Ben Goodwin

DS6371

HW12

Question 1)

Part A)

Chart, scatter chart

Description automatically generated

The non-log transformed data appears to have much of the data concentrated around the origin.

Chart, scatter chart

Description automatically generated

The log-transformed data follows the fit line with much more accuracy than the non-transformed data.

Judging by a scatterplot alone it seems the data reasonably falls along a straight horizontal line and nearly passes through the origin, it looks linear.

Part B)

I)

Chart, scatter chart

Description automatically generated

The non-log transformed data appears to have much of the data concentrated around the origin.

Chart, scatter chart

Description automatically generated

The log-transformed data follows the fit line with much more accuracy than the non-transformed data.

II)

Chart, scatter chart

Description automatically generated

The residuals from the non-transformed data indicate violation of assumptions.

Chart, scatter chart

Description automatically generated

The log-transformed data seems to be more evenly spread around the plot

III)

Chart, histogram

Description automatically generated

Non-transformed

Chart, histogram

Description automatically generated

Log-transformed looks much more normally distributed.

IV)

It seems based on my above responses, that a log-transformed model fits the assumptions better. The assumption of normality is better met with the log-transformed model, based on the histograms. Additionally, it seems that the assumption of linearity is better met with the log-transformed model based on the scatterplots.

The assumption of equal standard deviations looks substantially better from the log-transformed data, and this is based on the residual scatter plots.

Finally, I will make the assumption that the data is independent, we do not know exactly how the data was collected, but to proceed we will assume its independent. The data also seems like there are no outliers that will cause problems, although should one occur we can use a technique such as dfbetas to determine if the outlier is causing major problems.

Based on the information above, it seems the log transformed model will do best.

Part C)

Table

Description automatically generated

Part D)

Mu\_hat(log(Metabolic rate)|log(mass^3/4))=5.638330664+0.984991519\*mass^3/4

SAS output:

Table

Description automatically generated

Part E)

Interpretation: I found this difficult, mostly due to the log-transformation and the mass^3/4 parameter.

The median increases by 2^0.984991519 when mass^3/4 is doubled. Essentially when we double mass^3/4, the median increases by 98%.

A 95% CI:

(20.9467,21.0243)

Part F)

Based on the data we can say that 96.49% of the variability in the metabolic rate of the different species of mammals can be accounted for by mass.

SAS CODE for Question 1:

FILENAME REFFILE '/folders/myfolders/sasuser.v94/metabolic.csv';

PROC IMPORT DATAFILE=REFFILE

DBMS=CSV

OUT=metabolic;

GETNAMES=YES;

RUN;

\*Need to raise mass to 0.75 power;

data metabolic;

set metabolic;

powermass = mass\*\*0.75;

run;

\*Create model and view assumptions;

proc glm data = metabolic plots = all;

model metab = powermass/solution;

run;

\*Log-transform things;

data metabolic;

set metabolic;

logpowermass=log(powermass);

logmetab=log(metab);

run;

\*look at model with log transformed things;

proc glm data = metabolic plots = all;

model logmetab = logpowermass/solution clparm;

run;

Question 2)

Part A:

I)

R input:

#Part I

#Read in data

autismDat <- read.csv("autismDat.csv")

#Check Data

#head(autismDat)

#Create LM for Prevalence vs Year

autlm <- lm(Prevalence ~ Year, data = autismDat)

#Extract year data

yearX <- autismDat$Year

#Sort year for plotting purposes

sortYear <- sort(yearX)

#Confidence Interval (95%)

prd\_c <- predict(autlm, newdata= data.frame(Year = sortYear), interval=c("confidence"),

type = c("response"), level=0.95)

#Prediction Interval (95%)

prd\_p <- predict(autlm, newdata= data.frame(Year = sortYear), interval=c("prediction"), level=0.95)

