---
title: “STATS 201 Assignment 2”
author: ’My name:Wang Yingpai
My ID number: 2019210179’
date: ‘Due Date: 2021/11/7’
output:
pdf_document: default
word_document: default
html_document:
fig_caption: yes
number_sections: yes

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

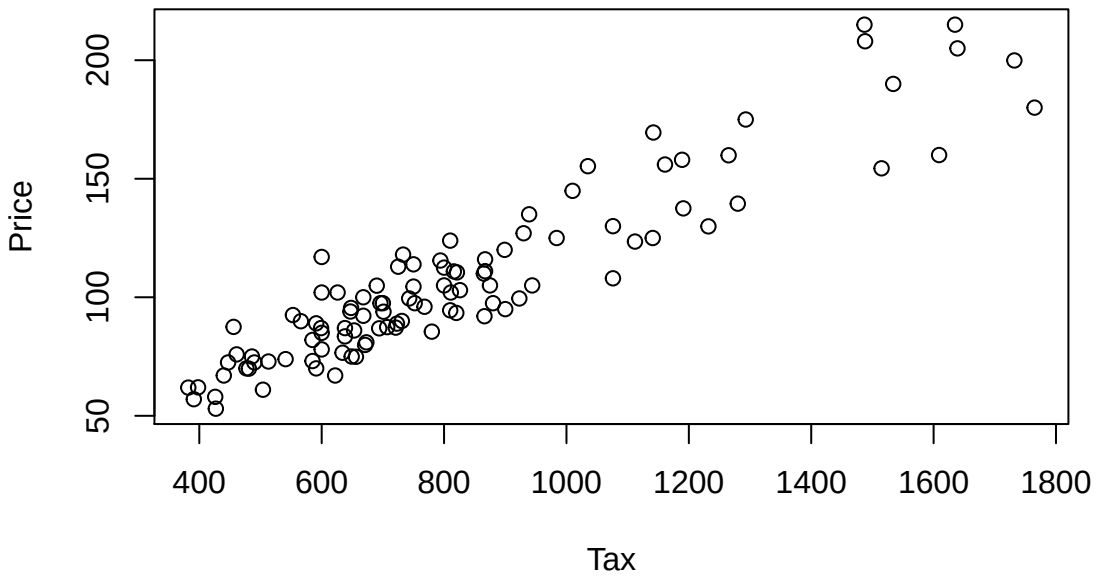
Question 1

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%.

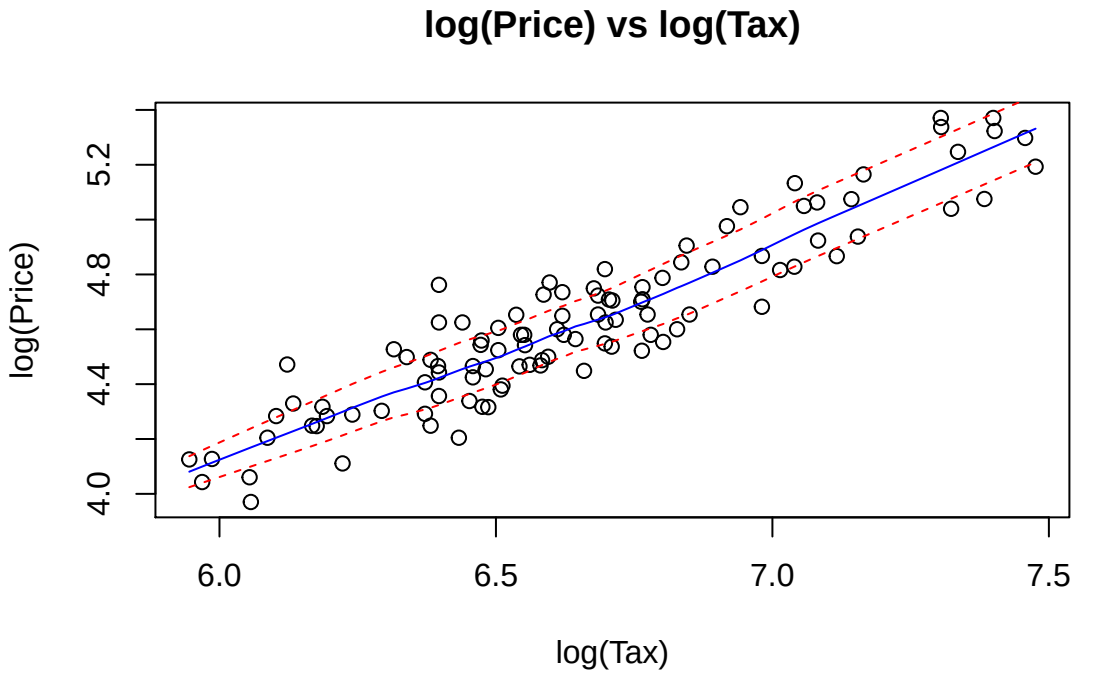
Read in and inspect the data:

```
hometax.df=read.csv("hometax.csv")
plot(Price~Tax,data=hometax.df)
```



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


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



We plot
a original data(x-axis=Tax, y-axis=Price) and make trend scatter to our original data and log-log(power)

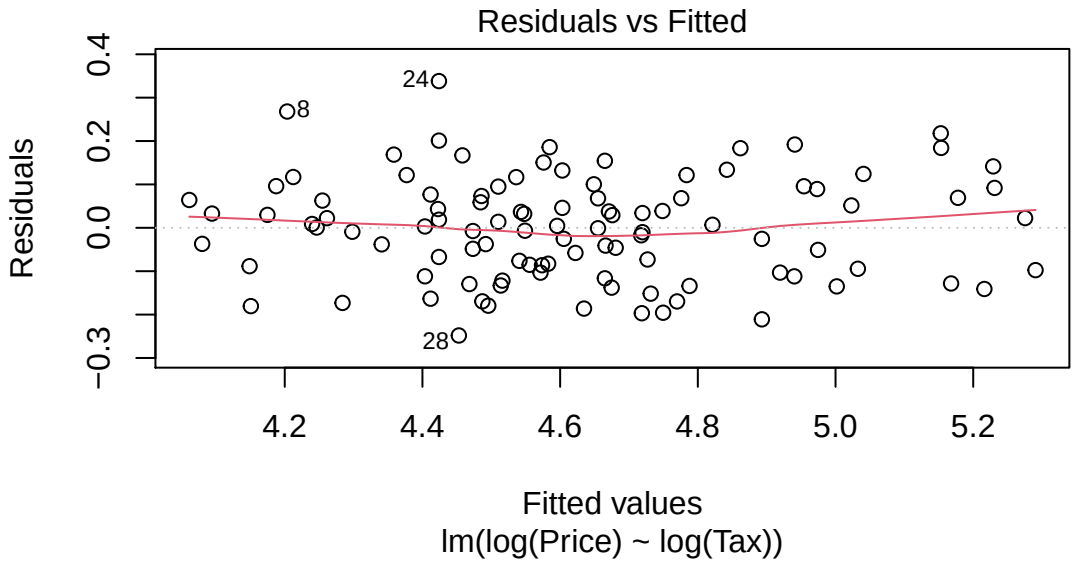
data. The first scatter plot suggest that it is a positive non-linear relation.And after make log-log to our data,it suggests a positive linear relation between log(Tax) and log(Price).

