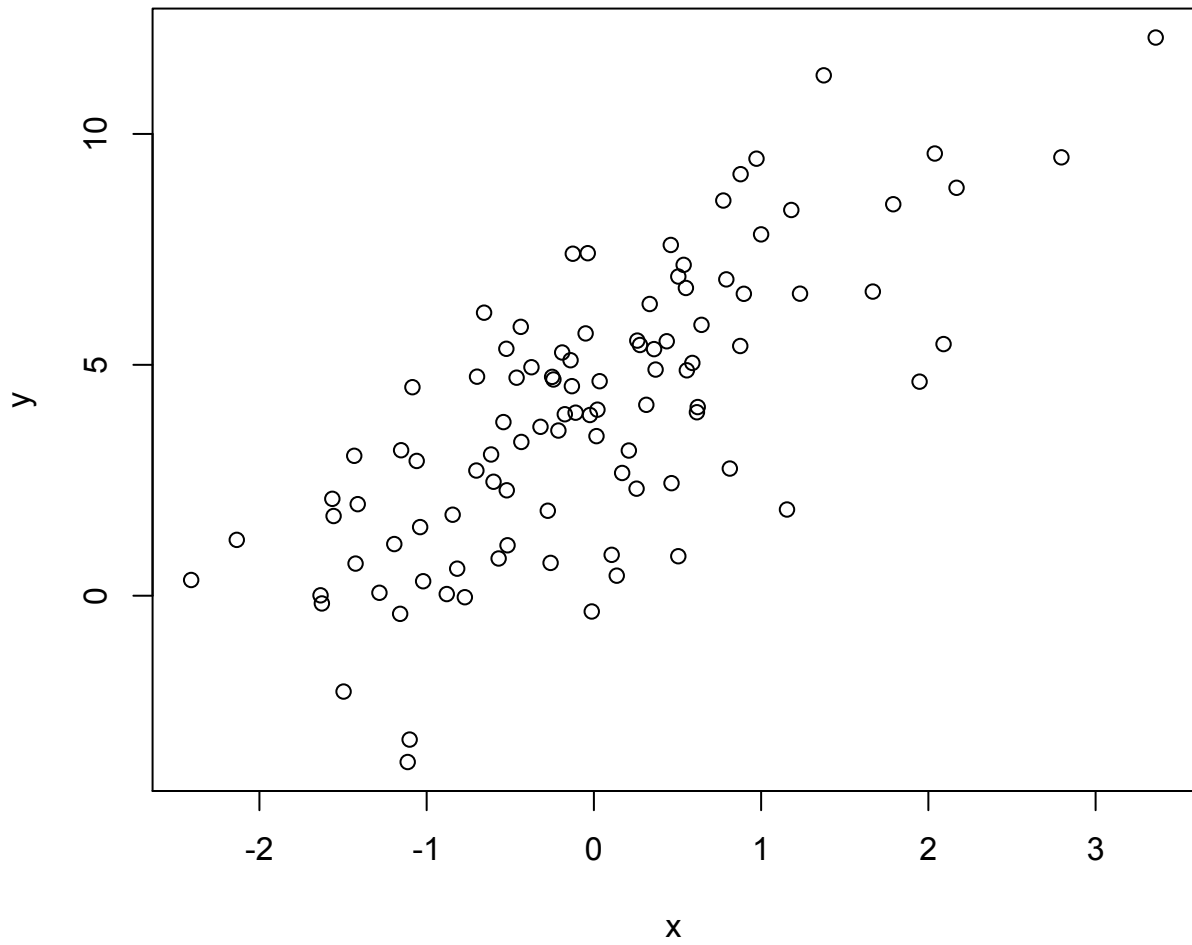


Correlation and Linear Regression

Quantitative Understanding in Biology, 2.1

Question:

**You are making paired measurements.
How do you know if the measurements
are related?**



There seems to be a relationship between x and y . We would like to quantify that relationship, often known as the **correlation.**

Question:

How do you calculate the correlation?

Pearson's Correlation Coefficient

$$r = \frac{1}{n} \sum_{i=1}^n \frac{(x_i - \bar{x})(y_i - \bar{y})}{SD_x SD_y}$$

Pearson's Correlation Coefficient

$$Z_x(x_i) = \frac{(x_i - \bar{x})}{SD_x}$$

Pearson's Correlation Coefficient

$$Z_x(x_i) = \frac{(x_i - \bar{x})}{SD_x}$$

Z-score

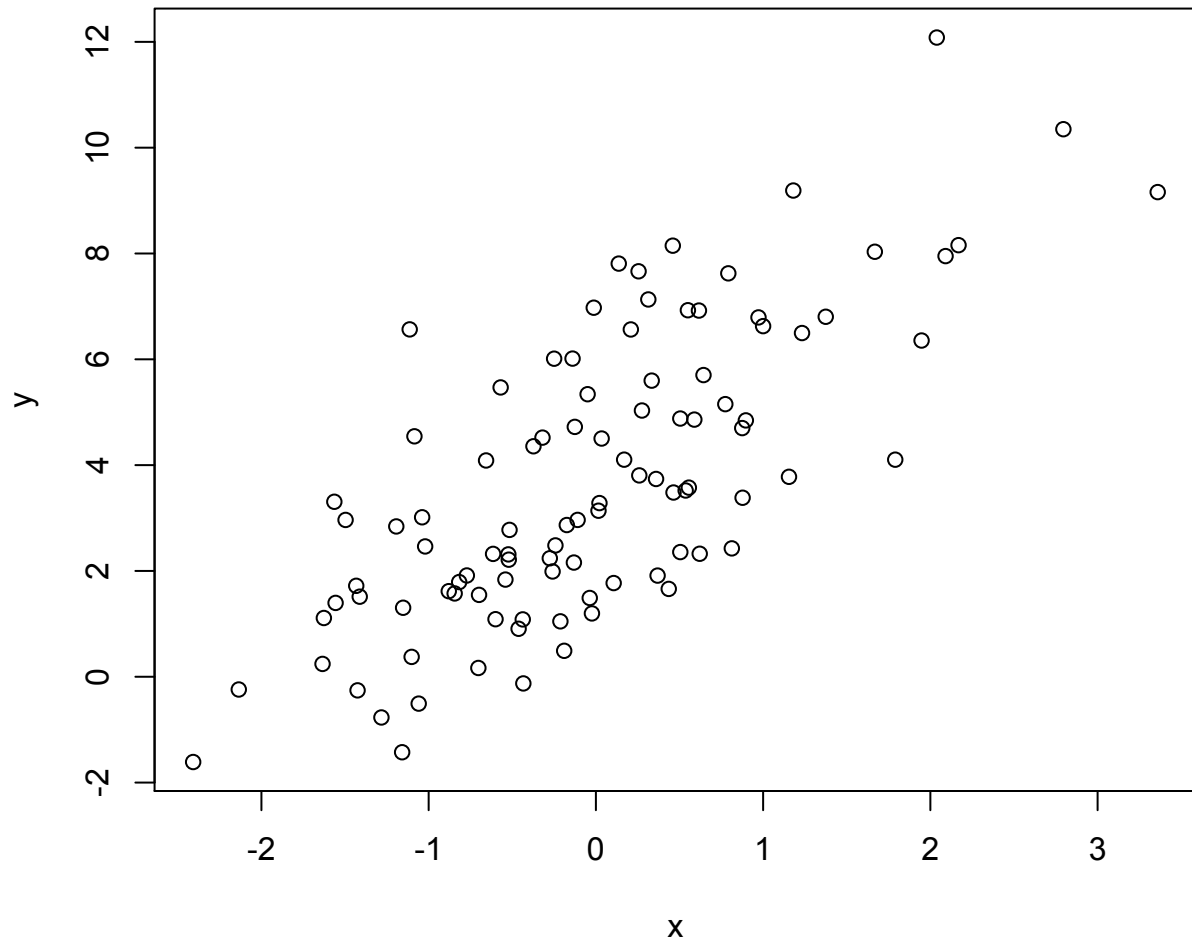


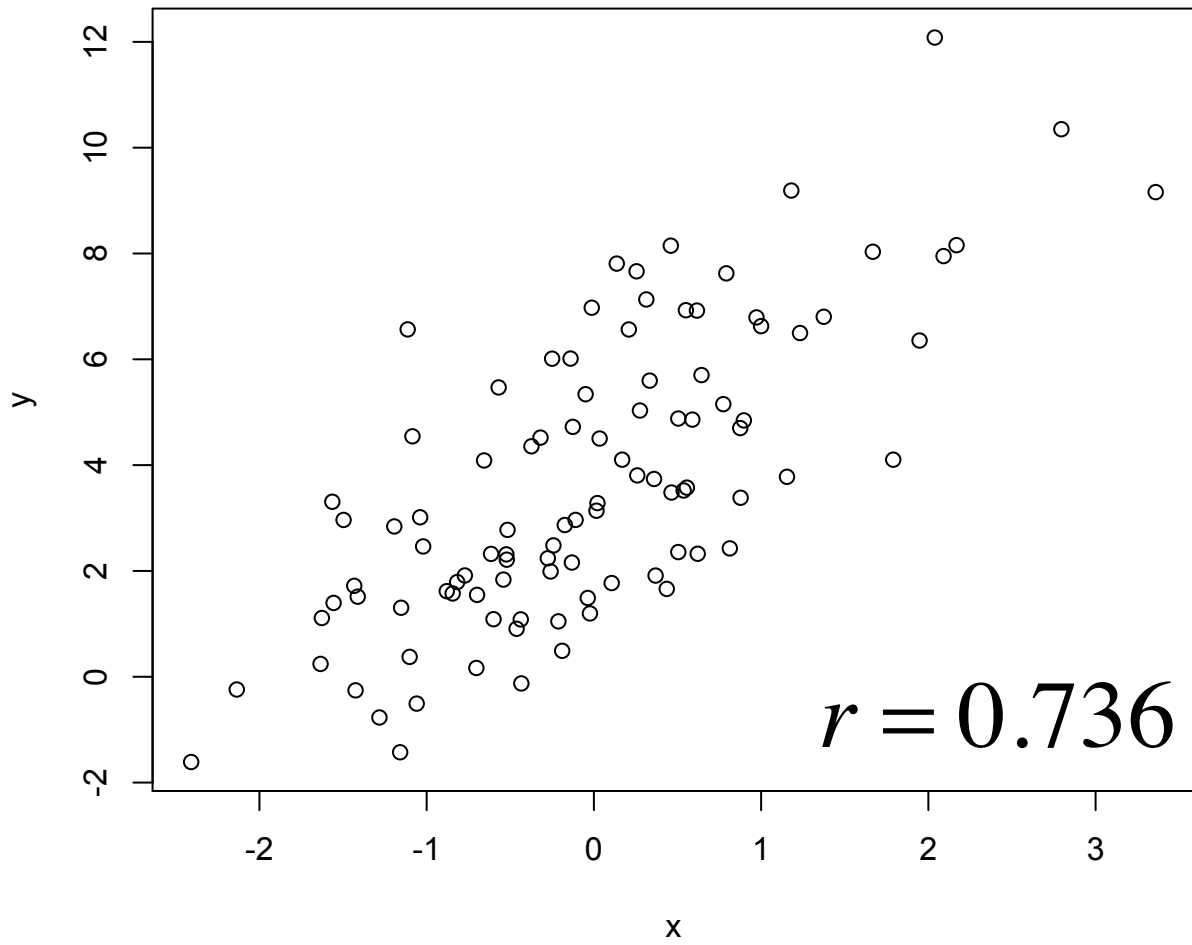
Pearson's Correlation Coefficient

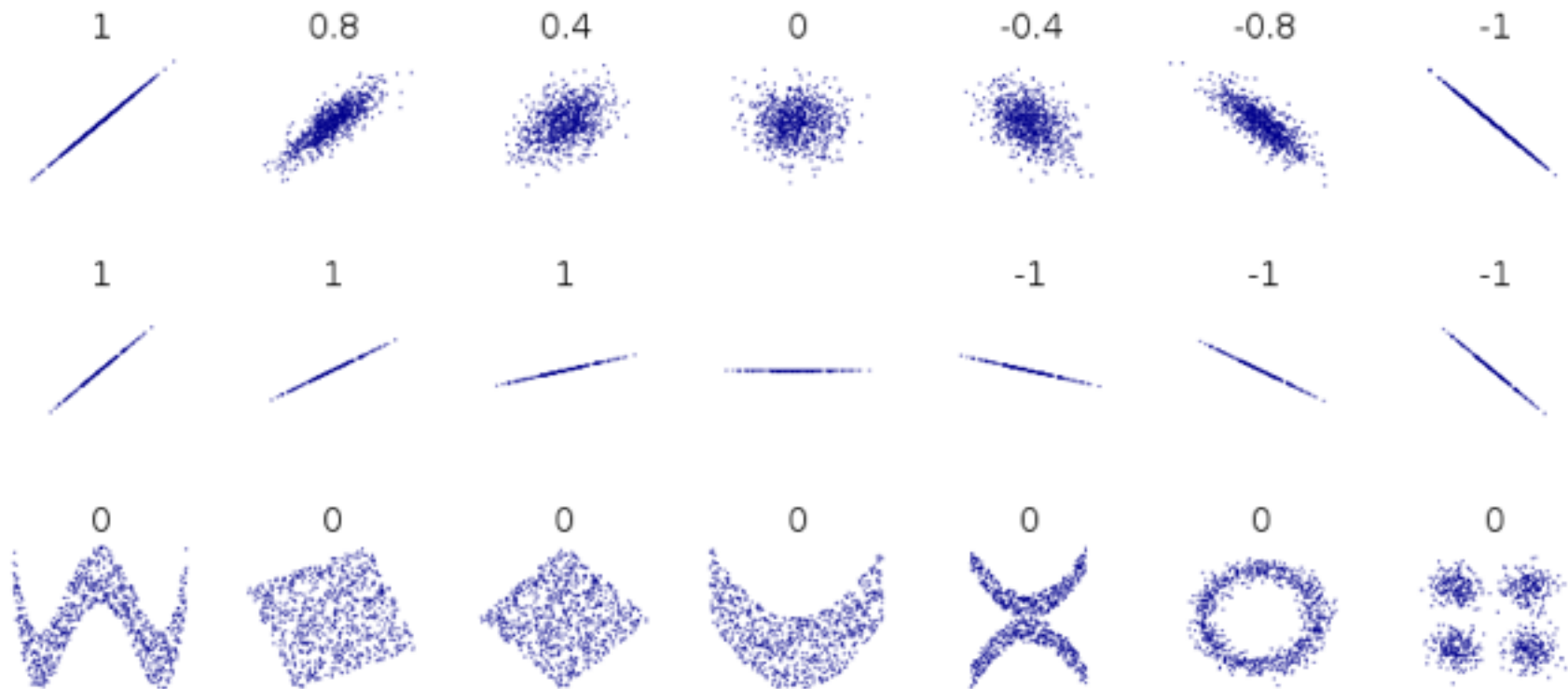
$$r = \frac{1}{n} \sum_{i=1}^n Z_x(x_i) Z_y(y_i)$$