#Plot with confidence and prediction intervals

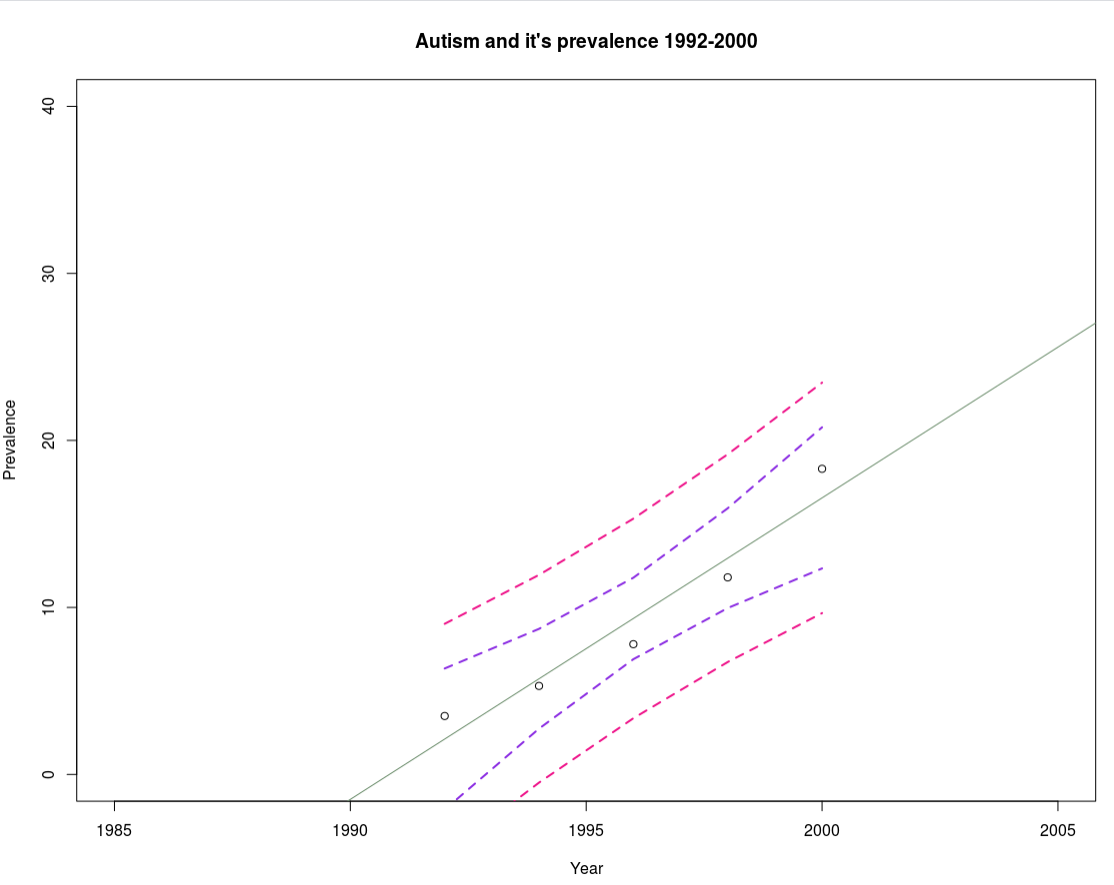
plot(autismDat[,1], autismDat[,2],xlim = c(1985,2005), ylim = c(0,40), xlab = "Year",

ylab = "Prevalence", main = "Autism and it's prevalence 1992-2000")

p <- abline(autlm, col = "darkseagreen4")+lines(sortYear,prd\_c[,2],col = "blueviolet",lty = 2, lwd = 2)+lines(sortYear,prd\_c[,3],col = "blueviolet", lty = 2, lwd = 2)+lines(sortYear,prd\_p[,2],col = "deeppink2", lty = 2, lwd = 2)+lines(sortYear,prd\_p[,3],col = "deeppink2", lty = 2, lwd = 2)

p

R output:



R Input for Log transformed Data:

#Data def looks non-linear, log transform

autismDat$log.prevalence <- log(autismDat$Prevalence)

#Create LM for Prevalence vs Year

autlmLog <- lm(log.prevalence ~ Year, data = autismDat)

#Confidence Interval (95%)

prd\_c <- predict(autlmLog, newdata= data.frame(Year = sortYear), interval=c("confidence"),

type = c("response"), level=0.95)

#Prediction Interval (95%)

prd\_p <- predict(autlmLog, newdata= data.frame(Year = sortYear), interval=c("prediction"), level=0.95)

