c2-worksheet Linear Regression with Transformed Variables

1. Download the data set **mammals** by calling for **library("MASS")**

a. Enter **mammals** and describe the data set. What are the units for the body and brain columns? *Hint: Look at the data for humans and ask if those values are in lbs, oz, kg, or grams.*

Units for the body column would be kg and the units for the brain column would be grams

```
# load a built in library in R
library("MASS")

# shows the first few rows of a data frame
head(mammals)
```

```
##                    body brain
## Arctic fox        3.385  44.5
## Owl monkey        0.480  15.5
## Mountain beaver   1.350   8.1
## Cow             465.000 423.0
## Grey wolf        36.330 119.5
## Goat             27.660 115.0
```

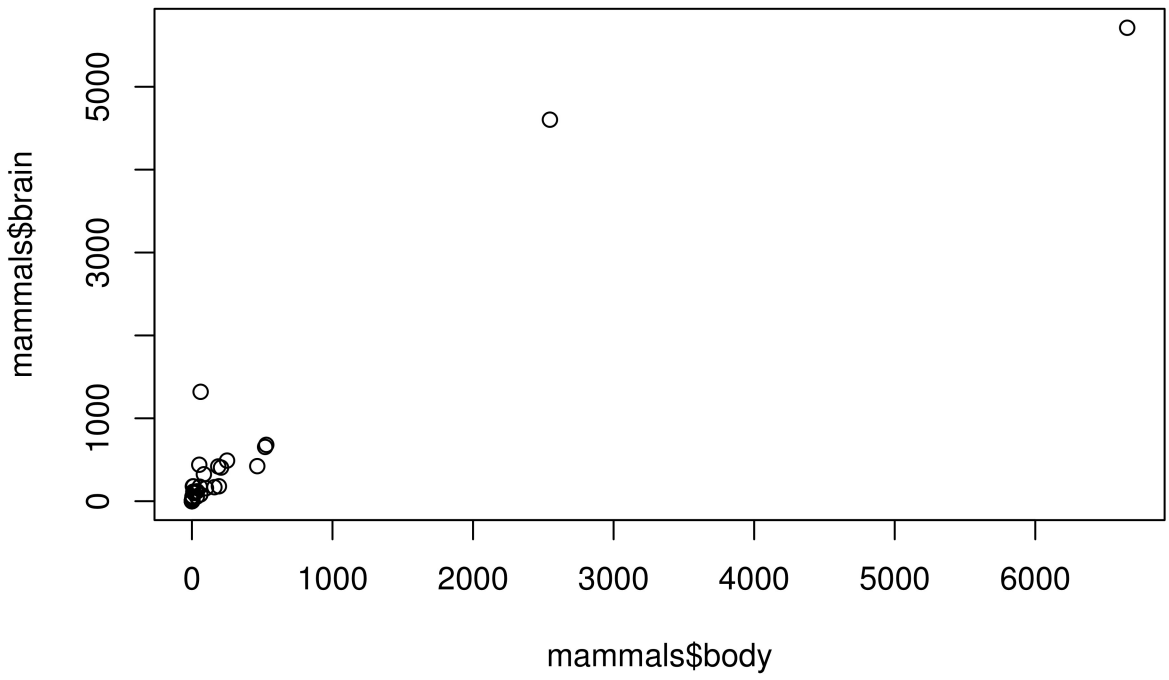
```
# you can also access rows by their name
mammals["Human",]
```

```
##       body brain
## Human   62  1320
```

b. Plot the data with **plot(mammals)** and describe the plot.

The plot doesn't look linear on first glance. All the data points on the graph are conglomerated to the bottom left corner with a couple of outliers.

```
plot(mammals$body, mammals$brain)
```



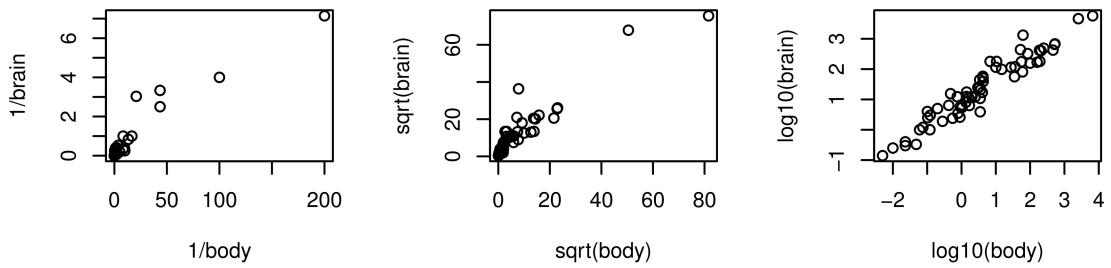
c. Below are scatterplots showing different transforms for the data. Which one is appropriate for linear regression? Explain why

The third plot (the log-log) plot is most appropriate for our linear regression since we can clearly see a linear relationship in our data. Also, this linear graph shows us a very strong positive correlation in the data we have.

```
par(mfrow=c(1,3))
# Reciprocal transform: 1/body vs 1/brain
plot(1/mammals$body,1/mammals$brain,xlab="1/body",ylab="1/brain")

# Square roots: sqrt(body) vs sqrt(brain)
plot(sqrt(mammals$body),sqrt(mammals$brain),xlab="sqrt(body)",ylab="sqrt(brain)")

# Logrithmic:  log10(body) vs log10(brain)
plot(log10(mammals$body),log10(mammals$brain),xlab="log10(body)",ylab="log10(brain)")
```



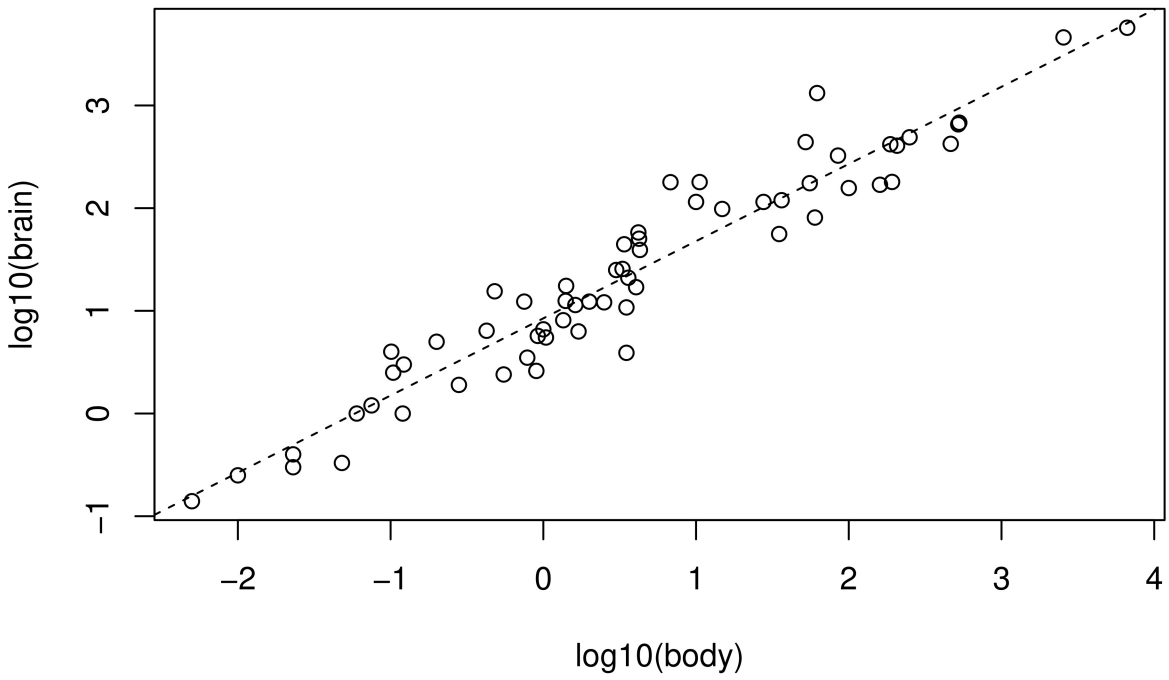
d. Give the coefficients in the regression line of the transformed variables. Do a fresh scatterplot using your chosen transform and add the regression line that plot.

EQUATION OF THE REGRESSION LINE: tbrain = 0.92713 + (tbody)0.75169

```
# The first line here binds the two new columns of the transformed data to the mammals data
 mammals <- cbind(mammals,tbody=log10(mammals$body),tbrain=log10(mammals$brain))
 mammals.lm <- lm(tbrain ~ tbody, data=mammals)
 summary(mammals.lm)
```

```
##
## Call:
## lm(formula = tbrain ~ tbody, data = mammals)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.74503 -0.21380 -0.02676  0.18934  0.84613
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.92713    0.04171   22.23   <2e-16 ***
## tbody        0.75169    0.02846   26.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3015 on 60 degrees of freedom
## Multiple R-squared:  0.9208, Adjusted R-squared:  0.9195
## F-statistic: 697.4 on 1 and 60 DF,  p-value: < 2.2e-16
```

