Homework 7

Ben Lieberman

November 19, 2019

```
library(knitr)
library(doBy)
library(msm)
```

```
##
## Attaching package: 'msm'
```

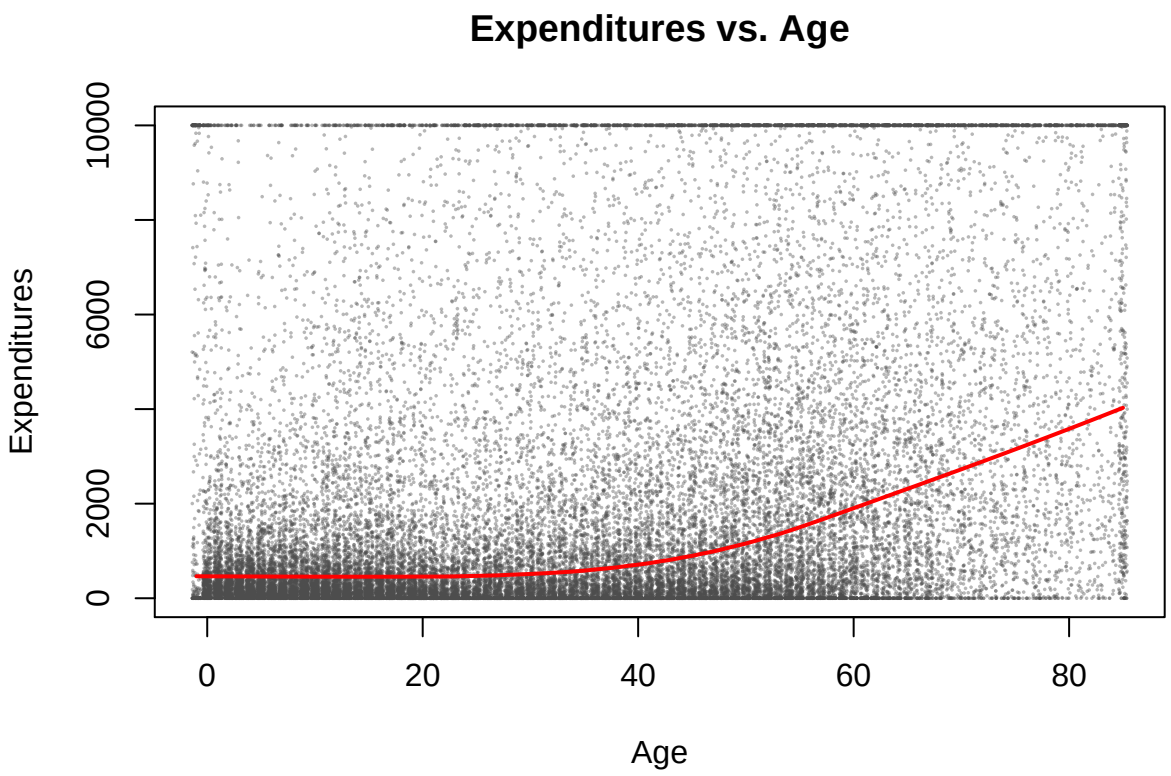
```
## The following object is masked from 'package:doBy':
##
##     fev
```

I. Advanced Inference for Linear Regression

```
load(url("http://people.stat.sc.edu/hoyen/Stat704/Data/h129.RData"))
dat = data.frame(h129$TOTEXP09, h129$SEX, h129$RACEX, h129$ASTHDX, h129$DOBMM, h129$DOBYY,h129$ADSMOK42
```

1.

```
plot.expenditures=NULL

# find and store expenditures near 10,000
for(i in 1:nrow(dat)){
  ifelse (dat$h129.TOTEXP09[i]>10000, plot.expenditures[i] <- 10001, plot.expenditures[i]<-dat$h129.TOT
}

# plot expenditures vs. age
plot(jitter(dat$h129.AGE09X,2),plot.expenditures, cex = .25, main="Expenditures vs. Age",
xlab = "Age", ylab = "Expenditures", col = rgb(0.3,0.3,0.3, 0.4), pch = 16)
# add the lowess smoothing
lines(lowess(dat$h129.AGE09X,dat$h129.TOTEXP09), lwd = 2, col = "red")
```



2. It looks like the slope of the smooth curve begins to change (significantly) at the age of 40, so let's add a spline term there. There does not seem to be a linear relationship between age and medical expenditures. The age coefficient of 17.944 suggests that on average, with every year that passes, ones medical expenditures increase by \$17.94. Once the person reaches age 40, their expenditures increase by an average \$17.94 + \$185.86 = \$203.62 every year older they get.

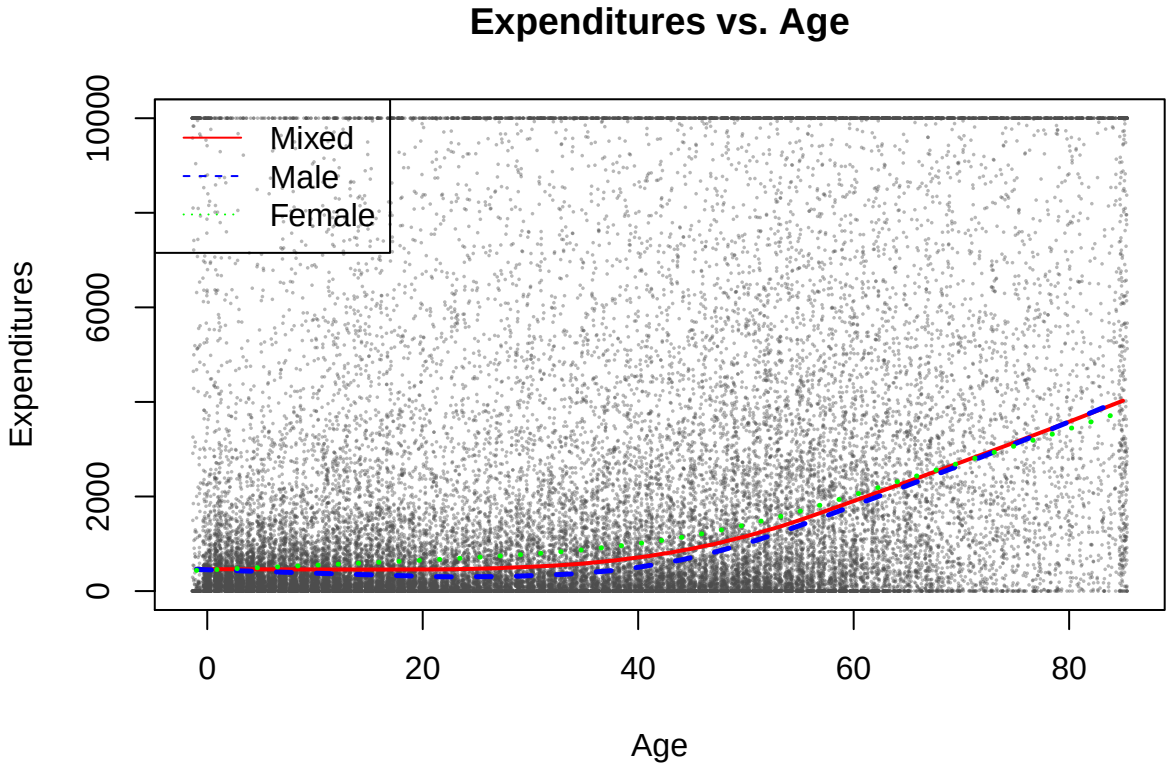
It is important to note that the units may be scaled, so it may be \$2,030 rather than \$203.

```
age40 = ifelse(dat$h129.AGE09X > 40, dat$h129.AGE09X-40, 0)
fitlsp = lm(dat$h129.TOTEXP09 ~ dat$h129.AGE09X + age40)
summary(fitlsp)
```

```
##
## Call:
## lm(formula = dat$h129.TOTEXP09 ~ dat$h129.AGE09X + age40)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -11477  -2306  -1687   -713 614953
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1588.177    126.600  12.545  < 2e-16 ***
## dat$h129.AGE09X   17.944      4.847   3.702 0.000214 ***
## age40            185.857      9.525  19.512  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10720 on 36852 degrees of freedom
## Multiple R-squared:  0.05047,    Adjusted R-squared:  0.05042
## F-statistic: 979.5 on 2 and 36852 DF,  p-value: < 2.2e-16
```