#Plot with confidence and prediction intervals

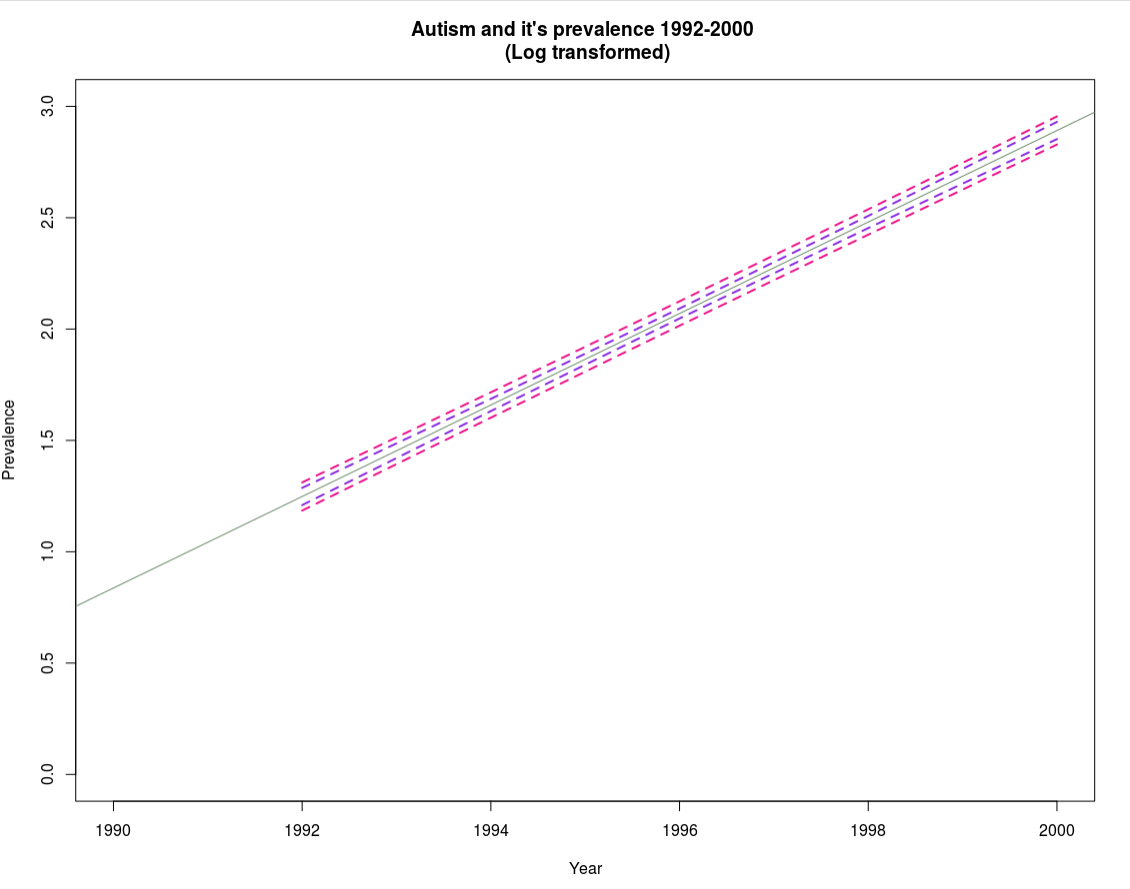
plot(autismDat[,1], autismDat[,2],xlim = c(1990,2000), ylim = c(0,3), xlab = "Year",

ylab = "Prevalence", main = "Autism and it's prevalence 1992-2000")

p <- abline(autlmLog, col = "darkseagreen4")+lines(sortYear,prd\_c[,2],col = "blueviolet",lty = 2, lwd = 2)+lines(sortYear,prd\_c[,3],col = "blueviolet", lty = 2, lwd = 2)+lines(sortYear,prd\_p[,2],col = "deeppink2", lty = 2, lwd = 2)+lines(sortYear,prd\_p[,3],col = "deeppink2", lty = 2, lwd = 2)

p

R output for Log transformed data:



Part II

R input:

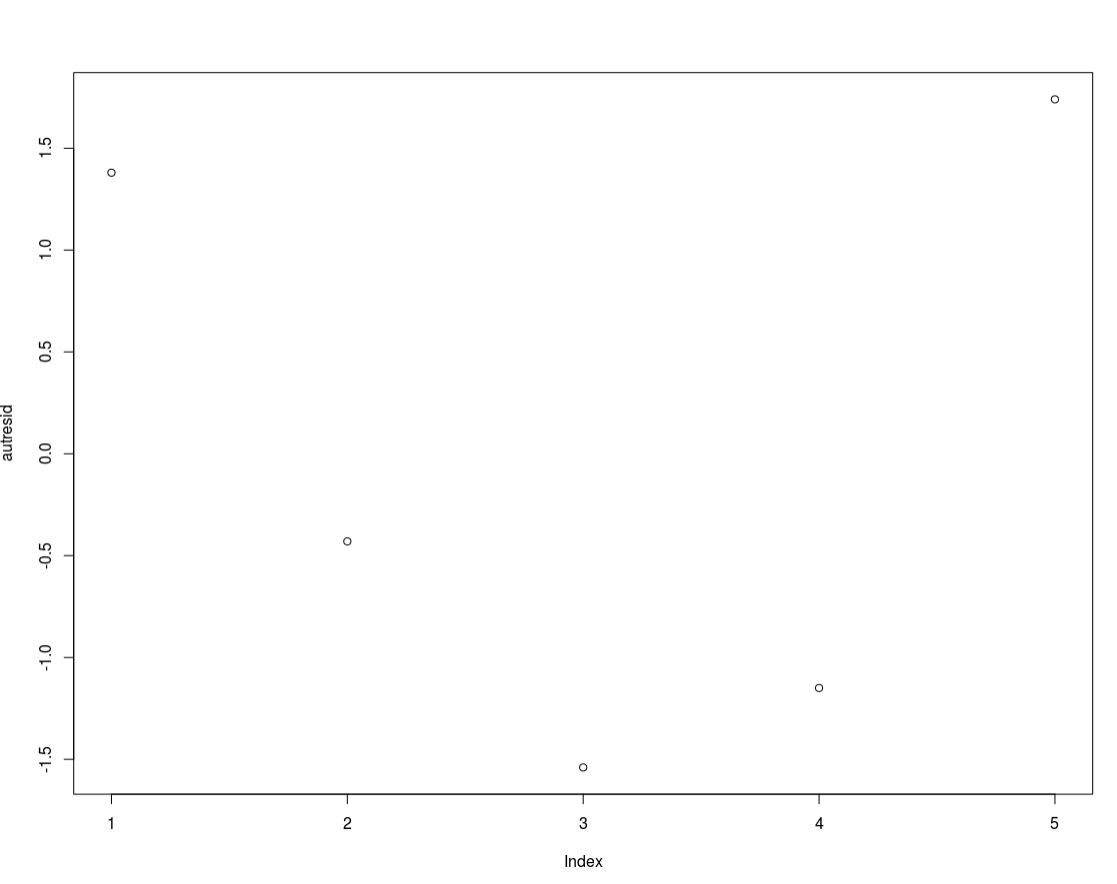
#Residuals

#residuals for original model

autresid <- resid(autlm)

plot(autresid)

R output:



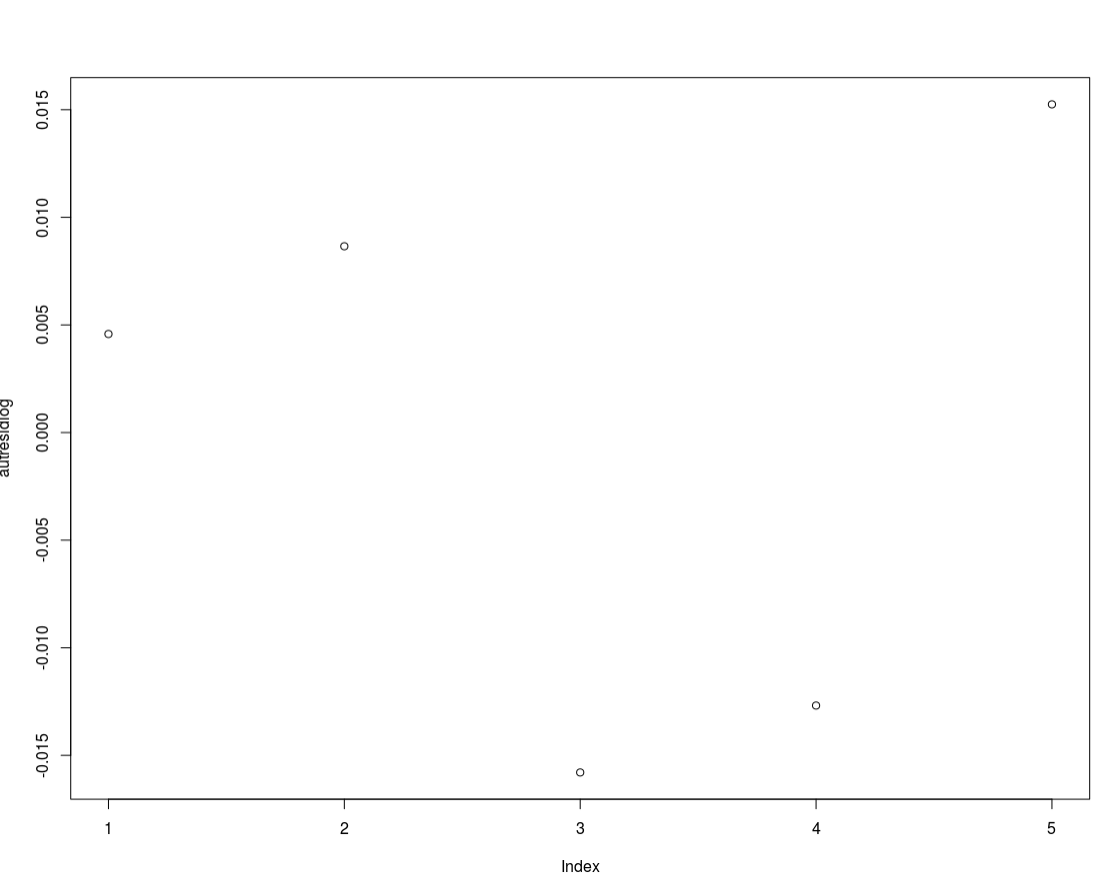
R input for log transformed data:

#residuals for log transforemd model

autresidlog <- resid(autlmLog)

plot(autresidlog)

R output for log transformed data:



Part III)

R input for A histogram of residuals with the normal distribution superimposed:

studentizedresaut <- rstudent(autlm)

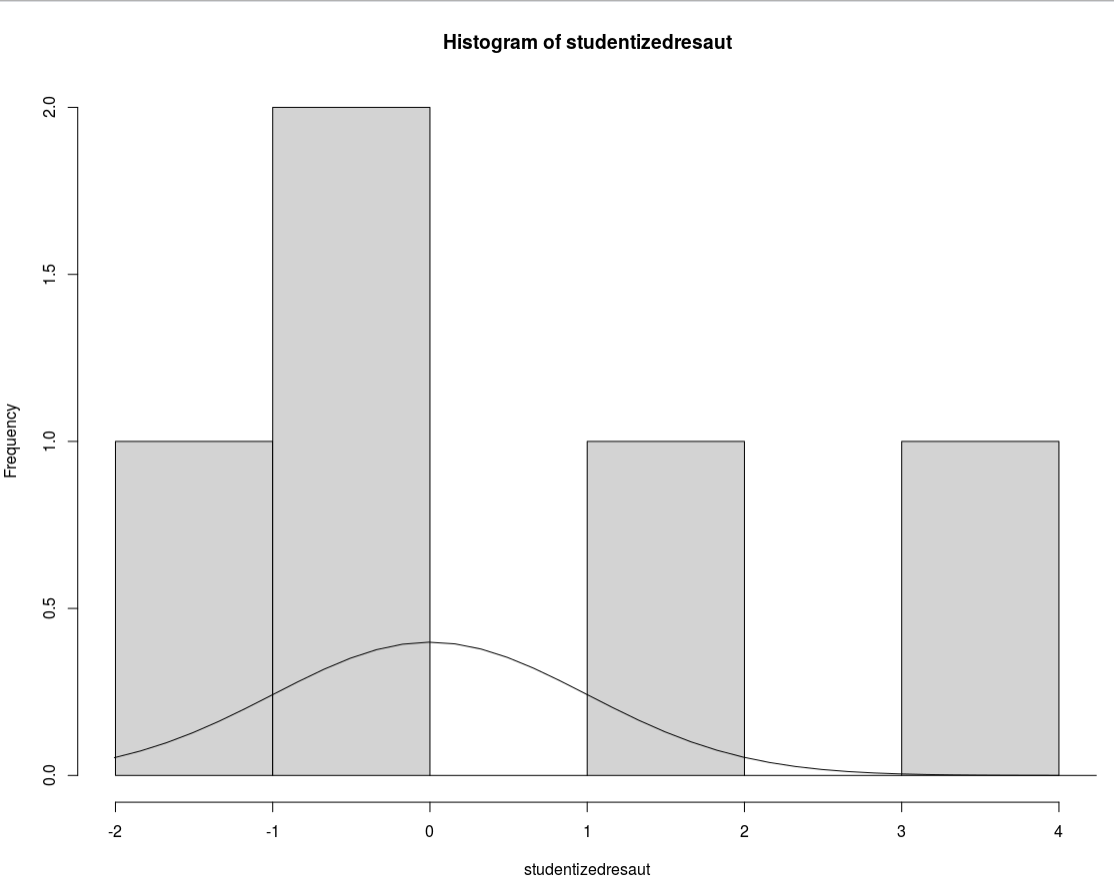
hist(studentizedresaut)

