STATS 201 Assignment 2

Li Ruoqi 2019220113

Due Date: 2021-11-07

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

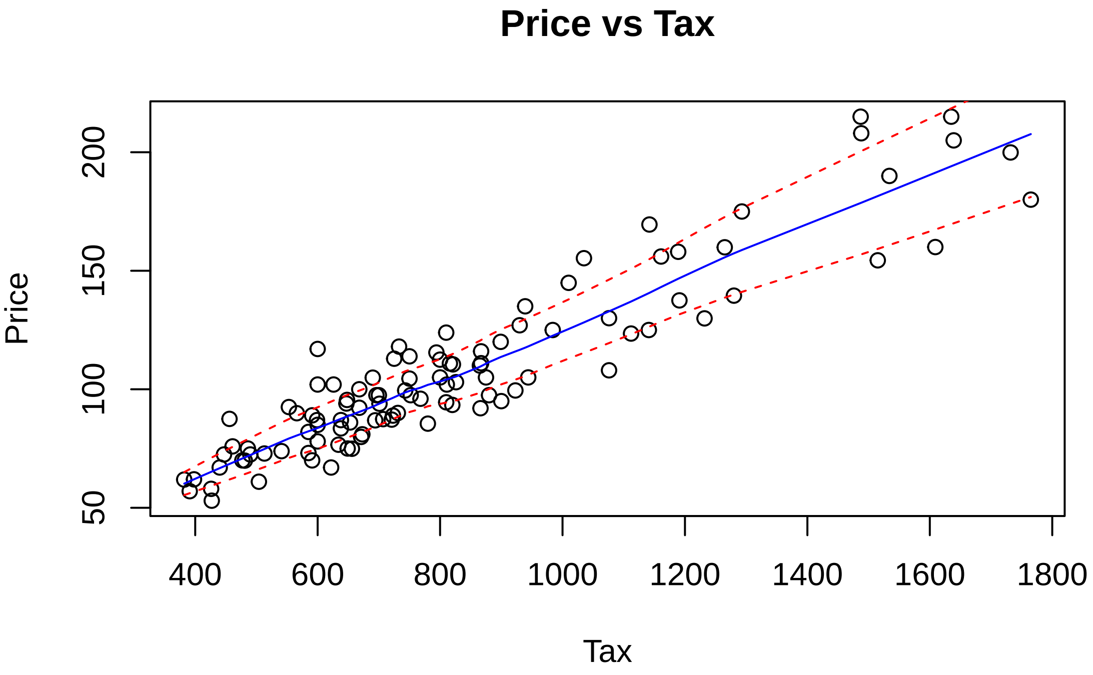
Question 1

Question of interest/goal of the study

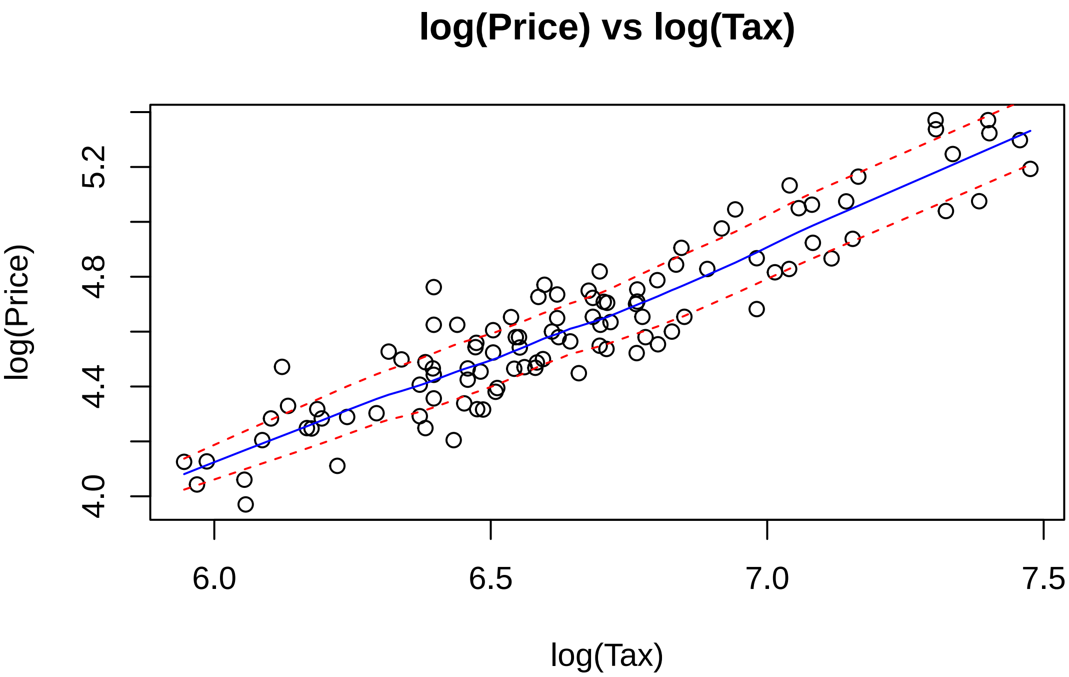
We want to build a model to explain the sale price of houses using their annual city tax bill (similar idea to rates in New Zealand) for houses in Albuquerque, New Mexico. In particular, we are interested in estimating the effect on sales price for houses which differ in city tax bills by 1% and 50%.

