STATS 201 Assignment 2

Model Answers

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

Question 1

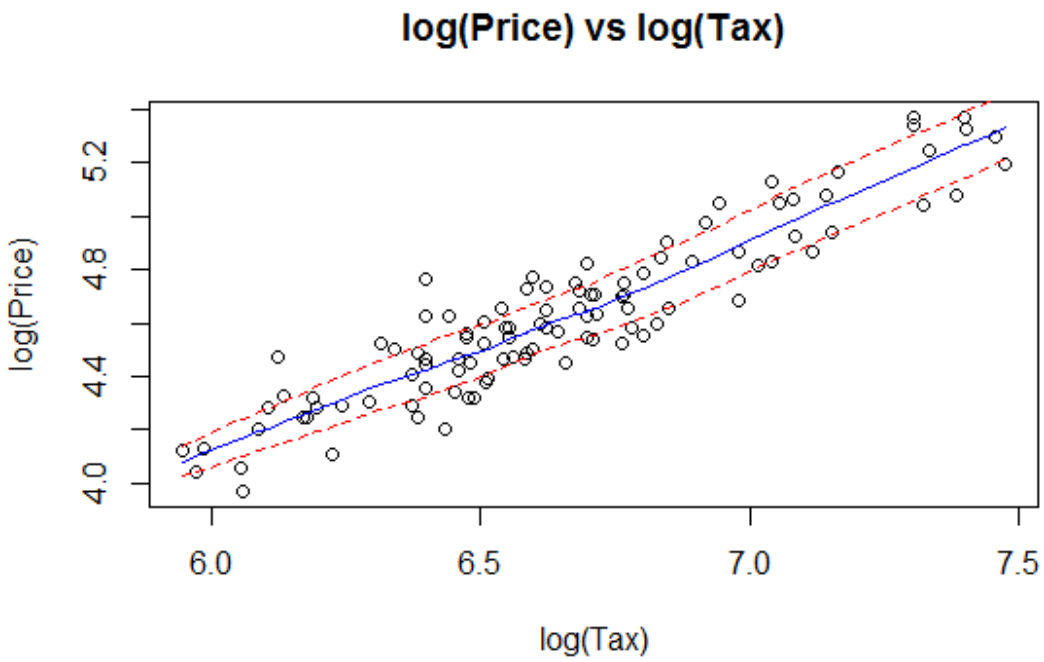
Question of interest/goal of the study

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)
```



There is a roughly linear increasing relationship between tax and price. Hovever, we can see that as tax increases the amount of variability in price also increases. Also, both tax and Price are positively (right) skewed, with most of the values being low and relatively few larger values. The plot of log(Price) versus log(Tax) shows an increasing linear relationship with roughly constant scatter.

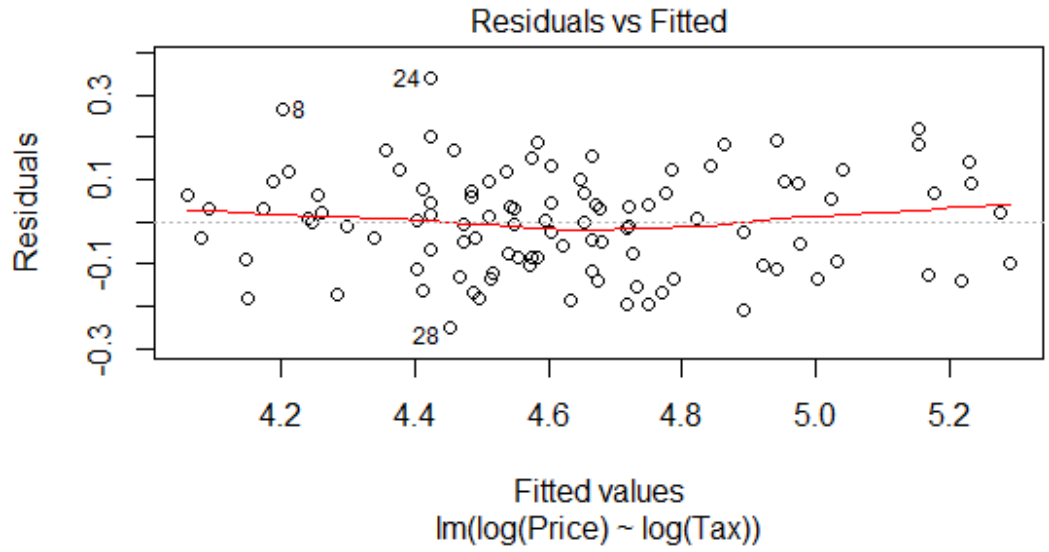
Justify why a log-log (power) model is appropriate here.

A log-log model can be justified several ways:

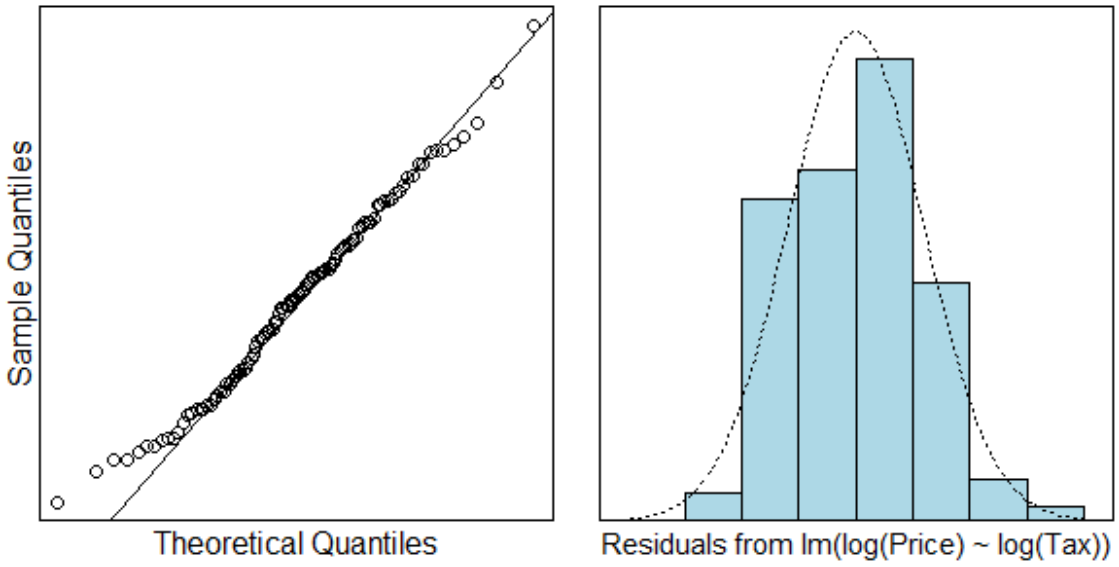
- we have commercial data where both variables (Tax and Price) are likely to have exponential distributions as they are based on property values, which will tend to have a long upper tail.
- the data reveals a pattern of a linear relationship with increasing scatter.
- both variables are right skewed.
- and a plot of the logged variables has solved any of our problems with the assumptions showing the model is appropriate.
- we want to interpret both variables in terms of percentage changes and the log-log plot appears to satisfy the linear models assumptions.

Fit model and check assumptions.

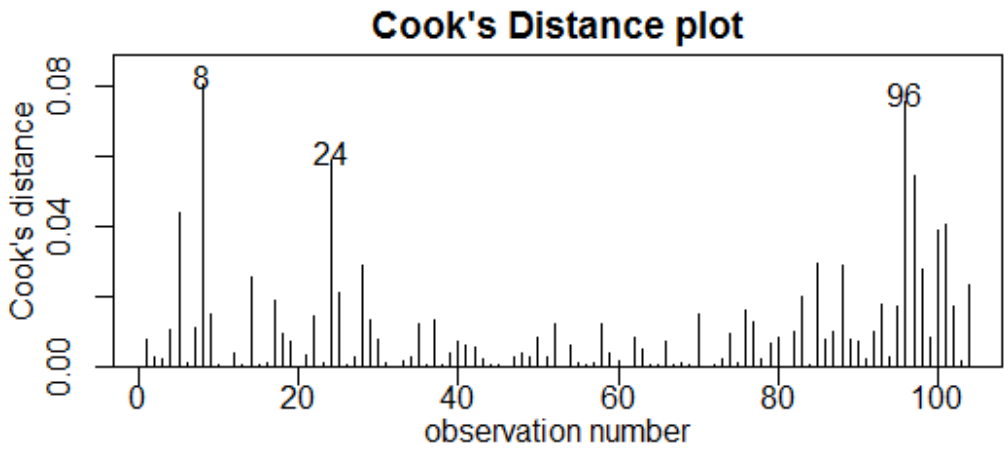
```
homefit1=lm(log(Price)~log(Tax),data=hometax.df)
plot(homefit1,which=1)
```




