TP: is Batman somewhere?

Gabriele Degola

December 2021

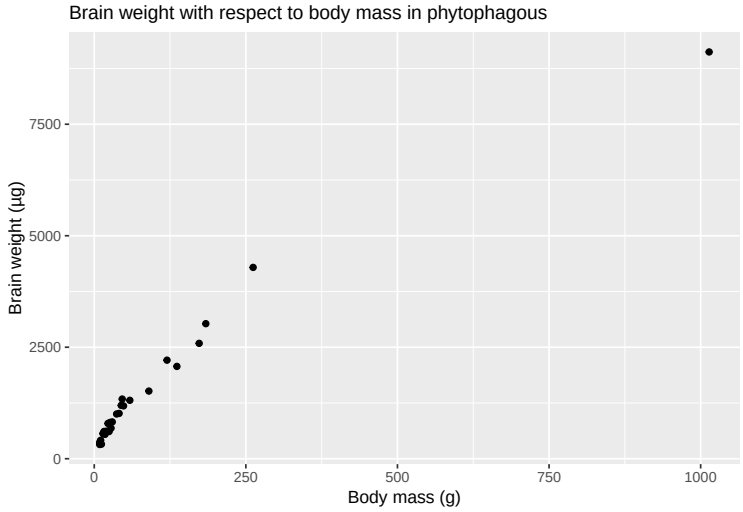
```
library(tidyverse)
```

```
myData <- read.table(file="../data/bats.csv", sep=";", skip=3, header=TRUE)
names(myData)
```

```
## [1] "Species" "Diet"    "Clade"   "BOW"     "BRW"     "AUD"     "MOB"
## [8] "HIP"
```

Study of the relationship between brain weight and body mass

```
phyto <- myData[(myData$Diet == 1),]
ggplot(phyto, aes(x=BOW, y=BRW)) + geom_point() +
  xlab("Body mass (g)") + ylab("Brain weight (µg)") +
  ggtitle("Brain weight with respect to body mass in phytophagous") +
  theme(plot.title = element_text(size=12))
```



It looks like body mass (in grams) has a clear influence on brain weight (in micrograms). This relation can be analysed fitting a linear model of brain weight as a function of body mass. The fitted line is also shown in next plot.

```
reg1 <- lm(BRW ~ BOW, data=phyto)
ggplot(phyto, aes(x=BOW, y=BRW)) + geom_point() + geom_smooth(method="lm") +
  xlab("Body mass (g)") + ylab("Brain weight (µg)") +
  ggtitle("Brain weight with respect to body mass in phytophagous (with regression
  ↪  line)") +
```

```
  theme(plot.title = element_text(size=12))
```

With this expression, R estimates the β coefficients of the formula $Y = \beta_1 + \beta_2 X + \epsilon$, where X is the vector containing the values of the measured body masses and Y contains the measured brain weights. ϵ is random noise.

```
summary(reg1)
```

```
##
## Call:
## lm(formula = BRW ~ BOW, data = phyto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -628.32 -233.94  -65.74  158.26 1308.59
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 623.4469    81.4762   7.652 3.14e-08 ***
## BOW           8.9999     0.3972  22.659  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 396.9 on 27 degrees of freedom
## Multiple R-squared:   0.95,  Adjusted R-squared:  0.9482
## F-statistic: 513.4 on 1 and 27 DF,  p-value: < 2.2e-16
```

In this case, a few results can be observed: - The intercept β_1 is estimated as 623.45. - The p-value for the employed test statistics is extremely small, lower than 2.2e-16. - The null hypothesis of this test is $H_0 : \beta_1 = \beta_2 = 0$. This hypothesis is therefore rejected. - Body mass has a significative relation with brain weight. The intercept is also important. - The coefficient of determination R^2 of the fitted model is 0.95, so a big portion of the global variation is explained by the model.

Next, analysis of variance can be performed.

