Case: Brain

Story

The dataset `brainweight.txt` contains measurements of the weight of both brain and body for different mammals.

Data

Variable	Description
art	species of mammal
body	weigth of body (kg)
brain	weight of brain (gram)

Exercise

1. Make a scatterplot of body against brain. Do you see any association between the variables?

   Answer:

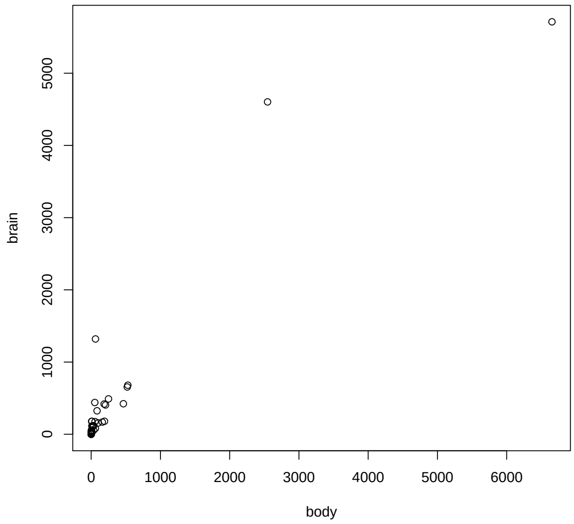
```
## Read data:
brain <- read.table("../exercises_data/brainweight.txt", header=TRUE, sep="")
str(brain)

## 'data.frame': 62 obs. of  3 variables:
##  $ art  : Factor w/ 61 levels "African.elephant",..: 3 42 36 12 26 22 52 28 59 11 ...
##  $ body : num  3.38 0.48 1.35 465 36.33 ...
##  $ brain: num  44.5 15.5 8.1 423 119.5 ...

head(brain)

##                art    body brain
## 1       Arctic.fox   3.385  44.5
## 2       Owl.Monkey   0.480  15.5
## 3 Mountain.beaver    1.350   8.1
## 4              Cow 465.000 423.0
## 5        Gray.wolf  36.330 119.5
## 6             Goat  27.660 115.0

## summary(brain)
plot(brain ~ body, data=brain)
```

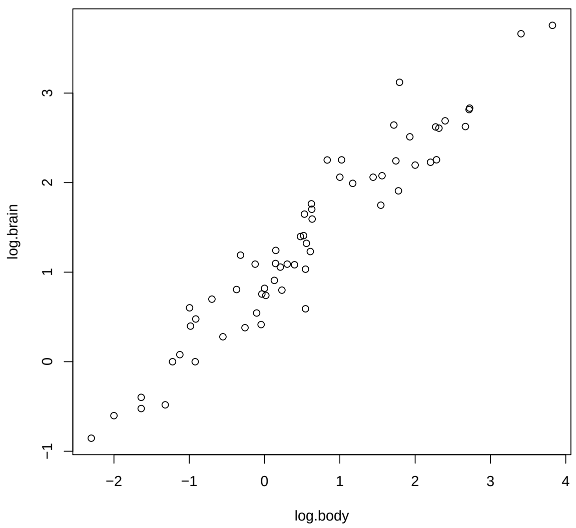


It is hard to see the association here since most data are bundled up in the lower left corner.

2. Make a log transform of both body and brain. Make a scatterplot of the transformed variables. Do you see any association now?

   Answer:

```
brain <- within(brain, {
  log.brain <- log(brain, 10)
  log.body <- log(body, 10)
})
plot(log.brain ~ log.body, data=brain)
```



Now it is much easier to see the relation between brain and body weight.

3. Fit a regression model describing `log(brain)` by `log(body)`

   Answer:

```
fm.brain <- lm(log.brain ~ log.body, data=brain)
summary(fm.brain)

##
## Call:
## lm(formula = log.brain ~ log.body, data = brain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7450 -0.2138 -0.0268  0.1893  0.8461
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.9271     0.0417    22.2   <2e-16 ***
## log.body      0.7517     0.0285    26.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.302 on 60 degrees of freedom
## Multiple R-squared:  0.921, Adjusted R-squared:  0.919
## F-statistic:  697 on 1 and 60 DF,  p-value: <2e-16
```

4. Write up the mathematical model for `log(brain)`. Then take the anti-log on both sides to obtain an expression for `brain` and simplify the equation.