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)
```



Do a trendscatter plot on Price and Tax, clearly indicate that the trend is linear but the scatter is not constant, the boundary becomes wider and wider(the distance between two red dash lines become bigger and bigger) as Tax increases, means we have a quite big difference scatter when Tax is bigger and we do not see a lot of variability left. If we get Tax is small, we see a lot of variability. In other words, if we keep going on Tax, we will not expect what will happen(big variance).
If we use trendscatter plot on log(Price) and log(Tax), we see the distance between the two dash red lines are roughly constant, and a little bit smaller when log(Tax) is small compare to what happens in the middle, that probably because we not have enough data. And the relationship between log(Price) and log(Tax) is linear, with fairly amount of scatters from the central(blue) line. The trend line is actually a straight line, try to represent the center(median) of Price conditional on Tax. So the plot makes sense.

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

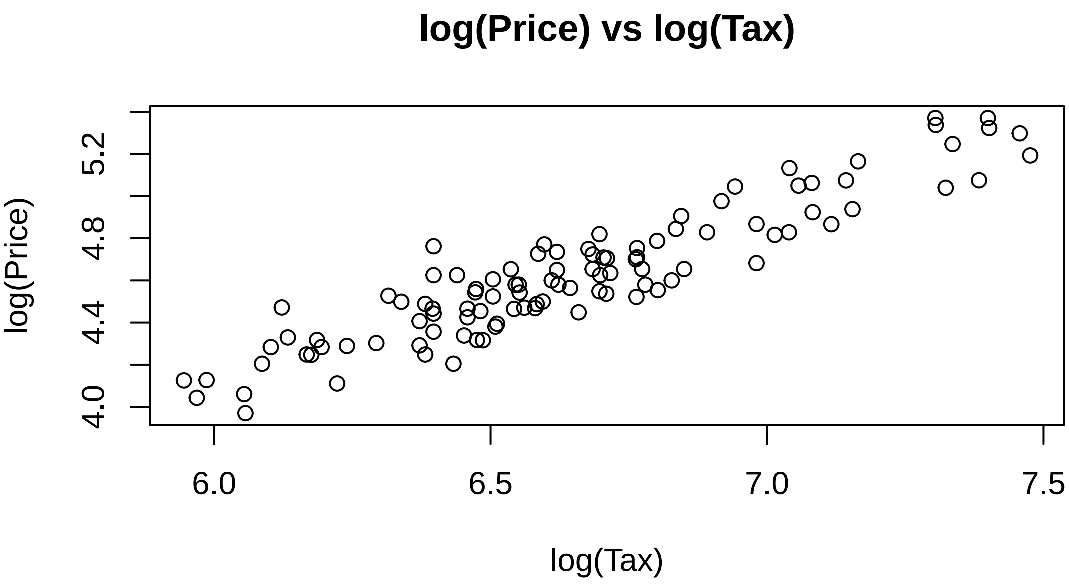
Justifying by looking at the two components we are interested in–trend and scatter. It is roughly observed that the original scatter plot has a straight line trend but the scatter is non-constant intuitively. The range between two red dash lines(variation) get bigger as the predictor Tax increases. For small log(Tax), the variance is small. But for large log(Tax), we have big variance.

If we look at the scatter plot for the log-log model, we see that we have a very clear straight line(linear) relationship between log(Price) and log(Tax), and scatter is roughly constant. To address the variance problem, we can connecting log(Price) to log(Tax) instead of connecting Price to Tax directly.

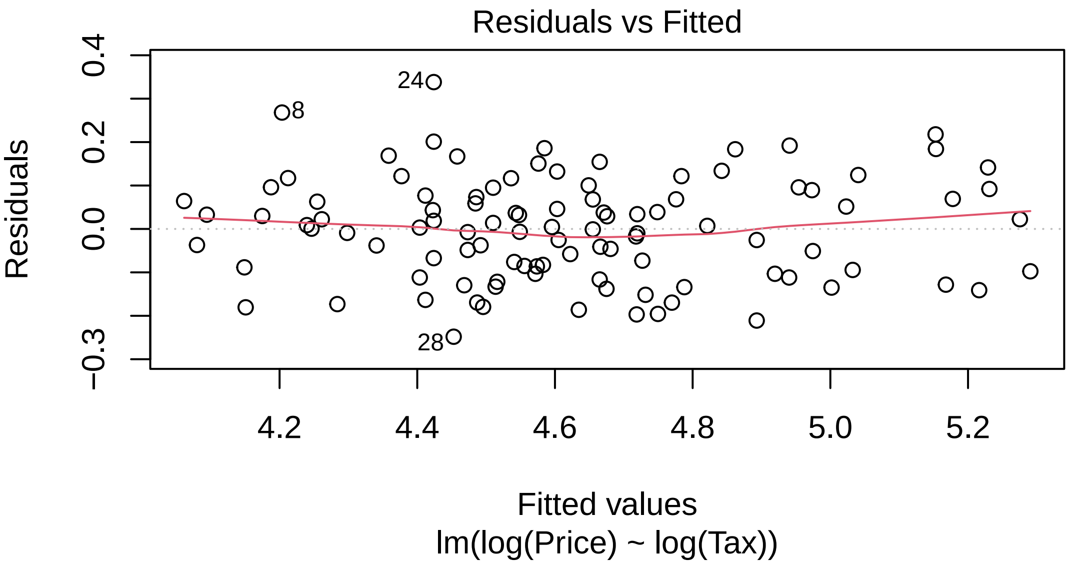
So we prefer to fitting the log-log model(power law model), it is more appropriate in this case.

Fit model and check assumptions.

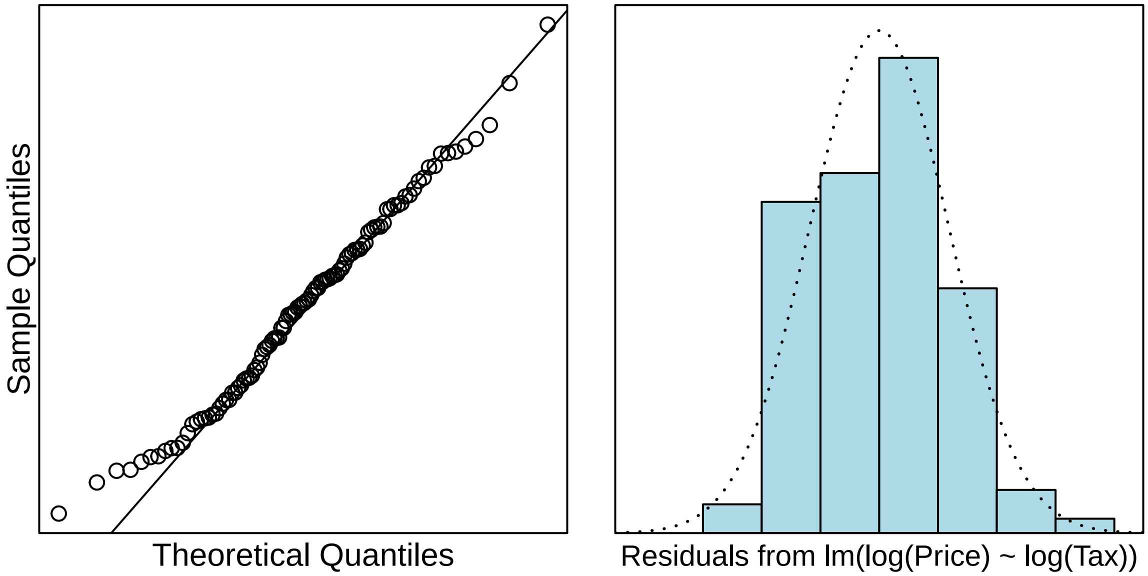
```
##fitting a linear model using log(Price) and log(Tax)
plot(log(Price)~log(Tax),main="log(Price) vs log(Tax)",data=hometax.df)
```



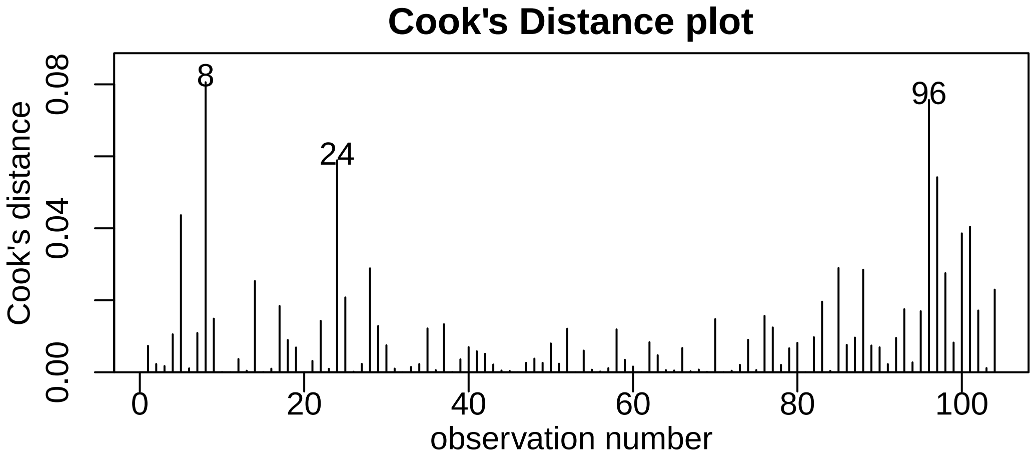
```
hometax.lm = lm(log(Price)~log(Tax), data=hometax.df)
plot(hometax.lm, which = 1)
```



```
##check the normality assumption
normcheck(hometax.lm)
```



```
##check for influential observations
cooks20x(hometax.lm)
```



```
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
```

```
##CI for beta0 and beta1
confint(hometax.lm)
```

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

```
##CI for beta1
confint(hometax.lm)[2,]
```

```
##     2.5 %    97.5 %
## 0.7385139 0.8677080
```

```
## 1% increase
1.01^confint(hometax.lm)[2,]
```

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

