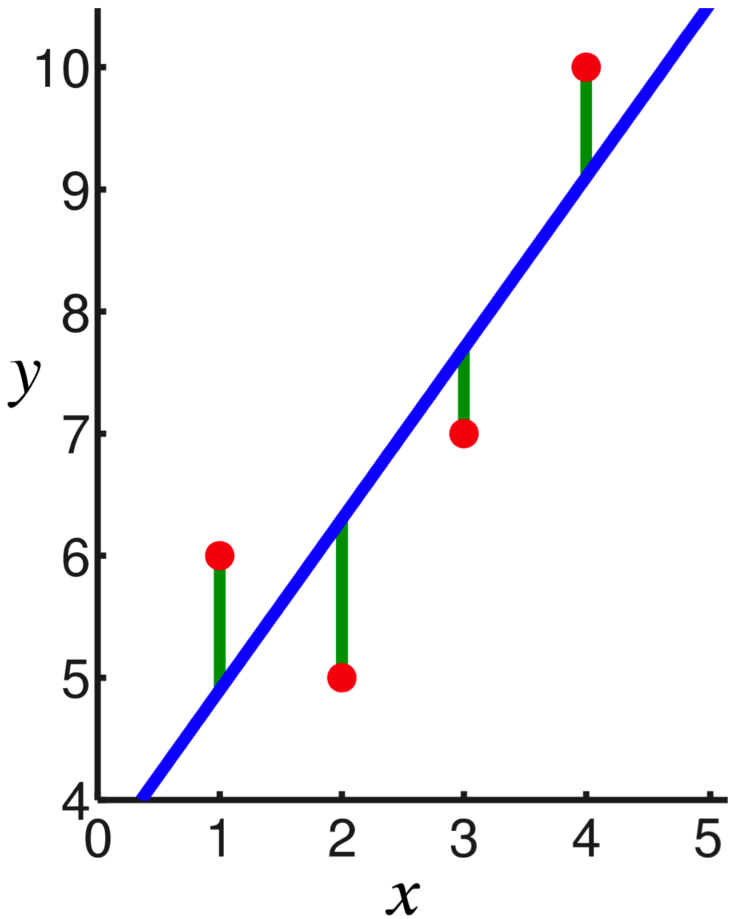
Intro to
Linear regression

STA9750 / Baruch College

Spring 2018

Goal: Find best line to predict *y* given *x*



Equation of a line: $y = b_0 + b_1 x$

Only need to specify b_0 and b_1

For each data point x_i, y_i:

Observed values: y_i

Predicted values: $\hat{y}_i = b_0 + b_1 x_i$

Prediction error: $e_i = y_i - \hat{y}_i$

Least squares line:

Minimize sum of squared prediction errors

That is, find $\mathbf{b_0}$ and $\mathbf{b_1}$ that minimize

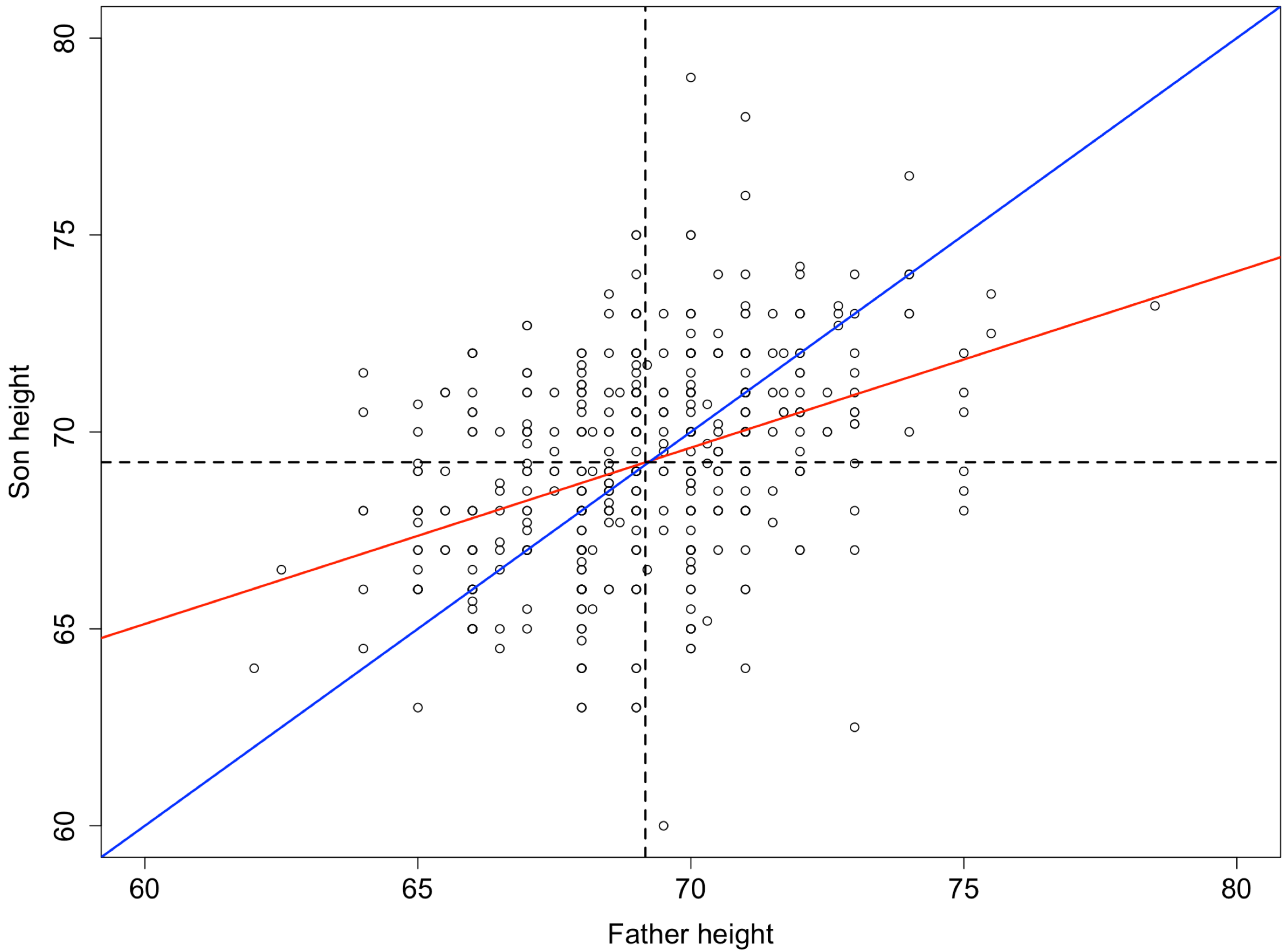
$$e_1^2 + e_2^2 + e_3^2 + e_4^2$$

Galton’s example

- In 1886, Galton published a study where he compared heights of fathers and sons

Red line: least squares line

Blue line: $y = x$ [Son height = Father height]



- If your father is tall, you're likely to be tall, but shorter than he is
- If your father is short, you're likely to be short, but taller than he is

That is, if your father is at the extremes, you're likely to "regress" to the overall population mean

Coefficient of determination: R^2

- R^2 is commonly used for quantifying the "strength" of the least squares line and it is simply

$$R^2 = r^2$$

- It can be interpreted as the fraction of the total variability in *y* that is explained by the regression line
- It is between 0 and 1 (perfect linear relationship)

