Model Diagnostics II: multicolinearity, outliers, influential observations

Emilija Perković

Dept. of Statistics
University of Washington

Multicollinearity

Two predictors x_1 and x_2 are collinear if

$$x_2 = \lambda_0 + \lambda_1 x_1,$$

for some $\lambda_0, \lambda_1 \in \mathbb{R}$.

K, $K \geq 2$, predictors $x_1, x_2 \ldots, x_K$ are multicollinear if

$$\lambda_0 + \lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_K x_K = 0,$$

for some $\lambda_0, \lambda_1, \ldots, \lambda_K \in \mathbb{R}$.
Multicollinearity refers to a situation in which two or more predictors variable in a multiple regression model are **highly linearly related**.

Above are examples of perfect collinearity and perfect multicollinearity.

Multicollinearity

We assume:

$$\underline{y} = X\underline{\beta} + \underline{\epsilon}$$

with $\underline{\epsilon} \sim \mathcal{N}(\underline{0}, \sigma^2 I_n)$.

We estimate $\underline{\hat{\beta}}$ using OLS:

$$\underline{\hat{\beta}} = (X'X)^{-1}X'\underline{y}$$

If perfect multicollinearity is present among columns of X, then $X'X$ is singular (the inverse $(X'X)^{-1}$ does not exist).

Multicollinearity

If there is no perfect multicollinearity among predictors $X'X$ will be invertible, but the values of $(X'X)^{-1}$ will be very large.(see Hw 2 for an example).

Then since,

$$\text{Var}[\underline{\hat{\beta}}] = \sigma^2(X'X)^{-1},$$

the variance of the estimated coefficients $\underline{\hat{\beta}}$ will also be very large. Hence, it may be difficult to say anything precise about the coefficients.

One example: the global F-Test (`anova()` - test comparing the full and empty model) turns out to be significant, but none of the individual predictors is significant.

Note however, that multicollinearity of predoctors does not violate model assumptions!

Diagnosing multicollinearity

Linear relationships between pairs of variables are fairly easy to diagnose:

- we make the pairs plot (pairwise scatter plot) of all the variables, and we see if any of them fall on a straight line, or close to one.

A multicollinear relationship involving three or more variables might be totally invisible on a pairs plot.

For instance, suppose x_1 and x_2 are independent Gaussians, of equal variance σ^2 and x_3 is their average:

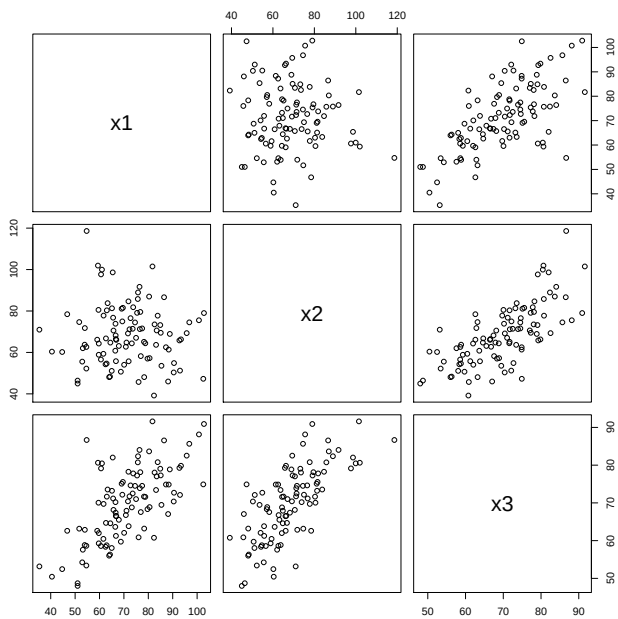
$$x_3 = \frac{x_1 + x_2}{2}.$$

Then the correlation of x_1 and x_3 is:

$$\begin{aligned}\text{Cor}[x_1, x_3] &= \frac{\text{Cov}[x_1, x_3]}{\sqrt{\text{Var}[x_1]\text{Var}[x_3]}} = \frac{\text{Cov}[x_1, (x_1 + x_2)/2]}{\sqrt{\sigma^2 \sigma^2/2}} \\ &= \frac{\sigma^2/2}{\sigma^2/\sqrt{2}} = \frac{1}{\sqrt{2}} \approx .71,\end{aligned}$$

which is quite a bit smaller than 1.

