = c (full = summary(full_int)$adj.r.squared,
                  age = summary(age.tc)$adj.r.squared,
                  age_2 = summary(age_2.tc)$adj.r.squared,
                  age.gender = summary(age_2_gender.tc)$adj.r.squared,
```

```
                  gender = summary(gender.tc)$adj.r.squared)
adj.r.square
```

```
##        full         age       age_2  age.gender      gender
##  0.17572079  0.12813910  0.14723695  0.14761138 -0.00112677
```

Full interaction model has the highest adjusted r-squared value.

```
root_MSE = c (full = summary(full_int)$sigma,
                  age = summary(age.tc)$sigma,
                  age_2 = summary(age_2.tc)$sigma,
                  age.gender = summary(age_2_gender.tc)$sigma,
                  gender = summary(gender.tc)$sigma)
root_MSE
```

```
##       full        age      age_2 age.gender     gender
##  0.9529996  0.9801198  0.9693257  0.9691129  1.0502679
```

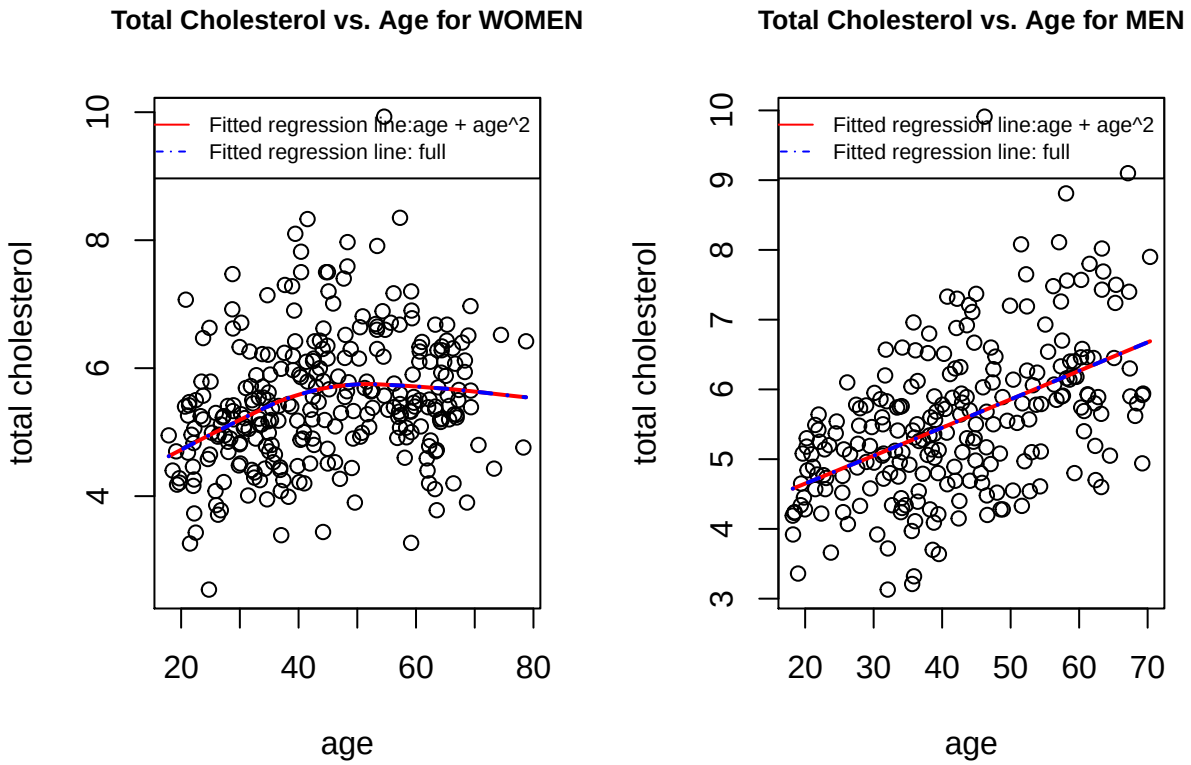
Full interaction has the smallest square-rooted MSE

#2(f)

