UNIT 11 HW

1. From Problem 26, Chapter 8:

The Metabolic data set has the average mass, metabolic rate, and average lifespan of 95 different species of mammals. Kleiber’s Law states that the metabolic rate of an animal species, on average, is proportional to its mass raised to the power ¾. Judge the adequacy of this theory with these data. Ultimately, for this problem, we want to find the best model. (At this point, you will limit the analysis to the two variables under study, though the data set has more variables.) In the current data set, assume that mass has not yet been raised to the power ¾.

* Use alpha = 0.05.
* Use **SAS** for this problem.
* Include **relevant** code and output. Make sure you directly answer the questions. Do NOT assume the answer is obvious from the output.

Specifically, provide/answer the following:

* + 1. Judging by a scatterplot alone, does it seem reasonable that the metabolic rate of an animal species, on average, is proportional to its mass raised to the power of ¾? (Recall that if some variable y is proportional to the variable x, then (with nonzero m) is a well-fitting model.) In other words, does the data (metabolic rate, mass3/4) reasonably fall along a straight (nonhorizontal) line and nearly pass through the origin?

Code:

PROC IMPORT OUT= WORK.metabolism

DATAFILE= "/home/marinfamily1010/sasuser.v94/Data/MetabolismDataProb26\_2\_2\_2.xlsx"

DBMS=xlsx REPLACE;

GETNAMES=YES;

DATAROW=2;

RUN;

data WORK.metabolism3quarter;

set work.metabolism;

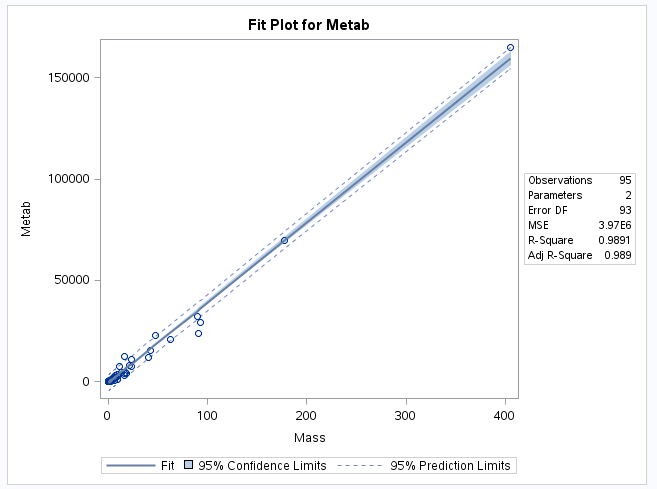
Mass= Mass\*\*(3/4);

run;

proc reg data = work.metabolism3quarter;

model Metab = Mass /cli;

run;

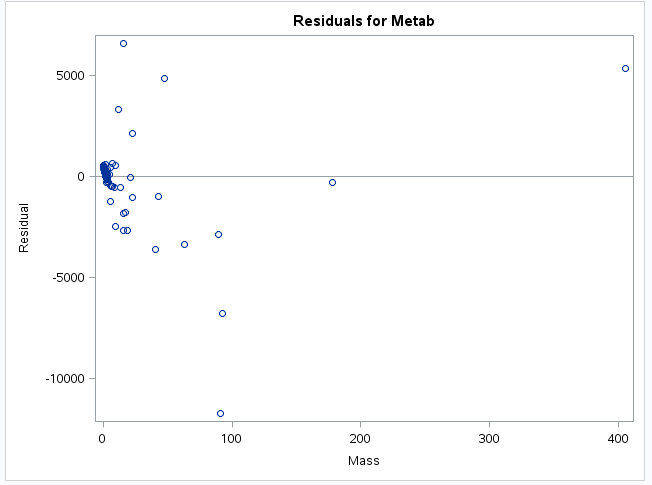


When looking at the scatter plot alone. The regression line seems to fit the data well when taking the Mass^(3/4). Also, the rsquared value is .989 which is extremely good suggesting a good fit. Note that there is a cluster of points right around zero that is hard to tell how well the fit is.