```
## 50% increase
1.5^confint(hometax.lm)[2,]
```

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

Fitting our model, it is linear on the log scale.

If we look at the residual plot, roughly is linearity, slightly get the negative residual between two extreme ends, for small fitted values and large fitted values, our residual tends to be positive. There is no pattern, it is a random scatter about 0 line, means linearity is probably OK. And if we look at the scatter, the scatter is roughly constant, we might have a little bit variance be small towards ends. That is probably because we do not have enough data towards the two ends rather than the variance not being constant.

Look at the normality. It is all good enough, seems to be OK.

Look at influential points, we don't have any problem, all of the data points has Cook's distance less than 0.4, indicating we can trust our model and summary.

The model assumptions look to be reasonably well satisfied, so we can use the fitted model to make statistical inference. P-value is significantly indicating that there is a relationship between Tax and Price.

Methods and assumption checks

The data seem to suggest between Price and Tax is a linear trend, but it has non-constant scatter. The trend of log(Price) and log(Tax) is linear with constant scatter.

So we have fitted a model with log(Price) and log(Tax)(power law model), with the relationship between log(Price) and log(Tax) is linear.

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

The residual plot on log scale shows roughly linearity, it is a random scatter about 0(zero) line constantly and a straight line has no pattern left between fitted values and residuals. We might have a little bit variance be small towards ends. That is probably because we do not have enough data towards the two ends rather than the variance not being constant. The normality is all good enough, seems to be OK. There is no unduly influential data points, all of the data points has Cook's distance less than 0.4, indicating we can trust our model and summary in this case.

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 86% of the total variation in the response variable, and so will be reasonable for prediction.

Executive Summary

Our aim is to find the relationship between sale price of houses and their annual city tax bill (similar idea to rates in New Zealand) for houses in Albuquerque, New Mexico. In particular, we want to estimate the effect on sales price for houses which differ in city tax bills by 1% and 50%.

