STATS 201 Assignment 2

LIYuzhuo-2019210146

Due Date: 07/11/2021

## Loading required package: s20x

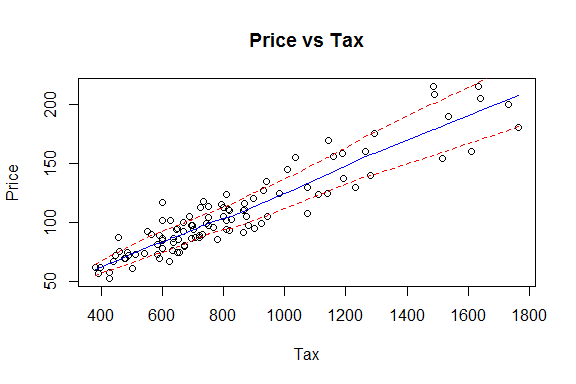
# Question 1

## Question of interest/goal of the study

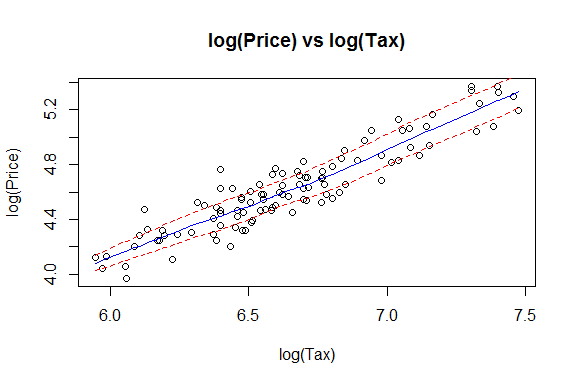
We want to build a model to explain the sale price of houses using their annual city tax bill (similar idea to rates in New Zealand) for houses in Albuquerque, New Mexico. In particular, we are interested in estimating the effect on sales price for houses which differ in city tax bills by 1% and 50%.

# 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)



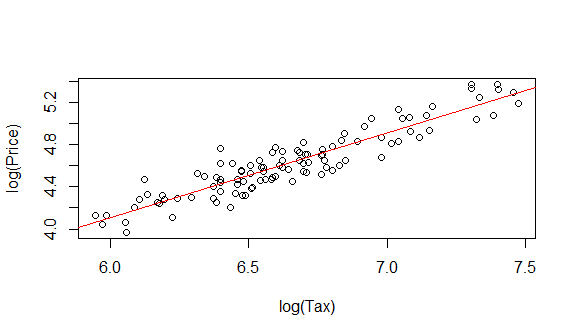
The trendscatter plot of price vs tax on their linear scales shows that residuals are non-constant. The plot shows small variance when tax is low and much greater variance when tax is high. As for the trendscatter plot of the log(price) vs log(tax), shows costant scatter.

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

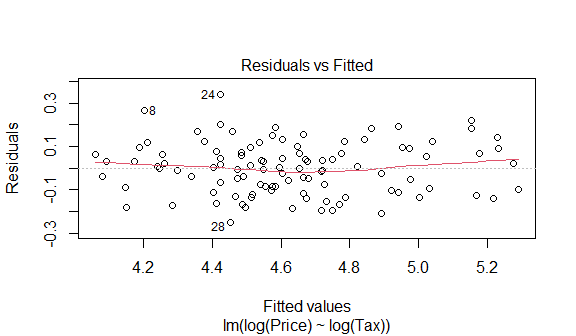
The log-log model is appropriate here because we are interested in estimating the effect on sales price for houses which differ in city tax bills by 1% and 50%. For the data we use is percentage change, it is known from common sense that a log-log (power) model is appropriate.

