MAST30025: Linear Statistical Models

Solutions to Week 6 Lab

1. In this question we consider the hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}^*$. The test statistic for this hypothesis is

$$\frac{(\mathbf{b}-\boldsymbol{\beta}^*)^T X^T X(\mathbf{b}-\boldsymbol{\beta}^*)/p}{SS_{Res}/(n-p)}.$$

(a) Show that

$$(\mathbf{b}-\boldsymbol{\beta}^*)^T X^T X(\mathbf{b}-\boldsymbol{\beta}^*) = (\mathbf{y}-X\boldsymbol{\beta}^*)^T(\mathbf{y}-X\boldsymbol{\beta}^*) - (\mathbf{y}-X\mathbf{b})^T(\mathbf{y}-X\mathbf{b}).$$

That is, it is the SS_{Res} for the null model minus the SS_{Res} for the full model.
Also show that

$$(\mathbf{b}-\boldsymbol{\beta}^*)^T X^T X(\mathbf{b}-\boldsymbol{\beta}^*) \neq \mathbf{y}^T X(X^T X)^{-1} X^T \mathbf{y} - \boldsymbol{\beta}^* X^T X \boldsymbol{\beta}^*.$$

That is, in this case we can not write it as the SS_{Reg} for the full model minus the SS_{Reg} for the model under H_0.
Solution: Consider the LHS first

$$\begin{aligned}
&(\mathbf{b}-\boldsymbol{\beta}^*)^T X^T X(\mathbf{b}-\boldsymbol{\beta}^*) \\
&= ((X^T X)^{-1} X^T \mathbf{y} - \boldsymbol{\beta}^*)^T X^T X((X^T X)^{-1} X^T \mathbf{y} - \boldsymbol{\beta}^*) \\
&= \mathbf{y}^T X(X^T X)^{-1} X^T X(X^T X)^{-1} X^T \mathbf{y} - 2\boldsymbol{\beta}^{*T} X^T X(X^T X)^{-1} X^T \mathbf{y} + \boldsymbol{\beta}^{*T} X^T X \boldsymbol{\beta}^* \\
&= \mathbf{y}^T X(X^T X)^{-1} X^T \mathbf{y} - 2\boldsymbol{\beta}^{*T} X^T \mathbf{y} + \boldsymbol{\beta}^{*T} X^T X \boldsymbol{\beta}^*.
\end{aligned}$$

Now for the RHS

$$(\mathbf{y}-X\boldsymbol{\beta}^*)^T(\mathbf{y}-X\boldsymbol{\beta}^*) = \mathbf{y}^T\mathbf{y} - 2\boldsymbol{\beta}^{*T} X^T \mathbf{y} + \boldsymbol{\beta}^{*T} X^T X \boldsymbol{\beta}^*$$

and

$$\begin{aligned}
(\mathbf{y}-X\mathbf{b})^T(\mathbf{y}-X\mathbf{b}) &= (\mathbf{y}-X(X^T X)^{-1} X^T \mathbf{y})^T(\mathbf{y}-X(X^T X)^{-1} X^T \mathbf{y}) \\
&= \mathbf{y}^T(I - X(X^T X)^{-1} X^T)^T(I - X(X^T X)^{-1} X^T)\mathbf{y} \\
&= \mathbf{y}^T(I - X(X^T X)^{-1} X^T)\mathbf{y} \\
&= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T X(X^T X)^{-1} X^T \mathbf{y}.
\end{aligned}$$

Thus the RHS equals the LHS.
We also see that the only way to get $(\mathbf{b}-\boldsymbol{\beta}^*)^T X^T X(\mathbf{b}-\boldsymbol{\beta}^*) = \mathbf{y}^T X(X^T X)^{-1} X^T \mathbf{y} - \boldsymbol{\beta}^* X^T X \boldsymbol{\beta}^*$ is to have $\boldsymbol{\beta}^{*T} X^T \mathbf{y} = \boldsymbol{\beta}^{*T} X^T X \boldsymbol{\beta}^*$, which will only hold in general if $\boldsymbol{\beta}^* = 0$.

