**Simple Linear Regression Lab – Working with Messy Data**

**Data Set: Weight**

The dataset, ***Weight***, contains the brain weights (y) and body weights (x) of 32 species.

1. Use statistical software to

a) create and interpret a scatterplot of **brain weights (y) versus body weights (x),**

> brain = c(5712,44.5,4603,179.5,0.3,25.6,440,423,419,680,115,1,406,325,119.5,4,5.5,655,157,0.25,1320,0.4,180,12.1,39.2,1.9,50.4,179,175,3.9)

> body = c(6654,3.385,2547,10.55,0.023,3.3,52.16,46,5187.1,529,27.66,0.12,207,85,36.33,0.101,1.04,521,100,0.01,62,0.023,192,2.5,4.288,0.28,4.235,6.8,55.5,3.5)

> plot(body,brain, ylab="brain weight", xlab="body weight", main="scatter plot")



> cor(body,brain)

[1] 0.7580949

1. find the least squares regression equation for this relationship,

> lm(brain~body)

Call:

lm(formula = brain ~ body)

Coefficients:

(Intercept) body

196.9876 0.6344

Y = 196.9876+0.6344X

Brian weight = 196.9876 + 0.6344 Body weight

1. plot residuals versus the explanatory variable, and

> plot(body,result$res)



1. construct a QQ plot of the residuals.

> qqnorm(result$res)

> qqline(result$res)



> summary(result)

Call:

lm(formula = brain ~ body)

Residuals:

Min 1Q Median 3Q Max

-3068.72 -192.83 -121.11 76.79 2790.18

Coefficients:

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

(Intercept) 196.9876 166.6541 1.182 0.247

body 0.6344 0.1031 6.151 1.22e-06 \*\*\*

---

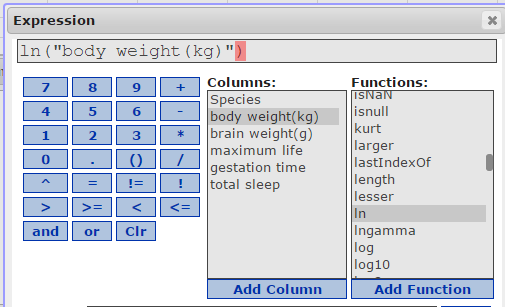
Residual standard error: 859.4 on 28 degrees of freedom

Multiple R-squared: 0.5747, Adjusted R-squared: 0.5595

F-statistic: 37.84 on 1 and 28 DF, p-value: 1.22e-06

2. We are now going to try a transformation of both the explanatory and response variable to create a better **linear** regression model.

i) Compute the natural log, ln, of both the brain weights (y) versus body weights (x) and store the results in StatCrunch. If using StatCrunch to transform the variables use DATA->COMPUTE->EXPRESSION then build – as shown in class.



ii) repeat the steps from above

a) create and interpret a scatterplot of **ln (brain weights) versus ln(body weights)** ,

a) find the least squares regression equation for this relationship,

b) plot residuals versus the explanatory variable, and

c) construct a QQ plot of the residuals.

Comment on your findings and compare the results to those found in question #1.

> brain = c(5712,44.5,4603,179.5,0.3,25.6,440,423,419,680,115,1,406,325,119.5,4,5.5,655,157,0.25,1320,0.4,180,12.1,39.2,1.9,50.4,179,175,3.9)

> body = c(6654,3.385,2547,10.55,0.023,3.3,52.16,46,5187.1,529,27.66,0.12,207,85,36.33,0.101,1.04,521,100,0.01,62,0.023,192,2.5,4.288,0.28,4.235,6.8,55.5,3.5)

> brain1 = log(brain)

> brain1

[1] 8.6503245 3.7954892 8.4344635 5.1901752 -1.2039728 3.2425924 6.0867747

[8] 6.0473722 6.0378709 6.5220928 4.7449321 0.0000000 6.0063532 5.7838252

[15] 4.7833164 1.3862944 1.7047481 6.4846352 5.0562458 -1.3862944 7.1853870

[22] -0.9162907 5.1929569 2.4932055 3.6686767 0.6418539 3.9199912 5.1873858

[29] 5.1647860 1.3609766

> body1= log(body)

> body1

[1] 8.80297346 1.21935391 7.84267147 2.35612586 -3.77226106 1.19392247

[7] 3.95431592 3.82864140 8.55393005 6.27098843 3.31998733 -2.12026354

[13] 5.33271879 4.44265126 3.59264385 -2.29263476 0.03922071 6.25575004

[19] 4.60517019 -4.60517019 4.12713439 -3.77226106 5.25749537 0.91629073

[25] 1.45582042 -1.27296568 1.44338333 1.91692261 4.01638302 1.25276297

> plot(body1,brain1)



> lm(formula = brain1 ~ body1)

Call:

lm(formula = brain1 ~ body1)

Coefficients:

(Intercept) body1

2.2414 0.7284

> result = lm(formula = brain1 ~ body1)

> plot(body1, result$res)



> qqnorm(result$res)

> qqline(result$res)



**In addition to brain weight and body weight, the *Weight* data set contains information on lifespan, gestation, and hours of sleep per day.**

3. Use statistical software to

a) create and interpret a scatterplot of **lifespan versus body weight**

a) find the least squares regression equation for this relationship,

b) plot residuals versus the explanatory variable, and

c) construct a QQ plot of the residuals.