## Fit model and check assumptions.

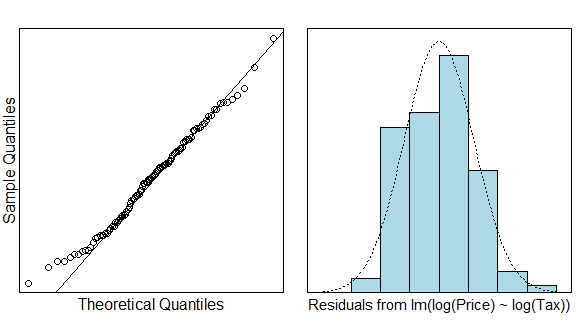
power.fit = lm(log(Price) ~ log(Tax),hometax.df)  
plot(log(Price) ~ log(Tax), hometax.df)  
abline(coef(power.fit), col='red')



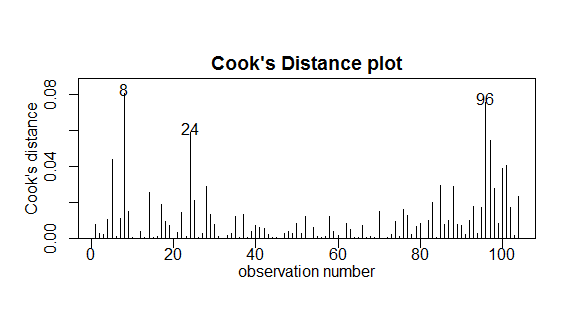
plot(power.fit, which =1)



normcheck(power.fit)



cooks20x(power.fit)



summary(power.fit)

##   
## 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

100\*(1.01 ^ confint(power.fit)[2,]-1)

## 2.5 % 97.5 %   
## 0.7375524 0.8671362

100\*(1.50 ^ confint(power.fit)[2,]-1)

## 2.5 % 97.5 %   
## 34.91053 42.16602

## Methods and assumption checks

We fit a log-log model because we were interested in estimating the percentage change in response when there is a percentage change in the predictor.

The data was collected from a random sample of 104 houses sold in Alburquerque, New Mexico. So the data is independent, and representation of our population.

The residual plot showed fairly patternless residals and constant scatter. The QQ plot shows that price follows a normal distribution for most of the dataset. The histogram also shows the data follows normally distributed. From the Cook’s Distance, we can conclud that there isn’t influencial point in the data. Overall, all the model assumptions are satisfied

Our model is: where

Our model explains approximately 86% of the total variation in Price.

## Executive Summary

Our aim is to explain the sale price of houses using their annual city tax bill. The population of interest were houses in Albuquerque, New Mexico. No data points were excluded.

We have found strong evidence that a relationship between the annual city tax bill of a house in Alburquerque and its sale price exists . Sale price was seen to follow a power-law relationship with respect to tax bill.

We estimate that a 1% increases in annual city tax bill corresponds to an increase in median house sale price of between 0.74% and 0.87%. Similarly, we estimate that a 50% increase in annual city tax bill corresponds to an increase in median house sale price of between 34.91% and 42.17%.

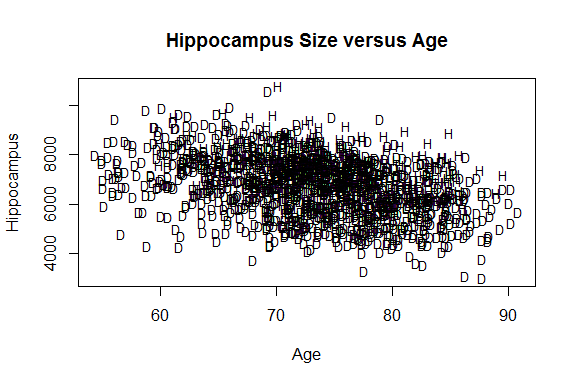
# Question 2

## Question of interest/goal of the study

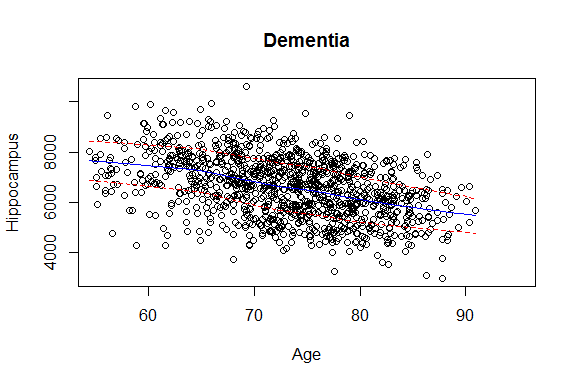
We want to explore the relationship between hippocampus size and age. In particular, we are interested in whether the relationship differs between healthy individuals and individuals with dementia related symptoms.