Pearson's Correlation Coefficient

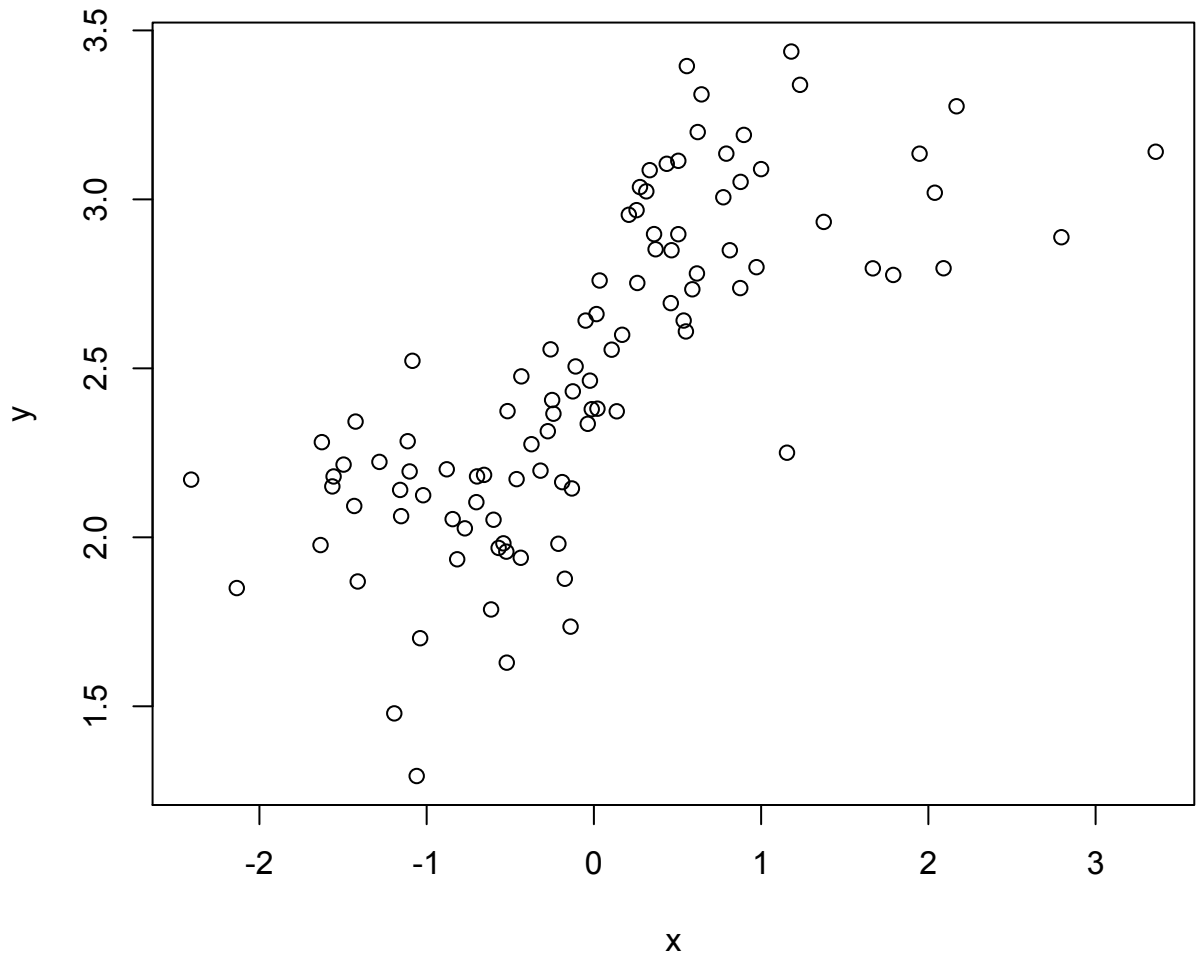
$$-1 \leq r \leq 1$$

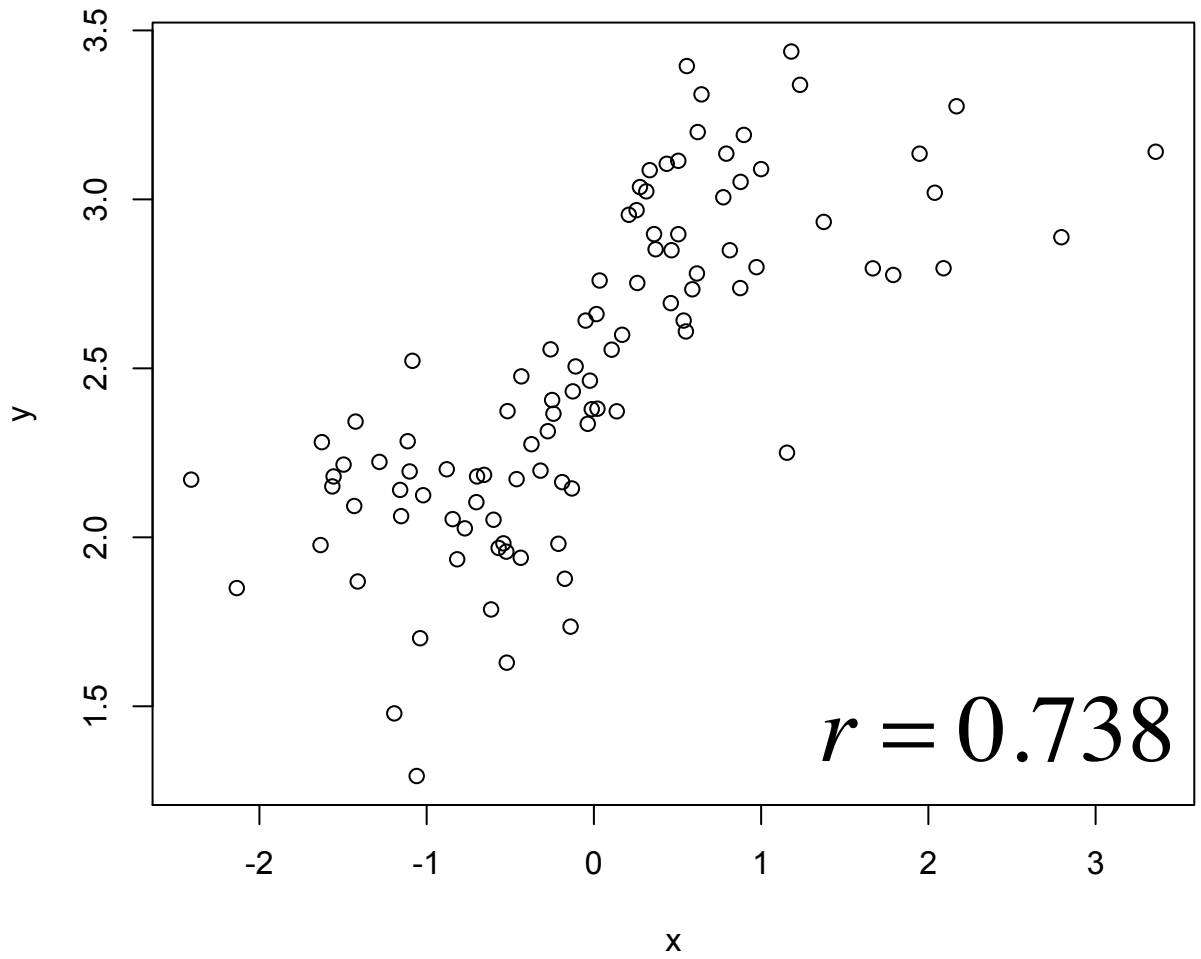


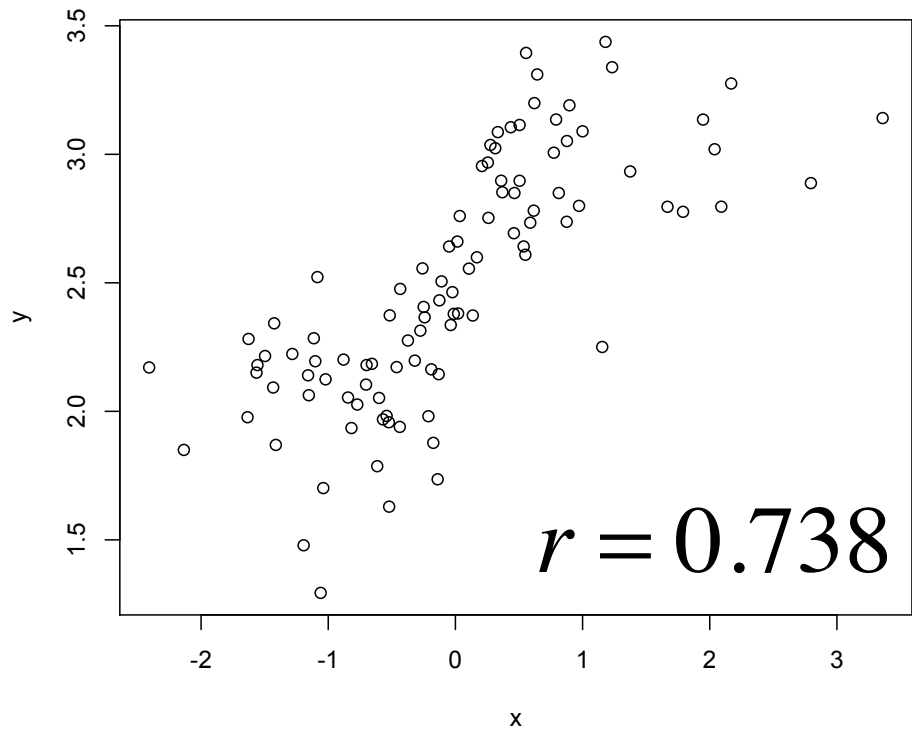
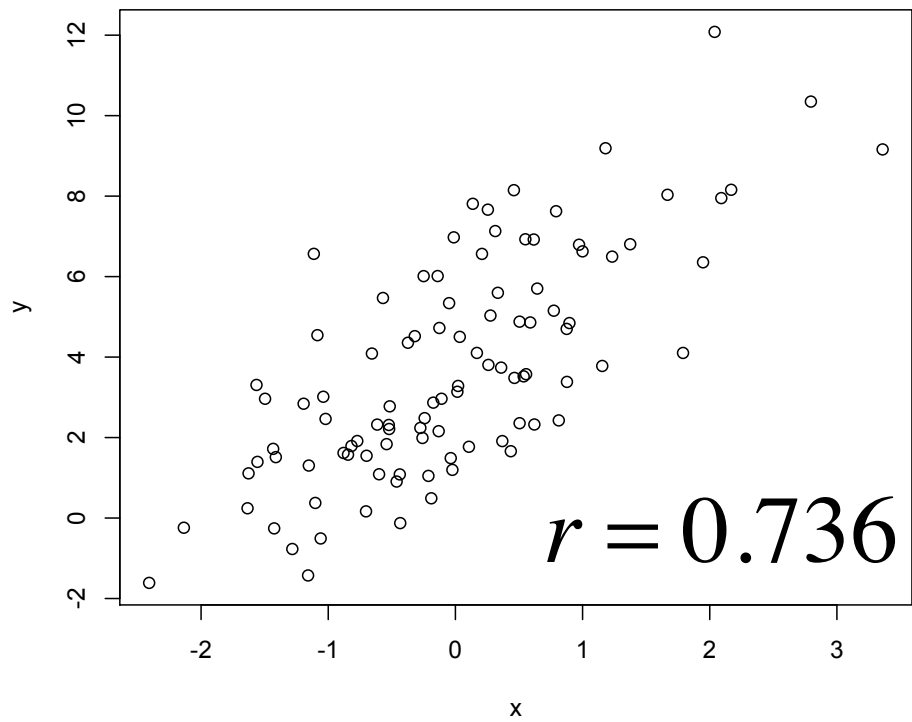




Pearson's correlation coefficient captures linear correlations.







Spearman's rank correlation coefficient captures non-linear correlations.

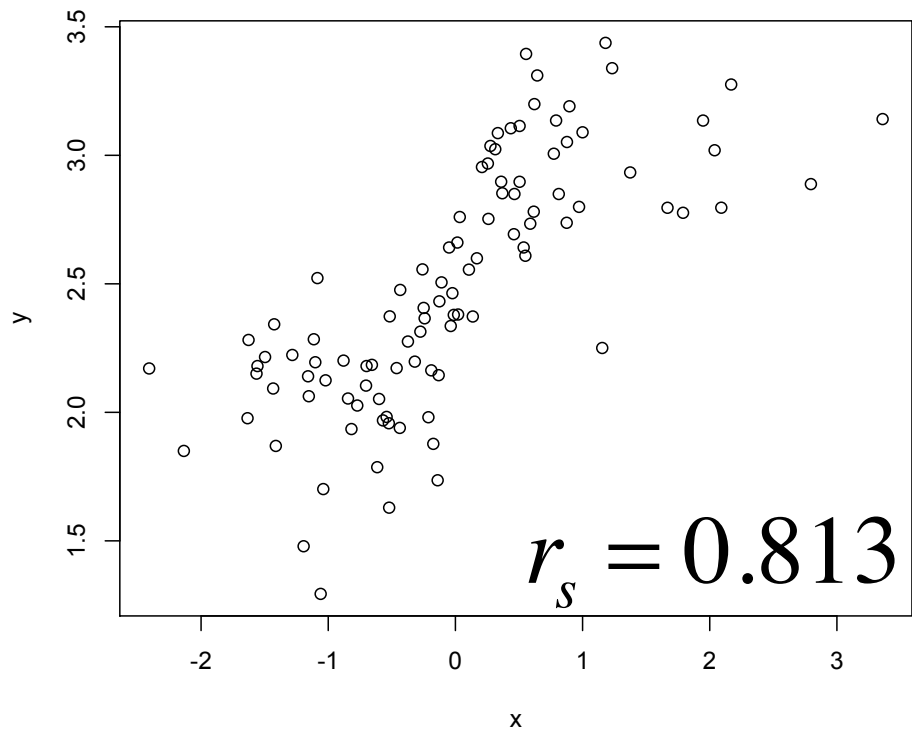
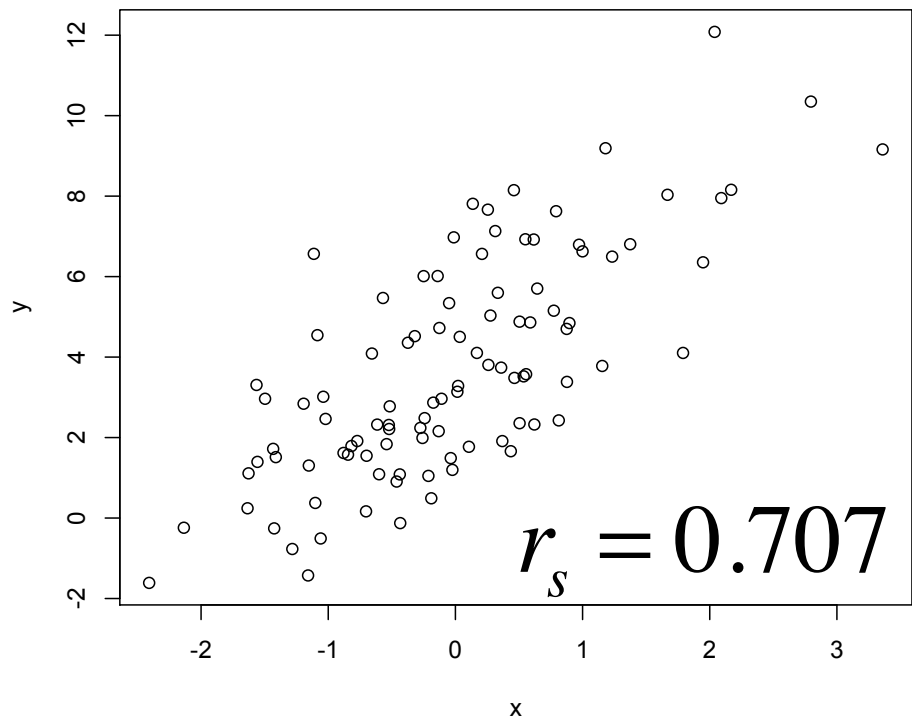
Spearman's Rank Correlation Coefficient

$$a = \text{rank}(x)$$

$$b = \text{rank}(y)$$

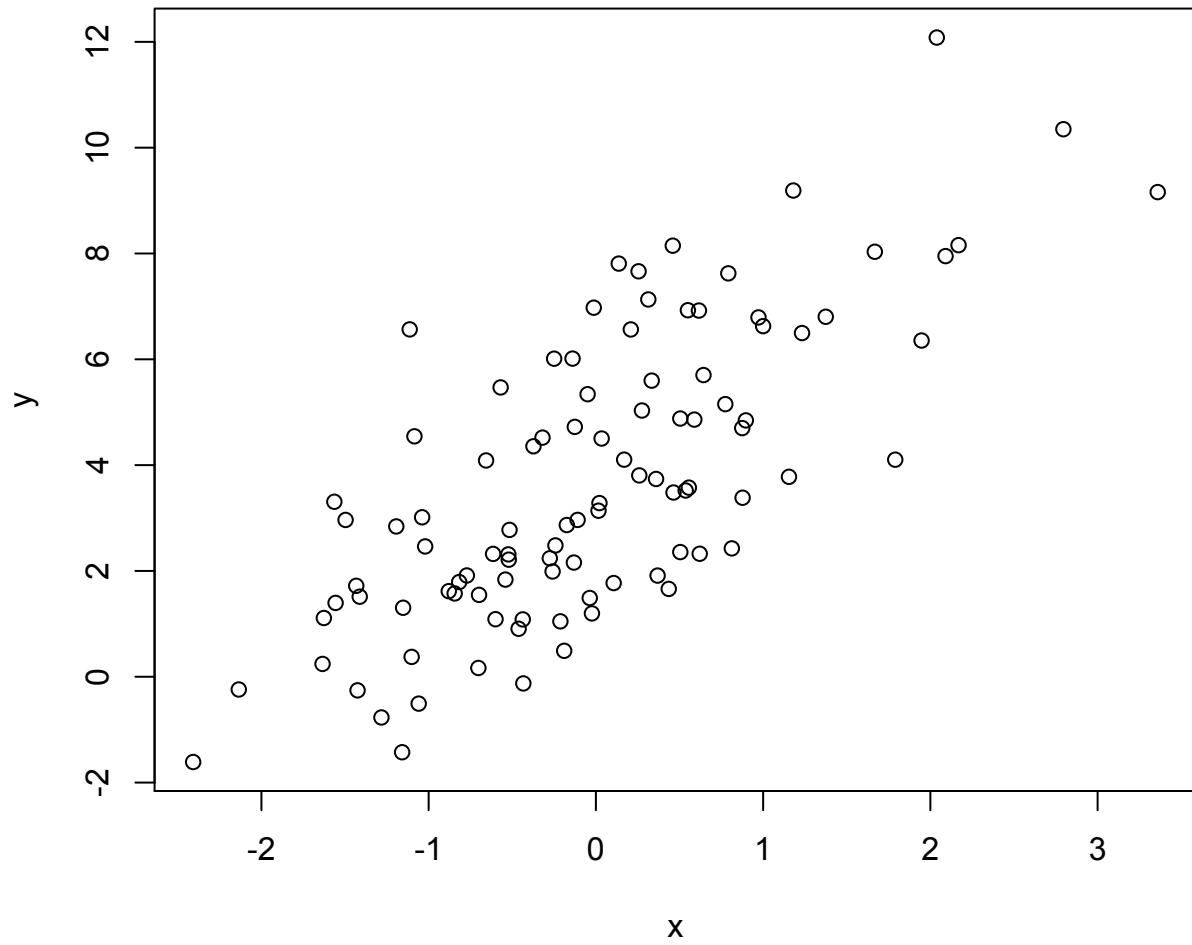
Spearman's Rank Correlation Coefficient

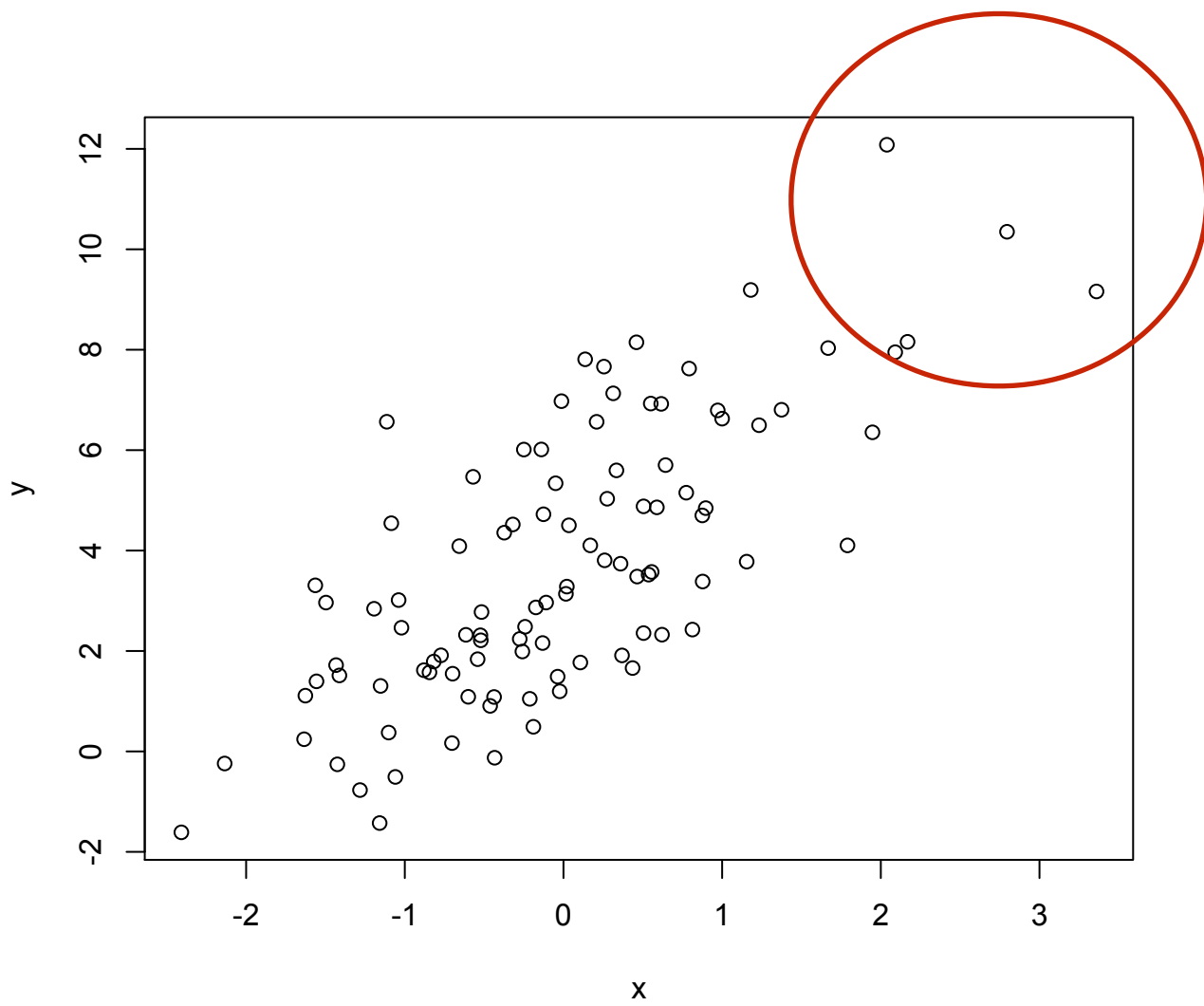
$$r_{\text{spearman}} = \frac{1}{n} \sum_{i=1}^n Z_a(a_i) Z_b(b_i)$$



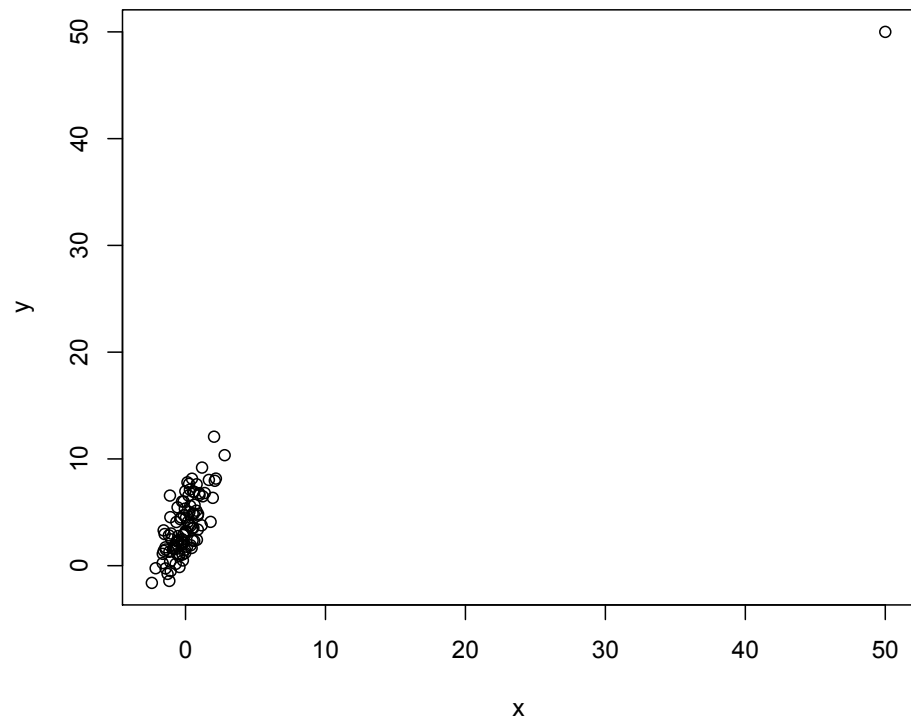
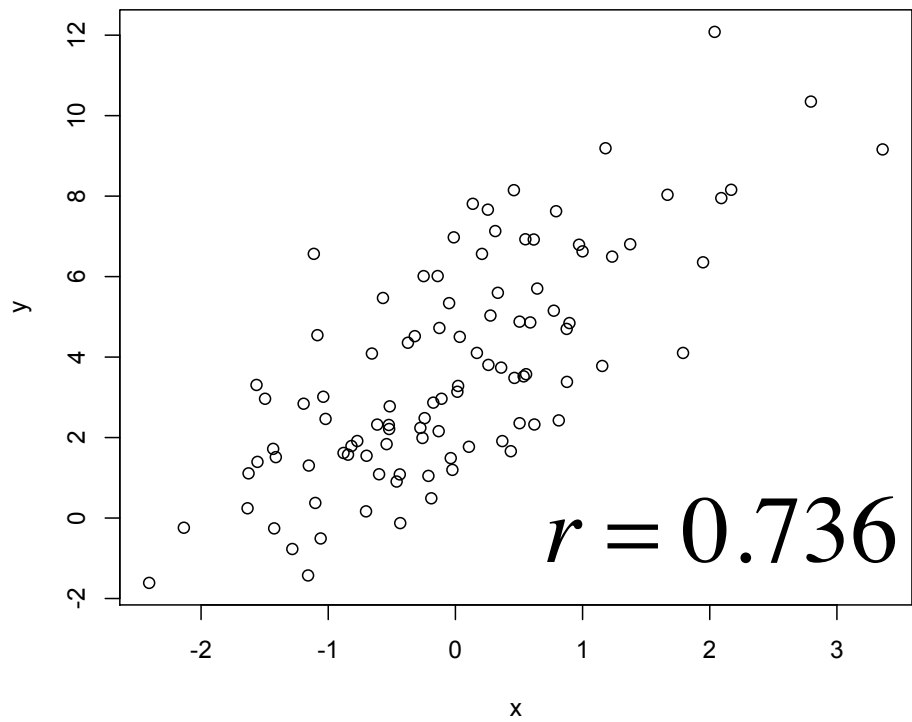
Question:

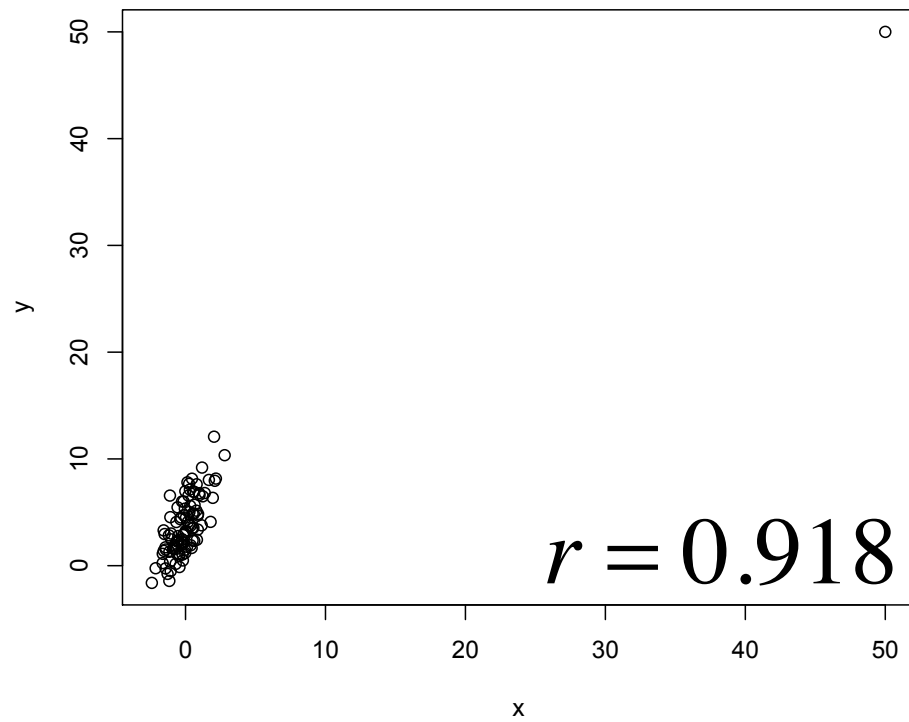
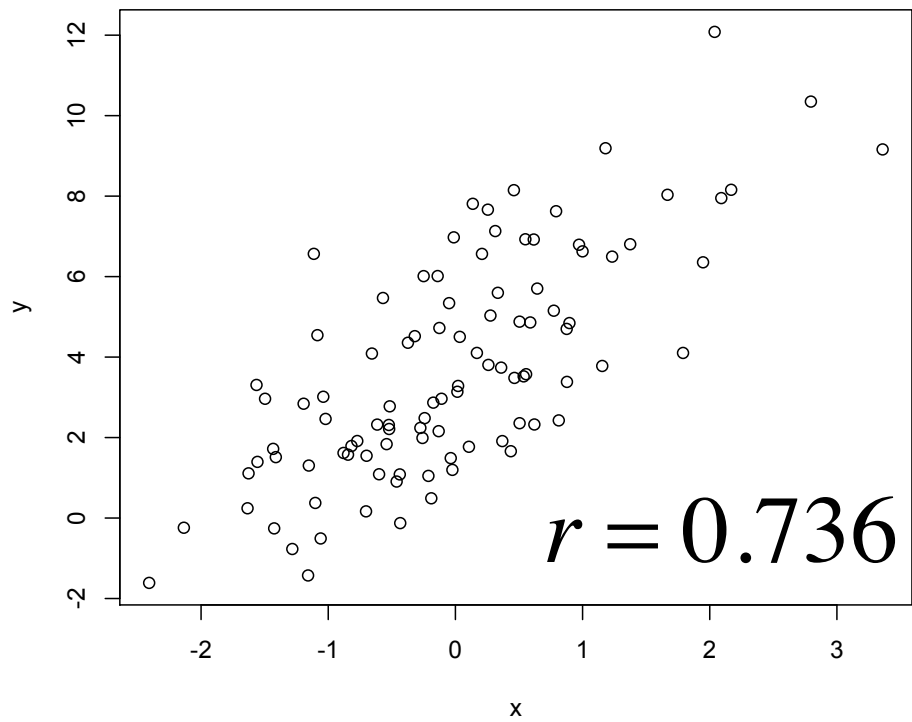
Why did the correlation decrease in our first example?



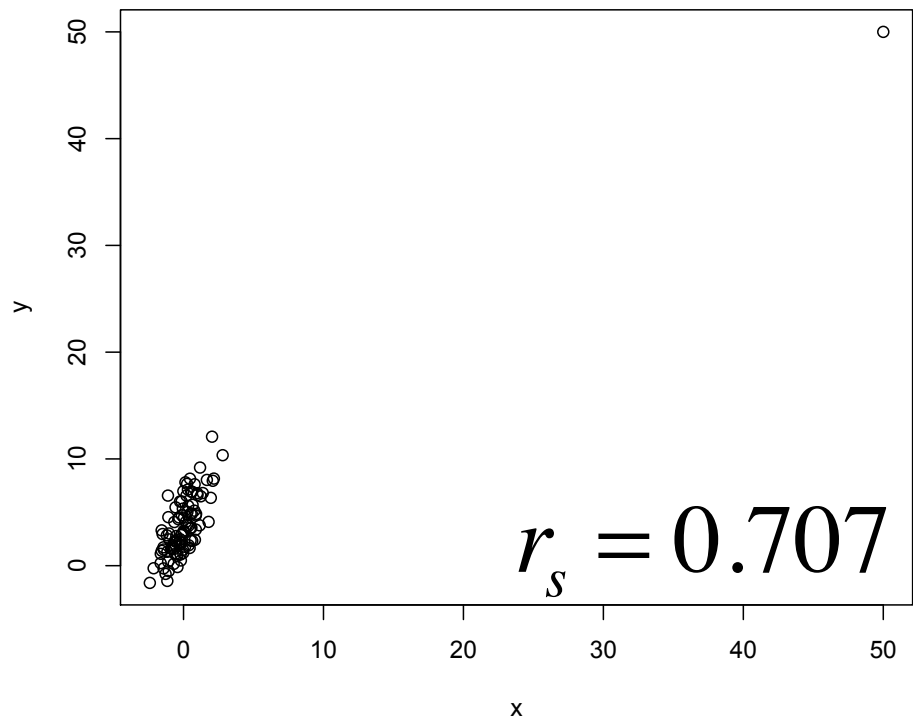
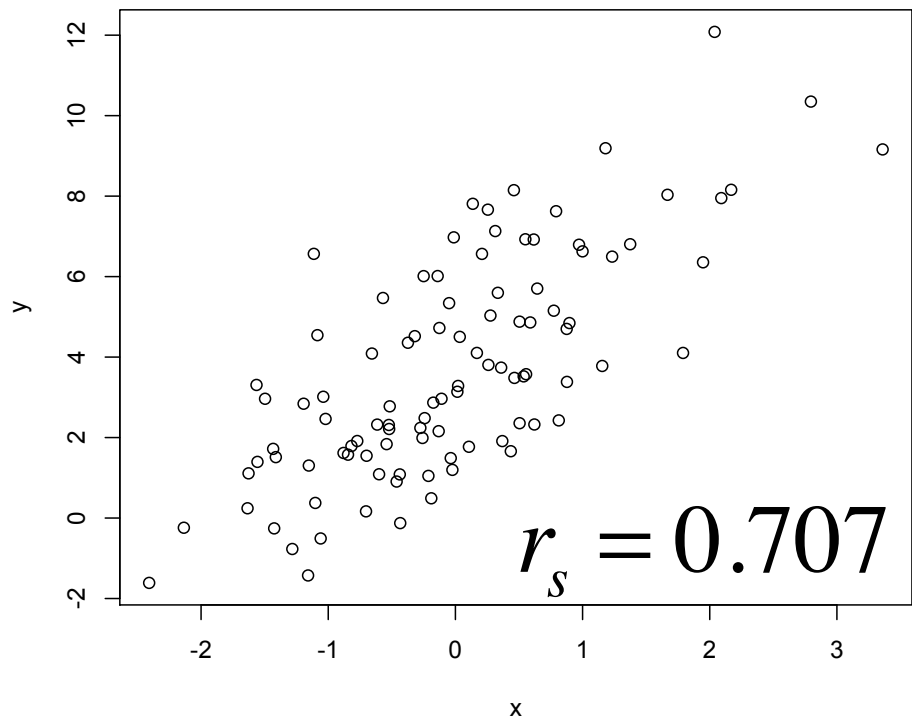


Outliers increase the Pearson's correlation coefficient.

