STATS 201 Assignment 2

Liu Siyuan 2019210173

Due Date: 2021.11.7

Question 1

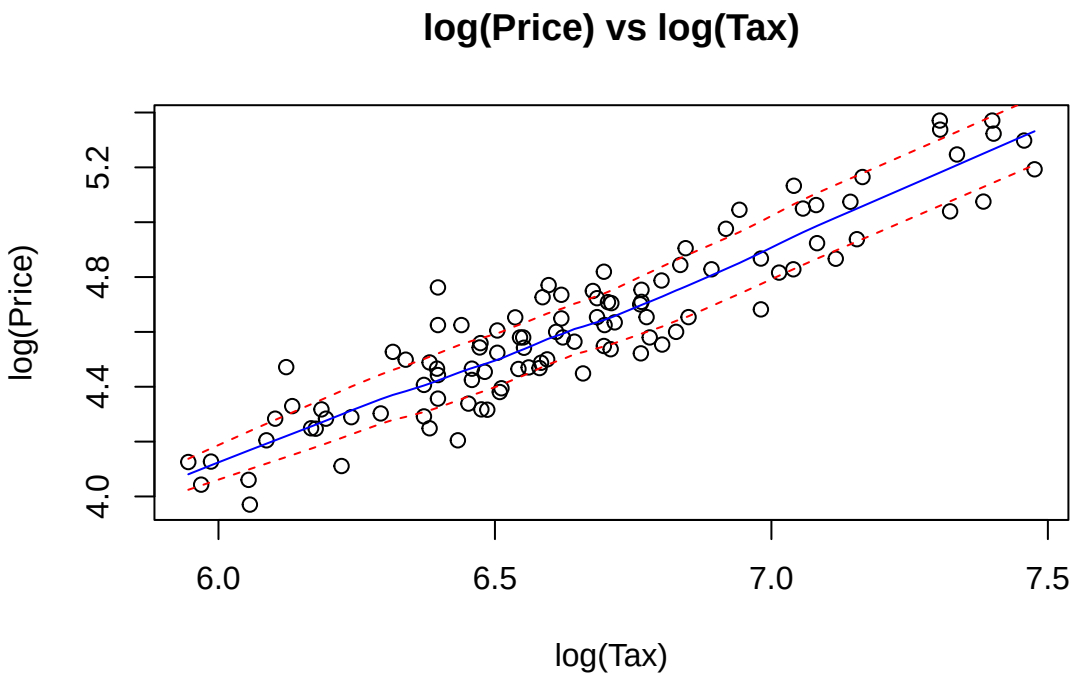
Question of interest/goal of the study

We want to build a model to explain the sale price of houses using their annual city tax bill (similar idea to rates in New Zealand) for houses in Albuquerque, New Mexico. In particular, we are interested in estimating the effect on sales price for houses which differ in city tax bills by 1% and 50%.

Read in and inspect the data:

```
hometax.df=read.csv("hometax.csv")

trendscatter(Price~Tax,main="Price vs Tax",data=hometax.df)
```

```
trendscatter(log(Price)~log(Tax),main="log(Price) vs log(Tax)",data=hometax.df)
```



Comment

Comparing the two images, it is obvious that the log log model data points are more evenly distributed next to the fitted curve, and most of them are within the confidence interval, while the linear fitting data points are unevenly distributed. The obvious data points on the left are more clustered, and the data points on the right are more scattered. The upper interval width is large and the lower interval width is small, so there will be too large variance.

Justify why a log-log (power) model is appropriate here.

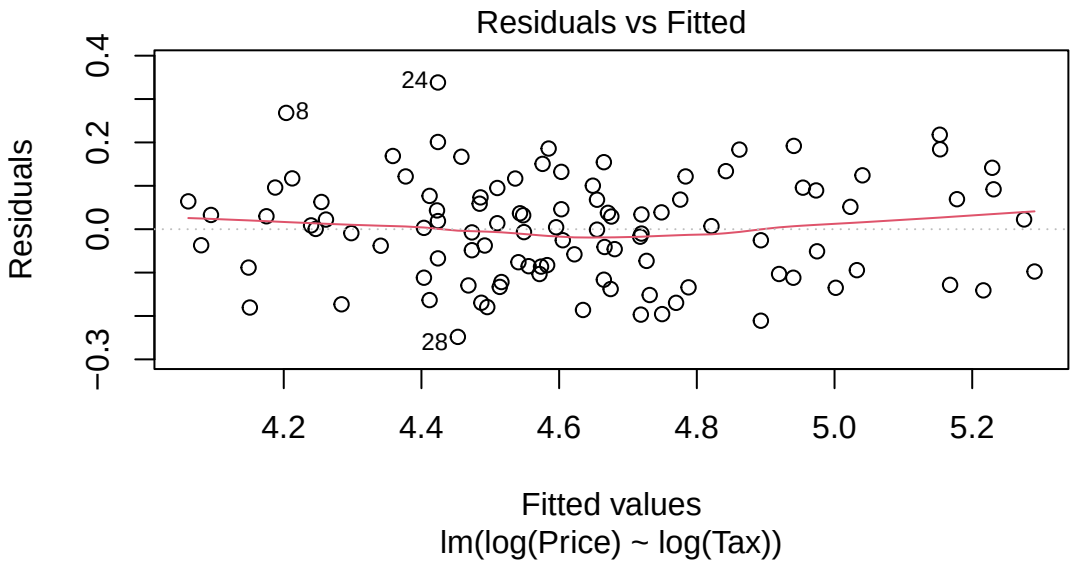
Comparing the two images, it is obvious that the log-log model data points are more evenly distributed next to the fitted line, and most of them are within the confidence interval, while the linear fitting distribution is uneven, and the variance will be too large. So the log-log model is more fit.

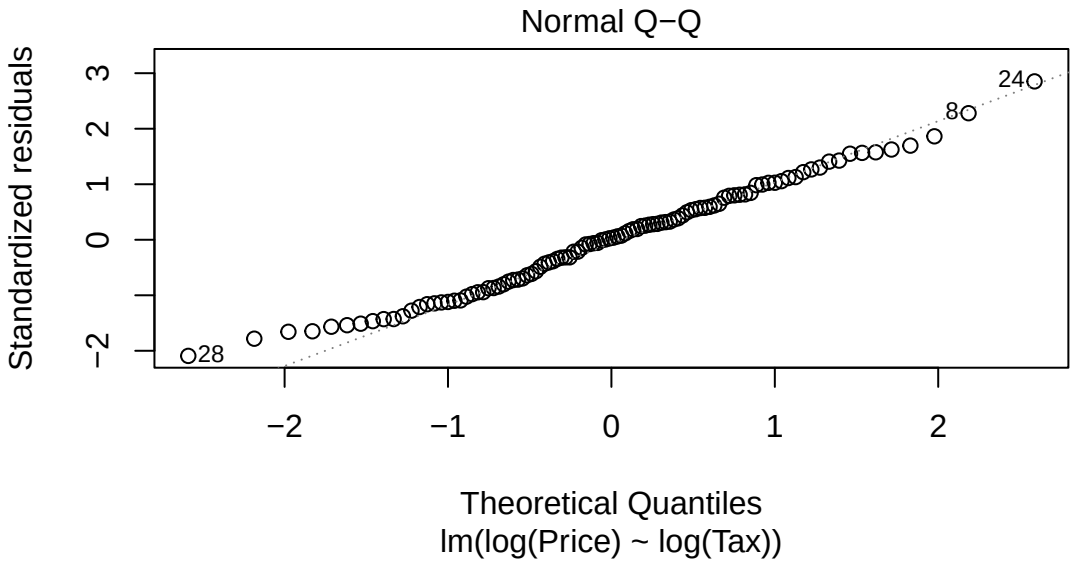
From the image distribution of linear fitting, X and Y show a right skew distribution. It is obvious that the data in the left area is more dense, the interval of independent variables is small, the interval of independent variables in the right area is large, and the growth range of dependent variables is large, which is more in line with the power-law model.

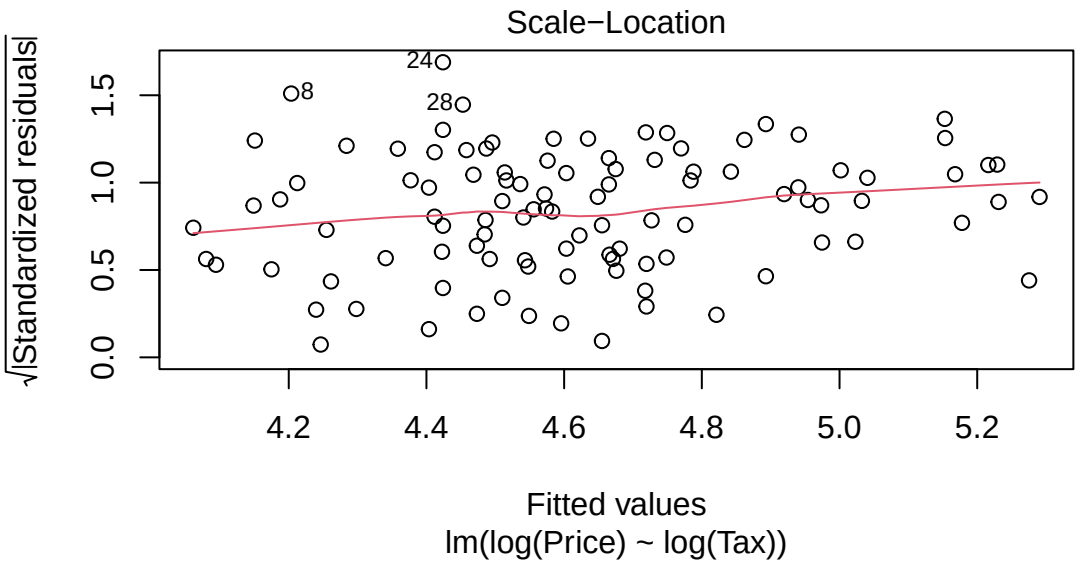
From the perspective of median, both X and Y show large median data, which is relatively scattered, which is in line with the application principle of power-law model.

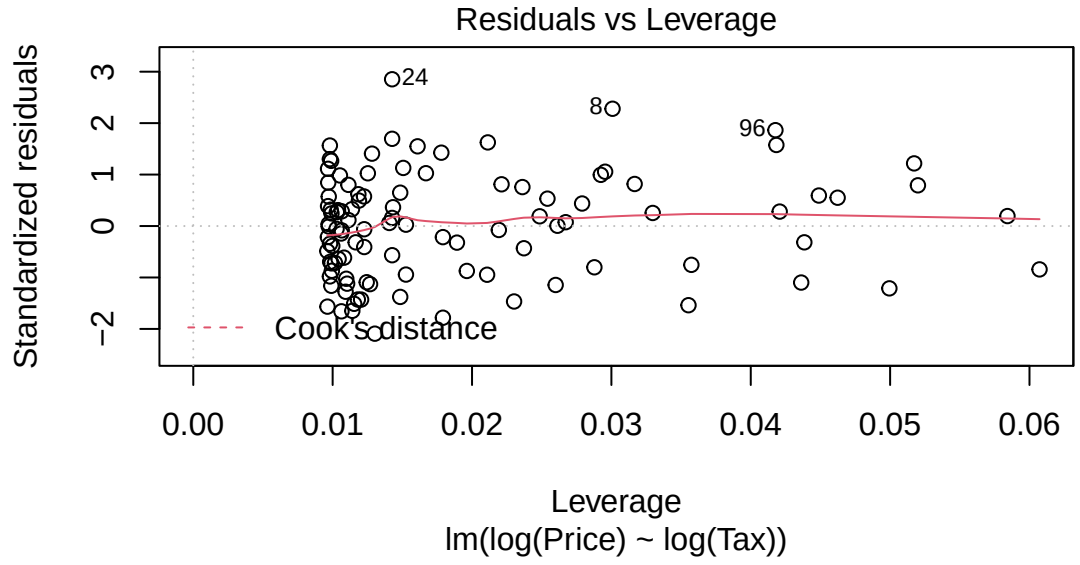
Fit model and check assumptions.

```
hometax.lm = lm(log(Price)~log(Tax), data = hometax.df)
plot(hometax.lm)
```