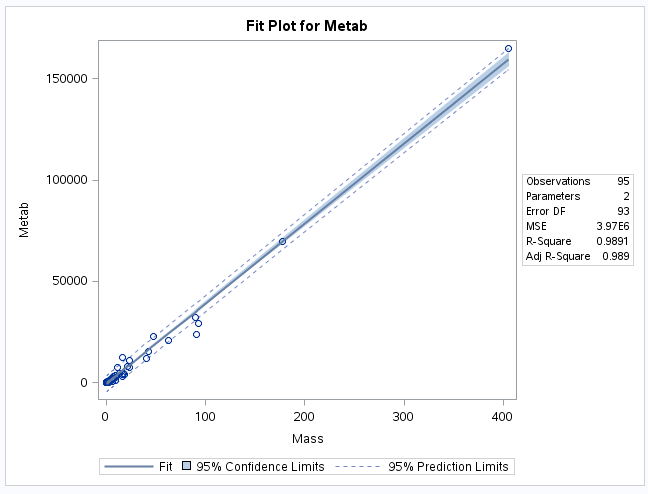
* + 1. We want to find the “best” model to predict metabolic rate from mass3/4 **and** make appropriate statistical inferences. Therefore, address all the assumptions prior to the analysis (using mass3/4). If the assumptions are not met, handle the data appropriately. If a transformation is used to satisfy the assumptions, address the assumptions again to ensure that the transformation is logical, and carry out your analysis on your newly transformed data. For example, you should include a scatter plot for the original data AND transformed data, etc. (Hint: if a transformation is necessary, try one of the transformations discussed in class first.) Either way, keep the “mass3/4” in the model; do not go back to regular “mass,” although mass3/4 may be transformed if it makes sense for the assumptions. At minimum, provide and interpret the following elements to address assumptions FOR THE ORIGINAL DATA AND ANY TRANSFORMED DATA (IF you use a transformation). You may include more graphs if you find them useful.

Assumptions:

* Linearity
  + A straight line may be inadequate.
  + Outliers may bias estimates.

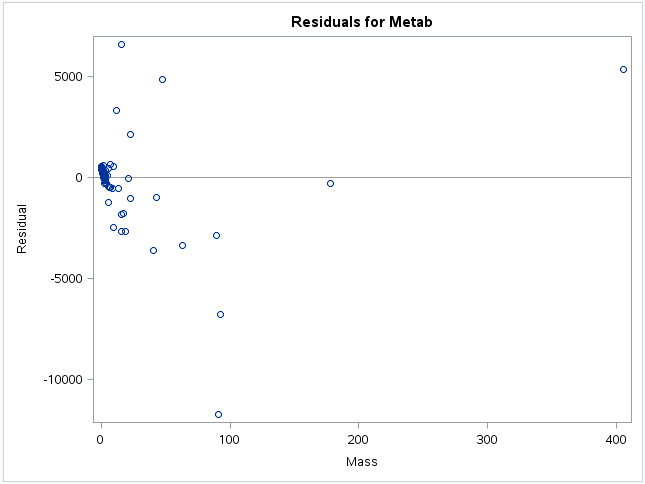


It appears that after subtracting the mean, we have evidence of outliers and some evidence of non-linearity.



However, in regular scatterplot, non-linearity is hard to see.

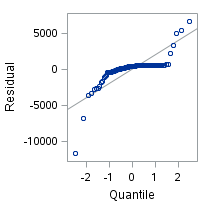
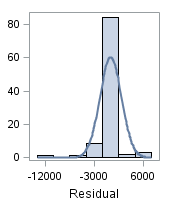
Constant Variance:



Doesn’t exactly look like a random cloud and there is some curvature beyond mass of 50.

Normality

When looking at the qqplot, the data doesn’t quite look normally distributed here. See below. However, the histogram looks a bit normal minus the value at -12000 being a pretty extreme outlier.

1. A scatterplot with the following included on the graph: regression line, confidence intervals of the regression line, and prediction intervals of the regression line.

ii. A scatterplot of residuals.

iii. A histogram of residuals with the normal distribution superimposed.

iv. A discussion supporting the use of the model you chose (support that the assumptions are met).