```
dat_woman = dat[dat$gender == 0, ]
dat_man = dat[dat$gender == 1, ]
age_2_woman = lm(tc~age + I(age^2), data = dat_woman)
age_2_man = lm(tc~age + I(age^2), data = dat_man)
par(mfrow = c(1,2))

# plot regression for women
plot(dat_woman$age, dat_woman$tc, xlab = "age", ylab = "total cholesterol", main = "Total Cholesterol vs
lines(loess.smooth(dat_woman$age, fitted(age_2_woman)), col= "red",lwd = 2)
lines(loess.smooth(dat$age[which(dat$gender == 0)], fitted(full_int)[which(dat$gender == 0)]), col = "bl
legend("topright", c("Fitted regression line:age + age^2", "Fitted regression line: full"), col = c("red

# plot regression for men
plot(dat_man$age, dat_man$tc, xlab = "age", ylab = "total cholesterol", main = "Total Cholesterol vs. Ag
lines(loess.smooth(dat_man$age, fitted(age_2_man)), col= "red",lwd = 2)
lines(loess.smooth(dat$age[which(dat$gender == 1)], fitted(full_int)[which(dat$gender == 1)]), col = "bl
legend("topright", c("Fitted regression line:age + age^2", "Fitted regression line: full"), col = c("red
```



Both models give the same prediction curve for women and for men separately.

I would choose the full interaction model. In this case, there are only two categories in gender variable, so we only need to separately generate two linear regression models for males and females. But in other cases, the categorical variable may have more categories, which make it harder to generate a separate linear regression model for each category.

The full interaction model incorporates all possible values for the categorical variable and generate a more comprehensive model.

Question 3

#3(a)

```
dat$BMI = dat$weight / (dat$height/100)^2
dat$BMI_categorical[dat$BMI < 18.5] = 0
dat$BMI_categorical[(dat$BMI >= 18.5) & (dat$BMI < 25)] = 1
dat$BMI_categorical[(dat$BMI >= 25) & (dat$BMI < 30)] = 2
dat$BMI_categorical[dat$BMI >= 30] = 3

newdat = dat[order(dat$BMI),]
BMI_cont.tc = lm(tc~BMI, data = newdat)
BMI_cat.tc = lm(tc~BMI_categorical, data = newdat)