```
normcheck(homefit1)
```




```
cooks20x(homefit1)
```




```
summary(homefit1)
```

```
##
## 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(homefit1)
```

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

```
1.01^confint(homefit1)[2,]
```

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

```
1.5^confint(homefit1)[2,]

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

Methods and assumption checks

We want to interpret both variables in terms of percentage changes and the log-log plot appears to satisfy the linear models assumptions so we fitted a log-log model to the data. The residual plot showed approximately constant variability and no trend. Normality looks good and no influential points were detected. A random sample of hosues was taken so independence is satisfied. Model assumptions are satisfied.

Our model is: $log(Price_i) = \beta_0 + \beta_1 \times log(Tax_i) + \epsilon_i$, where $\epsilon_i \sim iidN(0, \sigma^2)$

Our model explained 85.6% of the variability in the logged data.

Executive Summary

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.

We found strong evidence that the prices of houses increased as the city tax bill was higher. The increase followed a power-law relationship.

We estimate that a city tax bill that is 1% higher is associated with a median sale price of houses that is between 0.74 and 0.87% higher.

We estimate that a city tax bill that is 50% higher is associated with a median sale price of houses that is between 35 and 42% higher.

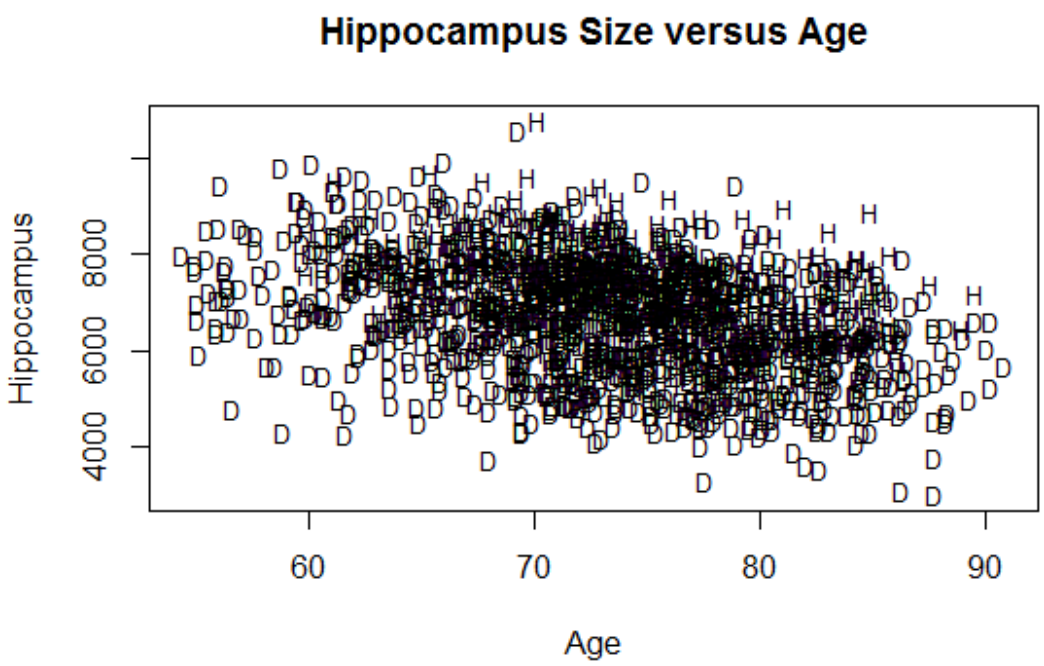
Question 2

Question of interest/goal of the study

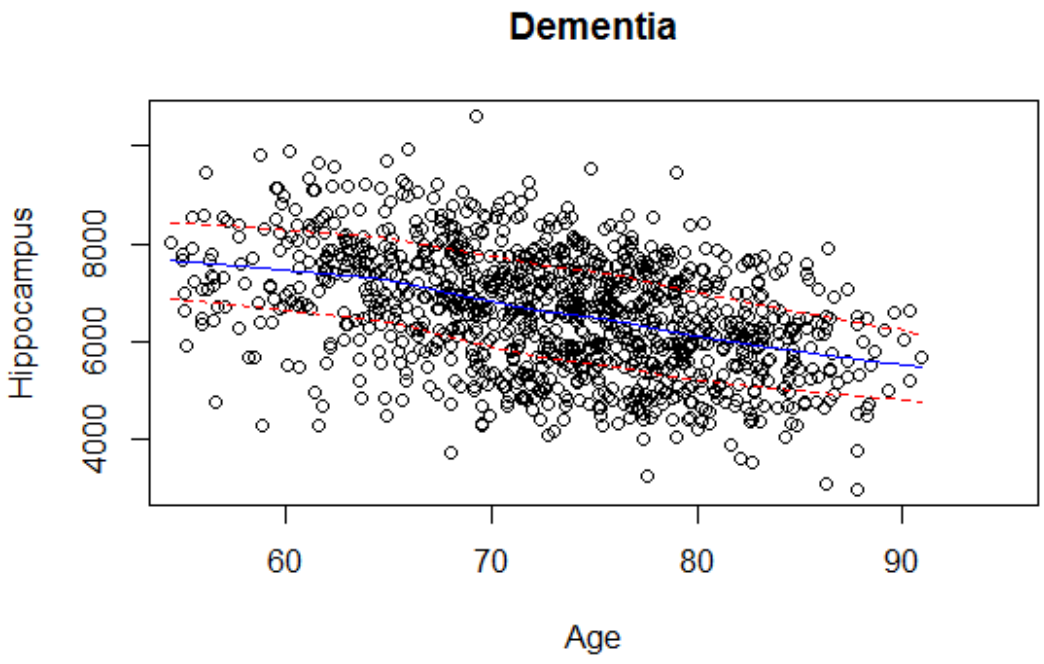
We want to explore the relationship between hippocampus size and age. In particular, we are interested in whether the relationship differs between healthy individuals and individuals with dementia related symptoms.