3.

```
# Separating men and women
male.data = dat[which(dat$h129.SEX == 1),]
female.data = dat[which(dat$h129.SEX ==2),]

# plot expenditures vs. age
plot(jitter(dat$h129.AGE09X,2),plot.expenditures, cex = .25, main="Expenditures vs. Age",
xlab = "Age", ylab = "Expenditures", col = rgb(0.3,0.3,0.3, 0.4), pch = 16)

# add the lowess smoothing

# mixed
lines(lowess(dat$h129.AGE09X,dat$h129.TOTEXP09), lwd = 2, col = "red")
# for men
lines(lowess(male.data$h129.AGE09X,male.data$h129.TOTEXP09), lwd = 2.5, lty = 2, col = "blue")
# for women
lines(lowess(female.data$h129.AGE09X,female.data$h129.TOTEXP09), lwd = 2.5, lty = 3, col = "green")
# add the legend
legend("topleft", c("Mixed", "Male", "Female"), col=c("red", "blue", "green"), lty=c(1,2,3))
```



4.

Based on the graph, at younger ages there does not seem to be much difference of a difference between male and female medical expenditures. But, as age increases, women seem to have higher average expenditures than men until around the age of 60.

The regression coeficients support this claim. Based on the regression, gender clearly impacts the average medical expenditure. Initially, females have lower expenditures than males, but before age 40, female expenditures increase by an average of $64.47 dollars per year more than males.

However, once after the age of 40, males average expenditure per year increases faster than females. This is supported in the graph because older men have a higher average expenditure than older females.

Intuitively this makes sense as men have a shorter life expectency, and end of life care is expensive (cancer, hospice, etc.). Also, perhaps due to child birth and the need for increased frequency for cancer screenings (specifically breast cancer), this may explain the increased mid-life medical cost for women.

```
dat$h129.SEX = as.factor(dat$h129.SEX)
# add gender as a spline interaction term
MLR.gSpline = lm(dat$h129.TOTEXP09 ~ dat$h129.AGE09X + age40 + dat$h129.SEX + dat$h129.AGE09X*dat$h129.S
summary(MLR.gSpline)
```

```
##
## Call:
## lm(formula = dat$h129.TOTEXP09 ~ dat$h129.AGE09X + age40 + dat$h129.SEX +
##     dat$h129.AGE09X * dat$h129.SEX + age40 * dat$h129.SEX)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -12405  -2694  -1516   -652 614847
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    1867.994    176.546  10.581  < 2e-16 ***
## dat$h129.AGE09X                 -16.749      6.925  -2.418   0.0156 *
## age40                           265.786     14.174  18.752  < 2e-16 ***
## dat$h129.SEX2                  -518.592    253.060  -2.049   0.0404 *
## dat$h129.AGE09X:dat$h129.SEX2   64.474      9.701   6.646 3.05e-11 ***
## age40:dat$h129.SEX2           -146.222     19.166  -7.629 2.42e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10710 on 36849 degrees of freedom
## Multiple R-squared:  0.05289,    Adjusted R-squared:  0.05276
## F-statistic: 411.6 on 5 and 36849 DF,  p-value: < 2.2e-16
```

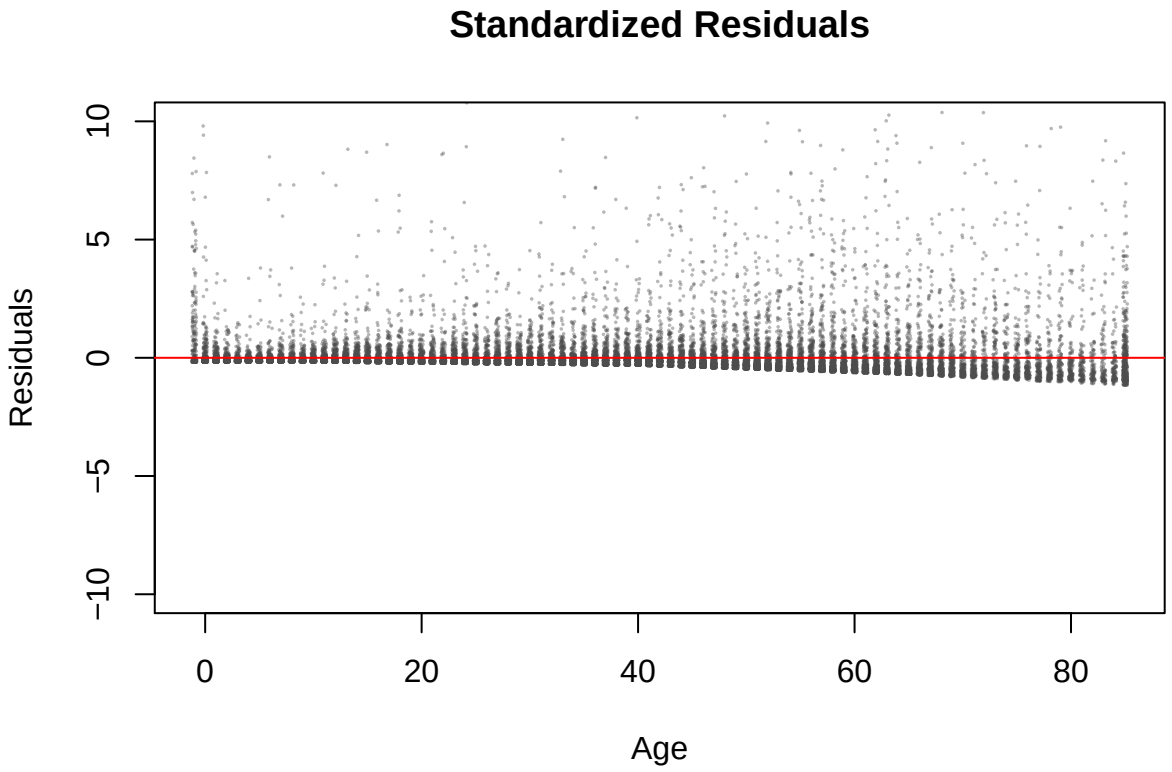
5. Normality: The QQ plot indicates that the residuals are not normally distributed. This assumption is clearly violated.

Equal variance: The residual plots indicates that we do not have constant variance accross all ages. This implies that we should consider a transformation in the data similar to those seen in KNNL 3.9.

