Intro to Linear regression

STA9750 / Baruch College

Spring 2019

Simple linear regression

Simple linear regression

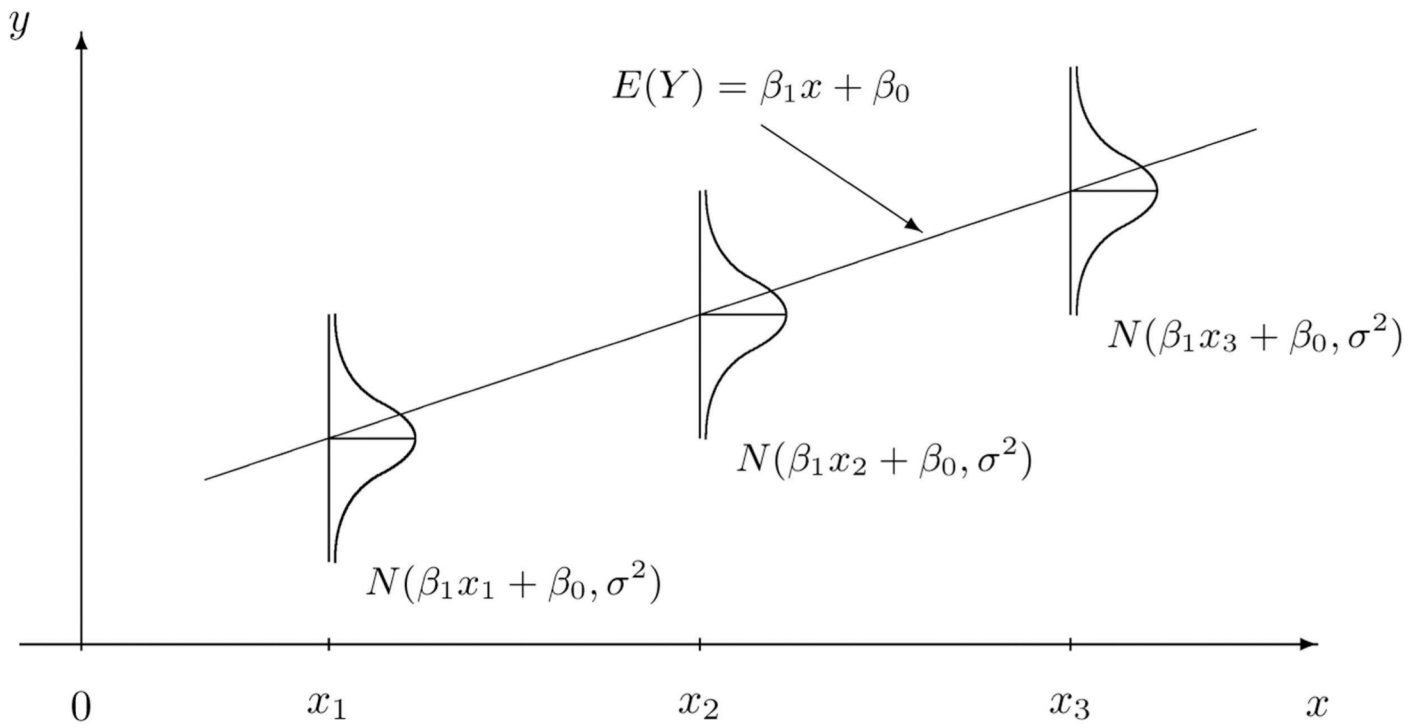
linear trend + normal noise

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

Assumptions on ε_i:

- Independence
- Normality
- Homoscedasticity: equal variance across observations, which doesn't depend on x_i

Also, linearity: *E(Y | X)* is a line



Source: https://stats.libretexts.org/Textbook_Maps/Introductory_Statistics/Book%3A_Introductory_Statistics_(Shafer_and_Zhang)

How do we check assumptions?