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

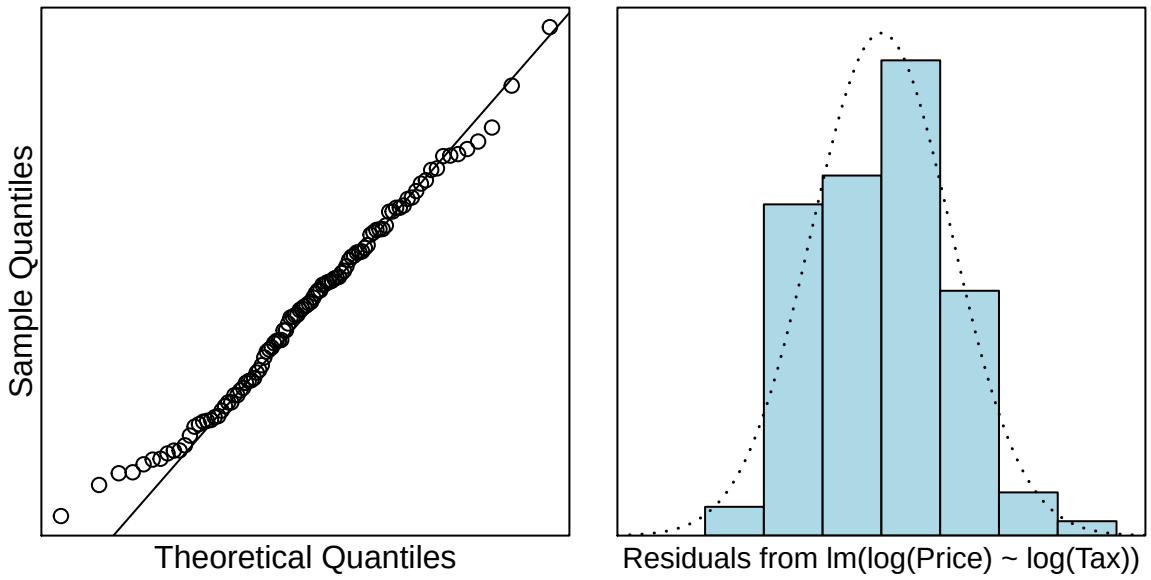
From the original plot,Tax on the x-axis and Price on the y-axis.It is non-linear relationship between Price and Tax. the data focus on the left, the data in right are decentralized(scattered).As the Tax change, the Price change in shape as grow a little larger(shape as if a little curve), so it's better to use a log-log(power) model.After make log-log to our data,the scatter plot suggest us it's a positive linear relation between log(Tax) and log(Price).

Fit model and check assumptions.

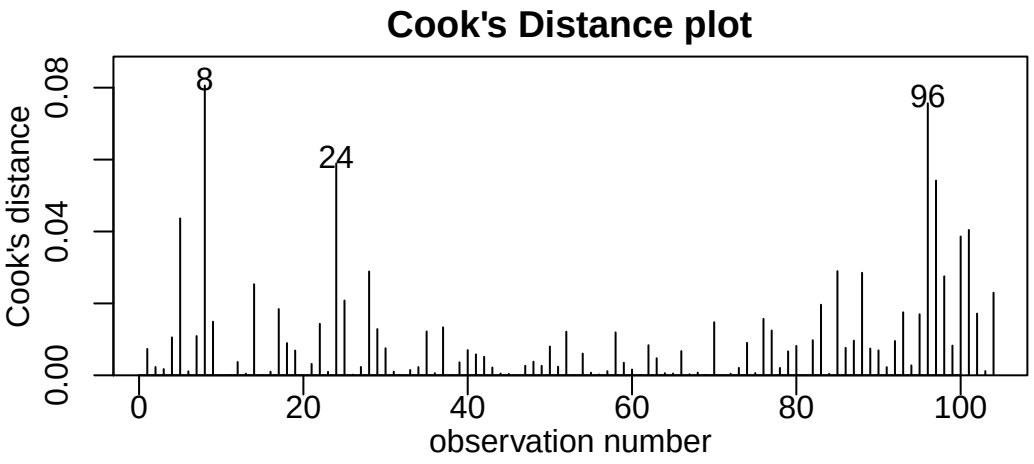
```
#fit a linear model with log-log(power) model to our data.
hometax.fit2 = lm(log(Price)~log(Tax),data=hometax.df)
plot(hometax.fit2,which = 1)
```




```
normcheck(hometax.fit2)
```




```
cooks20x(hometax.fit2)
```




```
summary(hometax.fit2)
```

```
##
## 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(hometax.fit2)
```

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

```
coef(hometax.fit2)[2] #for 1% Tax increase, how the Price change?(median value)
```

```
## log(Tax)
## 0.803111
```

```
(confint(hometax.fit2)[2,]) #for 1% Tax increase, how the Price change?(CI)
```

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

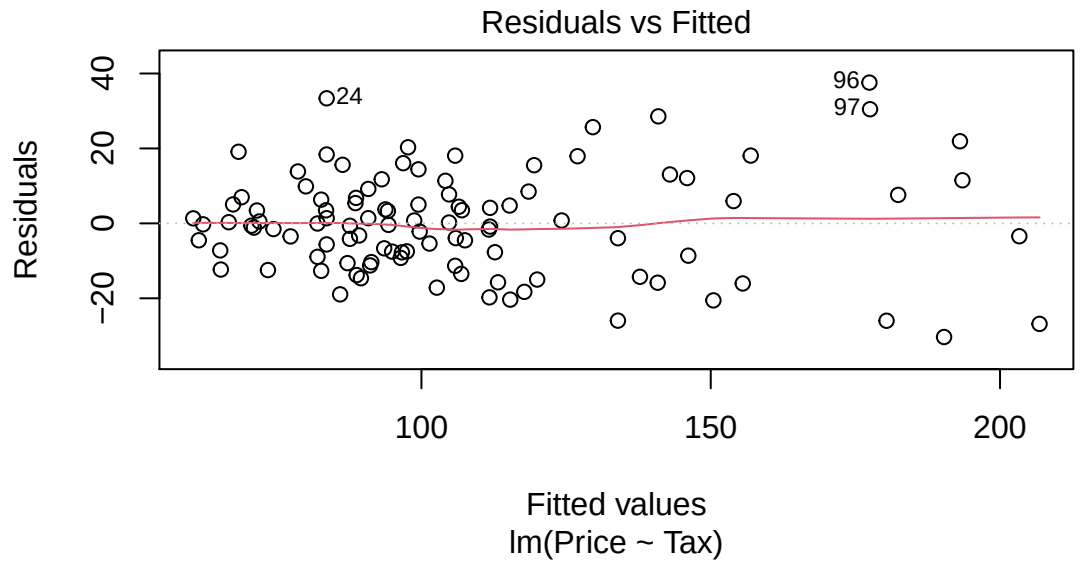
```
# 1%Tax
1.01^(coef(hometax.fit2)[2])
```

```
## log(Tax)
## 1.008023
```