Read in and inspect the data:

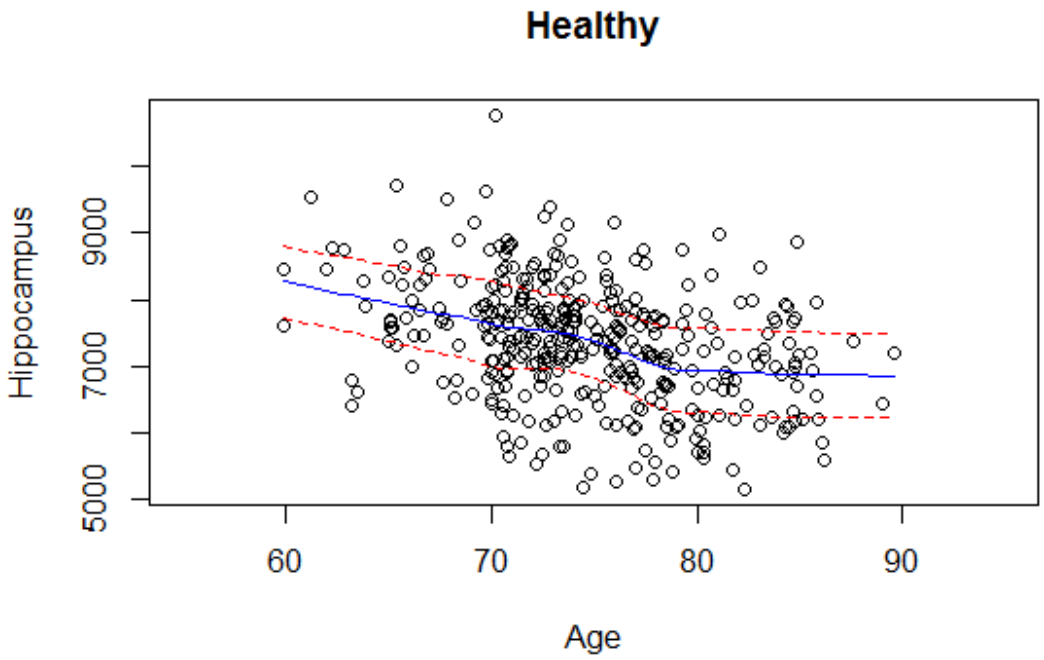
```
Hippocampus.df<-read.csv("Hippocampus.csv")
plot(Hippocampus~Age,main="Hippocampus Size versus
Age",type="n",data=Hippocampus.df)
text(Hippocampus.df$Age, Hippocampus.df$Hippocampus, Hippocampus.df$AD,
cex=.8)
```




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



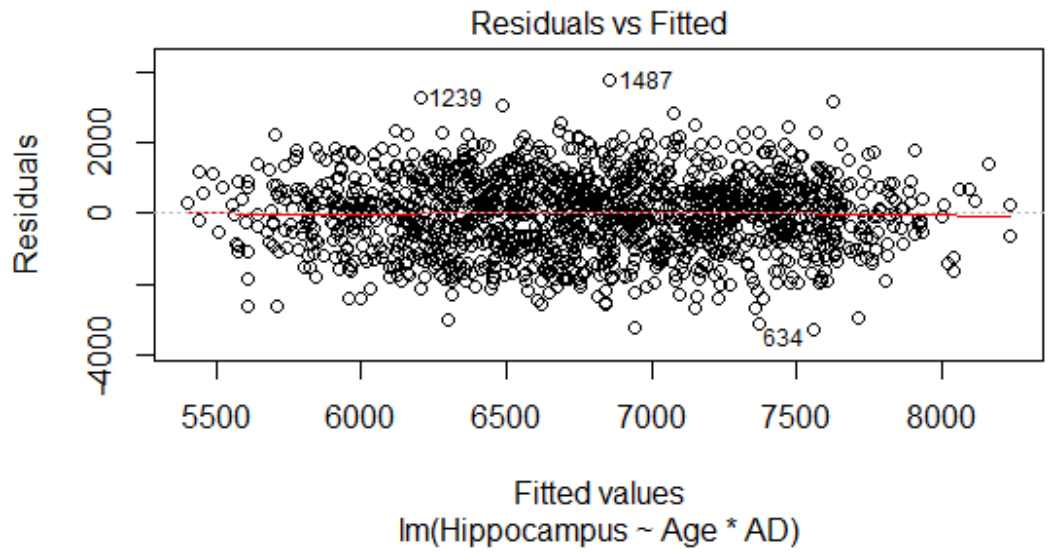

```
trendscatter(Hippocampus~Age,data=Hippocampus.df[Hippocampus.df$AD=="H",],xli
m=c(55,95),main="Healthy")
```



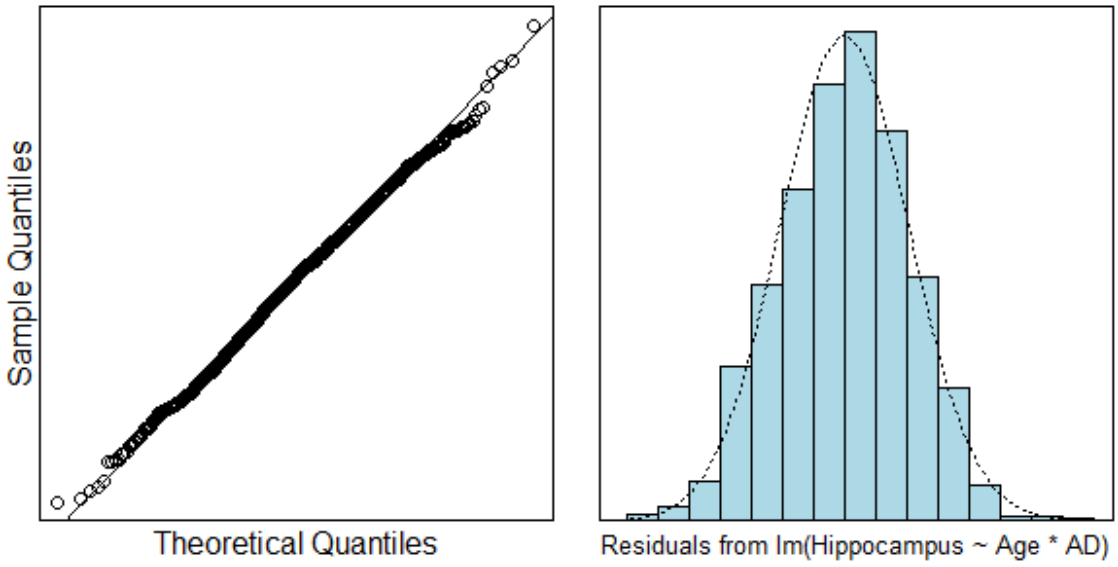
For both dementia and healthy subjects there is a decreasing linear relationship between age and Hippocampus volume. By looking at the plots, it looks like the two lines are roughly parallel, indicating there is might be any interaction between age and AD status. The scatter seems reasonably constant as age changes.