- Since

$$\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i) \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

… then, if the assumptions are satisfied:

$$e_i = y_i - (b_0 + b_1 x_i) \overset{\text{iid}}{\approx} N(0, s^2)$$

Assumptions:

1. Independence of outcomes y_i for i in 1:n (given the x_i).
2. Normality
3. Homoscedasticity (equal variance across observations, which doesn't depend on x_i)
4. Of course, linearity

How to check them:

1. Check if e_i are *strongly* correlated (e.g. serial correlation, if observations are taken over time)
2. Q-Q plot of e_i
3. Scatterplot of e_i vs $b_0+b_1x_i$
4. Scatterplot of e_i vs $b_0+b_1x_i$

Transformations

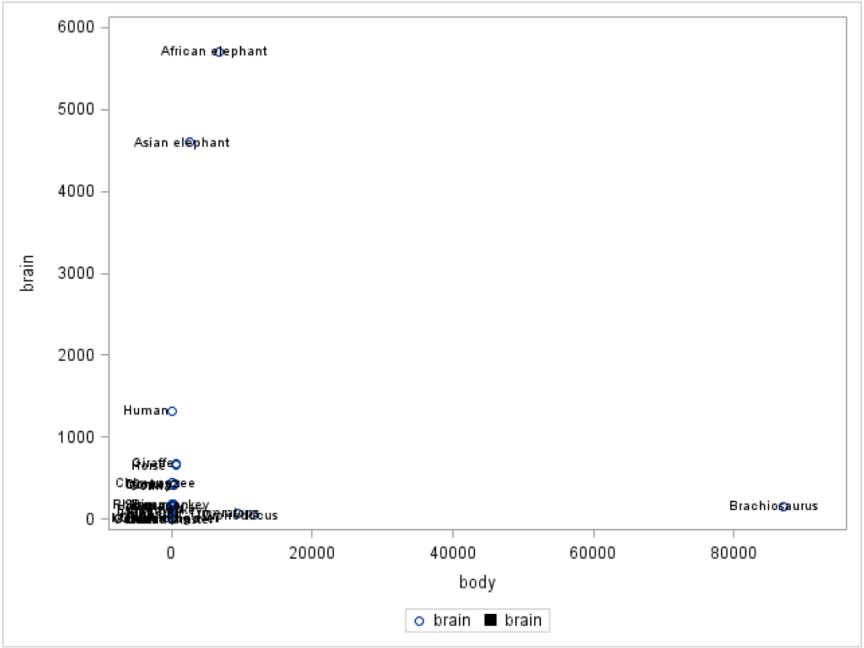
Transformations

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

- What should we do if the relationship in our scatterplot doesn't look linear?
- Take *y* and *x* to be functions (transformations) of the original variables of interest
- Most popular transformations:
  - log, square-root, square

Example:

- In "animals.csv", we want to predict brain weights given body weights
- Original relationship doesn't look linear

