Statistical analysis of expression data

Nicolas Delhomme

High Throughput Sequencing Data Analysis 2020

Learning Objectives

- By the end of the lecture, you will be able to:
  - reflect on the importance of the studies design for an experiment
  - describe the basic statistical concepts necessary for data analysis

Part I: Study Design

Model and assumptions: important aspects of studies design

- The following slides are from Patrik Rydén, from the Department of Mathematics and Mathematical statistics at Umeå University.

Can the weight of an animal's brain be explained by its body weight

- Patrik downloaded from an online resource the brain and body weight of species classified as animals.
- He defined the variable X and Y as the body and brain weight, respectively
- And asked himself the question: how can the relationship between X and Y be described?

In other words, what model?

Is there a “simple” relation between X and Y?

Some regression models that we can consider

$\boxed{Y=\alpha+\beta X}$ “simple linear regression”

or

$$Y=\alpha+\beta_1X+\beta_2X^2$$

or

$$\log(Y)=\alpha+\beta_1\log(X)+\beta_2\log(X)^2$$

Or .....

All models are wrong, but some are useful!

We can never prove that a model is correct, but we can reject bad models!

In other words, what model?

Is there a "simple" relation between X and Y?

Assumption #1
we expect a linear relationship

Some regression models that we can consider

$\boxed{Y=\alpha+\beta X}$ "simple linear regression"

or

$$Y=\alpha+\beta_1X+\beta_2X^2$$

or

$$\log(Y)=\alpha+\beta_1\log(X)+\beta_2\log(X)^2$$

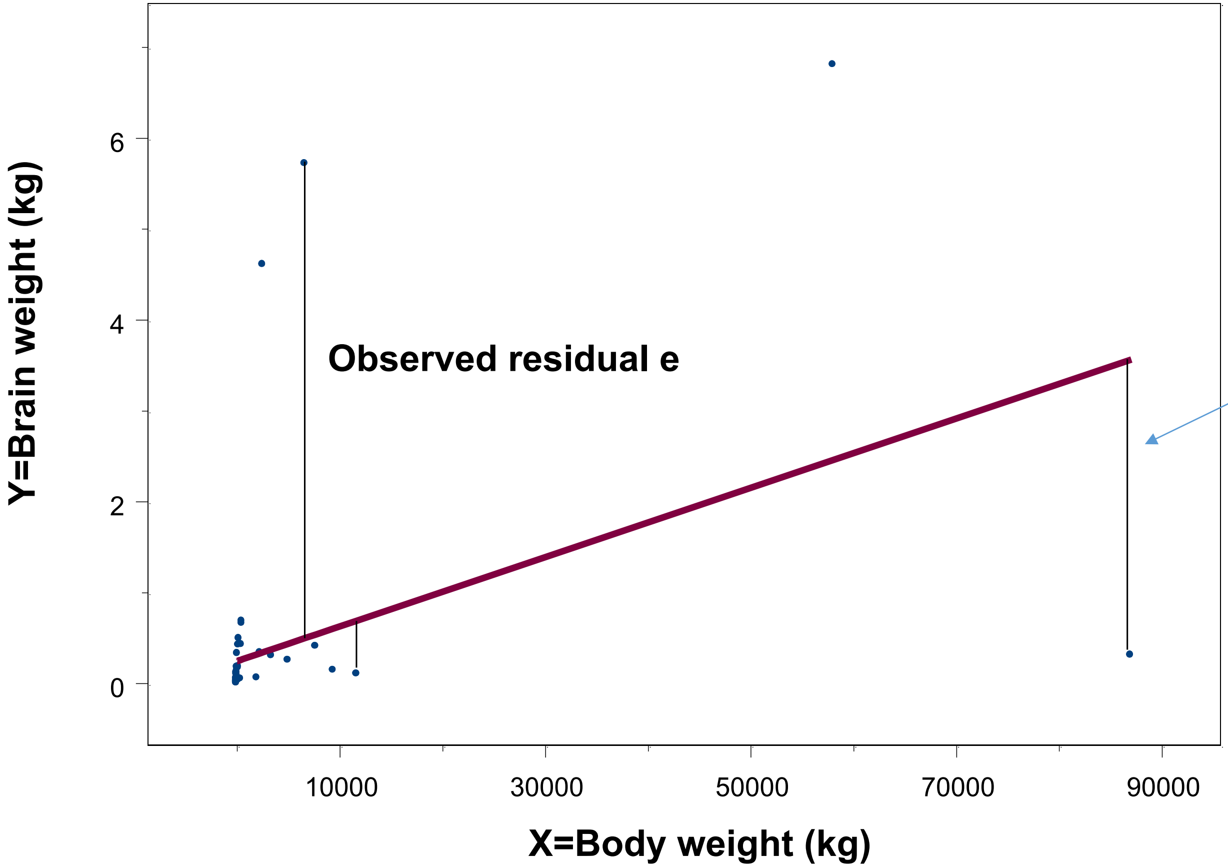
Or …..