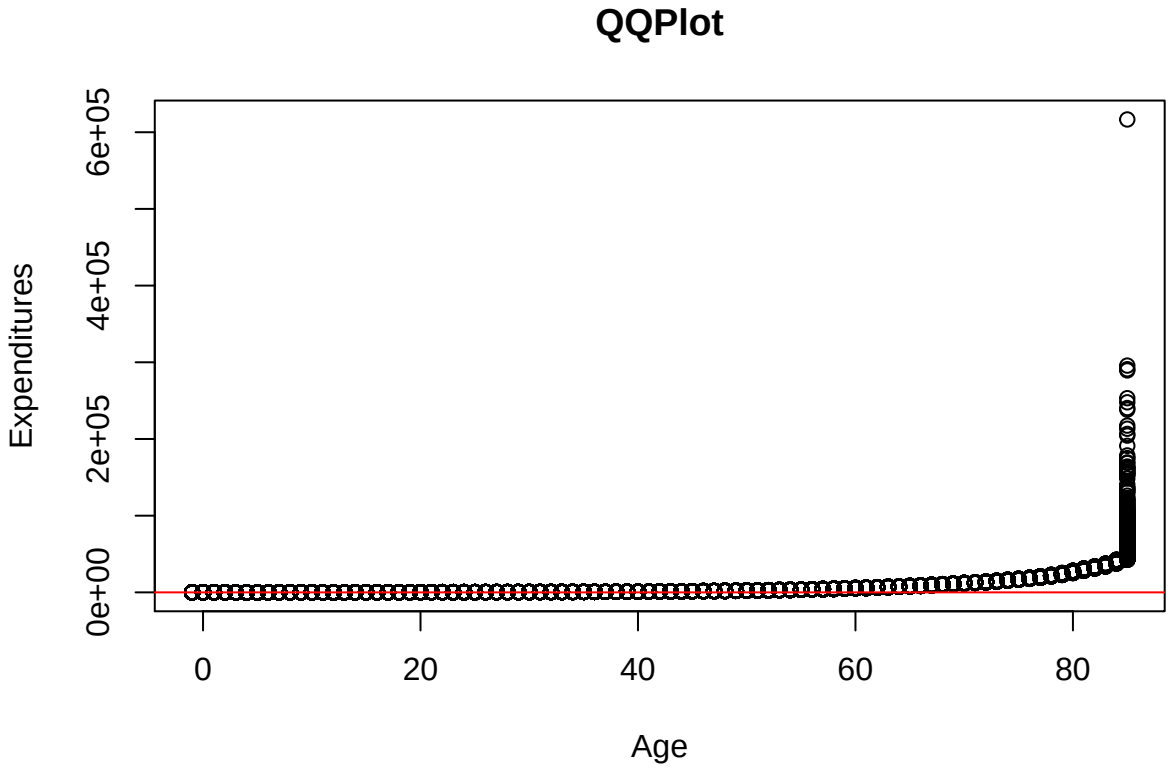
We also need to address the outlier issue, so perhaps one of the transformations might be the best idea moving forward.

```
#Using the MLR line with the spline and interaction terms
MLRresid = resid(MLR.gSpline)
plot(jitter(dat$h129.AGE09X), MLRresid,  cex = .25, ylim = c(-20000,20000),
  ylab="Residuals", xlab="Age",
   main="Expenditures and Age Residuals", col = rgb(0.3,0.3,0.3, 0.4), pch = 16)
abline(0, 0, col = "red")
```

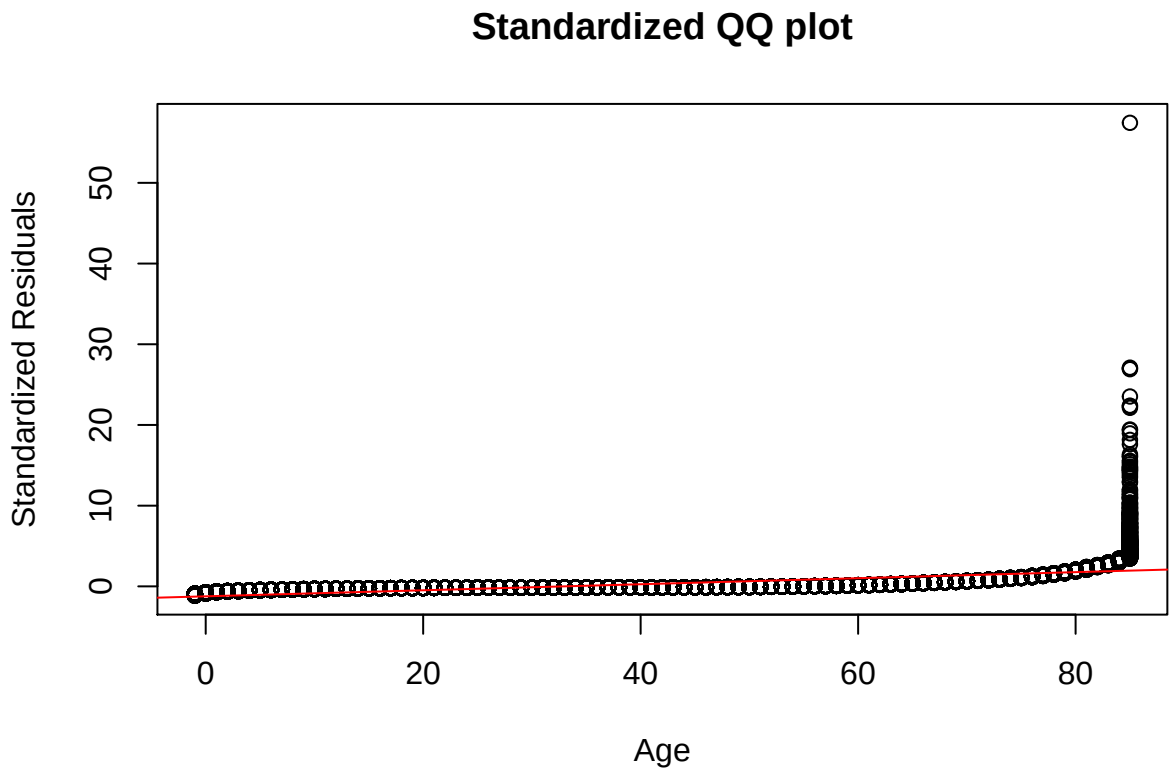

```
# standardized residuals
stand.resid = rstandard(MLR.gSpline)
plot(jitter(dat$h129.AGE09X), stand.resid,  cex = .25, ylim = c(-10,10),
  ylab="Residuals", xlab="Age",
   main="Standardized Residuals", col = rgb(0.3,0.3,0.3, 0.4), pch = 16)
abline(0, 0, col = "red")
```