xplot <- seq(min(studentizedresaut)-1, max(studentizedresaut)+1, length=40)

yplot <- (dnorm(xplot))

lines(xplot, yplot, ylim=c(0,0.5))

R output for A histogram of residuals with the normal distribution superimposed:



R input for A histogram of residuals with the normal distribution superimposed log transformed:

#Log transformed part

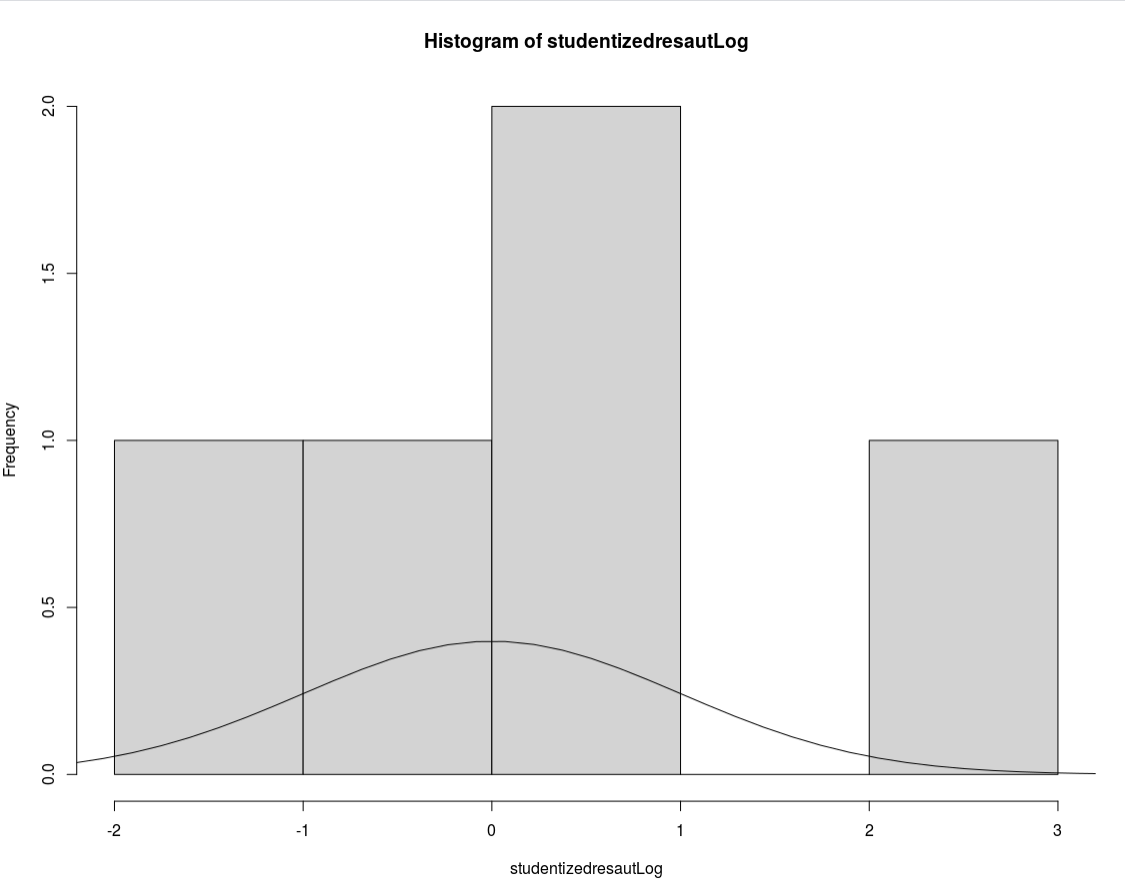
studentizedresautLog <- rstudent(autlmLog)

hist(studentizedresautLog)

xplotlog <- seq(min(studentizedresautLog)-1, max(studentizedresautLog)+1, length=40)

yplotlog <- (dnorm(xplotlog))lines(xplotlog, yplotlog, ylim=c(0,0.5))