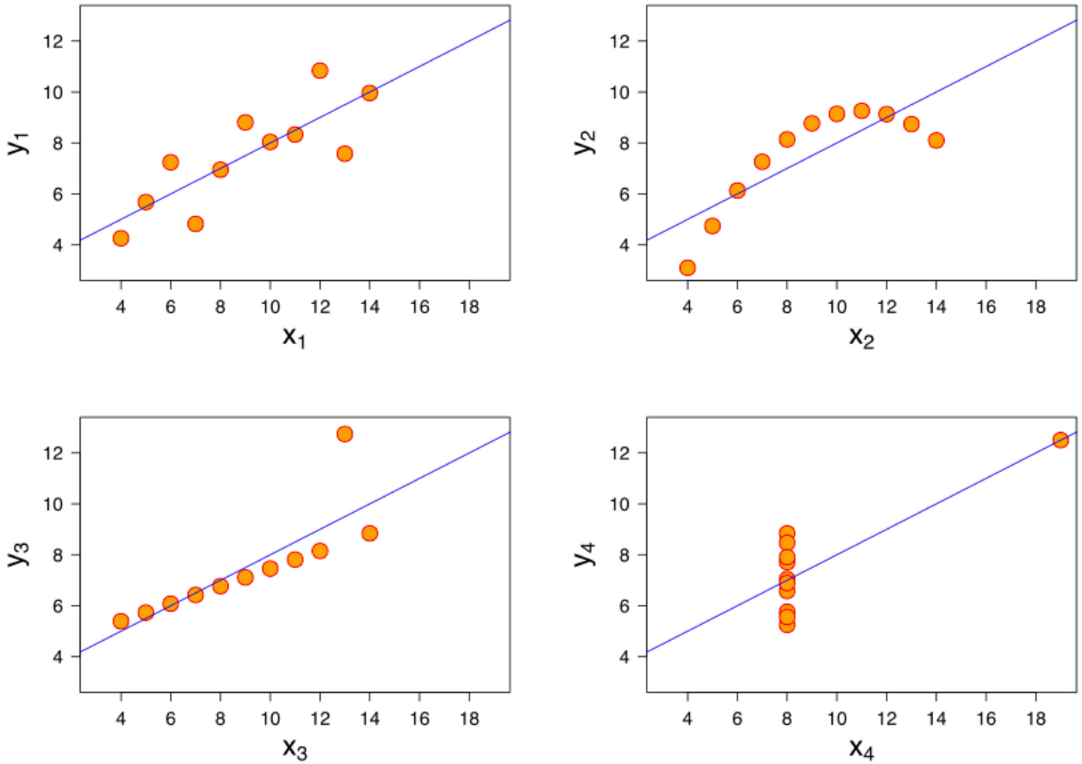
variability in y | squared pred. errors | variability in predictions

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

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

- It's easy to use: it goes from 0 to 1
- Tempting to use it as a "goodness-of-fit" statistic
- However, it can be highly deceptive when the relationship between y and x isn't linear

Anscombe's quartet



All datasets have $R^2 = 0.67$

… But vastly different stories!

https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Inference?

- So far, we haven’t made any distributional assumptions
- We just found the “best” line
- If we make some assumptions, we’ll be able to find predictive intervals and do hypothesis tests

Simple linear regression

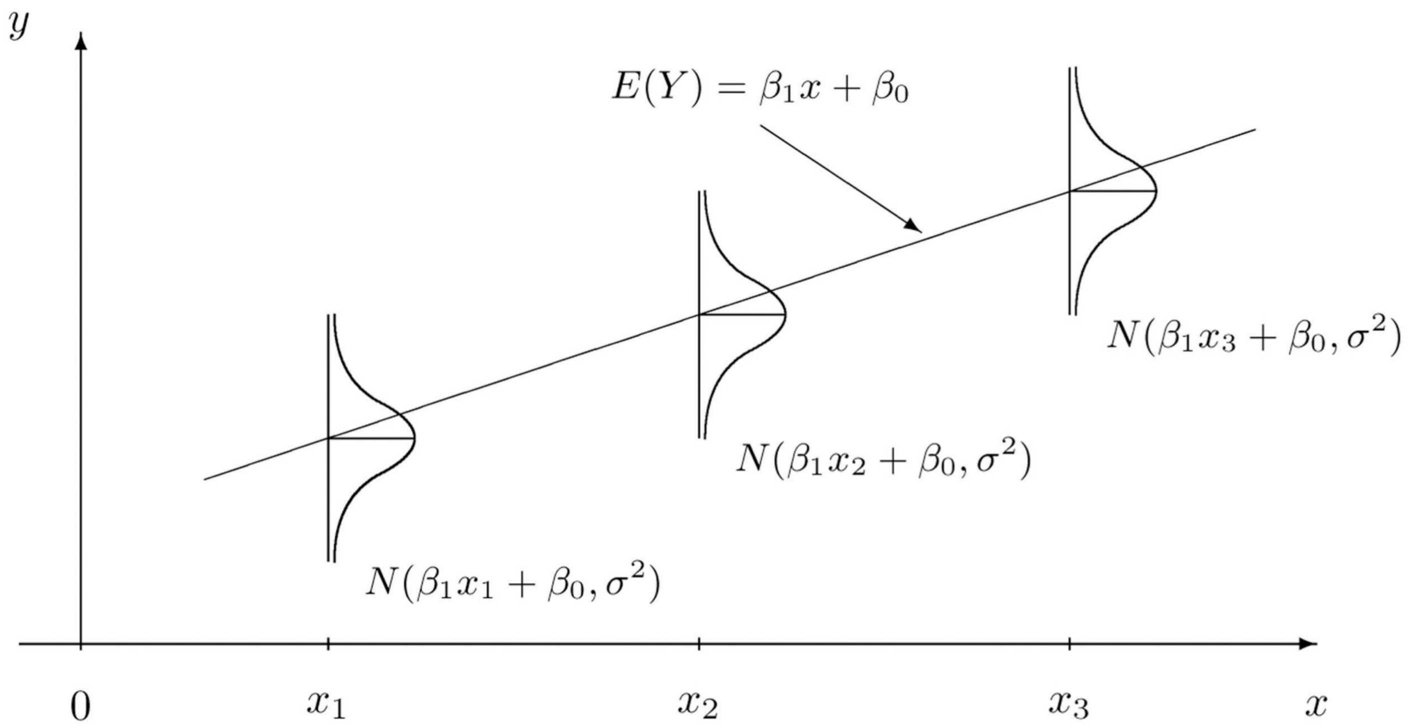
linear trend + normal noise

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

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:

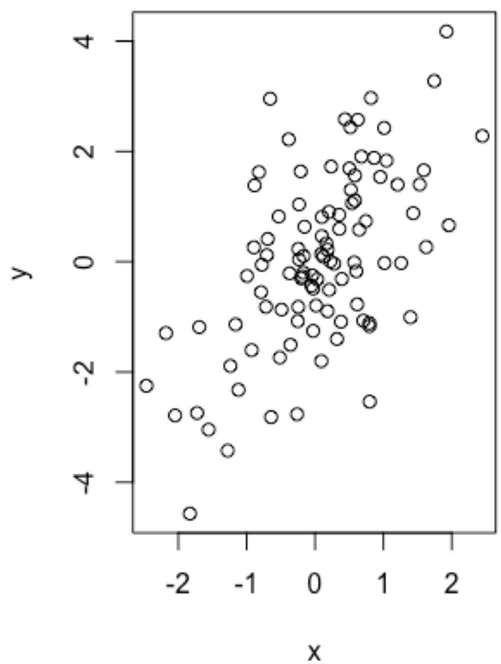
1. Independence of outcomes y_i for i in 1:n (given the x_i).
2. Normality
3. Homoscedasticity (equal variance across observations, which doesn't depend on x_i)
4. Of course, linearity

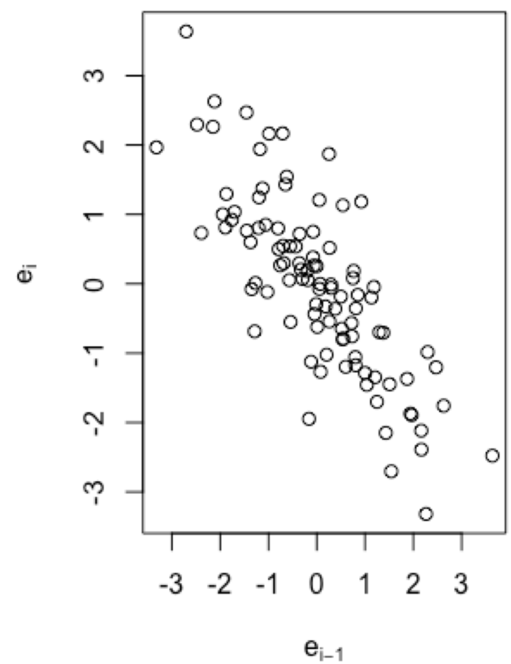
How to check them:

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 $b_0+b_1x_i$

Independence?