Fit model and check assumptions.

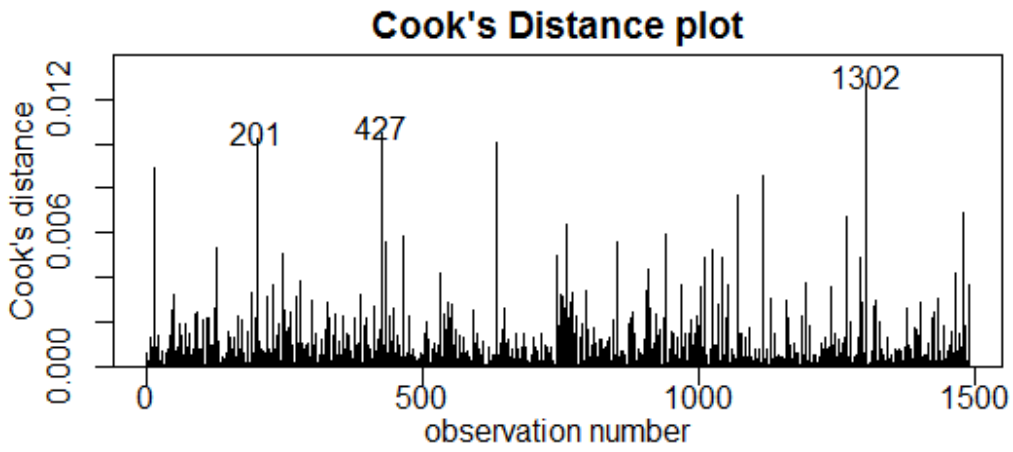
```
Hippocampusfit1=lm(Hippocampus~Age*AD,data=Hippocampus.df)
plot(Hippocampusfit1,which=1)
```