R output for A histogram of residuals with the normal distribution superimposed log transformed:



Part IV:

R input for Quantile-Quantile plots of non-transformed data

#Part IV

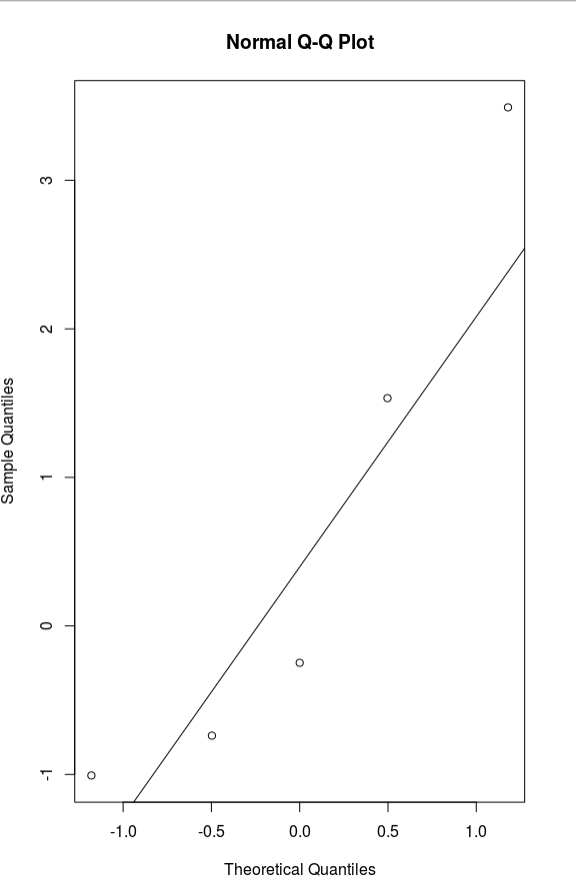
##Quantile-Quantile plots

par(mfrow=c(1,2))

p1 <- qqnorm(studentizedresaut)+qqline(studentizedresaut)

p2 <- qqnorm(studentizedresautLog)+qqline(studentizedresautLog)

