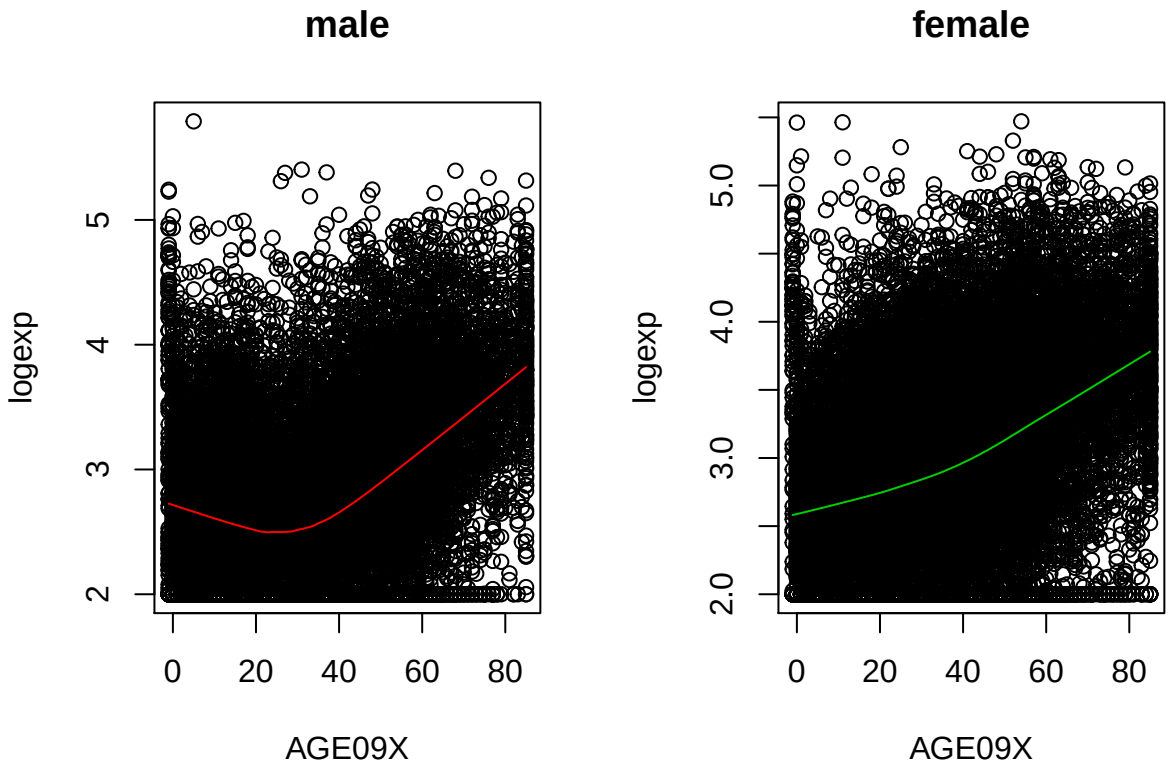
Homework7

ShanZhong

November 27, 2018

```
load(url("http://people.stat.sc.edu/hoyen/Stat704/Data/h129.RData"))

attach(h129)
dat<-data.frame(TOTEXP09, SEX,
                RACEX, ASTHDX, DOBMM,
                DOBYY,ADSMOK42, CALUNG, CHDDX,
                EDUCYR, POVCAT09, MARRY09X,
                SEATBE53, PERWT09F, AGE09X)

detach(h129)
rm(h129)

dat<-data.table(dat)
```

```
dat$logexp<-log(dat$TOTEXP09+100,base=10)

data<-dat[,c("AGE09X","logexp","SEX")]

par(mfrow = c(1,2))
plot(data[which(SEX==1)][,-c("SEX")],main="male")
lines(lowess(data[which(SEX==1)][,-c("SEX")]), col = 2)
plot(data[which(SEX==2)][,-c("SEX")],main="female")
lines(lowess(data[which(SEX==2)][,-c("SEX")]), col = 3)
```



Question2, we see there is an similar linear relationship between age and medical exp, as the expense increase as people getting older starting from 25 years old, also younger people less than 20 years old would have higher medical expense. Male are most healthy at 25 years old, female do not show decrease in medical exp.