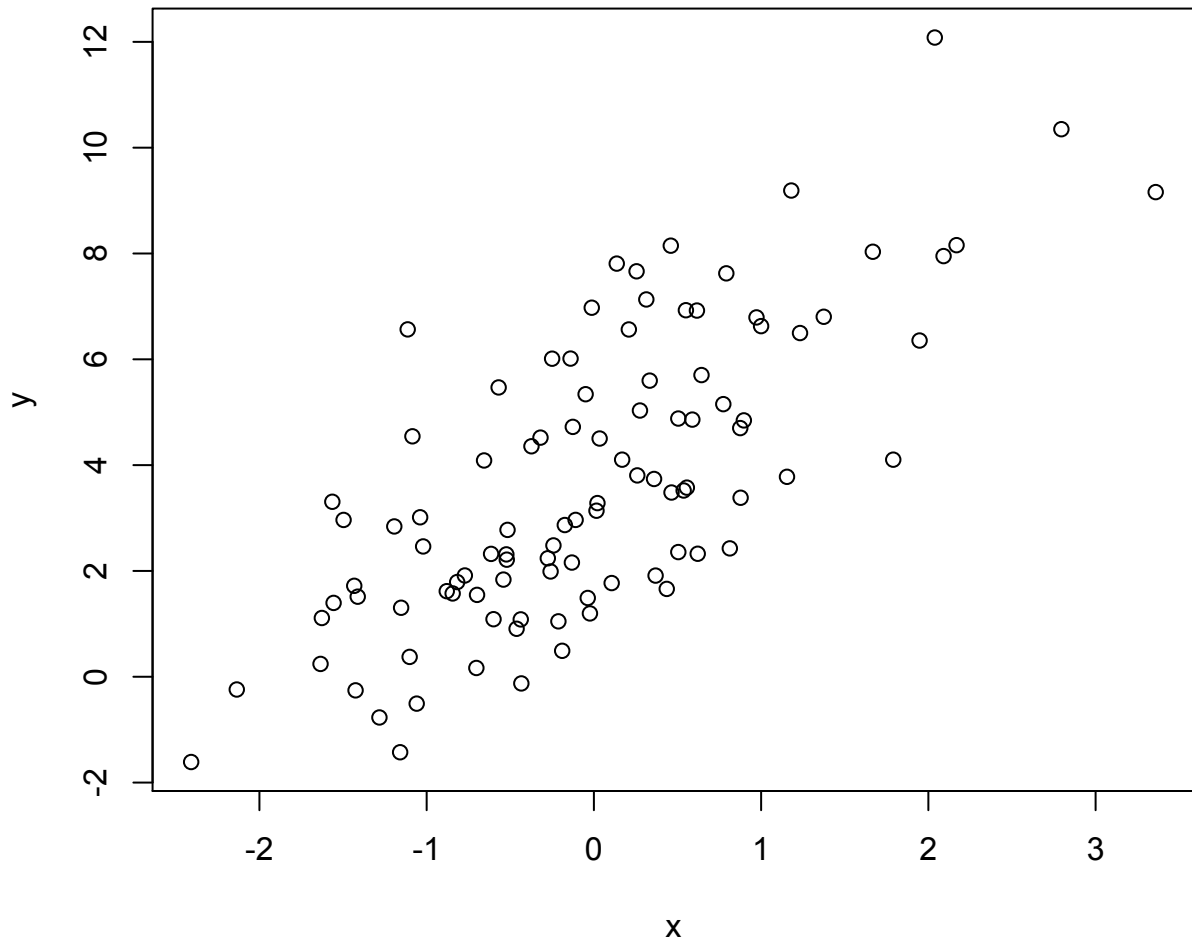


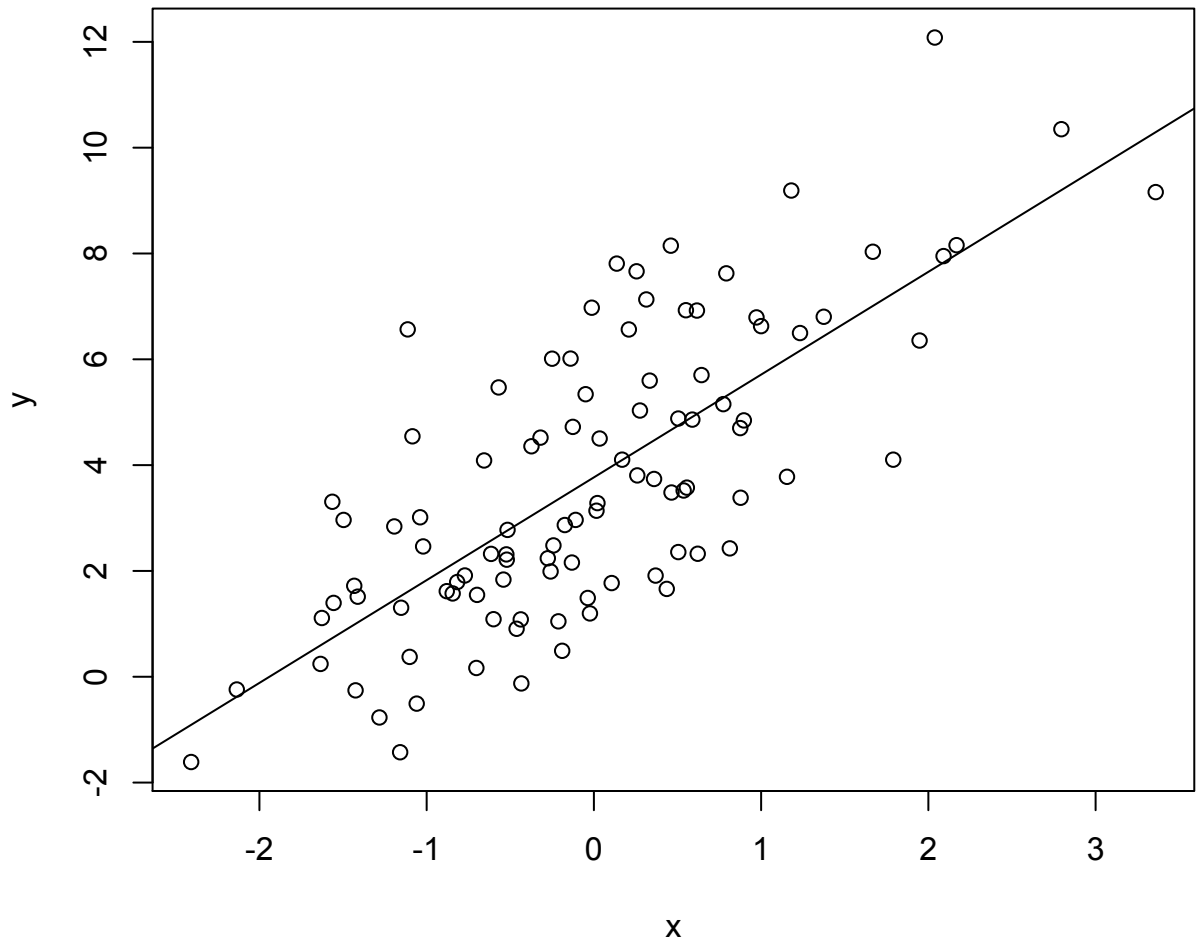


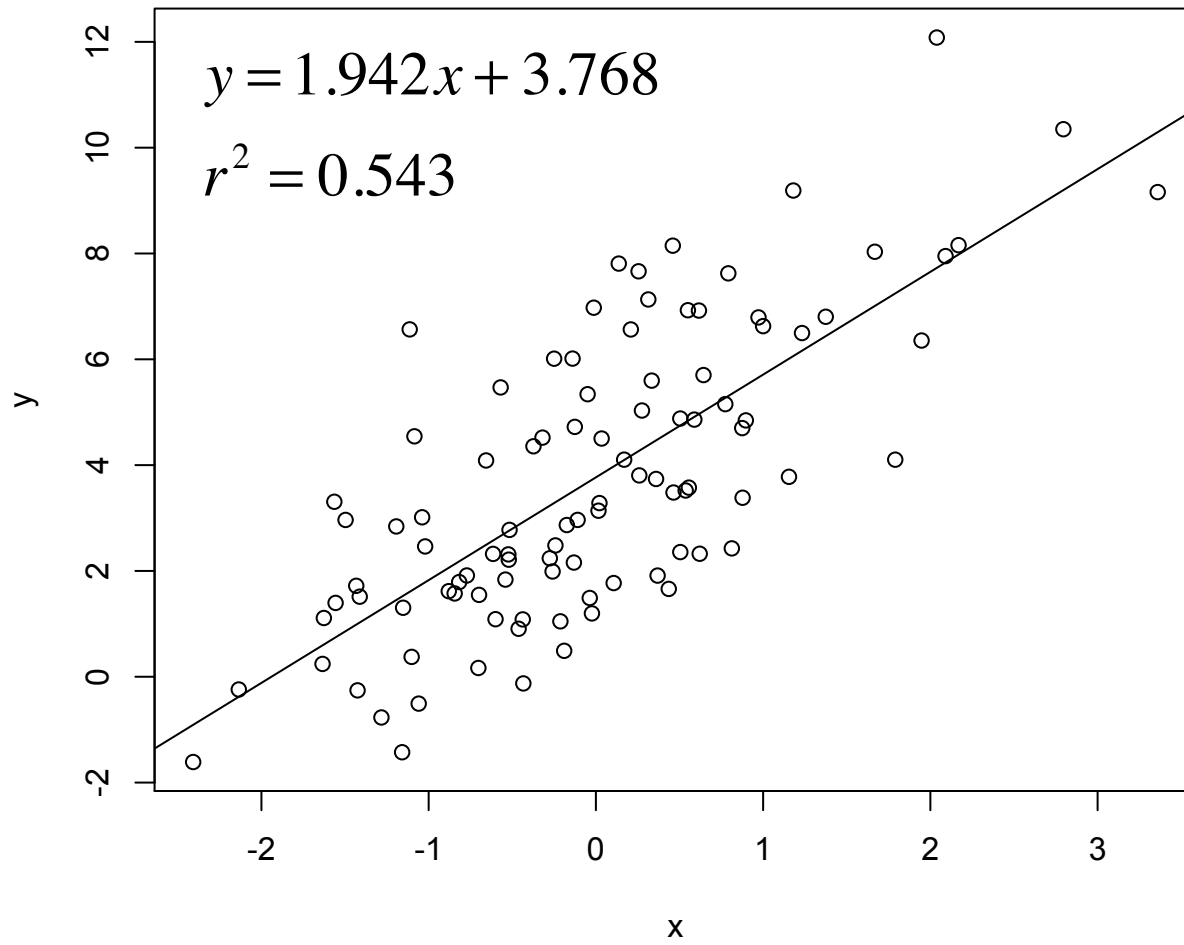
Outliers do not increase the Spearman's rank correlation coefficient.



[R Example]

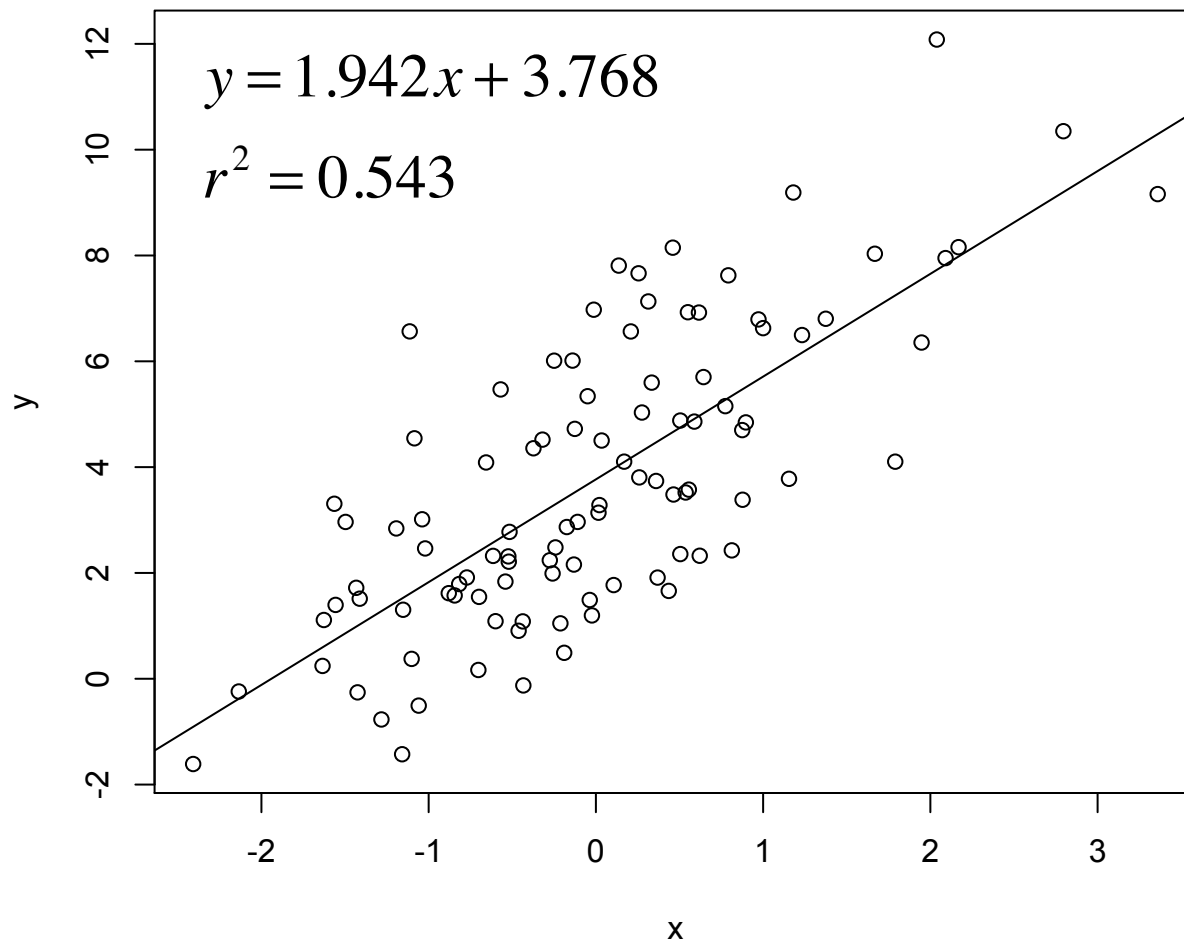




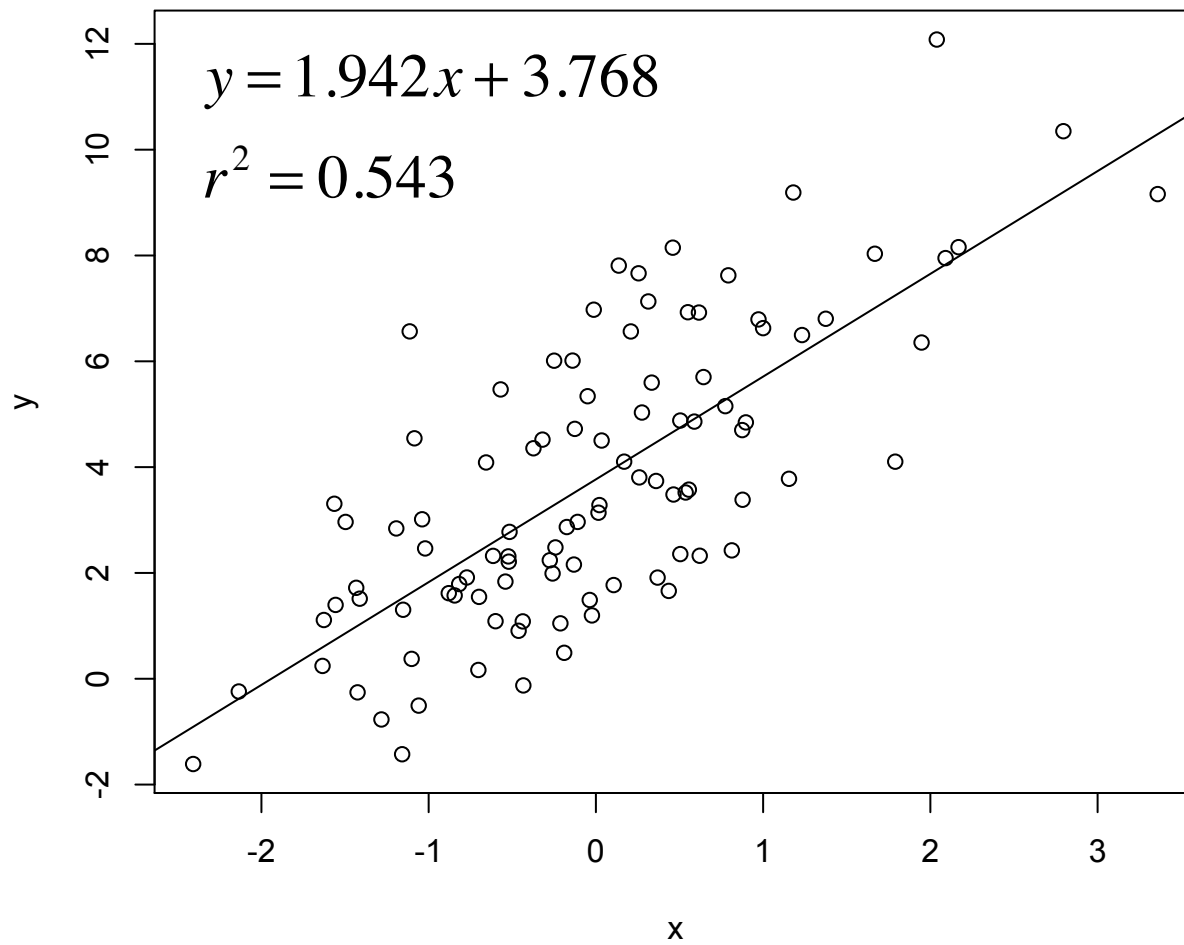


If there is a line of best fit, an equation, and the r is squared, you're looking at linear regression.

In linear regression, you fit a line that describes how one variable, y , is a function of another, x .



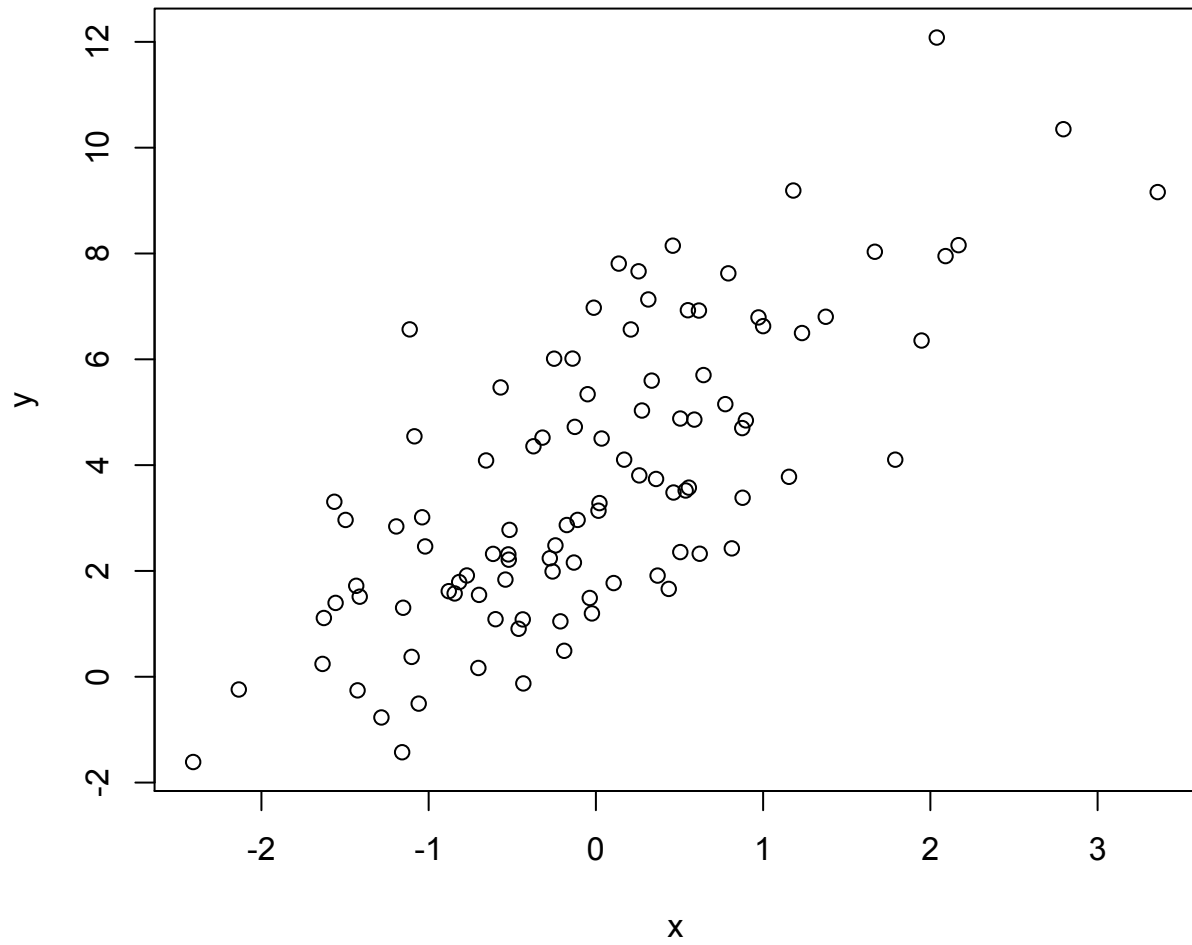
The “r-squared” quantifies the amount of variation in y can be attributed to x.

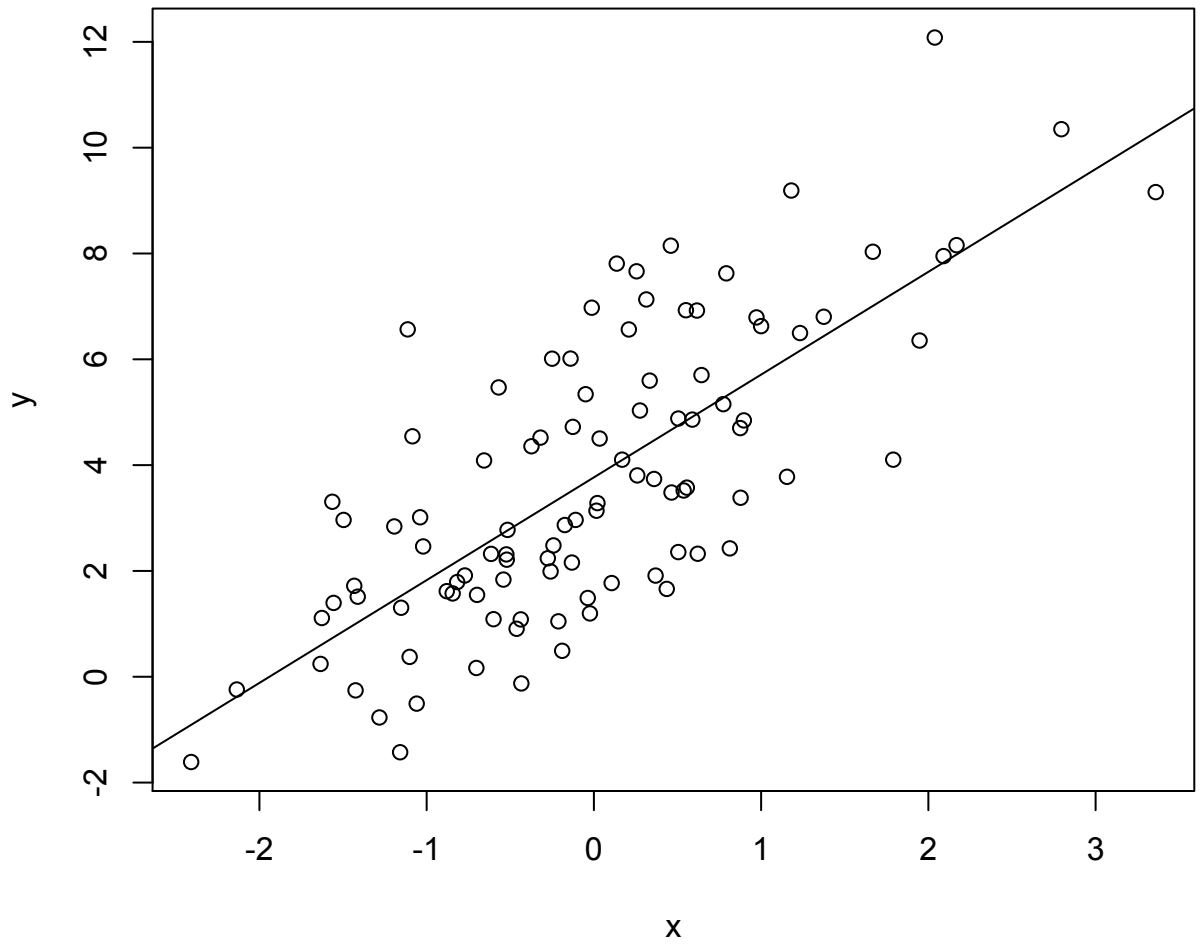


Question:

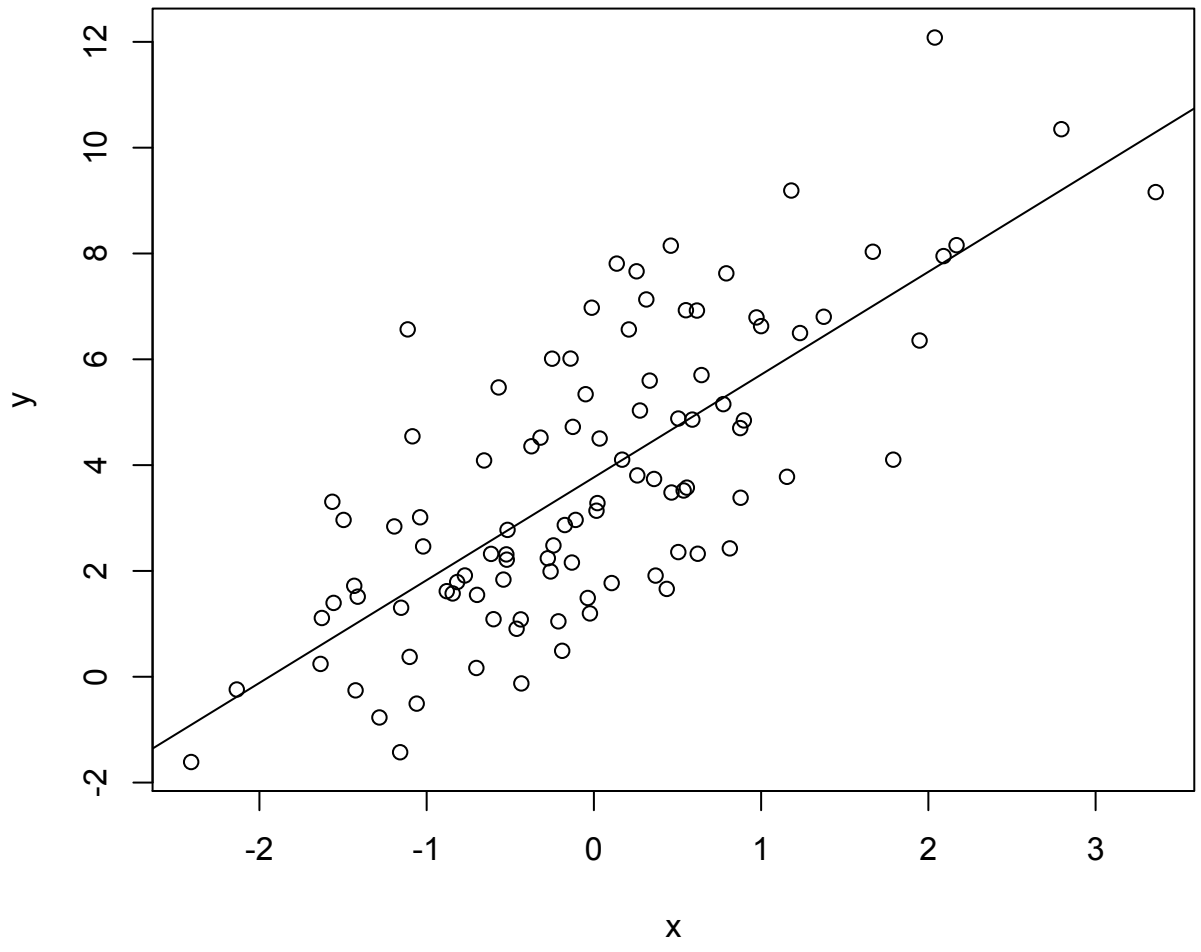
Why perform linear regression?

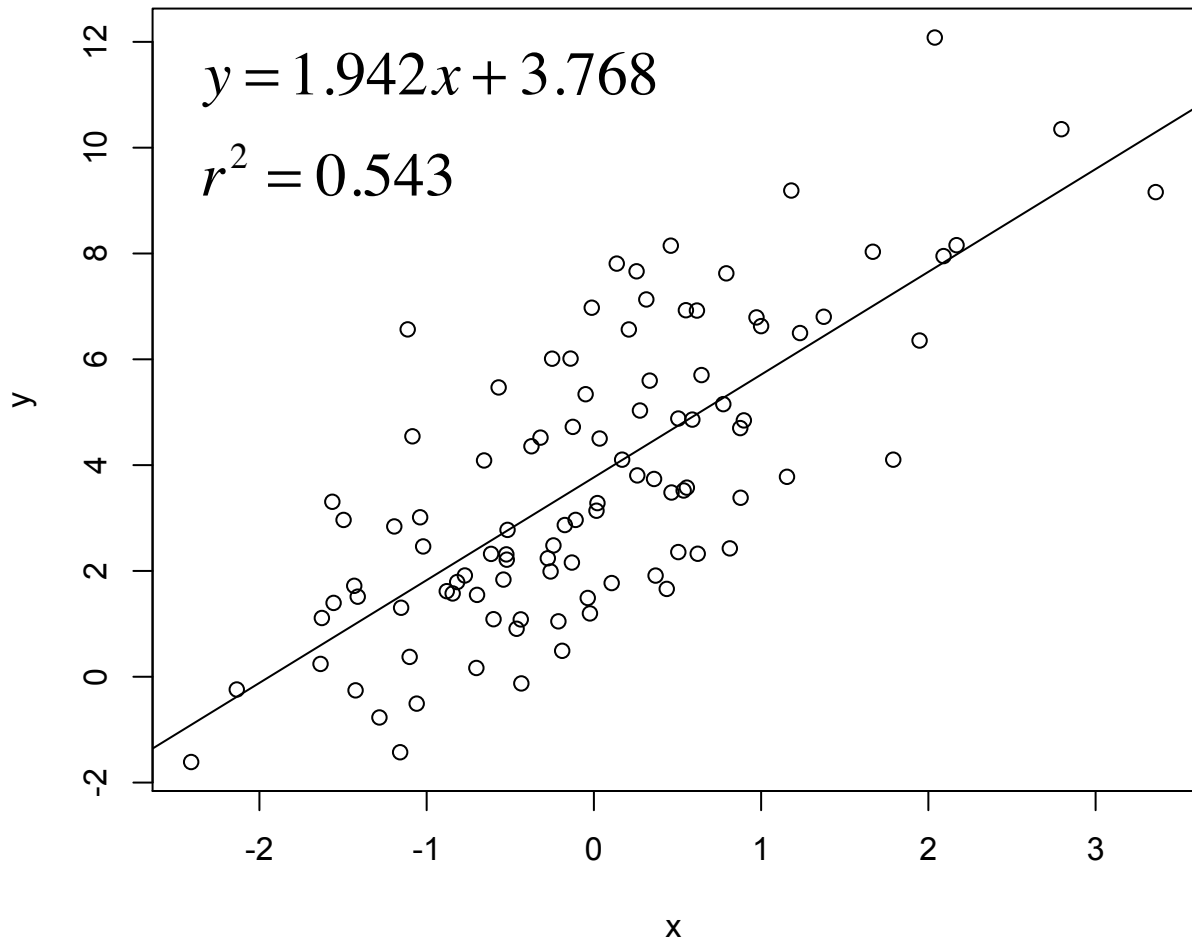
1. Artistic. You want to pull people's attention to the signal rather than the noise.