# Read in and inspect the data:

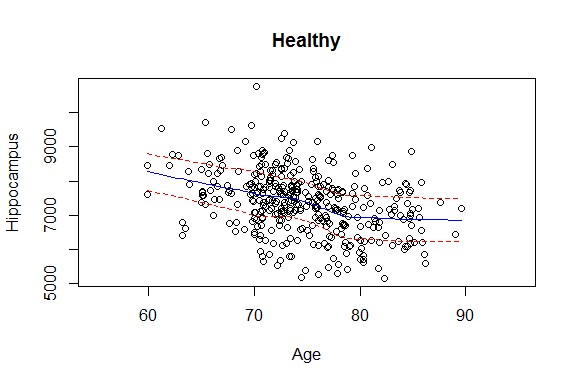
Hippocampus.df<-read.csv("Hippocampus.csv")  
plot(Hippocampus~Age,main="Hippocampus Size versus Age",type="n",data=Hippocampus.df)  
text(Hippocampus.df$Age, Hippocampus.df$Hippocampus, Hippocampus.df$AD, cex=.8)



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")



There appears to be a slight negative linear trend between age and hippocampus volume in both healthy individuals and individuals with dementia, which looks obviously in individuals with dementia. Also the data have quite large variance.

## Fit model and check assumptions.

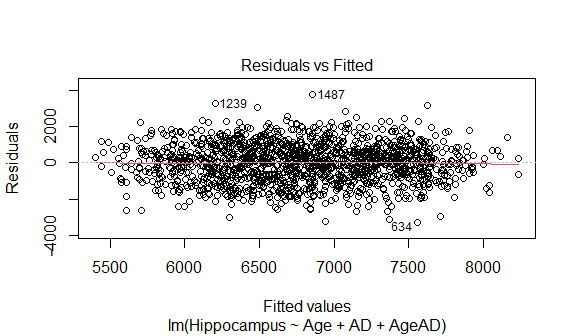
Hippocampus.df$AD = as.numeric(Hippocampus.df$AD == "H")  
Hippocampus.df$AgeAD = with(Hippocampus.df, {AgeAD = Age \* AD})  
fit <- lm(Hippocampus ~ Age + AD + AgeAD, Hippocampus.df)  
summary(fit)

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

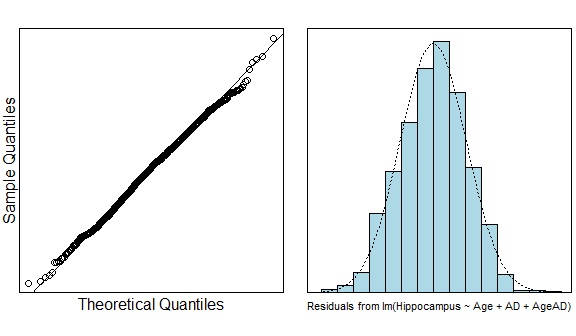
confint(fit)

## 2.5 % 97.5 %  
## (Intercept) 10917.36123 12108.97432  
## Age -75.31739 -59.10647  
## AD -1252.83698 1835.81096  
## AgeAD -13.06983 28.30305

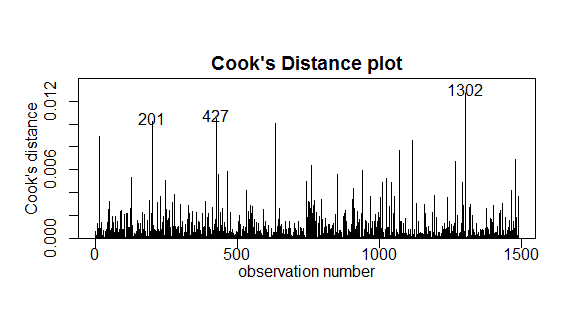
plot(fit, which = 1)



normcheck(fit)



cooks20x(fit)



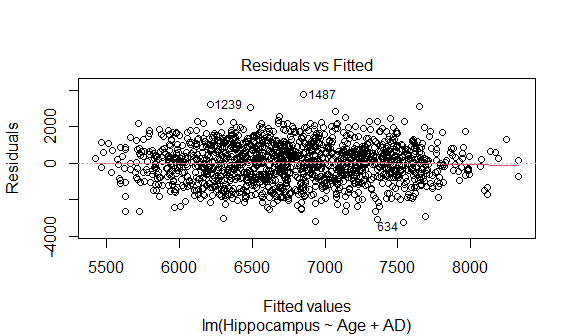
fit2 <- lm(Hippocampus ~ Age + AD, Hippocampus.df)  
summary(fit2)