We have strong evidence that suggest the relationship between sale price of houses and their annual city tax bill exists on log scale. The relationship looks like linear on log scale. The price of the houses was exponentially changing as the tax changes.

We estimated every 1% increases in Tax value results in a 0.73% to a 0.86% increase in the median value of sale price.

We estimated the increase our Tax by 50%(multiply 1.5), the increasing Tax by 50% corresponds to an increase in median sale price between 34.9% and 42.2%.

Our model explains 86% of the total variation in the response variable, and so will be reasonable for prediction.

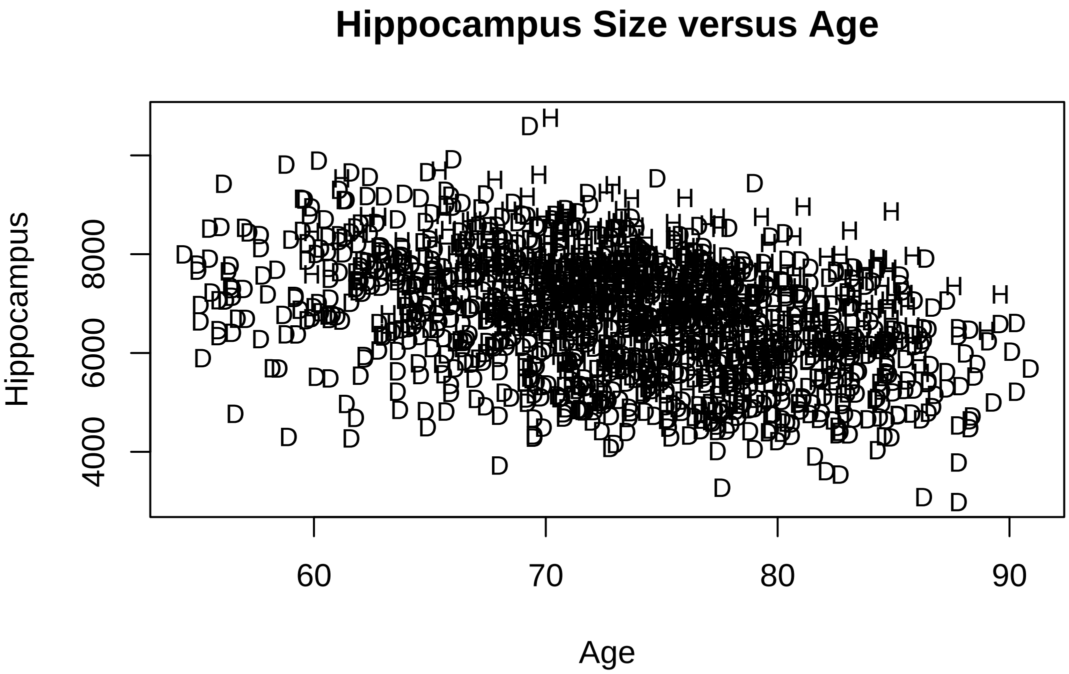
Question 2

Question of interest/goal of the study

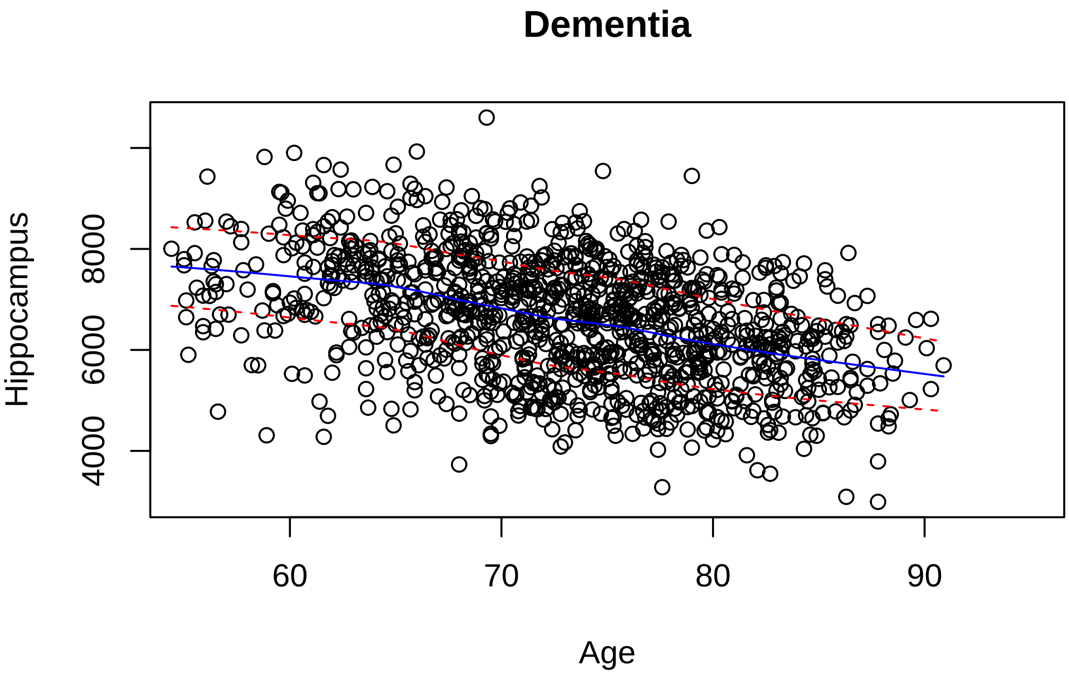
We want to explore the relationship between hippocampus size and age. In particular, we are interested in whether the relationship differs between healthy individuals and individuals with dementia related symptoms.

Read in and inspect the data:

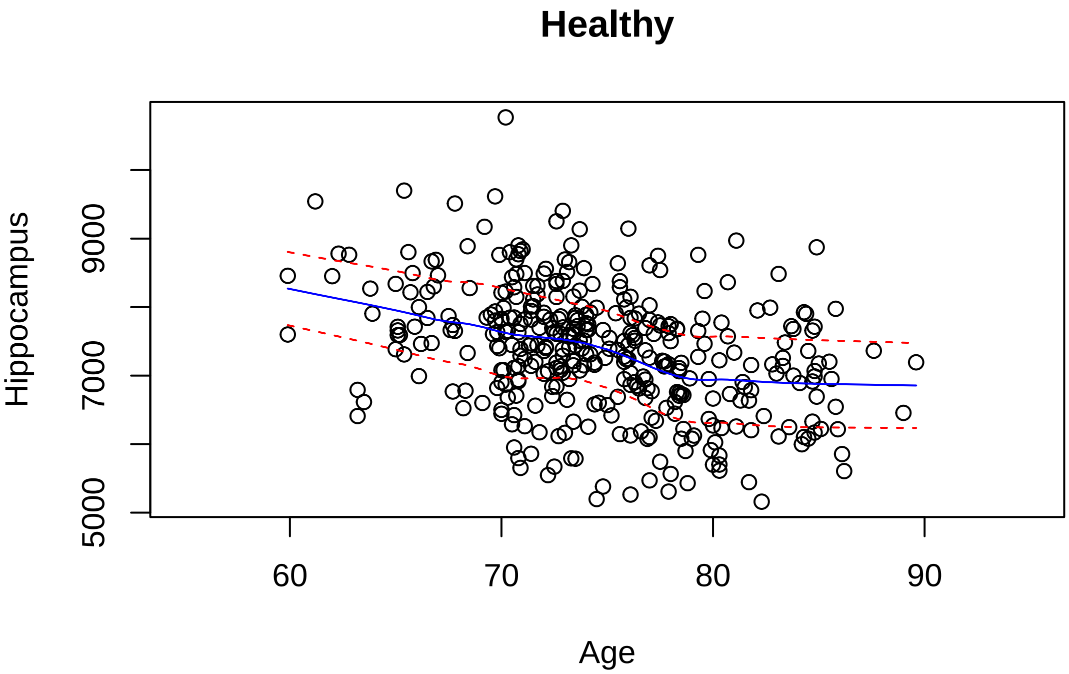
```
Hippocampus.df<-read.csv("Hippocampus.csv")
plot(Hippocampus~Age,main="Hippocampus Size versus Age",type="n",data=Hippocampus.df)
text(Hippocampus.df$Age, Hippocampus.df$Hippocampus, Hippocampus.df$AD, cex=.8)
```


Hippocampus Size versus Age

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


Dementia

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



Looking at the Hippocampus Size and Age plot, essentially use D and H, D is indicating the dementia individual, and H indicating the healthy individual. The plot is a bit cluttered, so we look at each plot for each individual type.

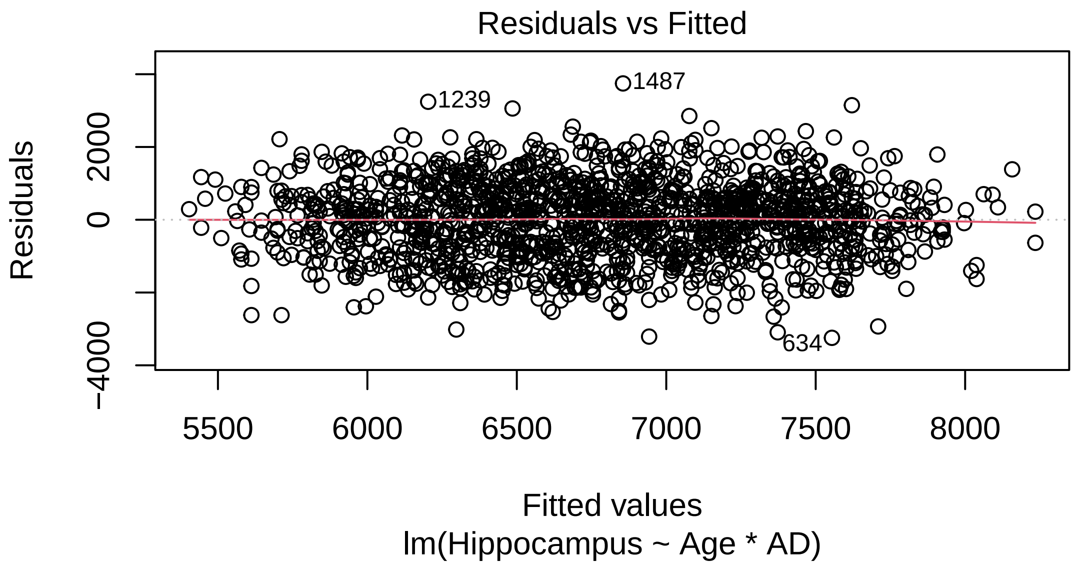
The both trendscatter plots seem to suggest that there is a negative relationship between Hippocampus and Age on both Healthy and Dementia, the relationship is not actually strong, therefore we do not know whether the relationship is true or not(actually exists or not). In other words, we do not know whether it is statistically significant or not. To address this question we need to construct model and do a hypothesis test.

Separately for the Healthy plot, we can see hint of non-linear relationship, looks like a little curvature, but not serious that can worry us.

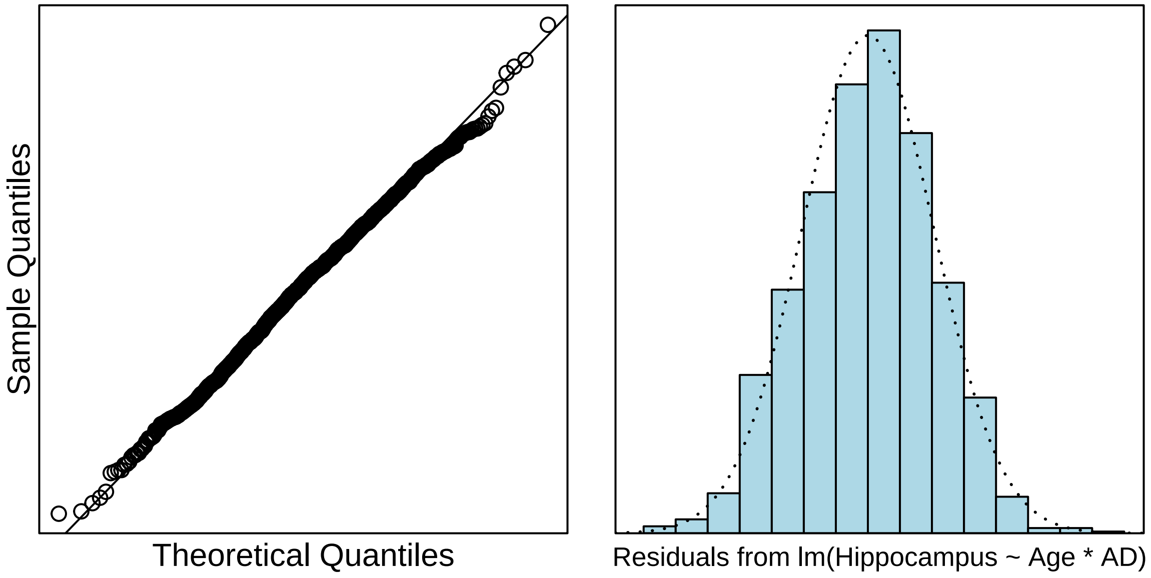
The relationship between Hippocampus Size and Age are not exactly the same for Healthy and Dementia individuals. These are what we have observed.

Fit model and check assumptions.

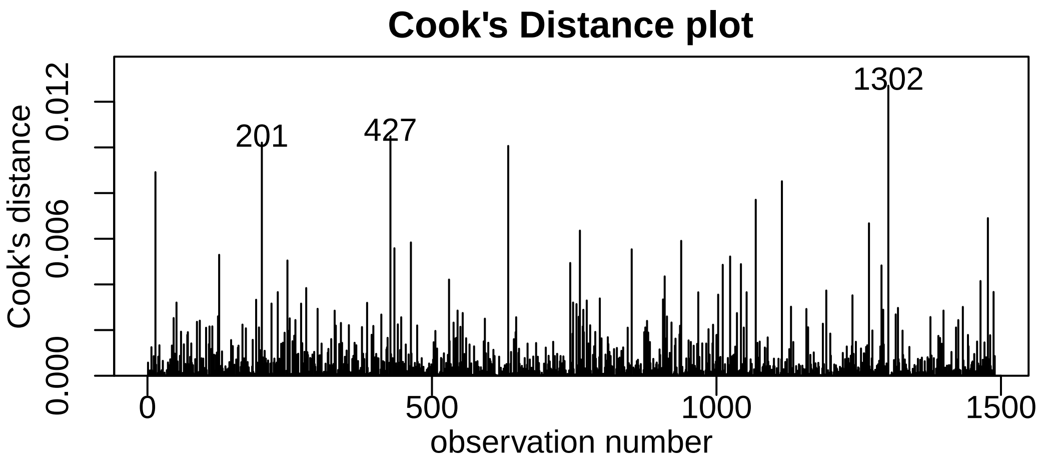
```
##use dummy variable D to fit model

