STATS 201 Assignment 2

Pang Bo 2019210176

Due Date: 2021/11/7

```
## Loading required package: s20x
```

Question 1

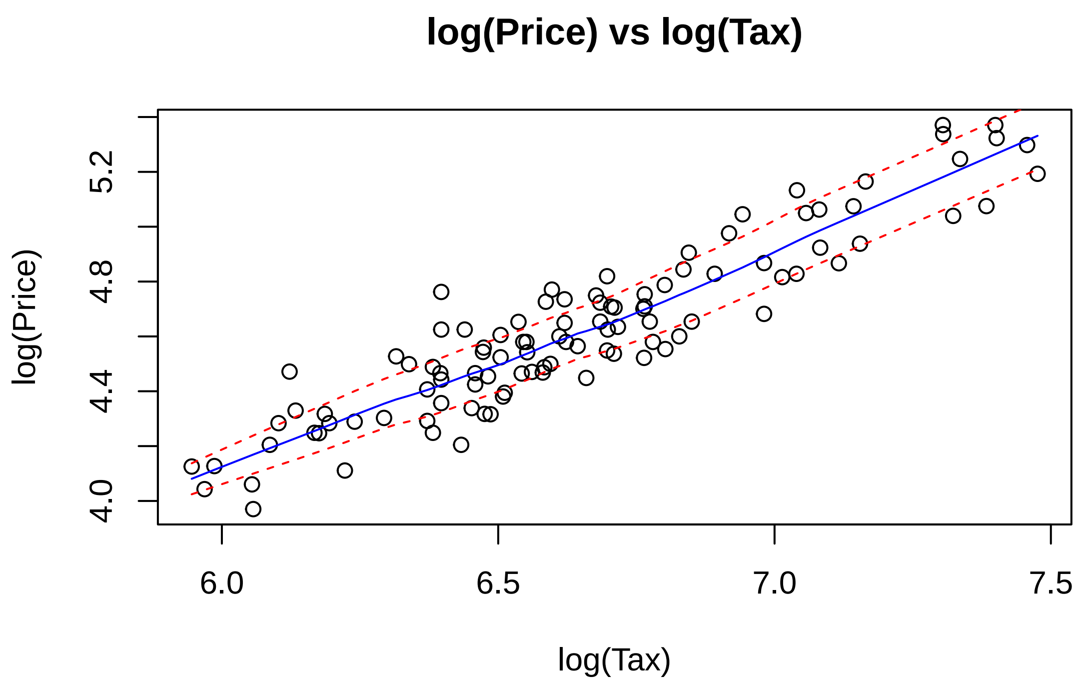
Question of interest/goal of the study

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)
```



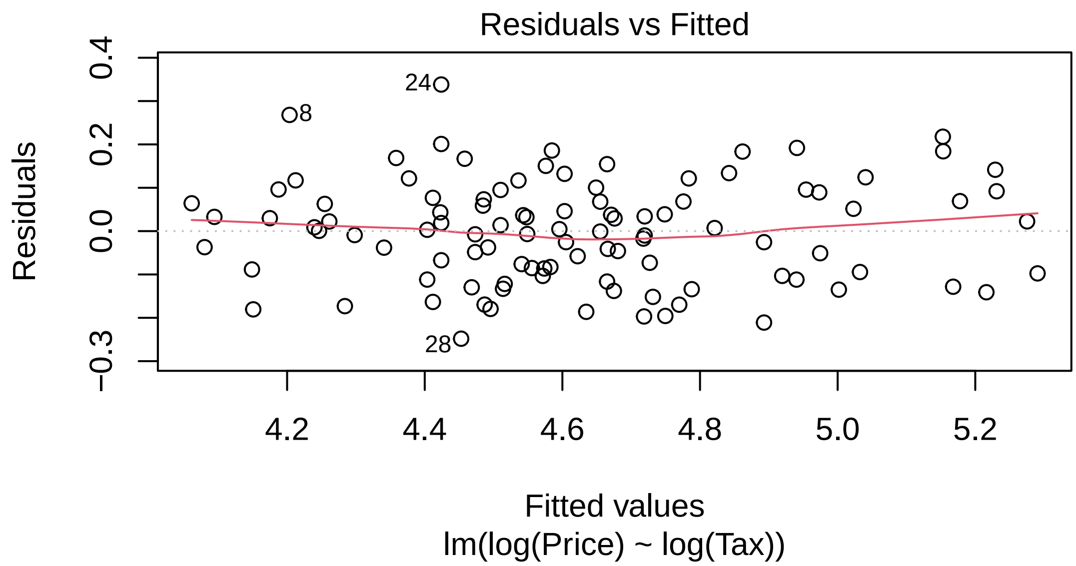
For initial plot, we can see the data is right-skewed. Because we find that some variables which have large explain variables have large response variables, it has characteristics of power law model.

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

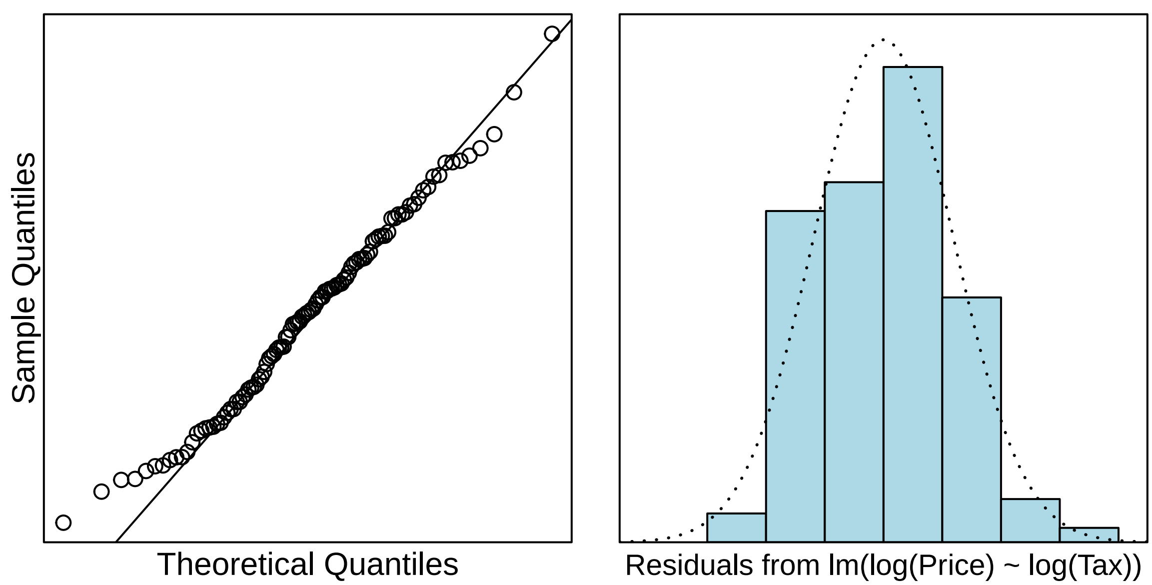
From the plot `Price vs Tax`, we could find that some points have a large explain variable value while they also have a large response variable value. It leads to the whole data set presents a right-skewed trend. It matches the characteristics of power law model. And when we add `log()` to each of the variables, the `trendscatter` plot seems to be perfect to fit using linear model. So we consider adopt power law model to fit this data set.

Fit model and check assumptions.

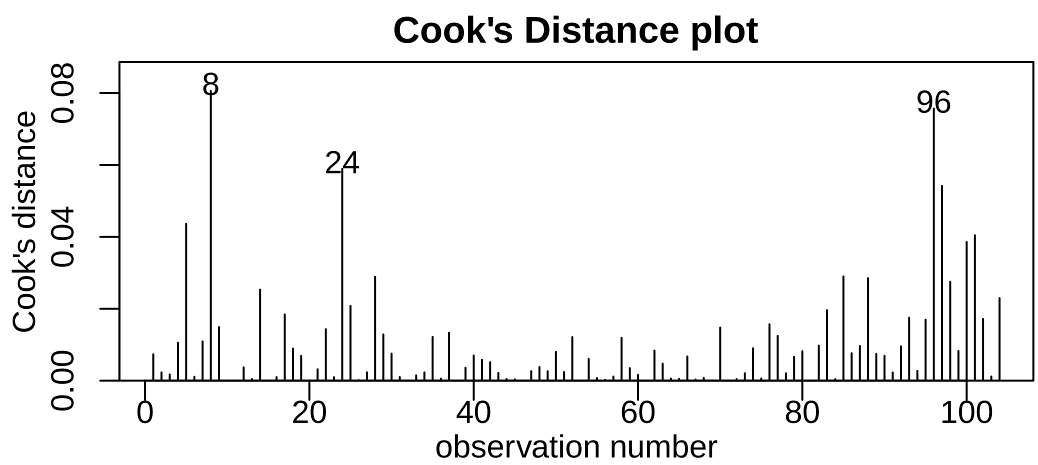
```
Price.lm = lm(log(Price) ~ log(Tax), data = hometax.df)