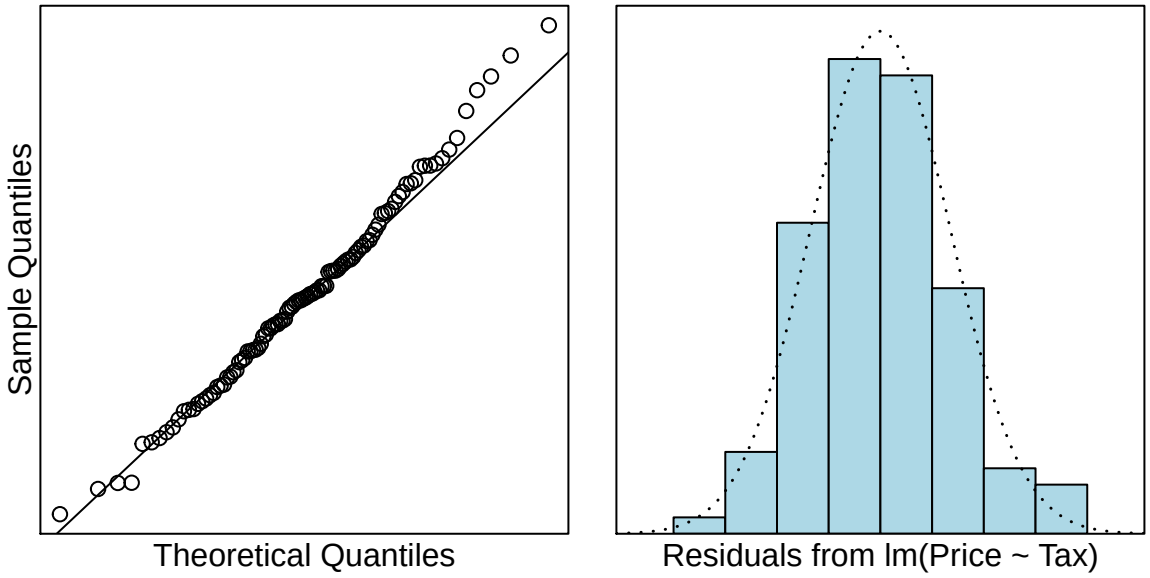
```
# 50%Tax
1.5^(coef(hometax.fit2)[2])
```

```
## log(Tax)
## 1.384908
```

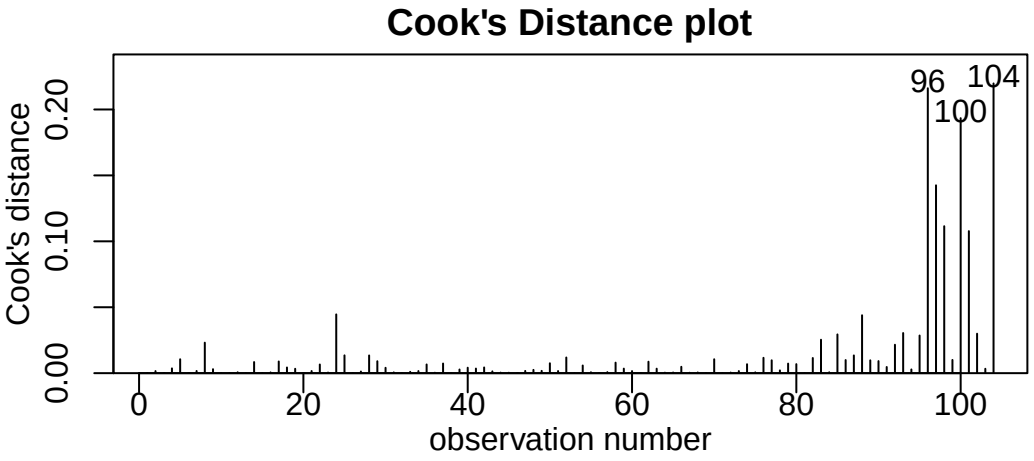
```
#(1.5^(coef(hometax.fit2)[2])-1.01^(coef(hometax.fit2)[2]))*100
#(1.5^(confint(hometax.fit2)[2,])-1.01^(confint(hometax.fit2)[2,]))*100

#fit a linear model with original data
hometax.fit1 = lm(Price~Tax,data=hometax.df)
plot(hometax.fit1,which = 1)
```




```
normcheck(hometax.fit1)
```




```
cooks20x(hometax.fit1)
```




```
summary(hometax.fit1)
```

```
##
## Call:
## lm(formula = Price ~ Tax, data = hometax.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.331  -9.005  -0.330   7.633  37.572
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.165675   3.673683   5.489 2.96e-07 ***
## Tax          0.105758   0.004155  25.451  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.57 on 102 degrees of freedom
## Multiple R-squared:  0.864,  Adjusted R-squared:  0.8626
## F-statistic: 647.8 on 1 and 102 DF,  p-value: < 2.2e-16
```

```
confint(hometax.fit1)
```

```
##                   2.5 %     97.5 %
## (Intercept) 12.87894219 27.4524084
## Tax          0.09751603  0.1140003
```

Methods and assumption checks

The scatter plot suggests that between log(Price) and log(Tax) is a linear trend on the whole. We have a random sample of 104 houses sold in Albuquerque, we assume the data are come from an independent and representative sample.The residual plot shows constant scatter and patternless residual, so no problems. The normaility check show slightly long tails on the left,but there no major problems. It's a good match

between the Sample Quantile and theoretical Quantile on the whole.So we may have a normal data.All the data's cooks' distance are under 0.2. So no strong influential points. We can trust our model. Our model is: $Log(Price_i) = \beta_0 \times \beta_1 \times Log(Tax_i) + \epsilon_i$ where $\epsilon_i \sim iid\ N(0, \sigma^2)$

Our model explain 85.64% of the total variation in the response variable,it's good for prediction. However, we use original data to fit a linear regression model, that model can explain 86.4% of the total variation.So our log-log(power) model is not good as our original linear regression model in our data.

Executive Summary

We want to find the relationship between Price and Tax.We fit a positive linear model to our log-log(power) data. We have strong evidence that the linear relationship exists between Log(Price) and Log(Tax) because the p-value is far less than 0.001. We are confident that 1% increase in the x = Tax value will result in a 0.74% to a 0.87% increase in the median value of y = Price. We estimate 1% tax bill will result in the median value of Price increase 0.8%,and 50% tax bill will result in the median value of Price increase 38%. Our model can explian 85% of the total variation,so it's useful for prediction.

We found a linear model to our original data is better than our linear model to our log-log(power) data.It can explain 86% of the total variation,but our model can explain 85% of the total variation. It's almost the same.So our model are useful for prediction.