##Hippocampus.df$D = as.numeric(Hippocampus.df$AD == "H")
##table(Hippocampus.df$AD, Hippocampus.df$D)
##Hippocampus.df$AgeD = with(Hippocampus.df, {AgeD = D * Age})
##Hippocampus.fit=lm(Hippocampus ~ Age + D + AgeD, data=Hippocampus.df)

##Prefer to do a simple way
Hippocampus.fit2 = lm(Hippocampus ~ Age * AD, data = Hippocampus.df)
plot(Hippocampus.fit2, which = 1)
```



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



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



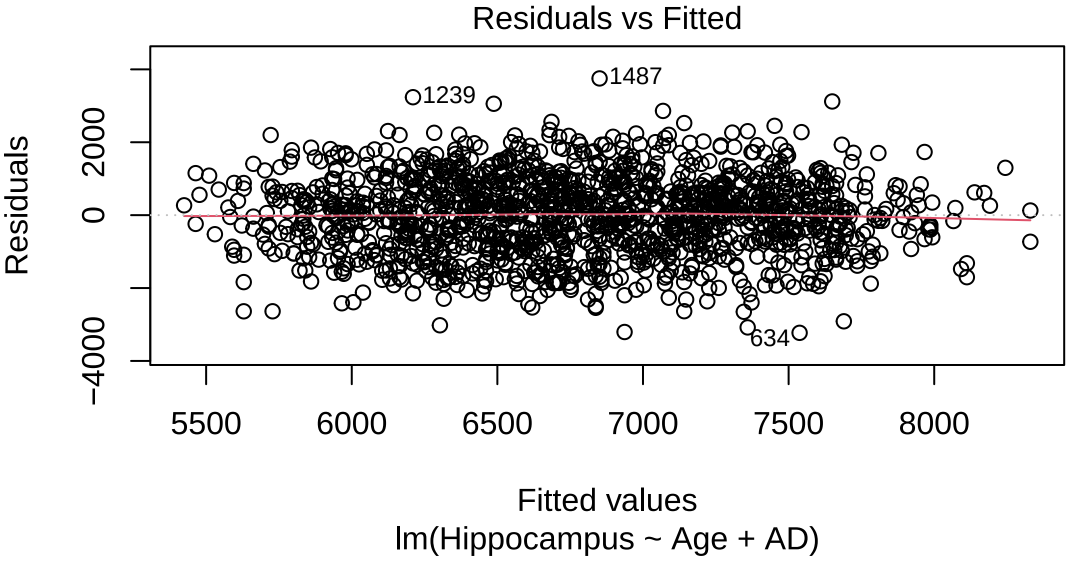
```
summary(Hippocampus.fit2)
```

```
##
## 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
```

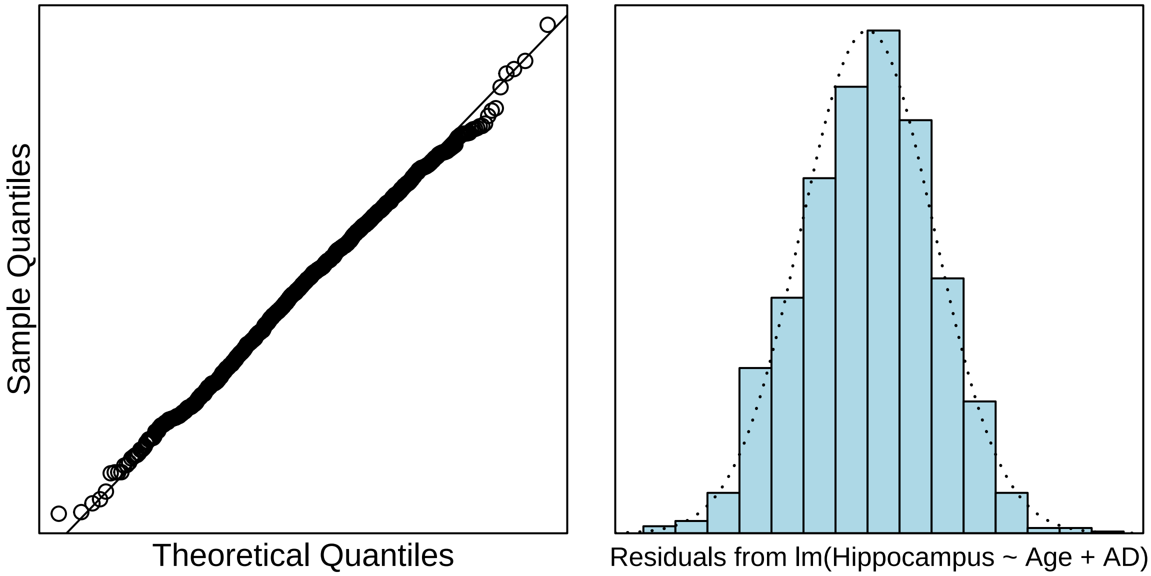
```
confint(Hippocampus.fit2)
```

```
##                    2.5 %      97.5 %
## (Intercept) 10917.36123 12108.97432
## Age           -75.31739   -59.10647
## ADH         -1252.83698  1835.81096
## Age:ADH       -13.06983    28.30305
```

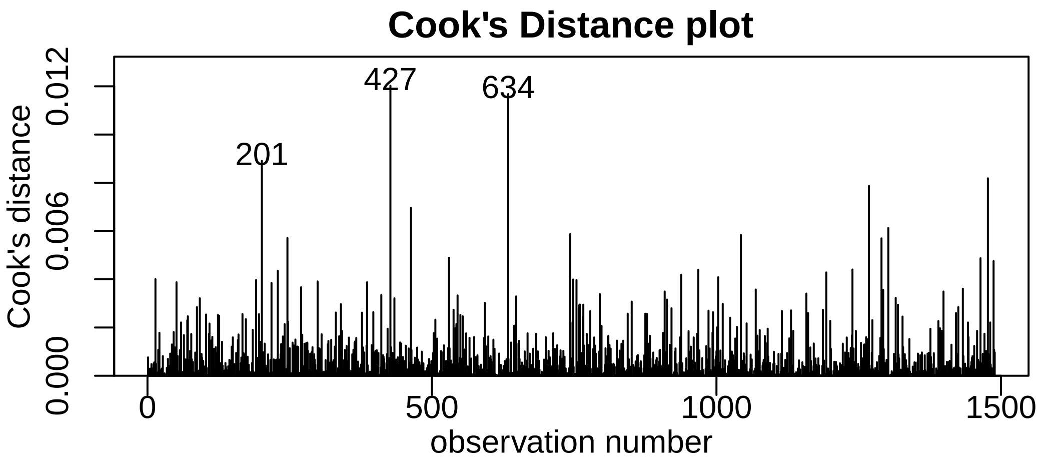
```
## fit a simpler model by removing the interaction term
Hippocampus.fit3 = lm(Hippocampus ~ Age + AD, data = Hippocampus.df)
plot(Hippocampus.fit3, which = 1)
```



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



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



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

```
##
## 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(Hippocampus.fit3)
```

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

```
anova(Hippocampus.fit3)
```

```
## Analysis of Variance Table
##
## Response: Hippocampus
##             Df    Sum Sq   Mean Sq F value    Pr(>F)
## Age          1 281920506 281920506  261.16 < 2.2e-16 ***
## AD           1 204153080 204153080  189.12 < 2.2e-16 ***
## Residuals 1486 1604149568   1079508
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For the linear model with interaction term, the residual plot looks reasonably good, and scatters seems to be fairly constant. The normality assumption is as good as we can hope for. Everything is all below 0.4, no unduly influential points in this case. but we have three really slightly bigger than the rest. We can trust the fitted linear model.

Summary shows Age is significant, for people who is dementia, the Age explains the Hippocampus size.

For the additional slope, seems to no evidence suggest the slope for who is health is bigger (p-value >0.05 on the 5% level). So it is not a good prediction ,we need to do a better model.

Look for CI for all the parameters in our first model, CI for beta2 and beta3 contain 0, indicating the hypothesis testing whether beta2=beta3=0, we do not have evidence to against so beta2 and beta3 maybe zero. In other words, whether people is dementia or not, not going to affect the intercept the slope, it seems unreasonable.

Because the term involving interaction is not significant (large P-value =0.47 for the ':'term), then we use a simpler (main effects) model. we simply replace * by + in the model formula.

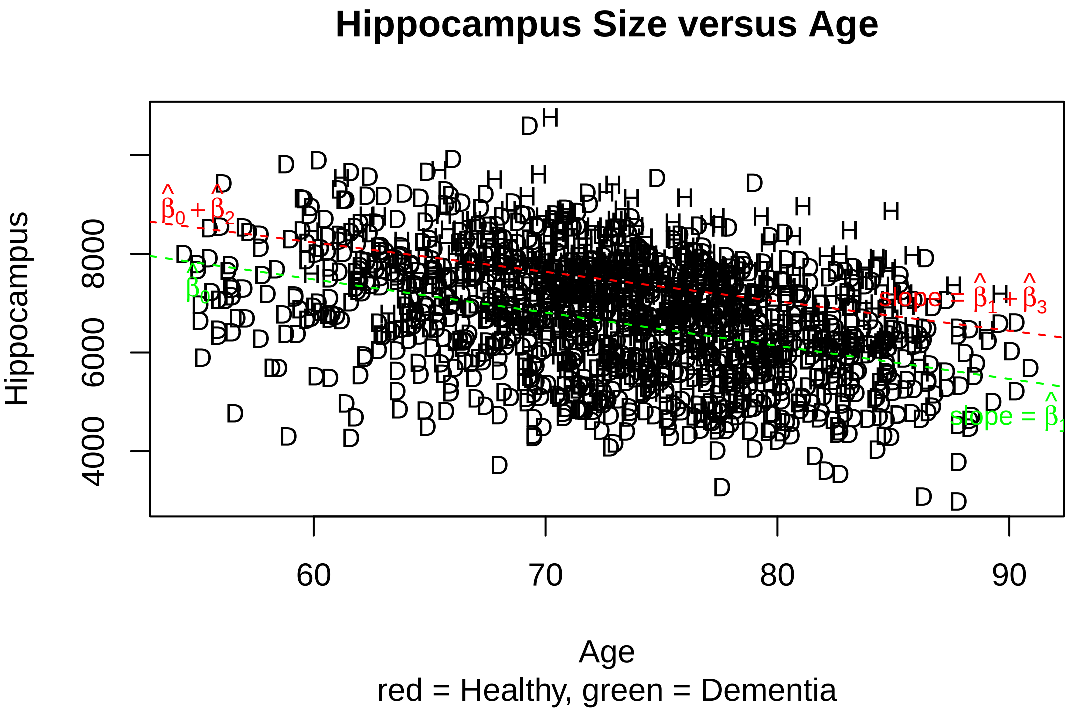
For our new model(the linear model without interaction term), all assumption is satisfied, and we can trust this new model, as all we have done is remove a term that was not significant. from the anova table,The P-value associated with the AD term is very small, so we conclude that the intercepts are different. We do have to fit different intercepts for health and dementia. Our instincts were right.