plot(Price.lm, which = 1)
```




```
normcheck(Price.lm)
```




```
cooks20x(Price.lm)
```




```
summary(Price.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
```

```
confint(Price.lm)
```

```
##                  2.5 %     97.5 %
## (Intercept) -1.1434829 -0.2834689
## log(Tax)     0.7385139  0.8677080
```

Methods and assumption checks

We used the power law model to find out the relationship between sale price of houses and their annual city tax bill because we found that the data is right-skewed distributed. It has clarified that data was collected from a random sample of 104 houses sold in Albuquerque. So there was no worries about the independence. The residual plot showed a patternless scatter with quite constant variability. The normcheck did not find serious problem. And there was no influential points. In conclusion, our model satisfies the problem.

Our model is:

$$log(Price_i) = \beta_0 + \beta_1 log(Tax_i) + \epsilon_i \text{ where } \epsilon \sim iid\ N(0, \sigma^2)$$

Our model explains 86% response variables, which means it is resonable for prediction.

Executive Summary

In order to specify the relation ship between sale price of houses and their annual city tax bill, we used a power law model to fit this data set. We have strong evidence that there exists some relationship between `Tax` and `Price`. And we estimate that a 1% increase in the `Tax` results in a 0.74% to a 0.87% increase in the median value of `Price`.

Our model explains 86% data in the data set. So it is reliable to predict using this model.

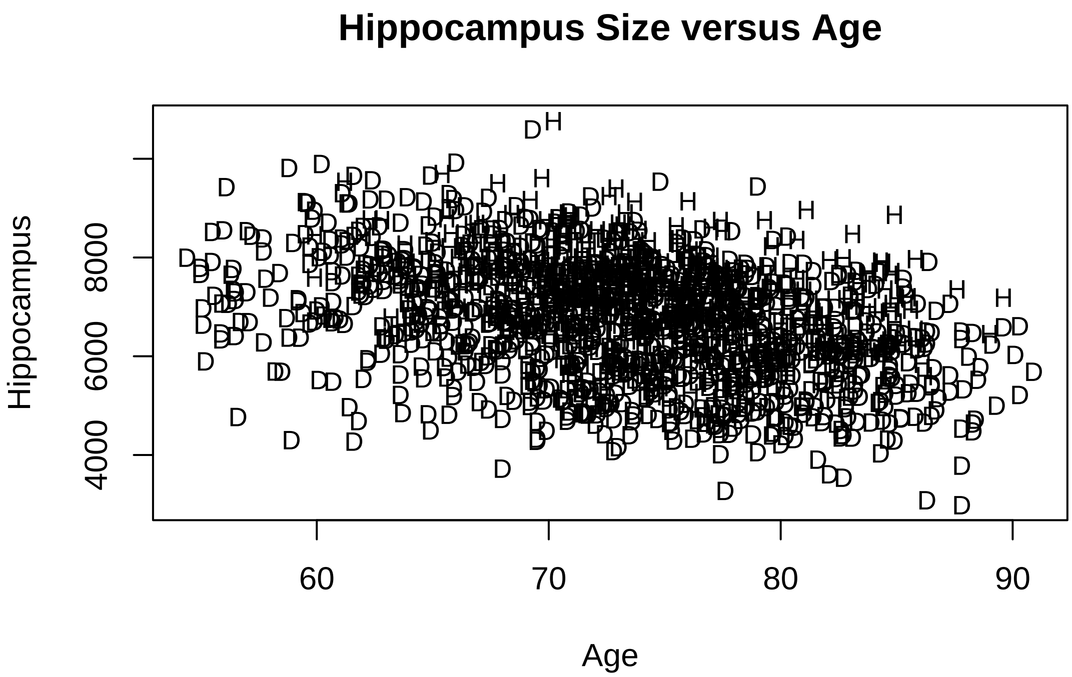
Question 2

Question of interest/goal of the study

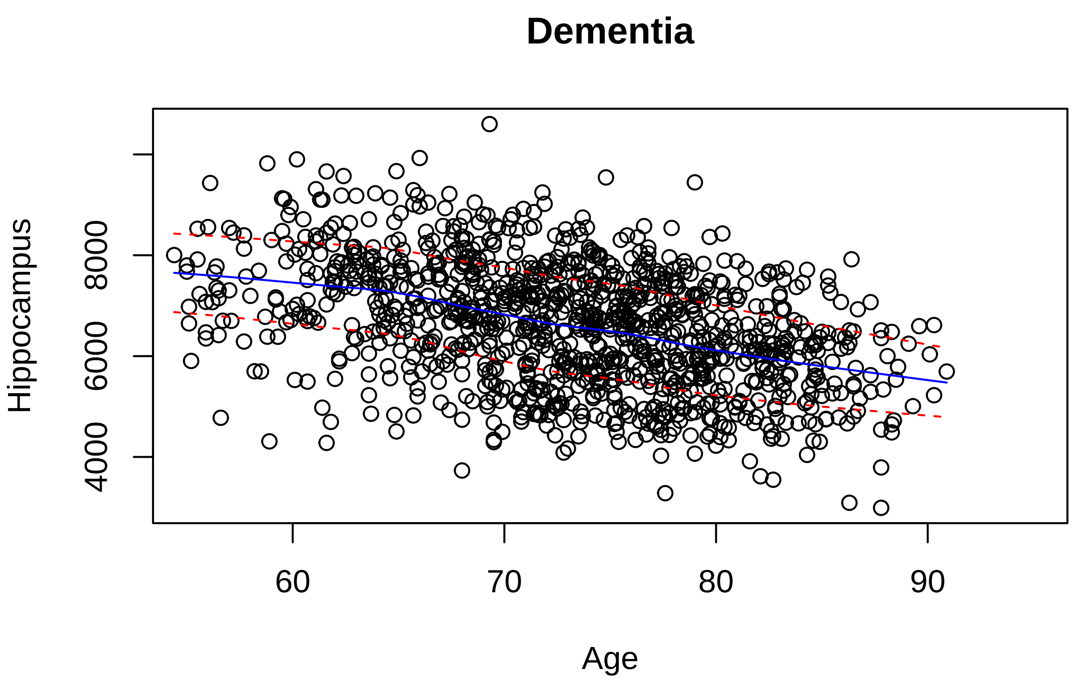
We want to explore the relationship between hippocampus size and age. In particular, we are interested in whether the relationship differs between healthy individuals and individuals with dementia related symptoms.

Read in and inspect the data:

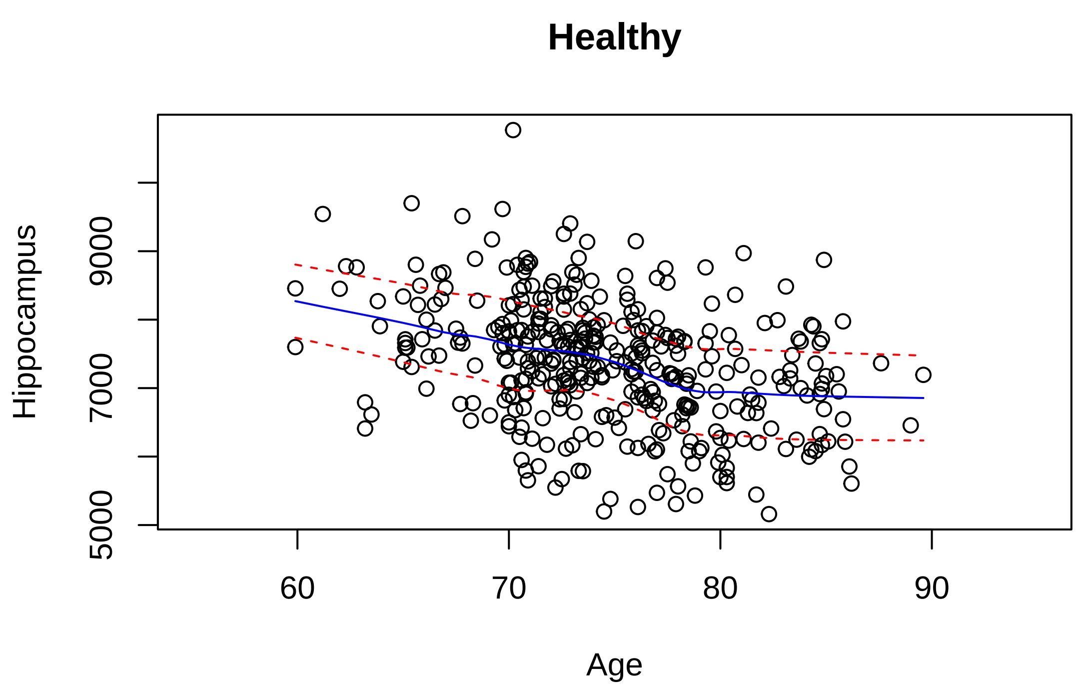
```
Hippocampus.df<-read.csv("Hippocampus.csv")
plot(Hippocampus~Age,main="Hippocampus Size versus Age",type="n",data=Hippocampus.df)
text(Hippocampus.df$Age, Hippocampus.df$Hippocampus, Hippocampus.df$AD, cex=.8)
```




