Reason for taking log of data.

The data has high variability, multiple outliers and is greatly skewed, as seen below, which makes it unsuitable for fitting linear models. Some of the outliers will have a greater effect in their raw form than if we took the logarithm.

> library(MASS)

> data(Animals, package= "MASS")

> plot(Animals[,1], Animals[,2],xlab= "body weight", ylab= "brain weight" )

> library(moments)

> skewness(Animals[,1]) #skewness

[1] 4.760097

> skewness(Animals[,2])

[1] 3.153176

> sd(Animals$body) # variability

[1] 16480.49

> sd(Animals$brain)

[1] 1334.929

> myboxplot <- boxplot(Animals) #outliers

> myboxplot$out

[1] 11700 2547 6654 9400 87000

[6] 4603 1320 5712

**Fitting a linear regression**

> y=log(Animals)

> lmfit=lm(brain~body, y)

**SUMMARY**

> summary(lmfit)

Call:

lm(formula = brain ~ body, data = y)

Residuals:

Min 1Q Median 3Q

-3.2890 -0.6763 0.3316 0.8646

Max

2.5835

Coefficients:

Estimate Std. Error

(Intercept) 2.55490 0.41314

body 0.49599 0.07817

t value Pr(>|t|)

(Intercept) 6.184 1.53e-06 \*\*\*

body 6.345 1.02e-06 \*\*\*

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’

0.05 ‘.’ 0.1 ‘ ’ 1

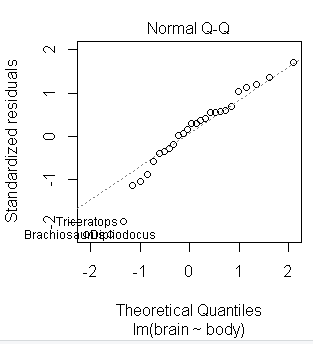
Residual standard error: 1.532 on 26 degrees of freedom

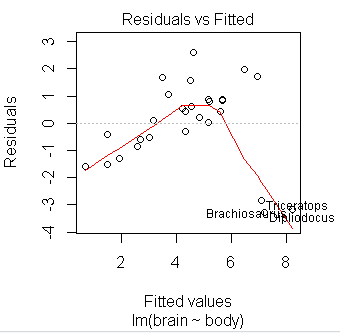
Multiple R-squared: 0.6076, Adjusted R-squared: 0.5925

F-statistic: 40.26 on 1 and 26 DF, p-value: 1.017e-06

For our linear model, we have an intercept of 2.55490 and a gradient of 0.49599. Variance of our response variable is 5.793649.

**Checking model fit**

> par(mfrow= c(1,2))

> plot(lmfit, 1:2)

* From the Residual VS fitted plot, the data seems to have a non-linear trend in the shape of a curve, which would suggest absence of homscedasticity, but we can clearly see that the three data points, the Brachiosaurus, Dipliodocus and triceratops are outliers on both plots.
* There is evidence to suggest that the residual follows a normal distribution because a majority of the data in the Q-Q plot lies on the QQ line.

**Identifying and excluding outliers then refitting regression model.**

> p=Animals[,1]

> order(p) # identifying the outliers

[1] 20 19 27 25 5 1 21 11

[9] 17 10 4 18 3 24 22 14

[17] 23 8 28 13 2 9 12 7

[25] 15 16 6 26

> z=y[c(-6,-16,-26),]

> lmfit2=lm(brain~body,z) #fitting the data

> summary(lmfit2)

Call:

lm(formula = brain ~ body, data = z)

Residuals:

Min 1Q Median 3Q

-0.9125 -0.4752 -0.1557 0.1940

Max

1.9303

Coefficients:

Estimate Std. Error

(Intercept) 2.15041 0.20060

body 0.75226 0.04572

t value Pr(>|t|)

(Intercept) 10.72 2.03e-10 \*\*\*

body 16.45 3.24e-14 \*\*\*

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’

0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7258 on 23 degrees of freedom

Multiple R-squared: 0.9217, Adjusted R-squared: 0.9183

F-statistic: 270.7 on 1 and 23 DF, p-value: 3.243e-14

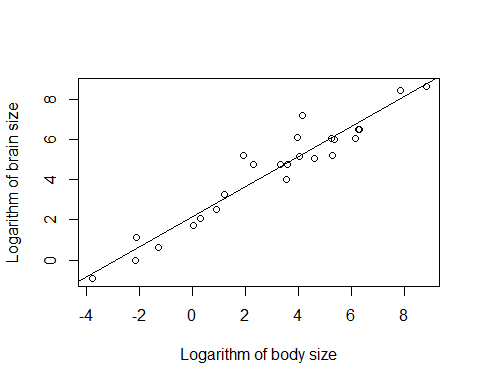
**Data plot**

> intercept=2.15041

> gradient=0.75226

> plot(z$body,z$brain, xlab= "body", ylab="brain")

> abline(intercept, gradient)



The plot show a high degree of correlation between the average weight of the brain and the body. This is backed by the result below.

cor(z)

body brain

body 1.0000000 0.9600516

brain 0.9600516 1.0000000

There is strong evidence to suggest that the weight of the brain is dependent on the body weight

**Finding confidence level from regression model.**

> a=data.frame(body=log(500))

> camel.brain.weighta=exp(predict(lmfit2, a, interval="confidence"))

> camel.brain.weight

fit lwr upr

1 920.9609 601.8355 1409.304