(b) Show directly that $(\mathbf{b}-\boldsymbol{\beta}^*)^T X^T X(\mathbf{b}-\boldsymbol{\beta}^*)$ and SS_{Res} are independent, that is without using our existing results that $\mathbf{b}$ and SS_{Res} are independent.
Hint: set $\mathbf{q} = \mathbf{y} - X\boldsymbol{\beta}^*$ then

i. Show that $(\mathbf{b}-\boldsymbol{\beta}^*)^T X^T X(\mathbf{b}-\boldsymbol{\beta}^*) = \mathbf{q}^T X(X^T X)^{-1} X^T \mathbf{q}$.
ii. Show that $SS_{Res} = \mathbf{q}^T[I - X(X^T X)^{-1} X^T]\mathbf{q}$ and hence that these two quadratic forms are independent.

Solution: We express both quantities as quadratic forms in $\mathbf{q}$. For the first we have

$$\begin{aligned}
\mathbf{q}^T X(X^T X)^{-1} X^T \mathbf{q} &= \mathbf{y}^T X(X^T X)^{-1} X^T \mathbf{y} - (\boldsymbol{\beta}^*)^T X^T X(X^T X)^{-1} X^T \mathbf{y} \\
&\quad -\mathbf{y}^T X(X^T X)^{-1} X^T X \boldsymbol{\beta}^* + (\boldsymbol{\beta}^*)^T X^T X(X^T X)^{-1} X^T X \boldsymbol{\beta}^* \\
&= \mathbf{b}^T X^T X\mathbf{b} - (\boldsymbol{\beta}^*)^T X^T X\mathbf{b} - \mathbf{b}^T X^T X\boldsymbol{\beta}^* + (\boldsymbol{\beta}^*)^T X^T X\boldsymbol{\beta}^* \\
&= (\mathbf{b}-\boldsymbol{\beta}^*)^T X^T X(\mathbf{b}-\boldsymbol{\beta}^*).
\end{aligned}$$

For the SS_{Res} note first that

$$\begin{aligned}\boldsymbol{\beta}^{*T}X^T[I - X(X^TX)^{-1}X^T]X\boldsymbol{\beta}^* &= \boldsymbol{\beta}^{*T}X^TX\boldsymbol{\beta}^* - \boldsymbol{\beta}^{*T}X^TX(X^TX)^{-1}X^TX\boldsymbol{\beta}^* \\ &= \boldsymbol{\beta}^{*T}X^TX\boldsymbol{\beta}^* - \boldsymbol{\beta}^{*T}X^TX\boldsymbol{\beta}^* \\ &= \mathbf{0}.\end{aligned}$$

Similarly $\mathbf{y}^T[I - X(X^TX)^{-1}X^T]X\boldsymbol{\beta}^* = \mathbf{0}$ and $\boldsymbol{\beta}^{*T}X^T[I - X(X^TX)^{-1}X^T]\mathbf{y} = \mathbf{0}$, so

$$\begin{aligned}&\mathbf{q}^T[I - X(X^TX)^{-1}X^T]\mathbf{q} \\ &= \mathbf{y}^T[I - X(X^TX)^{-1}X^T]\mathbf{y} - \boldsymbol{\beta}^{*T}X^T[I - X(X^TX)^{-1}X^T]\mathbf{y} \\ &\quad -\mathbf{y}^T[I - X(X^TX)^{-1}X^T]X\boldsymbol{\beta}^* + \boldsymbol{\beta}^{*T}X^T[I - X(X^TX)^{-1}X^T]X\boldsymbol{\beta}^* \\ &= \mathbf{y}^T[I - X(X^TX)^{-1}X^T]\mathbf{y} \\ &= SS_{Res}.\end{aligned}$$

Finally, we know that var $\mathbf{q} = \sigma^2 I$, so, using our theorem for the independence of quandratic forms

$$\begin{aligned}AVB &= X(X^TX)^{-1}X^T\sigma^2 I[I - X(X^TX)^{-1}X^T] \\ &= \sigma^2\left(X(X^TX)^{-1}X^T - X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T\right) \\ &= \sigma^2\left(X(X^TX)^{-1}X^T - X(X^TX)^{-1}X^T\right) \\ &= \mathbf{0}\end{aligned}$$

as required.