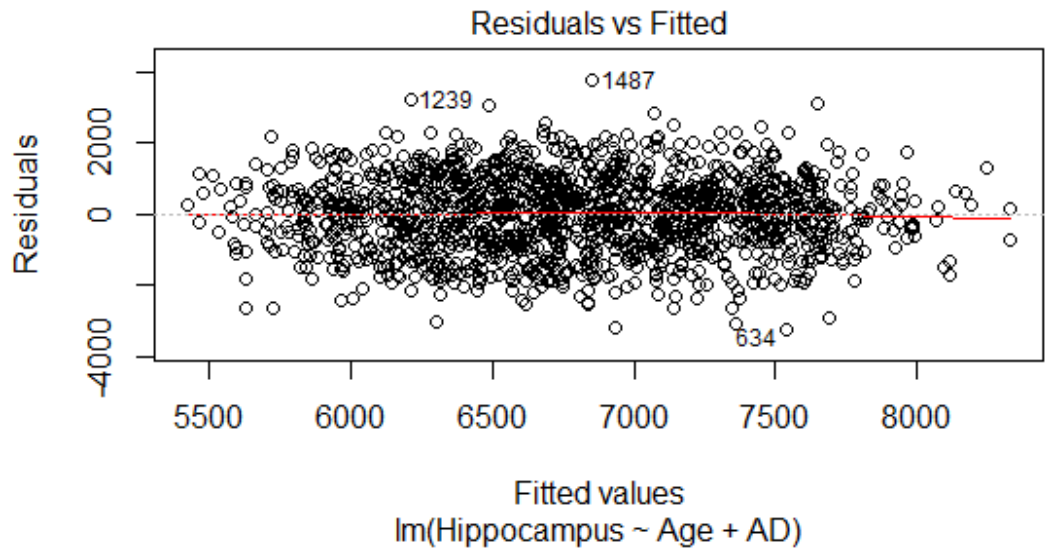

```
normcheck(Hippocampusfit1)
```



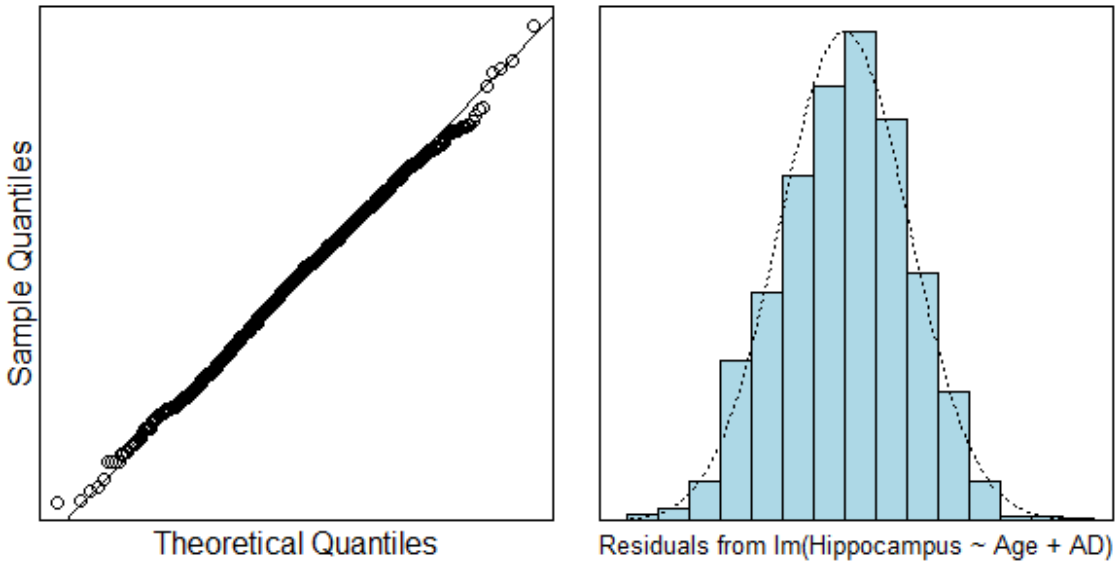

```
cooks20x(Hippocampusfit1)
```



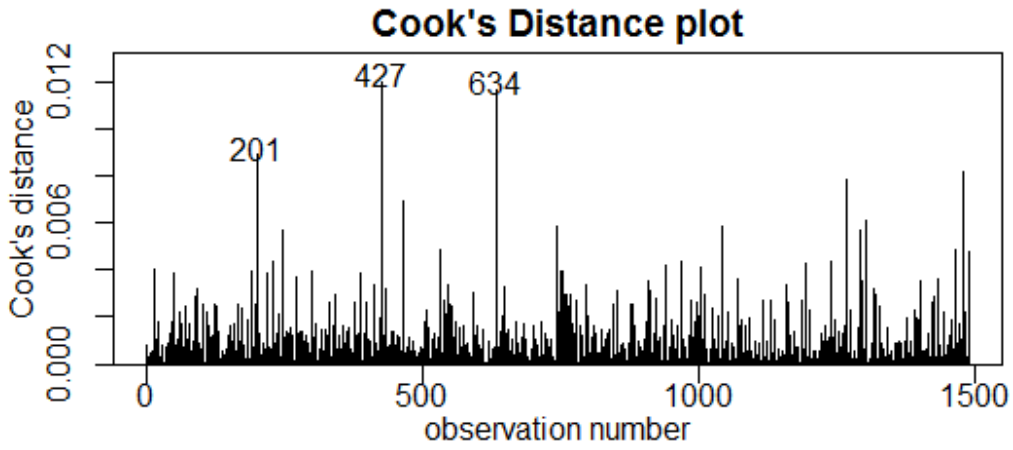

```
summary(Hippocampusfit1)

##
## Call:
## lm(formula = Hippocampus ~ Age * AD, 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 ***
## ADH           291.487    787.293   0.370    0.711
## Age:ADH         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

Hippocampusfit2=lm(Hippocampus~Age+AD,data=Hippocampus.df)
plot(Hippocampusfit2,which=1)
```




```
normcheck(Hippocampusfit2)
```



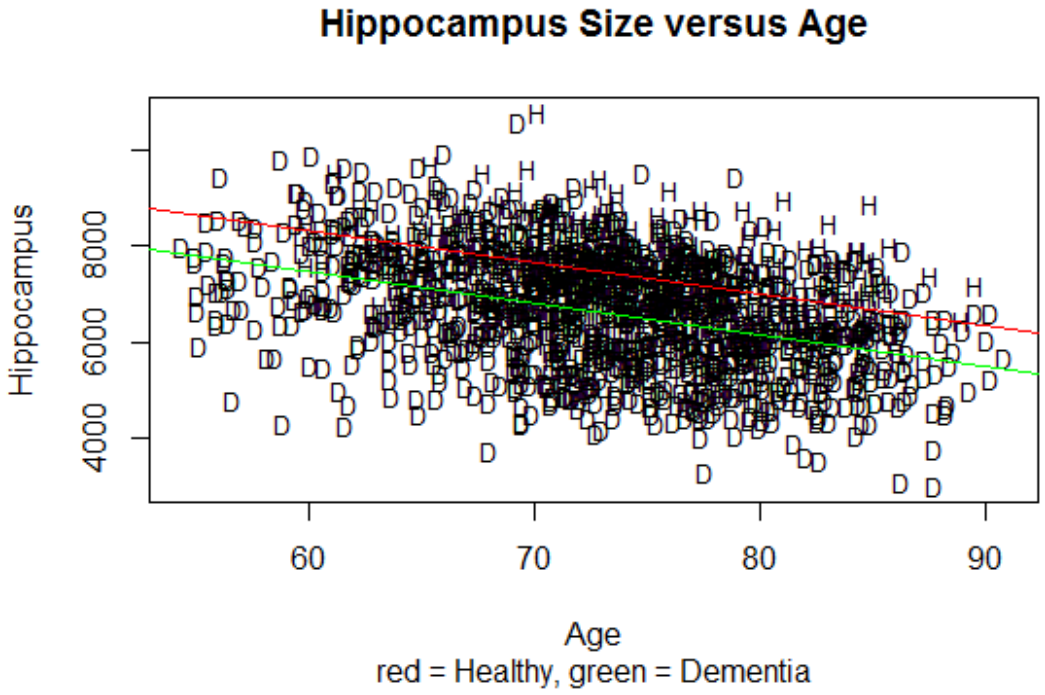

```
cooks20x(Hippocampusfit2)
```




```
summary(Hippocampusfit2)

##
## Call:
## lm(formula = Hippocampus ~ Age + AD, 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 ***
## ADH           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(Hippocampusfit2)

##                   2.5 %      97.5 %
## (Intercept) 10879.06602 11976.26257
## Age           -73.49872   -58.58644
## ADH           735.87923   980.73460
```