Will try a log transformation of Mass:

Code:

data work.metabolismxtransform;

set work.metabolism3quarter;

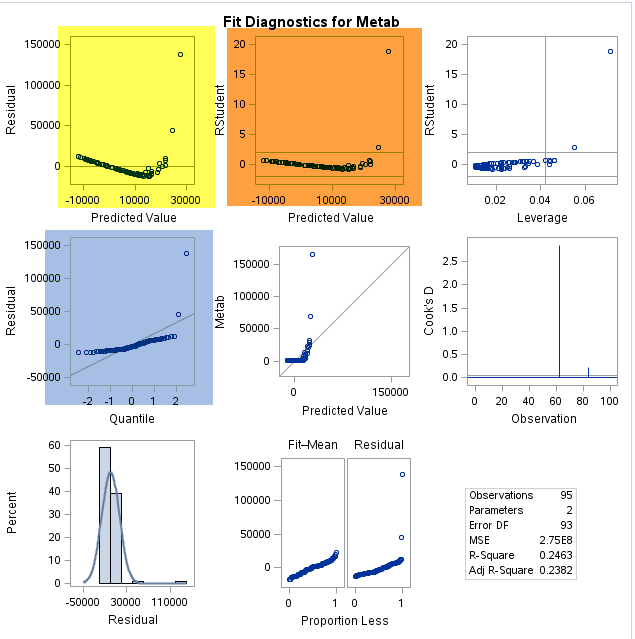
Mass = log(Mass);

run;

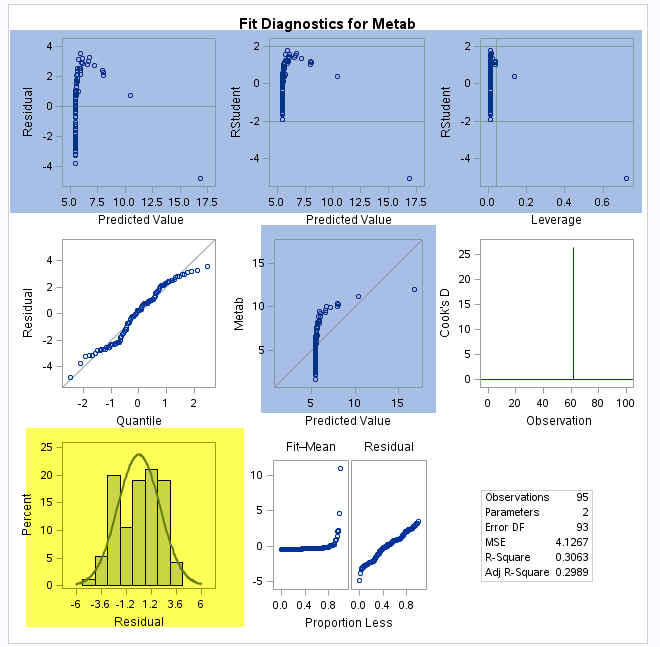
proc reg data = work.metabolismxtransform;

model Metab = Mass /cli;

run;

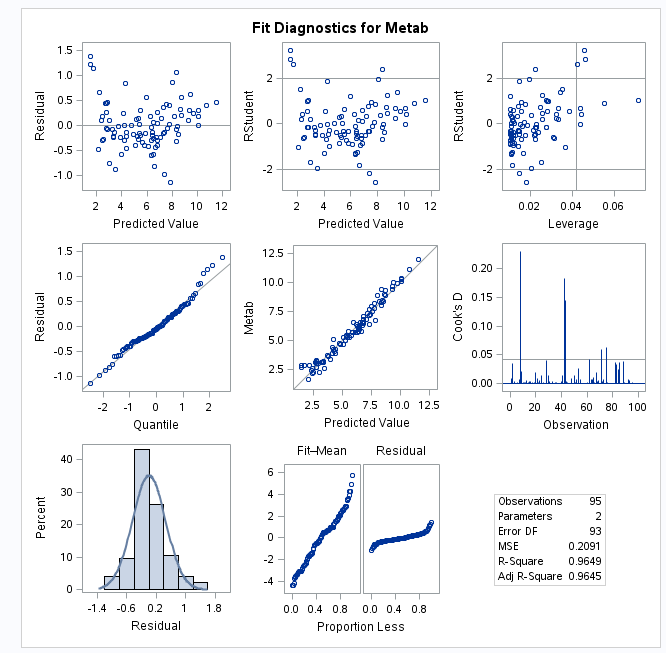


After doing the log transformation, the plots looks worse. These plots are non-linear, don’t have equal standard deviations, and not normally distributed. Will try a transformation of the y and revert x.