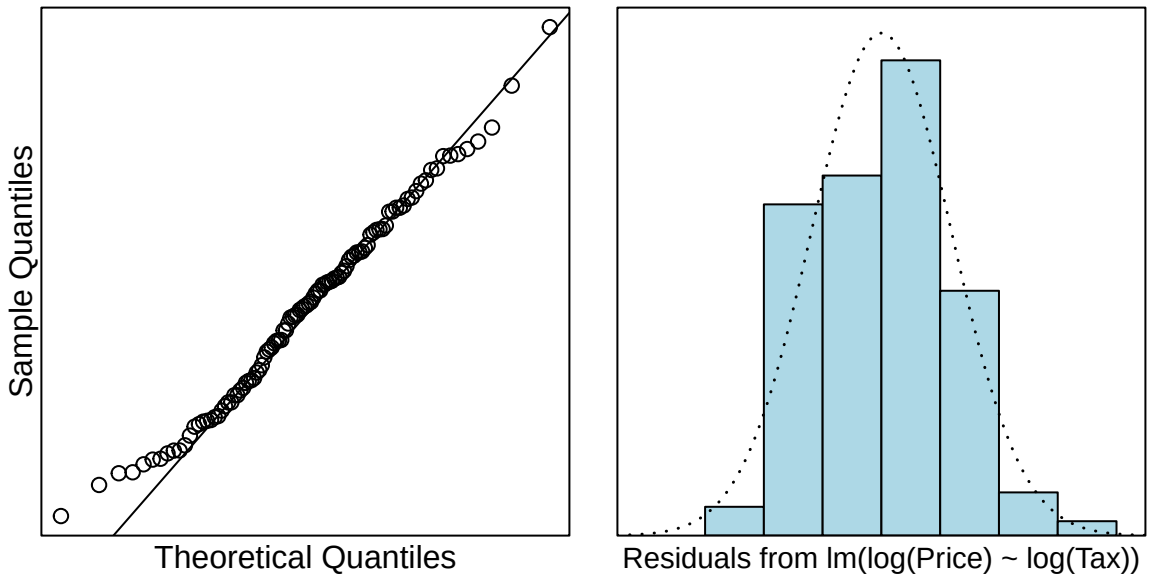




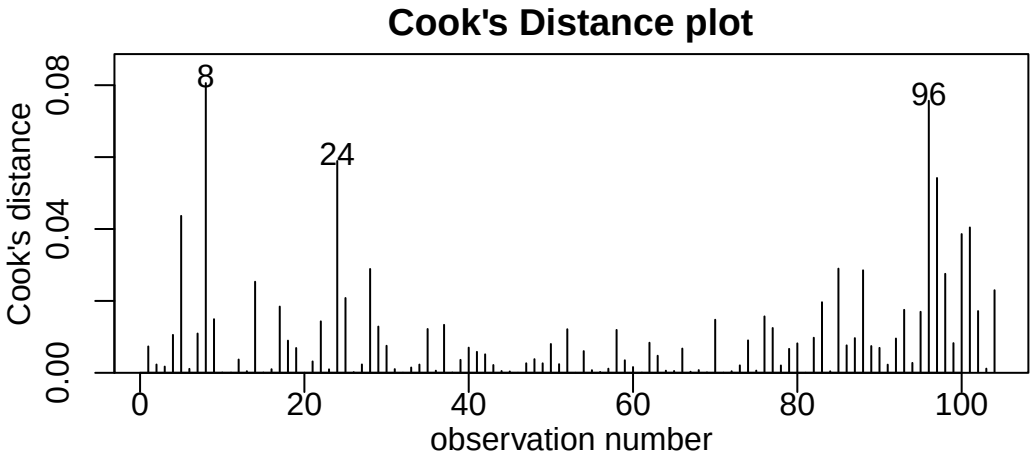




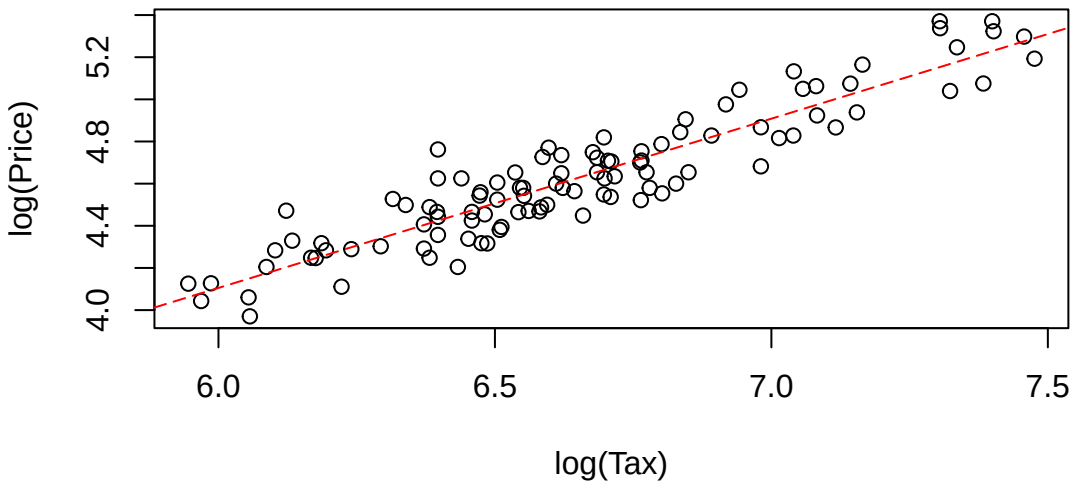

```
normcheck(hometax.lm)
```



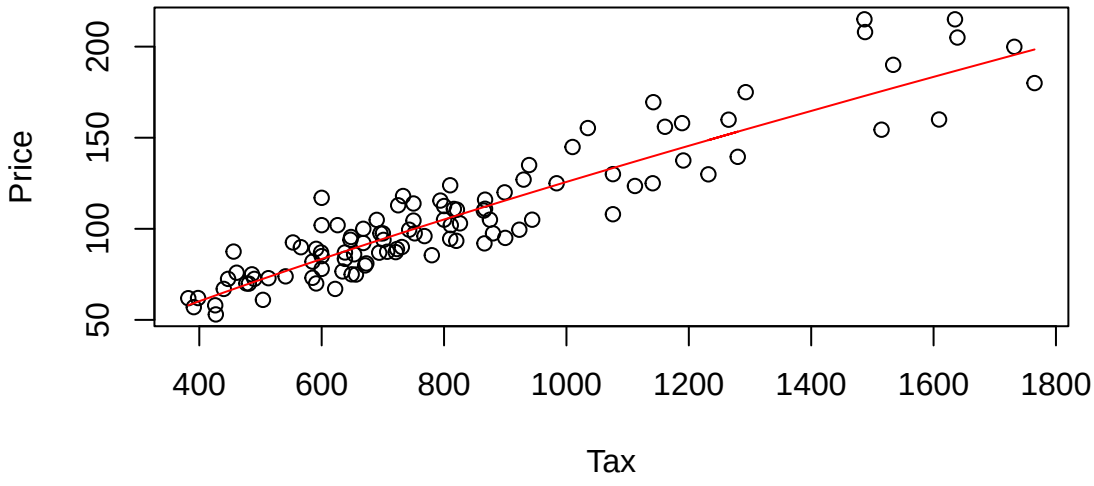

```
cooks20x(hometax.lm)
```




```
plot(log(Price)~log(Tax), data = hometax.df)
abline(coef(hometax.lm), lty = 5, col = "red")
```




```
plot(Price~Tax, data = hometax.df)
hometax.pred = exp(predict(hometax.lm, hometax.df))
lines(hometax.df$Tax, hometax.pred, col = "red")
```




```
summary(hometax.lm)
```

```
##
## Call:
## lm(formula = log(Price) ~ log(Tax), data = hometax.df)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.24820 -0.09519  0.00380  0.07994  0.33821
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.71348    0.21679  -3.291  0.00137 **
## log(Tax)     0.80311    0.03257  24.660  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1194 on 102 degrees of freedom
## Multiple R-squared:  0.8564, Adjusted R-squared:  0.855
## F-statistic: 608.1 on 1 and 102 DF,  p-value: < 2.2e-16
```

```
1.01 ^ confint(hometax.lm)[2, ]
```

```
##    2.5 %   97.5 %
## 1.007376 1.008671
```

```
1.5 ^ confint(hometax.lm)[2, ]
```

```
##    2.5 %   97.5 %
## 1.349105 1.421660
```

Methods and assumption checks

We have a random sample of 104 houses sold in Albuquerque, so we can assume they form an independent and representative sample.

The scatter plot of the Price vs Tax showed clear nonlinearity and an increase in variability with tax. Compared with the trendscatter plot of the log(Price) vs log(Tax), we believe that the power law linear model may more fit the sample.

Then, we plot the residual diagram for the EOV. The residuals show patternless scatter with fairly constant variability - so no problems. After this, the normaility checks don't show any major problems and the Cook's plot doesn't reveal any further unduly influential points. Overall, all the model assumptions are satisfied.

Our model is:

$$log(Price_i) = \beta_0 + \beta_1 \times log(Tax_i) + \epsilon_i$$

where$\epsilon_i \sim iid\ N(0, \sigma^2)$

Our model explains 85.64% of the total variation in the response variable, and so will be reasonable for prediction.

Executive Summary

We are interested in estimating the effect on sales price for houses which differ in city tax bills by 1% and 50%.

We fit a positive power law linear model to sample.

And we know that the p-value is less than 2e-16, which means we have extremely evidence that suggest the power law linear relationship between price and tax exists.

We have 95% of the confidence that increasing tax by 1% corresponds to an increase in median price between 0.73% and 0.86%.

We have 95% of the confidence that increasing tax by 50% corresponds to an increase in median price between 34.91% and 42.16%.

We can obtain the Multiple R-squared is equal to 0.8564.

In conclusion, our model explains 85.64% of the total variation in the response variable, and so will be reasonable for prediction.

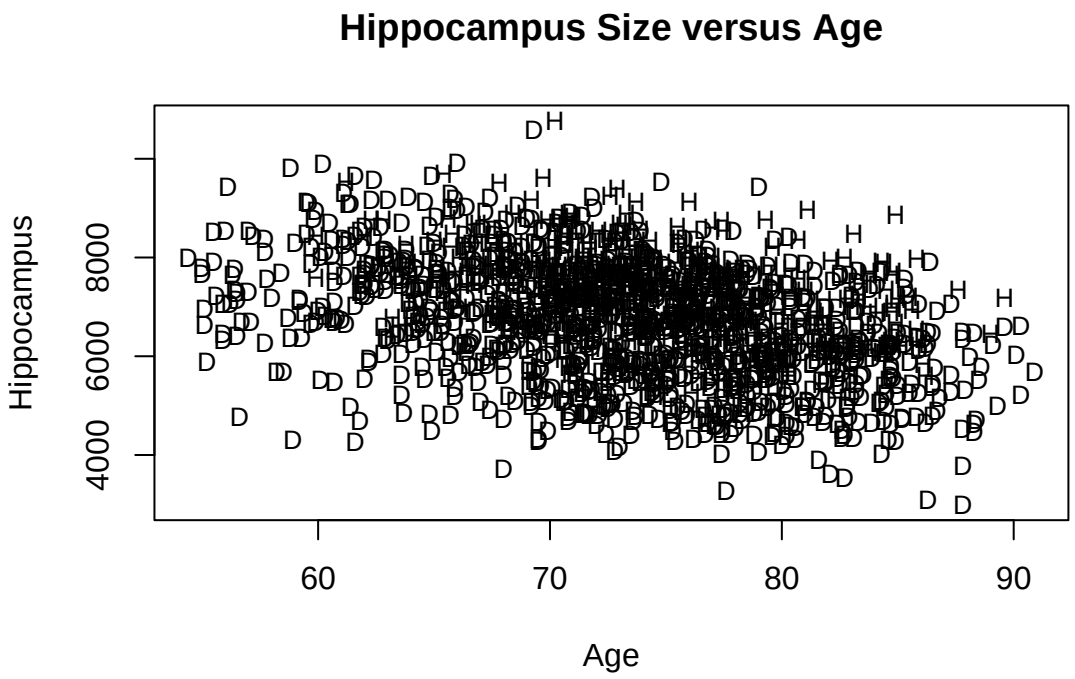
Question 2

Question of interest/goal of the study

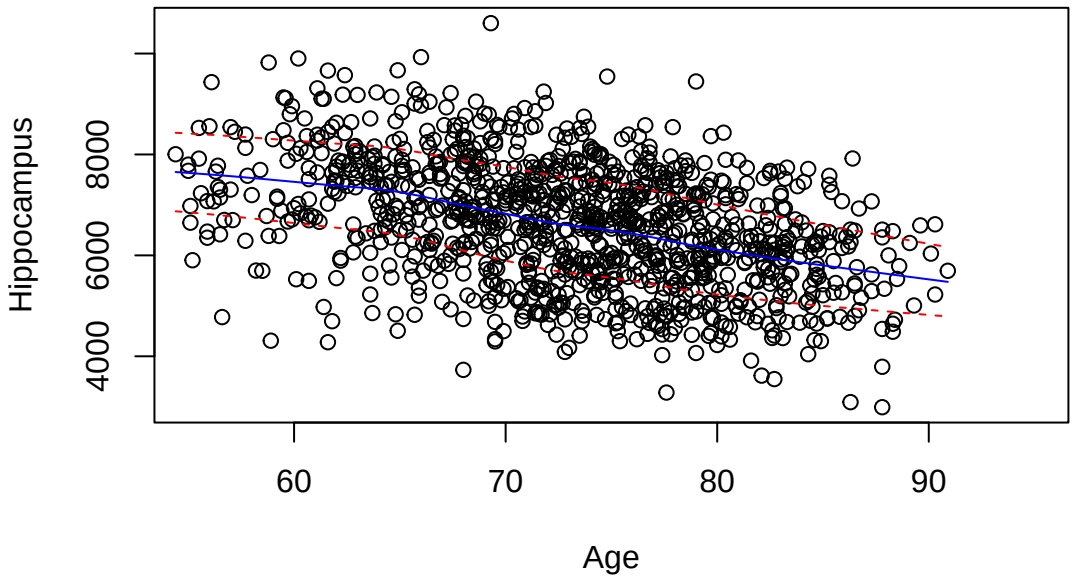
We want to explore the relationship between hippocampus size and age. In particular, we are interested in whether the relationship differs between healthy individuals and individuals with dementia related symptoms.

Read in and inspect the data:

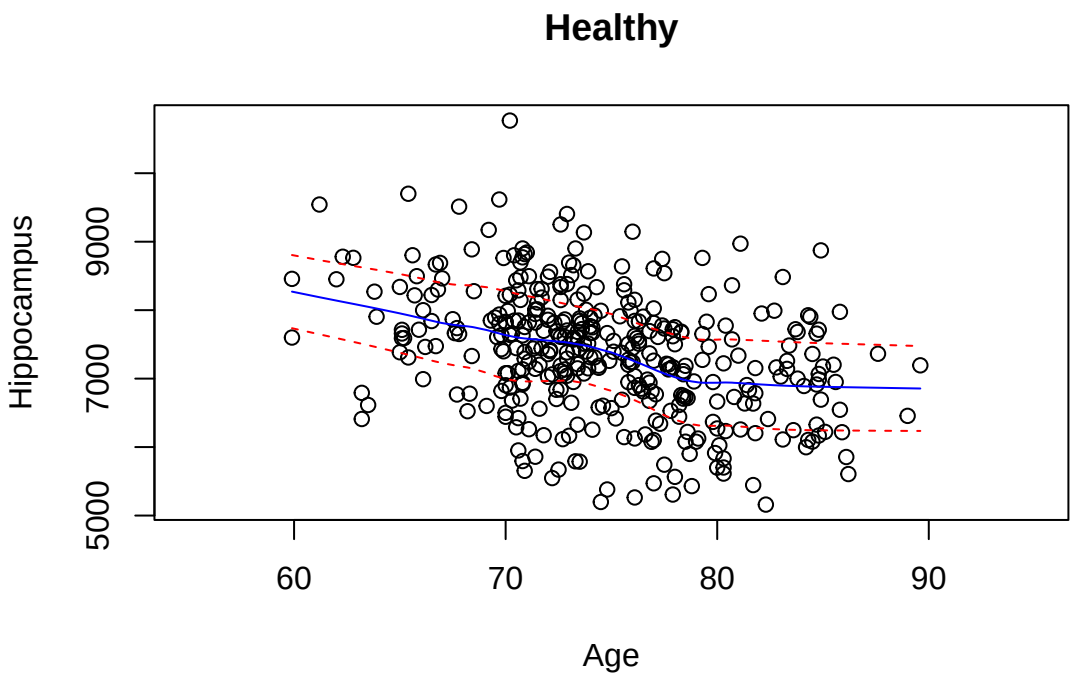
```
Hippocampus.df<-read.csv("Hippocampus.csv")
plot(Hippocampus~Age,main="Hippocampus Size versus Age",type="n",data=Hippocampus.df)
text(Hippocampus.df$Age, Hippocampus.df$Hippocampus, Hippocampus.df$AD, cex=.8)
```




```
trendscatter(Hippocampus~Age,data=Hippocampus.df[Hippocampus.df$AD=="D",],xlim=c(55,95),main="Dementia")
```



```
trendscatter(Hippocampus~Age,data=Hippocampus.df[Hippocampus.df$AD=="H",],xlim=c(55,95),main="Healthy")
```