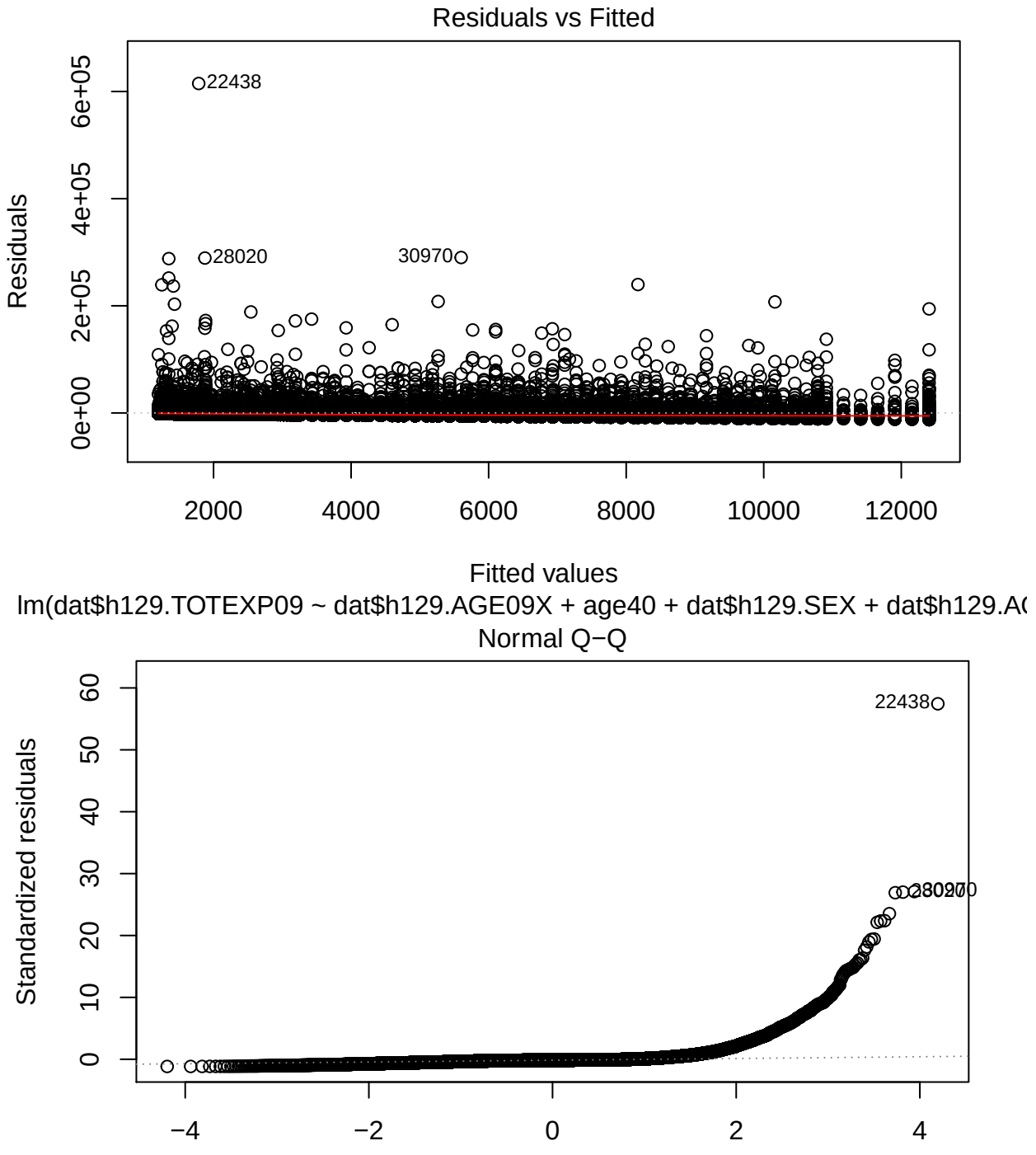

```
#QQ Plot using full data
qqplot(dat$h129.AGE09X,dat$h129.TOTEXP09, xlab = "Age",
       ylab = "Expenditures", main = "QQPlot")
qqline(dat$h129.AGE09X, datax = T, col = "red")
```



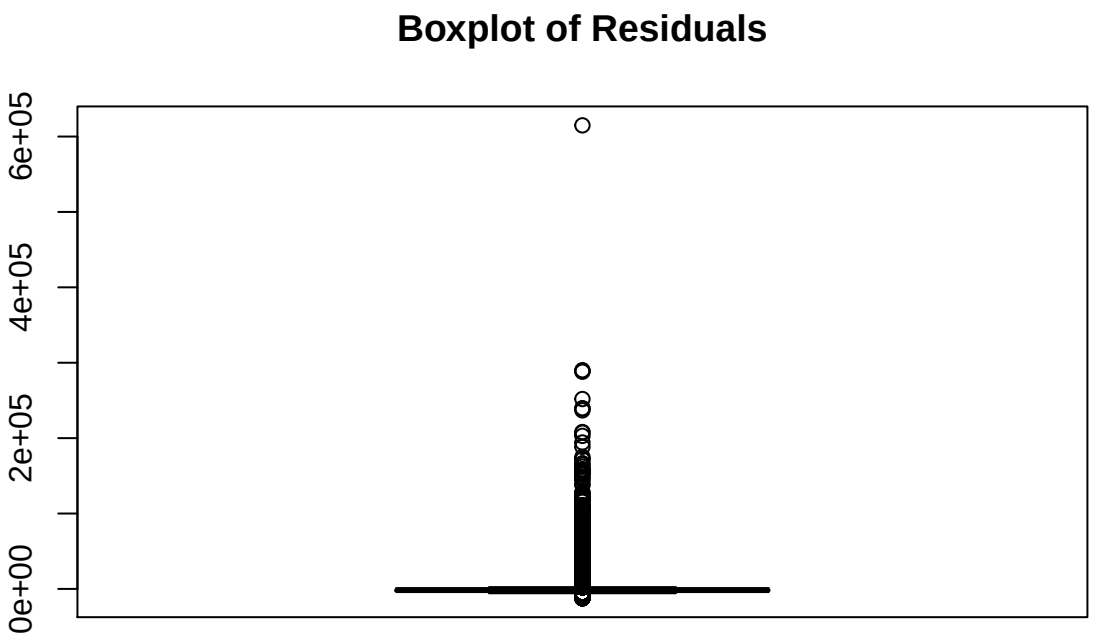
```r
#Standardized Resid QQ Plot
qqplot(dat$h129.AGE09X,stand.resid, xlab="Age", ylab="Standardized Residuals",
       main = "Standardized QQ plot")
qqline(dat$h129.AGE09X, datax = T, col = "red")
```




```r
# plot residuals v fitted and normal Q-Q
plot(MLR.gSpline, which = c(1,2))
```