What is the structure of the relation between brain and body weight? Is it linear?

Answer:

$$\begin{aligned} \log \text{brain} &= 0.93 + .75 \log \text{body} \\ \text{brain} &= 10^{0.93} \cdot \text{body}^{0.75} \\ \text{brain} &= 8.45 \cdot \text{body}^{0.75} \end{aligned}$$

where

```
10^coef(fm.brain)[1]

## (Intercept)
##       8.455
```

So the relation betwee brain and body weight is a power function; it is not linear!

5. Estimate confidence intervals for the parameters in the model for `brain`.

Answer:

```
CI <- confint(fm.brain)
CI[1, ] <- 10^CI[1, ]
round(CI, 2)

##              2.5 % 97.5 %
## (Intercept)  6.98  10.25
## log.body     0.69   0.81
```

6. Test the hypothesis that the brain-to-body ratio is constant (does not depend in body weight).

Answer:

Our model is on the form:

$$\text{brain} = a \cdot \text{body}^b$$

If $b = 1$, we may write the brain-to-body ratio as

$$\frac{\text{brain}}{\text{body}} = a$$

thus we are interested in assessing the evidence against $b = 1$. To do that we will calculate the p-value under the hypothesis that $b = 1$. Here the t-value is

$$t = \frac{\beta_1 - \beta_{\text{null}}}{\text{se}(\hat{\beta}_1)} = \frac{0.75 - 1}{0.028} = -8.72$$

And so we can calculate the p-value as $p = 2 \cdot P(T > |t|)$:

```r
B <- coef(summary(fm.brain))
(t <- (B[2, 1] - 1)/B[2, 2])

## [1] -8.724

pt(abs(t), df=60, lower.tail=FALSE)

## [1] 1.442e-12
```

Since the p-value is very small, there is strong evidence *against* a constant body-to-brain ratio.

7. Save the code that performs your analysis in a script and add plenty of comments to your code.

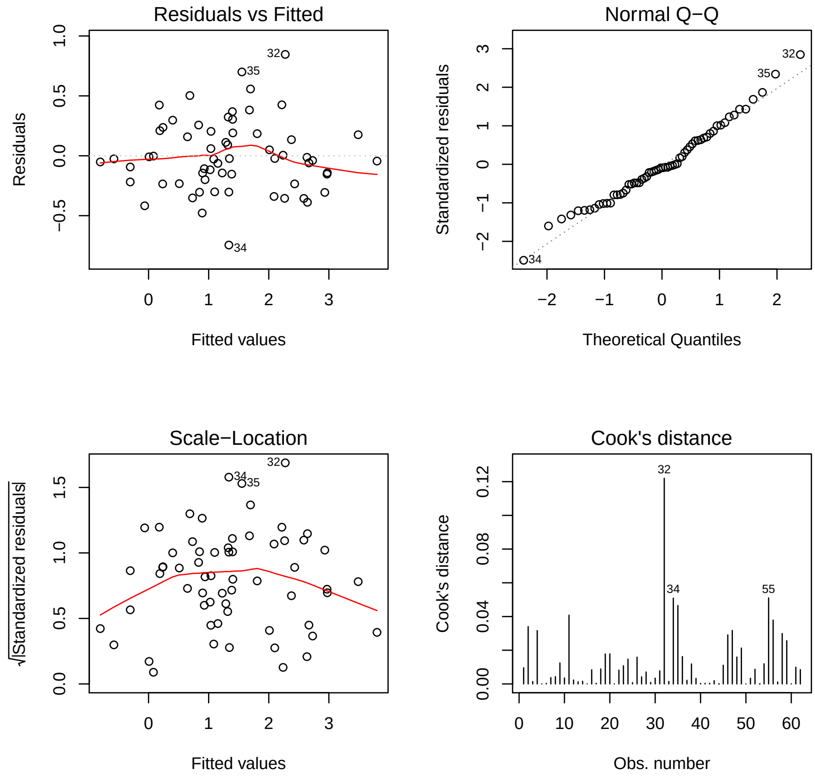
Extra

1. Are there any outliers in the data? Take a look at the observations that are the least well fitted by your model. Can you explain these observations?