##   
## 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 \*\*\*  
## AD 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

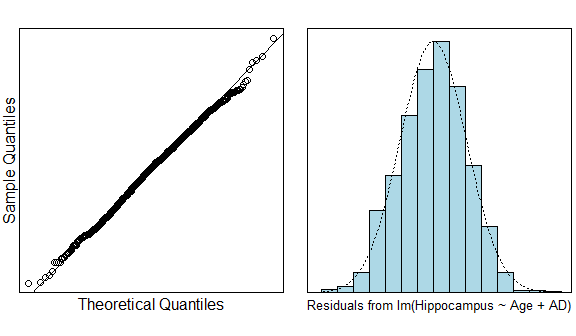
confint(fit2)

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

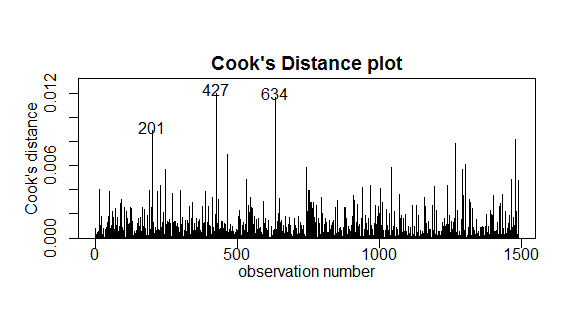
plot(fit2, which = 1)



normcheck(fit2)

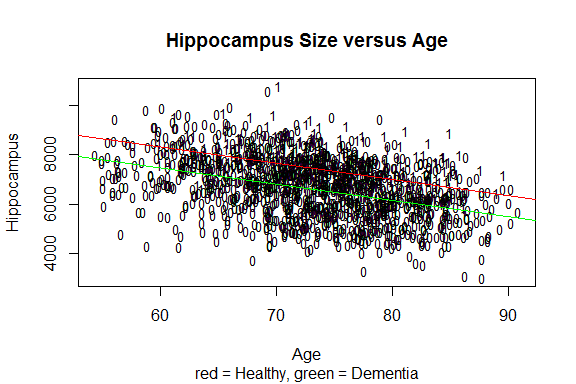


cooks20x(fit2)



## 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",data=Hippocampus.df)  
text(Hippocampus.df$Age, Hippocampus.df$Hippocampus, Hippocampus.df$AD, cex=.8)  
abline(fit2$coef[1],fit2$coef[2],col = "green")  
abline(fit2$coef[1]+fit2$coef[3],fit2$coef[2],col = "red")



## Methods and assumption checks

Since it appears to be a negative linear trend between age and hippocampus volume in both healthy individuals and individuals with dementia, and our dataset has one numeric predictor variable and a categorical predictor variable, the response variable is numeric. So we fit the data with a linear model with both continuous and factor variables. In the first model which has AGE \* AD, we find that the coefficient of AGE \* AD probably be zero, so we just drop this item in our model.

We have a random sample of individuals with and without dementia related symptoms between the ages of about 54 and 91, so that it is an independent and representative sample.

Our assumption checks all had good results. The residual plot showed constant and patternless scatter throughout the plot. The QQ plot and histogram show the dataset is very closely follows normally distributed. The Cook’s Distance plot shows there are no influencial points.

Our model is where and if the individual is healthy and if they have signs of dementia.

Our model explained 23% of the total variation in hippocampus size.

## Executive Summary

The purpose of this analysis was to study 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. No data points were excluded in this analysis.

We have found a linear relationship in hippocampus size and age . Our results show that hippocampus size and age follow the same relationship for healthy individuals and individuals with dementia related symptoms. However, healthy individuals have bigger hippocampus volume than those with dementia related symptoms on average.

We estimate that for every year any individual (healthy or otherwise) gets older their hippocampus will decrease on average by between 73.5 and 58.6 volume units. And we also find that the size of hippocampus of healthy individuals will be greater by 734 to 981 volume units than individuals with dementia related symptoms.