All models are wrong, but some are useful!

We can never prove that a model is correct, but we can reject bad models!

SOME DEFINITIONS

Fitted model: Y = 0.25 + 0.000038X, here R2=16.6%.



The distance **e** between the observed y-value and the y-value predicted by the model is called the **Residual.**

The line is obtained so that the sum of e^2 is minimized.

$$R^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n} e^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

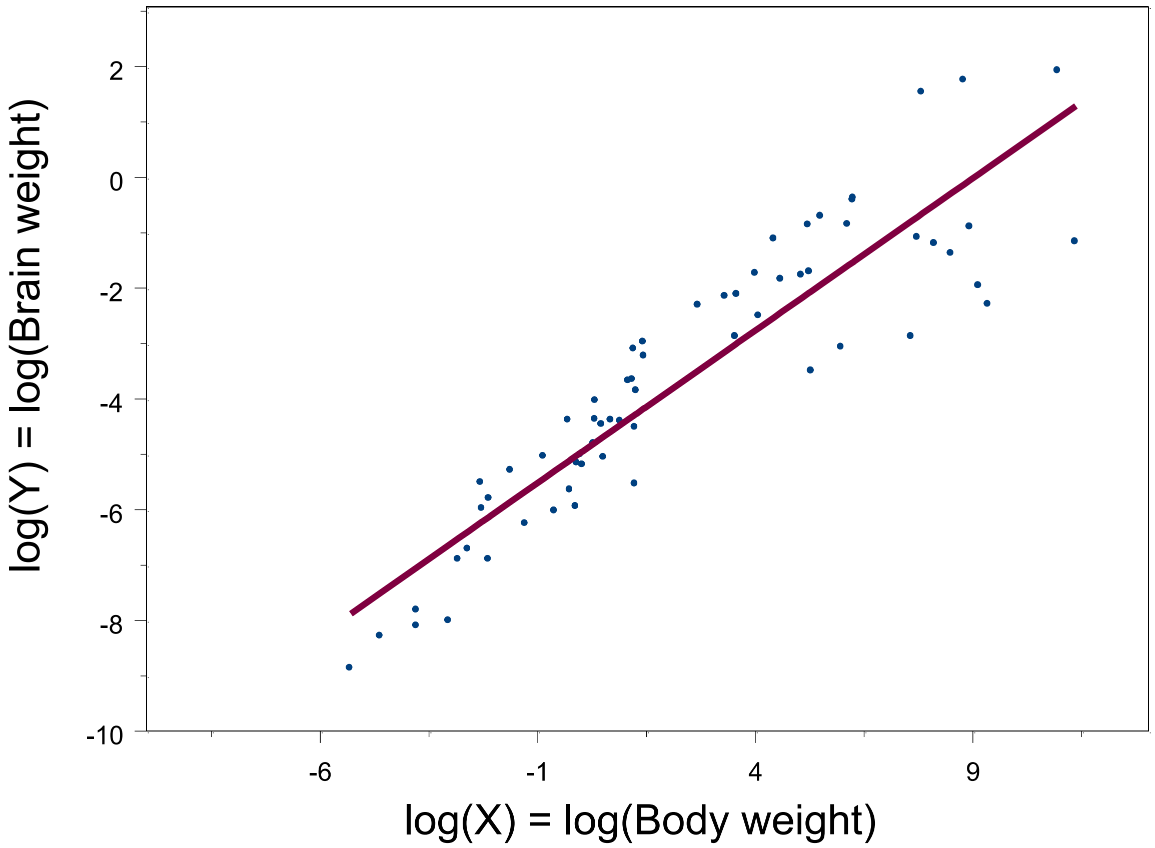
WHAT MAKES A MODEL "GOOD"?