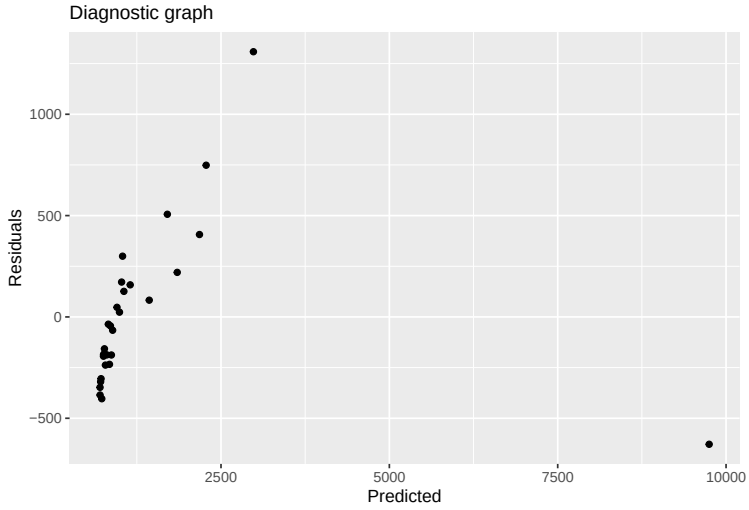
```
anova(reg1)
```

```
## Analysis of Variance Table
##
```

```
## Response: BRW
##           Df   Sum Sq  Mean Sq F value    Pr(>F)
## BOW        1 80888380 80888380  513.42 < 2.2e-16 ***
## Residuals 27  4253838   157550
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This confirms that body weight is highly significative. In addition, the previous table contains information about the model residual, with the sum of residual squares being 4253838. The residuals have been computed during model fitting and can be plotted against the predicted values.

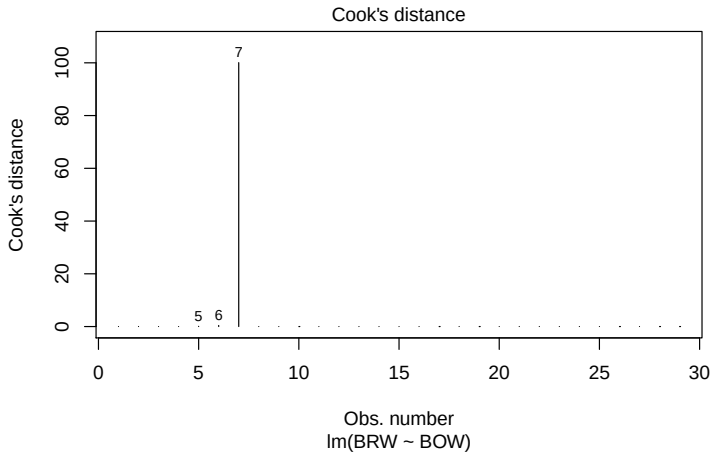
```
tmp <- data.frame(reg1$fitted.values, reg1$residuals)
ggplot(tmp, aes(x=reg1.fitted.values, y=reg1.residuals)) + geom_point() +
  xlab("Predicted") + ylab("Residuals") + ggtitle("Diagnostic graph") +
  theme(plot.title = element_text(size=12))
```



Clearly, something is not going well: most prediction values are below 3750 µg, except one with predicted brain weight around 10000 µg. As this last point is so far from the others, it influences the model results worsening the prediction for values in the middle. For now, we can consider it as an outlier.

Cook's distance can be used for outlier detection in the fitted model:

```
plot(reg1, 4)
```



Observation number 7 has distance much higher than the higher and probably corresponds to the previously identified point.

```
myData[7,]
```

```
##                Species Diet Clade  BOW  BRW   AUD    MOB    HIP
## 7  Pteropus  vampyrus     1     I 1014 9121 16.93 243.54 331.29
```

Indeed, Pteropus vampyrus has body mass 1014 g and brain weight 9121 µg.

We can therefore redo the previous analysis without taking it into account.