```
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")
```



Generally speaking, the `Hippocampus` volume is decreasing with the increase of `Age` for both dementia and healthy volunteers. By comparing the two plots, we could find the average `Hippocampus` volume is quite different between these two groups. `Healthy` volunteers have larger `Hippocampus` volume than `Dementia` volunteers.

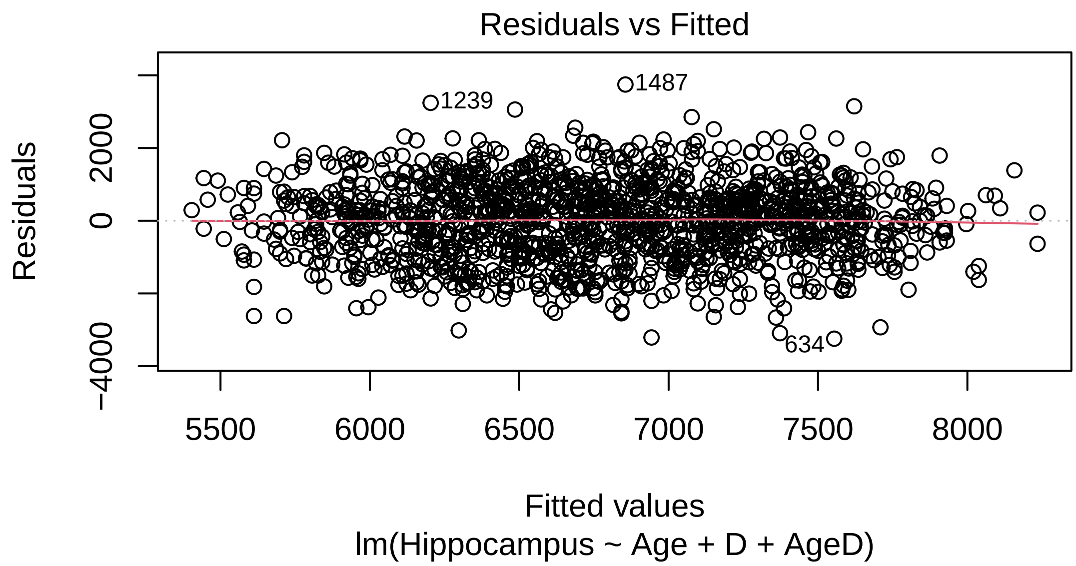
Fit model and check assumptions.

We define dummy variable `D` to represent healthy volunteers when `D=1` while baseline represents those volunteers who are dementia.

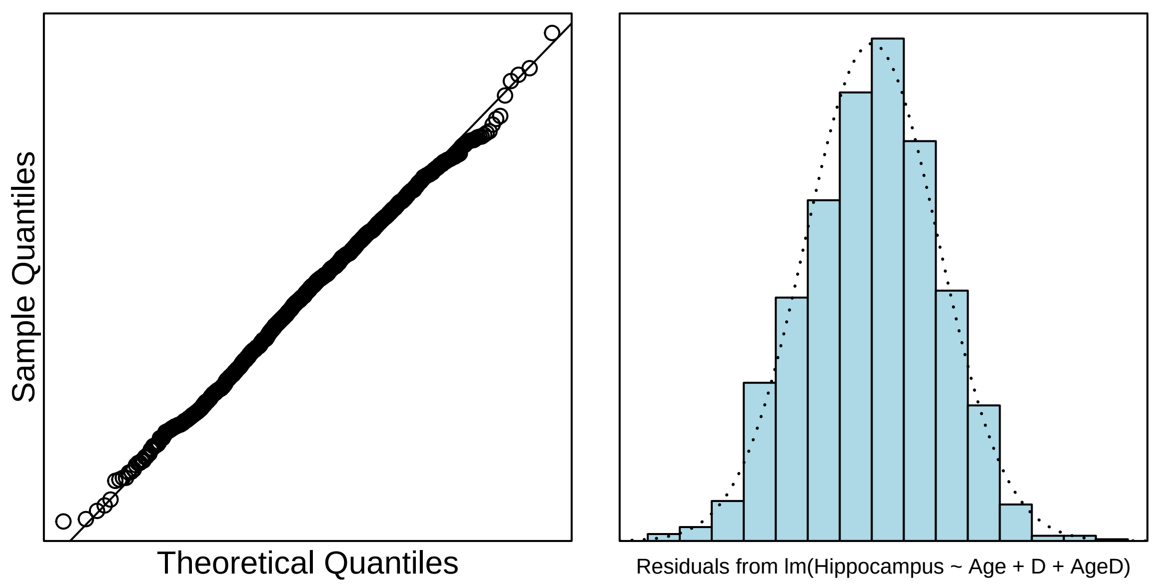
```
Hippocampus.df$D = as.numeric(Hippocampus.df$AD == "H")
table(Hippocampus.df$AD, Hippocampus.df$D)
```

```
##
##         0    1
##   D 1116    0
##   H    0  373
```

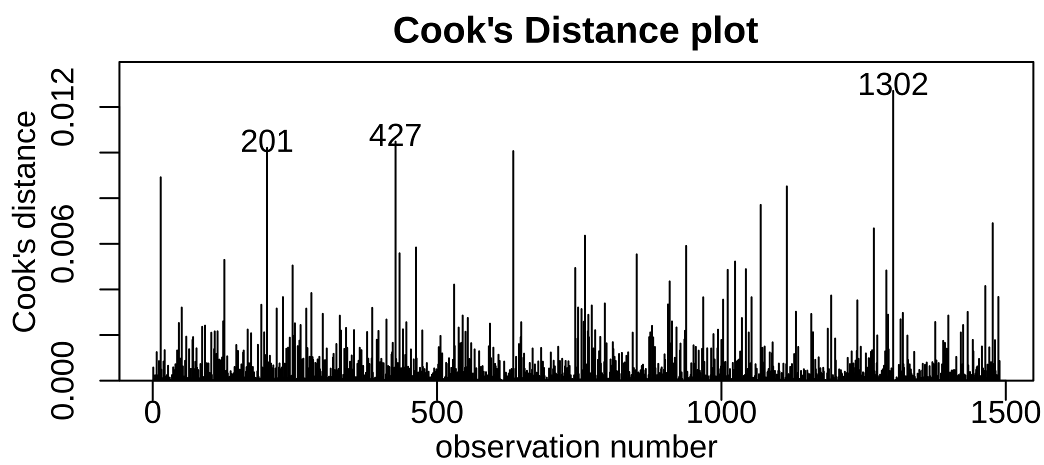
```
Hippocampus.df$AgeD = with(Hippocampus.df, {AgeD = D * Age})
AgeHippo.fit = lm(Hippocampus ~ Age + D + AgeD, data = Hippocampus.df)