- Hard to check unless data are collected over time or there are clear "groups" or variables that were not included in the regression

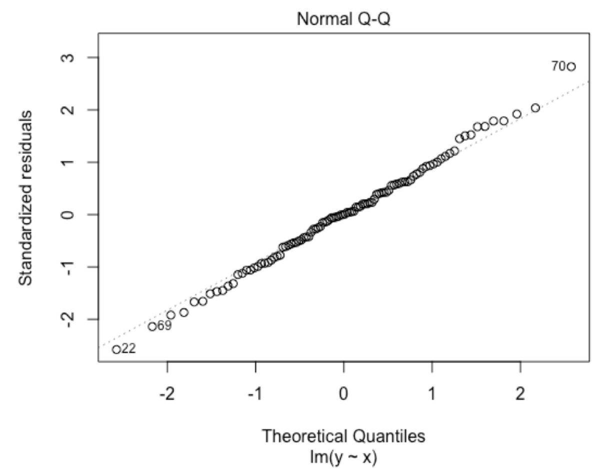


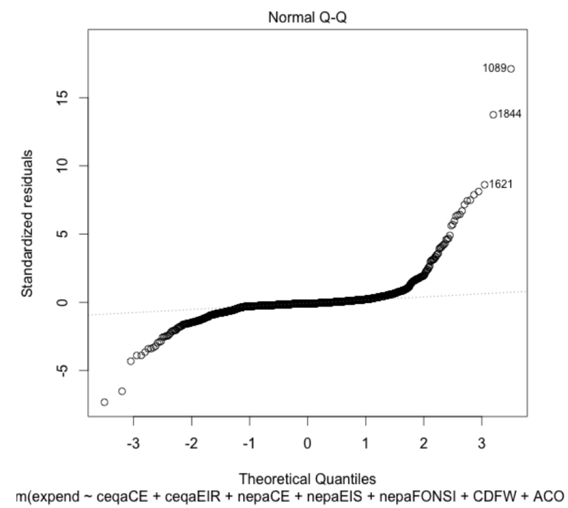


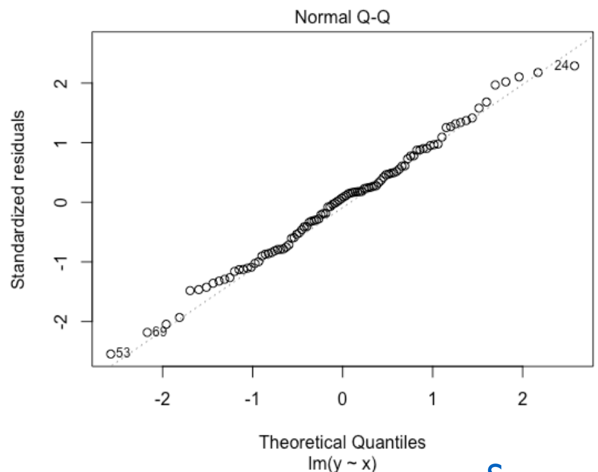
Normality? **Q-Q plot: see if it is roughly linear**