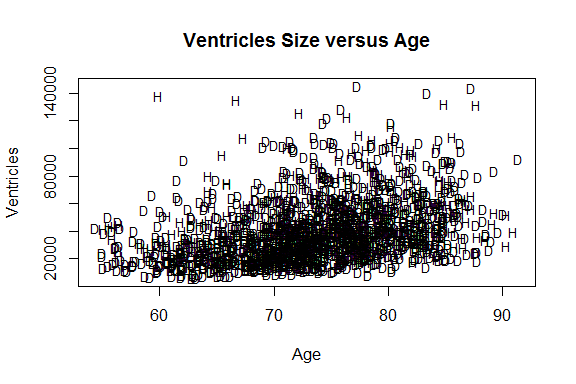
# Question 3

## Question of interest/goal of the study

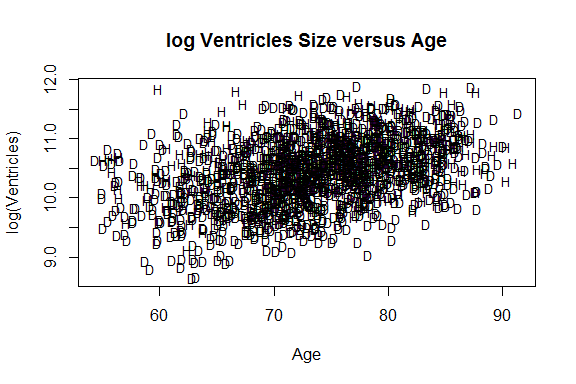
It is of interest to study the relationship between ventricles and age. In particular, we are interested in whether the relationship varies between healthy individuals and individuals with dementia related symptoms.

## Read in and inspect the data:

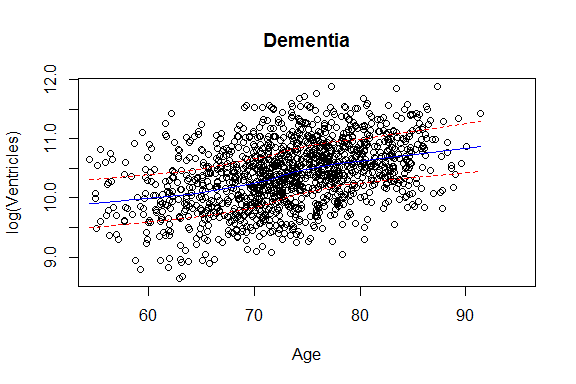
Ventricles.df=read.csv("Ventricles.csv")  
plot(Ventricles~Age,main="Ventricles Size versus Age",type="n",data=Ventricles.df)  
text(Ventricles.df$Age, Ventricles.df$Ventricles, Ventricles.df$AD, cex=.8)



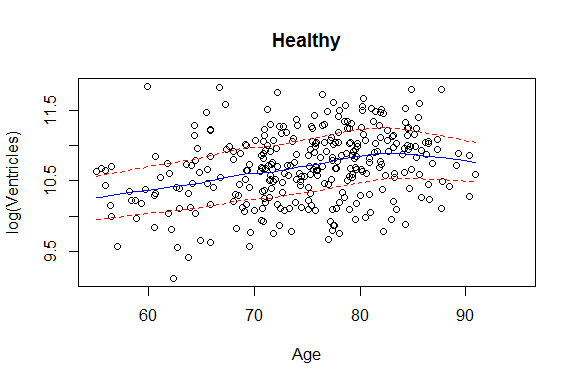
plot(log(Ventricles)~Age,main="log Ventricles Size versus Age",type="n",data=Ventricles.df)  
text(Ventricles.df$Age, log(Ventricles.df$Ventricles), Ventricles.df$AD, cex=.8)



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

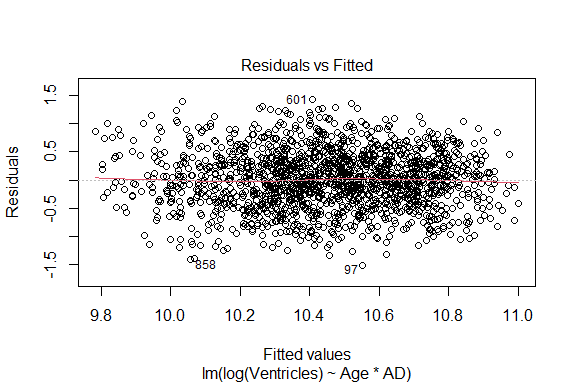


trendscatter(log(Ventricles)~Age,data=Ventricles.df[Ventricles.df$AD=="H",],xlim=c(55,95),main="Healthy")