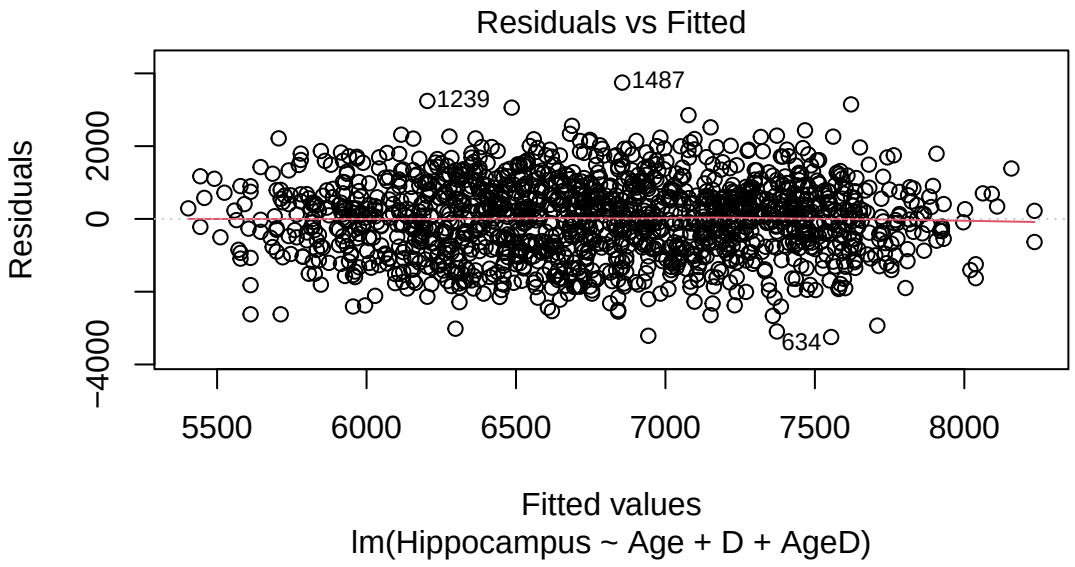
Comment

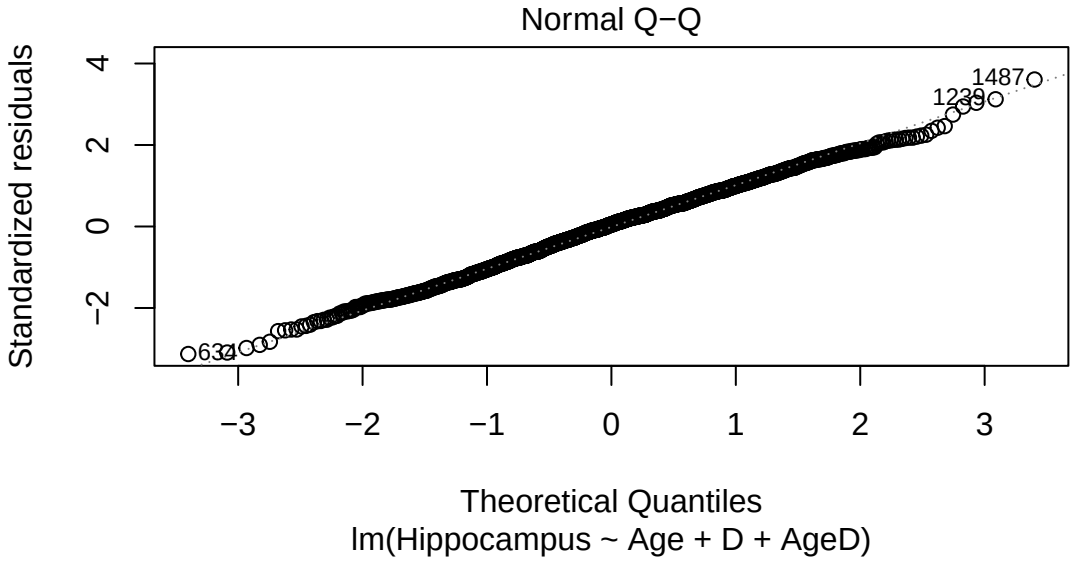
The first plot is a bit cluttered.

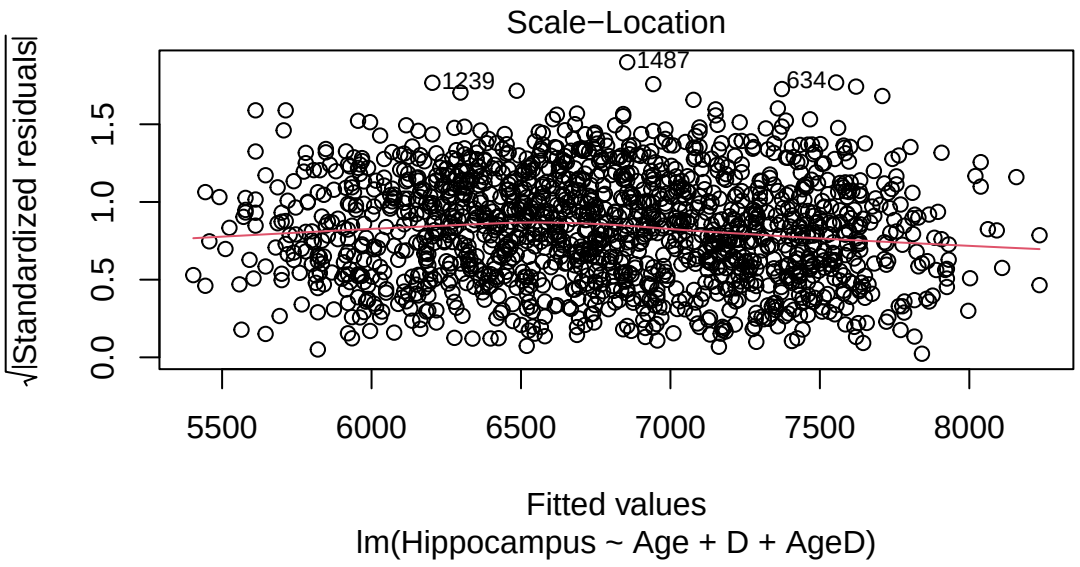
Looking at the two cases separately, the overall trend of the two graphs is like linear distribution (decreasing), but it is obvious that some other scattered points are distributed outside the confidence interval. Compared with the two graphs, the graphs with dementia category have more and more dense points, and the graphs with health category have fewer and more scattered points.

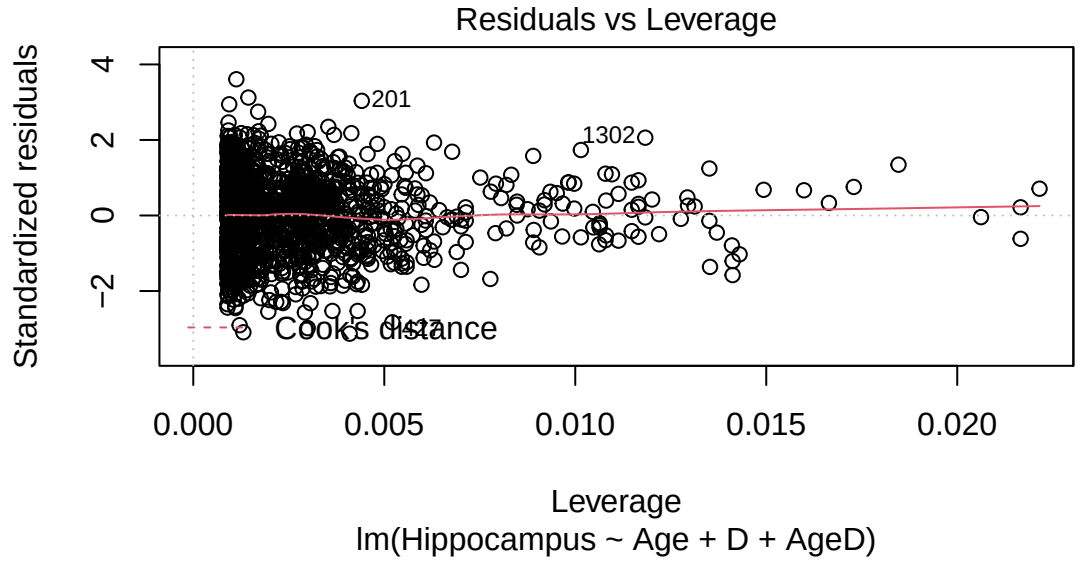
Fit model and check assumptions.

```
Hippocampus.df$D = as.numeric(Hippocampus.df$AD == "D")
Hippocampus.df$AgeD = with(Hippocampus.df, {AgeD = D * Age})

Hippocampus.fit = lm(Hippocampus~Age+D+AgeD, data = Hippocampus.df)

plot(Hippocampus.fit)
```

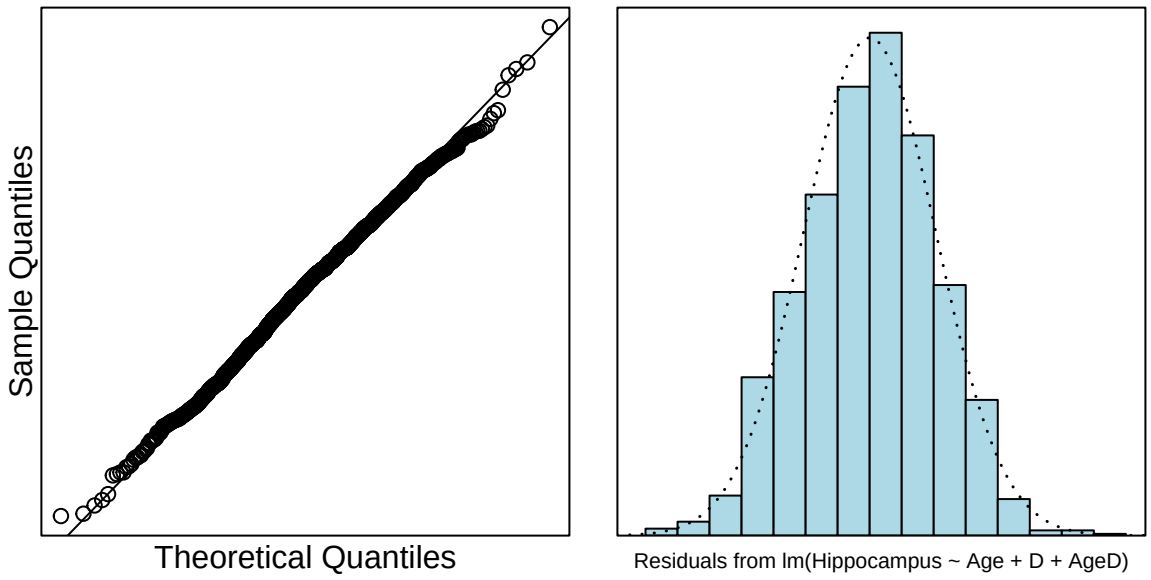

Residuals vs Fitted
Residuals
1239
1487
634
−4000
0
2000
5500
6000
6500
7000
7500
8000
Fitted values
lm(Hippocampus ~ Age + D + AgeD)


Normal Q−Q
Standardized residuals
1487
1239
634
−2
0
2
4
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(Hippocampus ~ Age + D + AgeD)

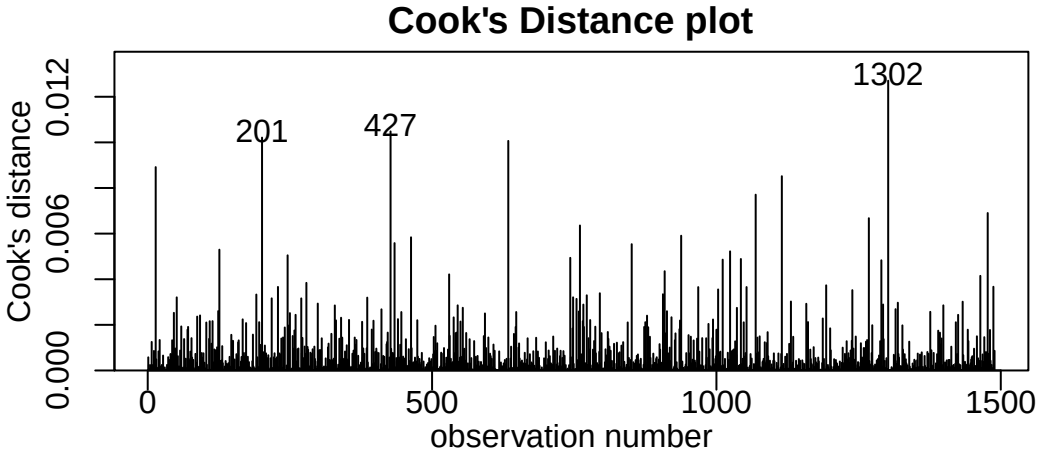





```
normcheck(Hippocampus.fit)
```




```
cooks20x(Hippocampus.fit)
```




```
summary(Hippocampus.fit)
```

```
##
## Call:
## lm(formula = Hippocampus ~ Age + D + AgeD, data = Hippocampus.df)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -3245.4  -729.8    52.1   701.9  3746.6
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11804.655    726.341  16.252  < 2e-16 ***
## Age           -59.595      9.703  -6.142 1.04e-09 ***
## D            -291.487    787.293  -0.370    0.711
## AgeD           -7.617     10.546  -0.722    0.470
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1039 on 1485 degrees of freedom
## Multiple R-squared:  0.2328, Adjusted R-squared:  0.2313
## F-statistic: 150.2 on 3 and 1485 DF,  p-value: < 2.2e-16
```