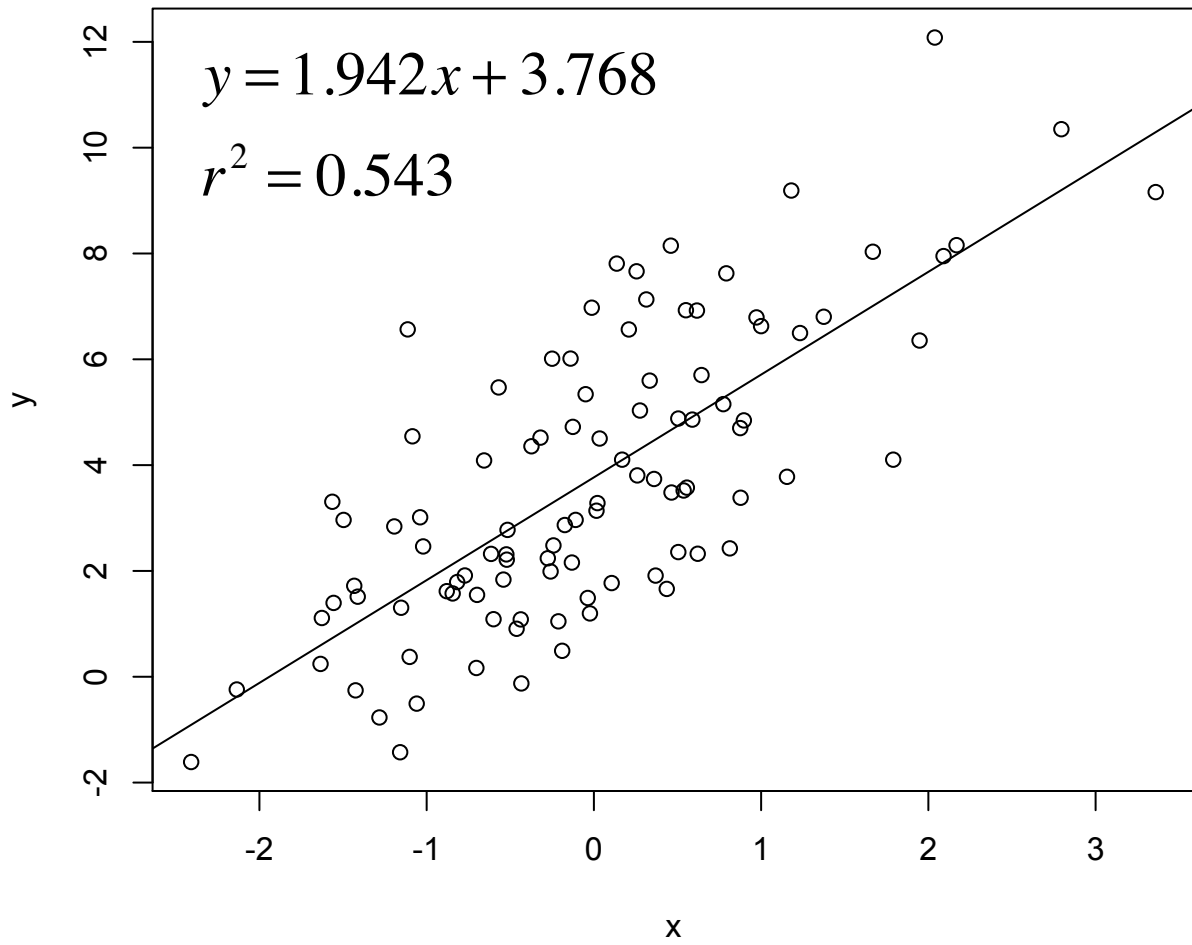


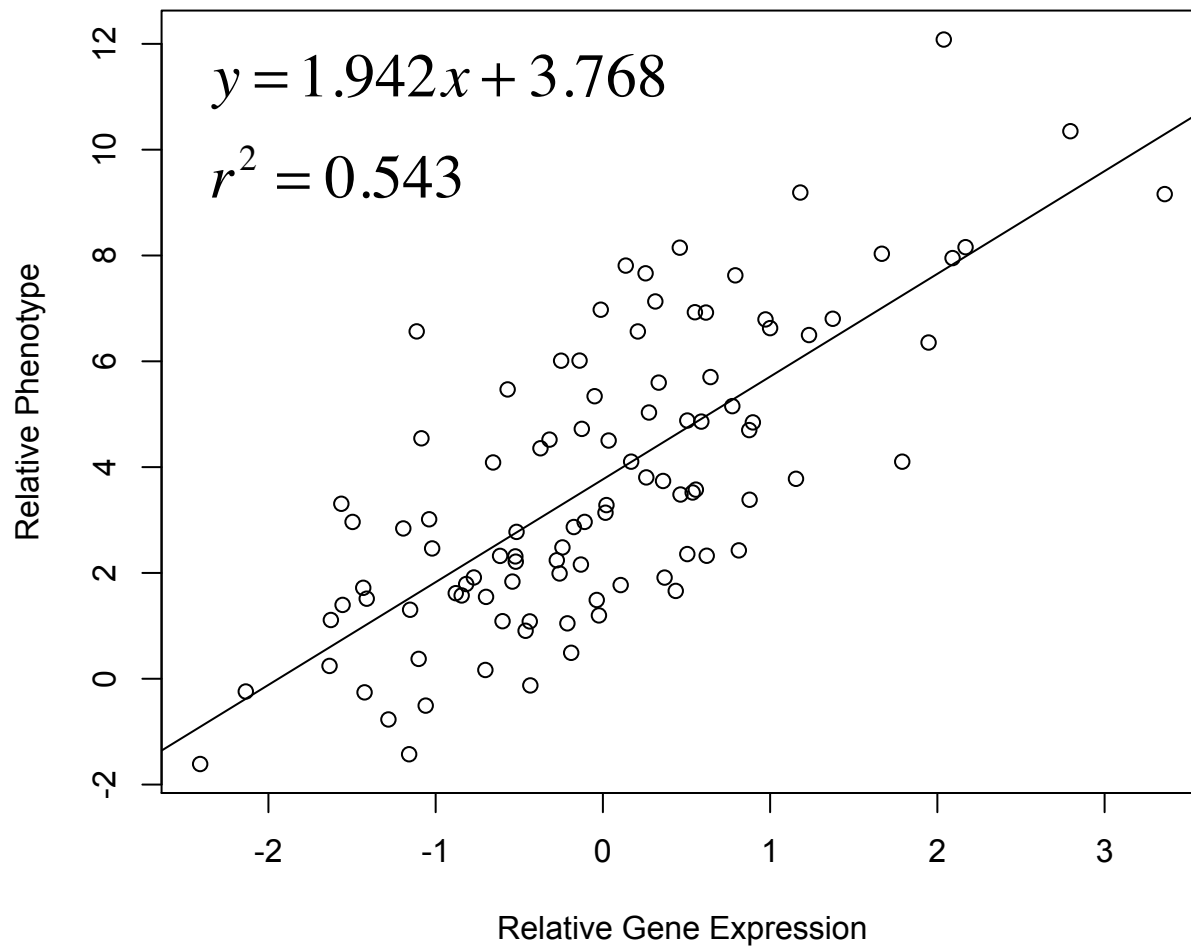
2. Predictive. You want to extrapolate your trend to future measurements.





3. Modeling. You want to generate a mechanistic model of the biology you're studying.





Question:

That model seems really simple. Can I make it more complex?

CORE PRINCIPLES IN RESEARCH



OCCAM'S RAZOR

"WHEN FACED WITH TWO POSSIBLE EXPLANATIONS, THE SIMPLER OF THE TWO IS THE ONE MOST LIKELY TO BE TRUE."



OCCAM'S PROFESSOR

"WHEN FACED WITH TWO POSSIBLE WAYS OF DOING SOMETHING, THE MORE COMPLICATED ONE IS THE ONE YOUR PROFESSOR WILL MOST LIKELY ASK YOU TO DO."

WWW.PHDCOMICS.COM

- **Fewer explanatory variables**

$$y = x$$

$$y = x_1 + x_2$$

- **Fewer explanatory variables**
- **Fewer parameters (model coefficients, etc.)**

$$y = x$$

$$y = mx$$

$$y = mx + b$$

- **Fewer explanatory variables**
- **Fewer parameters (model coefficients, etc.)**
- **Linear over non-linear relationships**

$$y = mx + b$$

$$y = \log_m (x)^2 + b$$

- **Fewer explanatory variables**
- **Fewer parameters (model coefficients, etc.)**
 - **Linear over non-linear relationships**
- **Monotonic vs. non-monotonic relationships**

$$y = mx + b$$

$$y = \sin(x) + b$$

- **Fewer explanatory variables**
 - **Fewer parameters (model coefficients, etc.)**
 - **Linear over non-linear relationships**
- **Monotonic vs. non-monotonic relationships**
 - **Fewer interactions among explanatory variables**

$$y = x_1 + x_2$$

$$y = x_1 + x_2 + x_1x_2$$

CORE PRINCIPLES IN RESEARCH



OCCAM'S RAZOR

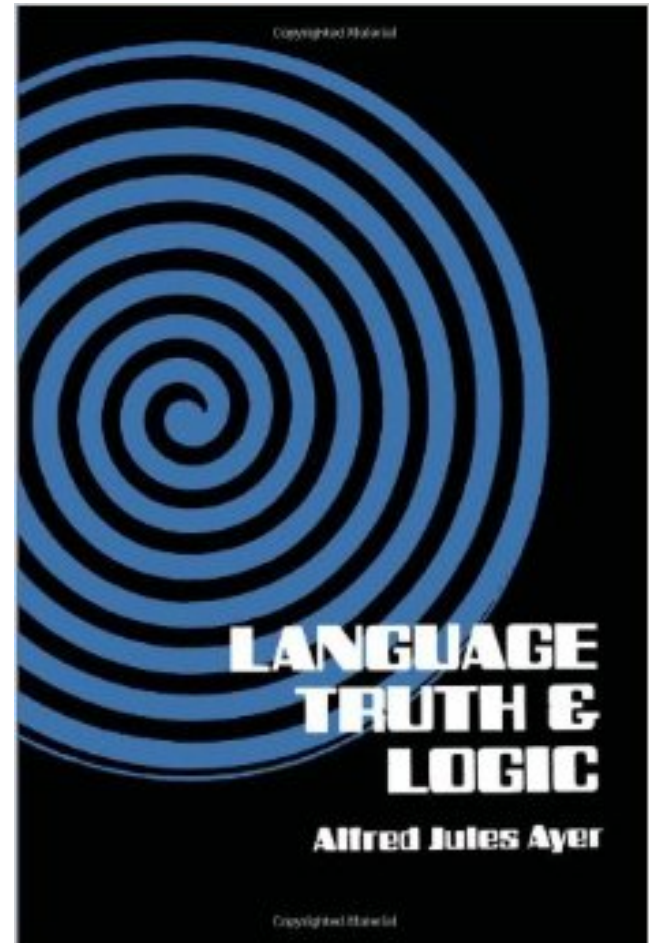
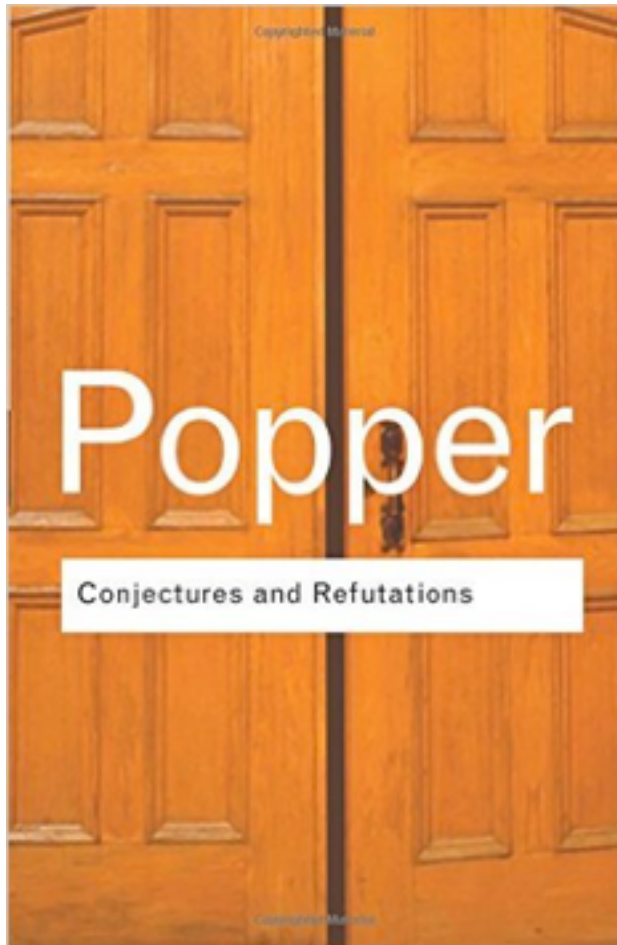
"WHEN FACED WITH TWO POSSIBLE EXPLANATIONS, THE SIMPLER OF THE TWO IS THE ONE MOST LIKELY TO BE TRUE."

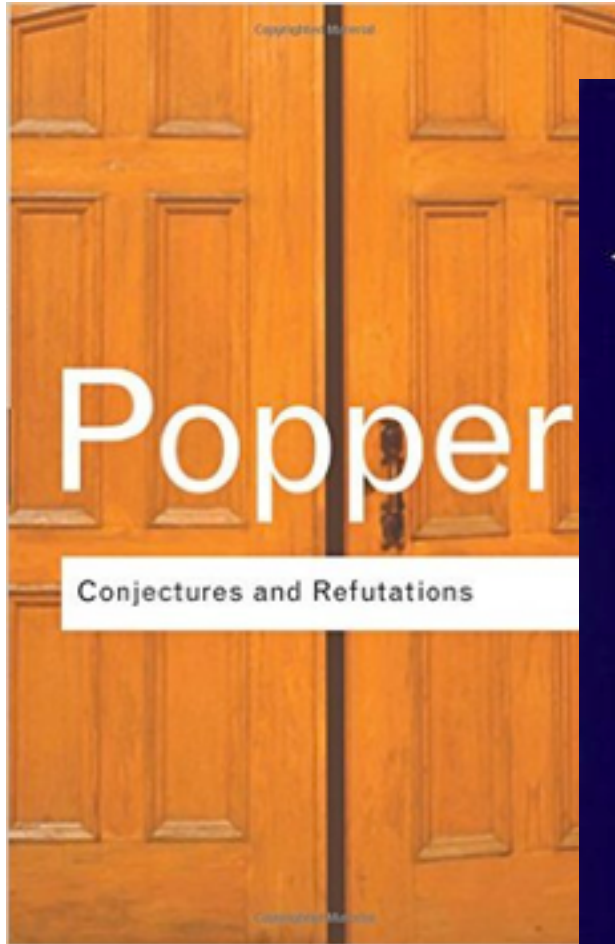


OCCAM'S PROFESSOR

"WHEN FACED WITH TWO POSSIBLE WAYS OF DOING SOMETHING, THE MORE COMPLICATED ONE IS THE ONE YOUR PROFESSOR WILL MOST LIKELY ASK YOU TO DO."