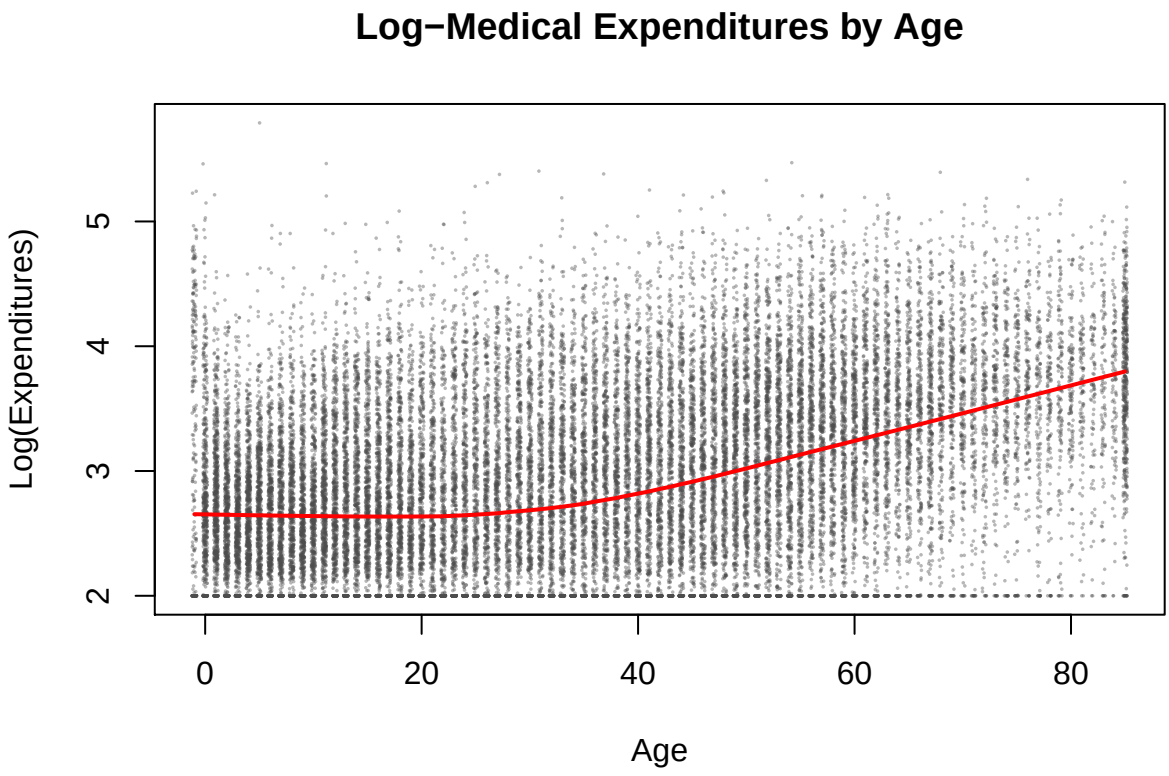



```r
# check for outliers
boxplot(MLRresid, main = "Boxplot of Residuals")
```



6.

```
lexp = log10(dat$h129.TOTEXP09 + 100)
plot(jitter(dat$h129.AGE09X),lexp, xlab = "Age", ylab = "Log(Expenditures)",
     main = "Log-Medical Expenditures by Age",
     cex=0.25, col = rgb(0.3,0.3,0.3, 0.4), pch = 16)
lines(lowess(dat$h129.AGE09X,lexp), lwd = 2, col= "red")
```




```
# let's try the splines, we see a
# significant change in slope in the lowess line around 30
```

```r
age30 = ifelse(dat$h129.AGE09X > 30, dat$h129.AGE09X-30, 0)


LogModel.fit = lm(lexp ~ dat$h129.AGE09X + age30 + dat$h129.SEX +
                    dat$h129.AGE09X*dat$h129.SEX + age30*dat$h129.SEX)
summary(LogModel.fit)
```

```
##
## Call:
## lm(formula = lexp ~ dat$h129.AGE09X + age30 + dat$h129.SEX +
##     dat$h129.AGE09X * dat$h129.SEX + age30 * dat$h129.SEX)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.78029 -0.51255 -0.05398  0.43982  3.08432
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    2.7616314  0.0121282 227.703  < 2e-16 ***
## dat$h129.AGE09X               -0.0111712  0.0005972 -18.705  < 2e-16 ***
## age30                          0.0357856  0.0009119  39.244  < 2e-16 ***
## dat$h129.SEX2                 -0.1151091  0.0174461  -6.598 4.22e-11 ***
## dat$h129.AGE09X:dat$h129.SEX2  0.0172378  0.0008399  20.523  < 2e-16 ***
## age30:dat$h129.SEX2           -0.0260573  0.0012530 -20.795  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6611 on 36849 degrees of freedom
## Multiple R-squared:  0.1801, Adjusted R-squared:   0.18
## F-statistic:  1619 on 5 and 36849 DF,  p-value: < 2.2e-16
```

7.

$\beta_0 = 2.76$ is the intercept. It can be intepreted as the log(expenditure) for those who are infants, i.e. age 0.

$\beta_1 = -0.011$ means that for every year increase in age, the average log(expenditure) decreases by \$0.01.

$\beta_2 = 0.036$ means that once someone reaches age 30, the average log(expenditure) increases \$0.04 than before age 30.The actual increase per year is $\beta_1 + \beta_2 = \$.025$ when someone is 30 or older.

$\beta_3 = -0.115$ is an indicator that only applies to females. As it's negative, it can be intrepreted to mean that women have lower log expenditures than men.

$\beta_4 = 0.017$ is an interaction term that is only included if the person is female. It shows the differing slopes between genders (difference in expenditures). It means that the average log(expenditure) increases \$0.02 faster as age increases for women than men, after controlling for other variables. Obviously, when actually calculating the difference, other coeficients and interaction terms are included.

$\beta_5 = -0.026$ is an interaction term that is only present when the person is 30+ year old female. The negative value means that the log expenditures of a 30+ y.o. female, decrease by \$-0.03 for every year. Again, other varaibles will effect her actual change in log expenditures, but when the requirements for this interaction term are met, this is how it will effect her log(Expenditure).

8. Since the p-value is $< 2.2 * 10^{-16}$, we reject $H_0 : \beta_4 = \beta_5 = 0$ and conclude that mean log-expenditures is different for men and women depending on age.

