# Assignment9

## #question1

library(tidyverse)

my\_lizards <- read.csv("lizards.csv")

set.seed(17053777)

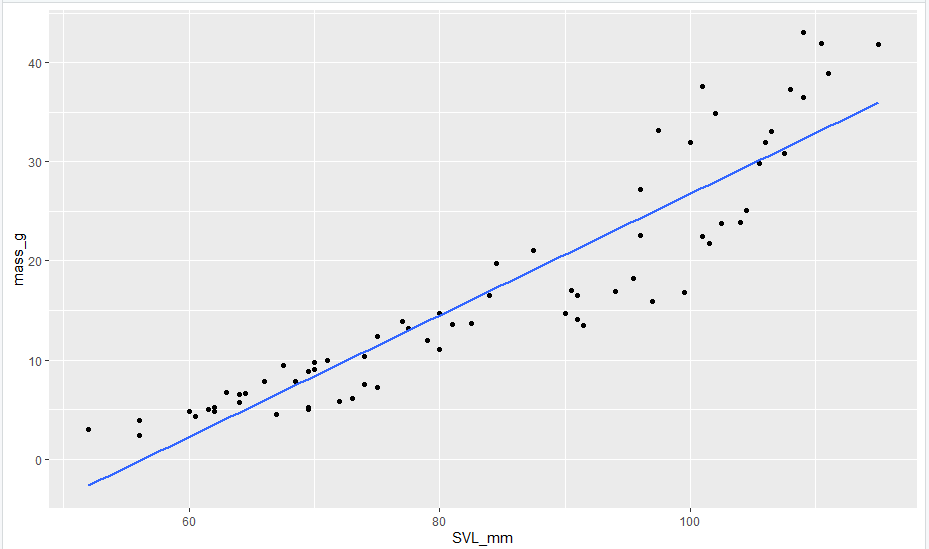
my\_lizards <- my\_lizards %>% sample\_n(70)

## #question2

ggplot(my\_lizards, aes(x = SVL\_mm, y = mass\_g)) +

geom\_point() +

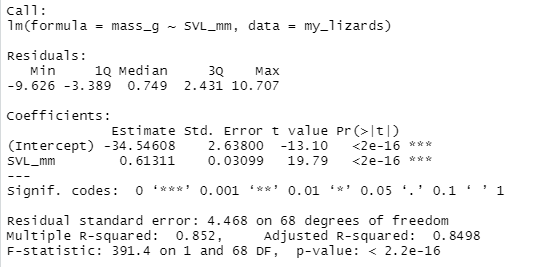
geom\_smooth(method = lm, se = FALSE)



## #question3

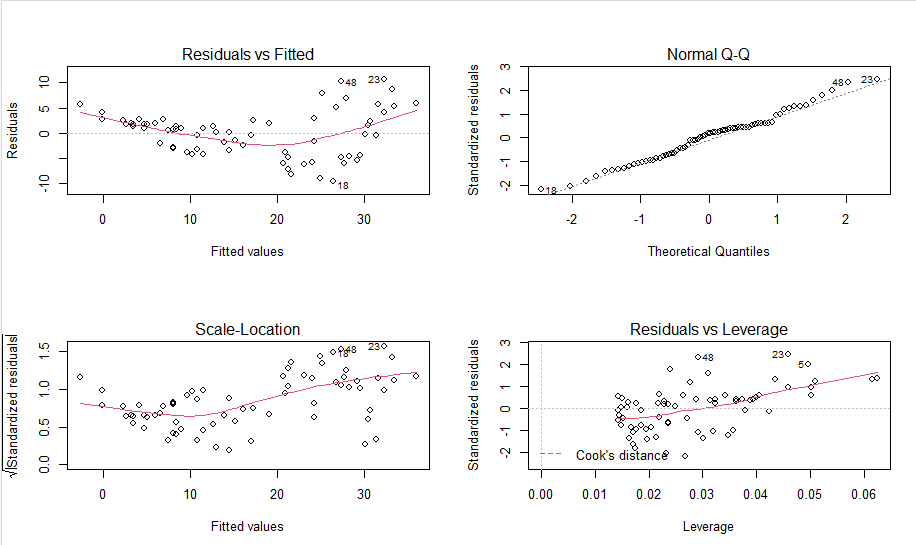
m1 <- lm(mass\_g ~ SVL\_mm, data = my\_lizards)

summary(m1)