Diagnosing multicollinearity: Example



See R script.

Variance inflation factor

If the predictors are uncorrelated, the variance of $\hat{\beta}_i$ would be

$$\text{Var}[\hat{\beta}_i] = \frac{\sigma^2}{\sum_{j=1}^{n}(x_{ji} - \overline{x_i})^2}. \tag{1}$$

With correlated predictors we have

$$\text{Var}[\hat{\beta}_i] = \sigma^2 (X'X)^{-1}_{i+1,i+1} = \sigma^2 \frac{1}{1-R_i^2} \frac{1}{\sum_{j=1}^{n}(x_{ji} - \overline{x_i})^2}, \tag{2}$$

where R_i^2 is the R-squared in the regression of x_i on all other predictors in X.

The R^2 (R-squared) of a multivariate linear regression fit with response y, predictors $x_1, \ldots, x_k$ and fitted values $\hat{\underline{y}} = X\hat{\underline{\beta}}$ is:

$$R^2 = 1 - \frac{\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}{\sum_{j=1}^{n}(y_j - \overline{y}_j)^2} = 1 - \frac{RSS}{SYY}.$$

The variance inflation factor (VIF) for the i-th coefficient is the ratio between equations (2) and (1), that is:

$$VIF_i = \frac{1}{1-R_i^2}.$$

Variance inflation factor

The variance inflation factor (VIF) for the i-th coefficient is

$$VIF_i = \frac{1}{1-R_i^2}.$$

Hence, $VIF_i \geq 1$, for any predictor $x_i$and VIF_i will increase as x_i becomes more correlated with some linear combination of the other predictors.

General heuristic:

- $VIF_i > 10$ indicates "serious" multicollinearity for a predictor.
- A $VIF_i \geq 5$, corresponds to a $R_i^2 \geq .8$ and $VIF_i \geq 10$ means that $R_i^2 \geq .9$.

Note however, that a large VIF_i does not violate any model assumptions!

Examples: Car seats

Data `seatpos` in R package `faraway`.

Car drivers like to adjust the seat position for their own comfort. Car designers would find it helpful to know where different drivers will position the seat depending on their size and age. Data on 38 drivers.

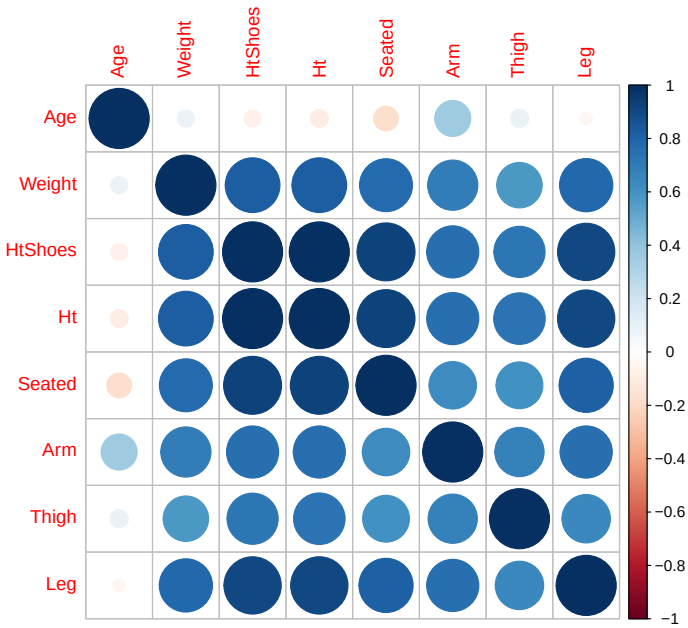
- hipcenter - horizontal distance of hips to steering wheel
- Age - age in years,
- Weight - weight in pounds,
- HtShoes, Ht, Seated - height w/o, w/ shoes, seated height,
- Arm, Thigh, Leg - arm, thigh and leg length.