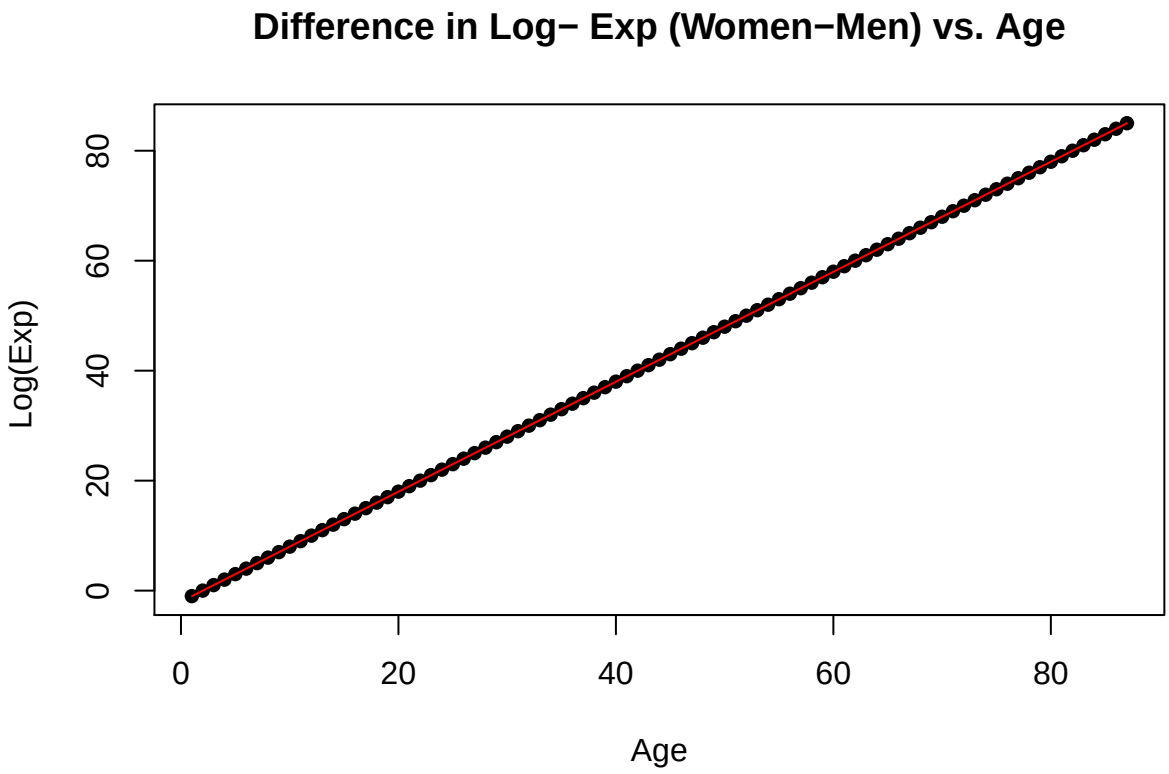
```r
fitFull = lm(lexp ~ dat$h129.AGE09X + age30 + dat$h129.SEX +
               dat$h129.AGE09X*dat$h129.SEX + age30*dat$h129.SEX)
```

```r
# reduced is full w/o interaction terms
fitRed = lm(lexp ~ dat$h129.AGE09X + age30 + dat$h129.SEX)
ftest = anova(fitRed,fitFull)
ftest
```

```
## Analysis of Variance Table
##
## Model 1: lexp ~ dat$h129.AGE09X + age30 + dat$h129.SEX
## Model 2: lexp ~ dat$h129.AGE09X + age30 + dat$h129.SEX + dat$h129.AGE09X *
##     dat$h129.SEX + age30 * dat$h129.SEX
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1  36851 16300
## 2  36849 16107  2    193.59 221.44 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

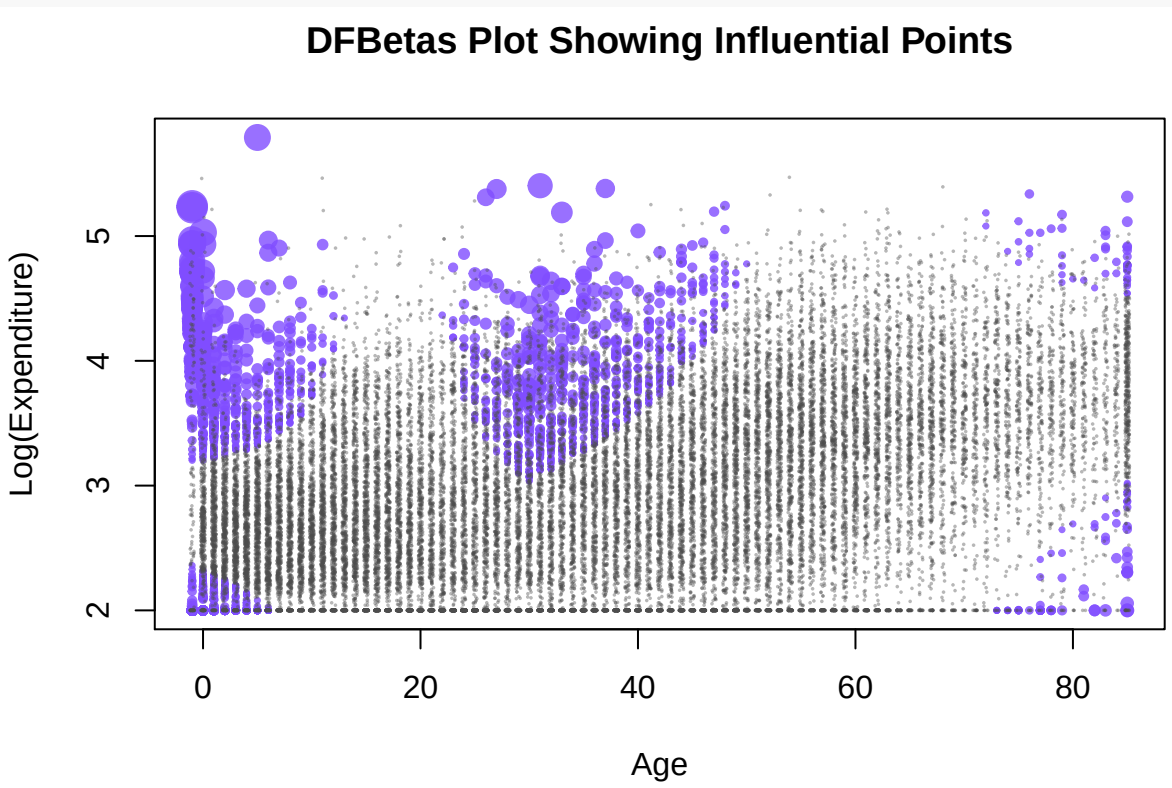
9.

```r
wmdif = c()
# -1 : 85 are ages
for(n in -1:85){
  lambda = c(0,0,0,1,n,ifelse(n>30,n-30,0))
  z = esticon(LogModel.fit, lambda)
  wmdif = c(wmdif,z$Estimate)
}
plot(-1:85, wmdif, xlab = "Age", ylab="Log(Exp) ", main="Difference in Log- Exp (Women-Men) vs. Age", p
lines(-1:85, wmdif, col = "red")
```



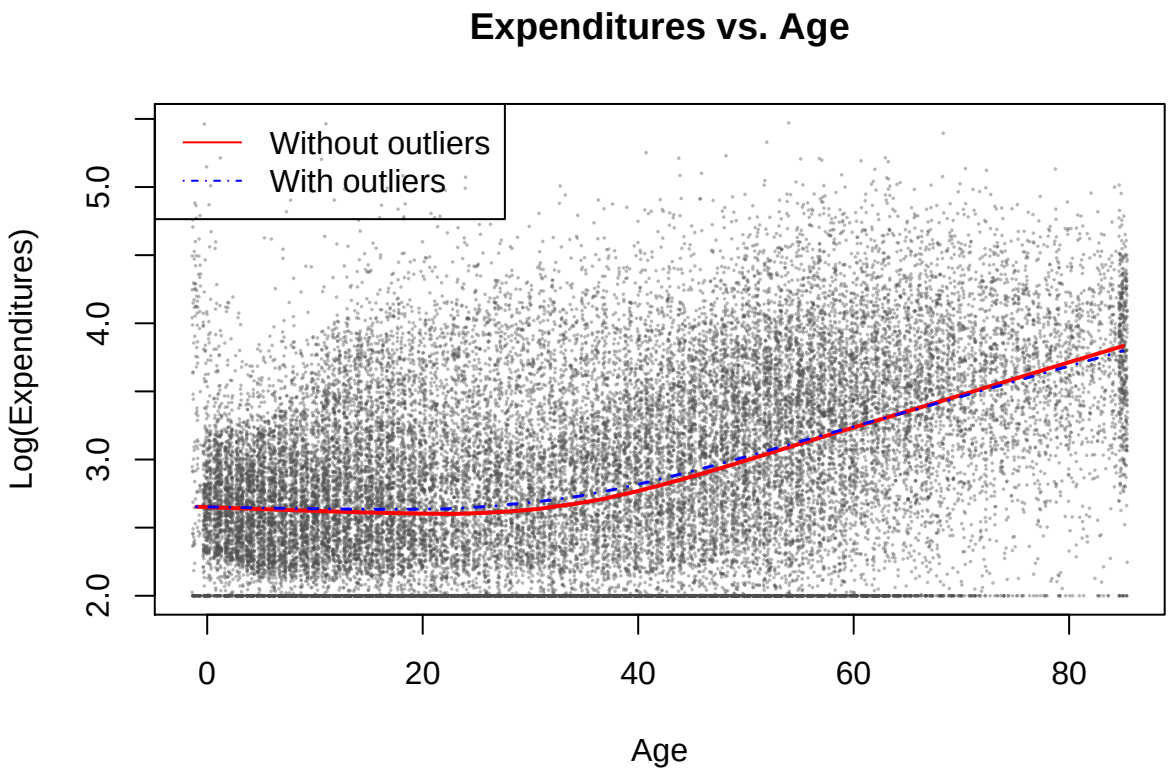
10. There are 1,784 highly influential observations.

```
fit = lm(lexp ~ dat$h129.AGE09X + age30 + dat$h129.SEX
         + dat$h129.AGE09X*dat$h129.SEX + age30*dat$h129.SEX)

# fit the dfbetas
fit.dfbetas = dfbetas(fit)

# set our cutoff
cutoff = 2 / sqrt(nrow(dat))

# now, which ones are outliers
outlierPos = which(fit.dfbetas[,2]>cutoff)
outlierNeg = which(fit.dfbetas[,2] < -1*cutoff)
outliers = c(outlierPos,outlierNeg)
outlier.dat = dat[outliers,]

# plot the influential points based on size
size = 33 * abs(fit.dfbetas[outliers,2])

plot(outlier.dat$h129.AGE09X,lexp[outliers], pch = 19, xlab= "Age", ylab="Log(Expenditure)", main = "DF

points(jitter(dat$h129.AGE09X[-outliers]),lexp[-outliers], xlab = "Age", pch = 16, cex = 0.25, col = rg
```




```
newdata = dat[-outliers,]

# plot the new data for an idea on spline location
plot(jitter(newdata$h129.AGE09X,2),lexp[-outliers], cex = .25, main="Expenditures vs. Age",
     xlab = "Age", ylab = "Log(Expenditures)", col = rgb(0.3,0.3,0.3, 0.4), pch = 16)
lines(lowess(newdata$h129.AGE09X,lexp[-outliers]), lwd = 2, col = "red")
lines(lowess(dat$h129.AGE09X,lexp), lwd = 1.5, lty = 4 , col= "blue")
```

```
legend("topleft", c("Without outliers", "With outliers"), col=c("red", "blue"), lty=c(1,4))
```




```
f.age30 = ifelse(newdata$h129.AGE09X > 30, newdata$h129.AGE09X-30, 0)

NoOutlierFit = lm(lexp[-outliers] ~ newdata$h129.AGE09X + f.age30 + newdata$h129.SEX + newdata$h129.AGE
summary(NoOutlierFit)
```

```
##
## Call:
## lm(formula = lexp[-outliers] ~ newdata$h129.AGE09X + f.age30 +
##     newdata$h129.SEX + newdata$h129.AGE09X * newdata$h129.SEX +
##     f.age30 * newdata$h129.SEX)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.69724 -0.43983 -0.04645  0.41689  2.81514
##
## Coefficients:
##                                              Estimate Std. Error t value
## (Intercept)                                 2.7881641  0.0134630 207.098
## newdata$h129.AGE09X                        -0.0175113  0.0006513 -26.886
## f.age30                                     0.0469476  0.0009681  48.494
## newdata$h129.SEX2                          -0.1416418  0.0179443  -7.893
## newdata$h129.AGE09X:newdata$h129.SEX2       0.0235778  0.0008581  27.476
## f.age30:newdata$h129.SEX2                  -0.0372192  0.0012642 -29.441
##                                            Pr(>|t|)
## (Intercept)                                 < 2e-16 ***
## newdata$h129.AGE09X                         < 2e-16 ***
## f.age30                                     < 2e-16 ***
## newdata$h129.SEX2                          3.03e-15 ***
```

```
## newdata$h129.AGE09X:newdata$h129.SEX2  < 2e-16 ***
## f.age30:newdata$h129.SEX2              < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6254 on 35065 degrees of freedom
## Multiple R-squared:  0.2305, Adjusted R-squared:  0.2304
## F-statistic:  2100 on 5 and 35065 DF,  p-value: < 2.2e-16
```

11.

```
# compute difference in expenditures (women - men)
difference = function(Age){
  lambda = c(0,0,0,1,Age,Age-30)
  results = esticon(NoOutlierFit, lambda)
  return(results)
}

# Create a matrix for kable, later
lexpend.mat = matrix(NA,3,3)
colnames(lexpend.mat) = c("Difference in Lexpend", "Std Error", "95% CI")
rownames(lexpend.mat) = c("40","65","80")

# Estimated diff
lexpend.mat[1, 1] = difference(40)[1,1]  # extract the estimate from diff
# SE
lexpend.mat[1,2] = difference(40)[1,2] # extract the SE from diff


CI1 = as.character(paste("(", round(difference(40)[1,7],digits = 4), round(difference(40)[1,8], digits =
lexpend.mat[1, 3] = CI1

# Estimated diff
lexpend.mat[2, 1] = difference(65)[1,1]
# SE
lexpend.mat[2,2] = difference(65)[1,2] #Std Error


CI2 = as.character(paste("(", round(difference(65)[1,7],digits = 4), round(difference(65)[1,8], digits =
lexpend.mat[2, 3] = CI2

# Estimated diff
lexpend.mat[3, 1] = difference(80)[1,1]
# SE
lexpend.mat[3,2] = difference(80)[1,2]

CI3 = as.character(paste("(", round(difference(80)[1,7],digits = 4), round(difference(80)[1,8], digits =
lexpend.mat[3, 3] = CI3

# Round all entries
for(i in 1:2){
  for(j in 1:3){
    lexpend.mat[j,i] = round(as.numeric(lexpend.mat[j,i]),digits = 4)
  }}
```

```
kable(lexpend.mat)
```

	Difference in Lexpend	Std Error	95% CI
40	0.4293	0.0095	(0.4106 0.448)
65	0.0882	0.0125	(0.0638 0.1127)
80	-0.1164	0.0195	(-0.1545 -0.0782)

12.

```
# Matrix to be used later for kable
MedEx.mat = matrix(NA,3,4)
colnames(MedEx.mat) = c("Median Diff", "Std Error (Delta)", "Std Error (bootstrap)", "95% CI")
rownames(MedEx.mat) = c("40","65","80")

############################################################
# follow same lambda idea as earlier
# where we have 1 for intercept, age, spline, etc.
#And a median estimate for men at a certain age and then subract
lambda = c(1,40,10,1,40,10)
women40 = esticon(NoOutlierFit, lambda)$estimate

lambda2 = c(1,40,10,0,0,0)
men40 = esticon(NoOutlierFit, lambda2)$estimate

MedEx.mat[1,1] = round(10^(women40) - 10^(men40), digits = 4)

lambda = c(1,65,35,1,65,35)
women65 = esticon(NoOutlierFit, lambda)$estimate

lambda2 = c(1,65,35,0,0,0)
men65 = esticon(NoOutlierFit, lambda2)$estimate

MedEx.mat[2,1] = round(10^(women65) - 10^(men65), digits = 4)

lambda = c(1,80,50,1,80,50)
women80 = esticon(NoOutlierFit, lambda)$estimate

lambda2 = c(1,80,50,0,0,0)
men80 = esticon(NoOutlierFit, lambda2)$estimate

MedEx.mat[3,1] = round(10^(women80) - 10^(men80), digits = 4)

############################################################

#Delta method to get Std Error
f.age30 = ifelse(newdata$h129.AGE09X > 30, newdata$h129.AGE09X-30, 0)