How can you tell which is the response and design variable?

The remaining questions use the 'sleep' dataset, which you can download from the course website. This dataset contains (among other things) data on the body weight (kg) and brain weight (g) of 62 mammals. Use the following commands to read the data:

```
> mammals <- read.csv("../data/sleep.csv")
> mammals$BodyWt <- log(mammals$BodyWt)
> mammals$BrainWt <- log(mammals$BrainWt)
```

This creates a data frame, `mammals`, with components (among others) named `BodyWt` and `BrainWt`, then applies a logarithmic transformation to both `BodyWt` and `BrainWt`.

2. Fit a linear model explaining brain weight from body weight, using the `lm` command.

   Display the summary of the fitted model, and then create a scatter plot of the data and superimpose the fitted regression line on it. Does it look like a reasonable fit?

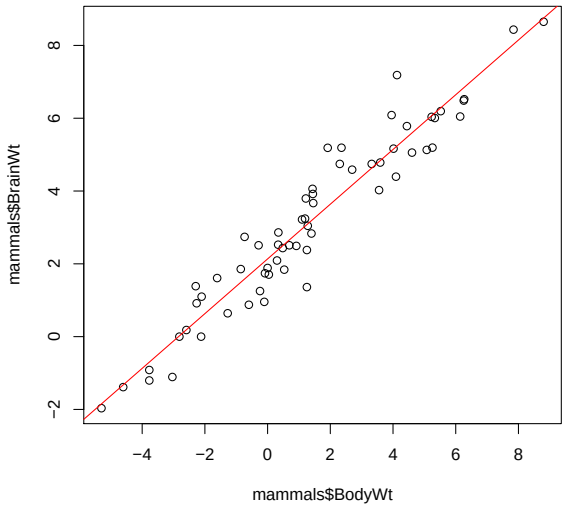
   Use diagnostic plots to assess if the model assumptions are satisfied.

   Solution:

```
> model <- lm(BrainWt ~ BodyWt, data = mammals)
> summary(model)

Call:
lm(formula = BrainWt ~ BodyWt, data = mammals)

Residuals:
     Min       1Q   Median       3Q      Max
-1.71550 -0.49228 -0.06162  0.43597  1.94829

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.13479    0.09604   22.23   <2e-16 ***
BodyWt       0.75169    0.02846   26.41   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

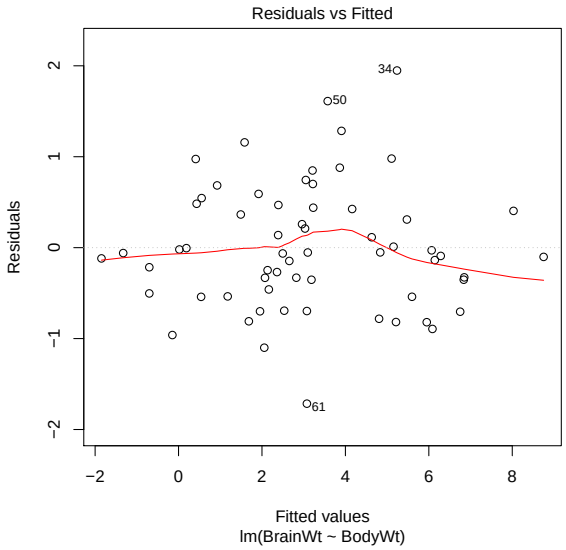
```
Residual standard error: 0.6943 on 60 degrees of freedom
Multiple R-squared:  0.9208,        Adjusted R-squared:  0.9195
F-statistic: 697.4 on 1 and 60 DF,  p-value: < 2.2e-16
```