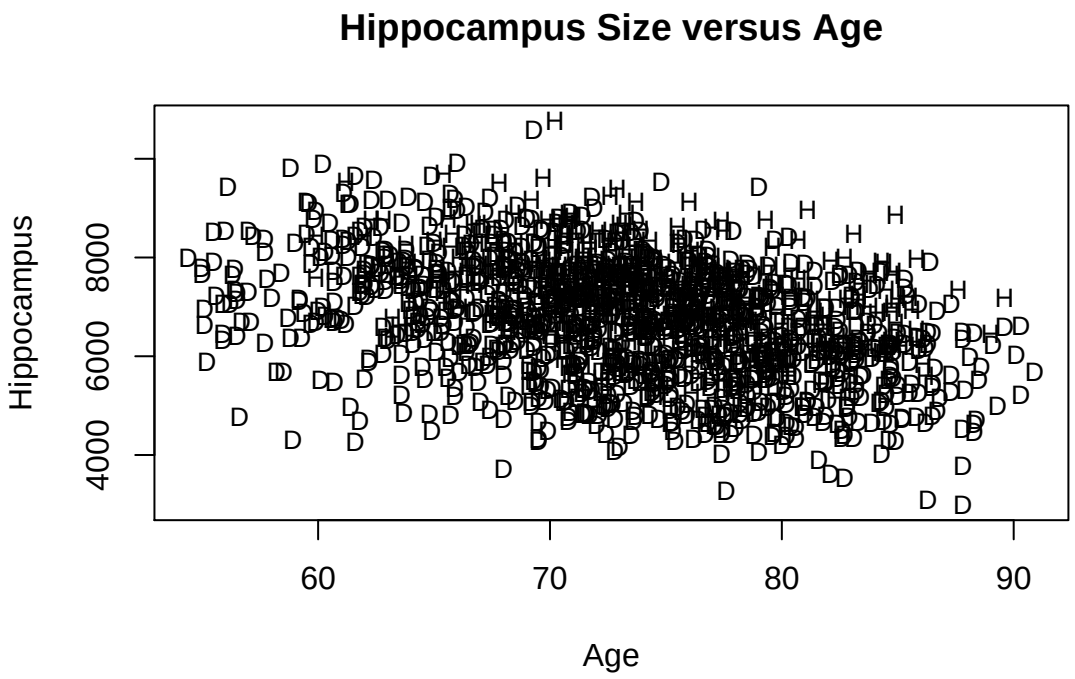
Question 2

Question of interest/goal of the study

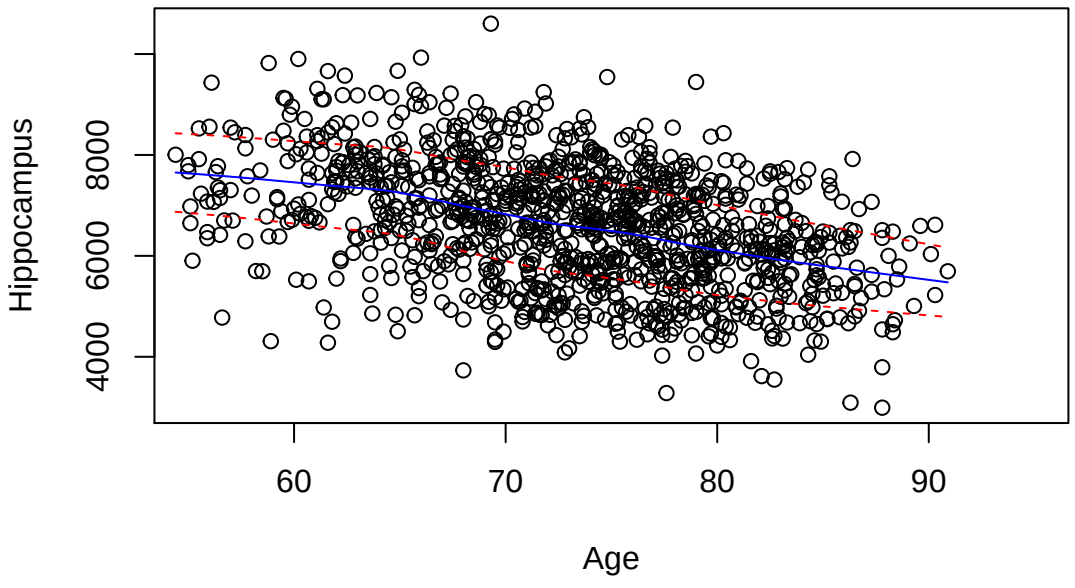
We want to explore the relationship between hippocampus size and age. In particular, we are interested in whether the relationship differs between healthy individuals and individuals with dementia related symptoms.

Read in and inspect the data:

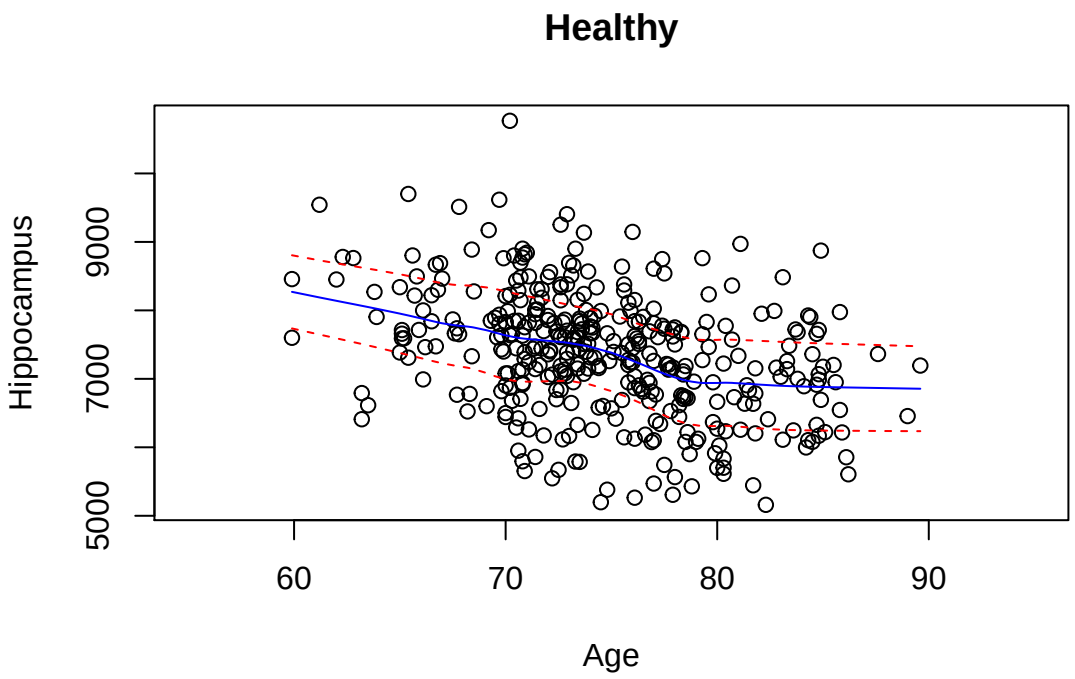
```
Hippocampus.df<-read.csv("Hippocampus.csv")
plot(Hippocampus~Age,main="Hippocampus Size versus Age",type="n",data=Hippocampus.df)
text(Hippocampus.df$Age, Hippocampus.df$Hippocampus, Hippocampus.df$AD, cex=.8)
```




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



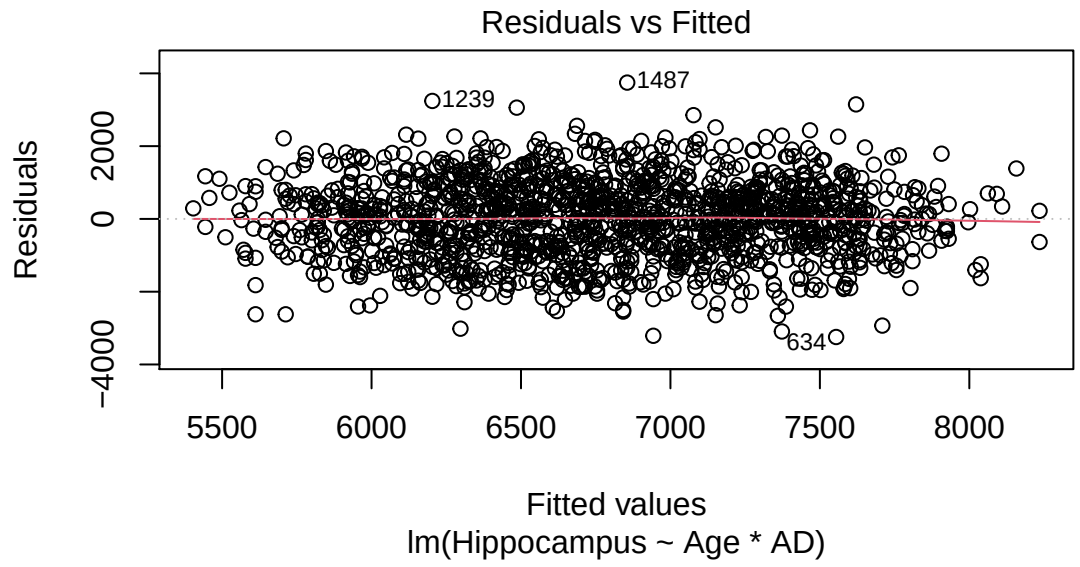

```
trendscatter(Hippocampus~Age,data=Hippocampus.df[Hippocampus.df$AD=="H",],xlim=c(55,95),main="Healthy")
```