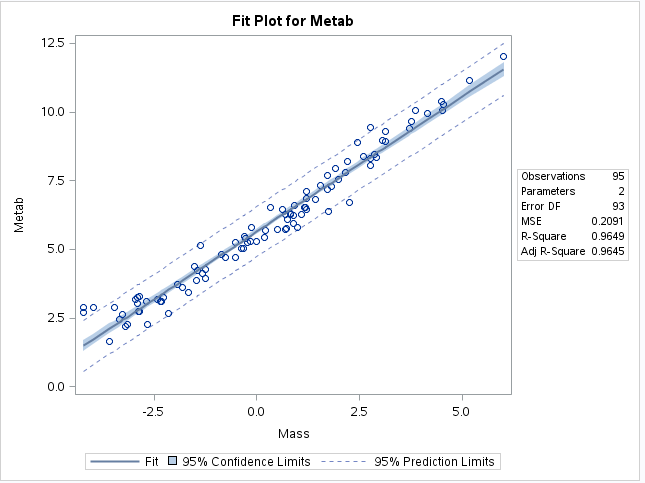
When transforming the y (log of y) the data appears to be normal, but not linear nor having a constant variance.

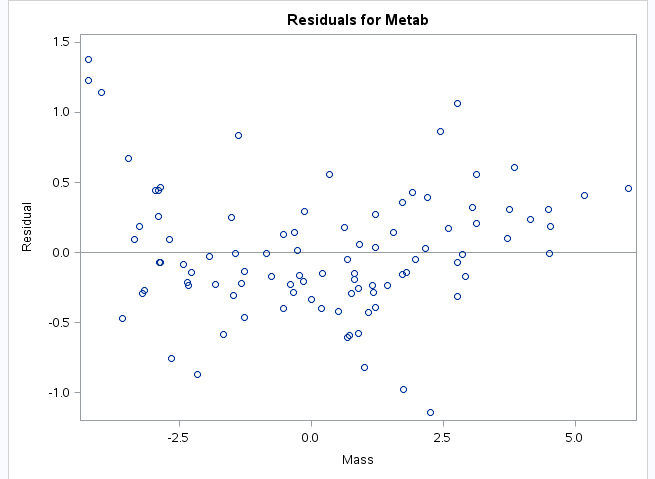
Will try a log transformation of x and y.



This seems to pass all assumptions.

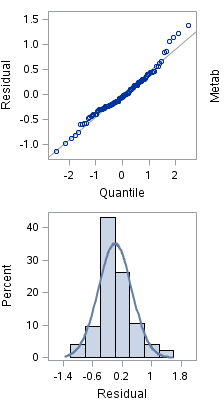
The with x and y log transformed, the data is linear with a r-square of .97.





We also have equal variance with even scatter and no curvature when looking at the residual plot.

When looking at the qq plot and histogram, the data suggests that they are normally distributed:



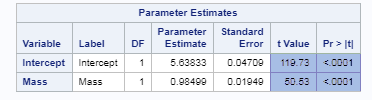
* + 1. Once a reasonable model is found (possibly using a transformation), provide a table showing the t-statistics and p-values for the significance of the regression parameters .

Code:

proc reg data = work.metabolismxytransform;

model Metab = Mass /cli;

run;



* + 1. The estimated regression equation. Make sure the dependent variable is noted as the predicted value or predicted mean value, not just the dependent variable.
    2. Interpretation of the model, paying special attention if you used a transformation (hint!). That is, interpret the slope as well as the **confidence interval**.

A 95% confidence interval for B1 is

.98499 +- 1.98580 \* .01949

= (.94,1.02)

Therefore, a 95% confidence interval for the multiplicative increase in median of Metab after taking Mass^(3/4) is (e^.985,e^1.02) = (2.55,2.77).

* + 1. A measure of the proportion of variation in the response that is accounted for by the explanatory variable. Interpret this measure clearly.

It is estimated that the amount of Mass^(3/4) is about 96.49% of the Metabolic rate.

1. From Problem 29, Chapter 8:

The autism data show the prevalence of autism per 10,000 ten-year-old children in the United States in each of five years. Analyze the data to describe the change in the distribution of autism prevalence per year during this time period.