Examples: Car seats

```
lm(formula = hipcenter ~ ., data = seatpos)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 436.43213  166.57162   2.620   0.0138
Age           0.77572    0.57033   1.360   0.1843
Weight        0.02631    0.33097   0.080   0.9372
HtShoes      -2.69241    9.75304  -0.276   0.7845
Ht            0.60134   10.12987   0.059   0.9531
Seated        0.53375    3.76189   0.142   0.8882
Arm          -1.32807    3.90020  -0.341   0.7359
Thigh        -1.14312    2.66002  -0.430   0.6706
Leg          -6.43905    4.71386  -1.366   0.1824
---
Residual standard error: 37.72 on 29 degrees of freedom
Multiple R-squared:  0.6866,Adjusted R-squared:  0.6001
F-statistic:  7.94 on 8 and 29 DF,  p-value: 1.306e-05
```

Note the global F-test p-value! Let's take a look at the correlation matrix.

Examples: Car seats



We notice very large pairwise correlations!

Examples: Car seats

Let's consider the variance inflation factors in this data set.

```
> vif(fit.carseat)
       Age     Weight    HtShoes         Ht
  1.997931   3.647030 307.429378 333.137832
    Seated        Arm      Thigh        Leg
  8.951054   4.496368   2.762886   6.694291
```

How to deal with multicollinearity?

Dealing with multicollinearity

- Amputation - remove all except one among the multicolinear predictors.
  - Potentially discarding valuable information.
  - Need to know which predictors are linearly dependent. Use domain knowledge.
- Transform predictors. Requires using domain knowledge.
- Principal component regression (not covered in the course).
- Ridge regression (hopefully covered later in the course).

Let's consider amputation and transforming the predictors on our example.

Examples: Car seats, amputation

Remove predictors: `HtShoes, Seated, Arm` and Leg. Fit the model again.

```
lm(formula = hipcenter ~ Age + Weight + Ht, data = seatpos)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 528.297729 135.312947   3.904 0.000426
Age           0.519504   0.408039   1.273 0.211593
Weight        0.004271   0.311720   0.014 0.989149
Ht           -4.211905   0.999056  -4.216 0.000174
---
Residual standard error: 36.49 on 34 degrees of freedom
Multiple R-squared:  0.6562,Adjusted R-squared:  0.6258
F-statistic: 21.63 on 3 and 34 DF,  p-value: 5.125e-08

> vif(fit.carseat2)
     Age   Weight       Ht
1.093018 3.457681 3.463303
```

Examples: Car seats, transformation

What if we transform the predictors instead?

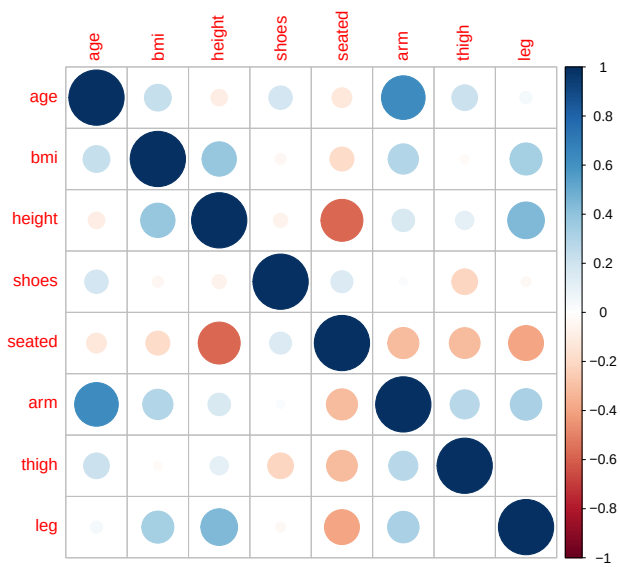
```
age    <- Age
bmi    <- (Weight*0.454)/(Ht/100)^2
shoes  <- HtShoes-Ht
seated <- Seated/Ht
arm    <- Arm/Ht
thigh  <- Thigh/Ht
leg    <- Leg/Ht

fit.carseat3 <- lm(hipcenter~.,data=seatpos.new)

> vif(fit.carseat3)
   age      bmi   height    shoes
 1.994473 1.408055 1.968447 1.155285
   seated      arm    thigh      leg
 1.851884 2.044727 1.284893 1.480397
```


Examples: Car seats, transformation

Correlation plot after transformation:



Outliers

An **outlier** is a data point which is very far, somehow, from the rest of the data.