spline model

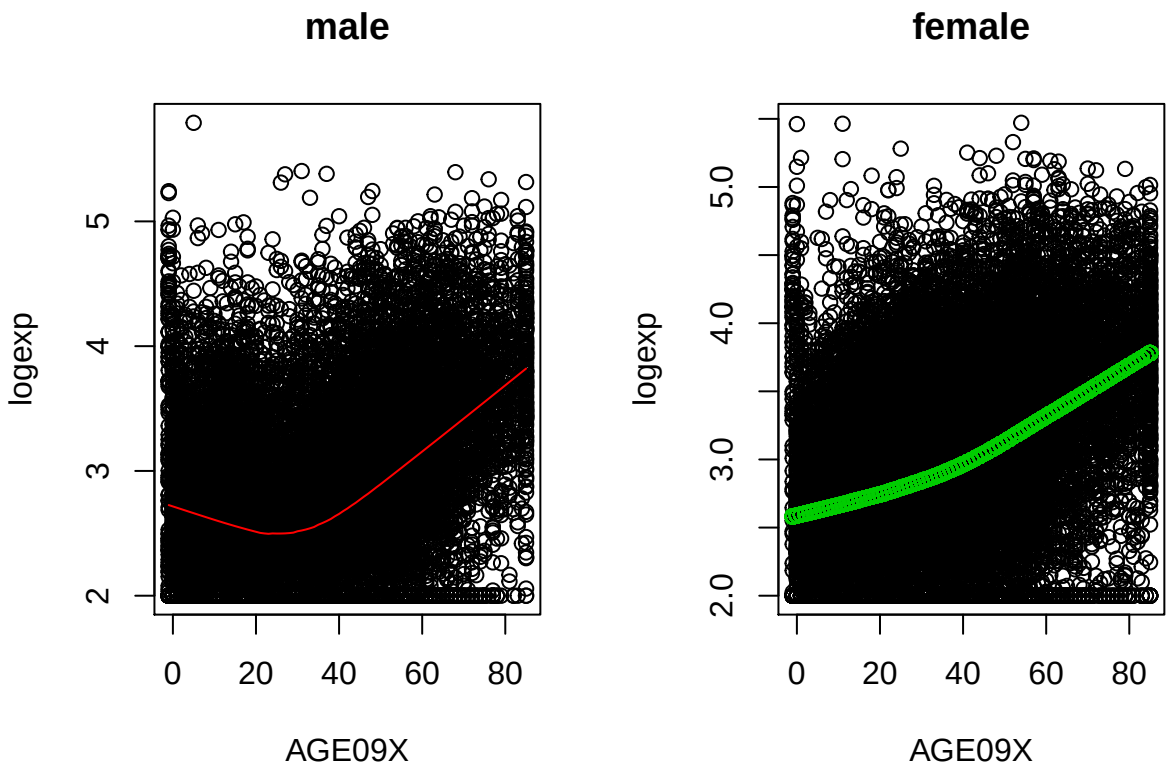
```
dat$logexp<-log(dat$TOTEXP09+100,base=10)

data<-dat[,c("AGE09X","logexp","SEX")]

data$age25<-ifelse(data$AGE09X > 25, data$AGE09X - 25, 0)
data$age40<-ifelse(data$AGE09X > 40, data$AGE09X - 40, 0)

summary(lm(logexp ~ AGE09X + age25 + age40 + SEX, data=data ))
```

```
##
## Call:
## lm(formula = logexp ~ AGE09X + age25 + age40 + SEX, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8303 -0.5566 -0.0556  0.4421  3.1760
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  2.4616470  0.0140672 174.992  < 2e-16 ***
## AGE09X      -0.0035939  0.0006161  -5.834 5.47e-09 ***
## age25        0.0158164  0.0013405  11.799  < 2e-16 ***
## age40        0.0085418  0.0011369   7.513 5.89e-14 ***
## SEX          0.1703845  0.0069450  24.534  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6647 on 36850 degrees of freedom
## Multiple R-squared:  0.1712, Adjusted R-squared:  0.1711
## F-statistic:  1903 on 4 and 36850 DF,  p-value: < 2.2e-16
```

Question 3

```
dat$logexp<-log(dat$TOTEXP09+100,base=10)

data<-dat[,c("AGE09X","logexp","SEX")]

par(mfrow = c(1,2))
plot(data[which(SEX==1)][,-c("SEX")],main="male")
lines(lowess(data[which(SEX==1)][,-c("SEX")]), col = 2)
plot(data[which(SEX==2)][,-c("SEX")],main="female")
lines(lowess(data[which(SEX==2)][,-c("SEX")]), col = 3,type="b")
```



Problem 4, gender does have effect on age-expenditure relationship

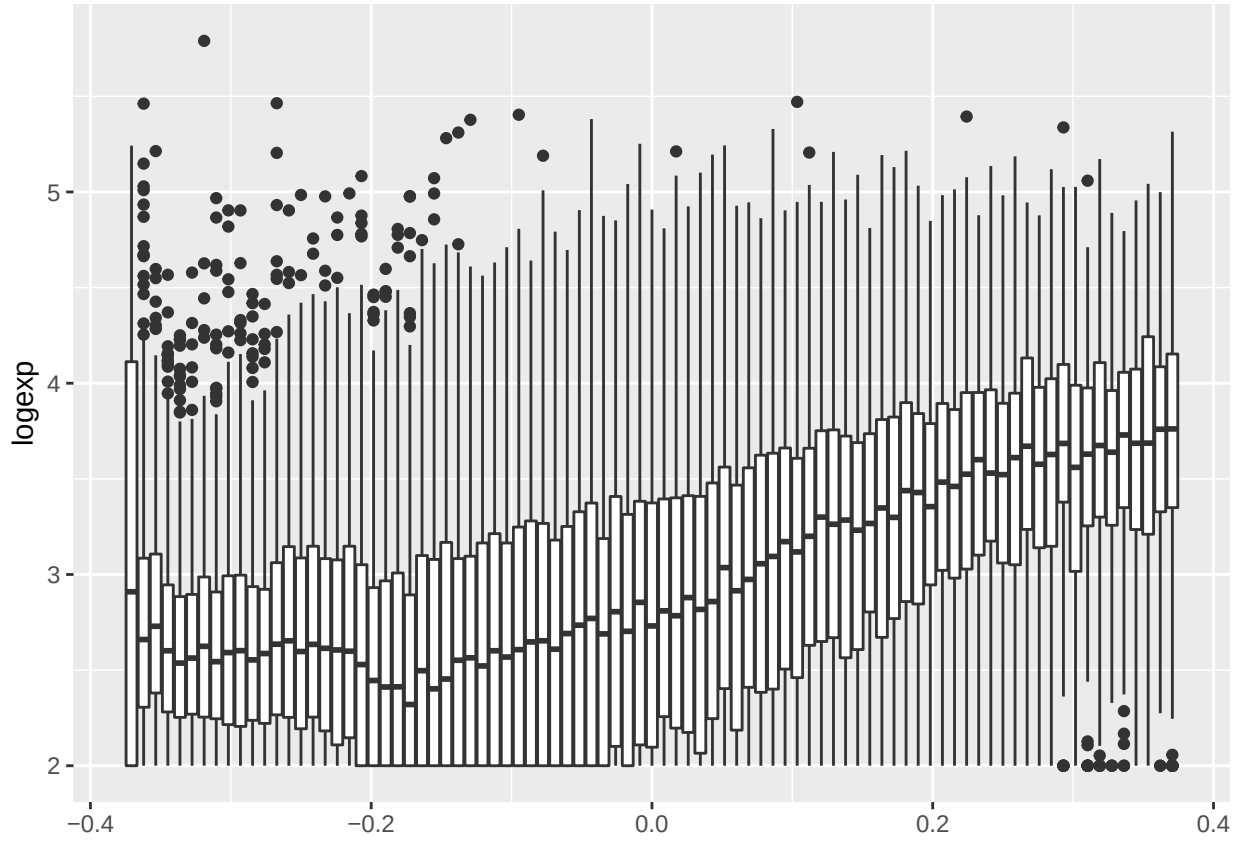
the results shows being female on average increase medical expense by 16.8%