```
phytobis <- phyto[which(phyto$BRW<8000),]
reg2 <- lm(BRW ~ BOW, data=phytobis)
ggplot(phytobis, aes(x=BOW, y=BRW)) + geom_point() + geom_smooth(method="lm") +
  xlab("Body mass (g)") + ylab("Brain weight (µg)") +
  ggtitle("Brain weight with respect to body mass in phytophagous (with regression
  ↪  line)") +
  theme(plot.title = element_text(size=12))
```

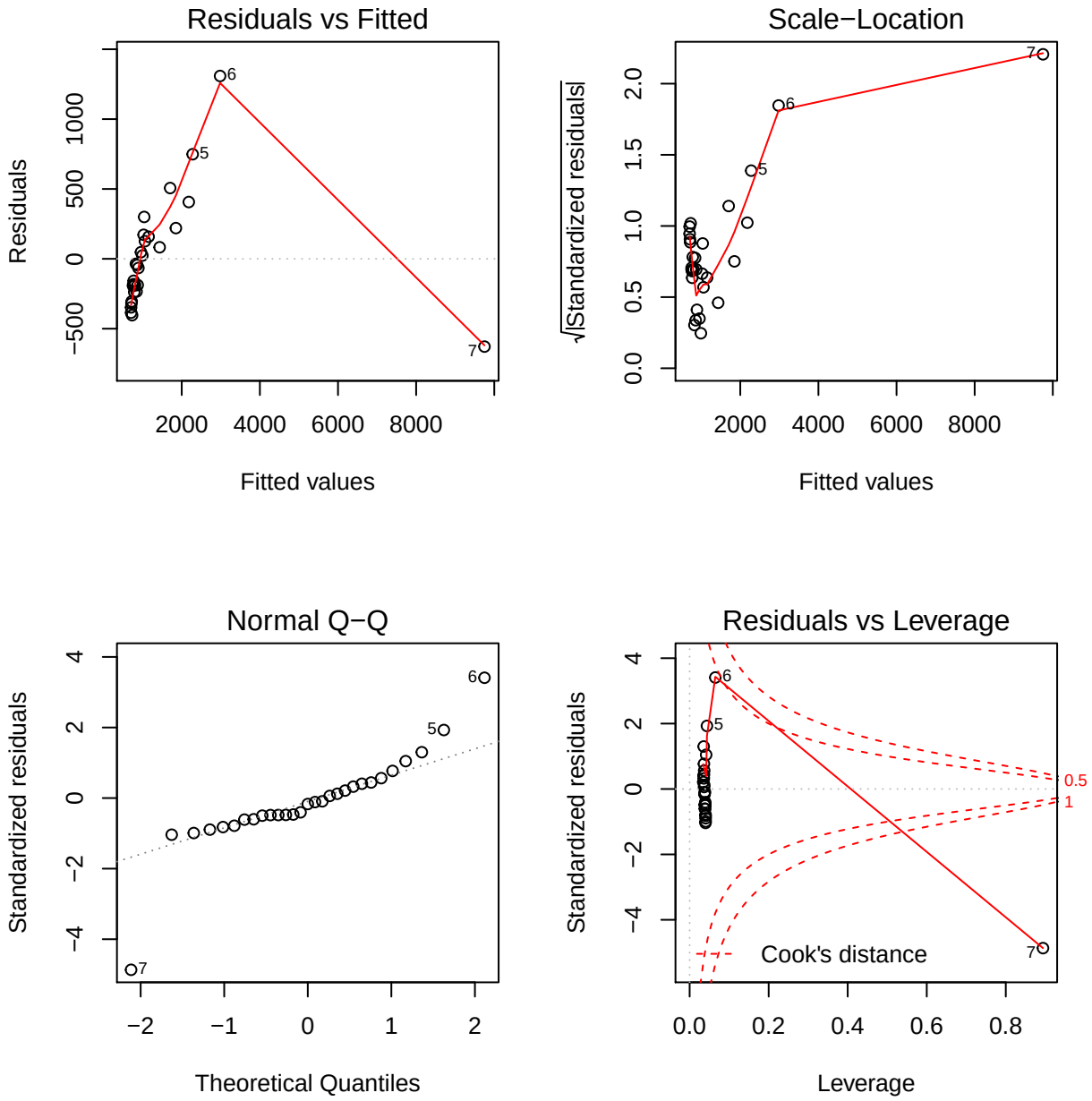