They are often worrisome, but not always a problem.

When we are doing regression modeling, in fact, we don't really care about whether some data point is far from the rest of the data, but whether it breaks a pattern the rest of the data seems to follow.

They can be easy to spot with only one predictor. Let's consider an example.

Example: Star

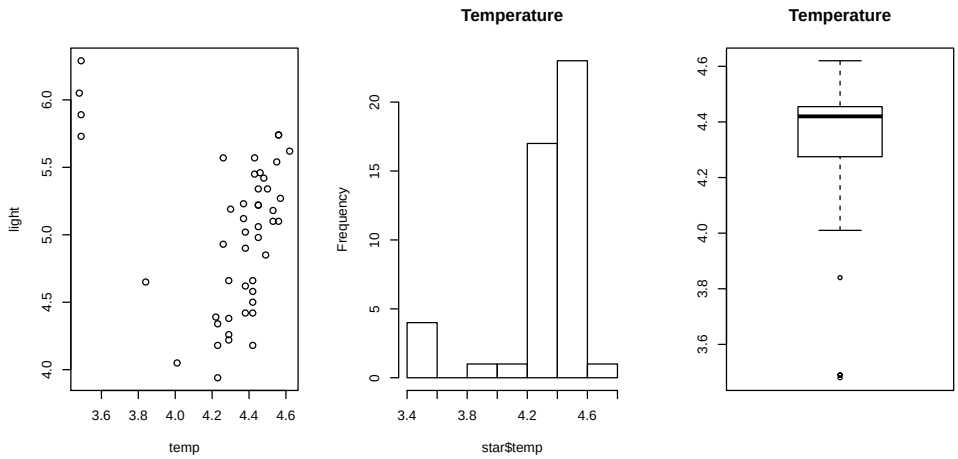
Data set `star` from R package `faraway`.
Data on the **log** of the surface temperature and the **log** of the light intensity of 47 stars in the star cluster CYG OB1, which is in the direction of Cygnus.

- index = a numeric vector
- temp = temperature
- light = light intensity

Response: light, predictor: temp

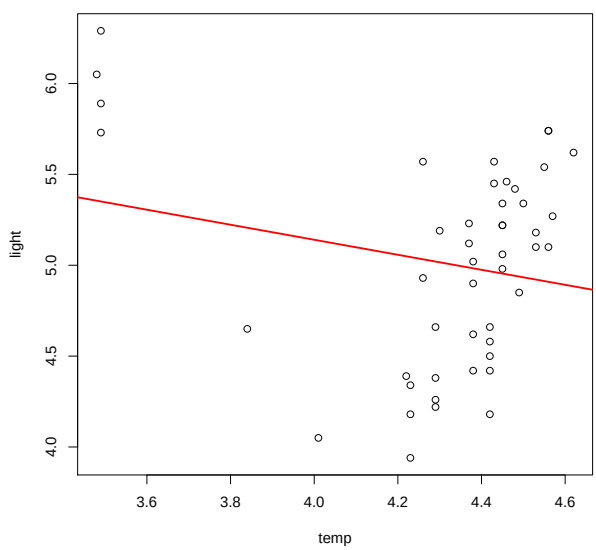
Let's look at the scatter plot, the histogram and boxplot of temp.

Example: Star



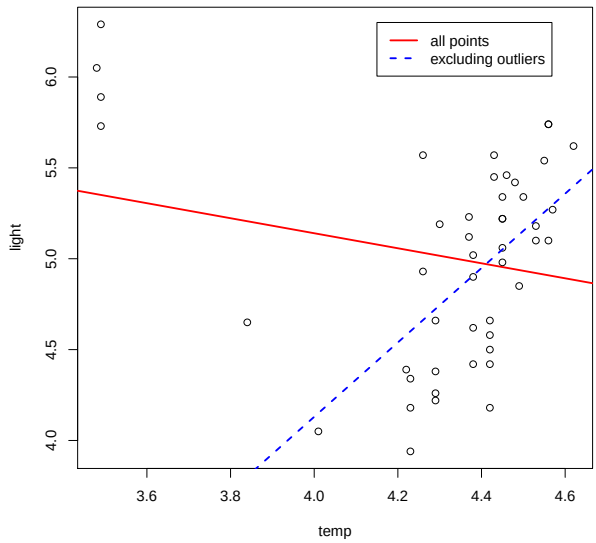
Do you notice anything unusual?Why is this an issue?