R output for Quantile-Quantile plots of non-transformed data



R input for Quantile-Quantile plots of transformed data

#Part IV

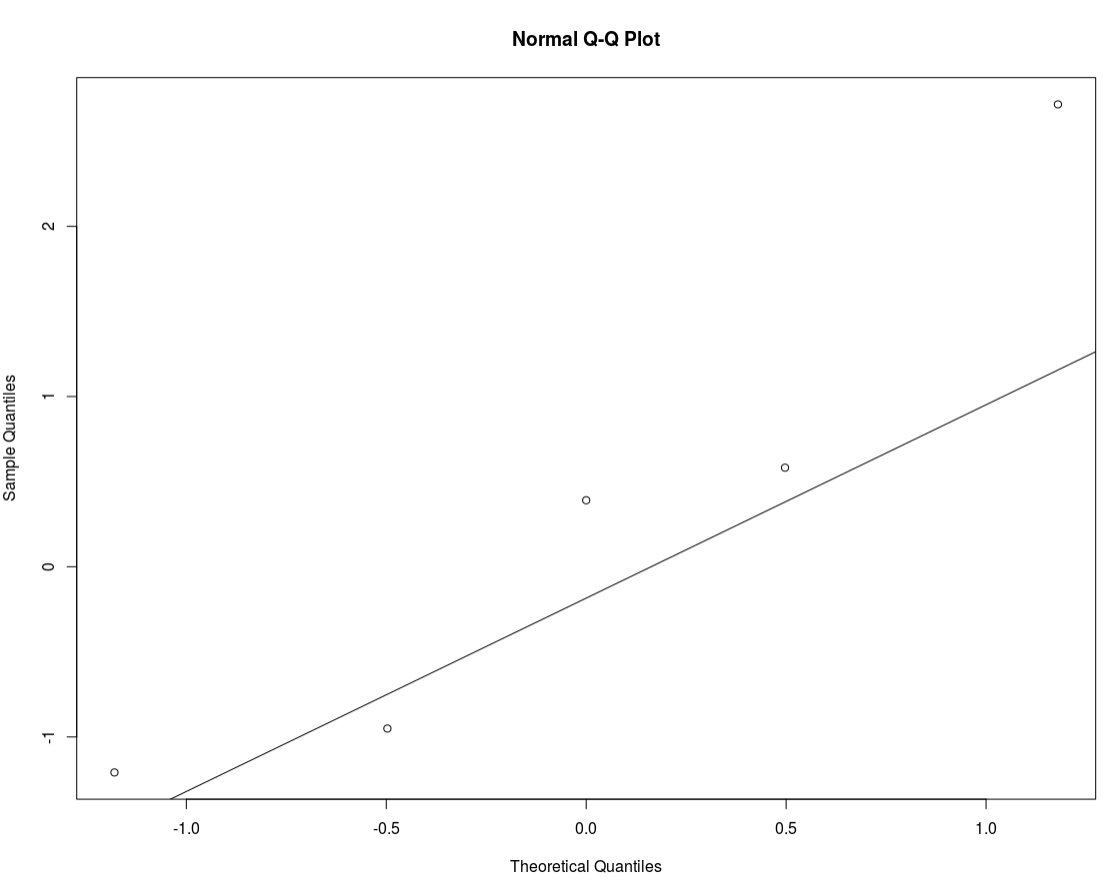
##Quantile-Quantile plots

par(mfrow=c(1,2))

p1 <- qqnorm(studentizedresaut)+qqline(studentizedresaut)

p2 <- qqnorm(studentizedresautLog)+qqline(studentizedresautLog)

R output for Quantile-Quantile plots of transformed data



Discussion of why I am choosing the log transformed model:

The normality assumption is better met with the log-transformed model, I feel this way because the histograms (although the tiny sample size), “looks better” than the non-transformed model. Additionally the Quantile-Quantile plots look much better on the log-transformed model.

Based on the scatterplots, the log-transformed model does better with the linearity assumption when compared to the non-transformed model. Again, “all models are wrong, some are useful.”

With regards to equal standard deviations, we really can’t tell if this assumption ISN’T met. After viewing the plots of the residuals, it would appear this assumption is met!

The data appears to be independent, there isn’t really a hard and fast rule for this, but I will assume these observations are independent.

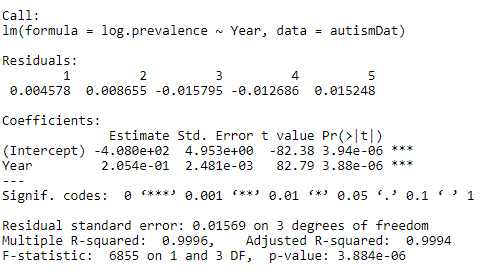
In general, the log-transformed model seems like the model to choose, the assumptions are better met and although we have a very tiny sample size, it seems like an okay decision to proceed with such a model. It’s also important to note that it doesn’t seem like there are any obvious outliers that we need to deal with.

Part B)

R input:

Summary(autlmLog)

R output:



Part C)

mu\_hat(log(Prevalence)|Year)=-4.080e^2+2.054e^-1\*Year

Part D)

I transformed the model, below is the interpretation:

For each additional year, we have a multiplicative change of 2.054E^-1 in the median. Furthermore we can estimate that each additional year our median increases by 2.054e^-1.

95% CI for the increase in the median is: (1.2234, 1.24432)

Part E)

We can measure the proportion of variation in the response that is accounted for by the explanatory variable is 0.996

So, we can say that about 99.6% of the variation in prevalence is accounted for by the explanatory variable year.

Bonus:

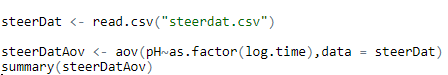
Part A)

H0: There is no difference between the linear regression model and the separate means model.

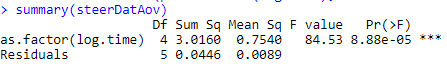
HA: There is a difference between the linear regression model and the separate means model.

Part B)

R input:



R output:



Part C)

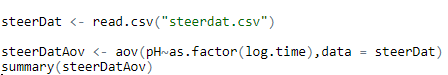
Reject H\_0

Part D)

There is sufficient evidence to conclude that there is a difference between the linear regression model and the separate means model (p-value <0.0001)

Part E)

R input:



R output:

