Lab Assignment 03

Beatriz Gascón Lourenço

Nov 22nd, 2018

Preamble

This assignment covers material from the following tutorials:

- Checking assumptions and data transformations
- Tutorial 17: Correlation analysis
- Tutorial 18: Regression analysis

These can all be found at the usual BIOL202 Tutorial homepage.

This assignment is worth **4% of your final grade**, and contributes to the "Lab" portion of your BIOL202 grade.

You have **one week** to complete this assignment. It must be submitted to Canvas prior to the beginning of your lab during the week of **November 19th, 2018**.

Late assignments will receive a **zero**, as will assignments that are not properly knitted.

Required packages

- `tigerstats`
- `visreg`

Be sure to load these packages!

Required data

- "booby.csv" which is described in example 16.1 in the text
- "mammals.csv", which includes the average body mass (kg) and brain mass (g) of 62 mammal species.

```
booby <-
read.csv(url("https://people.ok.ubc.ca/jpither/datasets/booby.csv"), header
= TRUE)
mammals <-
read.csv(url("https://people.ok.ubc.ca/jpither/datasets/mammals.csv"),
header = TRUE)
```

Instructions

For both questions, include in your assignment all R code in chunks, annotated in your main text (non-chunk areas) as follows:

I used a *such and such test* to test the hypothesis, and the following R code:

```
R chunk
function(~ variable, data = dataframe)
```

TIP: When constructing your answers, be sure to follow the lead of the example statements (e.g. concluding statements, interpretations of figures) in the tutorial materials. These statements are provided in larger, indented font in the tutorials.

Be sure to pay attention to the feedback you received regarding your last assignment from the teaching assistants.

Consult Tutorial_00 for instructions on how to prepare and submit you assignment.

Question 1

Use the booby data frame for this question.

Is there an association between the number of visits experienced by nestling boobies and the future behaviour of the same individuals as adults?

Hypotheses

H0 = There is no association between number of visits experienced by nestling boobies and the future behaviour of the same individuals as adults ($\rho = 0$)

HA = There is an association between number of visits experienced by nestling boobies and the future behaviour of the same individuals as adults ($\rho \neq 0$)

α level = 0.05

I used a *Pearson correlation analysis* to test the hypothesis in this question because the response variable (Y) is numerical, there is an explanatory variable (X) which is not categorical, and the goal is not to predict values of Y from X:

```
inspect (booby) #inspect the data

##
## quantitative variables:
##              name   class   min      Q1 median      Q3   max
```

```
mean
## 1 nVisitsNestling integer  1.00  8.7500   13.0 15.7500 31.00
13.12500
## 2  futureBehavior numeric -0.92 -0.2875   -0.1  0.1825  0.39 -
0.11875
##          sd  n missing
## 1 7.2069562 24       0
## 2 0.3739834 24       0
```

```
plot(nVisitsNestling ~ futureBehavior, data = booby,
     xlab = "Future behavior",
     ylab = "Number of visits",
     pch = 1,
     col = "blue",
     las = 1) # visualizing the data
```

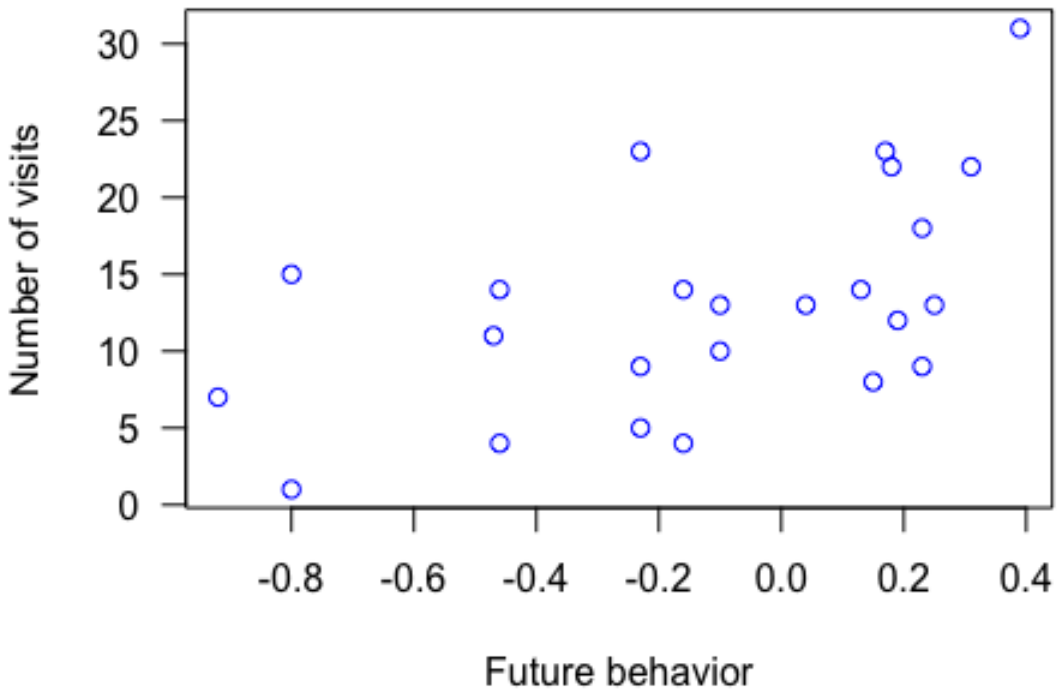

Figure 1.The association between Future behaviour and number of nestling visits (n = 24).

```
set.seed(246)

plot(jitter(nVisitsNestling,
```

```
     amount = 10) ~ futureBehavior,
  data = booby,
     xlab = "Future behavior",
     ylab = "Number of visits",
     pch = 1,
     col = "blue",
     las = 1) # visualizing the data
```

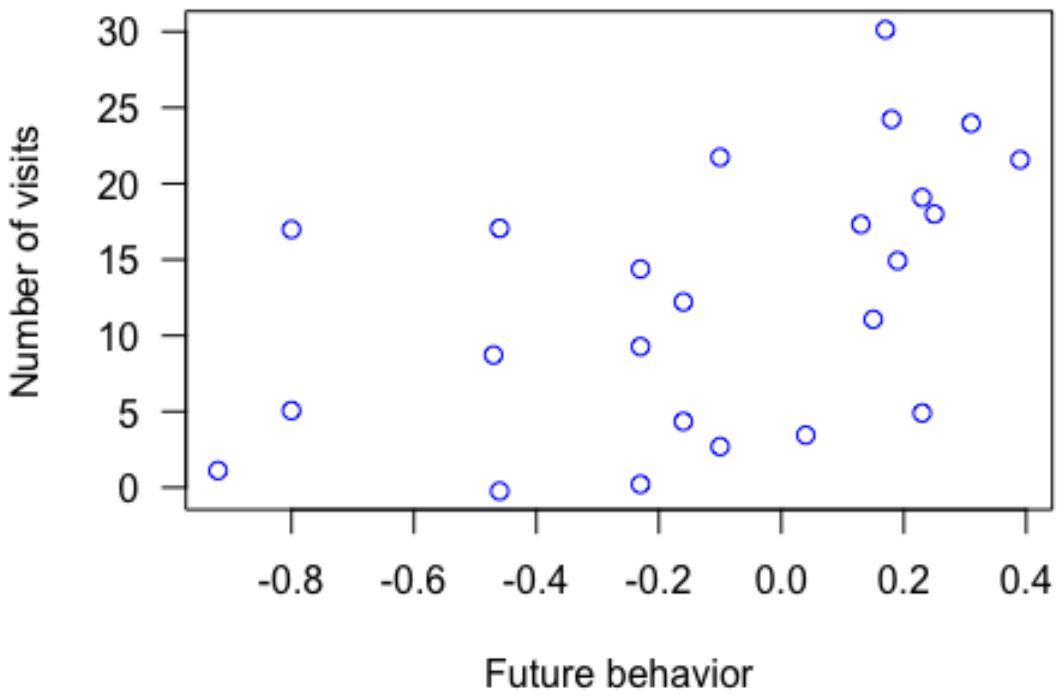

Figure 2.The association between Future behaviour and number of nestling visits (n = 24). Values have been shifted slightly in the Y direction to improve legibility.

We see in figure 2 that the associaton between future behaviour and number of nestling visits is positive, linear and moderate. There are one or two apparent outliers to the association.

Assumptions of correlation analysis

Correlation analysis assumes that:

- *the sample of individuals is a random sample from the population*;
- *the measurements have a bivariate normal distribution, which includes the following properties:*
- *the relationship between the two variables (X and Y) is linear*;
- *the cloud of points in a scatterplot of X and Y has a circular or elliptical shape*;
- *the frequency distributions of X and Y separately are normal*

Based on Figure 2, there doesn't seem to be any indications that the assumptions are not met, so we'll proceed with testing the null hypothesis.

```
booby.cor <- cor.test(nVisitsNestling ~ futureBehavior,
            data = booby,
            method = "pearson", conf.level = 0.95)
booby.cor #correlation analysis
```

```
##
##  Pearson's product-moment correlation
##
## data:  nVisitsNestling and futureBehavior
## t = 2.9603, df = 22, p-value = 0.007229
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1660840 0.7710999
## sample estimates:
##       cor
## 0.5337225
```

Concluding statement:

Nestling visits is significantly associated with future behaviour of the same individuals as adults (Pearson r = 0.53; 95% confidence limits: 0.167, 0.78; df = 22; P = 0.007).

Question 2

Use the `mammals` data frame for this question.

Can we predict average brain mass from average body mass for these mammal species?

NOTE: Although a Model-II regression would be more appropriate for this question, use a Model-I regression (which is what you learned in Tutorial 18).

Hypotheses

H0:β = 0 HA:β ≠ 0

α = 0.05

I used a *Least-squares linear regression* to test the hypothesis in this question because the response variable (Y) is numerical, there is an explanatory variable (X) which is not categorical, and the goal is to predict values of Y from X:

```
plot(brain_mass_g ~ body_mass_kg, data = mammals,
     xlab = "Body mass (kg)",
     ylab = "Brain mass (g)",
     pch = 1,
     col = "firebrick",
     las = 1)
```

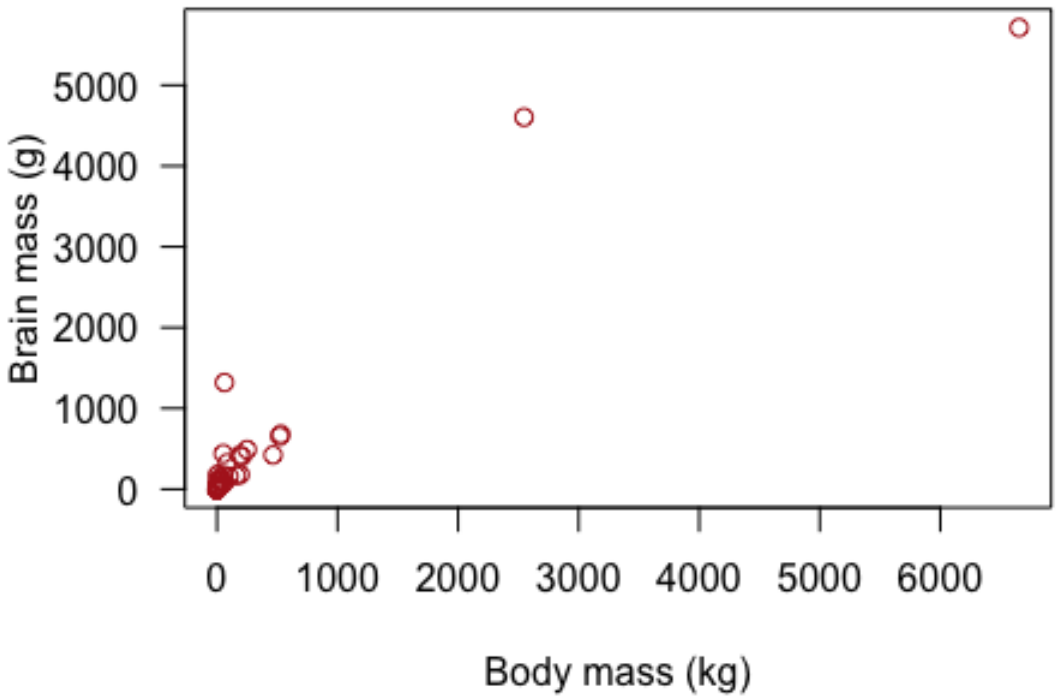

Figure 1. Scatterplot of association between Brain mass (g) and body mass (kg) of 62 mammals

---

We see in Figure 1 that there doesn't seem to be any association between brain mass and body mass of mammals. There is linearity within the data, however it is very weak and outliers are presence in highest body mass values.

Assumptions of regression analysis

Regression analysis assumes that:

- *the true relationship between X and Y is linear;*
- *for every value of X the corresponding values of Y are normally distributed;*
- *the variance of Y-values is the same at all values of X;*
- *at each value of X, the Y measurements represent a random sample from the population of possible Y values;*

Conduction of regression analysis:

```r brain_mass.lm <- lm(brain_mass_g ~ body_mass_kg, data = mammals)
mammals$brain_mass.lm.resids <- residuals(brain_mass.lm) ```
```
r { qqnorm(mammals$brain_mass.lm.resids, las = 1)
qqline(mammals$brain_mass.lm.resids) }
```

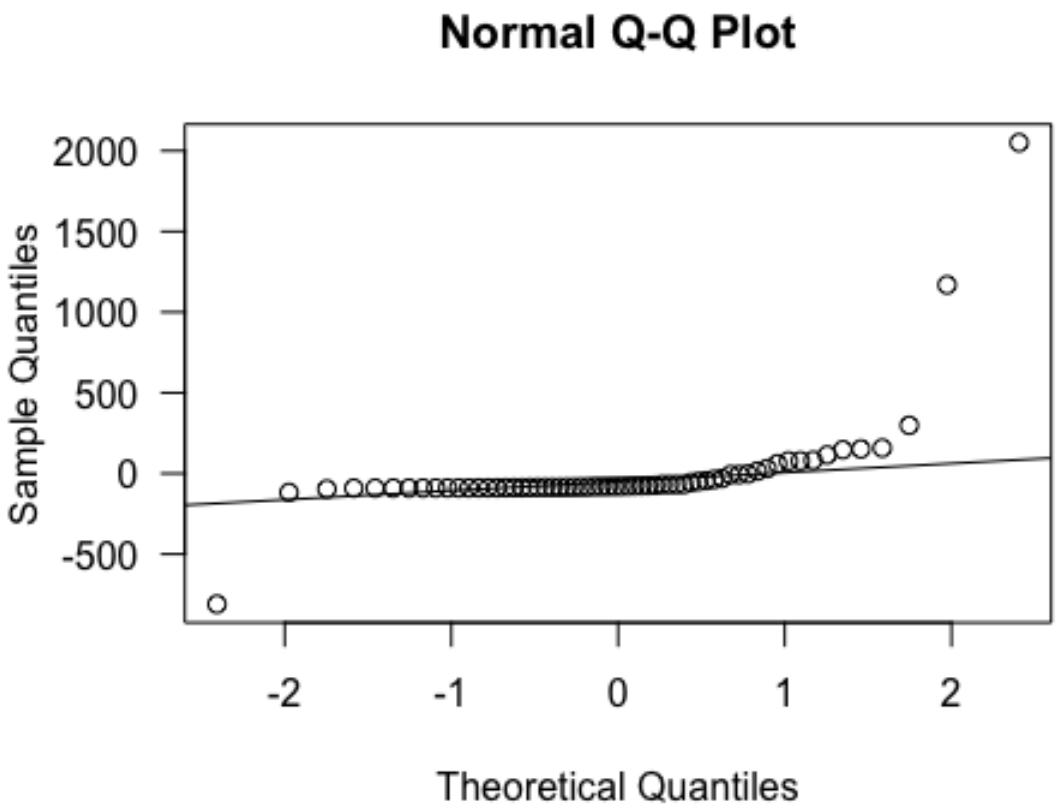


Figure 2: Normal quantile plot of the residuals from a regression of Brain mass on Body mass of 62 mammals.

*Figure 2 shows that the residuals don't fall consistently near the line in the normal quantile plot; there is slight curviture, and increasing values tend to fall further from the line. This suggests the need to log-transform the data.*

---

## Checking assumption "The variance of Y-values is the same at all values of X":

```
{
plot(brain_mass.lm.resids ~ body_mass_kg,
```
```

```
   data = mammals,
   ylab = "Residual",
   xlab = "Body mass (kg)",
   pch = 1,
   col = "firebrick",
   las = 1)
abline(0, 0, lty = 2)
}
```

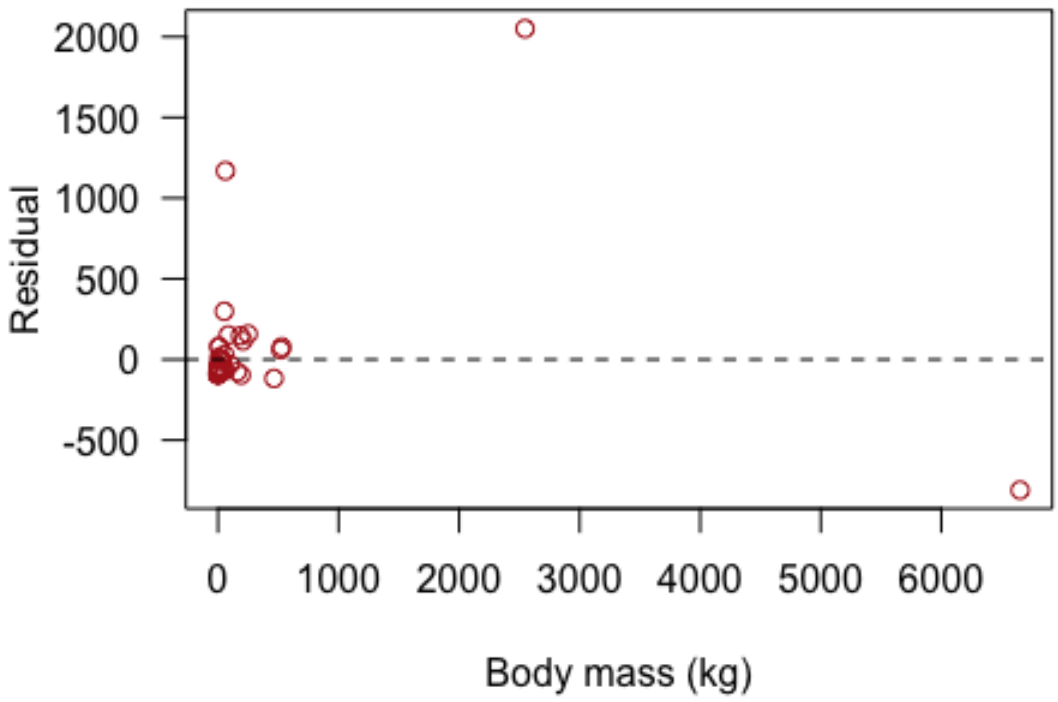

Figure 3.residual plot from a regression of brain mass (g) on body mass (g) for 62 plots.

Figure 3 shows that the variance in brain mass is very significant, specially in increasing body mass values. This suggests the need for a log transformation, possibly of both the explanatory and response variables.

Log transforming the data:

```
mammals$log.brain_mass <- log(mammals$brain_mass_g) #log-transform response variable

mammals$log.body_mass <-log(mammals$body_mass_kg) #log-transform explanatory variable

log.brain_mass.lm <- lm(log.brain_mass ~ log.body_mass,
             data = mammals) #create a linear model of the the transformed data

mammals$log.brain_mass.lm.resids <- residuals(log.brain_mass.lm) # create residual plots from the linear model of the log transformed data

{
  qqnorm(mammals$log.brain_mass.lm.resids, las = 1)
  qqline(mammals$log.brain_mass.lm.resids)
}
```

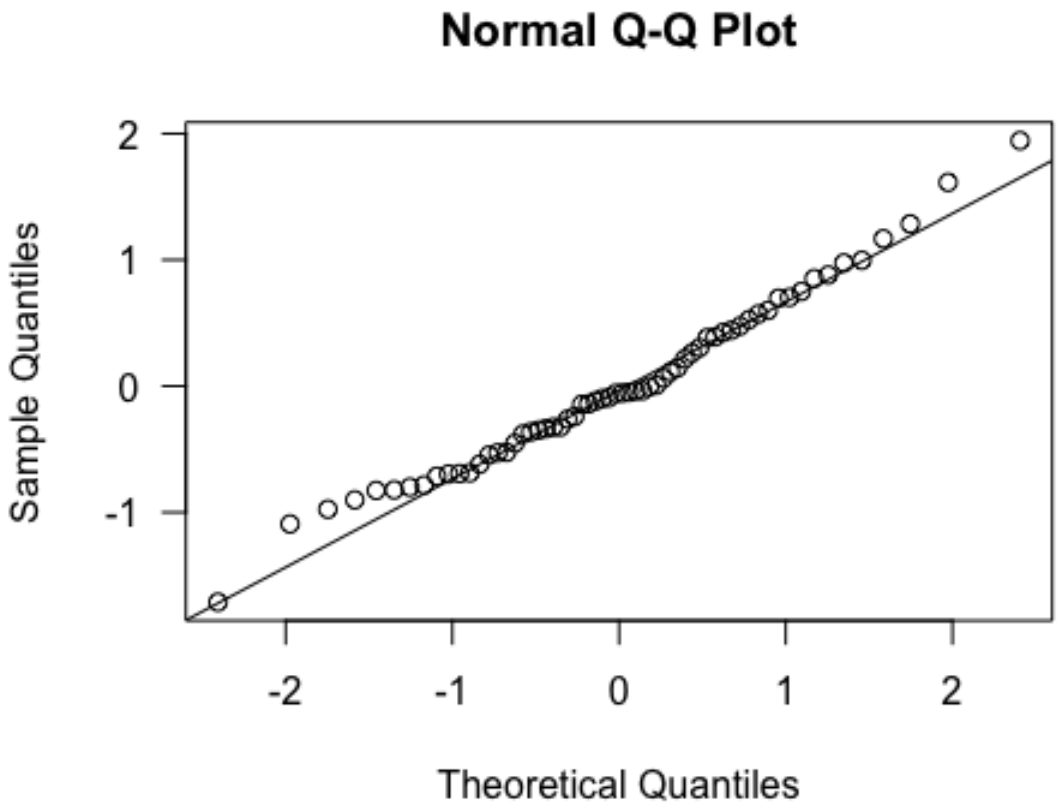

Figure 4. Normal quantile plot of the residuals from a regression of brain mass (log transformed) on body mass (log transformed) for 62 plots.

```
{
plot(log.brain_mass.lm.resids ~ log.body_mass, data = mammals,
      xlab = "Body mass (kg) (log transformed)",
      ylab = "Residual",
      pch = 1,
      col = "firebrick",
      las = 1) # plot log-transformed data
}
```