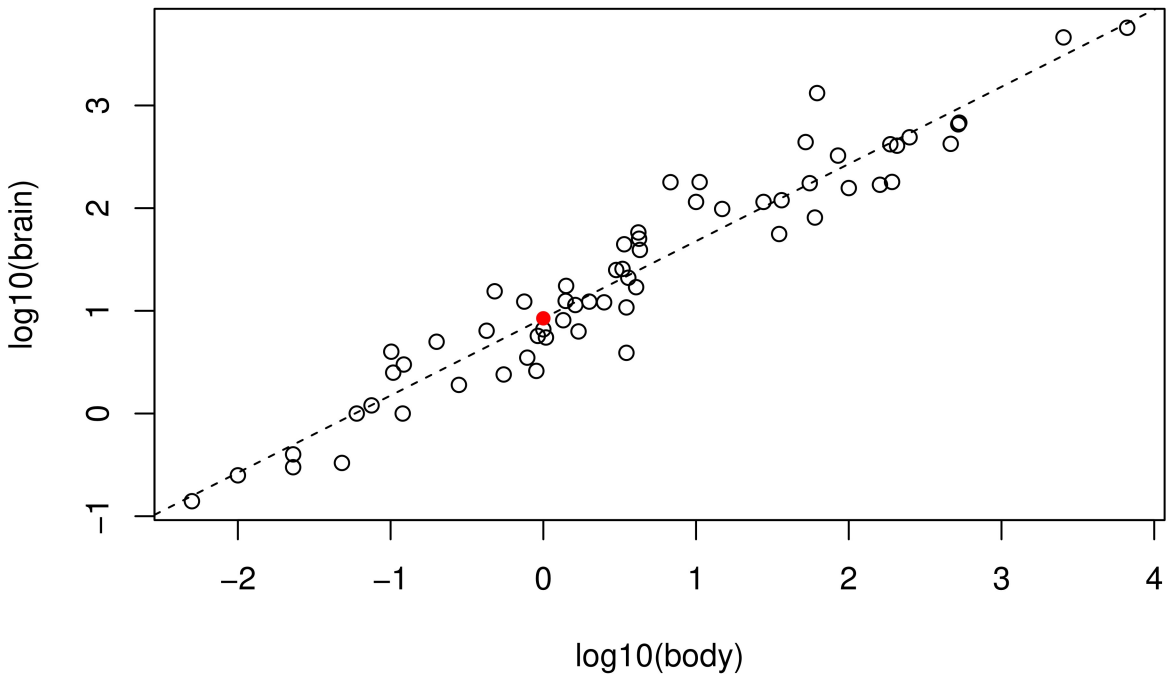
```
 plot(log10(mammals$body),log10(mammals$brain),xlab="log10(body)",ylab="log10(brain)")
 abline(mammals.lm,lty=2)
```



e. For an animal weighs 1.0 kg, what does the model predict for the mass of that animal's brain? Add the predicted value to the plot above. *Hint: What is the value of 1kg after doing the transformation you selected above? Use this to find the predicted brain mass.*

According to the equation used below to calculate the estimate which lies on the regression line, we know that log10(prediction) = 0.92719. This means the estimated prediction for the weight of an animal that weighs 1 kg is 10^(0.92719) = 8.45532 grams

```
# Don't forget to answer the question as well as produce the plot
 newmass <- log10(1)
 newpredbrain <- 0.92713  # 0.92719 + (0.75168)(0) = 0.92719
 plot(log10(mammals$body),log10(mammals$brain),xlab="log10(body)",ylab="log10(brain)")
 abline(mammals.lm,lty=2)
 points(newmass, newpredbrain, col="red", pch=16)
```



f. Find the 3 animals with the most negative residuals and the 3 animals with the most positive residuals. What do these groups of animals have in common?

```
mammals.res <- residuals(mammals.lm)
names(mammals.res) <- rownames(mammals)
mammals.sortres <- sort(mammals.res)
head(mammals.sortres, n=3)
```

```
## Water opossum        Tenrec    Musk shrew
##    -0.7450306    -0.4777583    -0.4173206
```

```
tail(mammals.sortres, n=3)
```

```
##        Baboon Rhesus monkey         Human
##     0.5577730     0.6999408     0.8461314
```

The animals with the most negative residuals: - Water opossum - Tenrec - Musk shrew

The animals with the most positive residuals: - Baboon - Rhesus monkey - Human

The animals with the most negative residuals are all members of the Rodent family whereas the animals with the most positive residuals are Primates which could be considered as the smartest family of animals on the planet.