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",data=Hippocampus.df)
text(Hippocampus.df$Age, Hippocampus.df$Hippocampus, Hippocampus.df$AD,
cex=.8)
abline(Hippocampusfit2$coef[1],Hippocampusfit2$coef[2],col="green")
abline(Hippocampusfit2$coef[1]+Hippocampusfit2$coef[3],Hippocampusfit2$coef[2
],col="red")
```



Methods and assumption checks

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 insignificant, so the interaction term was dropped, simplifying the model to a parallel lines model.

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: $Hippocampus_i = \beta_0 + \beta_1 \times Age_i + \beta_2 \times ADH_i + \epsilon_i$ 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 explained 23% of the variability in the data.

Executive Summary

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.

We don't have evidence suggesting that the relationship between hippocampus volume and age differed between healthy individuals and individuals with dementia related symptoms.(p-value = 0.47).

However, we have strong evidence suggesting that both age and whether or not someone has dementia related symptoms independently have an impact on hippocampus volume. As age is increasing, the size of the hippocampus is tending to decrease, while healthy people have, on average, larger hippocampus's than people with dementia related symptoms.

We estimated that for each additional year increase in age, the average hippocampus volume is decreased by somewhere between 59 and 73 units, regardless of an individual's disease status.

We estimated that the average hippocampus volume for an individual with some dementia symptoms is somewhere between 736 and 981 units lower than that of a healthy individual, for any given age of the individual.

(Note: These results apply **on average**. There is a large amount of variability in both groups leading to substantial overlap in hippocampus volume between the two groups at any age.)

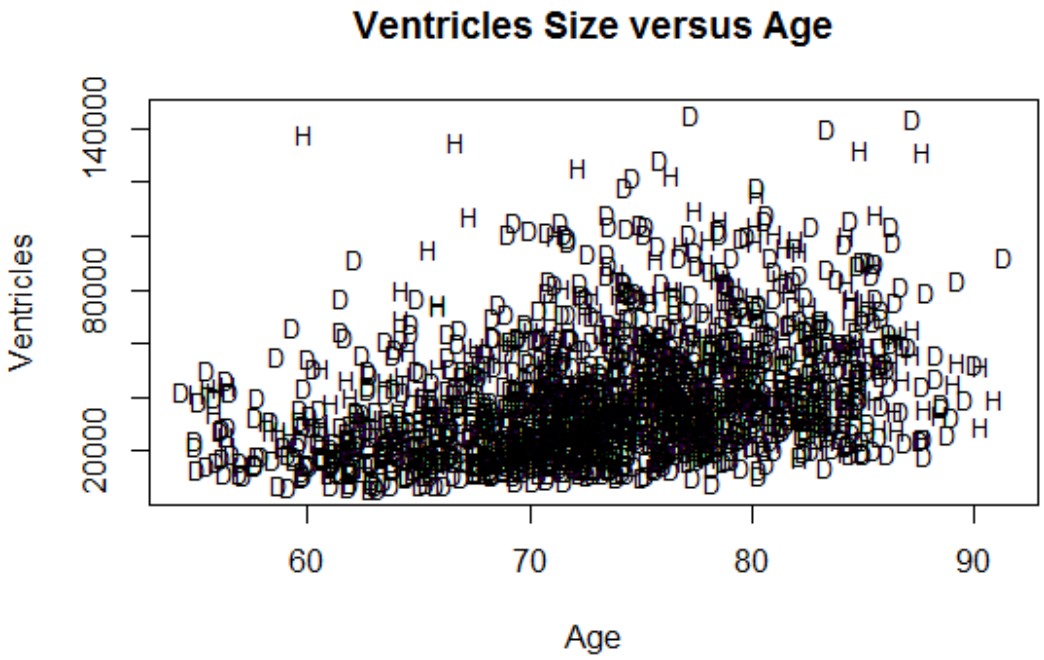
Question 3

Question of interest/goal of the study

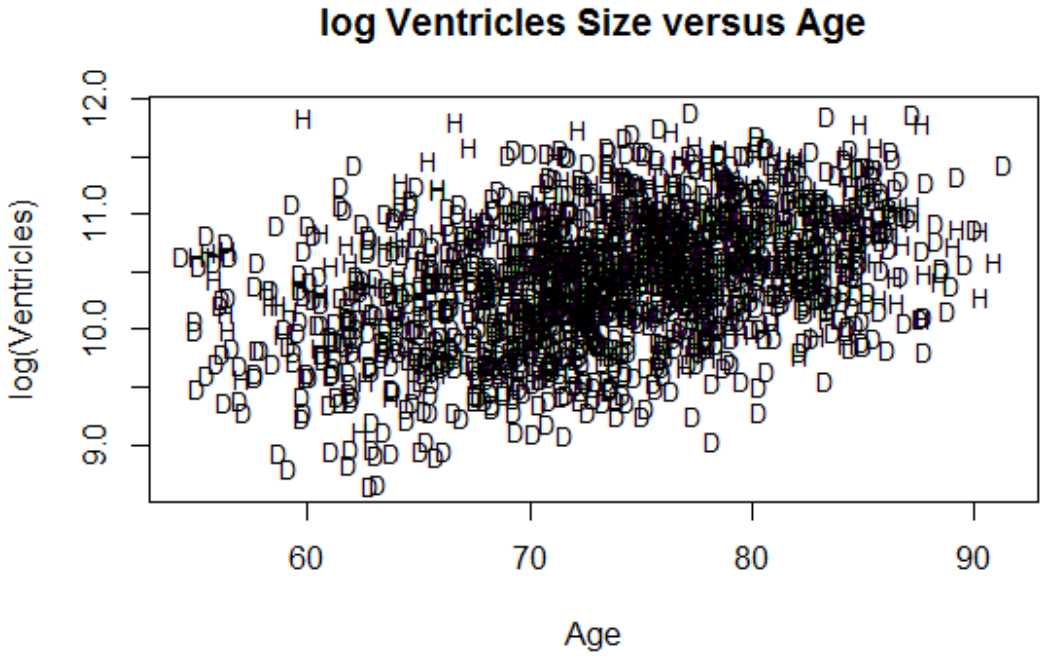
It is of interest to study the relationship between ventricles and age. In particular, we are interested in whether the relationship varies between healthy individuals and individuals with dementia related symptoms.

Read in and inspect the data:

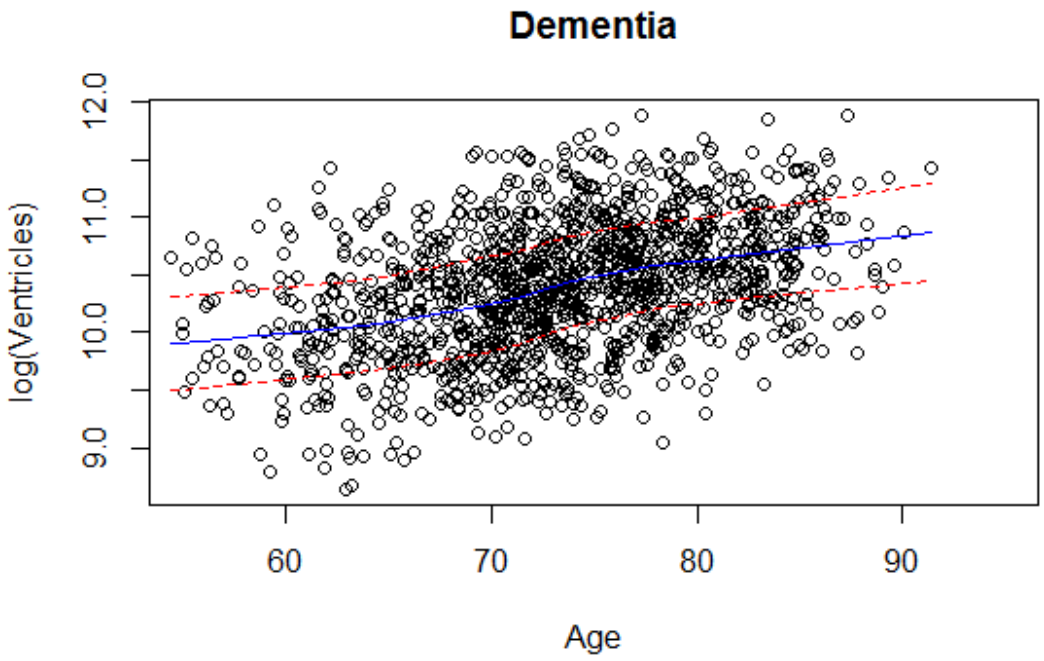
```
Ventricles.df=read.csv("Ventricles.csv")
plot(Ventricles~Age,main="Ventricles Size versus
Age",type="n",data=Ventricles.df)
text(Ventricles.df$Age, Ventricles.df$Ventricles, Ventricles.df$AD, cex=.8)
```



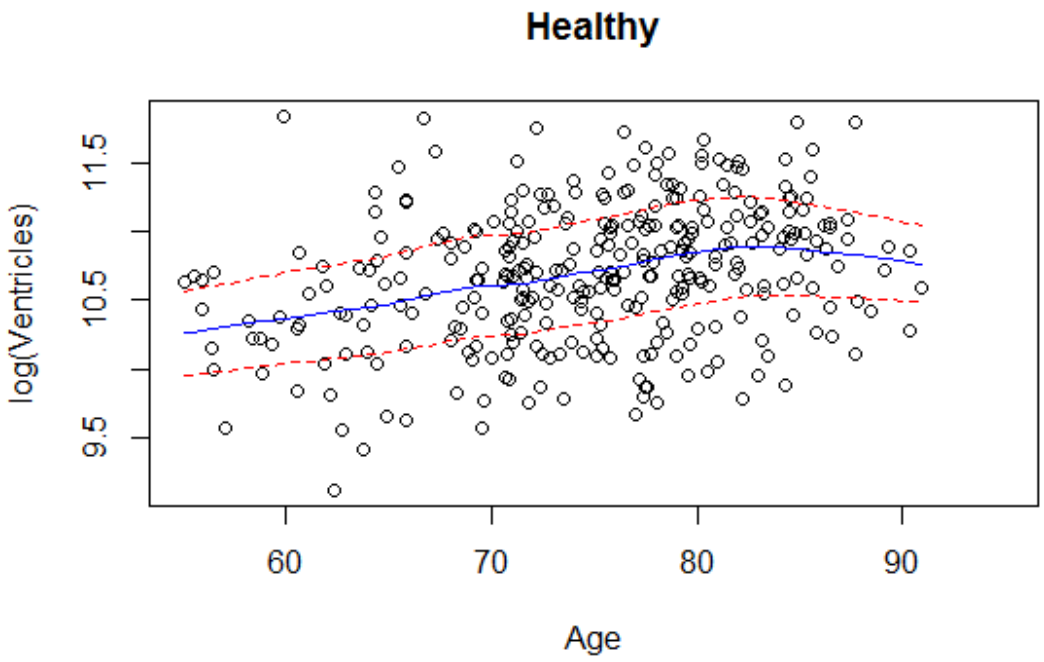