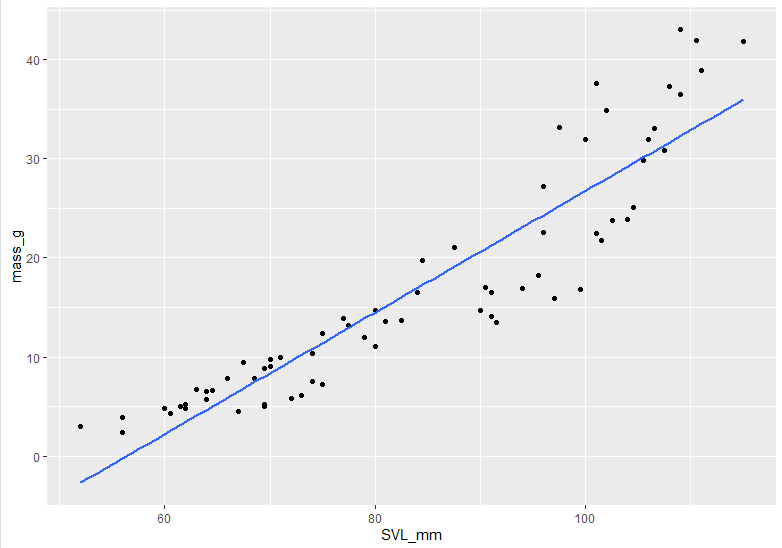
par(mfrow=c(2,2))

plot(m1)



ggplot(my\_lizards,aes(x=SVL\_mm,y=mass\_g))+

geom\_point()+

geom\_smooth(method = lm,se=FALSE) 

# From Residuals vs Fitted, we can see the red line does not match the 0 line, there is a deviation from linearity, so it does not fit the linear assumption.

# The Normal Q-Q graph helps to check the assumption that the errors are normally distributed. From the graph we can see there is some deviation from normality.

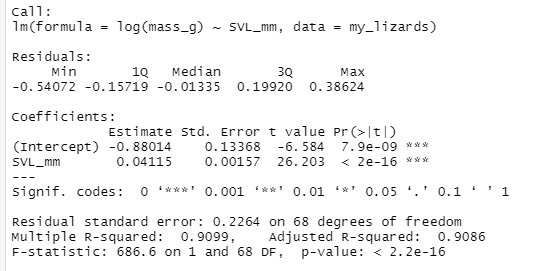
# The Scale-Location plots helps to check the assumption that the variance of the error is constant. From the graph we can see the variance is the similar.

# The residuals vs leverage plot does not suggest any highly influential observation since there is no outliers around the red dash line.

## #question4

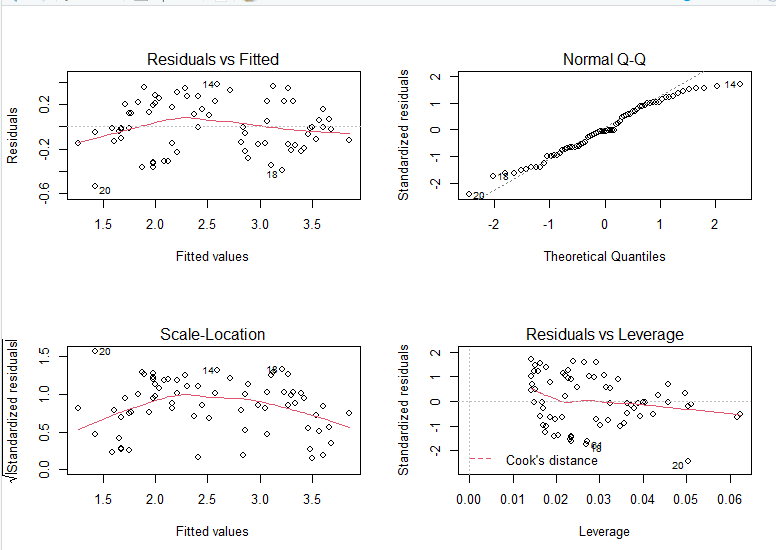
m2 <- lm(log(mass\_g) ~ SVL\_mm, data = my\_lizards)

summary(m2)



par(mfrow=c(2, 2))

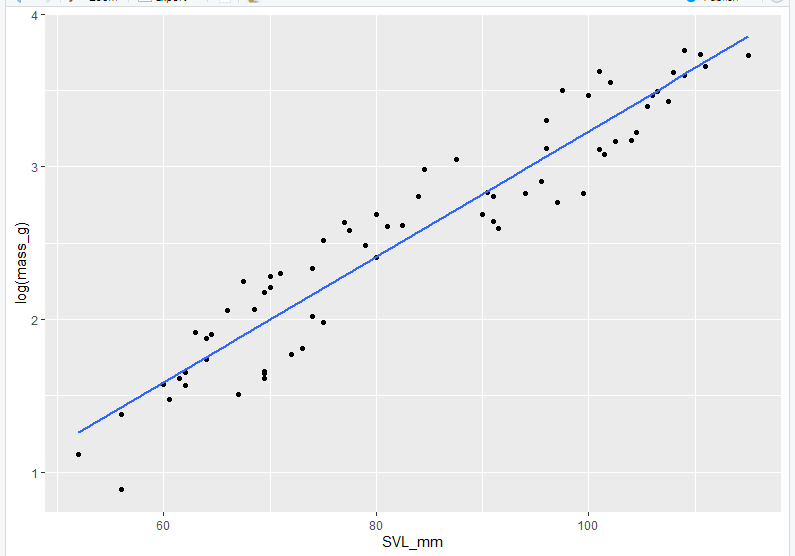
plot(m2)



ggplot(my\_lizards,aes(x=SVL\_mm,y=log(mass\_g)))+

geom\_point()+

geom\_smooth(method = lm,se=FALSE)



# From Residuals vs Fitted, we can see the red line still does not match the 0 line, but it becomes flatter than the red line in m1 so the deviation from linearity is smaller than the deviation in m1.

# The Normal Q-Q graph helps to check the assumption that the errors are normally distributed. From the graph we can see there is still some deviation from normality on the tail.

# The Scale-Location plots helps to check the assumption that the variance of the error is constant. From the graph we can see the variance is the similar, almost constant.

# The residuals vs leverage plot does not suggest any highly influential observation since there is no outliers around the red dash line.

## #question5

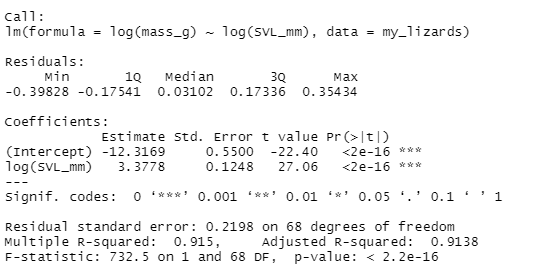
# An increase of 1 in SVL\_mm multiplies mass\_g by e ^ (0.04115).

# A useful approximation, if b1 with a smallish magnitude, an increase of 1 in SVL\_mm multiples mass\_g by approximately (1 + 0.04115)

## #question6

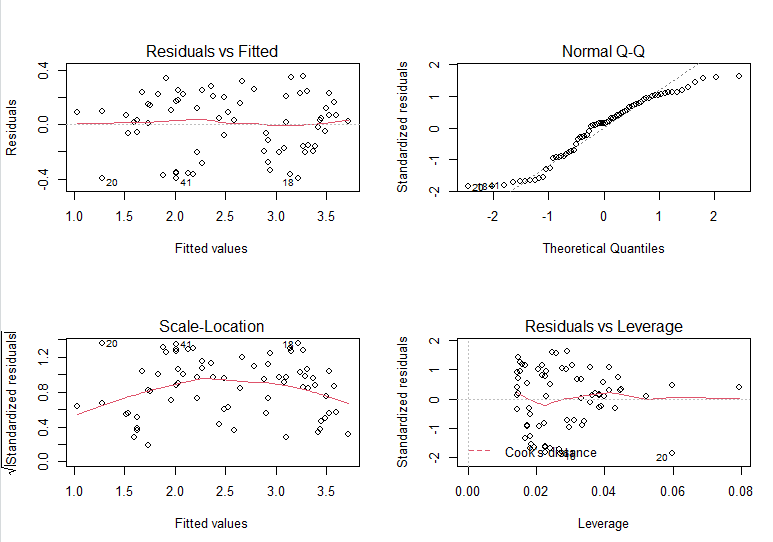
m3 <- lm(log(mass\_g) ~ log(SVL\_mm), data = my\_lizards)

summary(m3)

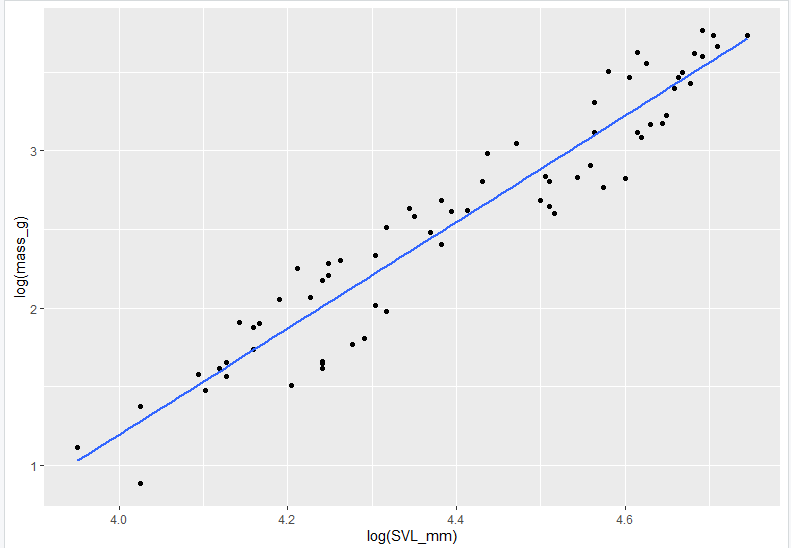


par(mfrow=c(2, 2))

plot(m3)



ggplot(my\_lizards, aes(x=log(SVL\_mm),y=log(mass\_g)))+geom\_point()+geom\_smooth(method = lm,se=FALSE)



# From residuals vs Fitted graph, we can see the red line is much flatter than the red line in m2. It fits the linearity assumption more, so linearity improves.

# From the Normal Q-Q, we can see there is still some deviation from normality.

# From Scale-Location, the variance is almost constant, so the model fits the constant residual variance assumption.

# From Residuals vs Leverage, there is not outlier around the red dash line.

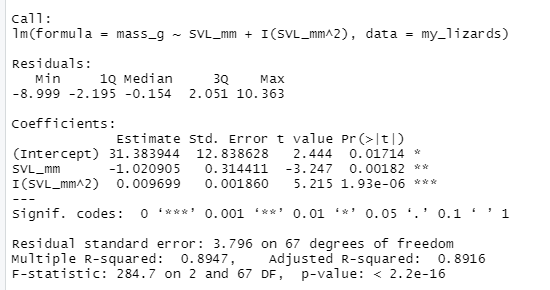
## #question7

# A useful approximation: if b1 with a smallish magnitude , an increase of 1% in SVL\_mm multiplies mass\_g by approximately (1 + 3.3778 ^ 0.01)

## # question8

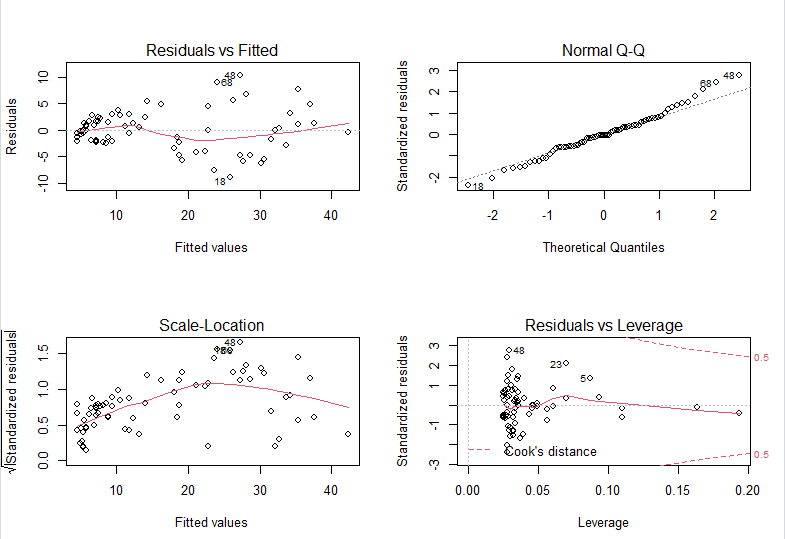
m4 <- lm(mass\_g ~ SVL\_mm + I(SVL\_mm ^ 2), data = my\_lizards)

summary(m4)



par(mfrow=c(2, 2))

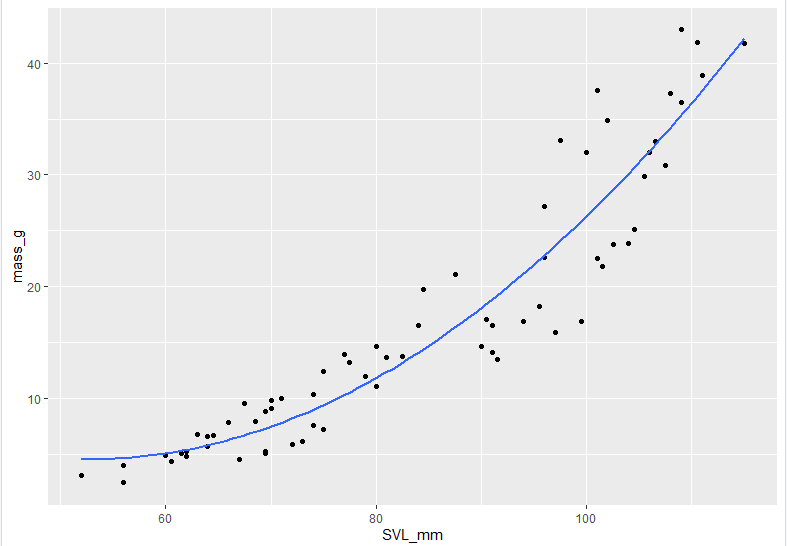
plot(m4)



ggplot(my\_lizards, aes(x=SVL\_mm,y=mass\_g))+

geom\_point()+

geom\_smooth(method = lm,formula = y~x + I(x ^ 2),se=FALSE)



# From Residuals vs Fitted, we can see the red line is as same as flat comparing to m3, so the model fits the linearity assumption.

# From Normal Q-Q, we can see the there are some deviation from the normality.

# From Scale-location graph, we can see the variance is not that constant than the variance in m3. there is a smaller variance at the beginning but then the variance increases.

# From residuals vs Leverage, we can see there is no outlier around the 0.5 red dash line.

## #question9

# the intercept 31.383944 is fitted value of mass\_g when SVL\_mm is 0

# b1 = -1.020905 is the slope of the tangent line to the curve at SVL\_mm = 0

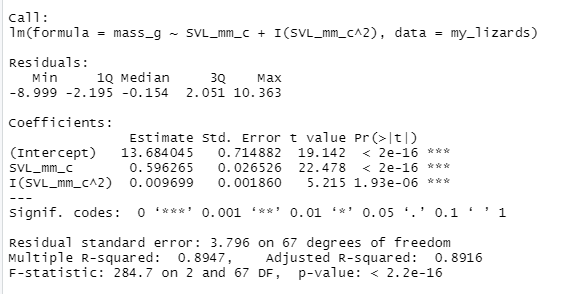
# b2 = 0.009699 is a measure of the increasing curvature.

## #question10

my\_lizards <- my\_lizards %>% mutate(SVL\_mm\_c = (SVL\_mm - mean(SVL\_mm)))

m5 <- lm(mass\_g ~ SVL\_mm\_c + I(SVL\_mm\_c ^ 2), data = my\_lizards)

summary(m5)



# the intercept 13.684045 is fitted value of mass\_g when SVL\_mm is equal to the mean value of SVL\_mm

# b1 = 0.596265 is the slope of the tangent line to the curve at SVL\_mm is equal to the mean of SVL\_mm

# b2 = 0.009699 is a measure of the increasing curvature.

# SVL\_mm means lizard's snout-vent length in millimetres which could not be 0 in real life, so in m4, the interpretation about the intercept does not make sense.

# but in m5, the interpretation is more meaningful since we can get the mean value of SVL\_mm, and SVL\_mm can be equal to the mean of SVL\_mm in the real life.

## #question11

# m1 and m4 could be compared using a nested model Ftest with anova,

# since m1 is the linear relationship of mass\_g and SVL\_mm and m4 includes all terms in m1 and plus an extra quadratic term.

# So m1 is nested to m4

## # question12

my\_lizards %>% ggplot(aes(x = SVL\_mm, y = mass\_g)) +

geom\_point() +

geom\_function( fun = function(x) coef(m1)[[1]] + coef(m1)[[2]] \* x, aes(color = "blue"), size = 1 ) +

# alternative for linear is geom\_smooth(method = "lm", se = FALSE, aes(color = "blue") )

geom\_function( fun = function(x) exp(coef(m2)[[1]]) \* exp(x \* coef(m2)[[2]]), aes(color = "green"), size = 1 ) +

geom\_function( fun = function(x) exp(coef(m3)[[1]]) \* x^coef(m3)[[2]], aes(color = "orange"), size = 1 ) +

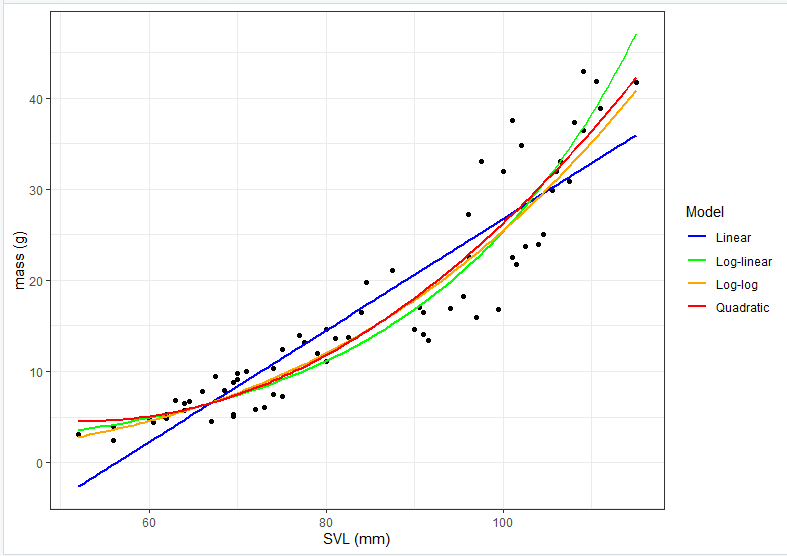
geom\_function( fun = function(x) coef(m4)[[1]] + coef(m4)[[2]] \* x + coef(m4)[[3]] \* x^2, aes(color = "red"), size = 1 ) +

labs(x = "SVL (mm)", y = "mass (g)") +

scale\_color\_identity( name = "Model",

breaks = c("blue", "green", "orange", "red"), labels = c("Linear", "Log-linear", "Log-log", "Quadratic"), guide = "legend" ) +

theme\_bw()



# I prefer m2(log-linear model)

# From the graph above, we can see the green line (log-linear) can fit the data well for small SVL and large SVL values, and it is a quadratic curve which fits the trend of data well.

# For orange(log-log) and red(quadratic) lines, we can see they fir the data well for small SVL but they don’t fit the data well for large SVL.

# For linear line, it is a straight line which does not fit the data at all comparing to the other lines.