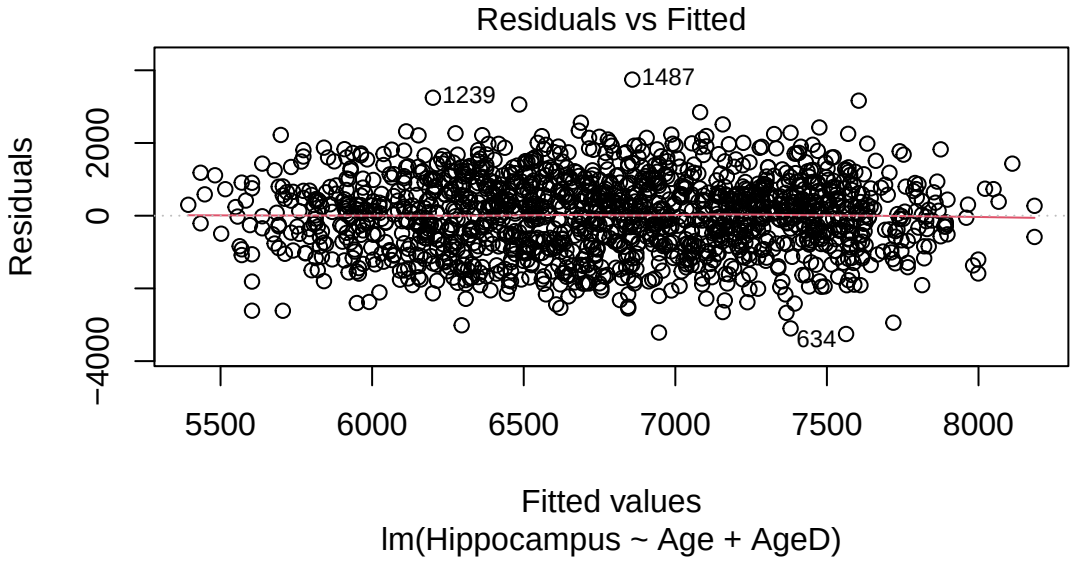
```
confint(Hippocampus.fit)
```

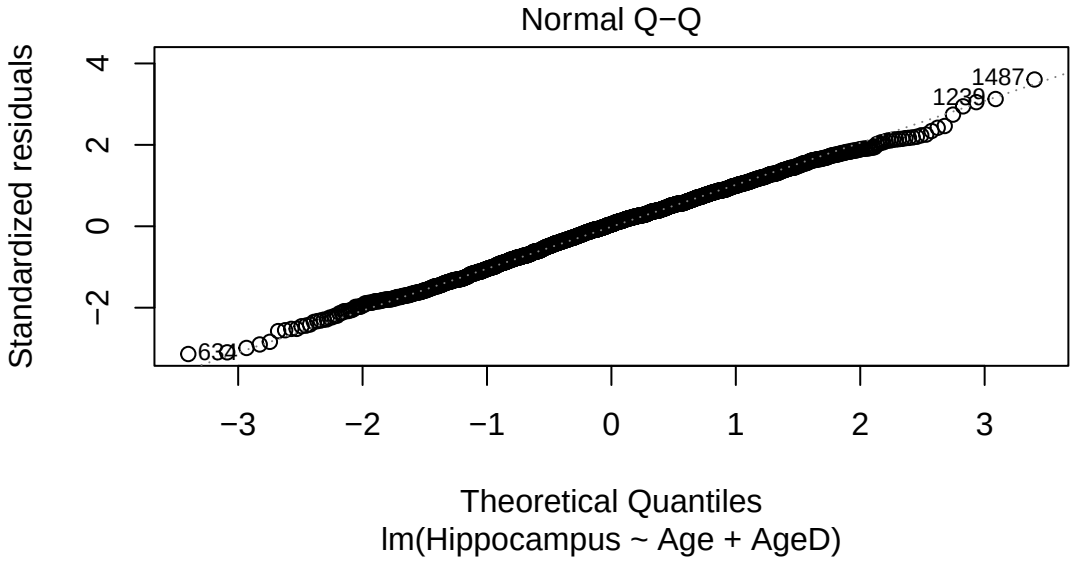
```
##                   2.5 %      97.5 %
## (Intercept) 10379.89125 13229.41829
## Age           -78.62767   -40.56297
## D           -1835.81096  1252.83698
## AgeD          -28.30305    13.06983
```

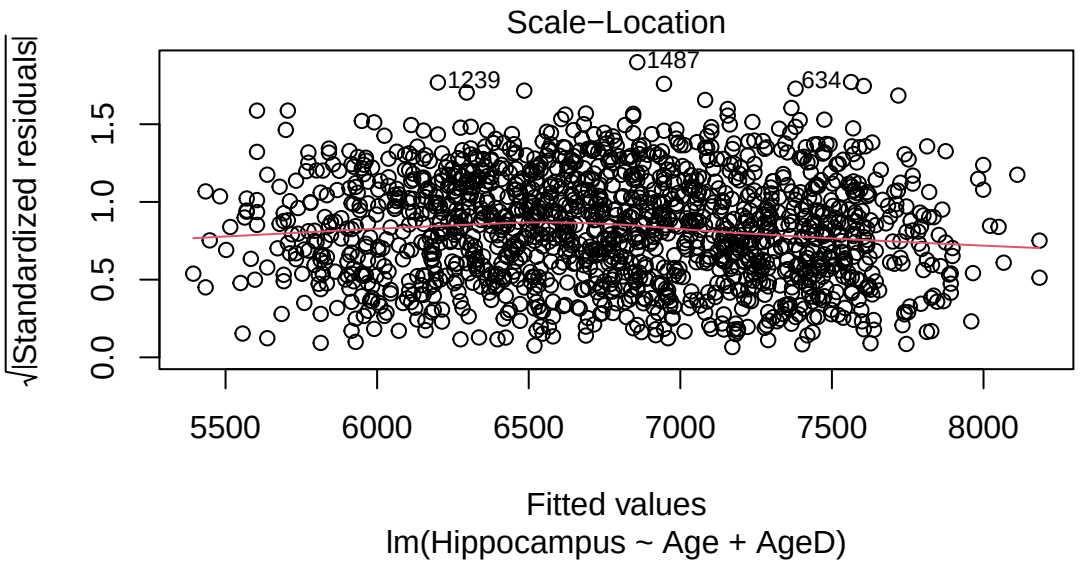
By observing the summary of model fitting, we can observed that we only explains 23.28% of the total variation in the response variable. The p- value of D coefficient and AgeD coefficient is relatively large, that is, it is more likely to be 0. Therefore, we consider deleting one of the parameters or deleting the two parameters, and then judge whether the model can be better explained by comparing the Multiple R-squared.

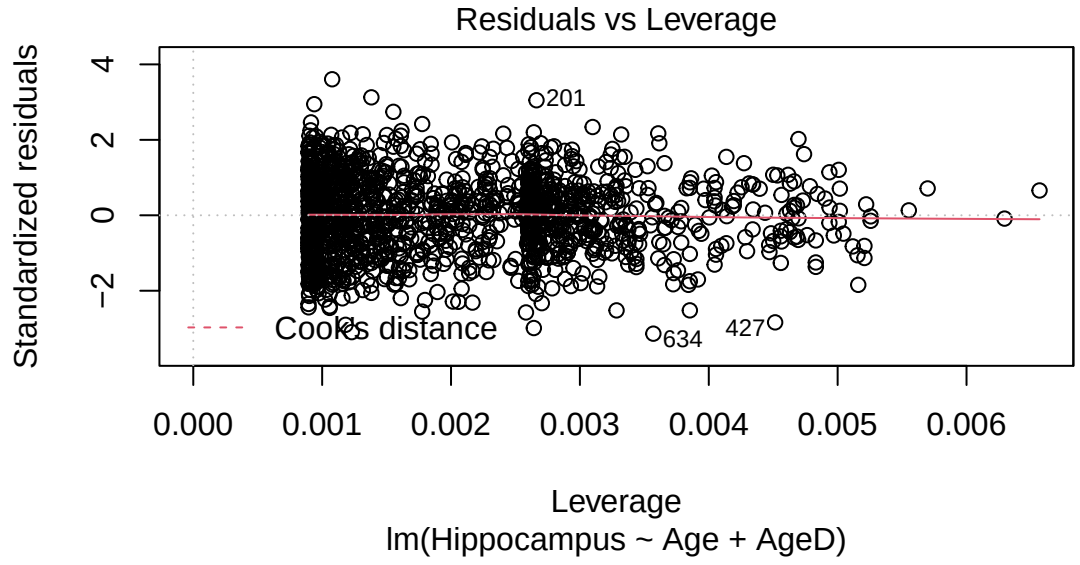
Fit model with delete D coefficient and check assumptions.

```
Hippocampus.fit1 = lm(Hippocampus~Age+AgeD, data = Hippocampus.df)

plot(Hippocampus.fit1)
```

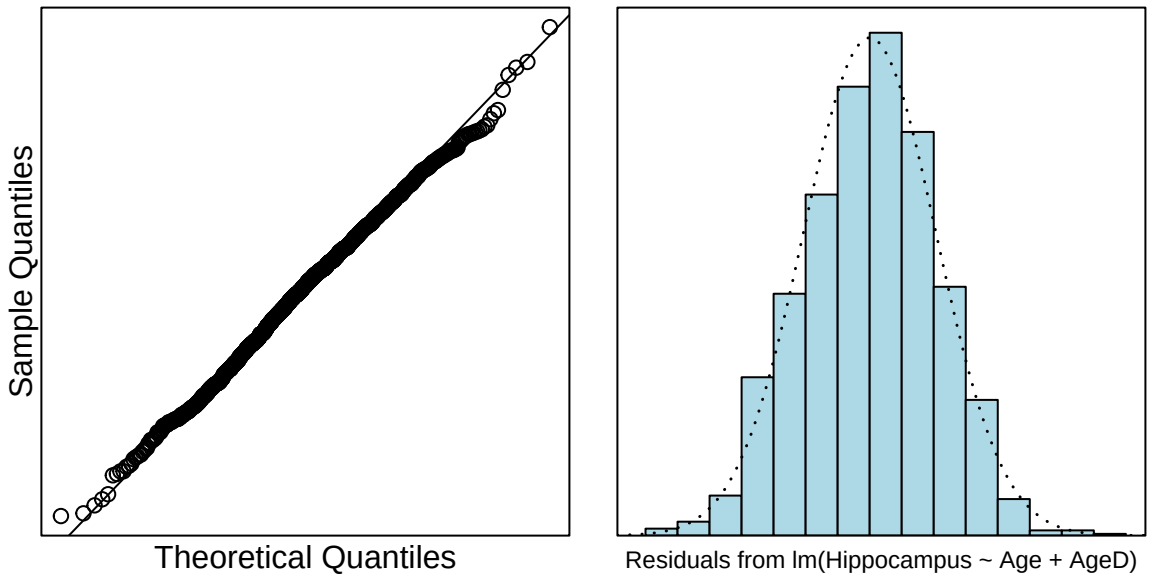

Residuals vs Fitted
Residuals
Fitted values
lm(Hippocampus ~ Age + AgeD)
1239
1487
634
−4000
0
2000
5500
6000
6500
7000
7500
8000


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Hippocampus ~ Age + AgeD)
1487
1239
634
−2
0
2
4
−3
−2
−1
0
1
2
3

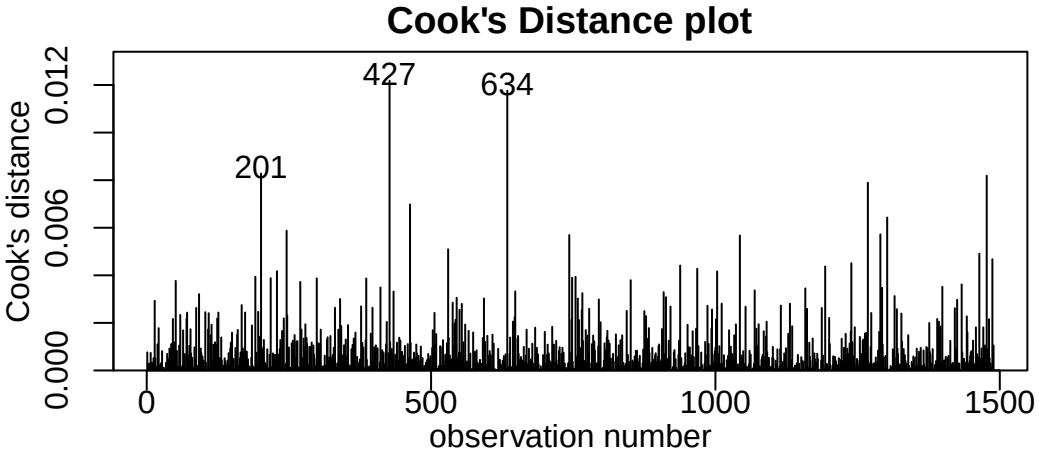





```
normcheck(Hippocampus.fit1)
```




```
cooks20x(Hippocampus.fit1)
```




```
summary(Hippocampus.fit1)
```

```
##
## Call:
## lm(formula = Hippocampus ~ Age + AgeD, data = Hippocampus.df)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -3254.2  -733.8    55.0   709.7  3743.9
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11556.5541   280.1440   41.25   <2e-16 ***
## Age           -56.2902     3.8005  -14.81   <2e-16 ***
## AgeD          -11.5088     0.8359  -13.77   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1039 on 1486 degrees of freedom
## Multiple R-squared:  0.2327, Adjusted R-squared:  0.2317
## F-statistic: 225.4 on 2 and 1486 DF,  p-value: < 2.2e-16
```

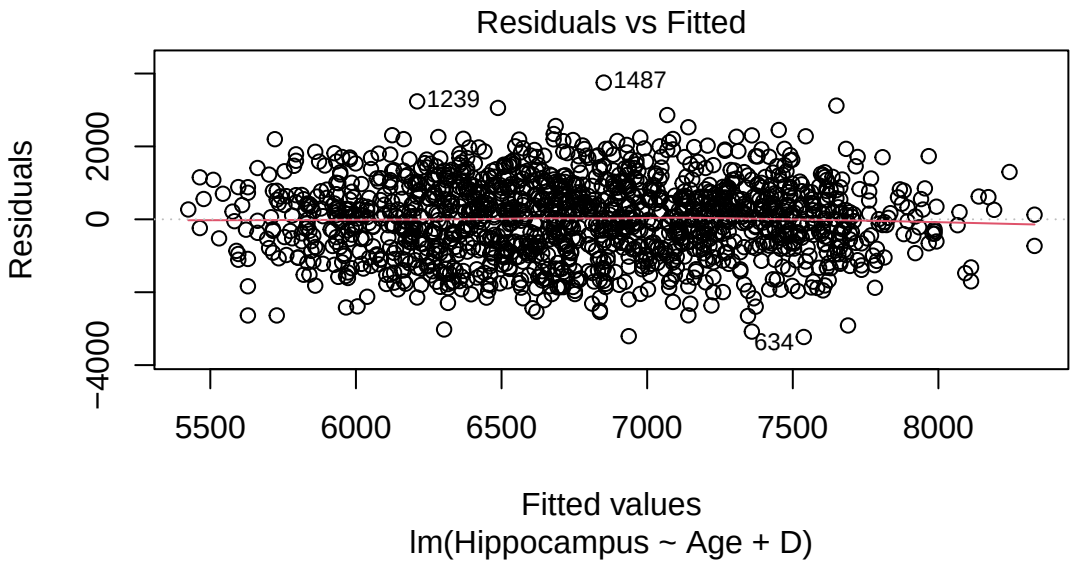
```
confint(Hippocampus.fit1)
```

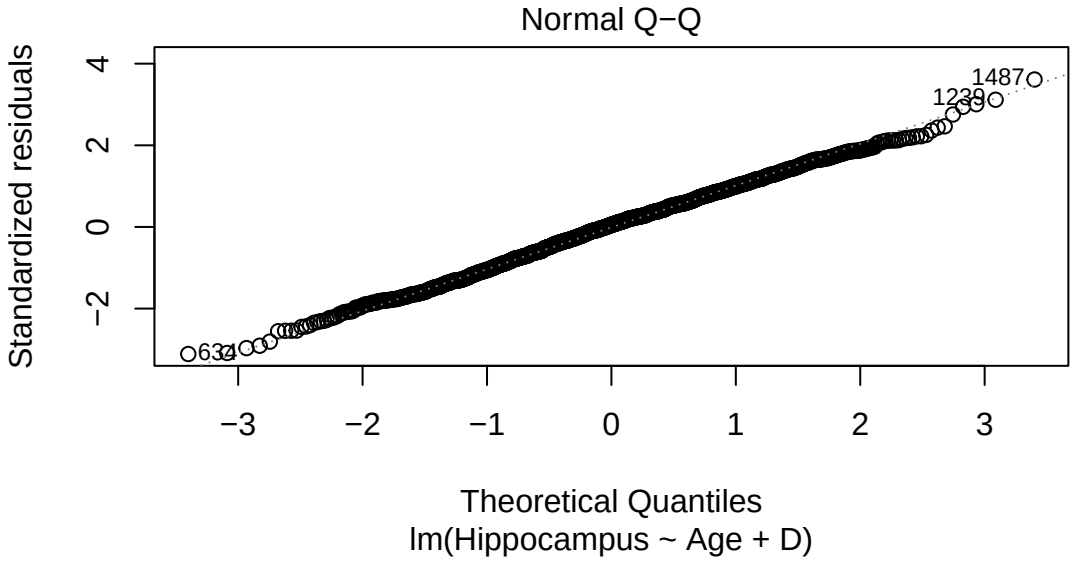
```
##                   2.5 %       97.5 %
## (Intercept) 11007.03440 12106.073810
## Age           -63.74514   -48.835329
## AgeD          -13.14855    -9.869105
```