   Answer:

```r
par(mfrow=c(2, 2))
fm.brain <- lm(log.brain ~ log.body, data=brain)
plot(fm.brain, which=1:4)
cbind(brain, brain.body.ratio=brain$brain / brain$body)[c(32, 34, 35), ]

##                art body  brain log.body log.brain brain.body.ratio
## 32           Human 62.0 1320.0   1.7924    3.1206           21.290
## 34 Water.opossum  3.5    3.9   0.5441    0.5911            1.114
## 35 Rhesus.monkey  6.8  179.0   0.8325    2.2529           26.324
```



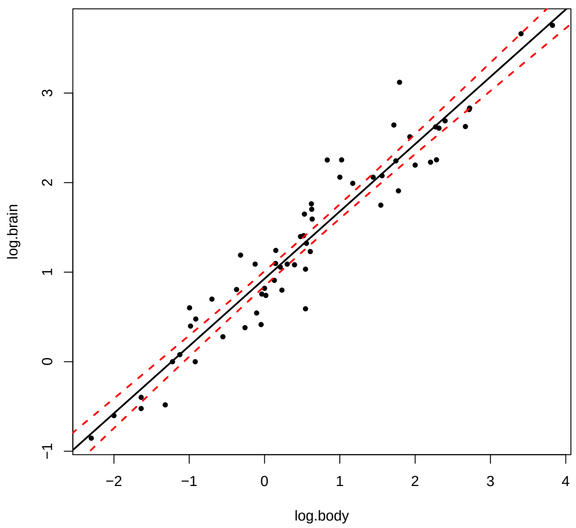
Note how humans show up here.

2. Perform model diagnostics on the model in question 3. and argue that each of the model assumptions are fulfilled.

3. Illustrate the fitted model on the data. Also quantify the model uncertainty.

   Answer:

```
par(mfrow=c(1, 1))
plot(log.brain ~ log.body, data=brain, pch=20)
abline(fm.brain, lwd=2)
xval <- seq(-3, 4.3, len=500)
pred <- predict(fm.brain, newdata=data.frame(log.body=xval),
                interval="confidence")
lines(xval, pred[, "lwr"], lty=2, col="red", lwd=2)
lines(xval, pred[, "upr"], lty=2, col="red", lwd=2)
```



4. What if a new mammal was discovered with a body weight of 5kg. What would the likely brain weight of this mammal be? To quantify the uncertainty in your estimate, provide an appropriate interval.

   Answer:

```
10^predict.lm(fm.brain, newdata=data.frame(log.body=log10(5)),
              interval="prediction", level=0.95)

##     fit  lwr upr
## 1 28.35 6.99 115
```

5. Investigate the quality of the data source and realize that one mammal appears twice. Which mammal is it and which of the entries is the right one (use google or wikipedia to get the facts)? Are there other problems with the data?

   Answer:

```
head(brain)

##                  art    body brain log.body log.brain
## 1        Arctic.fox   3.385  44.5   0.5296    1.6484
## 2        Owl.Monkey   0.480  15.5  -0.3188    1.1903
## 3 Mountain.beaver   1.350   8.1   0.1303    0.9085
## 4               Cow 465.000 423.0   2.6675    2.6263
## 5         Gray.wolf  36.330 119.5   1.5603    2.0774
## 6              Goat  27.660 115.0   1.4419    2.0607

which(table(brain$art) == 2)
```

```
## Rock.hyrax
##         51

subset(brain, art == "Rock.hyrax")

##           art body brain log.body log.brain
## 31 Rock.hyrax 0.75  12.3  -0.1249     1.090
## 50 Rock.hyrax 3.60  21.0   0.5563     1.322
```

From `http://en.wikipedia.org/wiki/Rock_hyrax` a Rock hyrax is about 4 kg, so observation 50 seems correct.