Summary shows Age and ADH both significant. Age indicating that for people who is dementia, the Age explains the Hippocampus size. ADH seems to suggest that it differs in intercept between dementia and health.

For our new model(, the CI for all parameters is really good(no 0 included). So we prefer the simpler (main effects) model we have fitted.

Plot the data with your appropriate model superimposed over it

```r
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)

coeffs = Hippocampus.fit2$coef
##for dementia
abline(coeffs[1:2], lty = 2, col = "green")
text(55, 7400, expression(hat(beta)[0]), col = "green", cex = 0.8)
text(90, 4800, expression("slope = " * hat(beta)[1]), col = "green", cex = 0.8)
##for healthy
abline(coeffs[1:2] + coeffs[3:4], lty = 2, col = "red")
text(55, 9000, expression(hat(beta)[0] + hat(beta)[2]), col = "red", cex = 0.8)
text(88, 7200, expression("slope = " * hat(beta)[1] + hat(beta)[3]), col = "red", cex = 0.8)
```



Methods and assumption checks

A plot of the data showed that Hippocampus size appear to decrease linearly with age increases, but with different lines for dementia and healthy individuals.

The model with interaction was fitted but the Age/ADH interaction was found to be not significant. After removing the term that was not significant, we have fitted a simpler model(non-interaction model/main effects model).

The individuals should be obviously independent, we have a random sample individuals so we can assume they form an independent and representative sample.

The residual plot looks reasonably good, and scatters seems to be fairly constant. The normality assumption is as good as we can hope for. And no unduly influential points in this case, but we have three really slightly bigger than the rest, not serious. All assumptions are satisfied.

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

Our model only explained 23% of the variability in the data.

Executive Summary

Our aim is to investigate 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 have strong evidence that suggest there is a clear linear relationship between Hippocampus size and age, but is differs in intercept between dementia and healthy individuals.

We estimated every 1 unit(each) additional age will decrease expected Hippocampus size of dementia individuals by 58.59 to 73.50.

We estimated the mean Hippocampus size of dementia individuals is on average smaller than the healthy individuals.

Our model only explained 23% of the variability in the data, and so will be not reasonable for prediction.

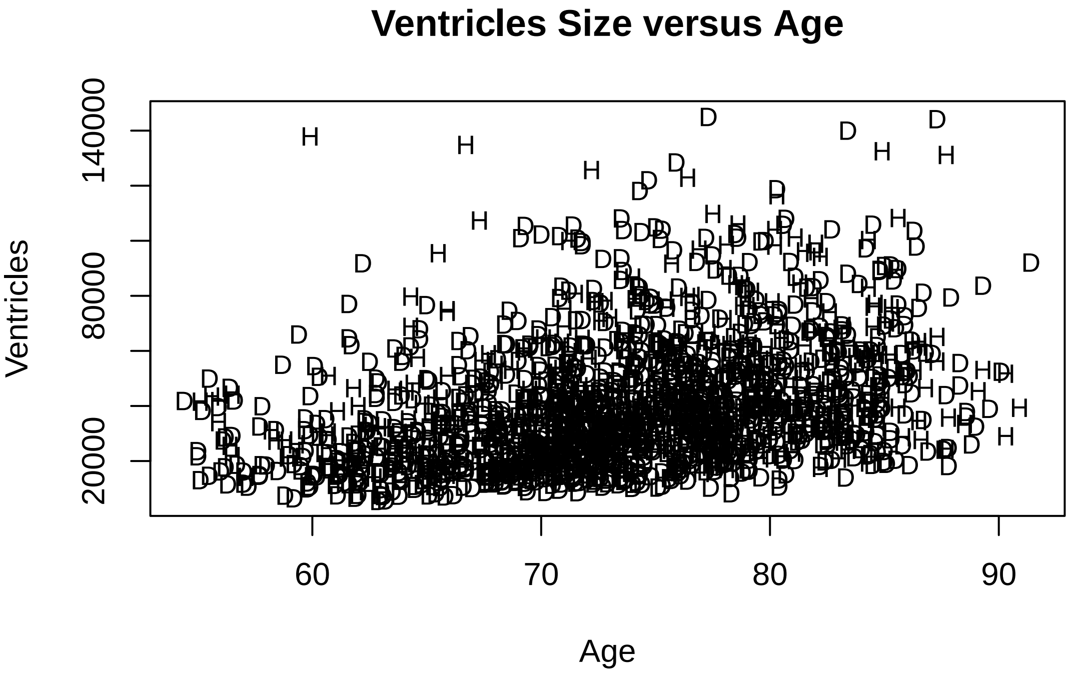
Question 3

Question of interest/goal of the study

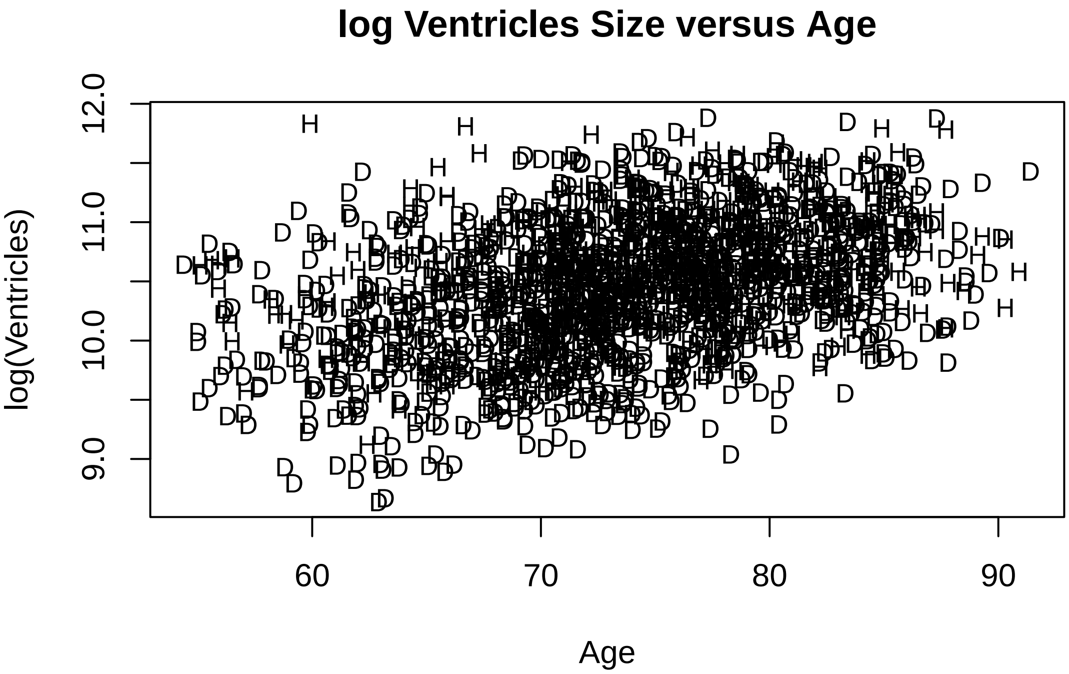
It is of interest to study the relationship between ventricles and age. In particular, we are interested in whether the relationship varies between healthy individuals and individuals with dementia related symptoms.