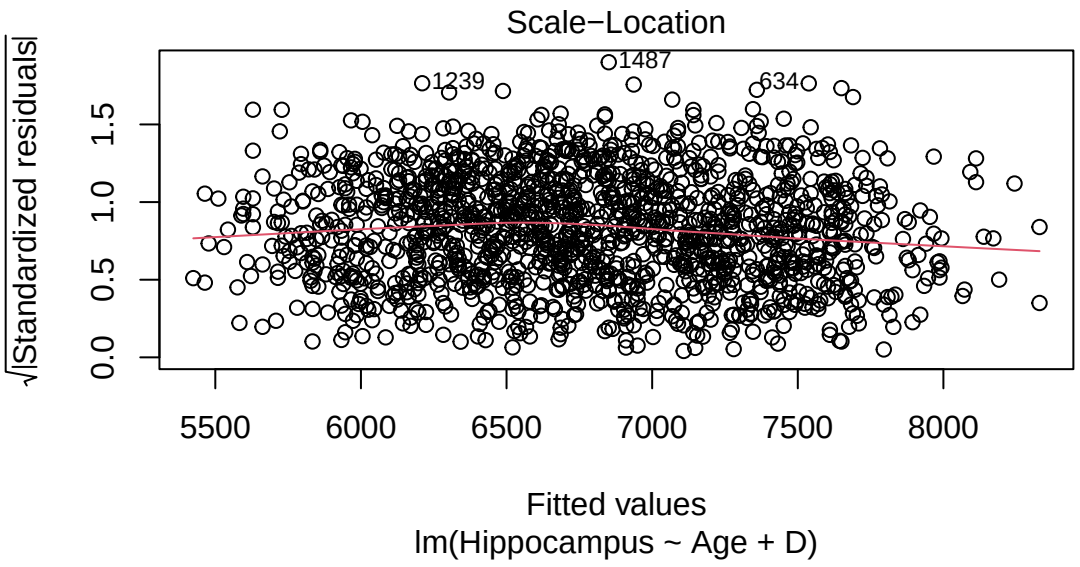
Observing the summary of deleting the D coefficient, the difference of Multiple R-squared is very small, but the p-value of coefficient AgeD becomes very small, indicating that the parameters of the model exist and the model fitting is relatively reasonable.

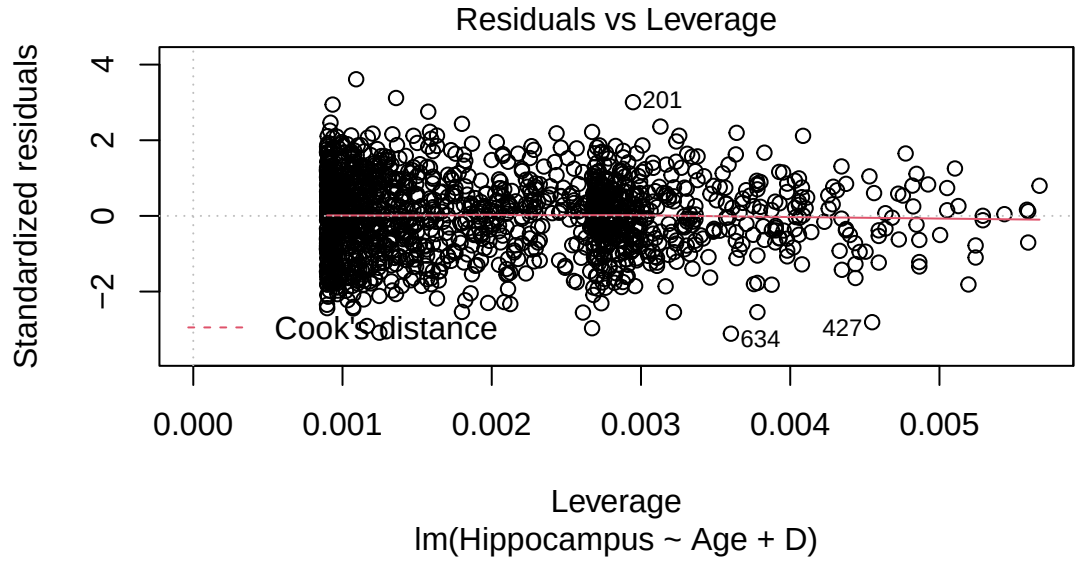
Fit model with delete AgeD coefficient and check assumptions.

```
Hippocampus.fit2 = lm(Hippocampus~Age+D, data = Hippocampus.df)

plot(Hippocampus.fit2)
```

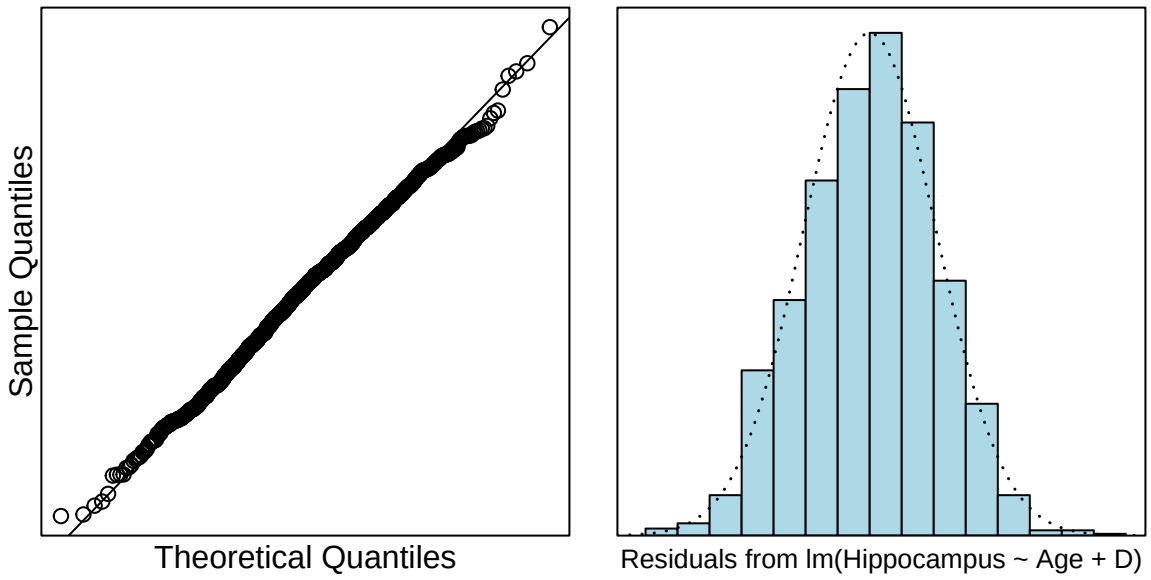

Residuals vs Fitted
Residuals
1239
1487
634
−4000
0
2000
5500
6000
6500
7000
7500
8000
Fitted values
lm(Hippocampus ~ Age + D)


Normal Q−Q
Standardized residuals
1487
1239
634
−2
0
2
4
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(Hippocampus ~ Age + D)

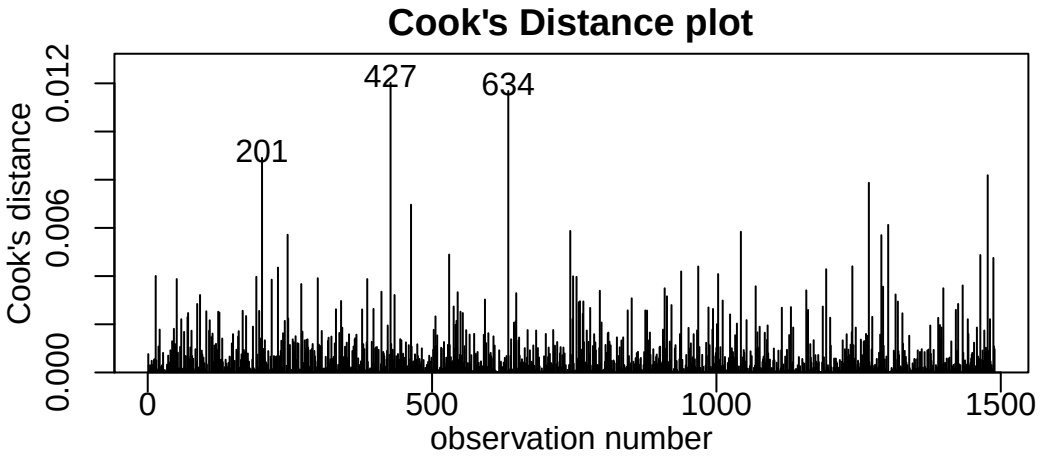





```
normcheck(Hippocampus.fit2)
```




```
cooks20x(Hippocampus.fit2)
```




```
summary(Hippocampus.fit2)
```

```
##
## Call:
## lm(formula = Hippocampus ~ Age + D, data = Hippocampus.df)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -3228.8  -727.2    54.5   705.0  3751.1
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12285.971    288.825   42.54   <2e-16 ***
## Age           -66.043      3.801  -17.37   <2e-16 ***
## D            -858.307     62.413  -13.75   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1039 on 1486 degrees of freedom
## Multiple R-squared:  0.2325, Adjusted R-squared:  0.2315
## F-statistic: 225.1 on 2 and 1486 DF,  p-value: < 2.2e-16
```

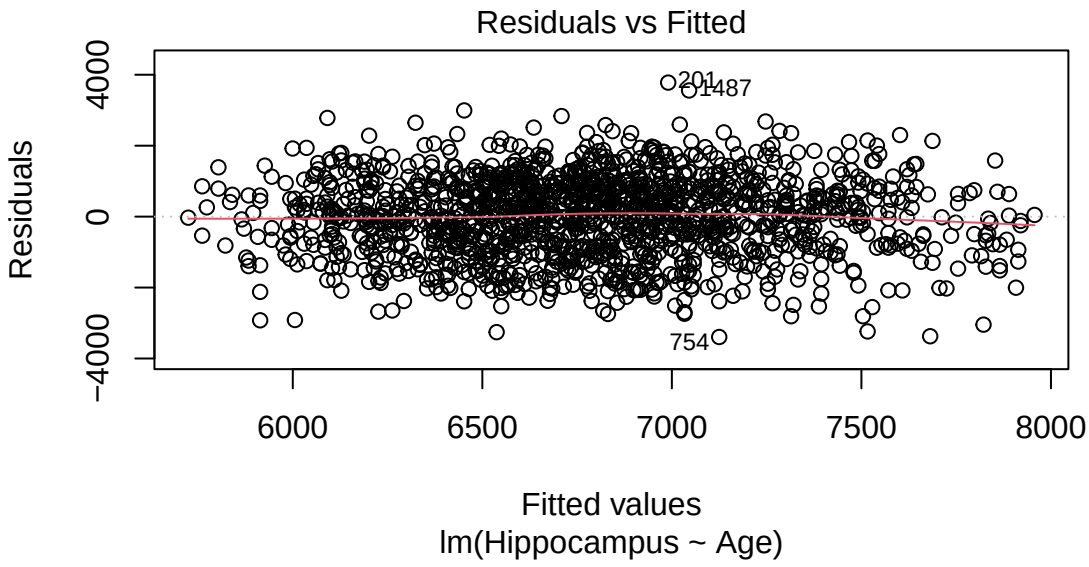
```
confint(Hippocampus.fit2)
```

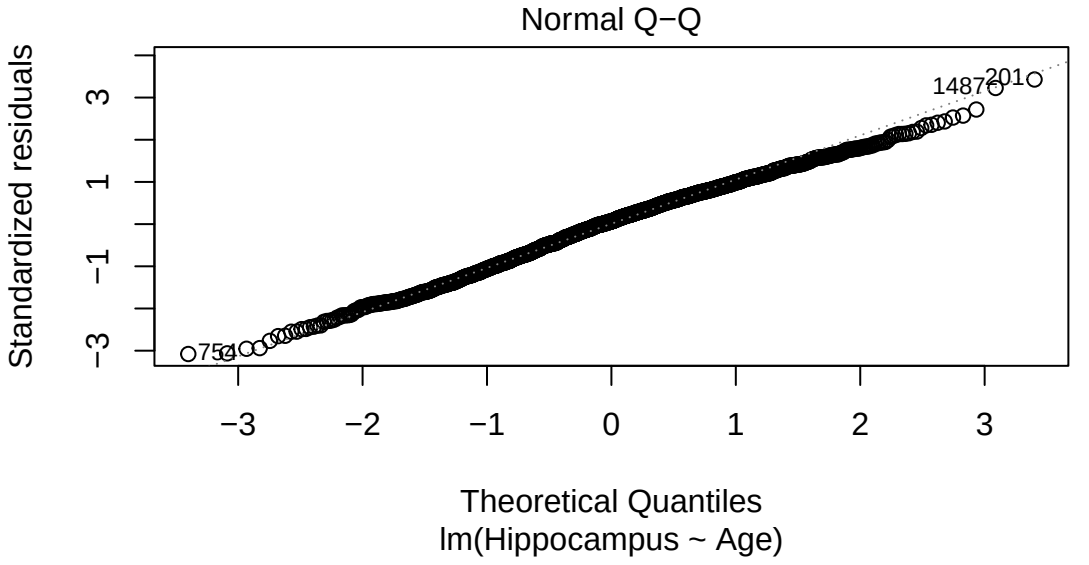
```
##                   2.5 %      97.5 %
## (Intercept) 11719.42285 12852.51958
## Age           -73.49872   -58.58644
## D            -980.73460  -735.87923
```