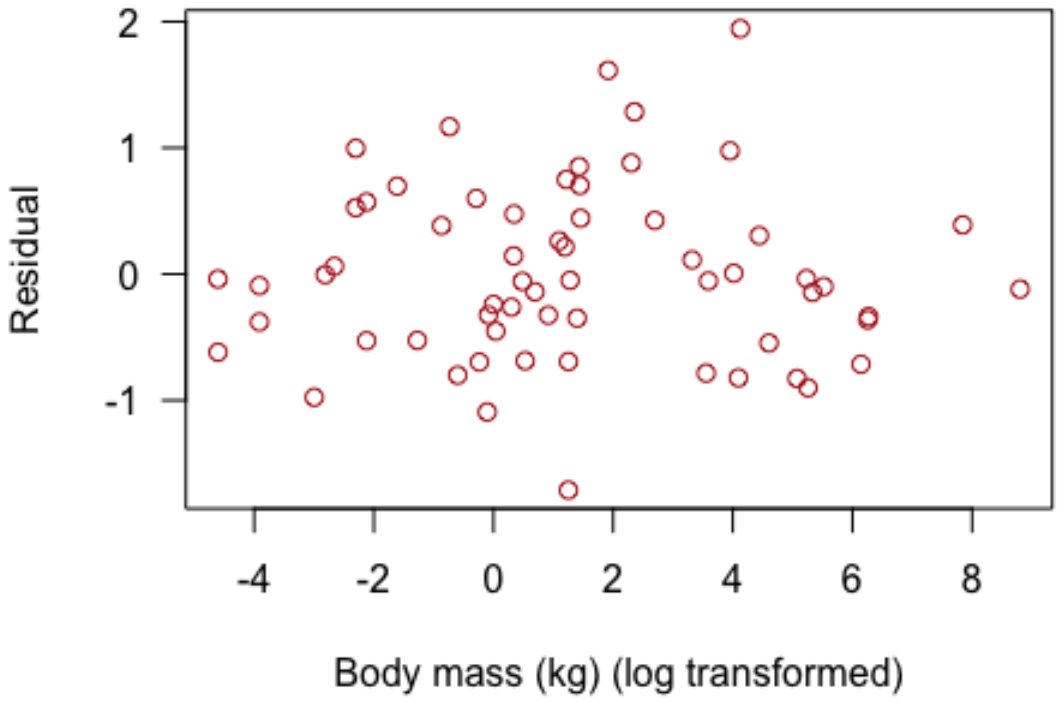

Figure 5. Residual plot from a regression of brain mass (kg) (log transformed) on body mass (log transformed) for 62 plots.

Figure 4 shows that the residuals are reasonably normally distributed, and Figure 5 shows no strong pattern of changing variance in Y along X, and nor is there an obvious curved pattern to the residuals. We therefore proceed with the analysis using the log-transformed response and explanatory variable.

```r
summary(log.brain_mass.lm)
```

```
##
## Call:
## lm(formula = log.brain_mass ~ log.body_mass, data = mammals)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.71143 -0.50667 -0.05606  0.43833  1.94425
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.12719    0.09682   21.97   <2e-16 ***
## log.body_mass  0.75451    0.02878   26.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.699 on 60 degrees of freedom
## Multiple R-squared:  0.9197, Adjusted R-squared:  0.9184
## F-statistic: 687.3 on 1 and 60 DF,  p-value: < 2.2e-16
```

```r
confint(log.brain_mass.lm)
```

```
##                   2.5 %    97.5 %
## (Intercept)   1.933525 2.3208556
## log.body_mass 0.696936 0.8120762
```

```r
visreg(log.brain_mass.lm,
    "log.body_mass",
    alpha = 0.05,
    line = list(col = "black"),
    points = list(cex = 1.5, pch = 1, col = "black", lwd = 1.75),
    las = 1,
    xlab = "Body mass (kg) (log transformed)",
    ylab = "Brain mass (g) (log transformed)")
```

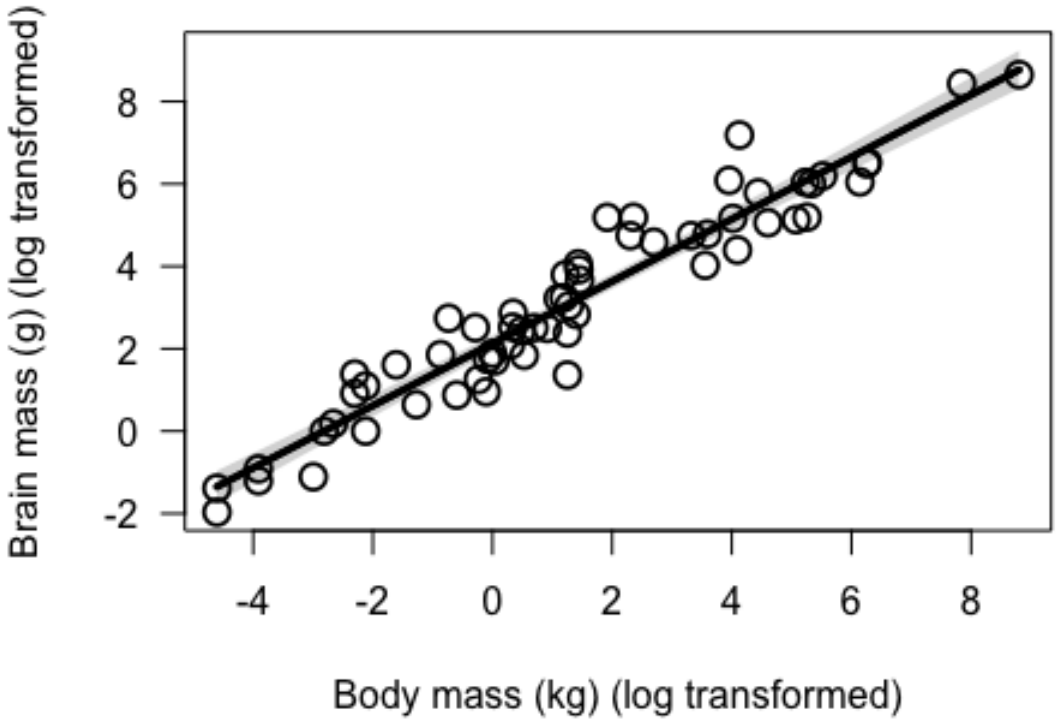

Figure 6. Brain mass (g) in 62 plots and their body mass (kg). Also shown is the significant least-square regression line (black solid line) and the 95% confidence bands (grey shading).