A model is “good” if:

- All explanatory variables in the model are significant (there are exceptions!)
- The model assumptions are correct
  - The residuals are normally distributed
  - The residuals are independent
  - The variance of the residuals does not depend on the explanatory variables
- The model explains a lot of the variation in the data – i.e. R2 is high
  (R2 is a number between 0 and 100%.
  R2=56% means that 56% of the variation in the Y-variable is explained by the model.

log(y) and log(x) were highly correlated (Pearsons r=0.91). Therefore we try

log(Y) = -4.96 + 0.55 log(X)



R^2=0.83
R=0.91

P-value for log(x)
0.000..

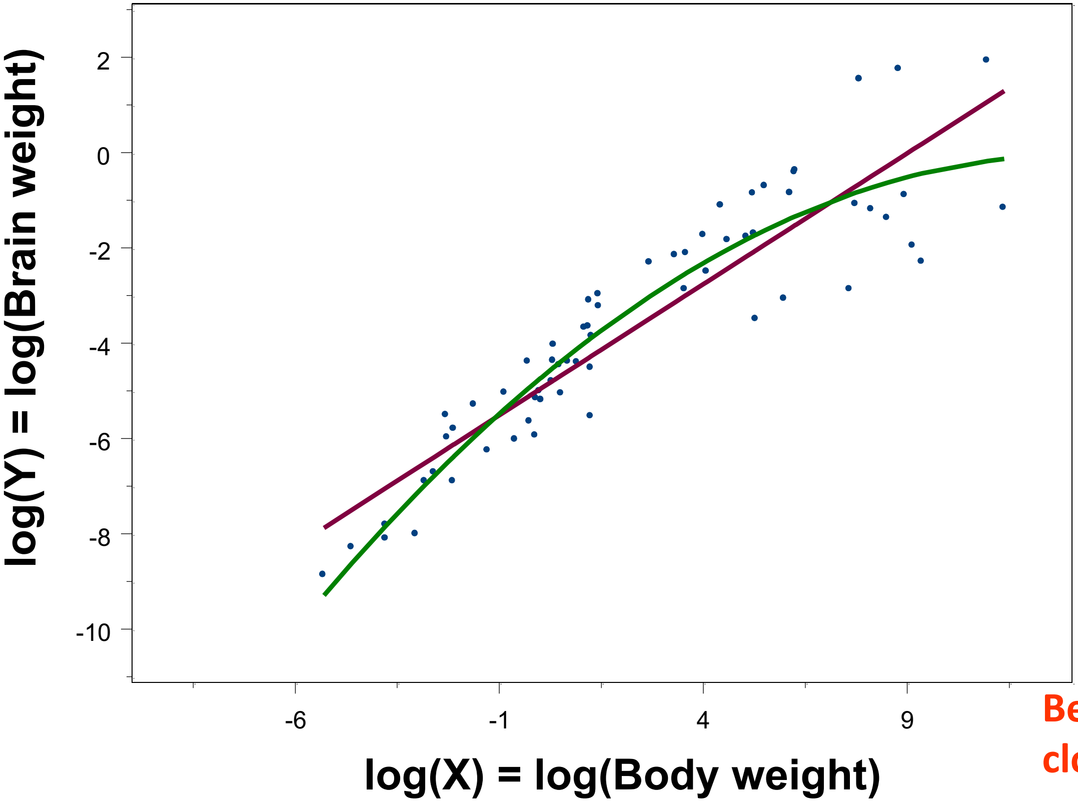
Residuals ?