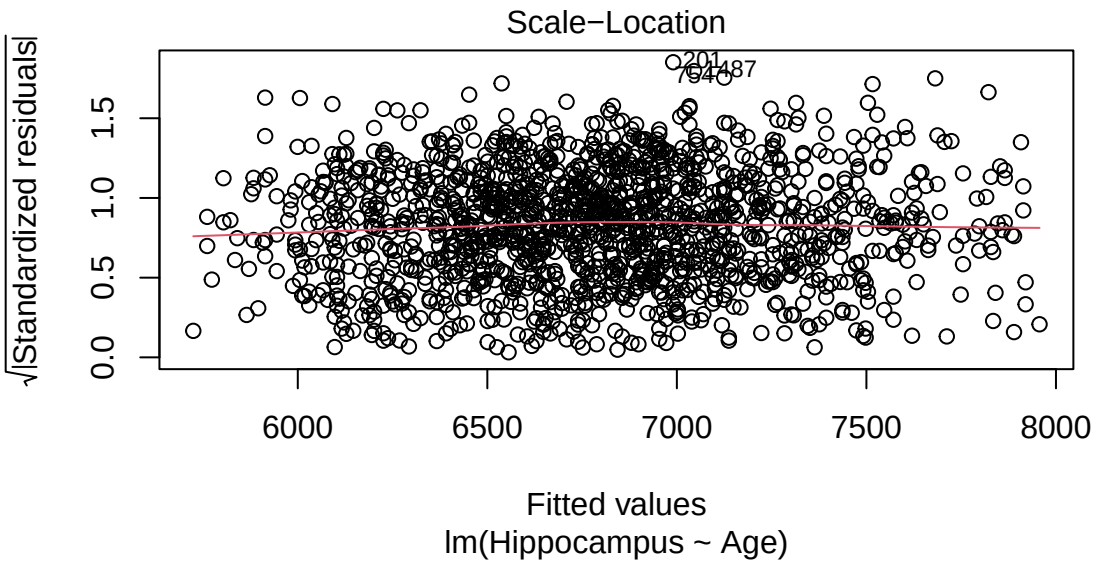
Observing the summary of deleting the AgeD coefficient, the difference of Multiple R-squared is very small, but the p-value of coefficient D becomes very small, indicating that the parameters of the model exist and the model fitting is relatively reasonable.

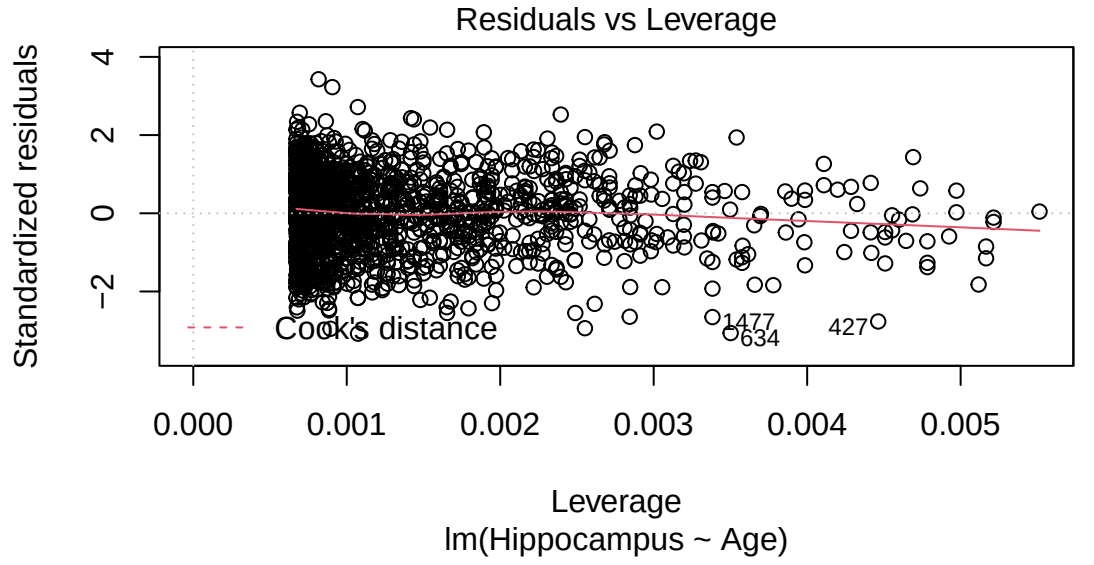
Fit model with delete both AgeD and D coefficient and check assumptions.

```
Hippocampus.fit3 = lm(Hippocampus~Age, data = Hippocampus.df)

plot(Hippocampus.fit3)
```

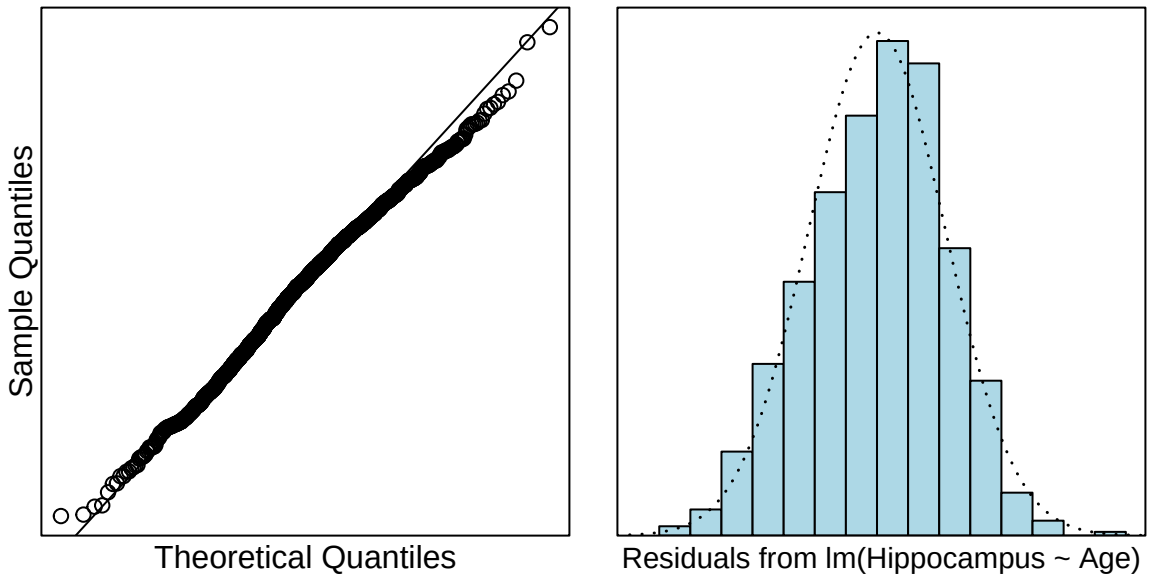

Residuals vs Fitted
Residuals
Fitted values
lm(Hippocampus ~ Age)
201
1487
754
4000
0
−4000
6000
6500
7000
7500
8000


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Hippocampus ~ Age)
201
1487
754
3
1
−1
−3
−3
−2
−1
0
1
2
3

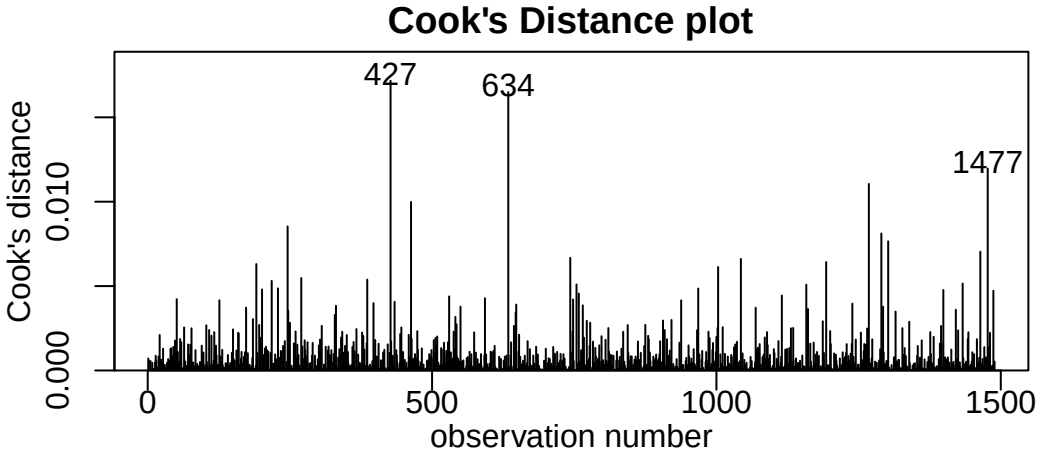





```
normcheck(Hippocampus.fit3)
```




```
cooks20x(Hippocampus.fit3)
```




```
summary(Hippocampus.fit3)
```

```
##
## Call:
## lm(formula = Hippocampus ~ Age, data = Hippocampus.df)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -3393.9  -769.2    76.1   788.1  3778.7
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11283.689    296.630   38.04   <2e-16 ***
## Age           -61.159      4.017  -15.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1103 on 1487 degrees of freedom
## Multiple R-squared:  0.1349, Adjusted R-squared:  0.1343
## F-statistic: 231.8 on 1 and 1487 DF,  p-value: < 2.2e-16
```

```
confint(Hippocampus.fit3)
```

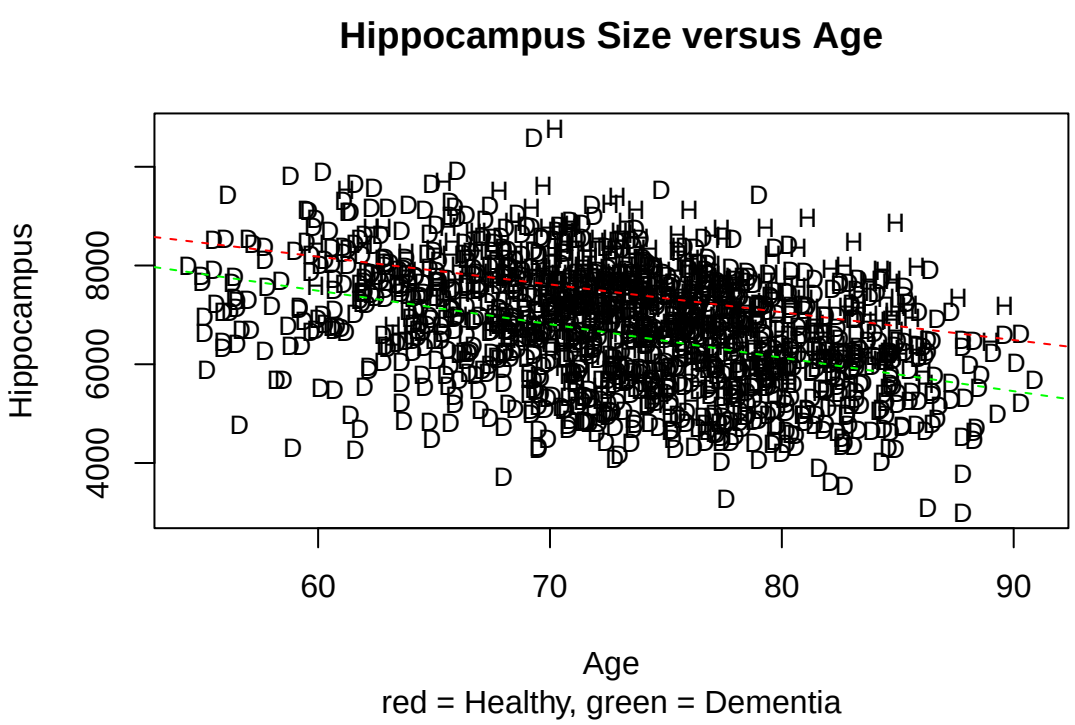
```
##                   2.5 %      97.5 %
## (Intercept) 10701.83170 11865.54699
## Age           -69.03786   -53.27964
```

Observing the summary of deleting both AgeD and D coefficient, the Multiple R-squared is smaller than the previous, so it is not used.

Lastly, relatively speaking, **the fitting degree of the model with D equal to 0 is better**, so next we choose this model for fitting and summary.

Plot the data with your appropriate model superimposed over it

```
plot(Hippocampus~Age,main="Hippocampus Size versus Age",sub="red = Healthy, green = Dementia",type="n",c
text(Hippocampus.df$Age, Hippocampus.df$Hippocampus, Hippocampus.df$AD, cex=.8)

b1 = coef(Hippocampus.fit1)
abline(b1[1:2], lty = 2, col = "red") #Health
abline(b1[1], b1[2] + b1[3], lty = 2, col = "green") #Dementia
```



Methods and assumption checks

We have a random sample of 1489 individuals who is Health or Dementia, so we can assume they form an independent and representative sample.

The scatter plot of the Hippocampus vs Age showed clear linearity and an decrease in variability with age. Hippocampus size are both influenced by numeric variable Age and the factor variable Healthy or Dementia. And we in order to find the best fit to the sample, we fitted four models, of which we selected a linear model with explanatory variables x1(Age), x2(AD * Age) (interaction) to our data.

Then, we plot the residual diagram for the EOV. The residuals show patternless scatter with fairly constant variability - so no problems. After this, the normaility checks don't show any major problems and the Cook's plot doesn't reveal any further unduly influential points. Overall, all the model assumptions are satisfied.

Our model is

$$Hippocampus_i = \beta_0 + \beta_1 \times Age_i + \beta_2 \times D_i \times Age_i + \epsilon_i$$

where $\epsilon_i \sim iid\ N(0, \sigma^2)$ The D is 1 when the individual is Dementia, and D is 0 when the individual is Healthy.

Our model explains 23.27% of the total variation in the response variable, and so it is not very good for prediction.

Executive Summary

In this question, we are interested in studying the relationship between Hippocampus size and age. In particular, we are interested in whether the relationship differs between healthy individuals and individuals with dementia related symptoms.

We fit a linear regression model (decreasing) both Healthy and Dementia to our sample with explanatory variables x1(Age), x2(AD * Age)(interaction).

And we know that the p-value is less than 2e-16, which means we have extremely evidence that suggest the linear relationship between Hippocampus size and Age exists.