* Use alpha = 0.05.
* Use **R** for this problem.
* Include **relevant** code and output. Make sure you directly answer the questions. Do NOT assume the answer is obvious in the output.

Specifically, provide/answer the following:

* + 1. Address all the assumptions for a linear regression model prior to the analysis. If the assumptions are not met, handle the data appropriately. If a transformation is used, address the assumptions again with the transformed data to ensure that the transformation is logical. The questions below should reflect this. For example, you should include a scatter plot for the original data AND transformed data, etc. (Hint: if a transformation is necessary, try one of the transformations discussed in class first.) At minimum, provide and interpret the following elements to address assumptions FOR THE ORIGINAL DATA AND ANY TRANSFORMED DATA (IF you use a transformation). You may include more graphs if you find them useful.

Code:

library(xlsx)

library(olsrr)

setwd("C:/Users/Marin Family/Desktop/Statistical Foundations for Data Science/Unit 11")

autism <- read.xlsx("AutismDataProb29\_2\_2\_2.xlsx", "Autism Data Prob 29")

model=lm(Year~Prevalence, data = autism)

plot(autism)

newx=autism$Prevalence

newx=sort(newx)

prd\_c=predict(model, newdata= data.frame(Prevalence = newx), interval=c("confidence"), type = c("response"), level=.95)

prd\_c

prd\_p=predict(model, newdata= data.frame(Prevalence = newx), interval=c("prediction"), type = c("response"), level=.95)

prd\_p

#Plot with confidence and prediction intervals

plot(autism[,2],autism[,1],xlim = c(0,20), ylim = c(1990,2000),xlab = "Prevelance",ylab = "Year", main = "Autism Prevelance per Year")

abline(model, col = "red")

lines(newx,prd\_c[,2],col = "blue",lty = 2, lwd = 2)

lines(newx,prd\_c[,3],col = "blue", lty = 2, lwd = 2)

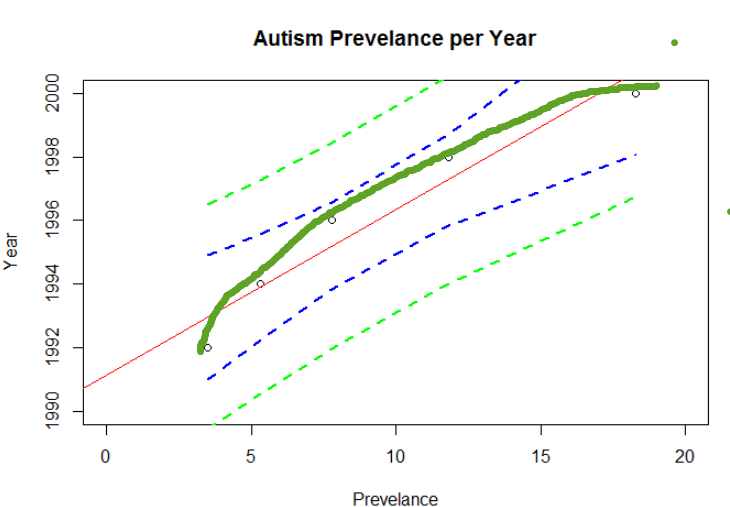
lines(newx,prd\_p[,2],col = "green", lty = 2, lwd = 2)

lines(newx,prd\_p[,3],col = "green", lty = 2, lwd = 2)

Assumptions:

Honestly, with 5 observations, I really don’t see a point. At work we would reject continuing with an analysis if we don’t have enough observations to begin with. Will continue and apply principles anyway.

1. A scatterplot with the following included on the graph: regression line, confidence intervals of the regression line, and prediction intervals of the regression line.