```
> plot(mammals$BodyWt, mammals$BrainWt)
> abline(model, col="red")
```



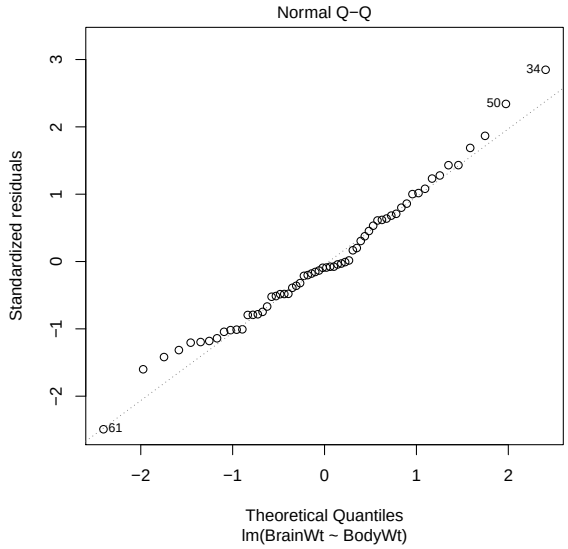
The fit is good.

```
> plot(model, which=1)
```



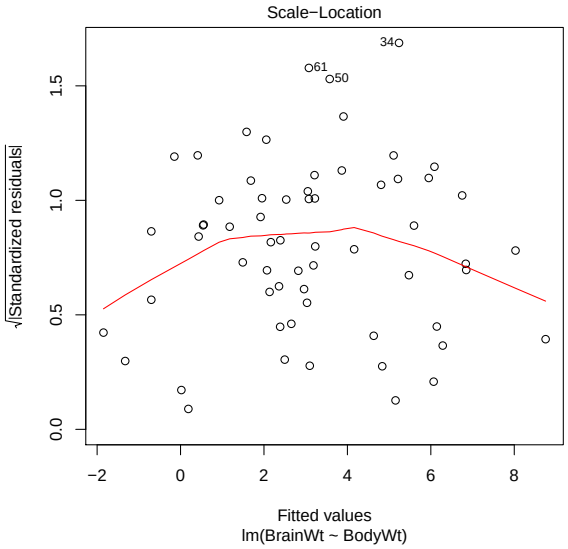
The residuals show a slight trend toward negativity as the fitted values increase, but not enough to be a problem.

```
> plot(model, which=2)
```



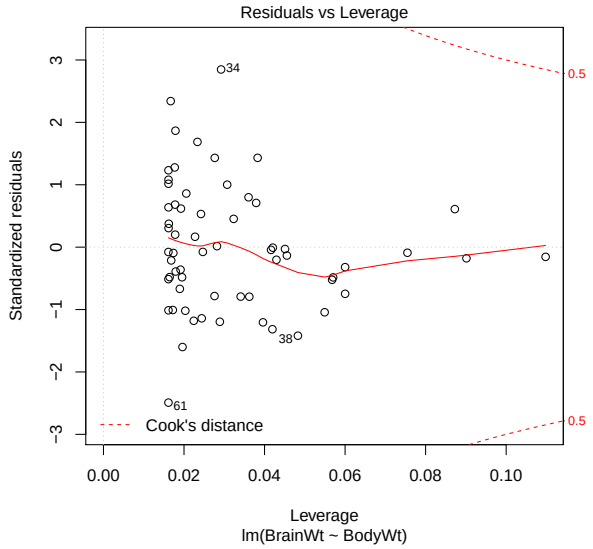
The Q-Q plot looks reasonably linear.

```
> plot(model, which=3)
```



The standardised residuals get smaller on both sides of the plot. This is not ideal, but the lack of a definite trend makes it difficult to correct.

```
> plot(model, which=5)
```



This plot is fine.

3. Using the fitted model or otherwise, obtain:

   (a) The least squares estimator of the parameters, $\mathbf{b}$;

   (b) The vector of residuals, $\mathbf{e}$;

   (c) The residual sum of squares, SS_{Res};

   (d) The regression sum of squares, SS_{Reg};

   (e) The estimator for the variance of the errors, s^2;

   (f) The standardised residuals;

   (g) The leverages of the points; and

   (h) The Cook's distances of the points.

Solution:

```
> model$coefficients # parameter estimates

(Intercept)      BodyWt
  2.1347887   0.7516859

> str(model$residuals) # first few residuals

 Named num [1:62] -0.102 -0.248 0.744 -0.332 0.404 ...
 - attr(*, "names")= chr [1:62] "1" "2" "3" "4" ...

> deviance(model) # residual sum of squares

[1] 28.92271

> sum(mammals$BrainWt^2) - deviance(model) # regression ss

[1] 947.5602

> deviance(model)/model$df.residual # sample variance

[1] 0.4820452

> str(rstandard(model)) # standardised residuals
```

```
 Named num [1:62] -0.155 -0.36 1.081 -0.482 0.61 ...
 - attr(*, "names")= chr [1:62] "1" "2" "3" "4" ...
```

```
> str(lm.influence(model)$hat) # leverages
```

```
 Named num [1:62] 0.1098 0.0191 0.0162 0.0195 0.0873 ...
 - attr(*, "names")= chr [1:62] "1" "2" "3" "4" ...
```

```
> str(cooks.distance(model)) # cook's distances
```

```
 Named num [1:62] 0.00148 0.00127 0.00958 0.00232 0.01777 ...
 - attr(*, "names")= chr [1:62] "1" "2" "3" "4" ...
```

4. Find a 95% confidence interval for a mammal weighing 50 kg.

   Solution:

```
> predict(model, data.frame(BodyWt = log(50)), interval = "confidence", level = 0.95)
```

```
       fit      lwr      upr
1 5.075401 4.846066 5.304736
```

5. Find a 95% prediction interval for a mammal weighing 50 kg.

   Solution:

```
> predict(model, data.frame(BodyWt = log(50)), interval = "prediction", level = 0.95)
```

```
       fit      lwr      upr
1 5.075401 3.667797 6.483006
```

6. Test the following hypotheses, using the `anova` function.

   (a) $H_0 : \boldsymbol{\beta} = 0$
   (b) $H_0 : \beta_1 = 0$
   (c) $H_0 : \beta_0 = 0$
   (d) $H_0 : \boldsymbol{\beta} = (2, 1)$

   Solution:

```
> null <- lm(BrainWt ~ 0, data = mammals)
> anova(null, model)

Analysis of Variance Table

Model 1: BrainWt ~ 0
Model 2: BrainWt ~ BodyWt
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     62 976.48
2     60  28.92  2    947.56 982.85 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> null <- lm(BrainWt ~ 1, data = mammals)
> anova(null, model)

Analysis of Variance Table

Model 1: BrainWt ~ 1
Model 2: BrainWt ~ BodyWt
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     61 365.11
2     60  28.92  1    336.19 697.42 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

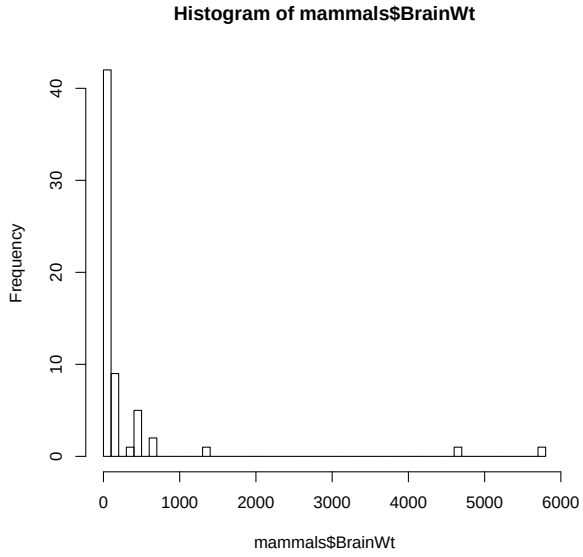
```
> null <- lm(BrainWt ~ 0 + BodyWt, data = mammals)
> anova(null, model)

Analysis of Variance Table

Model 1: BrainWt ~ 0 + BodyWt
Model 2: BrainWt ~ BodyWt
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     61 267.079
2     60  28.923  1    238.16 494.05 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> library(car)
> linearHypothesis(model,diag(2),c(2,1))

Linear hypothesis test

Hypothesis:
(Intercept) = 2
BodyWt = 1

Model 1: restricted model
Model 2: BrainWt ~ BodyWt

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     62 68.024
2     60 28.923  2    39.101 40.558 7.199e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We reject all null hypotheses.

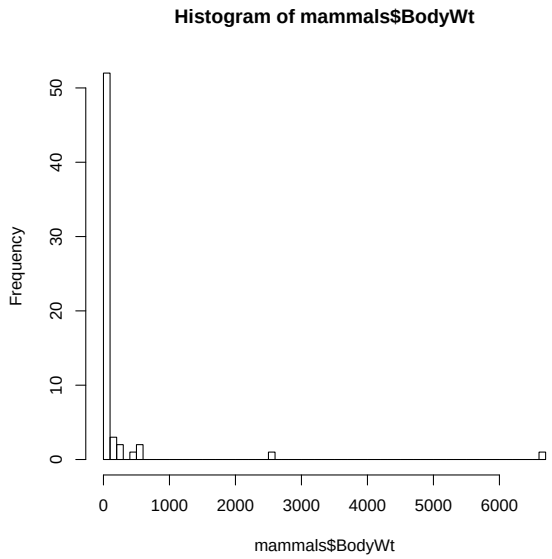
7. By visualising the raw data, justify the use of a double logarithmic transformation. Write down the final model for the (untransformed) brain weight vs. body weight.

Solution:

```
> mammals <- read.csv('../data/sleep.csv')
> hist(mammals$BrainWt,breaks=50)
```



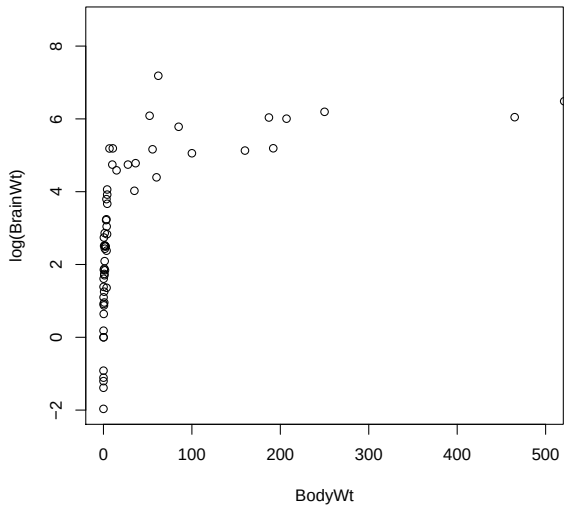
```
> hist(mammals$BodyWt,breaks=50)
```



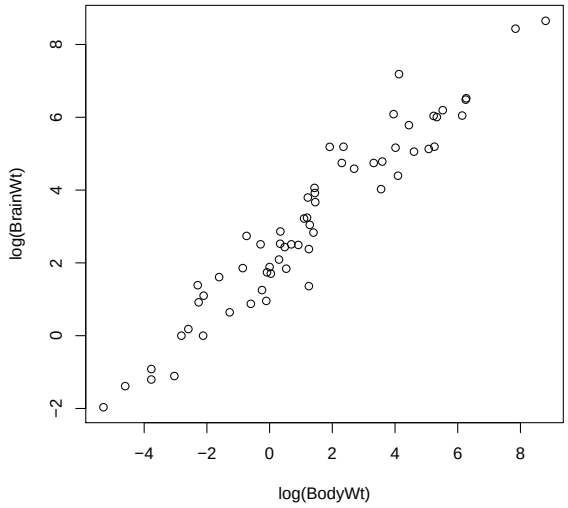
We see that both brain weight and body weight have extremely right-skewed distributions. This is one of the hallmarks of data which requires a logarithmic transformation, particularly in the response (to achieve normal errors). In addition, both variables are constrained to be positive, another indication that a logarithmic transformation may be required.

Merely being right-skewed would not be a strong enough case to transform the predictor, although the extreme nature of the skew results in some points with extremely high leverage/Cook's distance. However, transforming the brain weight alone does not result in a linear relationship, while transforming both brain and body weight results in an obviously linear relationship.

```
> plot(log(BrainWt)~BodyWt,data=mammals,xlim=c(0,500))
```




```
> plot(log(BrainWt)~log(BodyWt),data=mammals)
```



The final model is

$$\text{brain weight} = 8.46 \cdot (\text{body weight})^{0.75} \cdot \varepsilon.$$