```
plot(log(Ventricles)~Age,main="log Ventricles Size versus
Age",type="n",data=Ventricles.df)
text(Ventricles.df$Age, log(Ventricles.df$Ventricles), Ventricles.df$AD,
cex=.8)
```



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

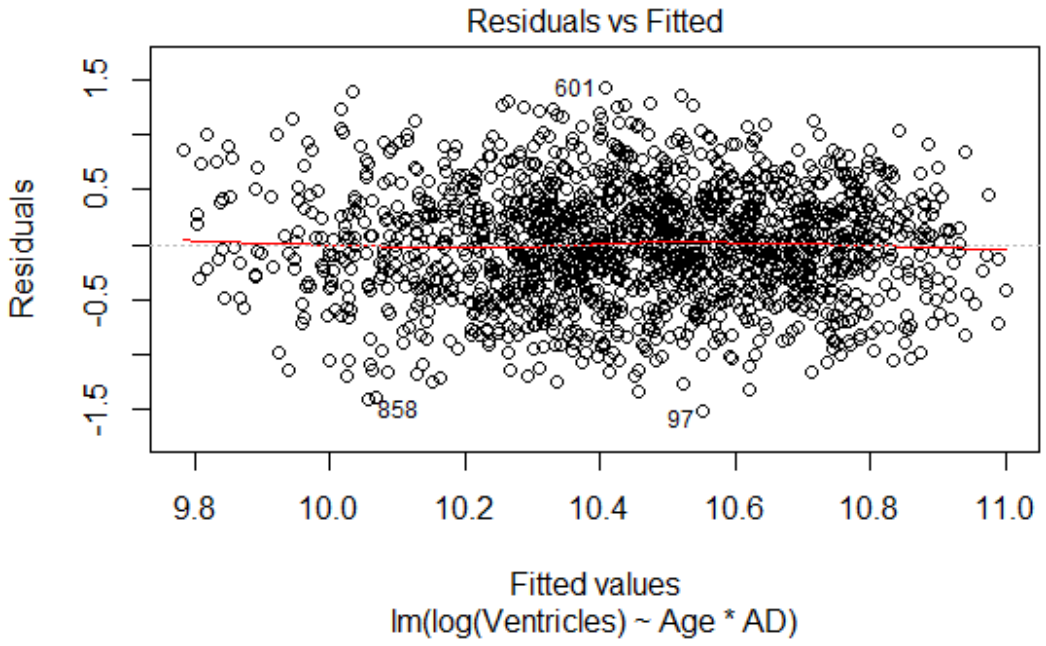



```
trendscatter(log(Ventricles)~Age,data=Ventricles.df[Ventricles.df$AD=="H",],xlim=c(55,95),main="Healthy")
```

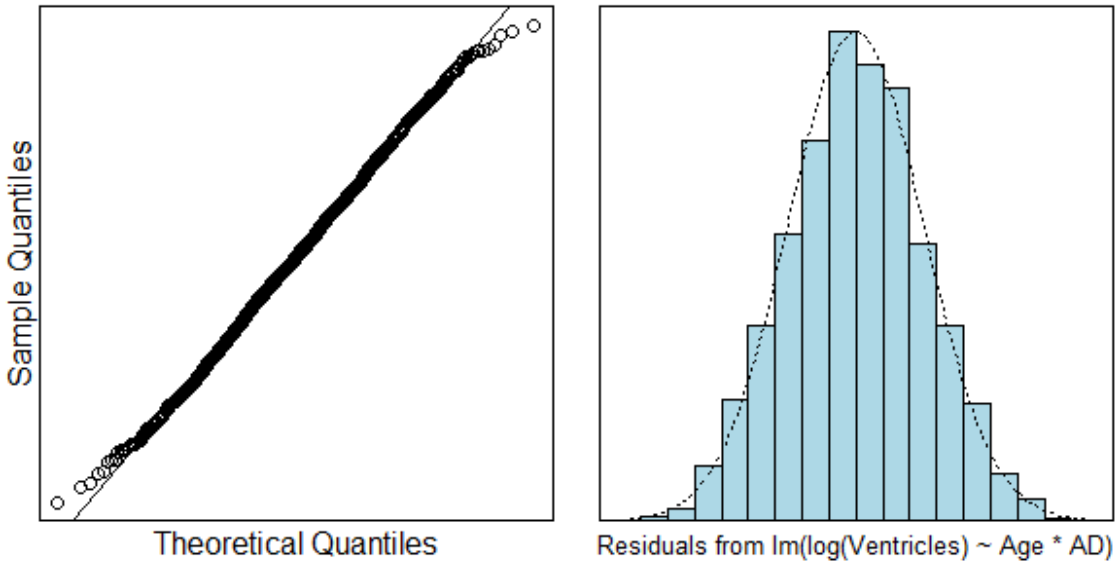


The relationship between Age and Ventricles is increasing, but as it increases, the scatter also increases. Logging the values of Ventricles evens out the scatter substantially. For both dementia and healthy subjects there is an increasing linear relationship between age and ventrical volume. By looking at the plots, it looks like the two lines have slightly different slopes, indicating there may be an interaction between age and AD status.

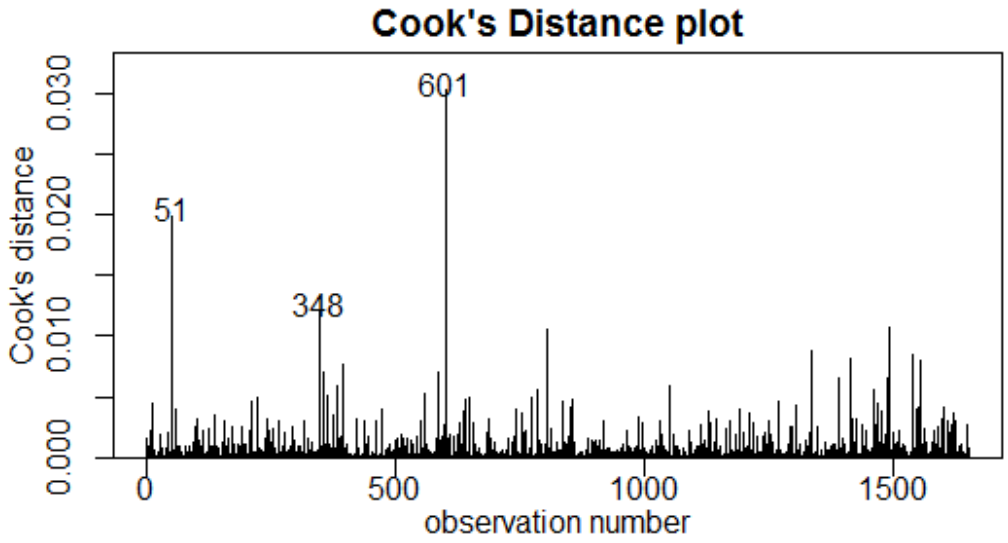
```
Ventriclesfit1=lm(log(Ventricles)~Age*AD,data=Ventricles.df)
plot(Ventriclesfit1,which=1)
```