Depends on log(X)

Solution 1:
include a
quadratic term.

log(Y) = -4.75 + 0.71 log(X) – 0.03 log(X)2

R^2 (adj)=0.87



p-value for
x and x^2
very low

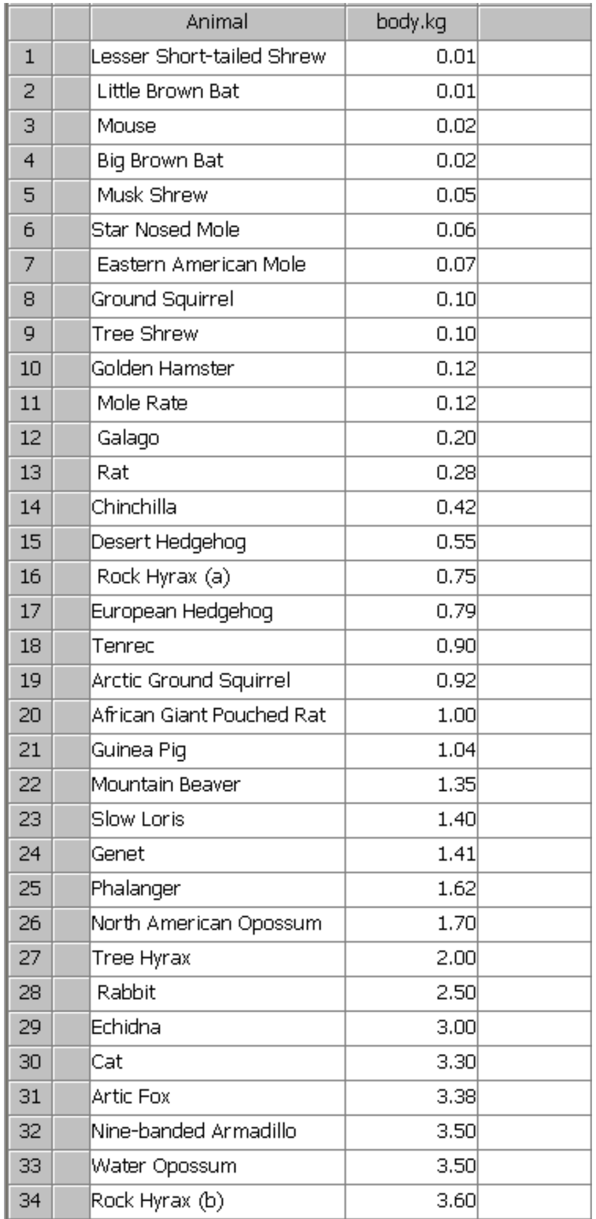
Residuals ?

Better, but the variance increases with log(X)!

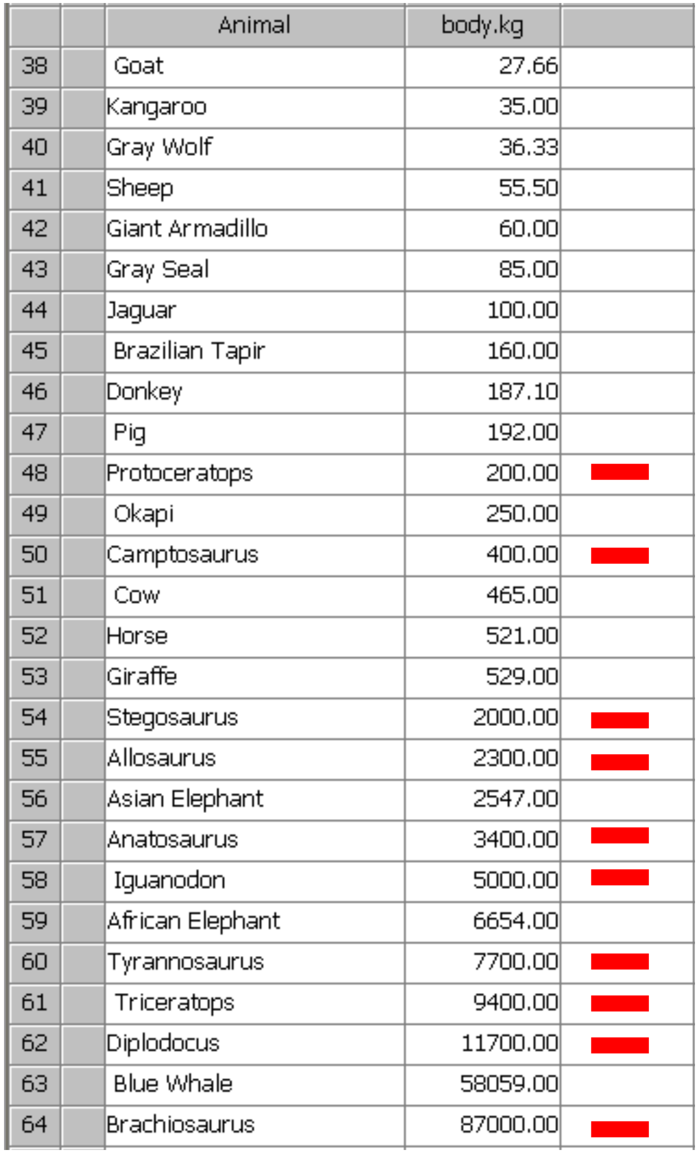
Problem with some of the heavy Animals!

Better take a closer look at the data

UN-CHALLENGED ASSUMPTION!
WHAT IS AN "ANIMAL"?

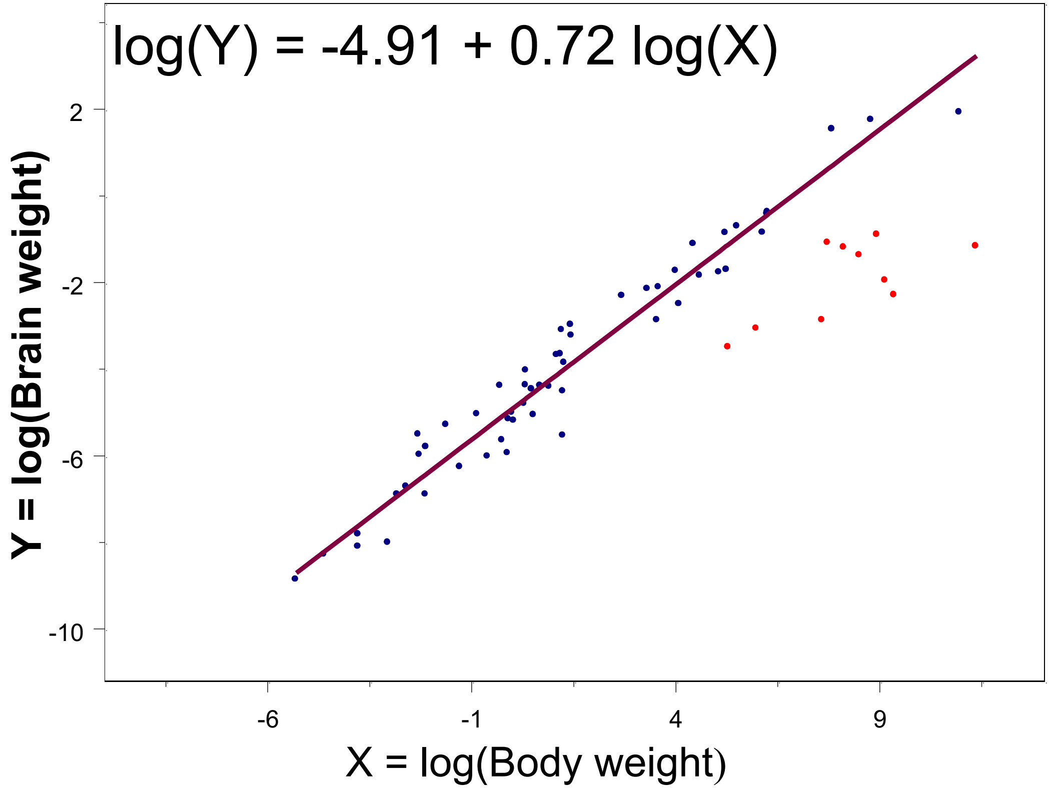


		Animal	body.kg	
1		Lesser Short-tailed Shrew	0.01	
2		Little Brown Bat	0.01	
3		Mouse	0.02	
4		Big Brown Bat	0.02	
5		Musk Shrew	0.05	
6		Star Nosed Mole	0.06	
7		Eastern American Mole	0.07	
8		Ground Squirrel	0.10	
9		Tree Shrew	0.10	
10		Golden Hamster	0.12	
11		Mole Rate	0.12	
12		Galago	0.20	
13		Rat	0.28	
14		Chinchilla	0.42	
15		Desert Hedgehog	0.55	
16		Rock Hyrax (a)	0.75	
17		European Hedgehog	0.79	
18		Tenrec	0.90	
19		Arctic Ground Squirrel	0.92	
20		African Giant Pouched Rat	1.00	
21		Guinea Pig	1.04	
22		Mountain Beaver	1.35	
23		Slow Loris	1.40	
24		Genet	1.41	
25		Phalanger	1.62	
26		North American Opossum	1.70	
27		Tree Hyrax	2.00	
28		Rabbit	2.50	
29		Echidna	3.00	
30		Cat	3.30	
31		Artic Fox	3.38	
32		Nine-banded Armadillo	3.50	
33		Water Opossum	3.50	
34		Rock Hyrax (b)	3.60	



		Animal	body.kg	
38		Goat	27.66	
39		Kangaroo	35.00	
40		Gray Wolf	36.33	
41		Sheep	55.50	
42		Giant Armadillo	60.00	
43		Gray Seal	85.00	
44		Jaguar	100.00	
45		Brazilian Tapir	160.00	
46		Donkey	187.10	
47		Pig	192.00	
48		Protoceratops	200.00	🟥
49		Okapi	250.00	
50		Camptosaurus	400.00	🟥
51		Cow	465.00	
52		Horse	521.00	
53		Giraffe	529.00	
54		Stegosaurus	2000.00	🟥
55		Allosaurus	2300.00	🟥
56		Asian Elephant	2547.00	
57		Anatosaurus	3400.00	🟥
58		Iguanodon	5000.00	🟥
59		African Elephant	6654.00	
60		Tyrannosaurus	7700.00	🟥
61		Triceratops	9400.00	🟥
62		Diplodocus	11700.00	🟥
63		Blue Whale	58059.00	
64		Brachiosaurus	87000.00	🟥

The data suggest that dinosaurs had, relatively to their size, small brains.
One single model can not explain the relationship, we need to exclude the dinosaurs.