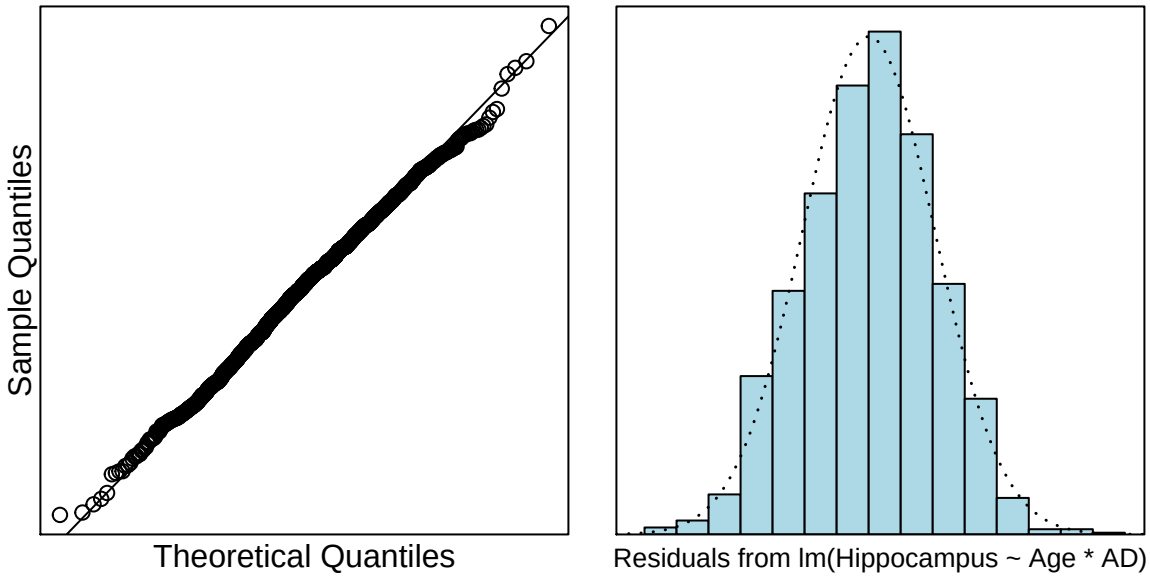
We have constant variable and categorical variable in our data.So we need to discuss it in different situation. The x-axis is Age and y-axis is Hippocampus size.The first plot,we can get that healthy people may have larger hippocampus size than demantia people on the whole.The second plot and the third plot suggests negative linear relationship between Age and Hippocampus size,but different lines.

Fit model and check assumptions.

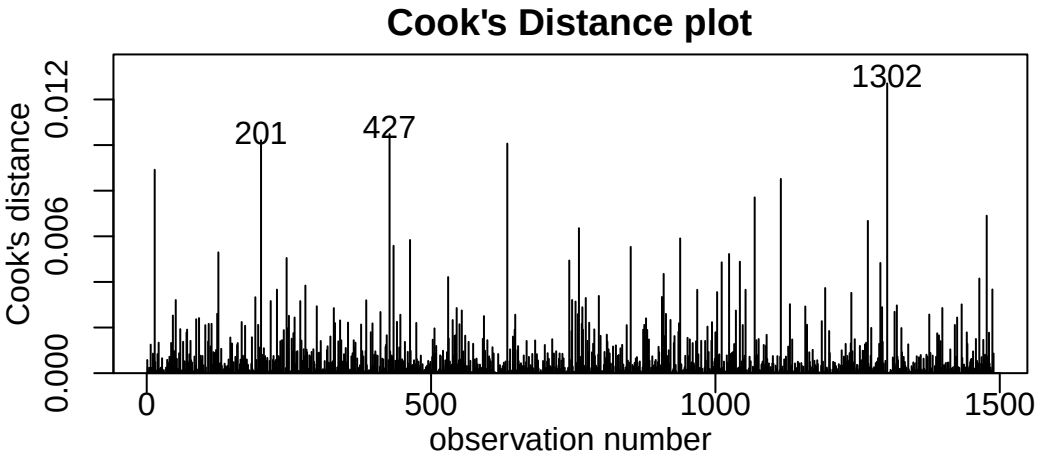
```
Hippocampus.fit = lm(Hippocampus~ Age * AD, data = Hippocampus.df)
plot(Hippocampus.fit,which=1)
```




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




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




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

```
##
## 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
```

```
coef(Hippocampus.fit)[2]
```

```
##       Age
## -67.21193
```

```
coef(Hippocampus.fit)[4]
```

```
##  Age:ADH
## 7.616612
```

```
confint(Hippocampus.fit)[2,]
```

```
##     2.5 %    97.5 %
## -75.31739 -59.10647
```

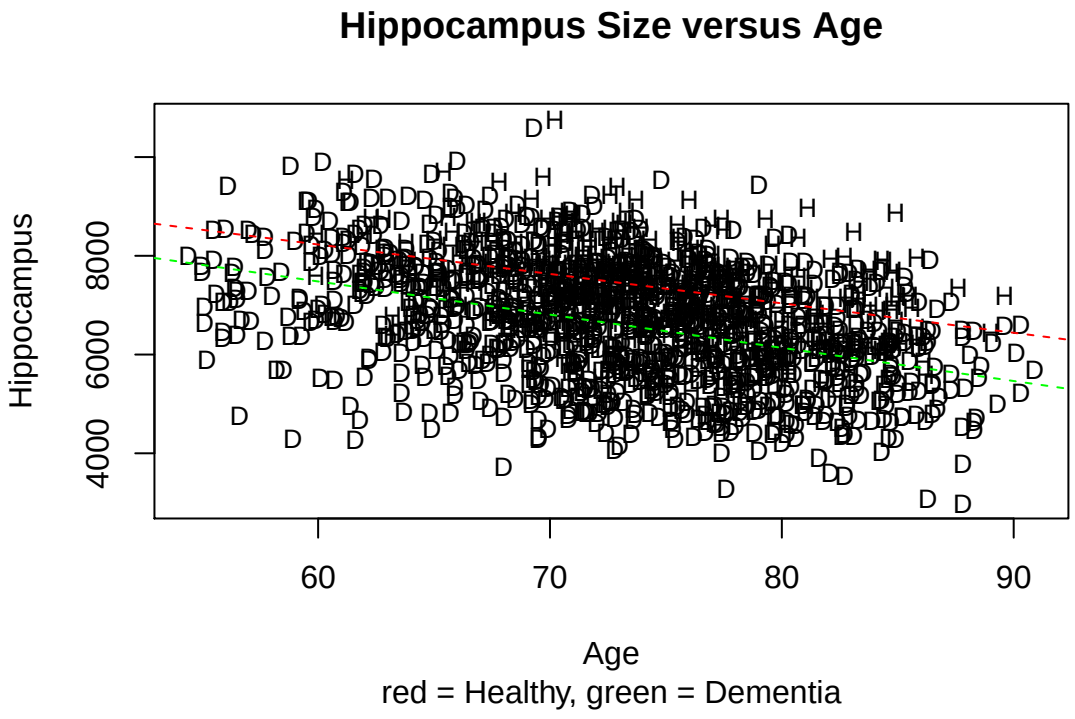
```
confint(Hippocampus.fit)[4,]
```

```
##     2.5 %    97.5 %
## -13.06983  28.30305
```

```
#Second way to fit model
#Hippocampus.df$D = as.numeric(Hippocampus.df$AD == "H")
#Hippocampus.df$HD = with(Hippocampus.df,{HD =  D * Age})
#Hippocampus.fit2 = lm(Hippocampus ~ Age + D + HD, data = Hippocampus.df)
#plot(Hippocampus.fit2,which = 1)
#normcheck(Hippocampus.fit2)
#cooks20x(Hippocampus.fit2)
#summary(Hippocampus.fit2)
```