```
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))
```

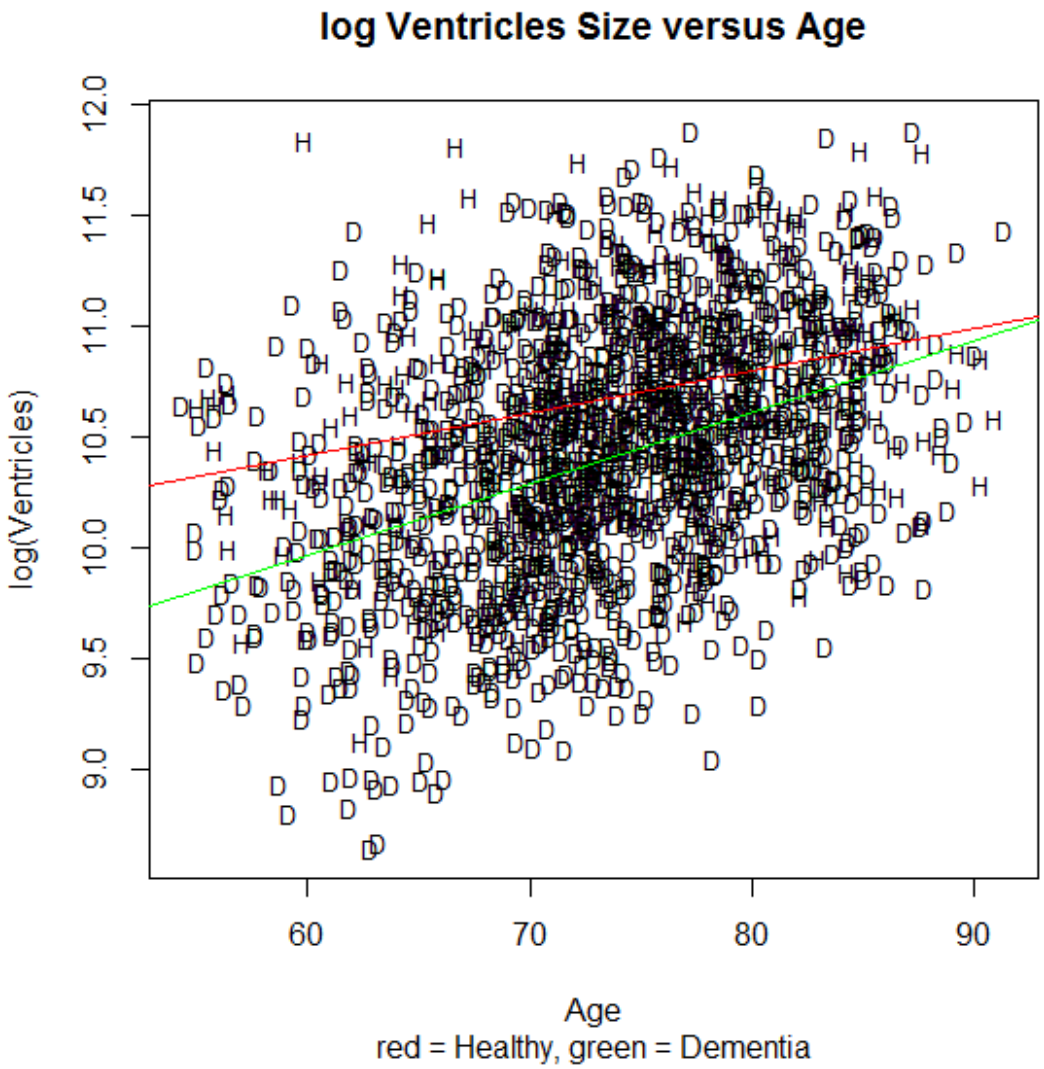
```
##                    2.5 %       97.5 %
## (Intercept) 6151.8258140 1.806983e+04
## Age            1.0120271 1.026627e+00
## ADflipD        0.1590953 5.388152e-01
## Age:ADflipD    1.0049031 1.021405e+00
```

```
(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",type="n",data=Ventricles.df)
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.

The coefficient is estimating the difference in slope between the line for Age versus log(Ventricles) when people have dementia symptoms (the base line) and the slope of the line for Age versus log(Ventricles) for healthy people.

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 cannot answer this as we have fitted a non-parallel lines model so the difference is not constant, it depends on the value for age. We can only estimate the difference at the intercept. (Which does not make a lot of sense as we are talking about new born babies with dementia.)

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

For healthy people, we estimate that for each additional year of age the median size of the ventricles increases by between 1.2 and 2.7%.

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

For people exhibiting dementia symptoms, we estimate that for each additional year of age the median size of the ventricles increases by between 2.9 and 3.7%.

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

There is initially a large difference in average ventricle size with healthy 55 year olds having larger ventricles on average. However, the rate of increase in size as age increases is larger for people with dementia symptoms so the difference in sizes slowly converges until there is little difference between the averages for the two groups when ages exceed 90.