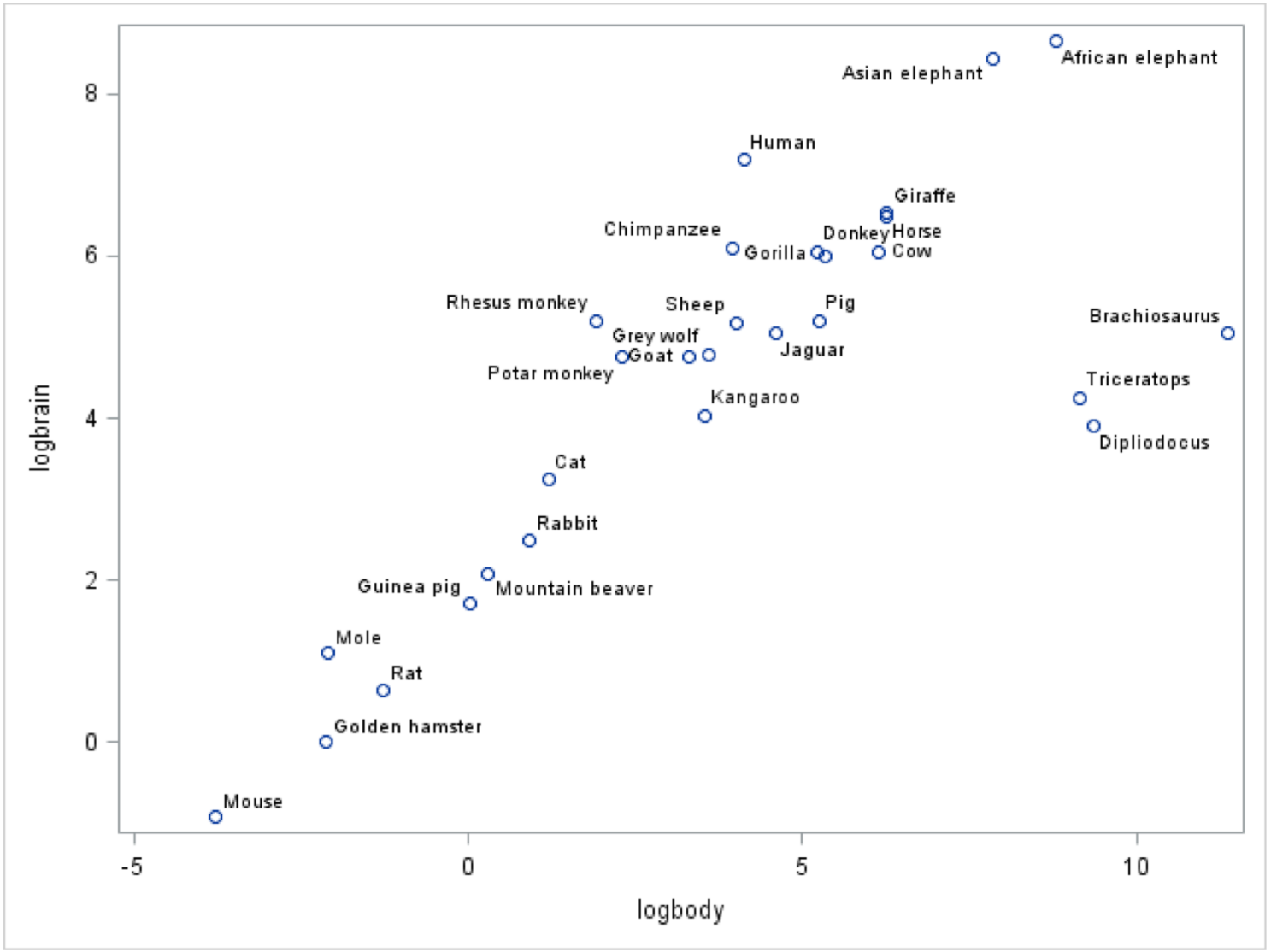


- **Goal:** find functions f and g such that

$$f(\text{brain weight}_i) = \beta_0 + \beta_1\, g(\text{body weight}_i) + \varepsilon_i$$

- If f(x) = g(x) = log(x)…

How to interpret models with logged outcomes and/or predictors? Click here



Influential and high-leverage observations, outliers

Influential observations

- Idea: how much does my fit change after taking out this observation?
- There are different ways to measure this
- For example: Cook's distance, DFFITS, etc.