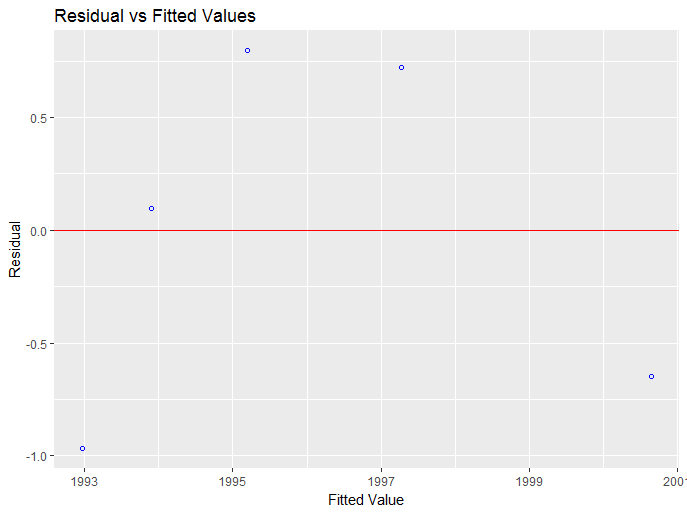


Linearity - The points don’t exactly look linear and looked slightly curved.

ii. A scatterplot of residuals.

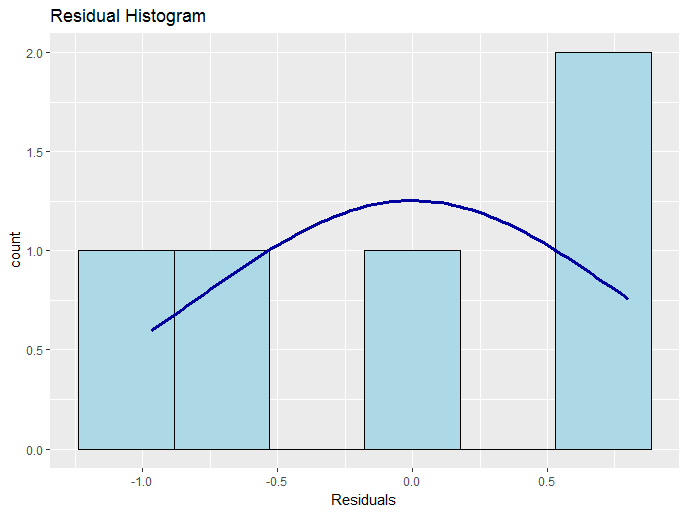
Additional Code:

ols\_rvsp\_plot(model)



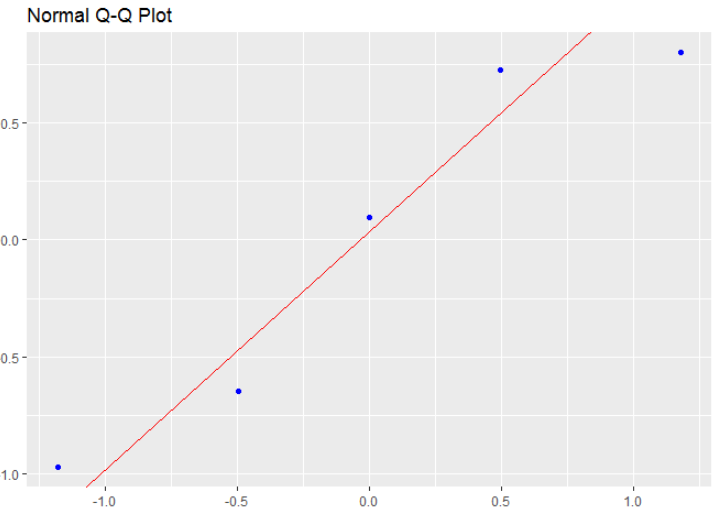
Not a lot of data here, but there appears to be curvature and doesn’t exactly look random.

iii. A histogram of residuals with the normal distribution superimposed.



This does not look normal either, but we don’t have a lot of data points.

Will look at qqplot:



Data isn’t quite within the line and doesn’t appear to be normal.

iv. A discussion supporting the use of the model you chose (support that the assumptions are met).

Taking the log of the years doesn’t make very much sense to me right now, going to take the natural log of Prevalence.