plot(AgeHippo.fit, which = 1)
```




```
normcheck(AgeHippo.fit)
```




```
cooks20x(AgeHippo.fit)
```



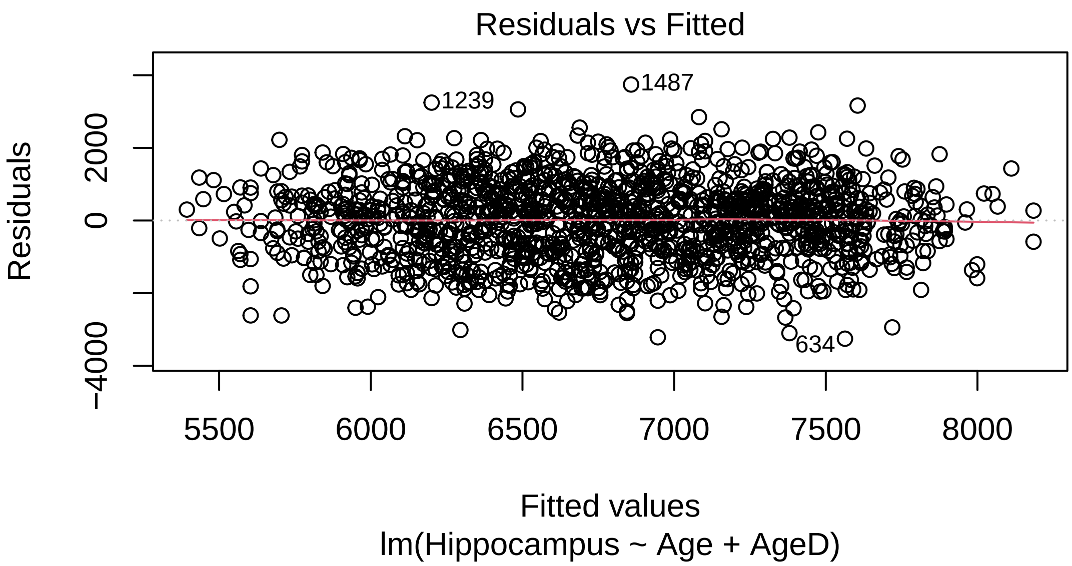

```
summary(AgeHippo.fit)
```

```
##
## 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) 11513.168    303.741  37.905   <2e-16 ***
## Age           -67.212      4.132 -16.266   <2e-16 ***
## 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
```

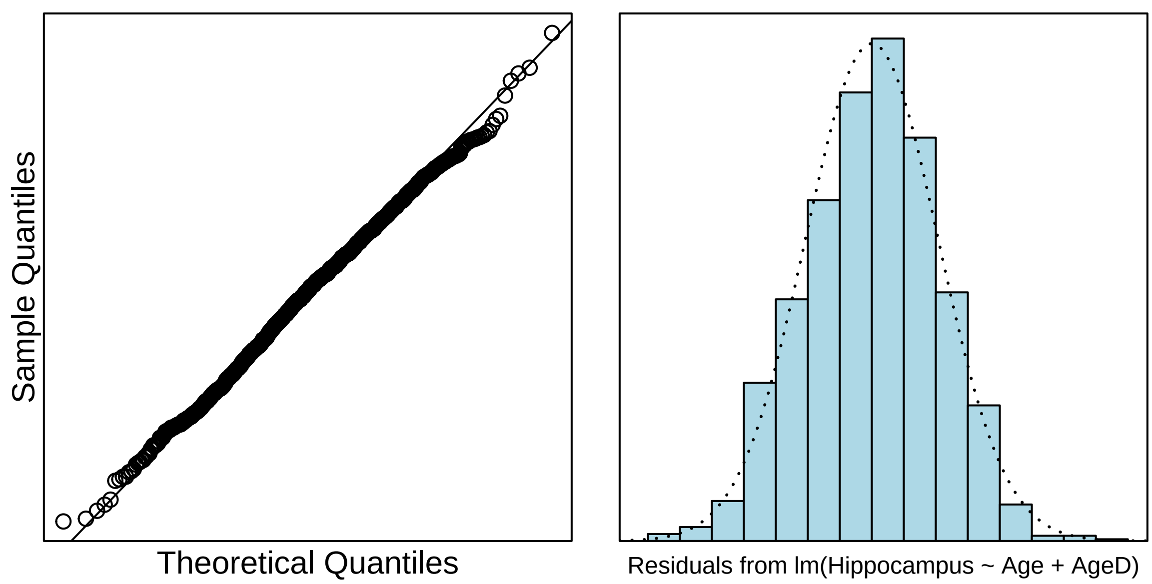
However, some coefficients seems may equal to 0. Let us drop some of the variables in the model.

Fit model and check assumptions AGAIN (DROP D).

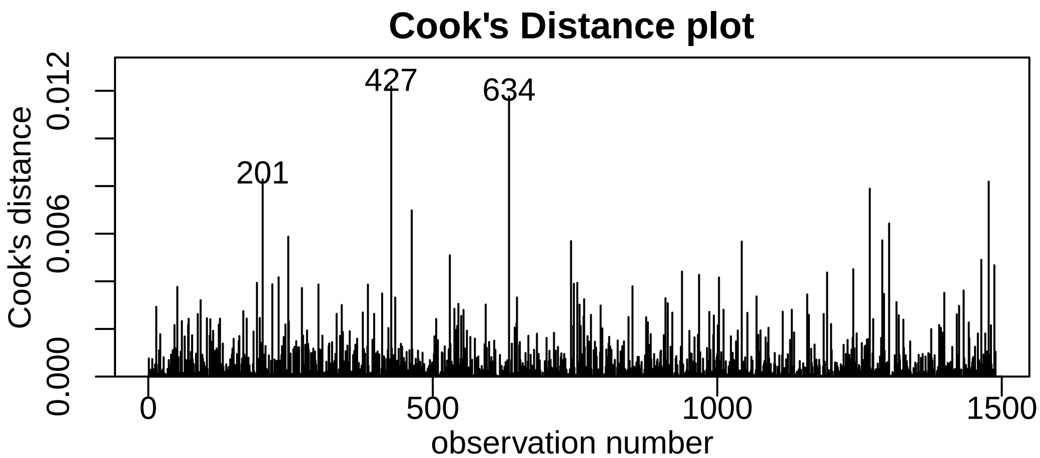
```
AgeHippo.fit2 = lm(Hippocampus ~ Age + AgeD, data = Hippocampus.df)

plot(AgeHippo.fit2, which = 1)
```




```
normcheck(AgeHippo.fit2)
```




```
cooks20x(AgeHippo.fit2)
```




```
summary(AgeHippo.fit2)
```

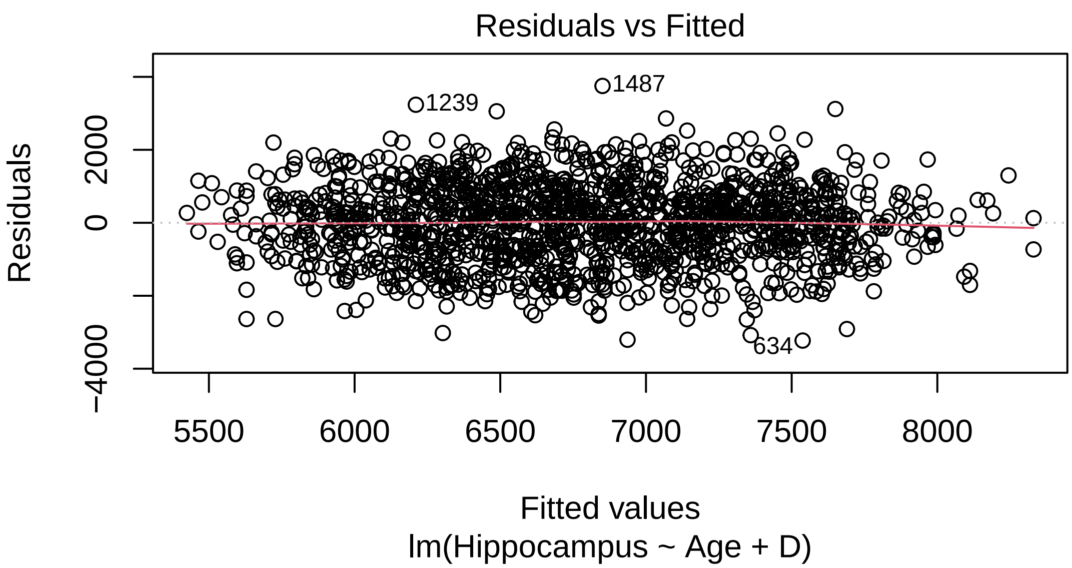
```
##
## Call:
## lm(formula = Hippocampus ~ Age + AgeD, data = Hippocampus.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3254.2  -733.8    55.0   709.7  3743.9
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11556.5541   280.1440   41.25   <2e-16 ***
## Age           -67.7991     3.8146  -17.77   <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(AgeHippo.fit2)
```

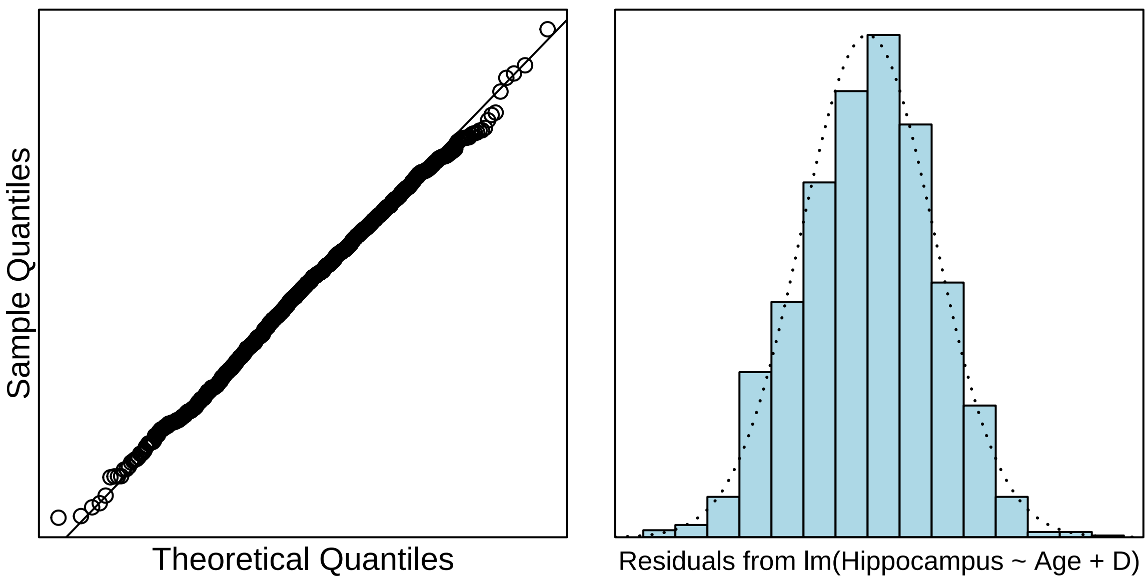
```
##                    2.5 %      97.5 %
## (Intercept) 11007.034397 12106.07381
## Age           -75.281680   -60.31645
## AgeD            9.869105    13.14855
```

Fit model and check assumptions AGAIN AND AGAIN (DROP AGED).

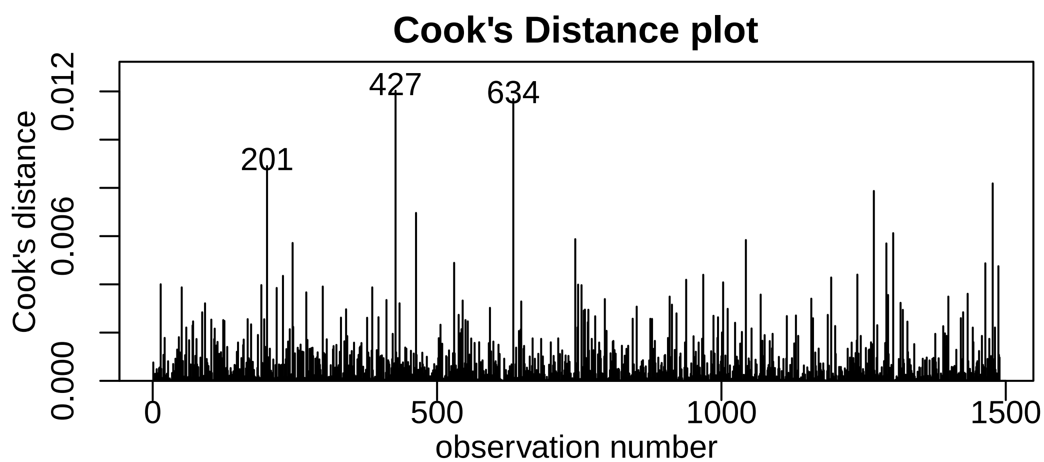
```
AgeHippo.fit3 = lm(Hippocampus ~ Age + D, data = Hippocampus.df)