Cook's distance

- Cook's distance of observation *i* is

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{y}_j - \hat{y}_{j(i)}\right)^2}{ps^2}$$

$\hat{y}_j$ predicted value for observation j with all the observations

$\hat{y}_{j(i)}$ predicted value for observation j after taking out the i-th observation

s^2 our usual estimator of the residual variance σ^2

- *How big is big?* Different recommendations… Some people say $D_i > 1$
- I recommend looking closely at any observation that seems to "stick out"

Leverage, outliers, and influence

- Leverage: measures how far away x_i is from the other x values [goes from 0 to 1, from "average x" to "very unusual x"]
- High leverage: unusual value of x_i, which may or may not be well predicted by our line
- Big residual $|e_i|$: point that is badly predicted by our line (outliers)
- Observations with high leverage and big residuals are highly influential… Cook's distance can be written as

$$D_i = \frac{e_i^2}{s^2 p}\left[\frac{h_i}{(1-h_i)^2}\right]$$

h_i: leverage of observation i
e_i: residual of observation i

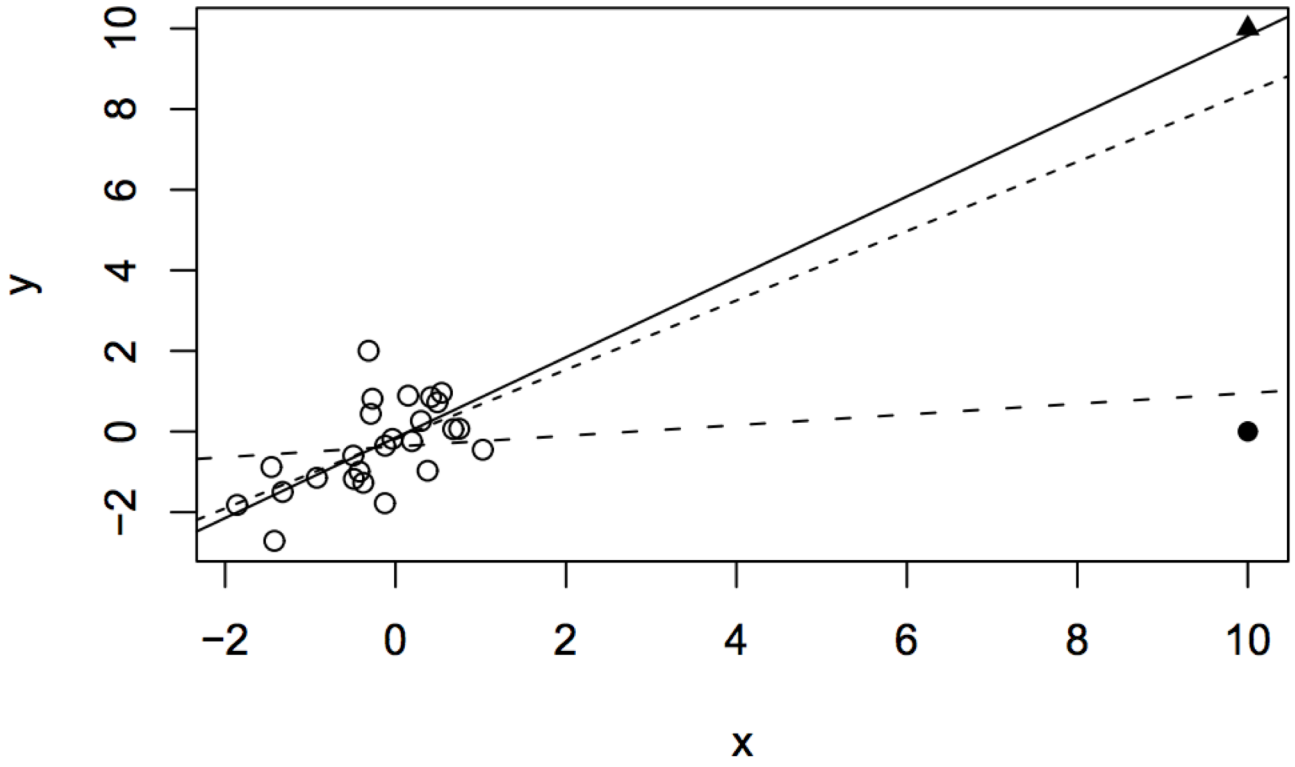


Figure 7.2: Outliers can conceal themselves. The solid line is the fit including the ▲ point but not the • point. The dotted line is the fit without either additional point and the dashed line is the fit with the • point but not the ▲ point.