```
summary(reg2)
```

```
##
## Call:
## lm(formula = BRW ~ BOW, data = phytobis)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -269.76  -93.33    8.73  112.93  322.55
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 346.5452    35.4920   9.764 3.48e-10 ***
## BOW          14.5099     0.4285  33.860  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 141.8 on 26 degrees of freedom
## Multiple R-squared:  0.9778, Adjusted R-squared:  0.977
## F-statistic:  1147 on 1 and 26 DF,  p-value: < 2.2e-16
```

Clearly, now the regression line better fits intermediate point and is not affected by the removed extreme observation. The predicted coefficient for the body weight predictor is higher than before, as well as the coefficient of determination.

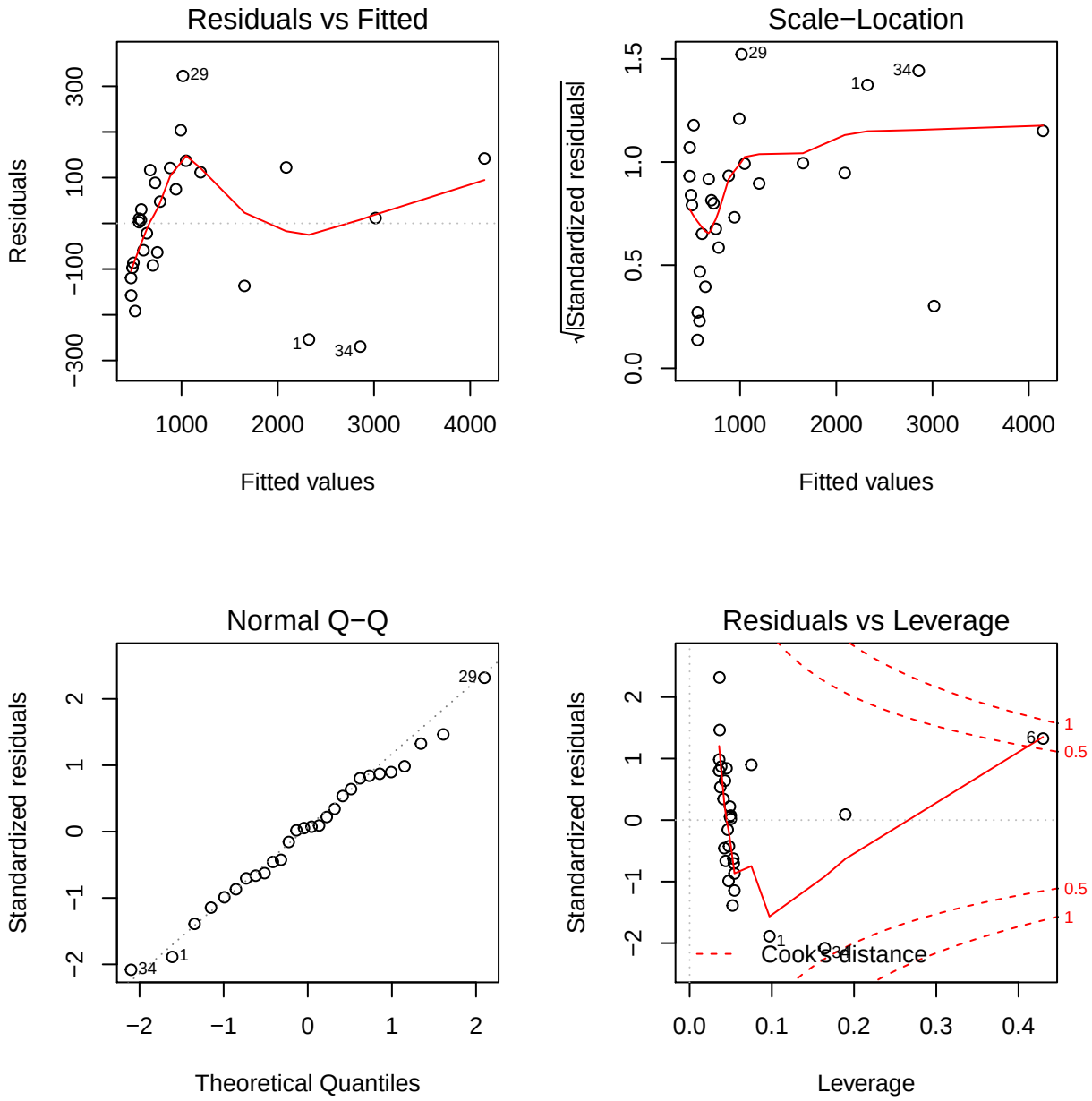
A few graphs can therefore plotted for model diagnosis, including the residuals agains predicted shown before. For the first model, including the "outlier":

```
par(mfcol=c(2,2))
plot(reg1)
```



If the "outlier" is removed:

```
par(mfcol=c(2,2))
plot(reg2)
```



Important differences can be noticed: - The plots related to residuals vs fitted and scale-location are more "flat" for the second model, so it provides more accurate predictions than the first one. - According to the Q-Q plot, it is more probable that the data used to fit the second model are drawn from a normal distribution, with respect to the first model.

Therefore, the second model better satisfies common assumptions of linear regression.