plot(AgeHippo.fit3, which = 1)
```




```
normcheck(AgeHippo.fit3)
```



```
cooks20x(AgeHippo.fit3)
```

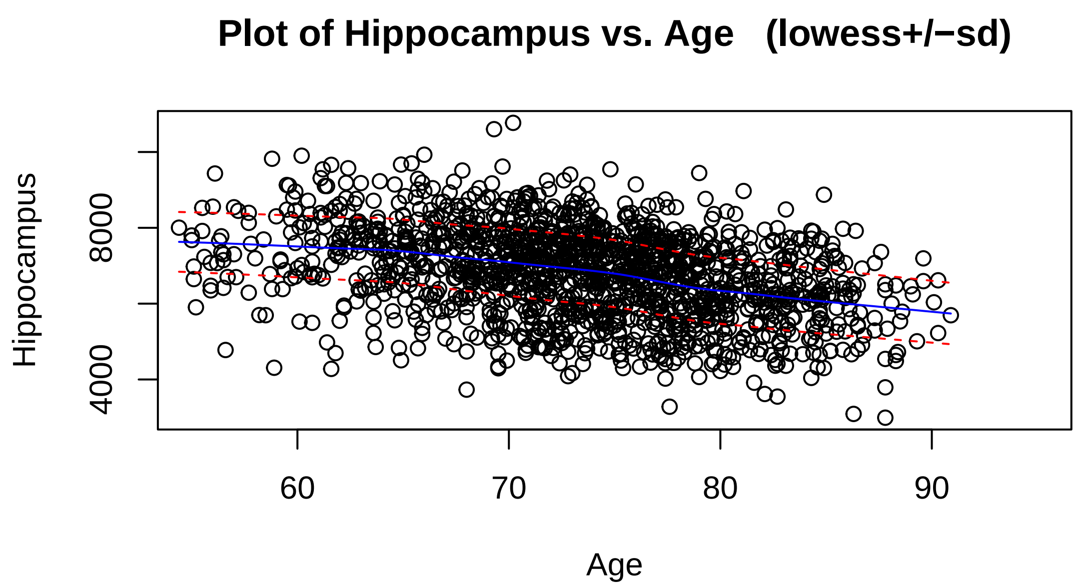



```
summary(AgeHippo.fit3)
```

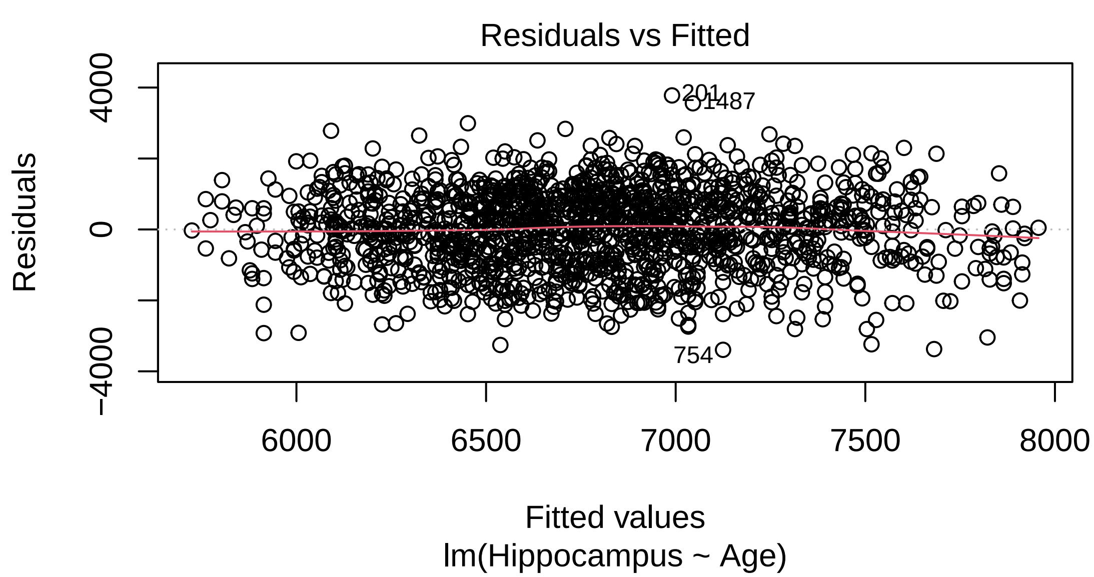
```
##
## Call:
## lm(formula = Hippocampus ~ Age + D, data = Hippocampus.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3228.8  -727.2    54.5   705.0  3751.1
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11427.664    279.674   40.86   <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
```

Fit model and check assumptions AGAIN AND AGAIN AND AGAIN (DROP D AND AGED).

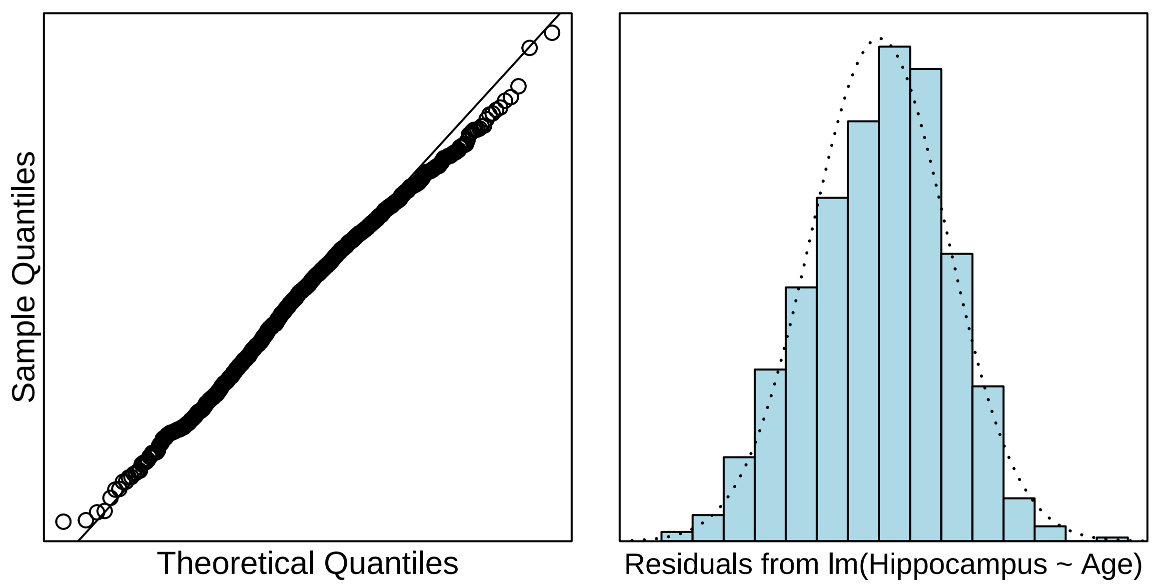
```
trendscatter(Hippocampus~Age,data=Hippocampus.df,xlim=c(55,95))
```



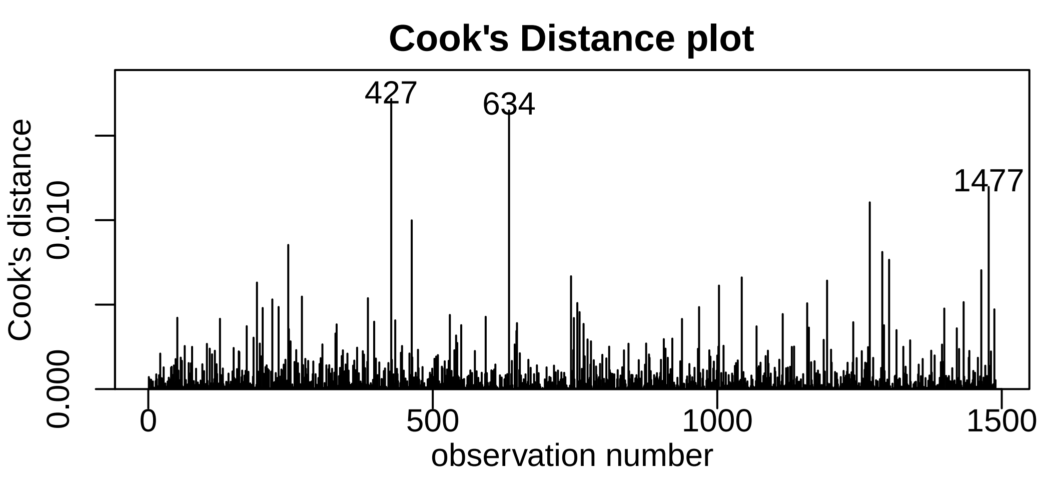

```
AgeHippo.fit4 = lm(Hippocampus ~ Age, data = Hippocampus.df)

plot(AgeHippo.fit4, which = 1)
```



```r
normcheck(AgeHippo.fit4)
```




```r
cooks20x(AgeHippo.fit4)
```




```r
summary(AgeHippo.fit4)
```

```
##
## 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
```

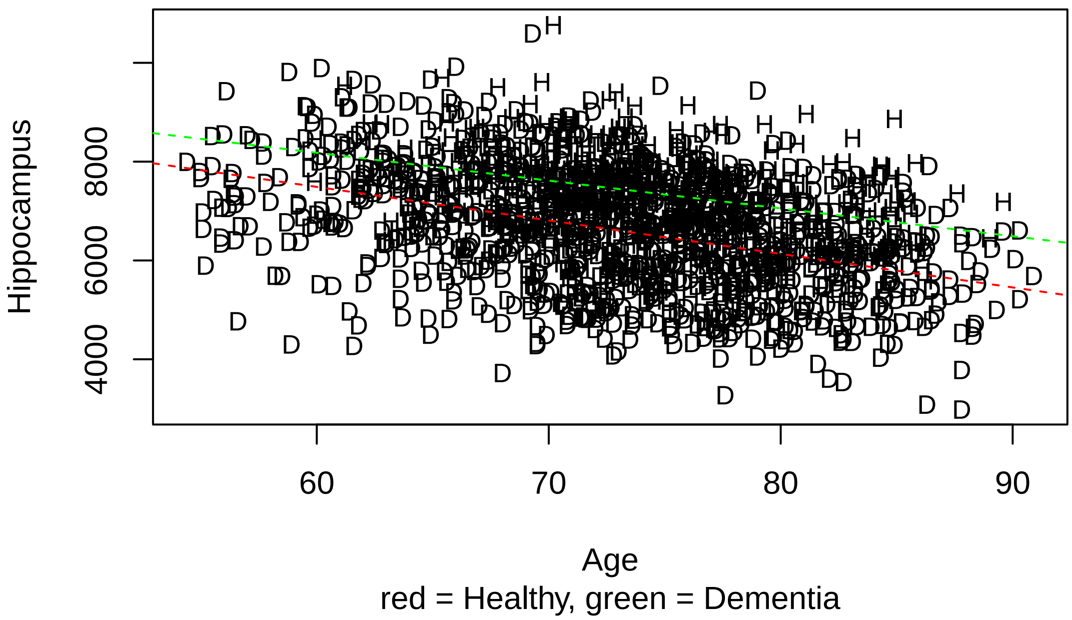
Plot the data with your appropriate model superimposed over it

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

b = coef(AgeHippo.fit2)
b
```

```
## (Intercept)         Age        AgeD
## 11556.55410   -67.79906    11.50883
```

```
abline(b[1:2], lty = 2, col = "red")
abline(b[1], b[2] + b[3], lty = 2, col = "green")
```

Hippocampus Size versus Age



We finally choose `AgeHippo.fit2` as our fit model.

Methods and assumption checks

In order to study the relationship between hippocampus size, age and whether the relationship differs between healthy individuals and individuals with dementia related symptoms, we used linear model with one dummy variable to fit the data set. We find that `Hippocampus` volume is decreasing as the increase of `Age`. The average volumes of `Healthy` volunteers' `Hippocampus` are much larger than those who `Dementia`. And `Dementia` volunteers' decreasing speed is faster than `Healthy` volunteers. The main problem we encountered is when we used `Hippocampus ~ Age + D + AgeD` to fit the data set, we found some coefficients with dummy variable `D` could equal to `0`. Then we are trying to remove some variable in our model to make all the coefficients have evidence existing. After removing `D`, `AgeD` and both `D` and `AgeD`, we found the plan which removes `D` variable can get the highest `Multiple R-squared` value. However, compared to not removing variable `D`, the `Multiple R-squared` value decreased about 0.01%, this does not affect the correctness of our model because dropping one useless variable makes our model more general. So we choose this linear model to fit. As the description of the question has stated that we can treat the data as if it came from random samples of subjects. So we do not have worries about independence. The residual plot showed a patternless scatter with quite constant variability. The normcheck did not find serious problem. And there was no influential points. In conclusion, our model can explain the data set.

Our model is:

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

where $D_i = 1$ 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 explains 23.27% data in the initial data set.

Executive Summary

We are going to study the relationship between hippocampus size, age and whether the relationship differs between healthy individuals and individuals with dementia related symptoms. We use linear model with numerical explain variable and categorical variable and we have extremely strong evidence that `Hippocampus` volume is decreasing as the increase of `Age`, the average volumes of `Healthy` volunteers' `Hippocampus` are much larger than those who `Dementia`, and `Dementia` volunteers' decreasing speed is faster than `Healthy` volunteers.

We estimate that for `Dementia` volunteers, their `Hippocampus` volumes decrease about $60 \sim 75$ units per year on average. For `Healthy` volunteers, their `Hippocampus` volumes decreasing speed could decrease $10 \sim 13$ units compared with `Dementia` volunteers per year on average.

Our model only explains about 23.27% variability in the initial data set, which means this model may not be very good for prediction. We estimate the reason may be the too scattered data points.

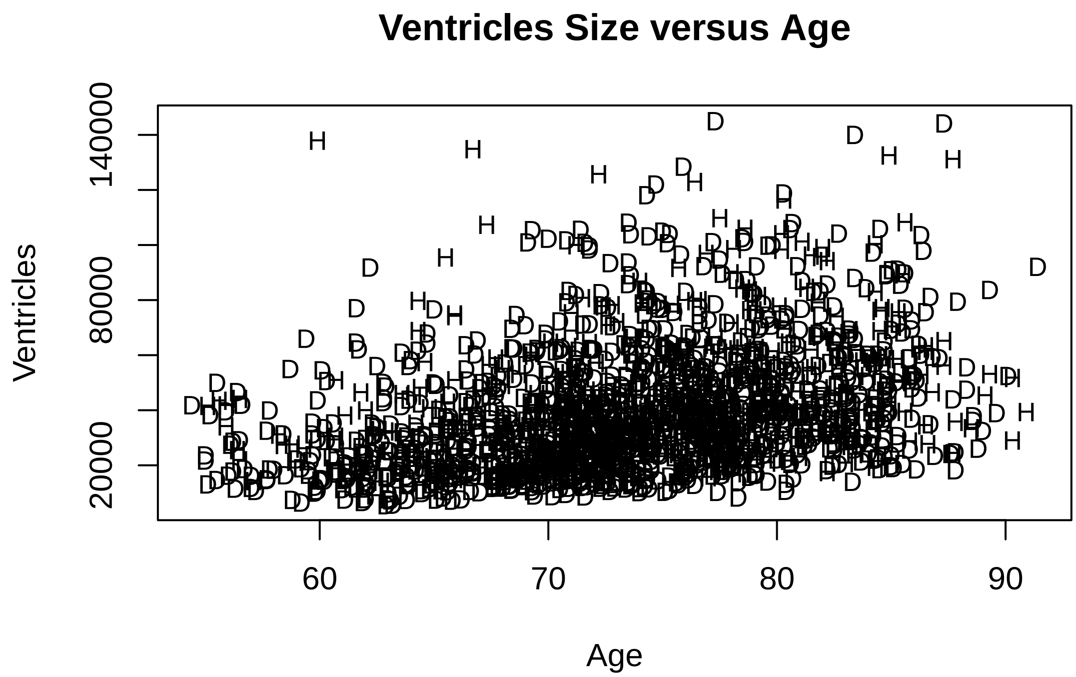
Question 3

Question of interest/goal of the study

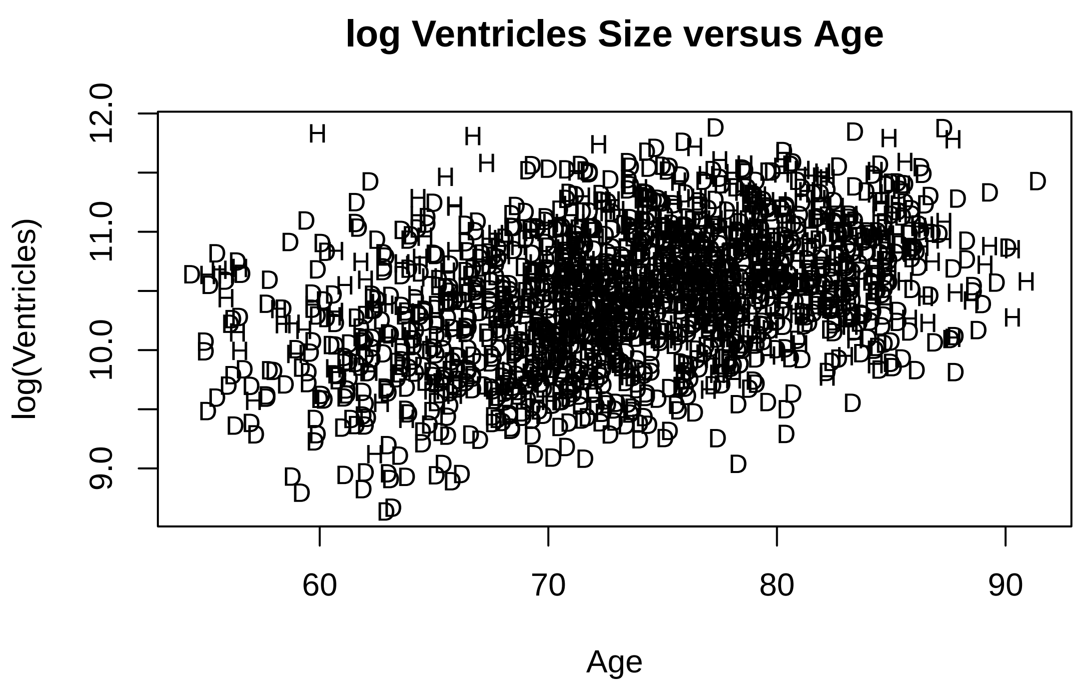
It is of interest to study the relationship between ventricles and age. In particular, we are interested in whether the relationship varies between healthy individuals and individuals with dementia related symptoms.

Read in and inspect the data:

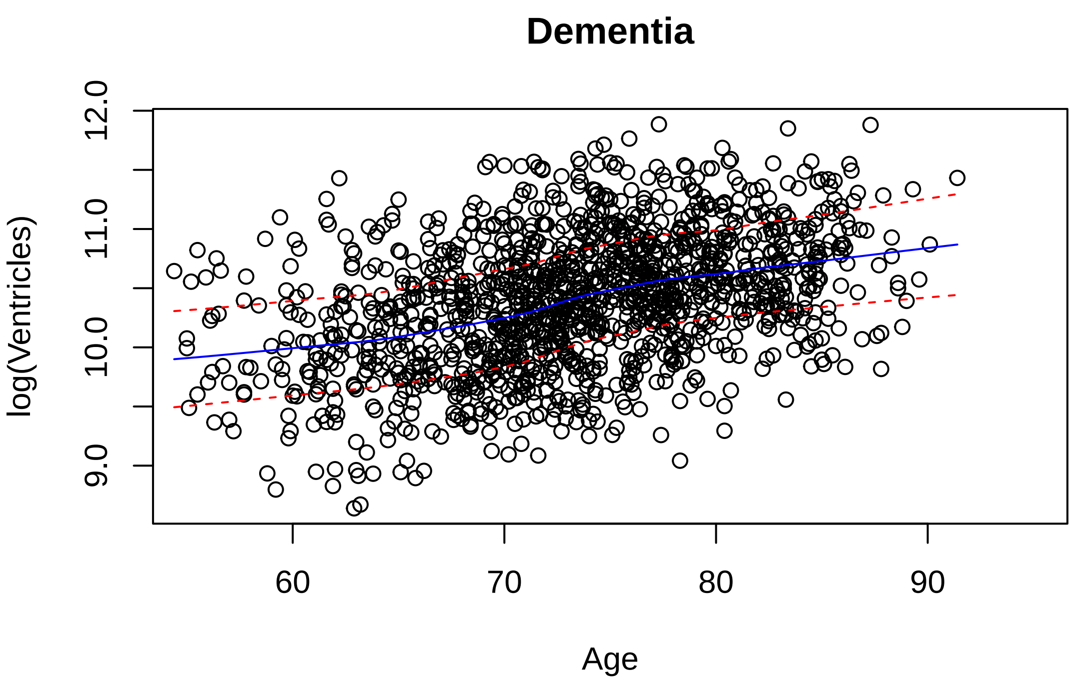
```
Ventricles.df=read.csv("Ventricles.csv")
plot(Ventricles~Age,main="Ventricles Size versus Age",type="n",data=Ventricles.df)
text(Ventricles.df$Age, Ventricles.df$Ventricles, Ventricles.df$AD, cex=.8)
```



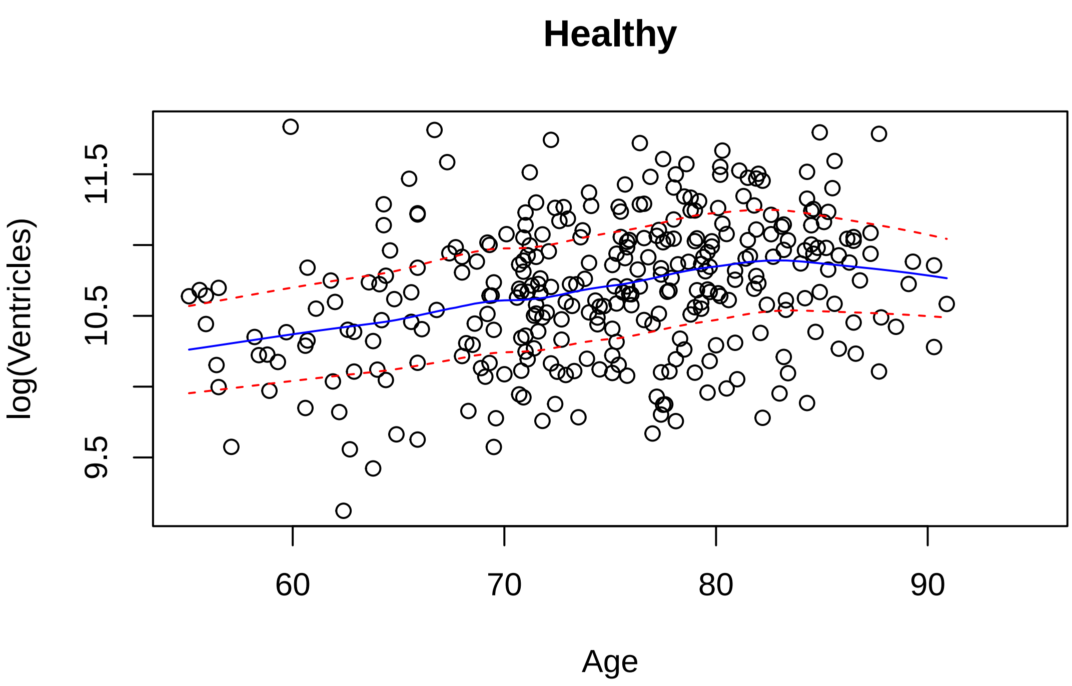

```
plot(log(Ventricles)~Age,main="log Ventricles Size versus Age",type="n",data=Ventricles.df)
text(Ventricles.df$Age, log(Ventricles.df$Ventricles), Ventricles.df$AD, cex=.8)
```



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



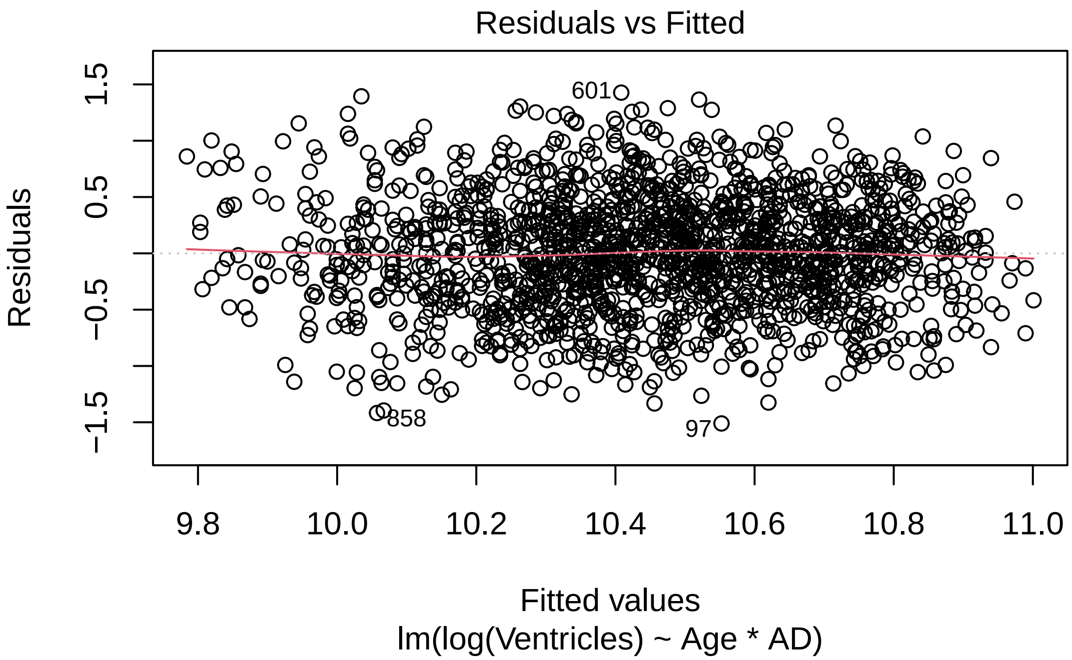

```
trendscatter(log(Ventricles)~Age,data=Ventricles.df[Ventricles.df$AD=="H",],xlim=c(55,95),main="Healthy
```