Comment on your findings.

> life = c(38.6,14,69,27,19,28,50,30,40,28,20,3.9,39.3,41,16.2,9,7.6,46,22.4,24,100,3.2,27,18,13.7,4.7,9.8,29,20,3)

> body = c(6654,3.385,2547,10.55,0.023,3.3,52.16,46,5187.1,529,27.66,0.12,207,85,36.33,0.101,1.04,521,100,0.01,62,0.023,192,2.5,4.288,0.28,4.235,6.8,55.5,3.5)

> plot(body,life)



> result = lm (fromula = life ~ body)

Error in terms.formula(formula, data = data) :

argument is not a valid model

> result=lm(formula=life~body)

> result

Call:

lm(formula = life ~ body)

Coefficients:

(Intercept) body

24.559063 0.003955

> plot(body,result$res)



> qqnorm(result$res)

> qqline(result$res)



> summary(result)

Call:

lm(formula = life ~ body)

Residuals:

Min 1Q Median 3Q Max

-21.573 -11.924 -4.724 4.168 75.196

Coefficients:

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

(Intercept) 24.559063 3.924257 6.258 9.16e-07 \*\*\*

body 0.003955 0.002429 1.628 0.115

---

Signif. codes: 0 ‚Äò\*\*\*‚Äô 0.001 ‚Äò\*\*‚Äô 0.01 ‚Äò\*‚Äô 0.05 ‚Äò.‚Äô 0.1 ‚Äò ‚Äô 1

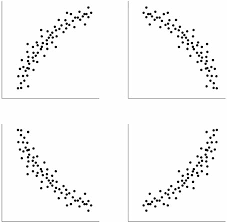
Residual standard error: 20.24 on 28 degrees of freedom

Multiple R-squared: 0.08651, Adjusted R-squared: 0.05389

F-statistic: 2.652 on 1 and 28 DF, p-value: 0.1146

4. Try various transformations of **lifespan versus body weight** to create a better regression model.

The chart below illustrates some common transformations.



Square root of y,

1/y, log(y), x2 , x3

1/

y2, y3 , x2 , x3

Square root of x, 1/x, log(x), y2 , y3

Square root of y, 1/y, log(y), square root of x, 1.x, log(x)

i) repeat the steps from above for **ONE** transformation that you think works best

a) create and interpret a scatterplot of the transformed variable(s) for lifespan and body weight,

a) find the least squares regression equation for this relationship,

b) plot residuals versus the explanatory variable, and

c) construct a QQ plot of the residuals.

Comment on your findings and compare the results to those found in question #3.

> life = c(38.6,14,69,27,19,28,50,30,40,28,20,3.9,39.3,41,16.2,9,7.6,46,22.4,24,100,3.2,27,18,13.7,4.7,9.8,29,20,3)

> body = c(6654,3.385,2547,10.55,0.023,3.3,52.16,46,5187.1,529,27.66,0.12,207,85,36.33,0.101,1.04,521,100,0.01,62,0.023,192,2.5,4.288,0.28,4.235,6.8,55.5,3.5)

> body1 = log(body)

> body1

[1] 8.80297346 1.21935391 7.84267147 2.35612586 -3.77226106 1.19392247

[7] 3.95431592 3.82864140 8.55393005 6.27098843 3.31998733 -2.12026354

[13] 5.33271879 4.44265126 3.59264385 -2.29263476 0.03922071 6.25575004

[19] 4.60517019 -4.60517019 4.12713439 -3.77226106 5.25749537 0.91629073

[25] 1.45582042 -1.27296568 1.44338333 1.91692261 4.01638302 1.25276297

> lot(body1,life)

Error: could not find function "lot"

> plot(body1,life)



> result = lm(life~body1)

> result

Call:

lm(formula = life ~ body1)

Coefficients:

(Intercept) body1

18.279 3.412

> plot(body1,result$res)



> qqnorm(result$res)

> qqline(result$res)



> summary(result)

Call:

lm(formula = life ~ body1)

Residuals:

Min 1Q Median 3Q Max

-19.553 -9.688 -5.275 5.281 67.639

Coefficients:

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

(Intercept) 18.2785 3.8245 4.779 5.07e-05 \*\*\*

body1 3.4121 0.8893 3.837 0.00065 \*\*\*

---

Signif. codes: 0 ‚Äò\*\*\*‚Äô 0.001 ‚Äò\*\*‚Äô 0.01 ‚Äò\*‚Äô 0.05 ‚Äò.‚Äô 0.1 ‚Äò ‚Äô 1

Residual standard error: 17.14 on 28 degrees of freedom

Multiple R-squared: 0.3446, Adjusted R-squared: 0.3212

F-statistic: 14.72 on 1 and 28 DF, p-value: 0.0006498