Source: https://cran.r-project.org/doc/contrib/Faraway-PRA.pdf

Multiple linear regression

Multiple linear regression

- The same, but with more variables
- Find the coefficients that minimize in-sample predictive error
- We can find CIs and hypothesis tests if we make assumptions
- We assume

$$y_i \overset{\text{ind}}{\sim} N(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1}, \sigma^2)$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \varepsilon_i, \ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

- Independence of outcomes y_i for i in 1:n (given the x_{ij}).
- Normality
- Homoscedasticity (equal variance across observations, which doesn't depend on x_{ij})
- Linearity (i.e. E[Y | X] is a linear comb. of the Xs)

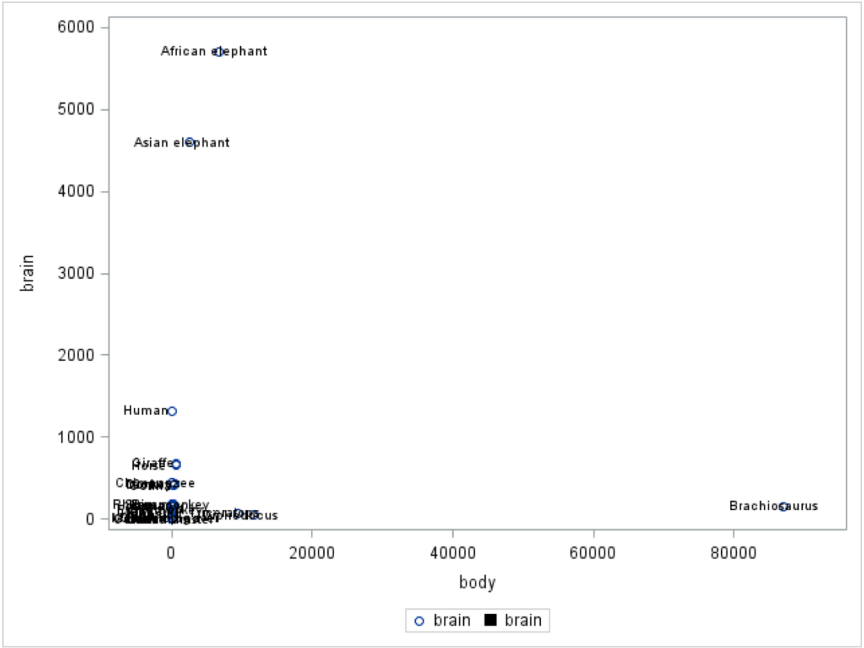
Transformations

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \ \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

- What should we do if the relationship in our scatterplot doesn't look linear?
- Take *y* and *x* to be functions (transformations) of the original variables of interest
- Most popular transformations:
  - log, square-root, square

Example:

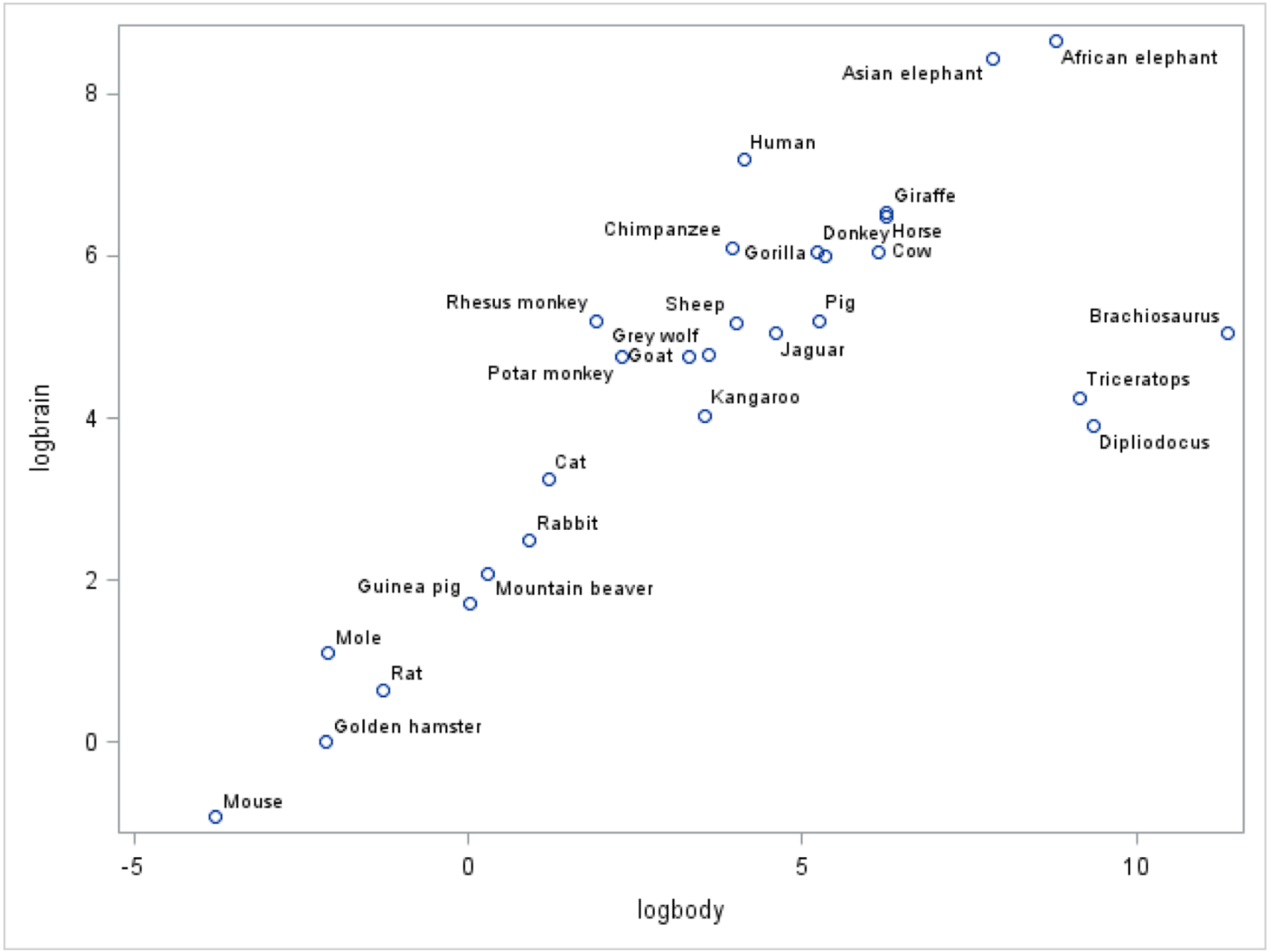
- In "animals.csv", we want to predict brain weights given body weights
- Original relationship doesn't look linear



- **Goal:** find functions f and g such that

$$f(\text{brain weight}_i) = \beta_0 + \beta_1\, g(\text{body weight}_i) + \varepsilon_i$$

- If f(x) = g(x) = log(x)…


African elephant
Asian elephant
Human
Giraffe
Chimpanzee
Donkey
Horse
Gorilla
Cow
Rhesus monkey
Sheep
Pig
Brachiosaurus
Grey wolf
Jaguar
Goat
Potar monkey
Triceratops
Kangaroo
Dipliodocus
Cat
Rabbit
Guinea pig
Mountain beaver
Mole
Rat
Golden hamster
Mouse
8
6
4
2
0
-5
0
5
10
logbrain
logbody