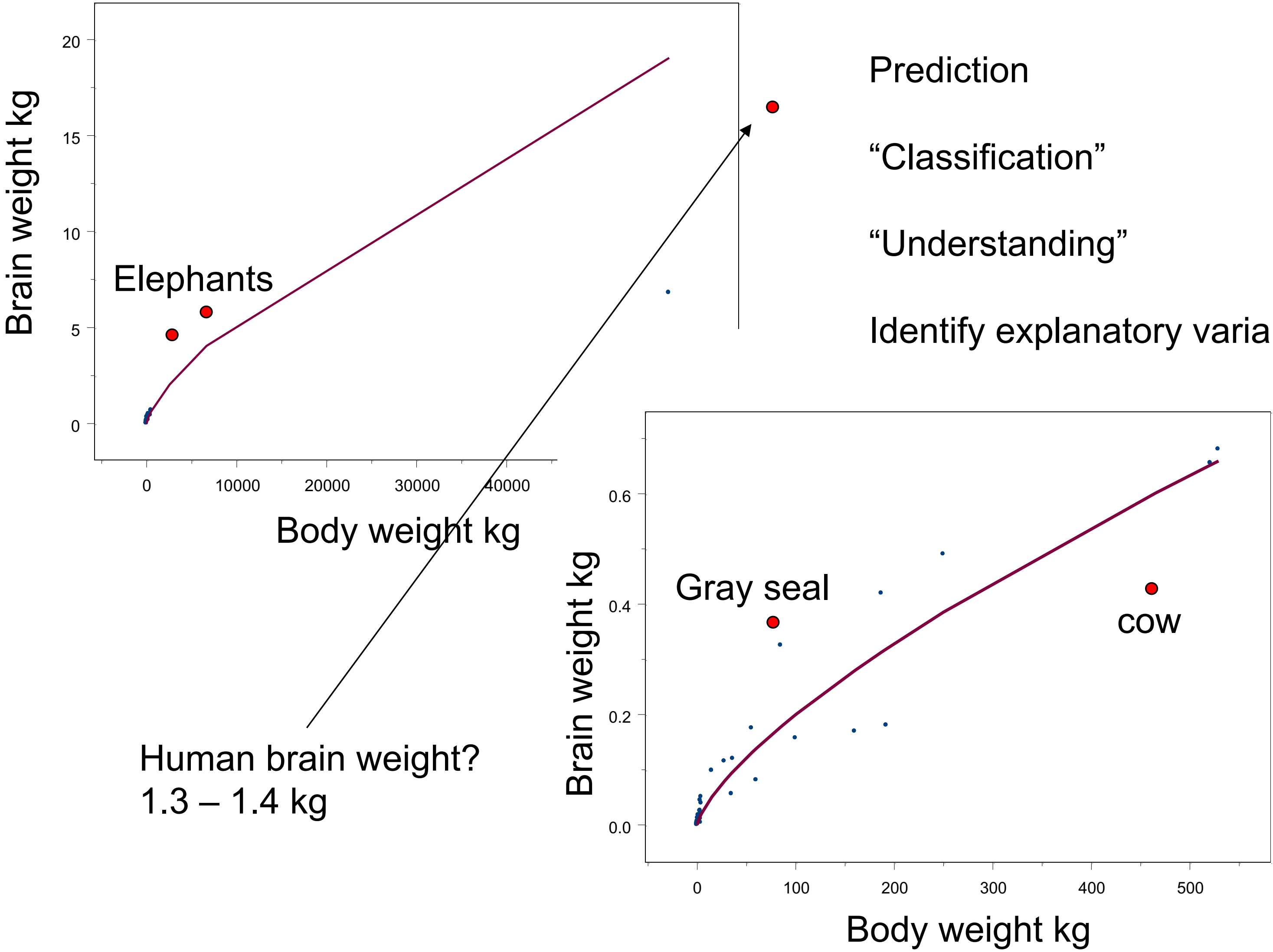


R^2 =0.95
Cor = R = 0.975

P-value for log(x) very low

The residuals looks good!

Use of a model



Human brain weight?
1.3 – 1.4 kg

Prediction

“Classification”

“Understanding”

Identify explanatory variables

Follow up: Sad but often true

- Watch the video:
  - Biologist talks to statistician: http://www.youtube.com/watch?v=Hz1fyhVOjr4
- Reflect on the issues related to study design as satirically presented in this video

Learning Objectives

- By the end of the lecture, you will be able to:
  - reflect on the importance of the studies design for an experiment
  - describe the basic statistical concepts necessary for data analysis