SansOutlierFit =  lm(lexp[-outliers] ~ newdata$h129.AGE09X + f.age30 + newdata$h129.SEX + newdata$h129.A

MedEx.mat[1,2] = round(deltamethod(~ 10^(x1+40*x2+10*x3+x4+40*x5+10*x6)
                                                    -10^(x1+40*x2+10*x3), coef(NoOutlierFit), vcov(NoOutli

MedEx.mat[2,2] = round(deltamethod(~ 10^(x1+65*x2+35*x3+x4+65*x5+35*x6)
```

```
                                                    -10^(x1+65*x2+35*x3), coef(NoOutlierFit), vcov(NoOutli

MedEx.mat[3,2] = round(deltamethod(~ 10^(x1+80*x2+50*x3+x4+80*x5+50*x6)
                                                    -10^(x1+80*x2+50*x3), coef(NoOutlierFit), vcov(NoOutli

############################################################

# easier to make a new mat for the bootstrap than use the kable one
newmat = cbind(newdata,lexp[-outliers],f.age30)
colnames(newmat)[16] = "lexp"
colnames(newmat)[2] = "sex"
colnames(newmat)[15] = "age"

# bootstrap function
boots = function(data, nboots=1000, age){
  diff = rep(NA,nboots)
  data[,"sex"] <- as.factor(data[,"sex"])
  for(i in 1:nboots){
    sam = sample(1:nrow(data), nrow(data), replace =T)
    fit = lm(lexp ~ age + f.age30 + sex + age*sex + f.age30*sex, data = data[sam,])
    a = round(predict(fit,newdata=data.frame(age=age, sex=as.factor(2), f.age30=ifelse(age>30,age-30,0)

    b = round(predict(fit,newdata=data.frame(age=age, sex=as.factor(1), f.age30=ifelse(age>30,age-30,0)
    diff[i]=round(10^a-10^b, digits = 4)
  }
  return(list(sd = sd(diff),lower = round(quantile(diff,0.025), digits = 4),upper = round(quantile(diff
}

b40 = boots(newmat,1000,40)
b65 = boots(newmat,1000,65)
b80 = boots(newmat,1000,80)

MedEx.mat[,3] = c(b40$sd,b65$sd,b80$sd)
MedEx.mat[,4] = c(paste(b40$lower,",",b40$upper), paste(b65$lower,",",b65$upper), paste(b80$lower,",",b8

kable(MedEx.mat)
```

	Median Diff	Std Error (Delta)	Std Error (bootstrap)	95% CI
40	608.5871	15.309	16.5869859086568	575.303 , 640.7873
65	442.4506	62.0527	69.7906404774064	311.6819 , 579.8881
80	-1275.9621	221.9534	221.231387773703	-1713.4251 , -828.8189

13. Initially, when trying to discover if men and women use roughly the same quantity of medical services, we tried to use the actual expenditure quantities; however, because the equivariance and normality assumptions are clearly violated, the actual expenditure data cannot be reliably analyzed without making a transformation to the log-expenditure data.

A sufficient transformation is taking log base ten of all of the expenditures. After this transformation, a reliable model can be attained.

Based on a scatterplot of age vs. log10(Expenditures), it is clear that a fitted model must allow for the slope to change around age 30, and since the focal-point of this study is to determine whether or not men and women use the same amount of medical services, the model must also take into account the subject's ginger.

Thus, the most appropriate model is:

$$\log(exp) = \beta_0 + \beta_1(Age) + \beta_2(Age + 30) + \beta_3(Gender) + \beta_4(Gender * Age) + \beta_5(Gender + s(Age + 30))$$

Where all, β_i are significant to the model.

Next, it became clear through the use of a DFBETAs plot that even with the transformation, the dataset has 1.784 influential outliers. Once eliminated, there were 35,071 people individuals left to analyze using the model. The question to answer remained the same, and we wanted to determine if men and women use the same amount of medical services.

With these premises in mind the results from the model indicate that there is indeed a difference between the amount of medical expenditures that men and women use across age groups. Initially, in their youth, men and women use roughly the same amount of medical expenditures, but as they age, the gap widens. This makes sense, as childhood medical services are consistent across the genders (vaccinations, check-ups, etc.). However, as women age, they begin to spend more than men. This difference peaks around age 30 (roughly when most women are having children) and then begins to taper off.

Then, at higher ages (around 80), the data suggests that men surpass women in medical expenditures. For example at age 40, the median expenditure amount for women is expected to be $629.09 more then men, having about 95% confidence that that value is between ($577, $639). However, later on, the median difference between males and females is -$1275.96 with a 95% confidence interval that the true median difference is between (-$1694 , -$836.). This means that men use more medical expenditures than women between 40 and 85 years of age.

Note that these median calculations are done by carefully transforming the model given above back to the original scale using an exponential back transform (section 3.9 and 6.8 notes reference e^{β_i}).

The $beta_3$ coefficient specifically allows for a one time change in average log(expenditures). The estimate for this coefficient is significant and has a value of -7.89. Transforming this variable back to the original scale $10\hat{}\beta_3 - 100$ gives the following result. That all other things held equal, by simply being a women you will have less medical expenditures. This is just an example (for illustrative purposes) because the other variables allow for changes in this median difference across different ages.

Overall, the key takeaway from this study is that across ages, men and women do not use the same quantity of medical services and that the average transformed medical expenditure data is not the same across all ages for men and women.

```
summary(NoOutlierFit)
```

```
##
## Call:
## lm(formula = lexp[-outliers] ~ newdata$h129.AGE09X + f.age30 +
##     newdata$h129.SEX + newdata$h129.AGE09X * newdata$h129.SEX +
##     f.age30 * newdata$h129.SEX)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.69724 -0.43983 -0.04645  0.41689  2.81514
##
## Coefficients:
##                                            Estimate Std. Error t value
## (Intercept)                               2.7881641  0.0134630 207.098
## newdata$h129.AGE09X                      -0.0175113  0.0006513 -26.886
## f.age30                                   0.0469476  0.0009681  48.494
## newdata$h129.SEX2                        -0.1416418  0.0179443  -7.893
## newdata$h129.AGE09X:newdata$h129.SEX2     0.0235778  0.0008581  27.476
## f.age30:newdata$h129.SEX2                -0.0372192  0.0012642 -29.441
```

```
##                                         Pr(>|t|)
## (Intercept)                              < 2e-16 ***
## newdata$h129.AGE09X                      < 2e-16 ***
## f.age30                                  < 2e-16 ***
## newdata$h129.SEX2                       3.03e-15 ***
## newdata$h129.AGE09X:newdata$h129.SEX2    < 2e-16 ***
## f.age30:newdata$h129.SEX2                < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6254 on 35065 degrees of freedom
## Multiple R-squared:  0.2305, Adjusted R-squared:  0.2304
## F-statistic:  2100 on 5 and 35065 DF,  p-value: < 2.2e-16
```

14.

```
multiple.regression = function(x, y) {

 # Matrix of feature variables from data
  X = as.matrix(x)

  # vector of ones with same length as rows in data
  intercept = rep(1, length(y))

  # Add intercept column to X
  X = cbind(intercept, X)

  # find betas
  betas = solve(t(X) %*% X) %*% t(X) %*% y

  # Round for easier viewing
  betas = round(betas, 4)

  return(betas)
}

multiple.regression(lexp, dat$h129.AGE09X)
```

```
##              [,1]
## intercept  2.2327
##           11.0144
```