We have 95% of the confidence that when each age increasing, the excepted value of Hippocampus size decrease are between 48.8 and 63.7 MRI and the estimated mean of their in Hippocampus size are between 11007 and 12106 MRI in Healthy individuals.

We have 95% of the confidence that when each age increasing, the excepted value of Hippocampus size decrease are between (48.8 + 9.87) and (63.7 + 13.1) MRI, that is (56.7, 76.8) in Dementia individuals.

We can obtain the Multiple R-squared is equal to 0.2327.

In conclusion, our model explains 23.27% of the total variation in the response variable, and so it is not very good for prediction.

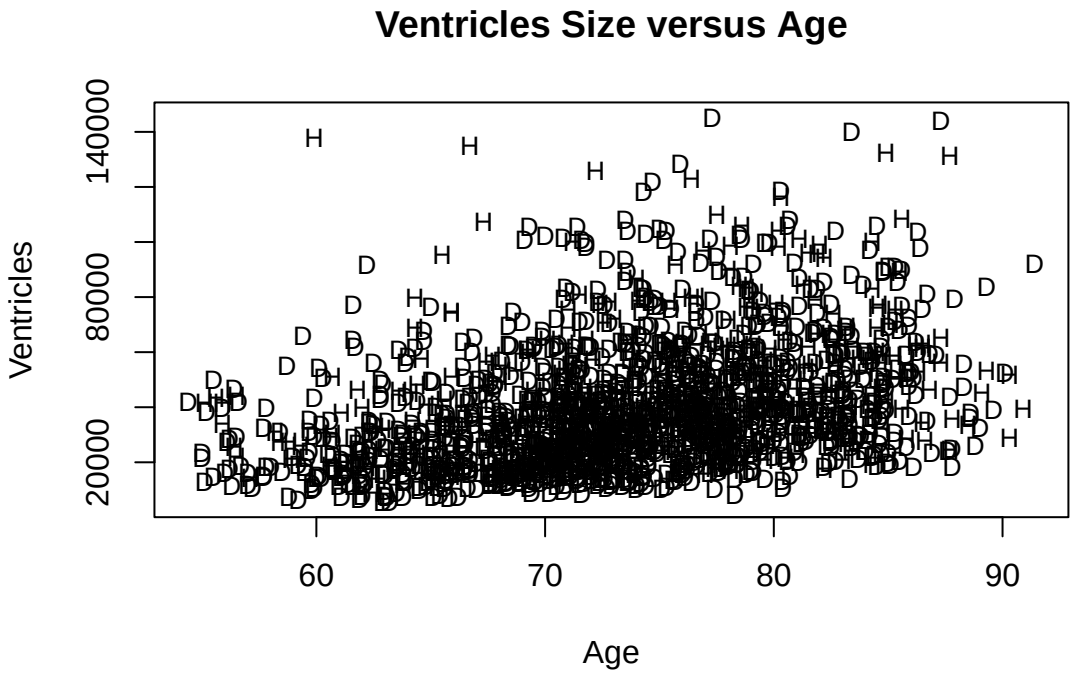
Question 3

Question of interest/goal of the study

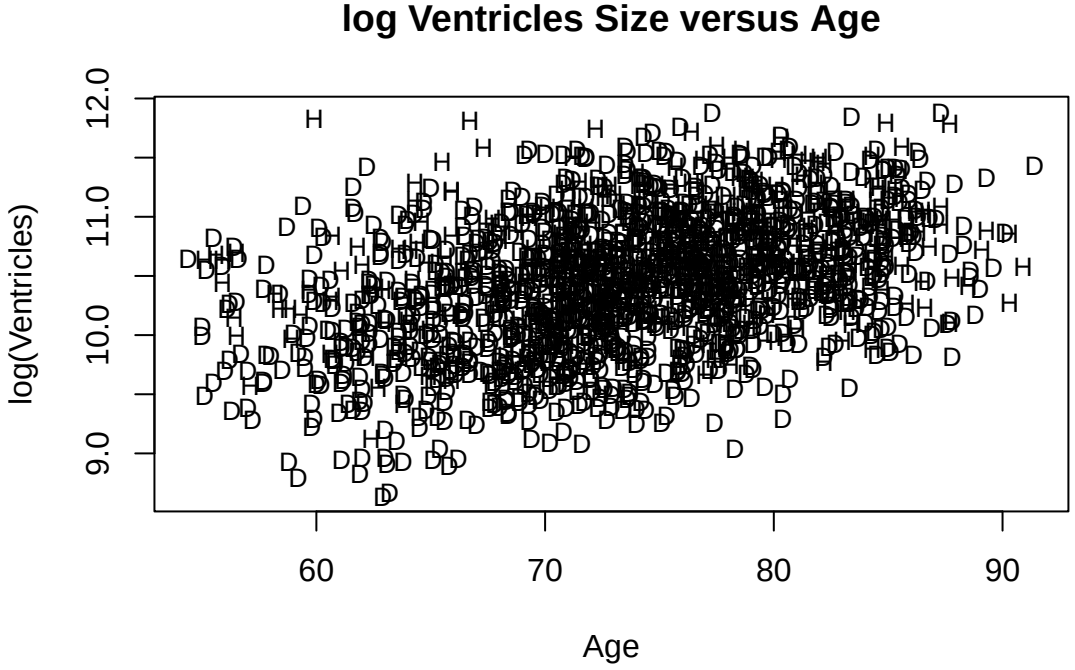
It is of interest to study the relationship between ventricles and age. In particular, we are interested in whether the relationship varies between healthy individuals and individuals with dementia related symptoms.

Read in and inspect the data:

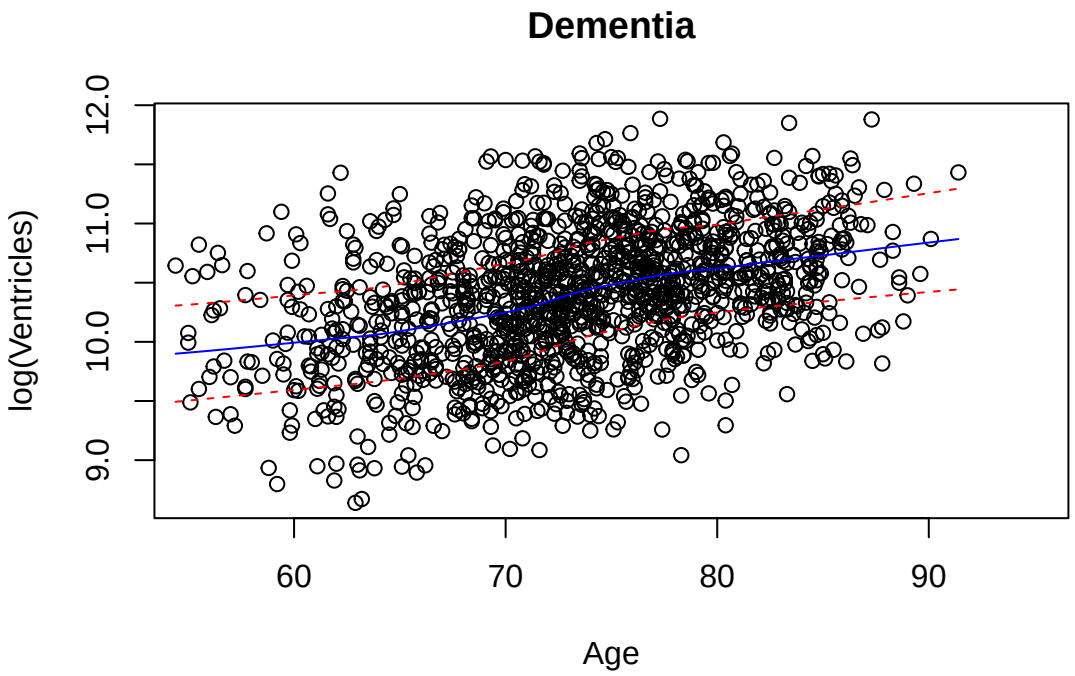
```
Ventricles.df=read.csv("Ventricles.csv")
plot(Ventricles~Age,main="Ventricles Size versus Age",type="n",data=Ventricles.df)
text(Ventricles.df$Age, Ventricles.df$Ventricles, Ventricles.df$AD, cex=.8)
```



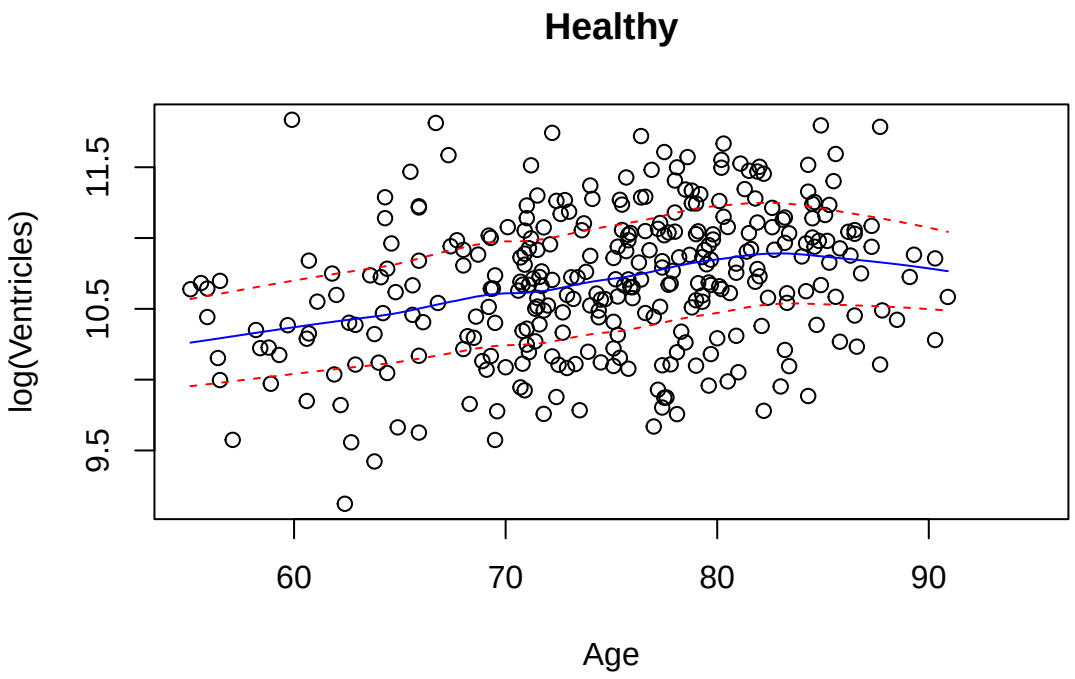

```
plot(log(Ventricles)~Age,main="log Ventricles Size versus Age",type="n",data=Ventricles.df)
text(Ventricles.df$Age, log(Ventricles.df$Ventricles), Ventricles.df$AD, cex=.8)
```



```
trendscatter(log(Ventricles)~Age,data=Ventricles.df[Ventricles.df$AD=="D",],xlim=c(55,95),main="Dementia
```




```
trendscatter(log(Ventricles)~Age,data=Ventricles.df[Ventricles.df$AD=="H",],xlim=c(55,95),main="Healthy
```



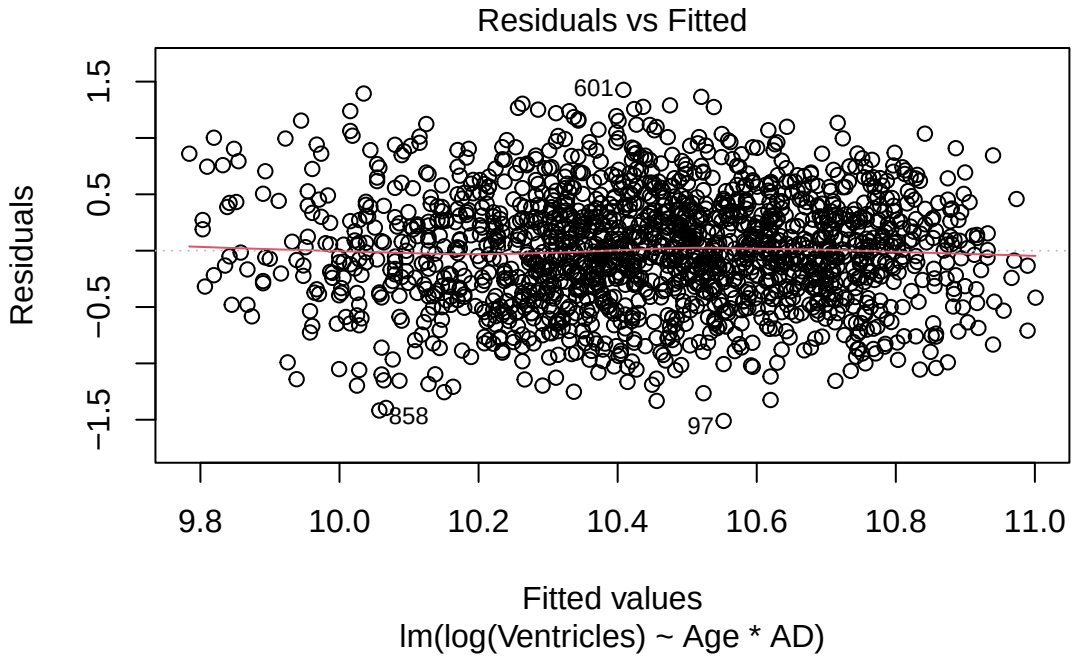
Comment

By observing the first graph, we found that most data points sink at the bottom of the graph, which does not fit the linear model, but the ventricle is also growing with age. In addition, it can be observed that the data points on the right are scattered, showing a right skew distribution, and the data with large median shows a dispersion trend. Perhaps the multiplication model can be used to fit the sample.