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$AD, cex=.8)
b = Hippocampus.fit$coef
abline(b[1:2],lty=2,col = "green")
abline(b[1] + b[3], b[2] + b[4],lty=2,col = "red")
```



Methods and assumption checks

The scatter plot suggest there is linear trend between Age and Hippocampus_size ,but with different lines for healthy people and demantia people. so we fit a negative linear model with explanatory variable $x_1(Age)$,$x_2(AD)$,$x_3(AD * Age)$(interaction) to our data. We treat the data as if came from random samples of subjects,so we assume our data come from an independent and representative sample. We have fitted a negative linear model.Then we go through model check with our model.The residuals plot show constant scatter and patternless residual,so no problems.The Sample Quantitles matches the Theoretical Quantitles well,so we may have normal data.All the cooks' distance of points are under 0.012,so it doesn't exist any strong influenced points.All the assumptions are satisfied.

Our model is: $Hippocampus_i = \beta_0 + \beta_1 \times Age_i + AD \times \beta_3 + \beta_4 \times AD \times Age_i + \epsilon_i$ where $\epsilon_i \sim iid\ N(0, \sigma^2)$ where AD = 1 if the sample is healthy,AD = 0 if the sample is demantia.

Our model explains 23.28% of the total variation in the response variable,it is not good for prediction.

Executive Summary

We wish to explain the relationship between hippocampus size and age in healthy people and demantia people. So we fit a linear regression model to our data with explanatory variable Age,AD and interaction $AD \times Age$. We have strong evidence that there exists a linear relationship between Hippocampus size and Age because the p-value of is too small. However, we have weak evidence that the explanatory variable AD and $AD \times Age$ are Statistical significance.It means that we have weak evidence that there exists difference between healthy people and demantia people. We estimate that for each age increase, the excepted value of hippocampus size decrease are 67.2(We are confident it is between 59 and 75) in demantia people, there is a further change of +7.6(We estimate it is between -13 and 28)for healthy people. We can only explain 23% of the total variation by our linear model,so it is not reasonable to predict.

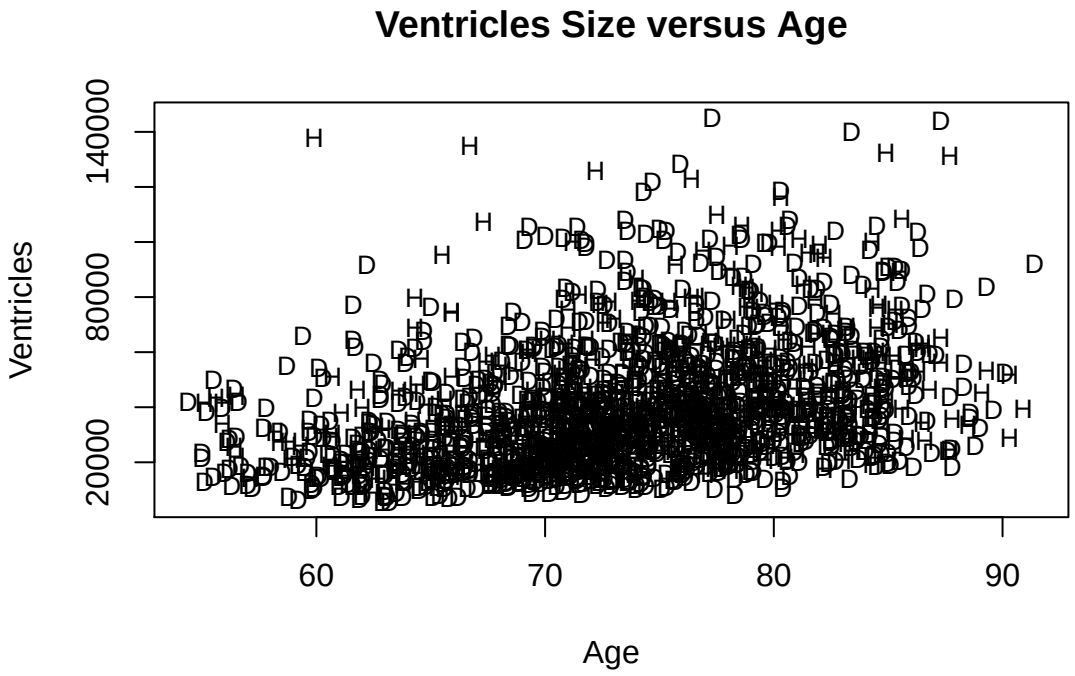
Question 3

Question of interest/goal of the study

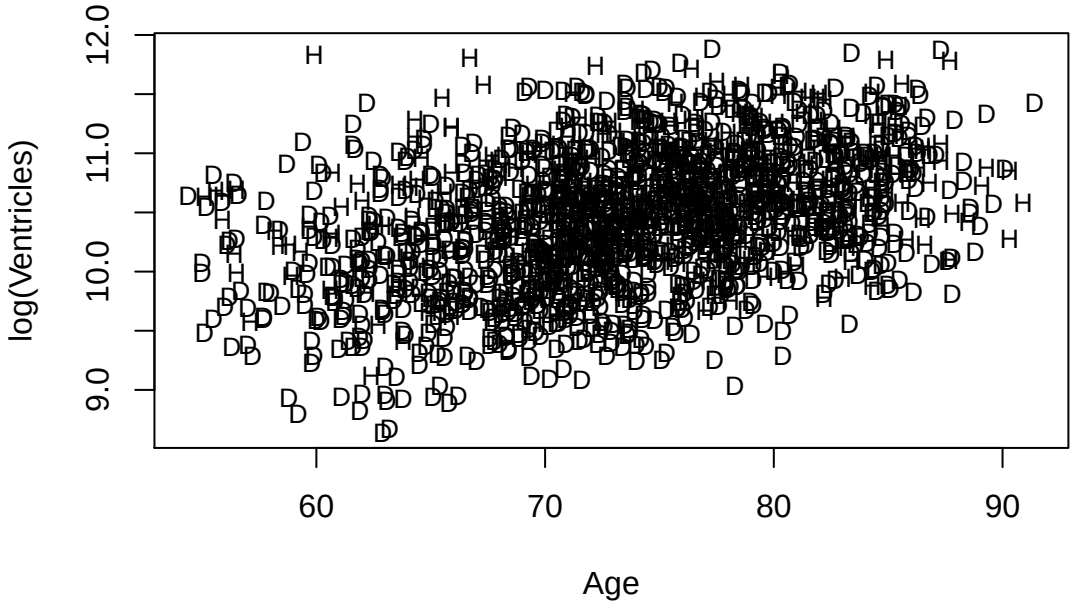
It is of interest to study the relationship between ventricles and age. In particular, we are interested in whether the relationship varies between healthy individuals and individuals with dementia related symptoms.

Read in and inspect the data:

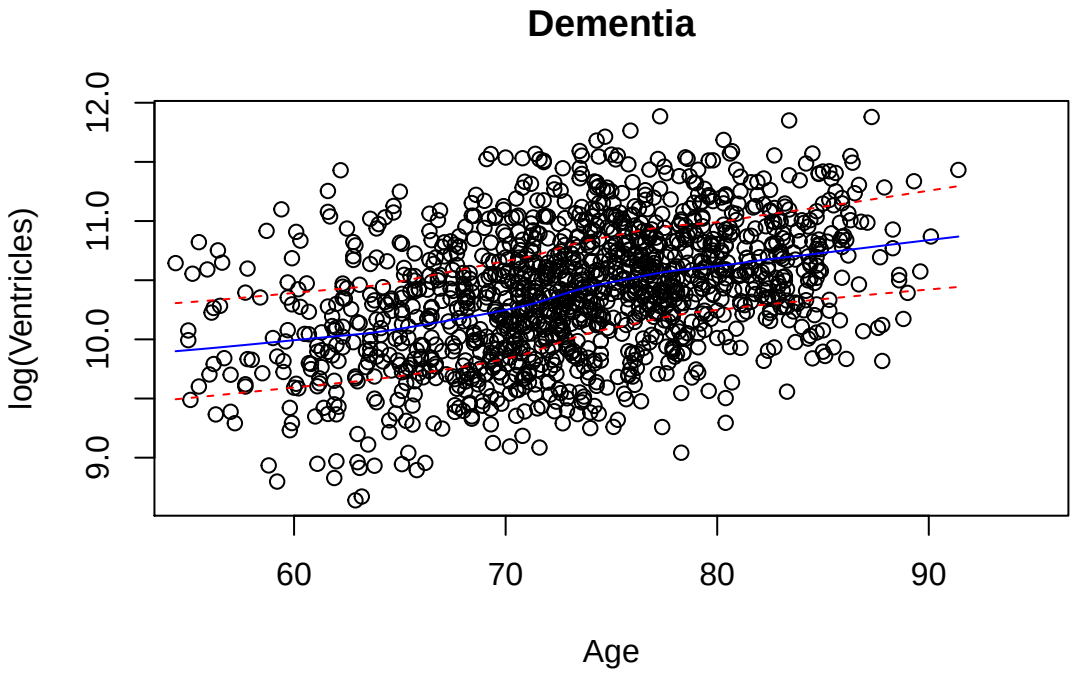
```
Ventricles.df=read.csv("Ventricles.csv")
plot(Ventricles~Age,main="Ventricles Size versus Age",type="n",data=Ventricles.df)
text(Ventricles.df$Age, Ventricles.df$Ventricles, Ventricles.df$AD, cex=.8)
```



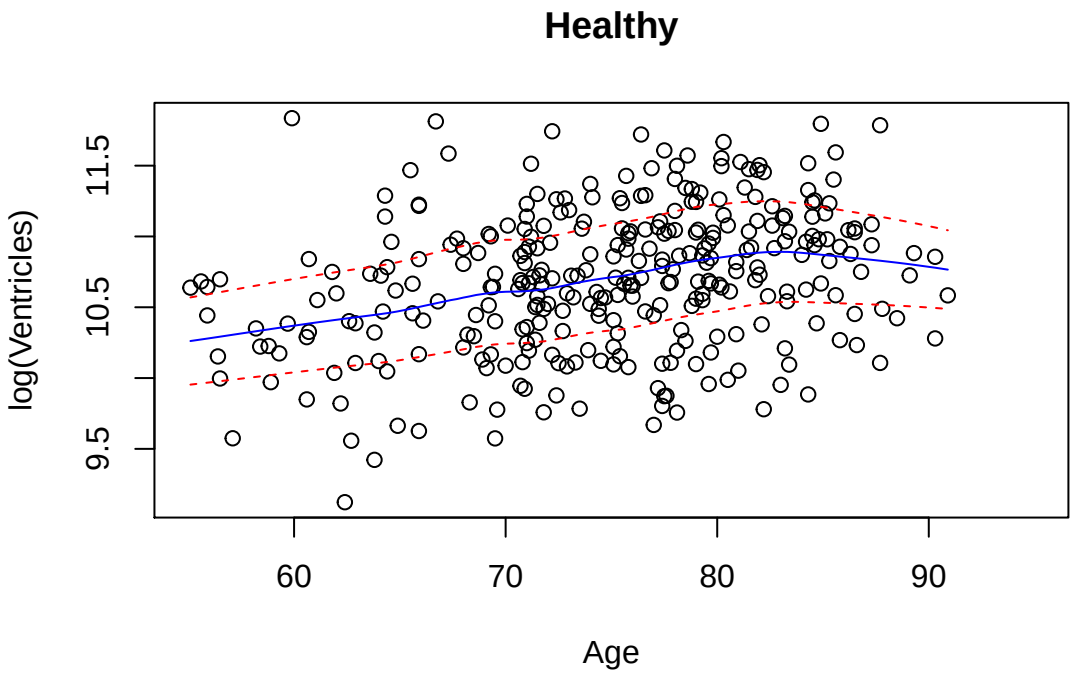