Code:

library(xlsx)

library(olsrr)

setwd("C:/Users/Marin Family/Desktop/Statistical Foundations for Data Science/Unit 11")

autism <- read.xlsx("AutismDataProb29\_2\_2\_2.xlsx", "Autism Data Prob 29")

autism$Prevalence <- log(autism$Prevalence)

#autism$Year <- log(autism$Year)

model=lm(Year~Prevalence, data = autism)

plot(autism$Year, autism$Prevalence)

newx=autism$Prevalence

newx=sort(newx)

prd\_c=predict(model, newdata= data.frame(Prevalence = newx), interval=c("confidence"), type = c("response"), level=.95)

prd\_c

prd\_p=predict(model, newdata= data.frame(Prevalence = newx), interval=c("prediction"), type = c("response"), level=.95)

prd\_p

#Plot with confidence and prediction intervals

plot(autism[,2],autism[,1],xlim = c(0,3), ylim = c(1990,2010),xlab = "Prevelance",ylab = "Year", main = "Autism Prevelance per Year")

abline(model, col = "red")

lines(newx,prd\_c[,2],col = "blue",lty = 2, lwd = 2)

lines(newx,prd\_c[,3],col = "blue", lty = 2, lwd = 2)

lines(newx,prd\_p[,2],col = "green", lty = 2, lwd = 2)

lines(newx,prd\_p[,3],col = "green", lty = 2, lwd = 2)

#qqplot

ols\_rsd\_qqplot(model)

##residual vs fitted test

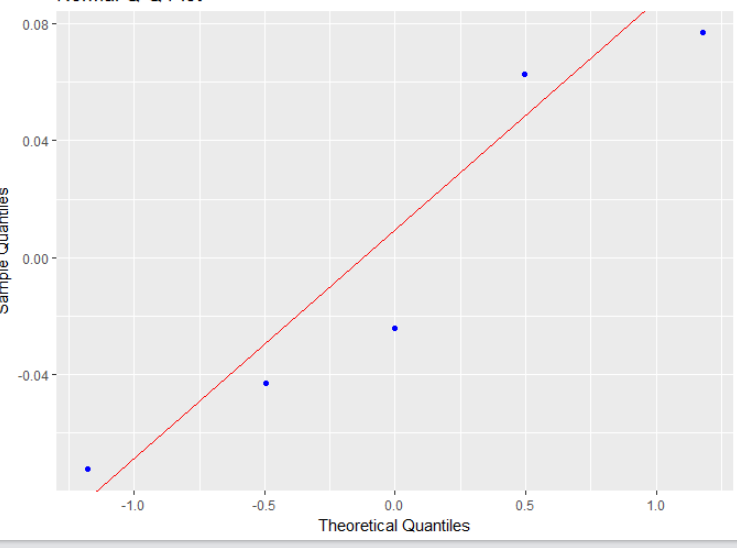
ols\_rvsp\_plot(model)

##histogram

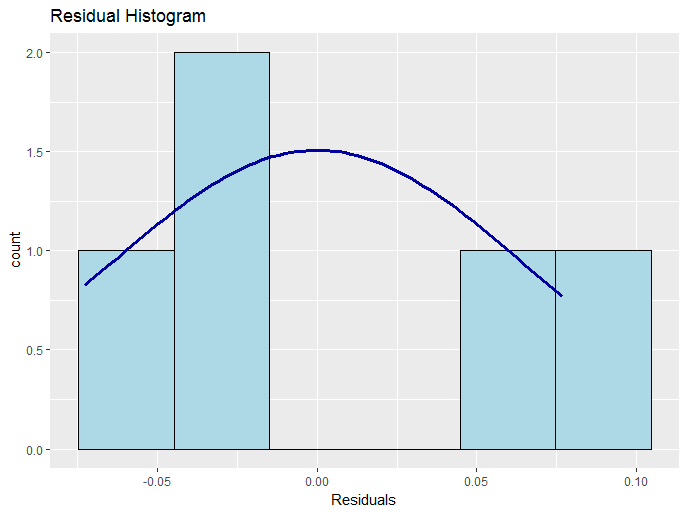
ols\_rsd\_hist(model)

ols\_rsd\_qqplot(model)