every one year increase in age increase medical expense by 1.1%

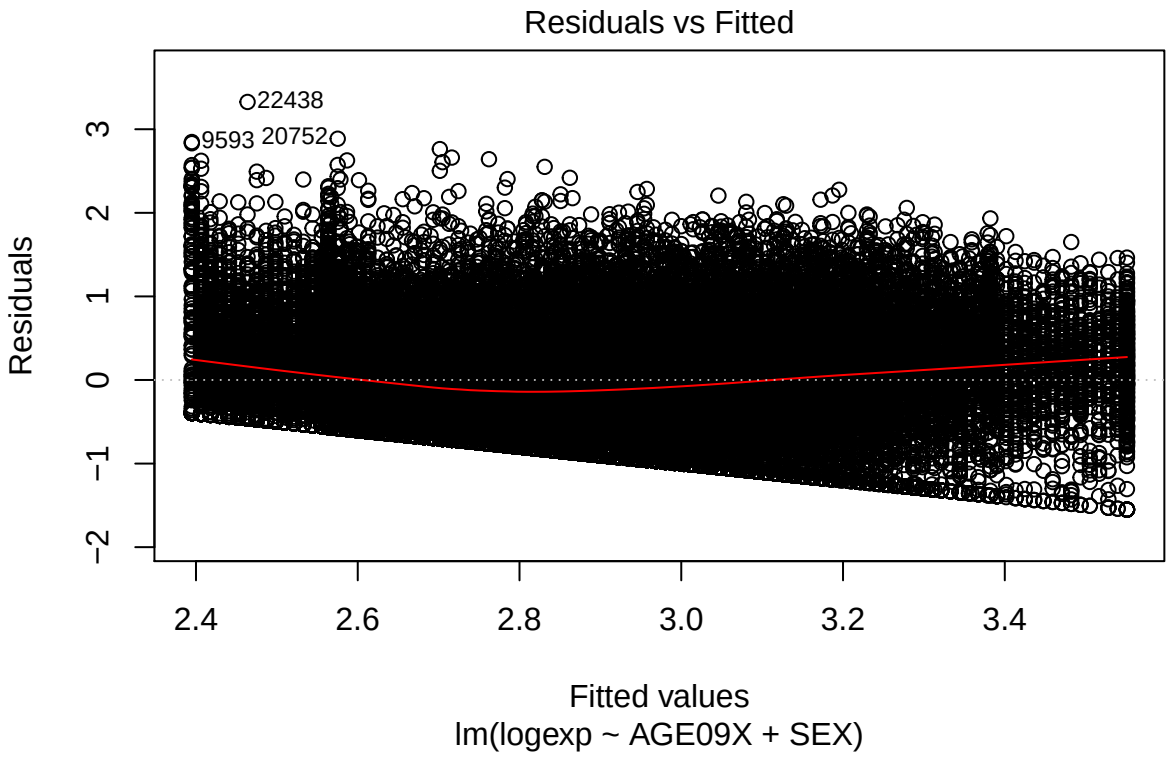
```
summary(lm(logexp~AGE09X+SEX,data=data))
```