summary(BMI_cont.tc)
```

```
##
## Call:
## lm(formula = tc ~ BMI, data = newdat)
##
## Residuals:
```

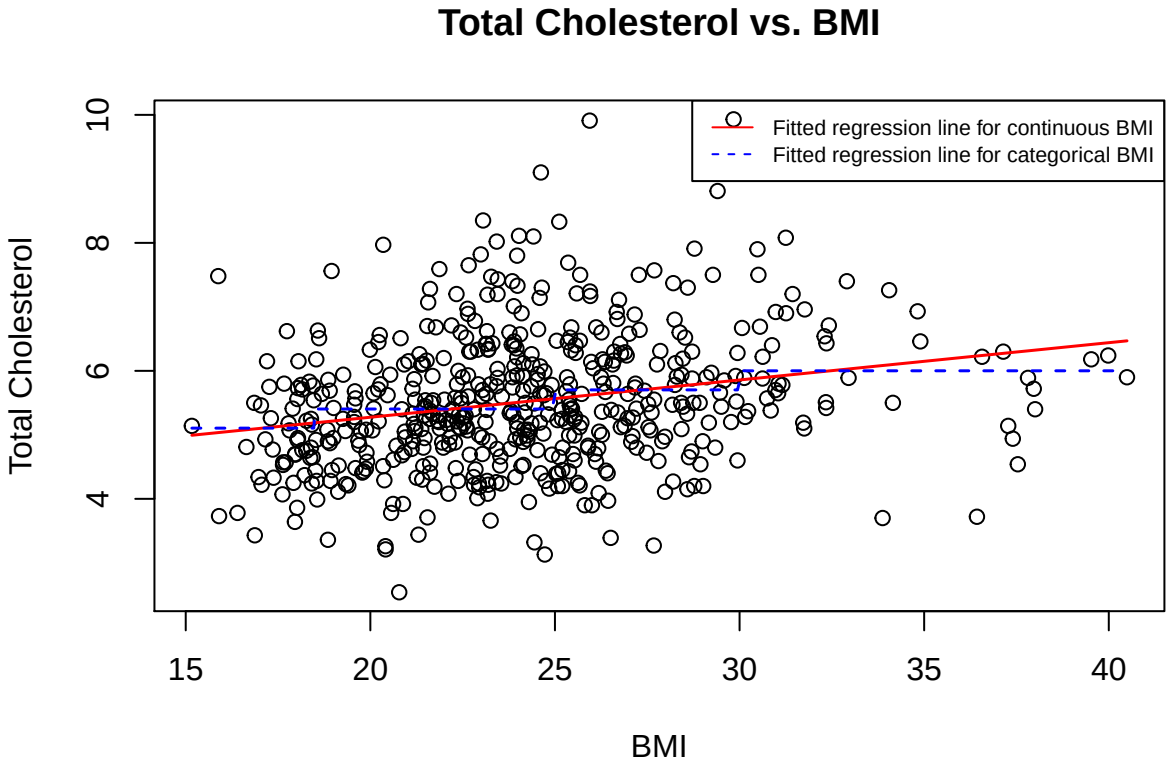
```
##     Min      1Q  Median      3Q     Max
## -2.7806 -0.6333 -0.1273  0.5735  4.2889
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.11005    0.25007  16.435  < 2e-16 ***
## BMI          0.05825    0.01019   5.719  1.8e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.019 on 525 degrees of freedom
## Multiple R-squared:  0.05864,    Adjusted R-squared:  0.05685
## F-statistic:  32.7 on 1 and 525 DF,  p-value: 1.803e-08
```

```
summary(BMI_cat.tc)
```

```
##
## Call:
## lm(formula = tc ~ BMI_categorical, data = newdat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8635 -0.6378 -0.1110  0.5972  4.2278
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.10475    0.09102  56.082  < 2e-16 ***
## BMI_categorical  0.29874    0.05743   5.202 2.83e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.025 on 525 degrees of freedom
## Multiple R-squared:  0.04902,    Adjusted R-squared:  0.04721
## F-statistic: 27.06 on 1 and 525 DF,  p-value: 2.83e-07
```

Comparing to continuous BMI model, categorical BMI model has a larger residual standard error, which indicates a higher residual value.

```
par(mfrow = c(1,1))
plot(newdat$BMI, newdat$tc, main = "Total Cholesterol vs. BMI", xlab = "BMI", ylab = "Total Cholesterol
lines(newdat$BMI, fitted(BMI_cont.tc), col = "red", lwd = 1.5)
lines(newdat$BMI, fitted(BMI_cat.tc), col = "blue", lwd = 1.5, lty = "dashed")
legend("topright", c("Fitted regression line for continuous BMI", "Fitted regression line for categorica
```



I prefer continuous BMI model because it has smaller residual value which means a relatively better fit. When transforming a continuous variable to a categorical variable, there will be loss of information which will leads to the inprecision of the prediction.

#3(b)

```
BMI_quad = lm(tc~BMI + I(BMI^2), data = newdat)
summary(BMI_quad)
```

```
##
## Call:
## lm(formula = tc ~ BMI + I(BMI^2), data = newdat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7664 -0.6523 -0.1061  0.5651  4.2049
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.619210   0.996621   1.625  0.10483
## BMI          0.257627   0.077911   3.307  0.00101 **
## I(BMI^2)    -0.003859   0.001495  -2.581  0.01012 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.014 on 524 degrees of freedom
## Multiple R-squared:  0.07046,    Adjusted R-squared:  0.06691
## F-statistic: 19.86 on 2 and 524 DF,  p-value: 4.859e-09
```

Adding a quadratic BMI term makes the model slightly better, R^2 improves from 0.05685 to 0.06691.

#3(c)

```
BMI_multi = lm(tc ~ age + I(age^2) + gender + BMI + I(BMI^2) + age*gender + I(age^2)*gender , data = da
summary(BMI_multi)
```

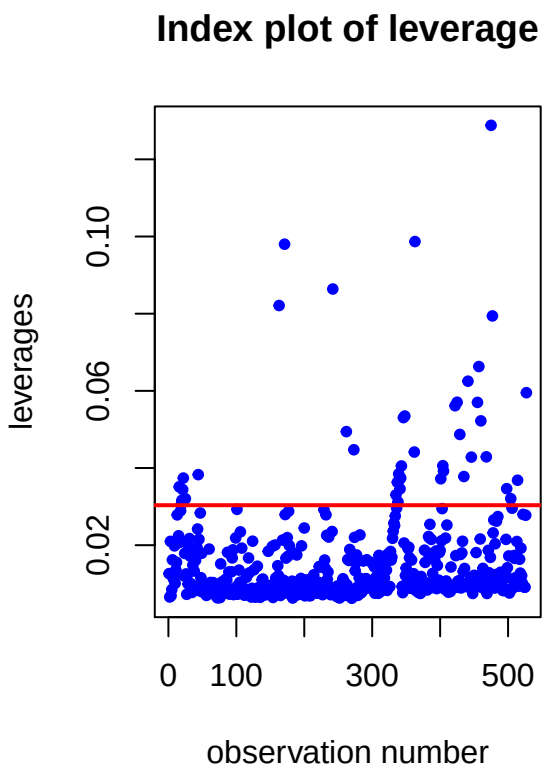
```
##
## Call:
## lm(formula = tc ~ age + I(age^2) + gender + BMI + I(BMI^2) +
##     age * gender + I(age^2) * gender, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5998 -0.5756 -0.0753  0.5622  4.1314
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.5236562  1.0452495   0.501 0.616593
## age               0.1008759  0.0248775   4.055 5.79e-05 ***
## I(age^2)         -0.0009564  0.0002700  -3.542 0.000433 ***
## gender            1.2081308  0.7719942   1.565 0.118205
## BMI               0.1813440  0.0751599   2.413 0.016177 *
## I(BMI^2)         -0.0029940  0.0014347  -2.087 0.037392 *
## age:gender       -0.0797282  0.0370736  -2.151 0.031974 *
## I(age^2):gender   0.0011347  0.0004149   2.735 0.006447 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9449 on 519 degrees of freedom
## Multiple R-squared:  0.2004, Adjusted R-squared:  0.1896
## F-statistic: 18.58 on 7 and 519 DF,  p-value: < 2.2e-16
```

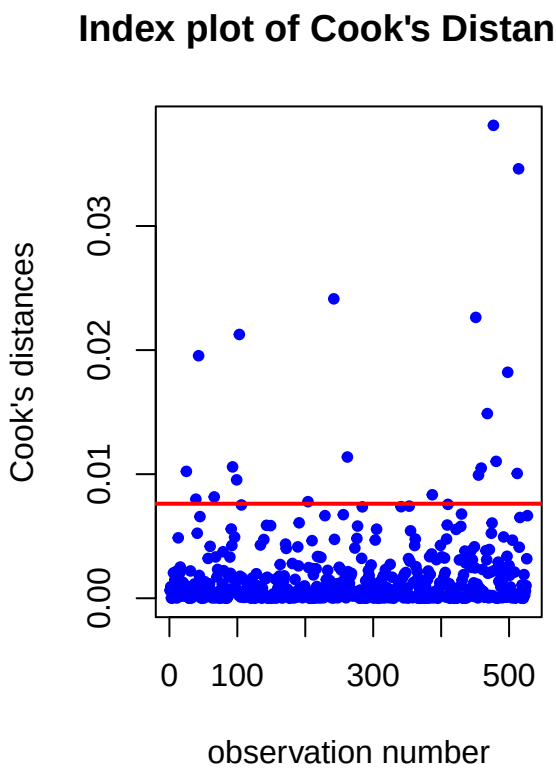
I choose the model with covariates: age, age^2, gender, BMI, BMI^2, and adding the effect modification of age&gender, age^2&gender. Because this model gives the least residual standard erro and the largest adjusted R-squared value among the models that I tried. Also the p-values for the interaction between BMI and other covariates are not significant, indicating BMI is not an effect modifier of the other covariates.

#3(d)

```
par(mfrow = c(1,2))
h = hat(model.matrix(BMI_multi))
plot(h, col="blue",
     main="Index plot of leverage",
     ylab="leverages", xlab="observation number", pch=20)
abline(h=2*(8)/nrow(dat), col="red", lwd=2)


cook <- cooks.distance(BMI_multi)
plot(cook, col="blue",
     main="Index plot of Cook's Distances",
     ylab="Cook's distances", xlab="observation number", pch=20)
abline(h=(4/(nrow(dat)-2)), col="red", lwd=2)
```






```
nrow(dat[h > ((2*(8))/nrow(dat)),]) # number of high leverage
```

```
## [1] 43
```

```
nrow(dat[cook > 4/(nrow(dat)-2),]) # number of high cook's distance
```

```
## [1] 20
```

```
nrow(dat[cook > 4/(nrow(dat)-2) & h > ((2*(8))/nrow(dat)),]) # number of high leverage $ high cook's di
```

```
## [1] 8
```

```
newdat2 = dat[!(cook > 4/(nrow(dat)-2) & h > ((2*(8))/nrow(dat))),]
BMI_multi_adj = lm(tc ~ age + I(age^2) + gender + BMI +  I(BMI^2) + age*gender + I(age^2)*gender, data =
summary(BMI_multi_adj)
```

```
##
## Call:
## lm(formula = tc ~ age + I(age^2) + gender + BMI + I(BMI^2) +
##     age * gender + I(age^2) * gender, data = newdat2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5917 -0.5605 -0.0896  0.5393  4.1487
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.7457494  1.0611656   0.703  0.48252
## age              0.1137647  0.0255104   4.460 1.01e-05 ***
## I(age^2)        -0.0011104  0.0002796  -3.971 8.17e-05 ***
## gender           1.3676349  0.7799332   1.754  0.08011 .
## BMI              0.1395277  0.0763945   1.826  0.06837 .
```

```
## I(BMI^2)        -0.0020751  0.0014607  -1.421  0.15603
## age:gender      -0.0862955  0.0378037  -2.283  0.02286 *
## I(age^2):gender  0.0011836  0.0004271   2.771  0.00579 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9289 on 511 degrees of freedom
## Multiple R-squared:  0.1965, Adjusted R-squared:  0.1855
## F-statistic: 17.86 on 7 and 511 DF,  p-value: < 2.2e-16
```

```
summary(BMI_multi)
```

```
##
## Call:
## lm(formula = tc ~ age + I(age^2) + gender + BMI + I(BMI^2) +
##     age * gender + I(age^2) * gender, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5998 -0.5756 -0.0753  0.5622  4.1314
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.5236562  1.0452495   0.501 0.616593
## age              0.1008759  0.0248775   4.055 5.79e-05 ***
## I(age^2)        -0.0009564  0.0002700  -3.542 0.000433 ***
## gender           1.2081308  0.7719942   1.565 0.118205
## BMI              0.1813440  0.0751599   2.413 0.016177 *
## I(BMI^2)        -0.0029940  0.0014347  -2.087 0.037392 *
## age:gender      -0.0797282  0.0370736  -2.151 0.031974 *
## I(age^2):gender  0.0011347  0.0004149   2.735 0.006447 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9449 on 519 degrees of freedom
## Multiple R-squared:  0.2004, Adjusted R-squared:  0.1896
## F-statistic: 18.58 on 7 and 519 DF,  p-value: < 2.2e-16
```

43 data points have high leverage, 20 data points have high Cook's distance. 8 of them have both high leverage and high Cook's distance. After dropping these 8 points from the dataset, the coefficients of the covariates change, adjusted R-squared decreases, and the residual standard error decreases.