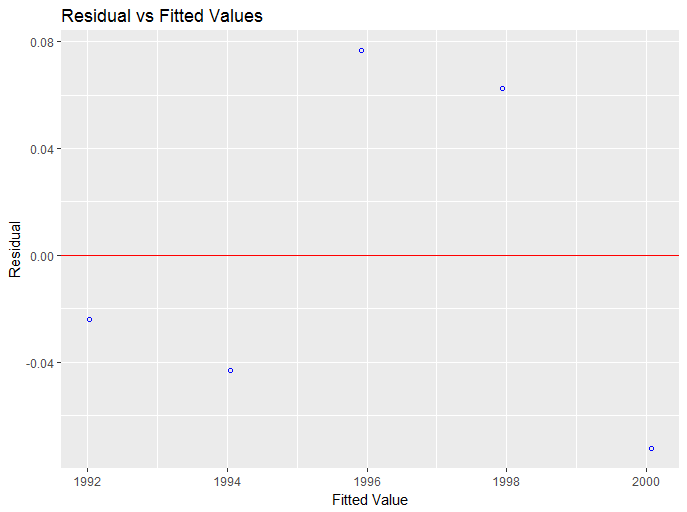
Assumptions again:



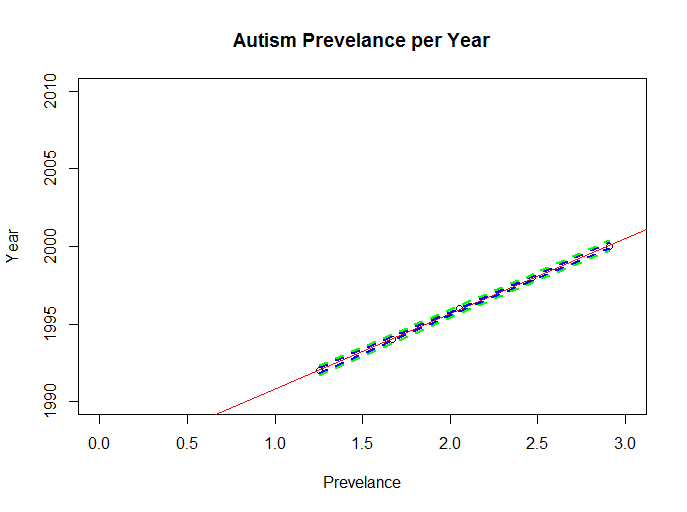
QQPlot doesn’t seem all that different to me. It doesn’t look at normal, but one can argue that it can pass.



With very little data here, this really doesn’t appear normal, but there is little data again.



In regards to equal variance, the data looks a little more random in the residual plot and there appears to be equal variance here with evidence of a random cloud.



When creating a regression model with the natural log of Prevalence, we have a near perfect linear model with a R^2 of .9996.

I’m going to go ahead and use this model with the natural log transformation since we have little data, supported equal variance, and linearity.

* 1. Once a reasonable model is found (possibly using a transformation), provide a table showing the t-statistics and p-values for the significance of the regression parameters .

Code:

Summary$model

Coefficients:

Estimate Std. Error t value Pr(>|t|)

B0:(Intercept) 1.986e+03 1.263e-01 15717.78 5.68e-13 \*\*\*

B1:Prevalence 4.866e+00 5.877e-02 82.79 3.88e-06 \*\*\*

* 1. The estimate regression equation. Make sure the dependent variable is noted as the predicted value or predicted mean value, not just the dependent variable.

Code:

model$cofficients

> model$coefficients

(Intercept) Prevalence

1985.928547 4.865627

* 1. Interpretation of the model, paying special attention if you used a transformation (hint!). That is, interpret the slope as well as the **confidence interval**.

A 95% confidence interval for B1 is

4.86563 +- 3.18245\* .05877

= (4.67,5.05)

Therefore, a 95% confidence interval for the multiplicative increase in median of Prevalence for years is (e^4.67,e^5.05) = (106.69,156.02).

* 1. A measure of the proportion of variation in the response that is accounted for by the explanatory variable. Interpret this measure clearly.

It is estimated that the Year explains about 99.96% of variation in the Prevalence.

1. Bonus! Consider the steer data in Display 7.3 on page 179 (Chapter 7) of the textbook (third edition). Perform a lack of fit test comparing the regression model and a separate means model. Because we have at least two points in at least one group (replication to estimate the variance), we can perform ANOVA. (ANOVA does not make sense if no values of the independent variable are repeated.) During live session, we already addressed the assumptions and determined that a linear-log model is best for regression. Perform this lack of fit test (all parts) on the transformed data. Use the software of your choice. Specifically, include the following:
   1. Hypotheses
   2. The ANOVA table you created
   3. Decision

d. Conclusion in nonstatistical terms

e. Code and relevant output

**Darn…just don’t have time for this**