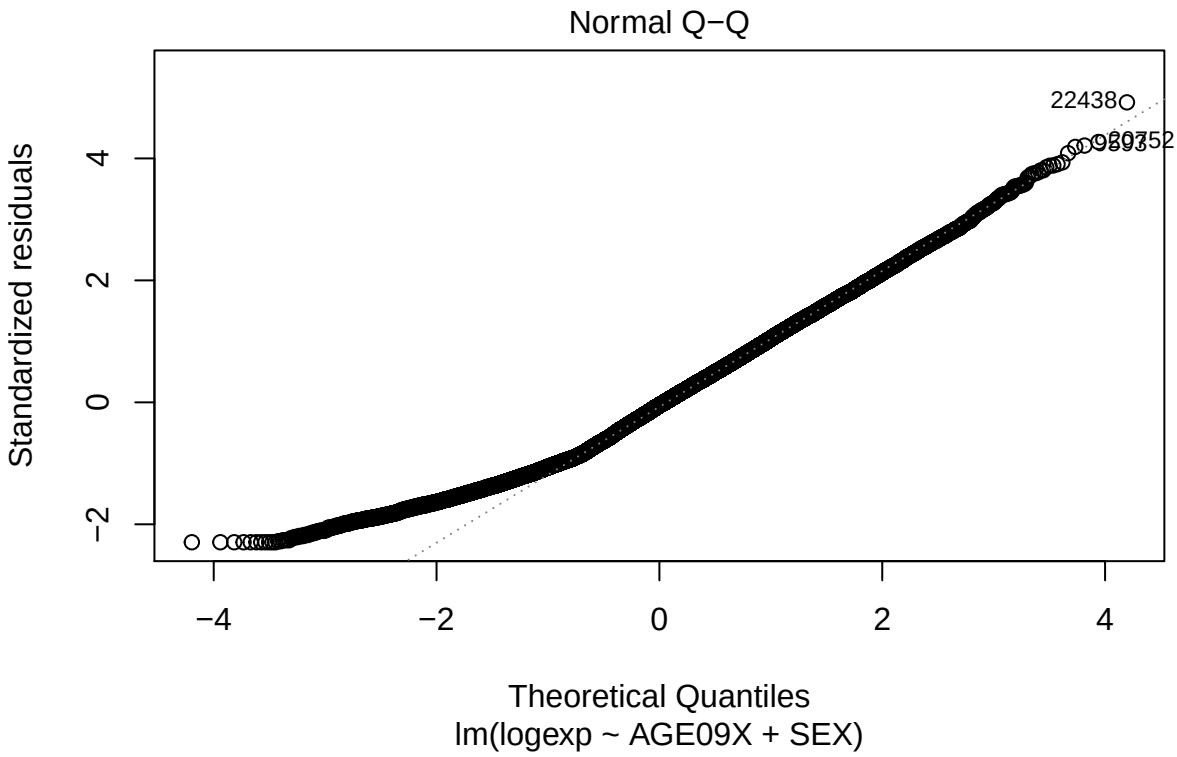
```
##
## Call:
## lm(formula = logexp ~ AGE09X + SEX, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5510 -0.5581 -0.0324  0.4580  3.3264
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.2374510  0.0122364  182.85   <2e-16 ***
## AGE09X      0.0114801  0.0001576   72.85   <2e-16 ***
## SEX         0.1688938  0.0070644   23.91   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6762 on 36852 degrees of freedom
## Multiple R-squared:  0.1423, Adjusted R-squared:  0.1422
## F-statistic:  3056 on 2 and 36852 DF,  p-value: < 2.2e-16
```

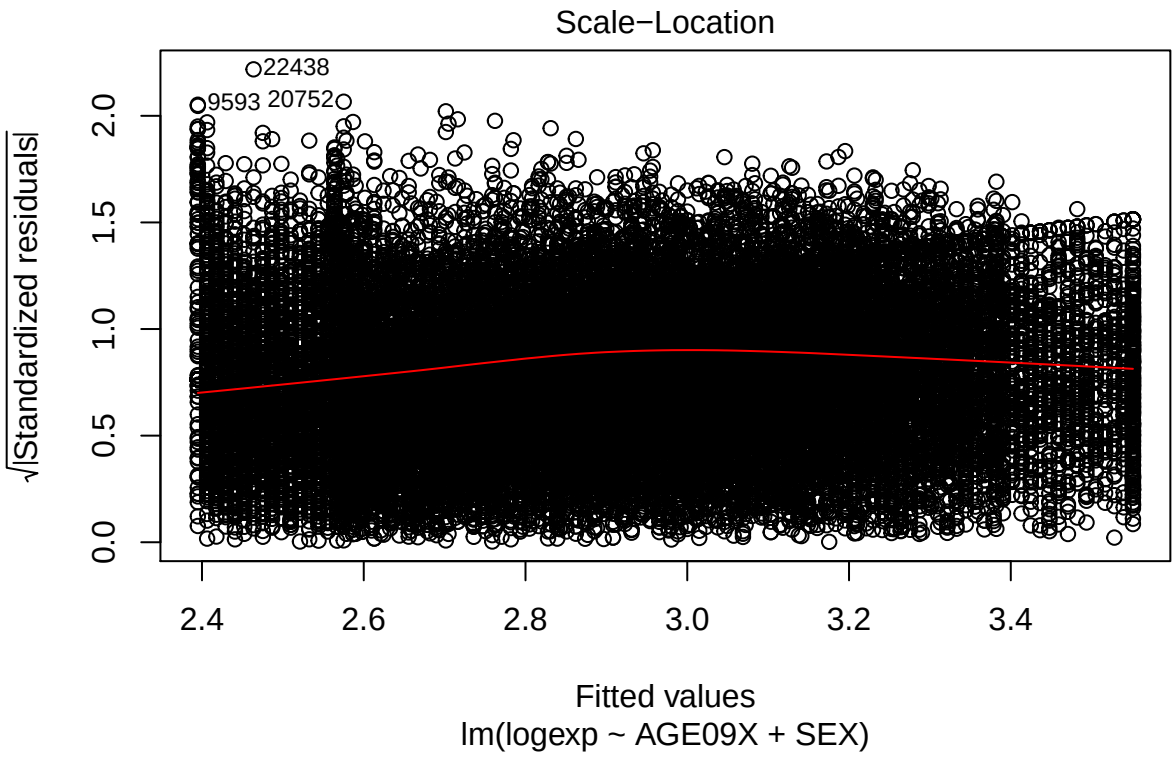
```
model<-lm(logexp~AGE09X+SEX,data=data)

###box plot
ggplot(data=data,aes(group=AGE09X,y=logexp))+geom_boxplot()
```

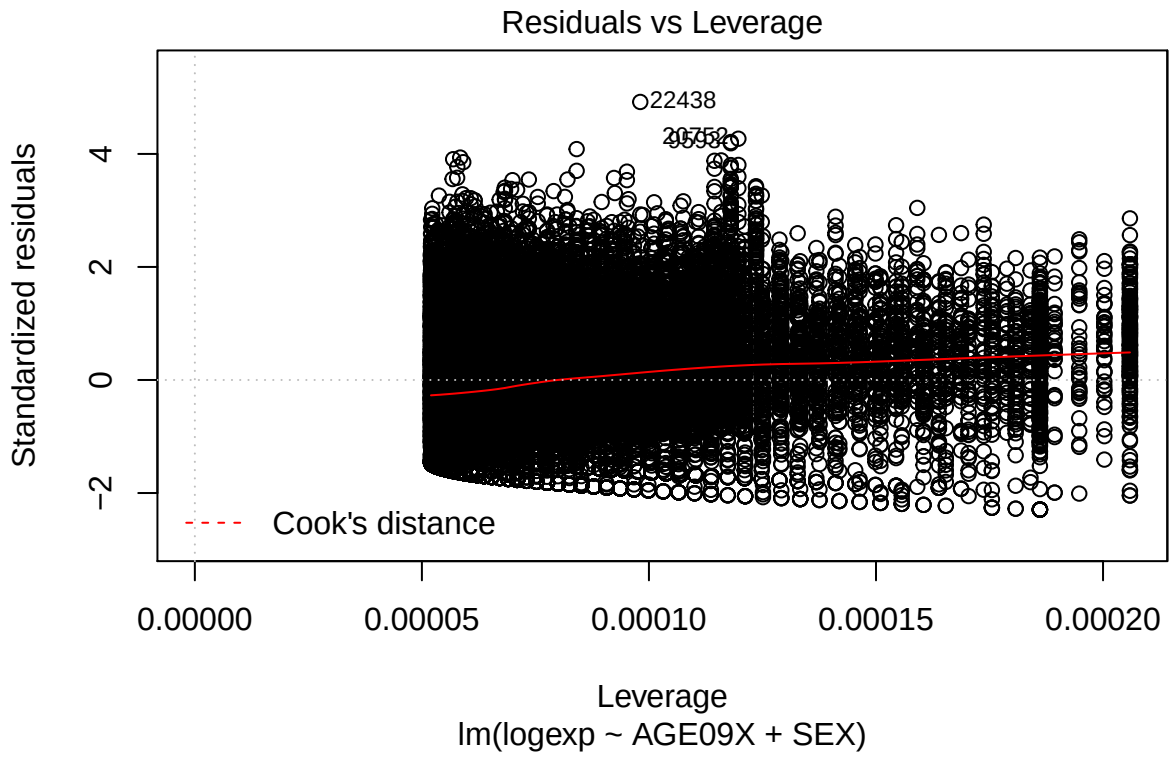



```
### residual and qq plot
plot(model)
```


Residuals vs Fitted
22438
9593
20752
Residuals
3
2
1
0
−1
−2
2.4
2.6
2.8
3.0
3.2
3.4
Fitted values
lm(logexp ~ AGE09X + SEX)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(logexp ~ AGE09X + SEX)
22438
−4
−2
0
2
4


Scale−Location
22438
9593
20752
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.4
2.6
2.8
3.0
3.2
3.4
Fitted values
lm(logexp ~ AGE09X + SEX)



###question 6, we need it to be on an log scale

```
### use base to be natural log
#dat$logexp<-log(dat$TOTEXP09+100,base=10)
dat$logexp<-log(dat$TOTEXP09+100)
data<-dat[,c("AGE09X","logexp","SEX")]

data$age25<-ifelse(data$AGE09X > 25, data$AGE09X - 25, 0)
data$age40<-ifelse(data$AGE09X > 40, data$AGE09X - 40, 0)

summary(lm(logexp ~ AGE09X + age25 + age40 + SEX, data=data ))
```

```
##
## Call:
## lm(formula = logexp ~ AGE09X + age25 + age40 + SEX, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.214 -1.282 -0.128  1.018  7.313
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.668152   0.032391 174.992  < 2e-16 ***
## AGE09X      -0.008275   0.001419  -5.834 5.47e-09 ***
## age25        0.036419   0.003087  11.799  < 2e-16 ***
## age40        0.019668   0.002618   7.513 5.89e-14 ***
## SEX          0.392325   0.015991  24.534  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.531 on 36850 degrees of freedom
## Multiple R-squared:  0.1712, Adjusted R-squared:  0.1711
## F-statistic:  1903 on 4 and 36850 DF,  p-value: < 2.2e-16
```

question 7 use base e instead of 10, we see the coefficient are the same

on average being female increase medical exp by 39% percent, every year older than 25 increase medical exp by 3.6%, while over age 40 this amount increased to (3.6%+1.9%). Also for counter this, every year increase in age decrease the medical expenditure by 0.8%.

8 new model:

```
### use base natural lof
#dat$logexp<-log(dat$TOTEXP09+100,base=10)
dat$logexp<-log(dat$TOTEXP09+100)
data<-dat[,c("AGE09X","logexp","SEX")]

data$age25<-ifelse(data$AGE09X > 25, data$AGE09X - 25, 0)
data$age40<-ifelse(data$AGE09X > 40, data$AGE09X - 40, 0)
data$SEX<-data$SEX-1



full=lm(logexp ~ AGE09X + age40 + SEX+ SEX*AGE09X+SEX*age25, data=data )
reduced=lm(logexp ~ AGE09X + age25,data=data)

anova(reduced,full)
```

```
## Analysis of Variance Table
##
## Model 1: logexp ~ AGE09X + age25
## Model 2: logexp ~ AGE09X + age40 + SEX + SEX * AGE09X + SEX * age25
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1  36852 87878
## 2  36848 85284  4    2593.5 280.14 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
### H0: reduced model
### H1L full model
```

the F test show we should accept the full model

```
### use base natural lof
#dat$logexp<-log(dat$TOTEXP09+100,base=10)
dat$logexp<-log(dat$TOTEXP09+100)
data<-dat[,c("AGE09X","logexp","SEX")]

data$age25<-ifelse(data$AGE09X > 25, data$AGE09X - 25, 0)
data$age40<-ifelse(data$AGE09X > 40, data$AGE09X - 40, 0)
data$SEX<-data$SEX-1
```