Read in and inspect the data:

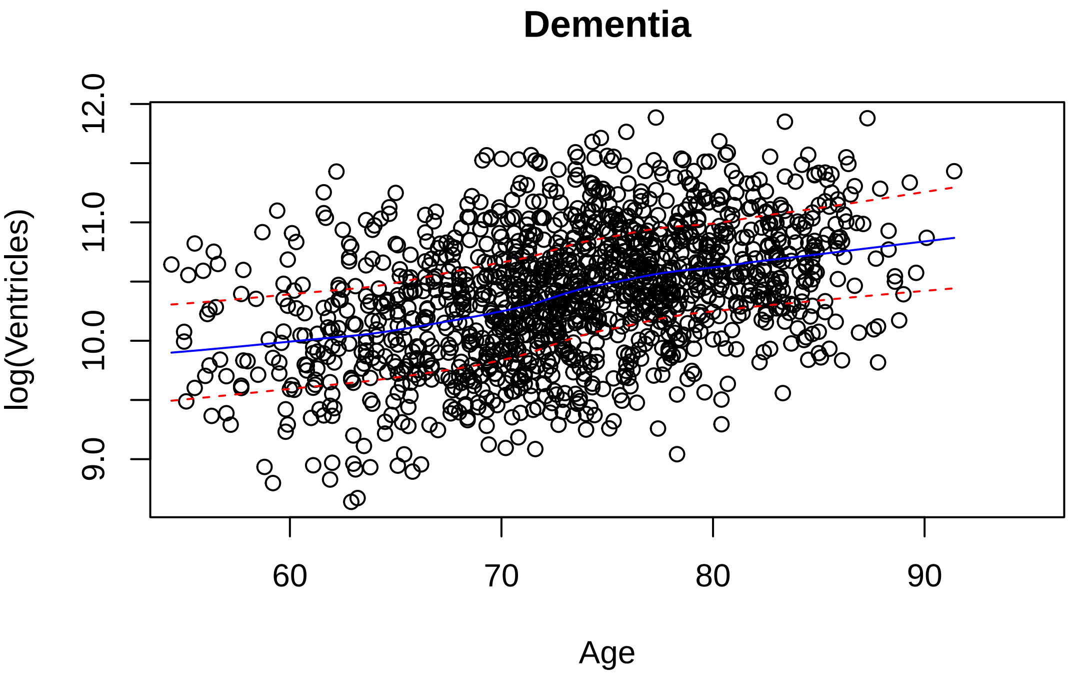
```
Ventricles.df=read.csv("Ventricles.csv")
plot(Ventricles~Age,main="Ventricles Size versus Age",type="n",data=Ventricles.df)
text(Ventricles.df$Age, Ventricles.df$Ventricles, Ventricles.df$AD, cex=.8)
```



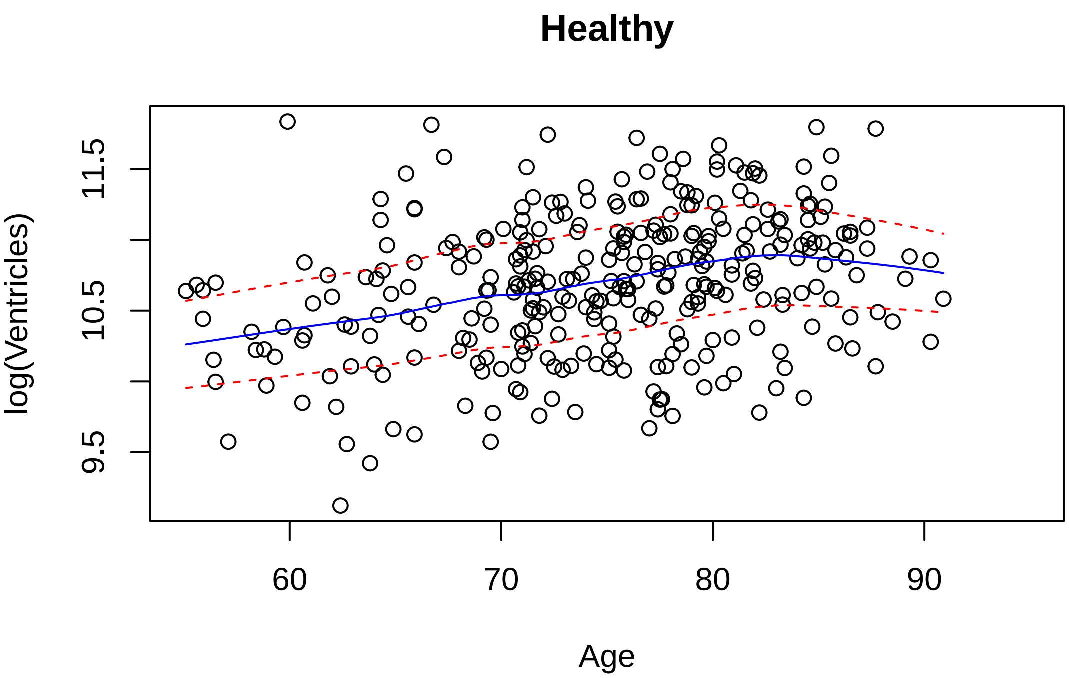
```
plot(log(Ventricles)~Age,main="log Ventricles Size versus Age",type="n",data=Ventricles.df)
text(Ventricles.df$Age, log(Ventricles.df$Ventricles), Ventricles.df$AD, cex=.8)
```



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



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



For plots, D is indicating the dementia individual, and H indicating the healthy individual.

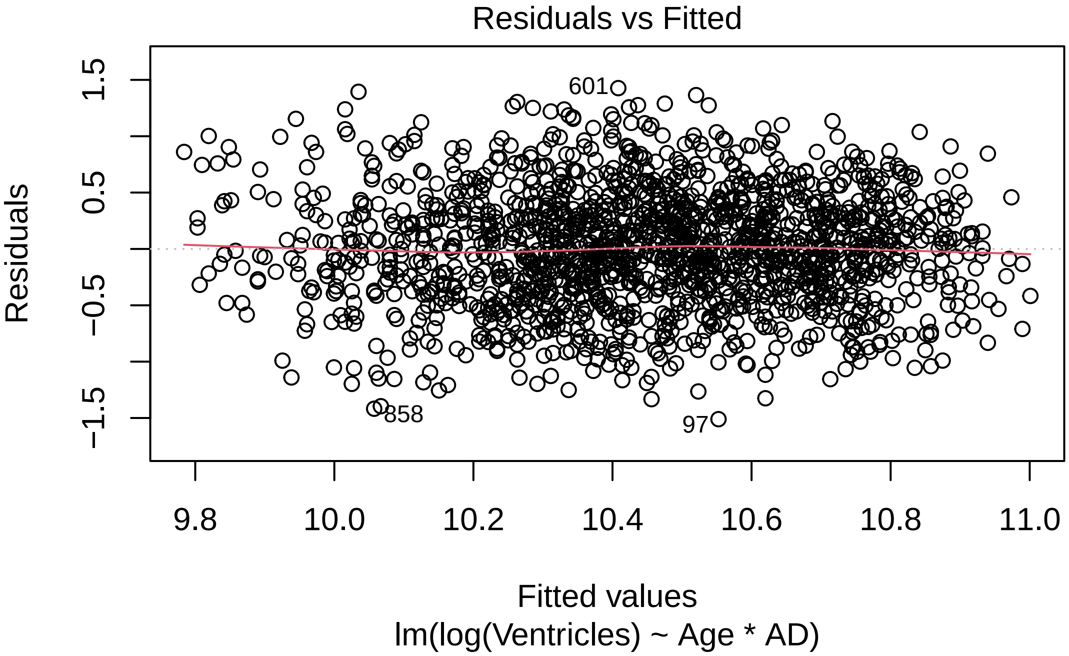
Look at the Ventricles Size and Age plot, notice that the relationship is weak. We do not know whether the relationship exists or not. To solve our problems, we logging the response, trying to convert a relationship to additive one. We can see the relationship is quite clear than previous and need trendscatter plot to make more details.

The both trendscatter plots seem to suggest that there is a positive relationship between Ventricles Size and Age on both Healthy and Dementia, the relationship is not actually strong, therefore we do not know whether the relationship is actually exists or not. In other words, we do not know whether it is statistically significant or not. To address this question we need to construct model.