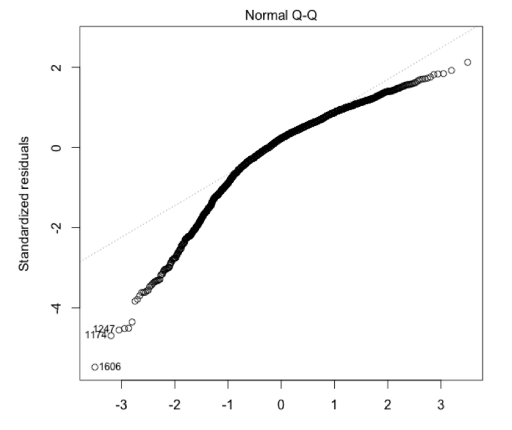
OK

Bad







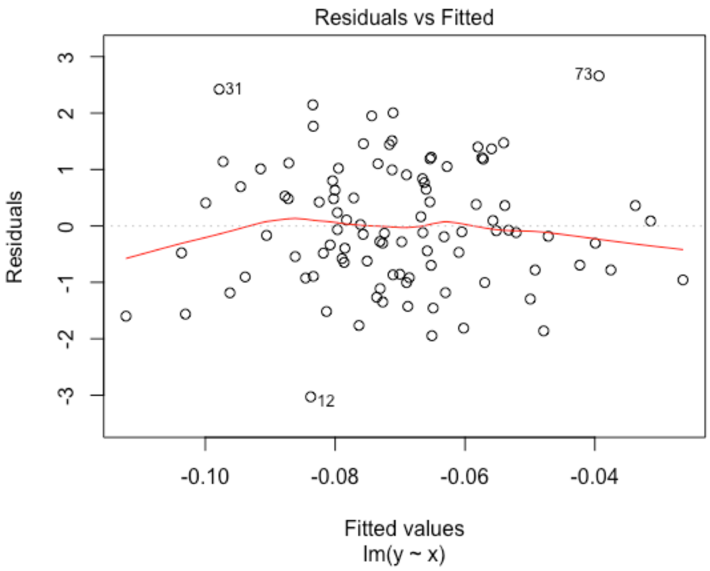


[Source of bad QQ-plots: https://stats.stackexchange.com/questions/160562/what-to-do-if-residual-plot-looks-good-but-qq-plot-doesnt-after-transforming-t](https://stats.stackexchange.com/questions/160562/what-to-do-if-residual-plot-looks-good-but-qq-plot-doesnt-after-transforming-t)

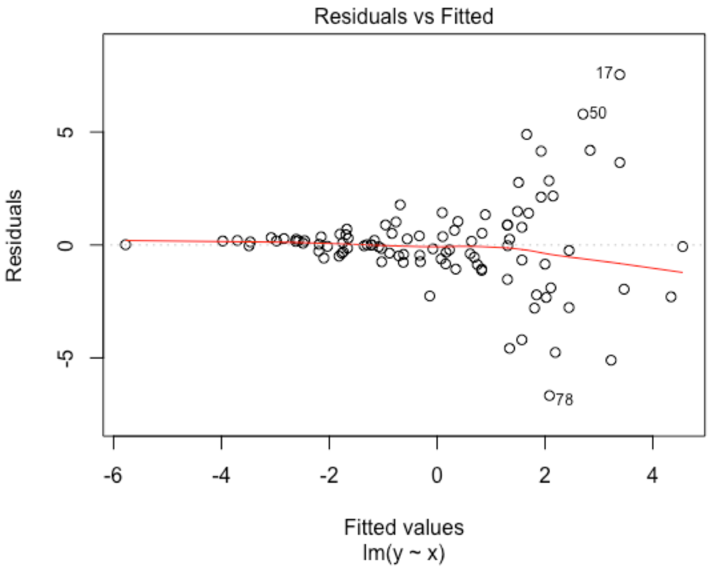
Homoscedasticity?