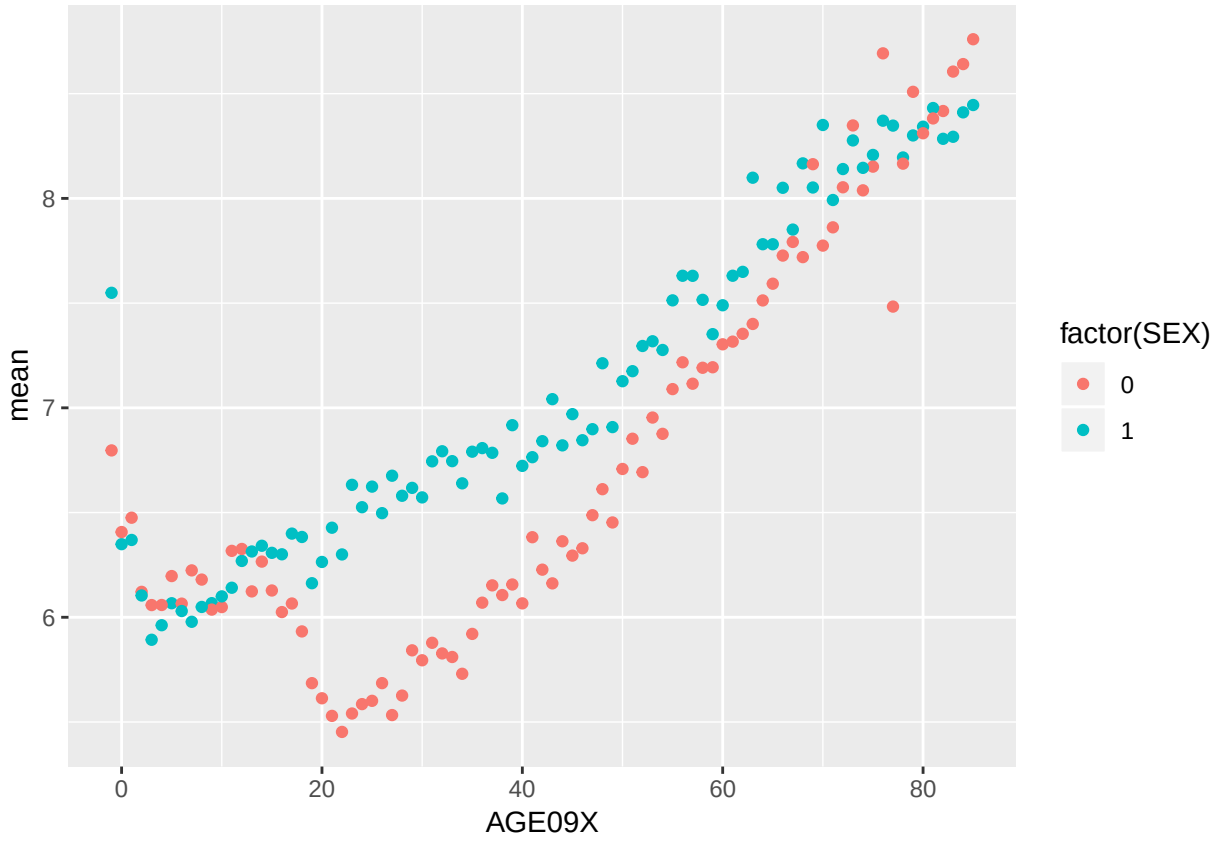
```r
model=lm(logexp ~ AGE09X + age25+SEX,data=data)
summary(model)
```

```
##
## Call:
## lm(formula = logexp ~ AGE09X + age25 + SEX, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0827 -1.2331 -0.1215  1.0197  7.2993
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.102441   0.022922  266.22   <2e-16 ***
## AGE09X      -0.013905   0.001205  -11.54   <2e-16 ***
## age25        0.056233   0.001605   35.04   <2e-16 ***
## SEX          0.393415   0.016003   24.58   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.532 on 36851 degrees of freedom
## Multiple R-squared:  0.1699, Adjusted R-squared:  0.1698
## F-statistic:  2514 on 3 and 36851 DF,  p-value: < 2.2e-16
```

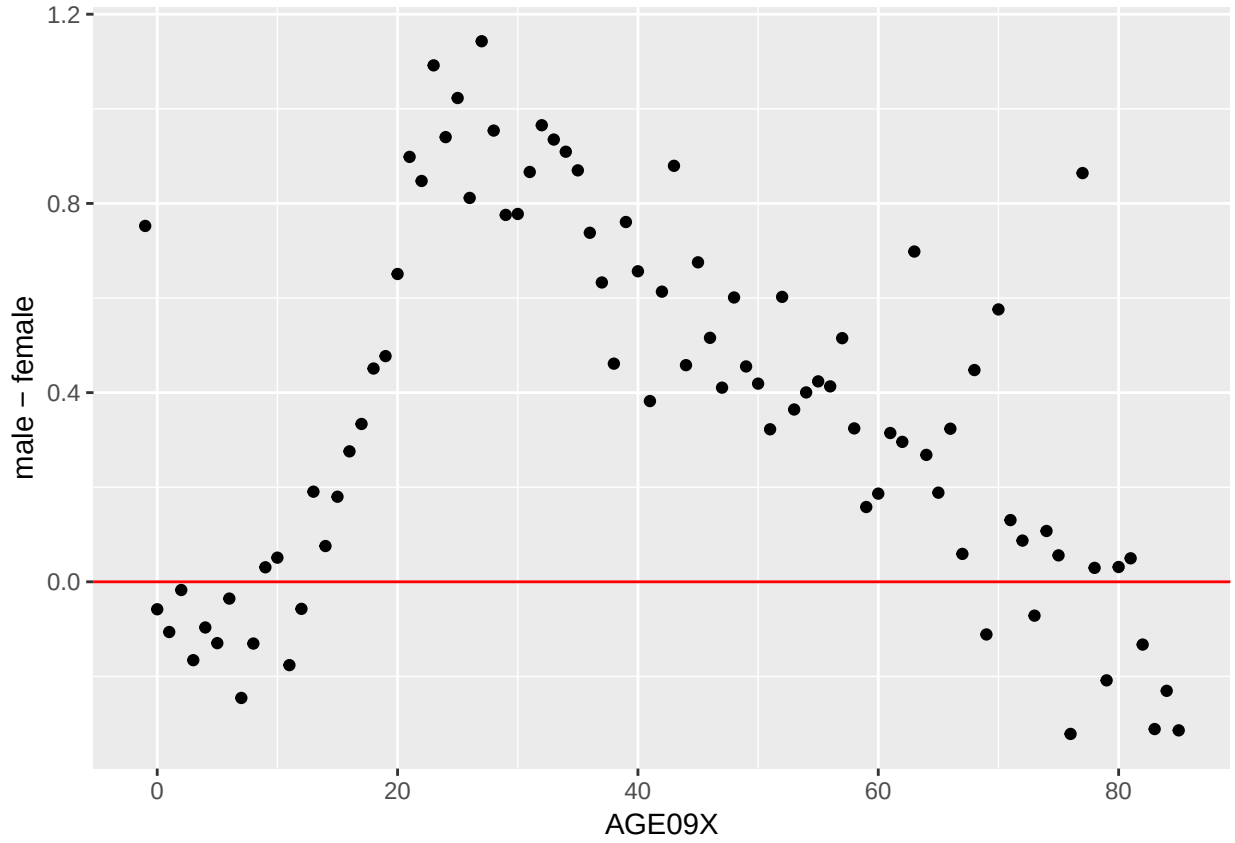
```r
### H0 SEX =0
### H1 SEX is not 0
### it shows there are difference between women and man in their mean
```