Example: Star



What if these point are measurement error?

Example: Star



What if these point are **NOT** measurement error?

Do not just discard unusual points!

It is dangerous to exclude outliers automatically.

One case from history:

- NASA's satellite that recorded atmospheric information automatically discarded extremely low ozone observations for data analyses for several years.
- Those observations were later attributed to the Antarctic ozone hole.

Let us first consider how to identify outliers when we have more than one predictor.

Recall: Hat matrix

The fitted values $\underline{\hat{y}}$ can be written as a function of the observed values $\underline{y}$:

$$\underline{\hat{y}} = H\underline{y},$$

$H = X(X'X)^{-1}X'$ is called the hat matrix and $\sum_{i=1}^{n} h_{ii} = p+1$, h_{ii} is the (i,i) element of H.

We also saw that

$$\underline{\hat{\epsilon}} \sim \mathcal{N}(0, \sigma^2(I_n - H)),$$

So that $\text{Var}[\hat{\epsilon}_i] = \sigma^2(1-h_{ii})$.

- h_{ii} is called the leverage of point i.
- since $\sum_{i=1}^{n} h_{ii} = p+1$, the average leverage is $(p+1)/n$.
- What happens to $\text{Var}[\hat{\epsilon}_i]$ if a point has high leverage? If, for example, h_{ii} is close to 1?
- What does that mean for the regression line?

Leverage

When h_{ii} is close to 1:

- $\text{Var}[\hat{\epsilon}_i]$ will be close to zero, independent of the observed y_i value; in other words, the regression fit will be "forced" close to y_i.
- Thus, points with a high leverage (sometimes called leverage points) have the potential to influence the fit.
- If in addition, observation i has a large residual, this can be concerning.
- To identify these points and get a fuller picture of the regression fit examine the standardized residuals.
- Recall:

$$r_i = \frac{\hat{\epsilon}_i}{SE(\hat{\epsilon}_i)}.$$

Rule of thumb: examine all points with leverage $> 2(p+1)/n$, (twice the average leverage).

Cook's distance

If we omit one observation, how **big** of a change do we notice in the **fitted regression**?

Of course, omitting one observation will change not only one but **all fitted values**.

This can be measured by a statistic called Cook's distance:

$$D_i = \frac{\sum_{i=1}^{n}(\hat{y}_i - \hat{y}_{i(i)})^2}{\hat{\sigma}^2(p+1)}$$

Where: $\hat{y}_{(i)}$ denotes the fitted values of an OLS regression after excluding the i-th observation from the original data set.

Cook's distance

In fact,

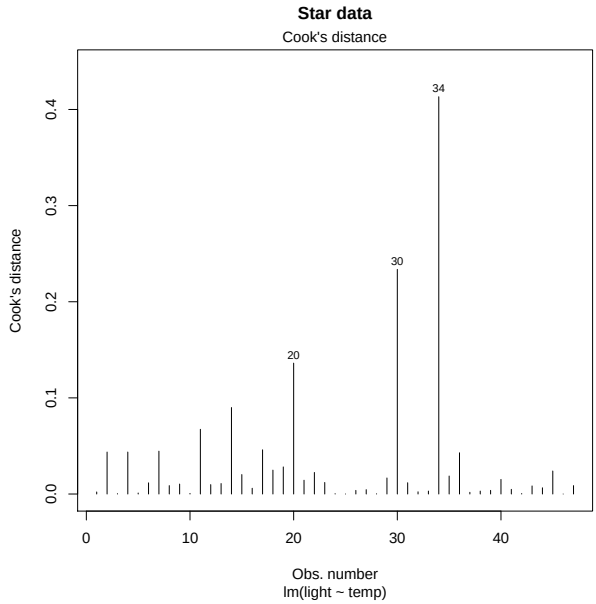
$$D_i = \frac{1}{\hat{\sigma}(p+1)}\hat{\epsilon}_i^2 \frac{h_{ii}}{(1-h_{ii})^2}.$$

Notice that $h_{ii}/(1-h_{ii})^2$ is a growing function of h_{ii}. Hence, the total "influence" of an observation i over all fitted values grows with both its leverage - h_{ii} -and the size of its residual - $\hat{\epsilon}_i$ - (when it is included in the OLS fit).

A distribution theory for D_i is not easy to obtain. However, it has been found useful to relate D_i to the $F_{p+1,n-p-1}$ distribution and ascertain the corresponding percentile value. If the percentile value is 50 percent or more, the fitted values obtained with and without the i-th case should be considered to differ substantially.
Rule of thumb: Investigate cases with $D_i \geq F_{0.5,p+1,n-p-1}$. For large n (and a large enough p) this will be close to 1.
In R see function `cooks.distance()`,`plot(lm(.), which =4)` and `plot(lm(.),which=5)`.

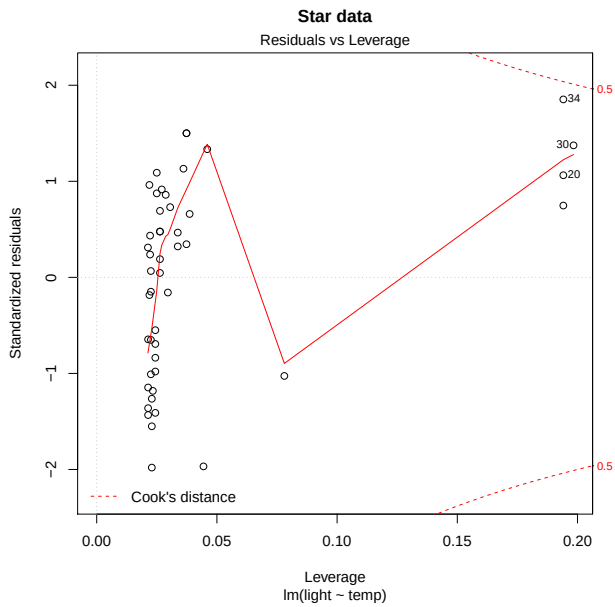
Example: Star data

Note that $F_{0.5,p+1,n-p-1} = F_{0.5,2,45} \approx 0.704$.



Example: Star data

R usually considers 0.5 and 1 as heuristics for a large Cook's distance.



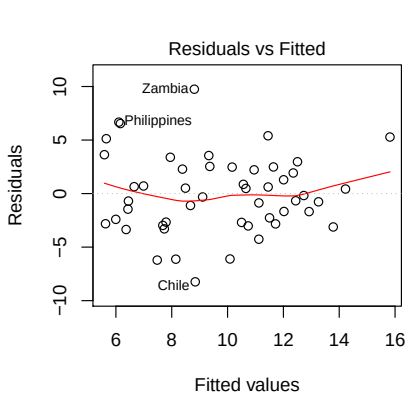
Example: Savings

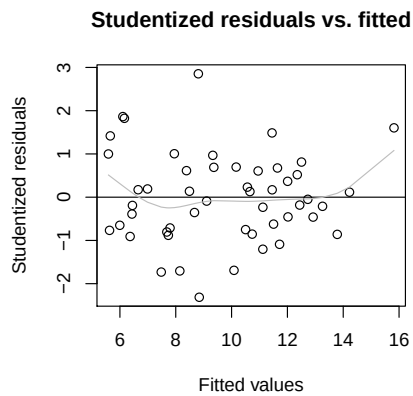
Consider `savings` data from R package `faraway`.
Savings rates from 50 countries, averaged over the period 1960-1970.

- sr - personal saving divided by disposable income
- pop15 - percent population under age of 15
- pop75 - percent population over age of 75
- dpi - per-capita disposable income in dollars
- ddpi - percent growth rate of dpi

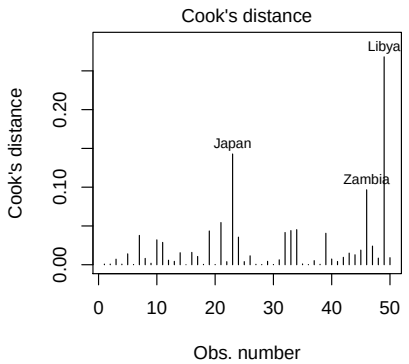
Let's look examine the leverage, standardized residuals and Cook's distance.