Constant spread in scatterplot of e_i vs $b_0+b_1x_i$

OK

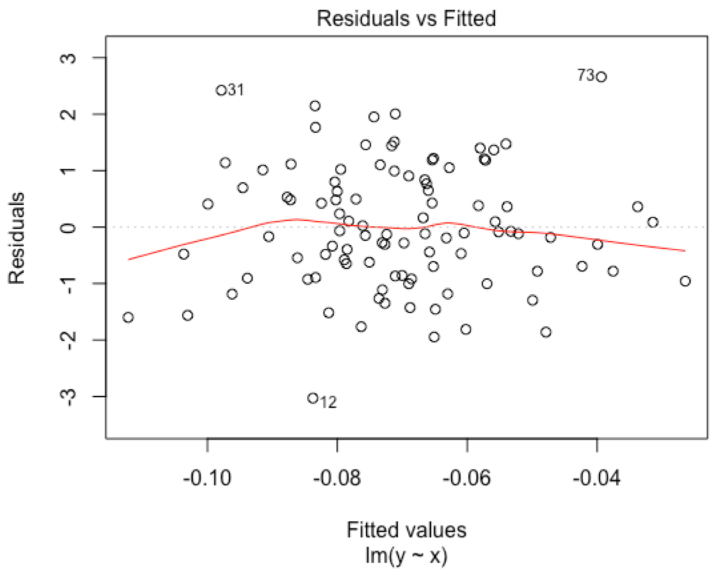


Bad

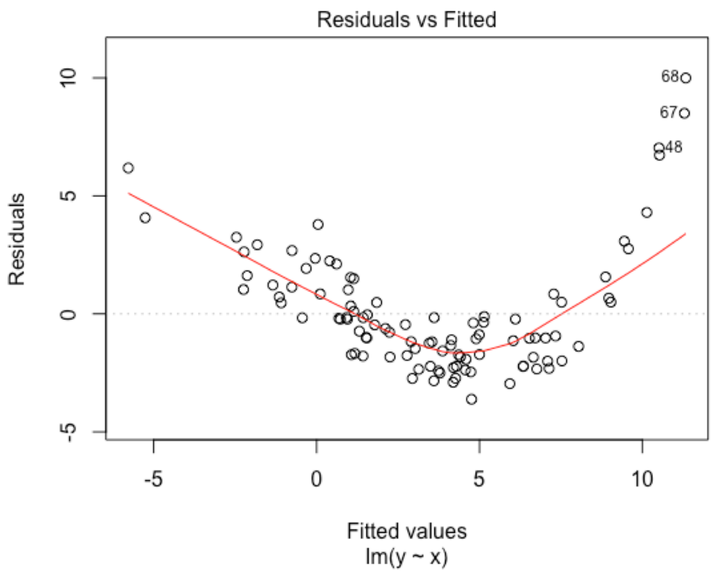


Linearity? No obvious patterns in scatterplot of e_i vs $b_0+b_1x_i$

OK



Bad



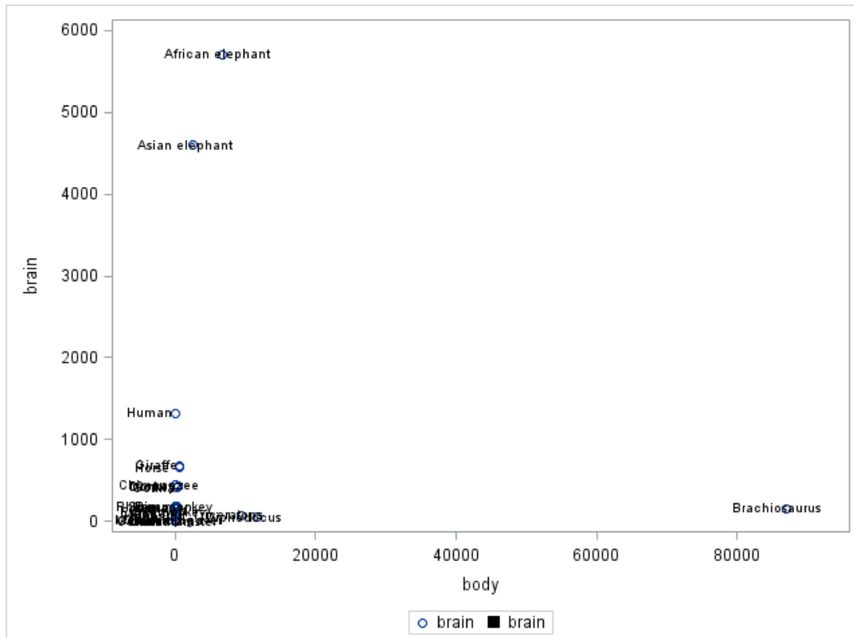
Transformations

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

- 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:

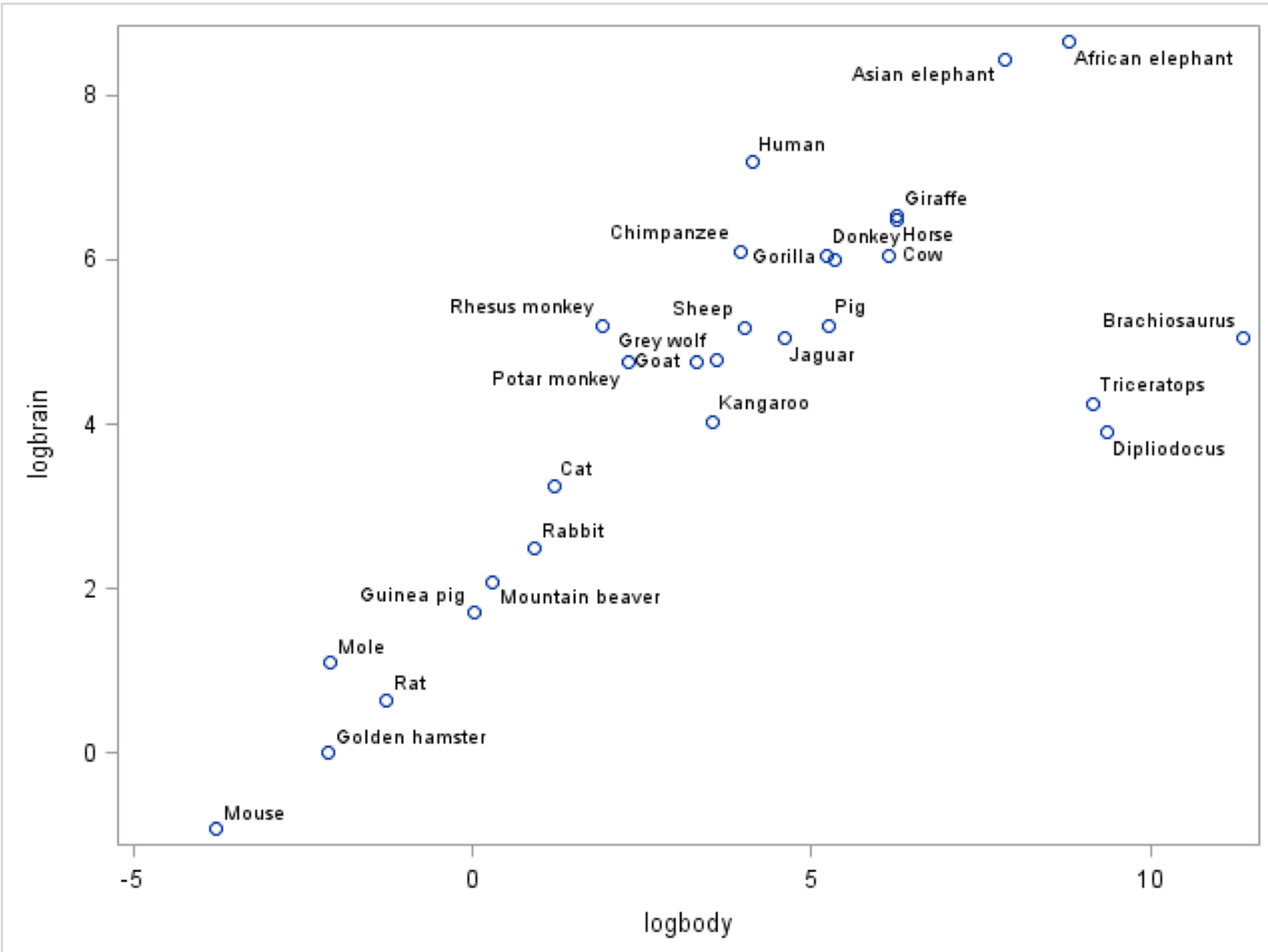
- In "animals.csv", we want to predict brain weights given body weights
- Original relationship doesn't look linear



- **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
Goat
Jaguar
Potar monkey
Triceratops
Kangaroo
Dipliodocus
Cat
Rabbit
Guinea pig
Mountain beaver
Mole
Rat
Golden hamster
Mouse
logbrain
logbody
8
6
4
2
0
-5
0
5
10