difference plot

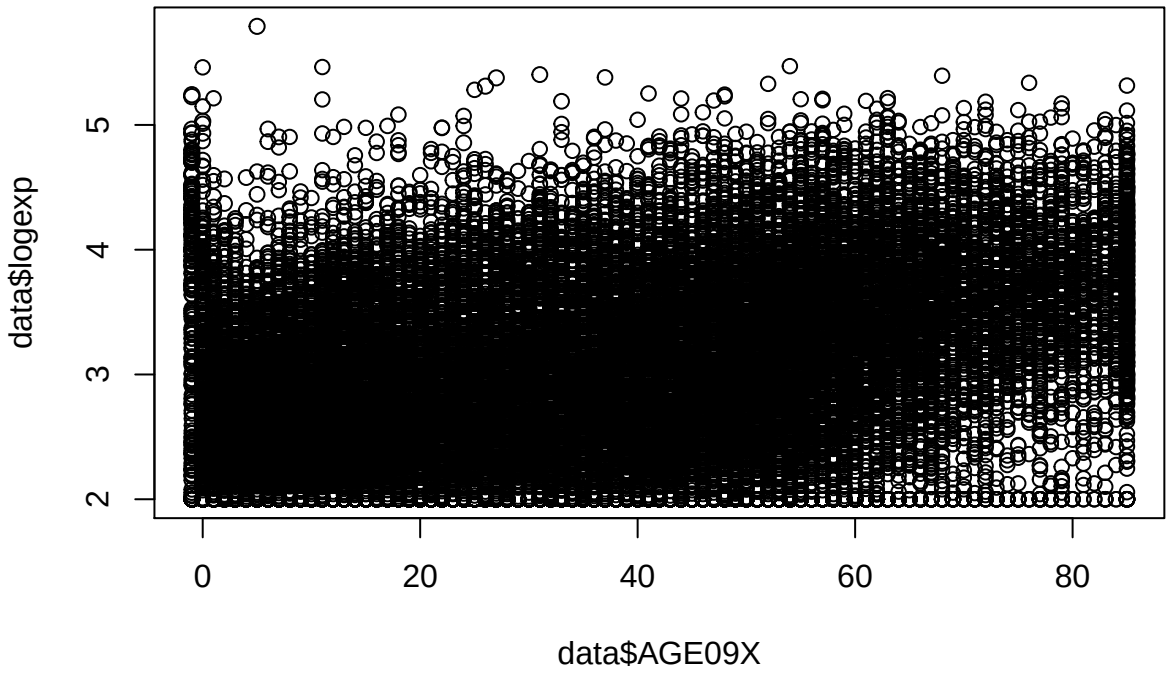
```r
data1<-data[,.(mean=mean(.SD$logexp)),by=c("AGE09X","SEX")]

### average for women and man
ggplot(data1,aes(x=AGE09X,y=mean,color=factor(SEX)))+geom_point()
```




```
### difference:

female<-data1[SEX=="0",][order(AGE09X)][,-c("SEX")]
names(female)[2]<-"female"

male<-data1[SEX=="1",][order(AGE09X)][,-c("SEX")]
names(male)[2]<-"male"

data1<-male[female,on="AGE09X"]
data1$difference<-data1$male-data1$female

ggplot(data1,aes(x=AGE09X,y=male-female))+geom_point()+geom_hline(yintercept = 0,col="red")
```




```
dat$logexp<-log(dat$TOTEXP09+100,base=10)

data<-dat[,c("AGE09X","logexp","SEX")]

data$age25<-ifelse(data$AGE09X > 25, data$AGE09X - 25, 0)
data$age40<-ifelse(data$AGE09X > 40, data$AGE09X - 40, 0)
data$SEX<-data$SEX-1

model<-lm(logexp ~ AGE09X + age40 + SEX+ SEX*AGE09X+SEX*age25, data=data )

infm<-influence.measures(model)
size<-abs(infm$infmat[,2])+1

plot(data$AGE09X,data$logexp,cex=size)
```




```
difference40<-c(0,0,0,1,0,40,15)
a<-esticon(model,L= difference40)

difference65<-c(0,0,0,1,0,65,30)
b<-esticon(model,L= difference65)

difference80<-c(0,0,0,1,0,80,55)
c<-esticon(model,L= difference80)

d<-rbind(a,b,c)[,c(2,3,7,8)]
d
```

```
##        Estimate  Std.Error      Lower  Upper
## [1,]  0.2824024  0.0087626  0.2652274 0.2996
## [2,]  0.3907451  0.0149218  0.3614979 0.4200
## [3,] -0.0114636  0.0191733 -0.0490438 0.0261
```

```
data$`AGE09X:SEX` <- data$AGE09X*data$SEX
data$`SEX:age25` <- data$age25*data$SEX
data$intercept<-1
predictmatrix<-data[,c("intercept","AGE09X","age40","SEX","age25","AGE09X:SEX","SEX:age25")]

### median difference for age 40
a<-10^(as.matrix(predictmatrix[which(AGE09X==40&SEX==1),][1]) %*% model$coefficients)-
10^(as.matrix(predictmatrix[which(AGE09X==40&SEX==0),][1]) %*% model$coefficients)

### median difference for age 65
```

```
b<-10^(as.matrix(predictmatrix[which(AGE09X==65&SEX==1),][1]) %*% model$coefficients)-
  10^(as.matrix(predictmatrix[which(AGE09X==65&SEX==0),][1]) %*% model$coefficients)

### median difference for age 80
c<-10^(as.matrix(predictmatrix[which(AGE09X==80&SEX==1),][1]) %*% model$coefficients)-
  10^(as.matrix(predictmatrix[which(AGE09X==80&SEX==0),][1]) %*% model$coefficients)

bootstrap<-data
bootstrap$Exp<-10^bootstrap$logexp

difference<-rep(0,1000)

for(i in 1:1000){
  male<-sample(bootstrap[AGE09X==40&SEX==1]$Exp,1)
  female<-sample(bootstrap[AGE09X==40&SEX==0]$Exp,1)
  difference[i]<-male-female
}

d<-sd(difference)

difference<-rep(0,1000)

for(i in 1:1000){
  male<-sample(bootstrap[AGE09X==65&SEX==1]$Exp,1)
  female<-sample(bootstrap[AGE09X==65&SEX==0]$Exp,1)
  difference[i]<-male-female
}

e<-sd(difference)

difference<-rep(0,1000)

for(i in 1:1000){
  male<-sample(bootstrap[AGE09X==80&SEX==1]$Exp,1)
  female<-sample(bootstrap[AGE09X==80&SEX==0]$Exp,1)
  difference[i]<-male-female
}

f<-sd(difference)

cbind(c(a,b,c),c(d,e,f))
```

```
##           [,1]     [,2]
## [1,]  420.7463 13569.92
## [2,]  495.2140 14662.02
## [3,] -119.9576 15991.14
```

13 on average, female spend 39$ than male, and their expense relation with age differs. Male are most healthy at age 25, where their medical expenditure is minimized at that time. While female's medical expenditure increase as age increase all over their life.

```
data$`AGE09X:SEX` <- data$AGE09X*data$SEX
data$`SEX:age25` <- data$age25*data$SEX
```

```r
data$intercept<-1
predictmatrix<-data[,c("intercept","AGE09X","age40","SEX","age25","AGE09X:SEX","SEX:age25")]

X<-as.matrix(predictmatrix)
Y<-as.matrix(data$logexp)

theta<-solve(t(X)%*%X)%*%(t(X))%*%Y
theta
```

```
##                     [,1]
## intercept   2.788794026
## AGE09X     -0.014370089
## age40       0.009522931
## SEX        -0.153754004
## age25       0.029874665
## AGE09X:SEX  0.021854245
## SEX:age25  -0.029200895
```

```r
t(Y-X%*%theta)%*%(Y-X%*%theta)/length(Y)
```

```
##          [,1]
## [1,] 0.436457
```

```r
summary(model)
```

```
##
## Call:
## lm(formula = logexp ~ AGE09X + age40 + SEX + SEX * AGE09X + SEX *
##     age25, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.78835 -0.50706 -0.05435  0.43796  3.07315
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.7887940  0.0131770 211.641   <2e-16 ***
## AGE09X      -0.0143701  0.0007983 -18.001   <2e-16 ***
## age40        0.0095229  0.0011319   8.413   <2e-16 ***
## SEX         -0.1537540  0.0187130  -8.216   <2e-16 ***
## age25        0.0298747  0.0014917  20.028   <2e-16 ***
## AGE09X:SEX   0.0218542  0.0010418  20.978   <2e-16 ***
## SEX:age25   -0.0292009  0.0013904 -21.001   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6607 on 36848 degrees of freedom
## Multiple R-squared:  0.1812, Adjusted R-squared:  0.1811
## F-statistic:  1359 on 6 and 36848 DF,  p-value: < 2.2e-16
```

```r
library(knitr)
purl("Homework7.Rmd")
```