Figure 6 shows that brain mass (log transformed) is positively and strongly linear (associated) to body mass and there is no variation that remains unexplained by the least-square regression model such as outliers.

Concluding statement:

*As seen in Figure 6, body mass is a significant predictor of brain mass: Brain mass (log) = 2.12 + 0.75 (log body mass); F = 687.3; n = 62; 95% confidence limits for the slope: 0.70, 0.81; R2 = 0.92; P< 0.001). Given the relatively high R2 value, predictions from this regression model will be accurate.

```r
set.seed(224)

new.X.data <- data.frame(log.body_mass = sample(1:16, 62,
                replace = T))

mammals$predicted.vals <- predict.lm(log.brain_mass.lm,
                          newdata = new.X.data)
{
 plot(log.brain_mass ~ log.body_mass,
     data = mammals,
     ylab = "Brain mass (log) (g)",
     xlab = "Body mass (log) (kg)",
     pch = 1,
     col = "firebrick",
     las = 1)
 points(new.X.data$log.body_mass,
       mammals$predicted.vals, pch = 16)
}
```

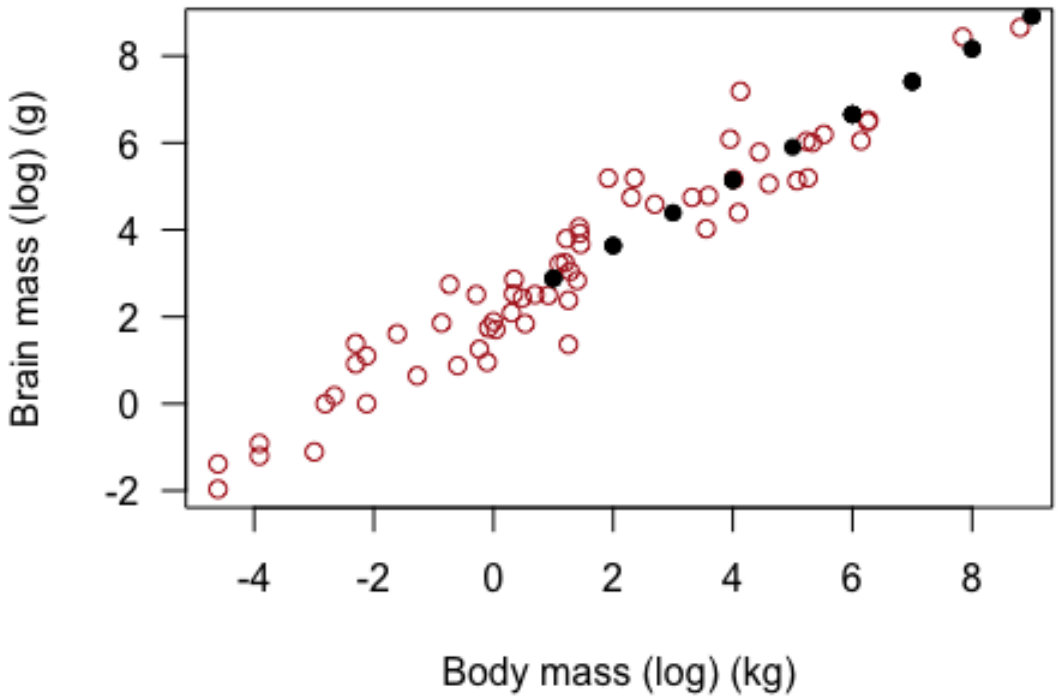

Figure 7. Measured brain mass (g) (log transformed) in 62 plots and their respective body mass (kg) (log transformed). Also shown are predicted values of Y across values of X.

Figure 7 shows the overall positive association between brain mass and body mass of mammals and it also shows no considerable scatter in the relationship, hence the relatively high R2 value.