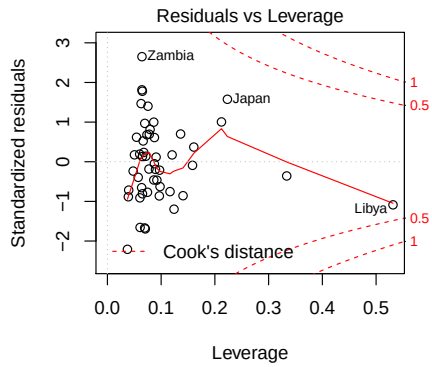
Example: Savings





Example: Savings





Summary: How to identify outliers?

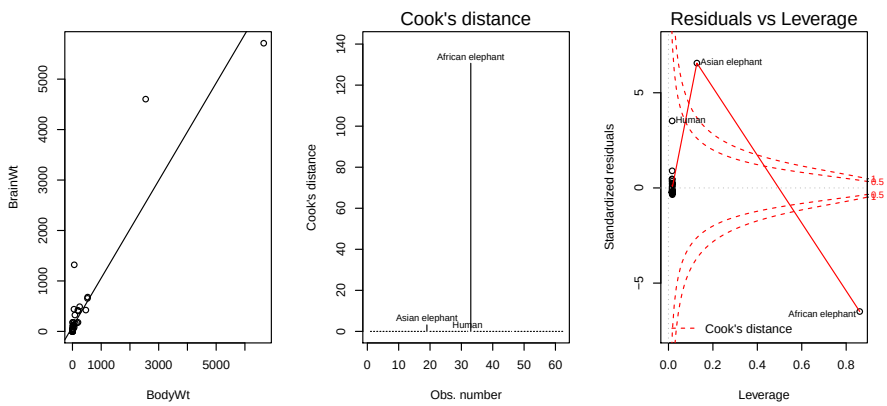
We have three ways of looking at whether points are outliers:

- We can look at their leverage, which depends only on the value of the predictors.
- We can look at their standardized residuals (or studentized), which depend on how far they are from the regression line.
- We can look at their Cook's statistics, which say how much removing each point shifts all the fitted values; it depends on the both the leverage and the residuals.

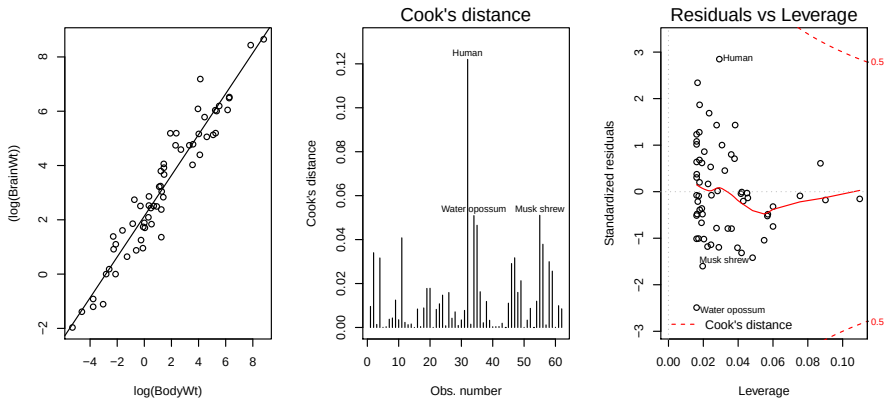
How to deal with outliers

- Check for a data entry error first. These are relatively common.Unfortunately, the original source of the data may have been lost.
- Examine the physical context - why did it happen? Sometimes, the discovery of an outlier may be of singular interest.
  Some scientific discoveries spring from noticing unexpected aberrations (remember NASA example).
  Another example of the importance of outliers is in the statistical analysis of credit card transactions. Outliers in this case may represent fraudulent use.
- Outliers can also serve as an indication that your model is wrong. After changing the model, the outliers may not be outliers anymore.Example: brain data (also star data).

Brain data, before transformation



Brain data, after transformation



No outliers after transformation!