```
plot(log(Ventricles)~Age,main="log Ventricles Size versus Age",type="n",data=Ventricles.df)
text(Ventricles.df$Age, log(Ventricles.df$Ventricles), Ventricles.df$AD, cex=.8)
```



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



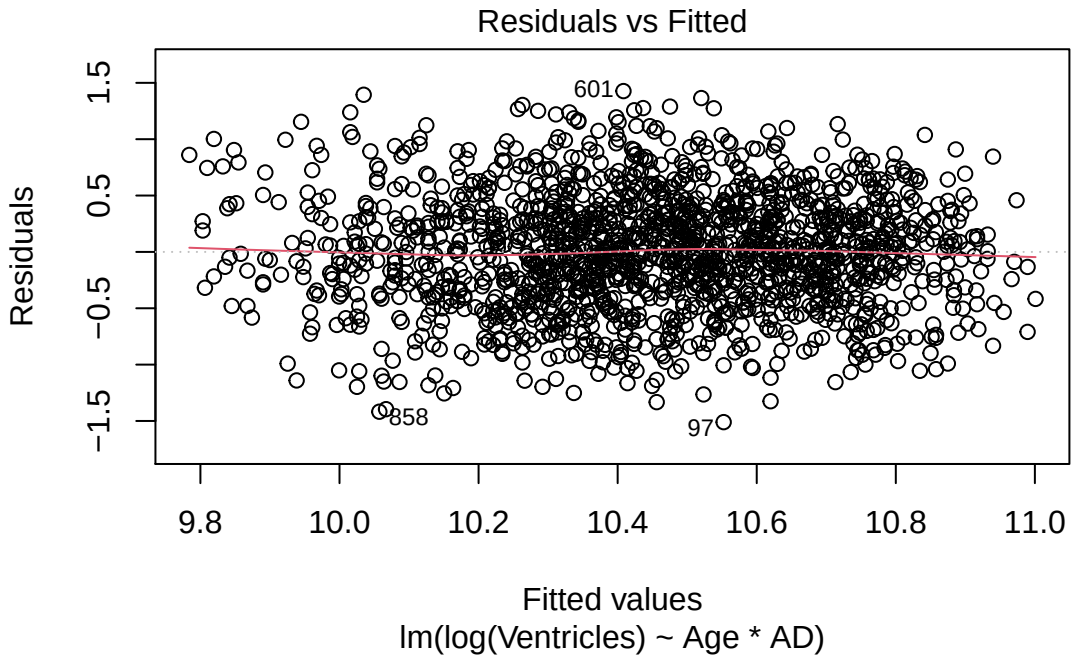