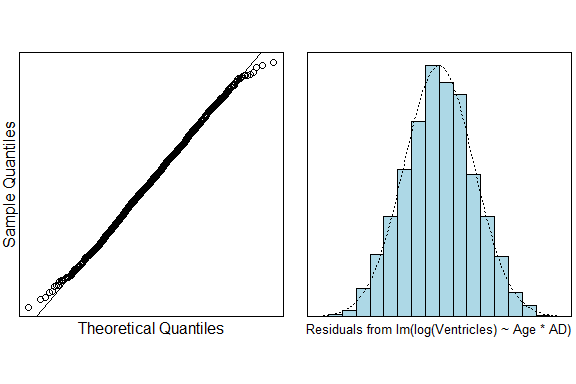


There isn’t clear trend was seen in the scatter plot of ventricular size versus age. After log-transformation of ventricular size, a linear trend in the data became apparent, and there appears to be a positive relationship between age and ventricle size for those with dementia related symptoms. For healthy individuals, there is not a simple positive trend as there is a sudden decrease in the mean of logged ventricles size past approximately 80 years.

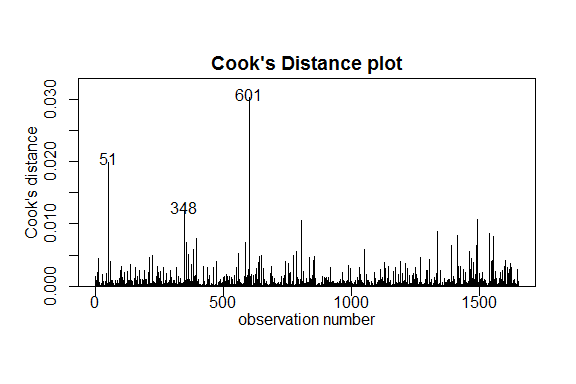
Ventriclesfit1=lm(log(Ventricles)~Age\*AD,data=Ventricles.df)  
plot(Ventriclesfit1,which=1)



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))

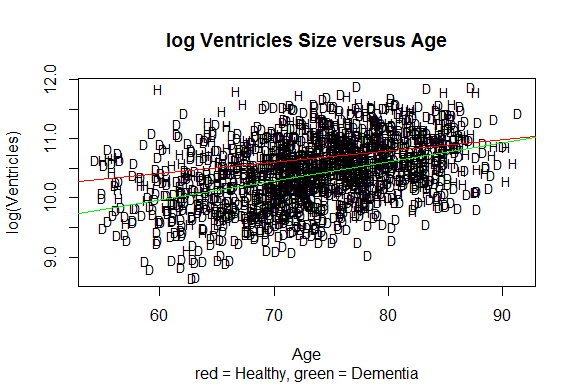
## 2.5 % 97.5 %  
## (Intercept) 6151.8258140 1.806983e+04  
## Age 1.0120271 1.026627e+00  
## ADflipD 0.1590953 5.388152e-01  
## Age:ADflipD 1.0049031 1.021405e+00

(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: where if the th subject is healthy and 0 if they have signs of dementia, and

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 estimated coefficient Age:ADH is estimating the amount of change in the increase in ventricular size in healthy individuals compared to dementia individuals for each year of age increase in individuals of the same age but different health status.

## 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 question from the R-output given because the healthy people and those exhibiting dementia symptoms follow a different relationship when using age to explain ventricles size. The difference in relationship we can concluded from the different slopes for the two groups, which shows they have different trend.

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

For each additional years aging on healthy people, the median size of ventricles increase by 1.2% to 2.7%.

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

For each additional years aging on people exhibiting dementia symptoms, the median size of ventricles increase by 1.2% to 2.7%.

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

At age before 90, the median ventricle size is larger in healthy individuals, but this difference becomes smaller as age becoming larger. The median ventricular size increases at a greater annual rate in people who develop dementia-related symptoms than in healthy individuals, and maybe beyond healthy individuals.