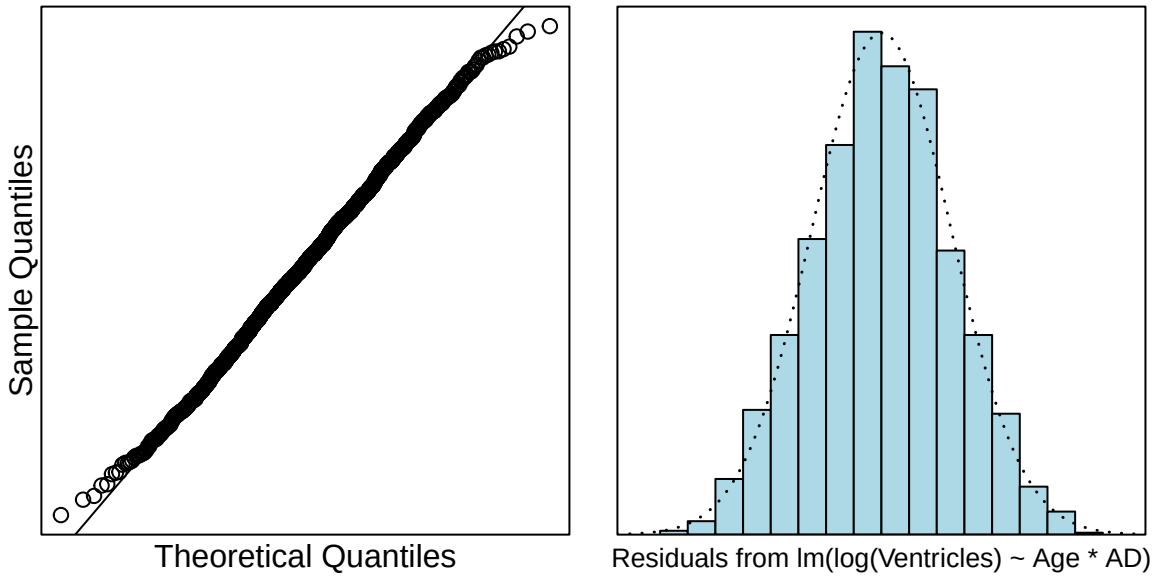
By observing the second graph, after taking the log function of Ventricles, there exists a clear linear positive relationship between Ventricles and Age.

By observing the third and fourth graphs, we find both Dementia and healthy has the positive linear relationship, and almost the points are in the confint intervals, though in the intervals between 80 and 90, the log(Ventricles) in healthy plot has a little drop trend. This makes us more confident to fit samples with multiplication model.

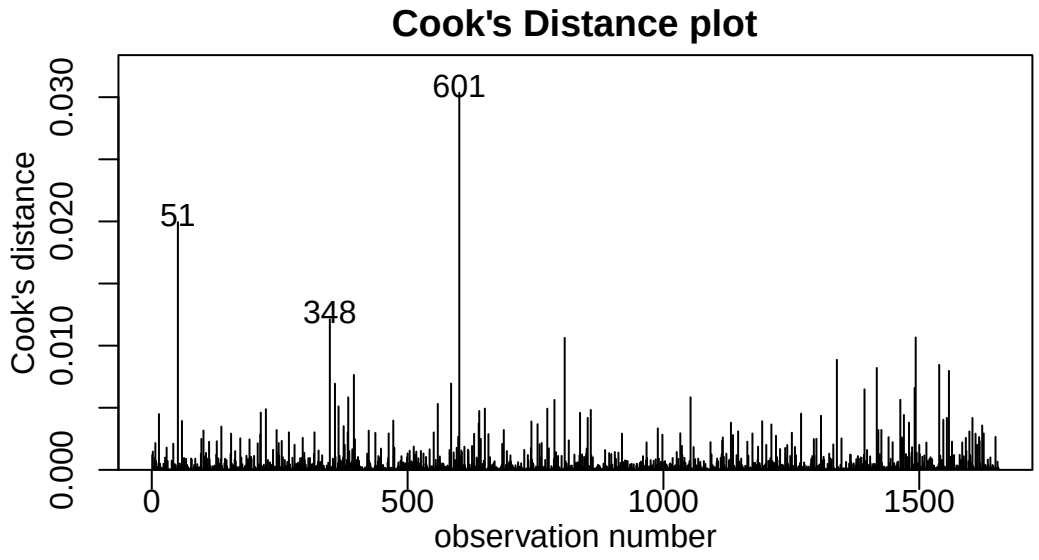
```
Ventriclesfit1=lm(log(Ventricles)~Age*AD,data=Ventricles.df)
plot(Ventriclesfit1,which=1)
```




```
normcheck(Ventriclesfit1)
```




```
cooks20x(Ventriclesfit1)
```



```
summary(Ventriclesfit1)
```

```
##
## Call:
## lm(formula = log(Ventricles) ~ Age * AD, data = Ventricles.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.51040 -0.34077  0.00086  0.33883  1.42693
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.034934   0.145792  55.112  < 2e-16 ***
## Age          0.032152   0.001977  16.262  < 2e-16 ***
## ADH          1.228317   0.310969   3.950 8.15e-05 ***
## Age:ADH     -0.013035   0.004152  -3.139  0.00172 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5053 on 1651 degrees of freedom
## Multiple R-squared:  0.1877, Adjusted R-squared:  0.1862
## F-statistic: 127.2 on 3 and 1651 DF,  p-value: < 2.2e-16
```

```
confint(Ventriclesfit1)
```

```
##                   2.5 %       97.5 %
## (Intercept)  7.74897653  8.320891845
## Age          0.02827412  0.036030020
## ADH          0.61838254  1.838252182
## Age:ADH     -0.02117928 -0.004891143
```

```
exp(confint(Ventriclesfit1))
```

```
##                    2.5 %       97.5 %
## (Intercept) 2319.1975659 4108.8228069
## Age            1.0286776    1.0366870
## ADH            1.8559237    6.2855427
## Age:ADH        0.9790434    0.9951208
```

```
(exp(confint(Ventriclesfit1))-1)*100
```

```
##                     2.5 %        97.5 %
## (Intercept) 231819.756593  4.107823e+05
## Age              2.867762  3.668697e+00
## ADH             85.592372  5.285543e+02
## Age:ADH         -2.095657 -4.879201e-01
```

```
# rotate factor
Ventricles.df=within(Ventricles.df,{ADflip=factor(AD,levels=c("H","D"))})
Ventriclesfit2=lm(log(Ventricles)~Age*ADflip,data=Ventricles.df)
summary(Ventriclesfit2)
```

```
##
## Call:
## lm(formula = log(Ventricles) ~ Age * ADflip, data = Ventricles.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.51040 -0.34077  0.00086  0.33883  1.42693
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.263252   0.274675  33.724  < 2e-16 ***
## Age          0.019117   0.003651   5.236 1.85e-07 ***
## ADflipD     -1.228317   0.310969  -3.950 8.15e-05 ***
## Age:ADflipD  0.013035   0.004152   3.139  0.00172 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5053 on 1651 degrees of freedom
## Multiple R-squared:  0.1877, Adjusted R-squared:  0.1862
## F-statistic: 127.2 on 3 and 1651 DF,  p-value: < 2.2e-16
```

```
confint(Ventriclesfit2)
```

```
##                    2.5 %      97.5 %
## (Intercept)  8.724504197  9.80199889
## Age          0.011955341  0.02627837
## ADflipD     -1.838252182 -0.61838254
## Age:ADflipD  0.004891143  0.02117928
```

```
exp(confint(Ventriclesfit2))
```

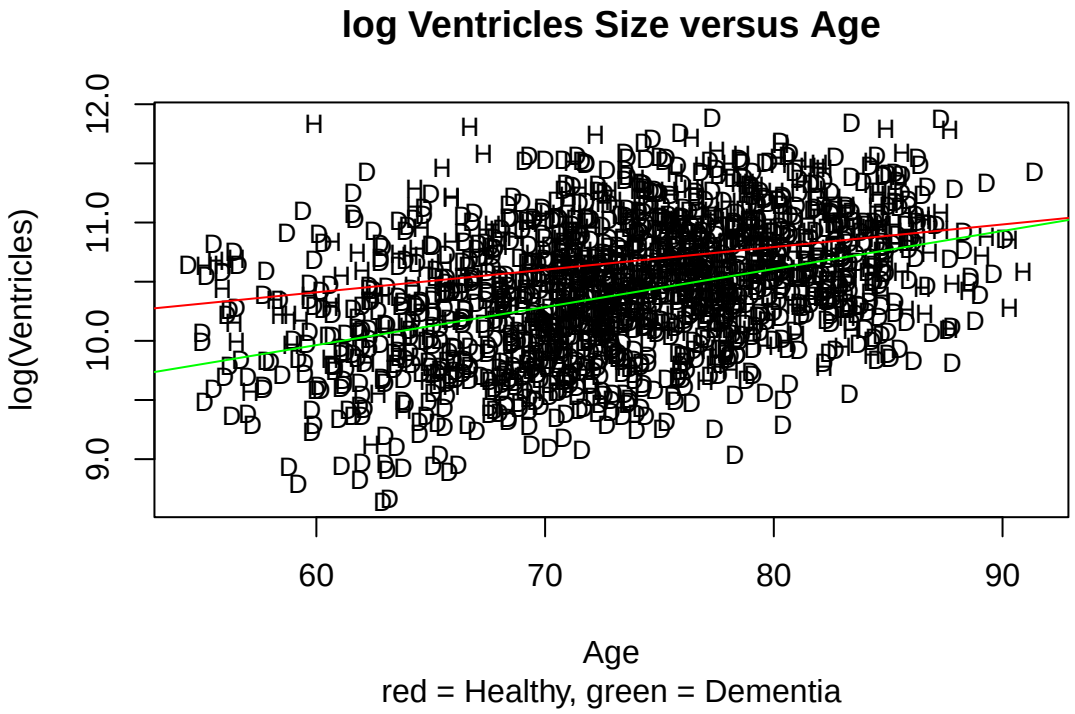
```
##                    2.5 %       97.5 %
## (Intercept) 6151.8258140 1.806983e+04
## Age            1.0120271 1.026627e+00
## ADflipD        0.1590953 5.388152e-01
## Age:ADflipD    1.0049031 1.021405e+00
```

```
(exp(confint(Ventriclesfit2))-1)*100
```

```
##                     2.5 %        97.5 %
## (Intercept)  6.150826e+05  1.806883e+06
## Age          1.202709e+00  2.662669e+00
## ADflipD     -8.409047e+01 -4.611848e+01
## Age:ADflipD  4.903124e-01  2.140515e+00
```

Plot the data with your appropriate model superimposed over it

```
plot(log(Ventricles)~Age,main="log Ventricles Size versus Age",sub="red = Healthy, green = Dementia",ty
text(Ventricles.df$Age, log(Ventricles.df$Ventricles), Ventricles.df$AD, cex=.8)
abline(Ventriclesfit1$coef[1],Ventriclesfit1$coef[2],col="green")
abline(Ventriclesfit1$coef[1]+Ventriclesfit1$coef[3],
       Ventriclesfit1$coef[2]+Ventriclesfit1$coef[4],col="red")
```




```
# or abline(Ventriclesfit2$coef[1],Ventriclesfit2$coef[2],col="red")
```

Methods and assumption checks

As the size of the ventricles increased the variabilty also increased so we logged the Ventricles data, this evened out the scatter. We have two explanatory variables, a grouping explanatory variable with two levels and a numeric explanatory variable, so have fitted a linear model with both variables and included an interaction term. The test for the interaction term proved to be significant, so the interaction term was kept and the model could not be simplified further.

Checking the assumptions there are no problems with assuming constant variability; looking at normality we see no issues and the Cook's plot doesn't reaveal any points of concern; as we have assumed the people were randomly sampled, independence is satisfied. The model assumptions are satisfied.

Our model is: $log(Ventricles_i) = \beta_0 + \beta_1 \times Age_i + \beta_2 \times ADH_i + \beta_3 \times Age_i \times ADH_i + \epsilon_i$ where $ADH_i = 1$ if the ith subject is healthy and 0 if they have signs of dementia, and $\epsilon_i \sim iid\ N(0, \sigma^2)$

Our model only explained 19% of the variability in the data.

Pretreatment

According to the model we fitted, we can disassemble the formula into two cases.

When $ADH_i = 0$,

$$Ventricles_i = e^{\beta_0+\beta_1\times Age_i+\epsilon_i} \times [ADH_i == 0]$$

When $ADH_i = 1$,

$$Ventricles_i = e^{\beta_0+\beta_2+(\beta_1+\beta_3)\times Age_i+\epsilon_i} \times [ADH_i == 1]$$

where$\epsilon_i \sim iid\ N(0, \sigma^2)$

In terms of slopes and/or intercepts, explain what the coefficient of Age:ADH is estimating.

Age:ADH is a interaction variable. It means that the healthy individuals compared to dementia individuals, when the age increase one, the median expected value of the size of Ventricles get additional times $e^{Age:ADH}$.

For each of the following, either write a sentence interpreting a confidence interval to estimate the requested information or state why we cannot answer this from the R-output given:

-in general, the difference in size of ventricles between healthy people and those exhibiting dementia symptoms.

We can deduce the formula, we assume that $Age^* = Age+1$, and we can deduce the variation of $\hat{Ventricles}^*$.

when $ADH_i = 0$, corresponding Dementia

$$\begin{aligned}\hat{Ventricles_1}^* &= e^{\hat{\beta_0}+\hat{\beta_1}\times Age^*} \\ &= e^{\hat{\beta_0}+\hat{\beta_1}\times Age+\hat{\beta_1}} \\ &= \hat{Ventricles_1} \times e^{\hat{\beta_1}}\end{aligned}$$

So when $ADH_i = 1$, corresponding Healthy

$$\begin{aligned}\hat{Ventricles_2}^* &= e^{\hat{\beta_0}+\hat{\beta_2}+(\hat{\beta_1}+\hat{\beta_3})\times Age^*} \\ &= e^{\hat{\beta_0}+\hat{\beta_2}+(\hat{\beta_1}+\hat{\beta_3})\times Age+(\hat{\beta_1}+\hat{\beta_3})} \\ &= \hat{Ventricles_2} \times e^{(\hat{\beta_1}+\hat{\beta_3})}\end{aligned}$$

So we can from the two formulas conclude that the difference in size of ventricles between healthy people and those exhibiting dementia symptoms is about when the age increase one, the median expected ventricular size of healthy individuals is $(1 + e^{\beta_3})\%$ of dementia.

And we can get this answer from `(exp(confint(Ventriclesfit1))-1)*100`. We can find that the Healthy individuals' expected median of the size of Ventricles could decrease by 0.49% to 2.10% than Dementia individuals'.

-the effect on the size of ventricles for each additional years aging on healthy people.

By observing the formula we deduced in the first question, and the R command `(exp(confint(Ventriclesfit1))-1)*100`, the Healthy individuals' the expected median of the size of Ventricles for each additional years could increased by about (2.87% - 2.10%, 3.67% - 0.49%), that is (0.77%, 3.18%).

-the effect on the size of ventricles for each additional years aging on people exhibiting dementia symptoms.

By observing the formula we deduced in the first question, and the R command `(exp(confint(Ventriclesfit1))-1)*100`, the Dementia individuals' the expected median of the size of Ventricles for each additional years could increased by about (2.87%, 3.67%).

Looking at the plot with the model superimposed, describe what seems to be happening.

For Healthy individuals, they almost have larger the size of Ventricles when they are young, and when they getting older, their the size of Ventricles will grow at a slower rate.

For dementia individuals, they almost have smaller the size of Ventricles when they are young, and when they getting older, their the size of Ventricles will grow at a faster rate.

And finally the Dementia and Healthy lines will coincide somewhere.