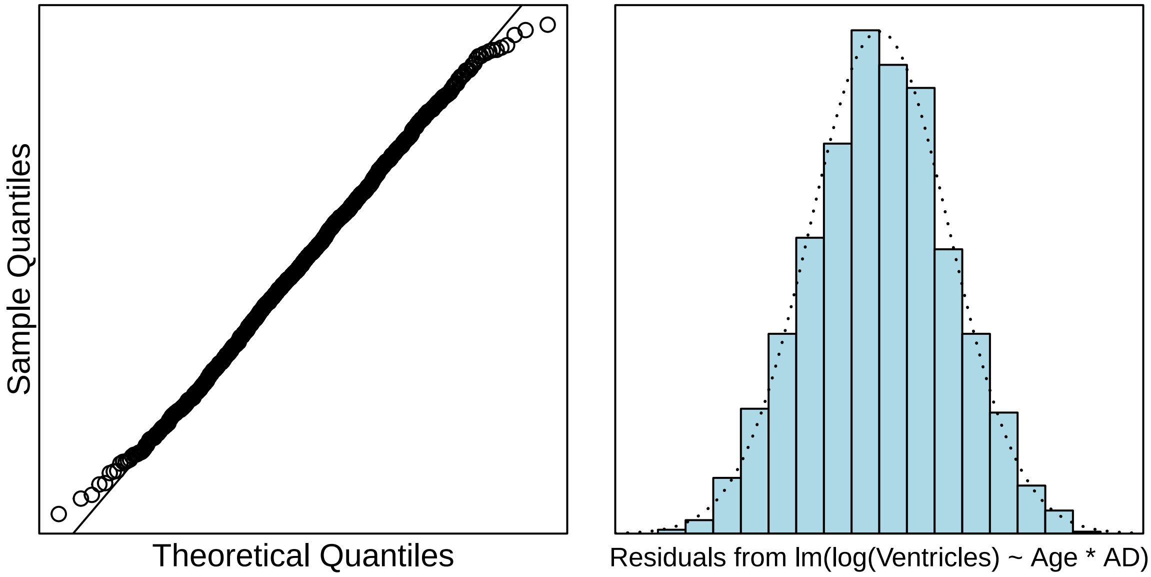
Separately for the Healthy plot, the red dash line indicating the range which actually the median of Ventricles given the Age. we can surely the relationship looks like some curvature, that might because the age is quite big and the variance becomes smaller. These trendscatter plots are useful.

The relationship between Ventricles Size and Age are not exactly the same for Healthy and dementia individuals.

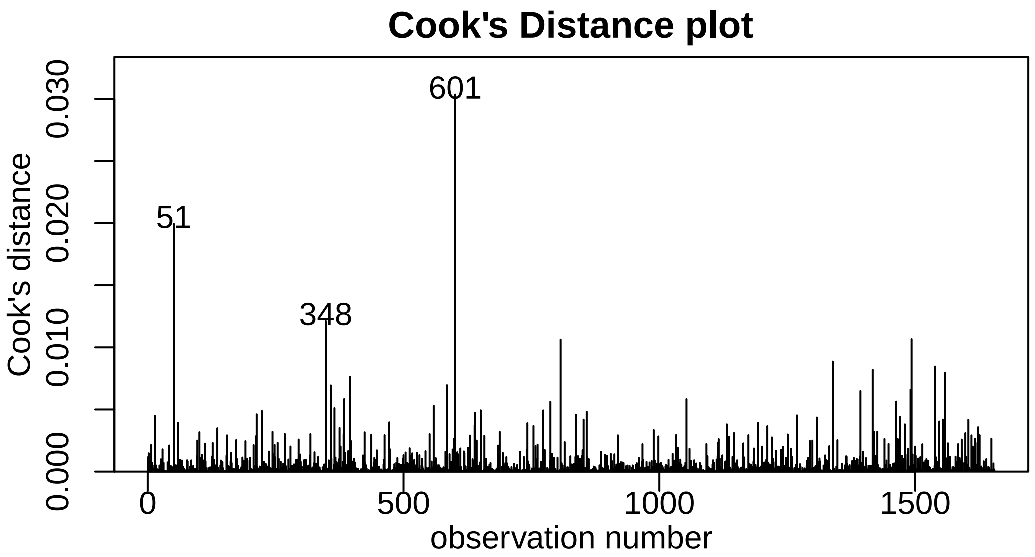
```
Ventriclesfit1=lm(log(Ventricles)~Age*AD,data=Ventricles.df)
plot(Ventriclesfit1,which=1)
```



```
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
```

Summary shows,

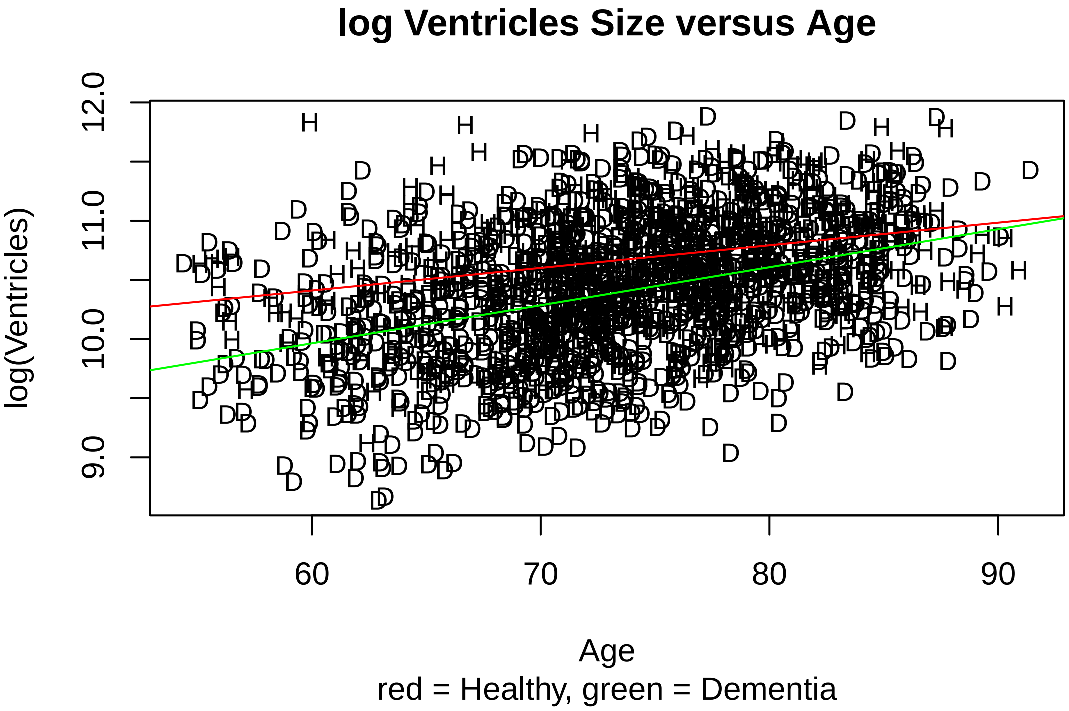
Age is significant, for people who is dementia, age explains the ventricles.

ADH seems to suggest the mean of ventricles of healthy individuals is slightly bigger than dementia individuals.

Age:ADH term suggest the slope for who is healthy is slightly smaller. and our model confirms that beta3 is negative and is significant at 5% level(p-value<0.05), means if the individual is healthy, and every 1 unit(each) additional age will increase expected ventricles.

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 variability 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 of term Age:ADH is estimating a(an) further(additional) number of slope for the healthy individuals.

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.

The R-output given that every 1% increase in the Age results in median size of ventricles between 85.6% and 528.6%. But the difference in size of ventricles between healthy people and those exhibiting dementia symptoms in general is decreasing actually, should be close to 0(zero) as the age increases, so we cannot answer this from the R-output given.

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

we estimate that each additional years aging on healthy people will increase the median size of ventricles by 1.20% to 2.66%.

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

we estimate that each additional years aging on people exhibiting dementia symptoms will increase the median size of ventricles by 2.87% to 3.67%.

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

The plot where green line is essentially dementia individuals, and red line is for healthy individuals. Immediately we notice the intercept for healthy is actually bigger than dementia, but the slope for healthy individuals is smaller. This is a visual confirmation of our intuition that the healthy individuals gets slower changes of ventricles as the age increases. And visually confirm that both two categories(dementia and healthy) will get the same value of log(ventricles) as the age increases to a point, that might be the age is quite old and notice that our model only explained 19% of the variability in the data.