WWW.PHDCOMICS.COM



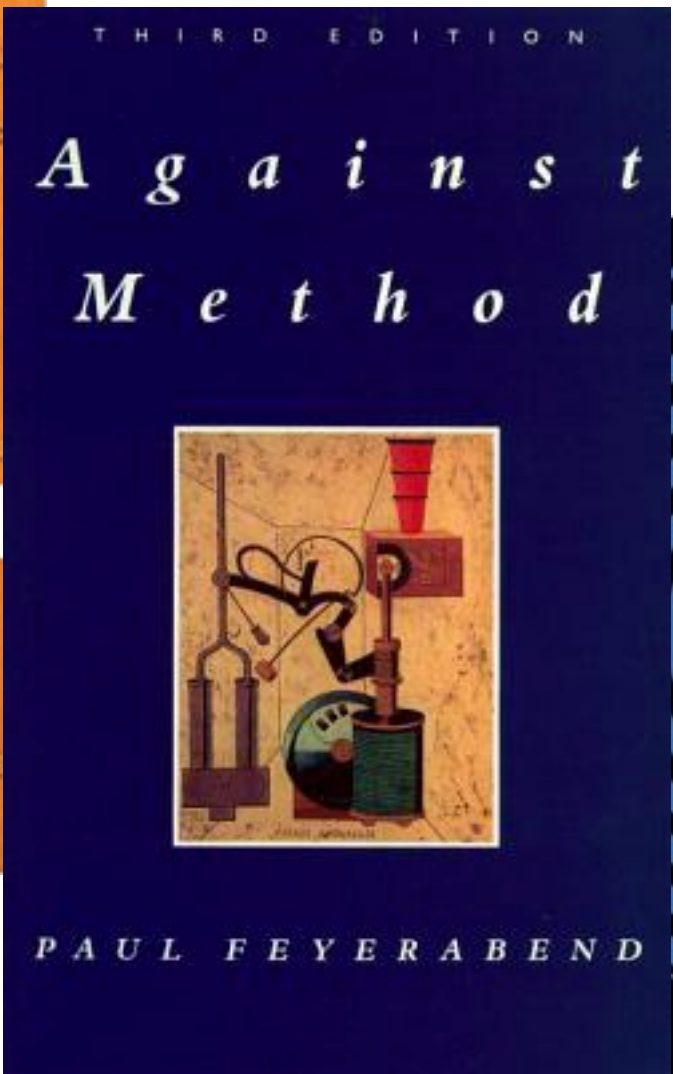


Copyrighted Material

Popper

Conjectures and Refutations

Copyrighted Material

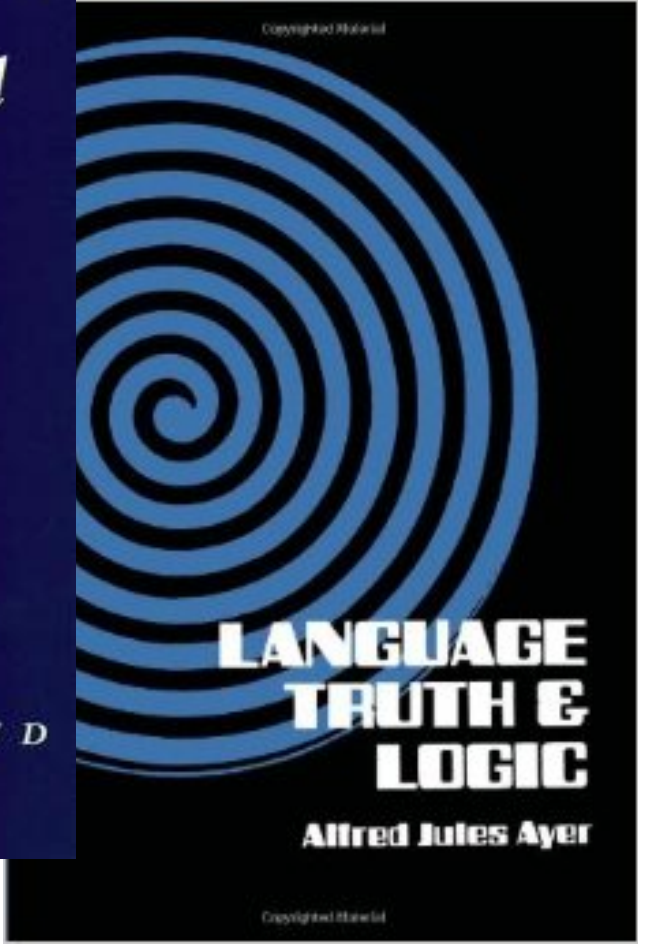


THIRD EDITION

Against Method



PAUL FEYERABEND



Copyrighted Material

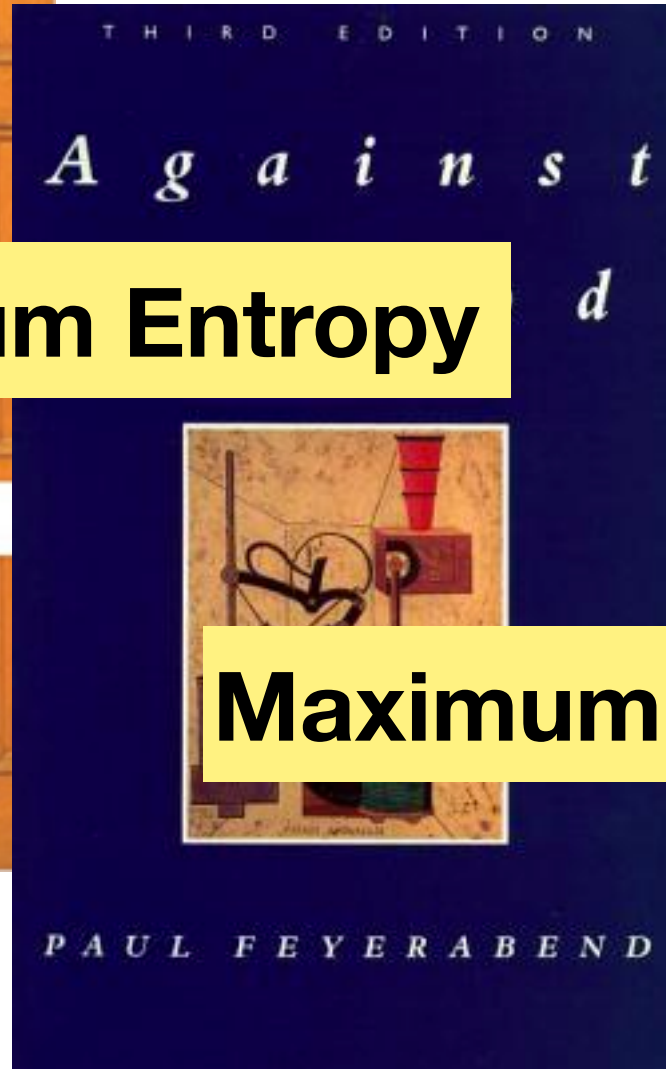
LANGUAGE TRUTH & LOGIC

Alfred Jules Ayer

Copyrighted Material



Maximum Entropy



Maximum Likelihood





THIRD EDITION

A g a i n s t

M e t h o d

nature

PHYSICS

Is String Theory Science?

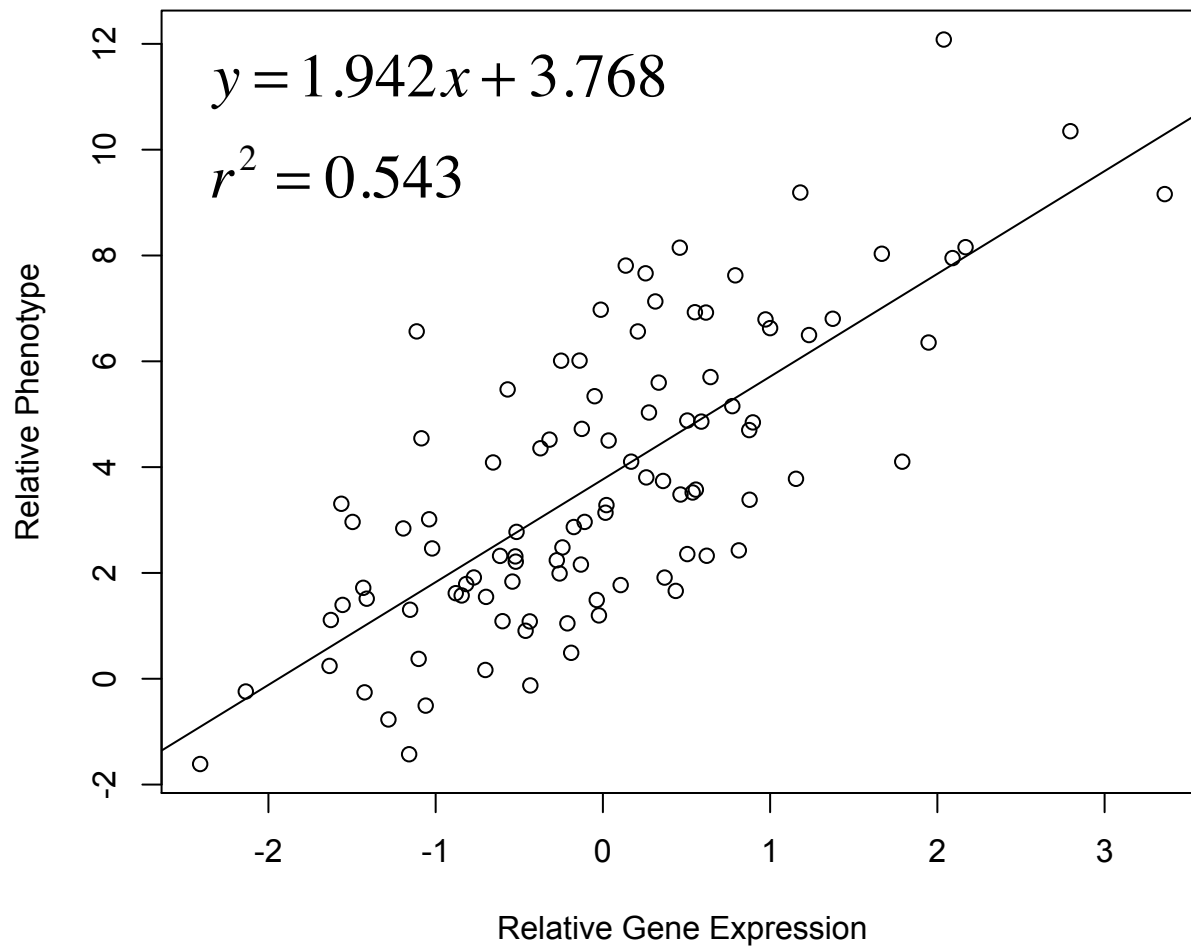
A debate between physicists and philosophers could redefine the scientific method
and our understanding of the universe

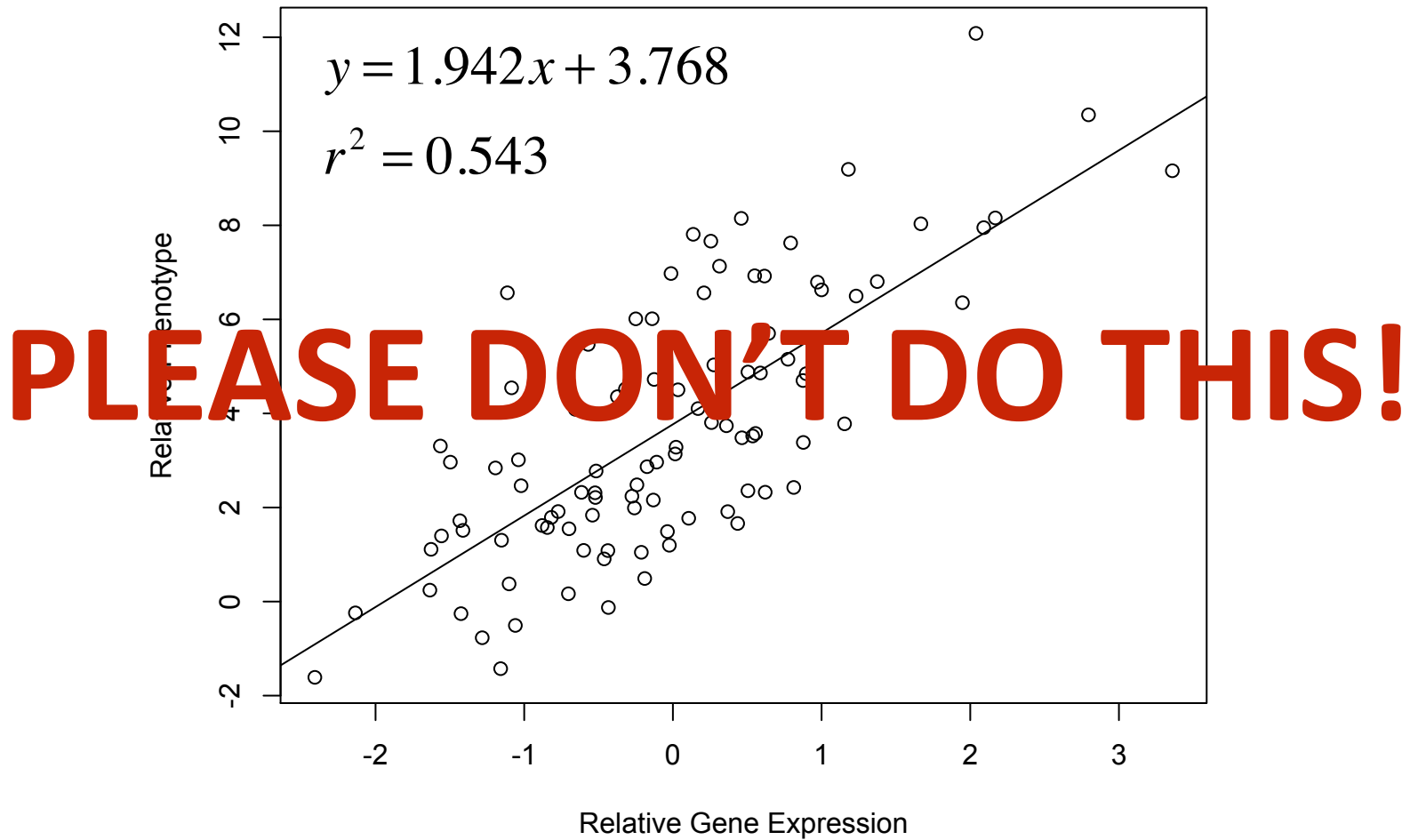


PAUL FEYERABEND

**LANGUAGE
TRUTH &
LOGIC**

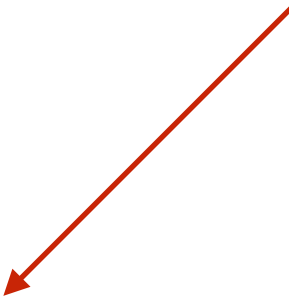
Alfred Jules Ayer





$$y = 1.942x + 3.768$$

dependent variable


$$y = 1.942x + 3.768$$

dependent variable

$$y = 1.942x + 3.768$$

independent variable

The independent variable should not have any measurement error associated with it. If it was sampled in a manner that is out of your control, it isn't really an independent variable.

dependent variable

$$y = 1.942x + 3.768$$

independent variable

phenotype

$$y = 1.942x + 3.768$$

drug concentration

[R Example]