```
trendscatter(log(Ventricles)~Age,data=Ventricles.df[Ventricles.df$AD=="H",],xlim=c(55,95),main="Healthy
```



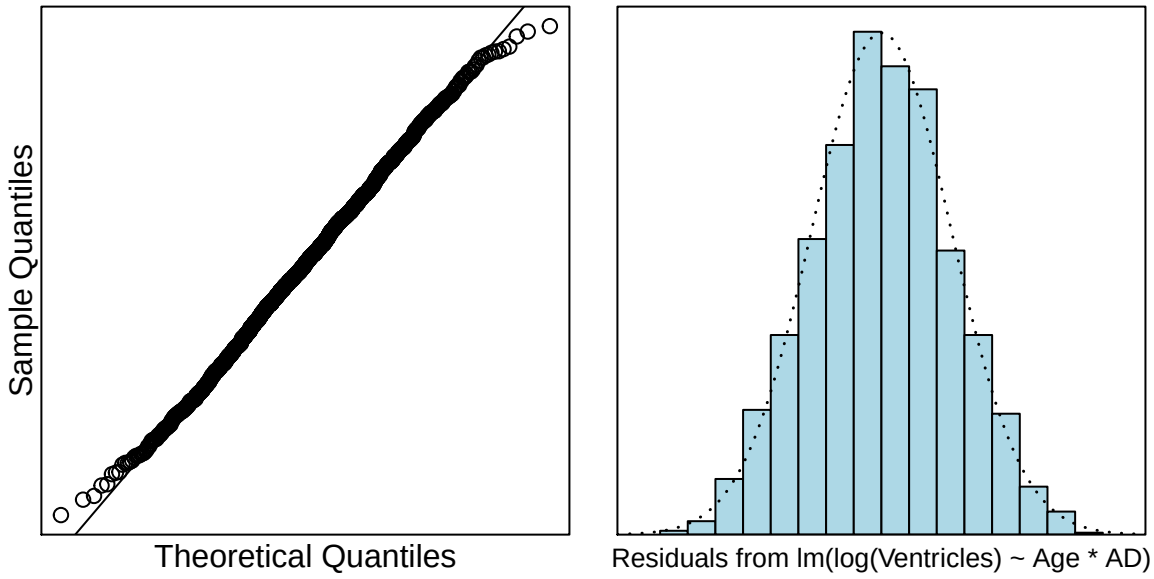
We have constant variable and categorical variable with our data.We want to find the relation between Age and

Ventricles sizes with healthy people and dementia people. We have four plot. The first plot is our original plot(x-axis=Age,y=Ventricles),the data focus on the bottom.The second plot is log(Ventricles) and Age,it's more concentrated and uniform.The third scatter and the fourth scatter both suggests a positive linear relation between Age and log(Ventricles),but a little drop trend in healthy people who Age between 80 and 90.

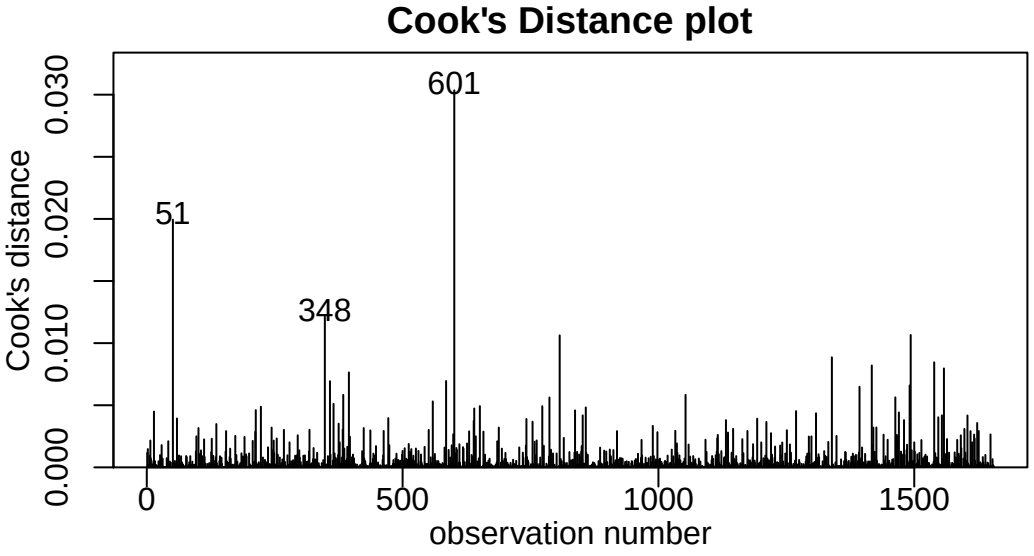
```
Ventriclesfit1=lm(log(Ventricles)~Age*AD,data=Ventricles.df)
plot(Ventriclesfit1,which=1)
```




```
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
```

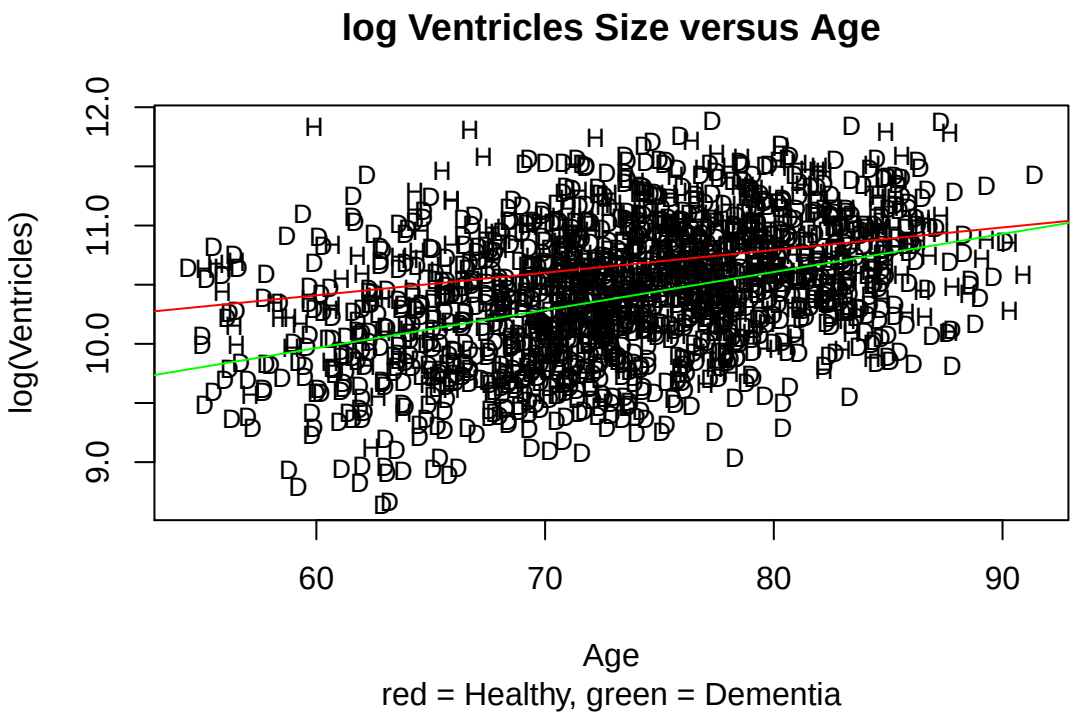
```
(exp(confint(Ventriclesfit2))-1)*100
```

```
##                     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
```

```
#((exp(confint(Ventriclesfit1))-1)*100)[2,]#baseline is unhealthy people
#((exp(confint(Ventriclesfit2))-1)*100)[2,]#baseline is healthy people
```

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.

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

Age:ADH is a interaction variable, it's coefficient means that healthy individual compared to dementia individual,for each age increase, the further median excepted value of ventricles' size change(positive means increase and negative means decrease).

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.

In general, healthy people's ventricles sizes is bigger than dementia people.However, with Age increase, the difference value are decreasing.

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

We estimate that for every additional years, the median value of size of ventricles on healthy people will increase between 1.2% and 2.66%

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

We estimate that for every additional years, the median value of size of ventricles on people exhibiting dementia symptoms will increase between 2.86% and 3.66%.

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

Looking at the plot,x-axis(x is between 50 to 100,so our sample are old people) is Age and y-axis is log(Ventricles size).On healthy people and demantia people are both positive linear relationship between Age and log(Ventricles size).So the relation between age and ventricles size is also positive.We know that the slope of healthy line are smaller than demantia line.So for each age increase, the healthy people's ventricles size increase slower than dementia people.With Age increasing,the difference between healthy people and dementia people are decreasing. In general, the median value of healthy people's ventricles size always bigger than dementia people.