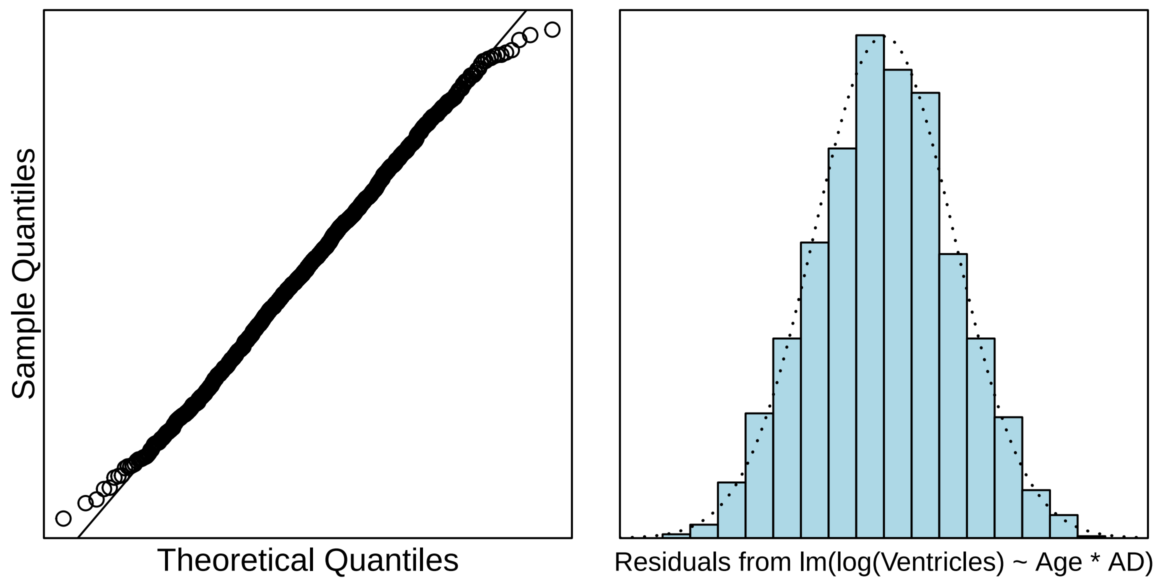
In the

`trendscatter` plots above, we could find that the initial data presents a trend that right-skewed. So we can use multiplicative model to fit this model. After transforming response variable into its logarithm, the data presents a linear relationship. However, in `Healthy` group, when age larger than 83, the trend seems to be different. It could be ignored because we have too little data in the interval.

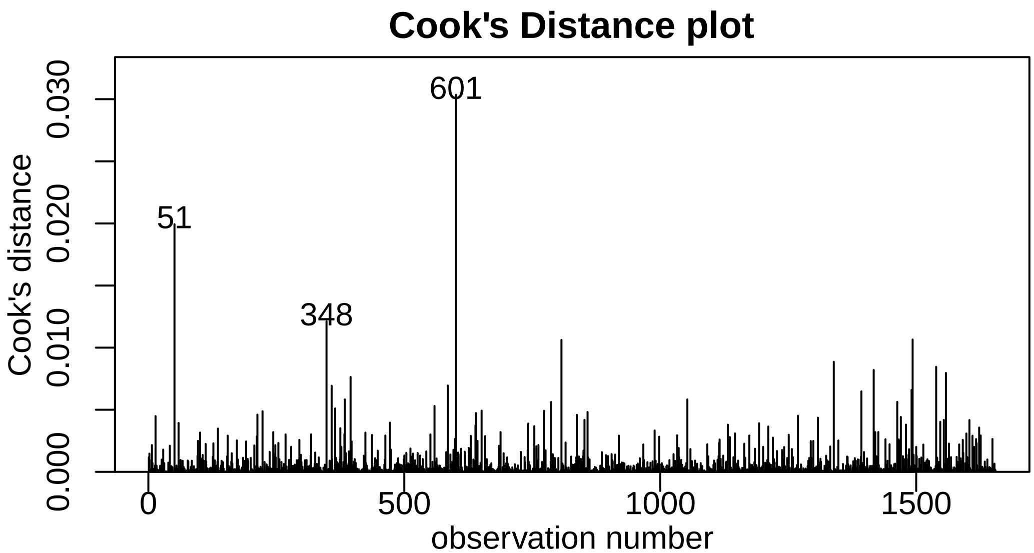
```
Ventriclesfit1=lm(log(Ventricles)~Age*AD,data=Ventricles.df)
plot(Ventriclesfit1,which=1)
```




```
normcheck(Ventriclesfit1)
```




```
cooks20x(Ventriclesfit1)
```



```r
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
```

```r
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
```

```r
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
```

```r
(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
```

```r
# 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))
```

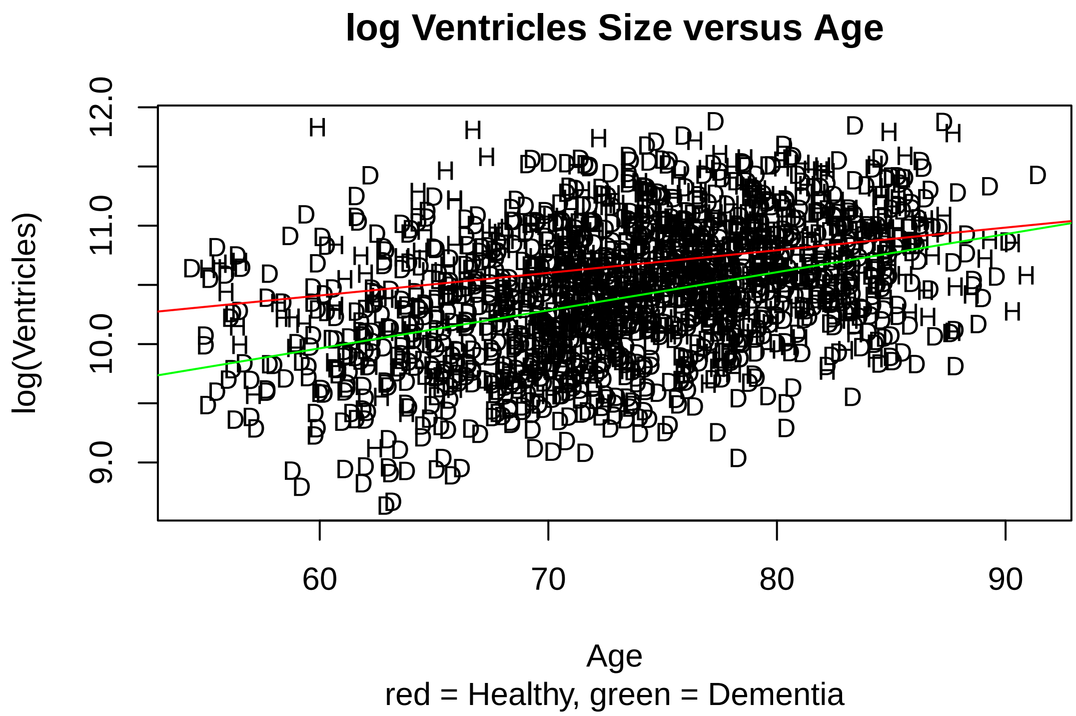
```
##                    2.5 %       97.5 %
## (Intercept) 6151.8258140 1.806983e+04
## Age            1.0120271 1.026627e+00
## ADflipD        0.1590953 5.388152e-01
## Age:ADflipD    1.0049031 1.021405e+00
```

```
(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",typ
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.

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

For slopes, `Age:ADH` means `Healthy` volunteers could get additional times of $e^{Age:ADH}$ median increasement every year than `Dementia` volunteers.

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 see in the output of `(exp(confint(Ventriclesfit1))-1)*100`, `ADH` is estimated between 85.59 and 528.55, which means the `Healthy` volunteers' median of `Ventricles` volumes could increased by 85.59% to 528.55% than `Dementia` volunteers.

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

We can see in the output of `(exp(confint(Ventriclesfit1))-1)*100`, `Age` is estimated between 2.87 and 3.67, which means the `Dementia` volunteers' median of `Ventricles` volumes for each additional years could increased by 2.87% to 3.67%. And `Age:ADH` is estimated between -2.10 and -0.49, which means the `Healthy` volunteers' median of `Ventricles` volumes for each additional years could decreased by 2.10% to 0.49% than the estimated median of `Ventricles` volumes of `Dementia` volunteers.

So, the effect on the size of ventricles for each additional years aging on healthy people could be estimated by $2.87 - 2.10 \sim 3.67 - 0.49$. That is, the `Healthy` volunteers' median of `Ventricles` volumes for each additional years could increased by about 0.7% to 3.2%.

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

We can see in the output of `(exp(confint(Ventriclesfit1))-1)*100`, `Age` is estimated between 2.87 and 3.67, which means the `Healthy` volunteers' median of `Ventricles` volumes could increased by 2.87% to 3.67% every year.

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

For `Healthy` volunteers, they usually have larger `Ventricles` volumes when they are young. However, when they getting old, they have a low speed increment of `Ventricles` volumes. For `Dementia` volunteers, they usually have less `Ventricles` volumes when they are young. However, when they getting old, they have a high speed increment of `Ventricles` volumes.