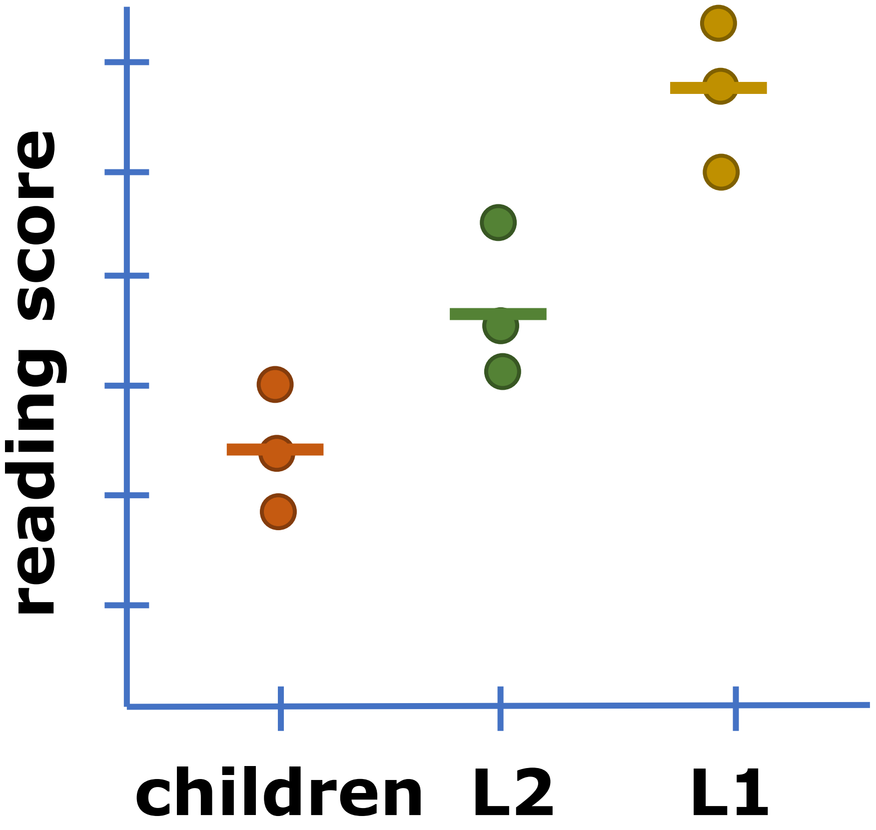
Department of Linguistics and Translation
香港城市大學
City University of Hong Kong

Fundamentals of Statistics for Language Sciences LT2206

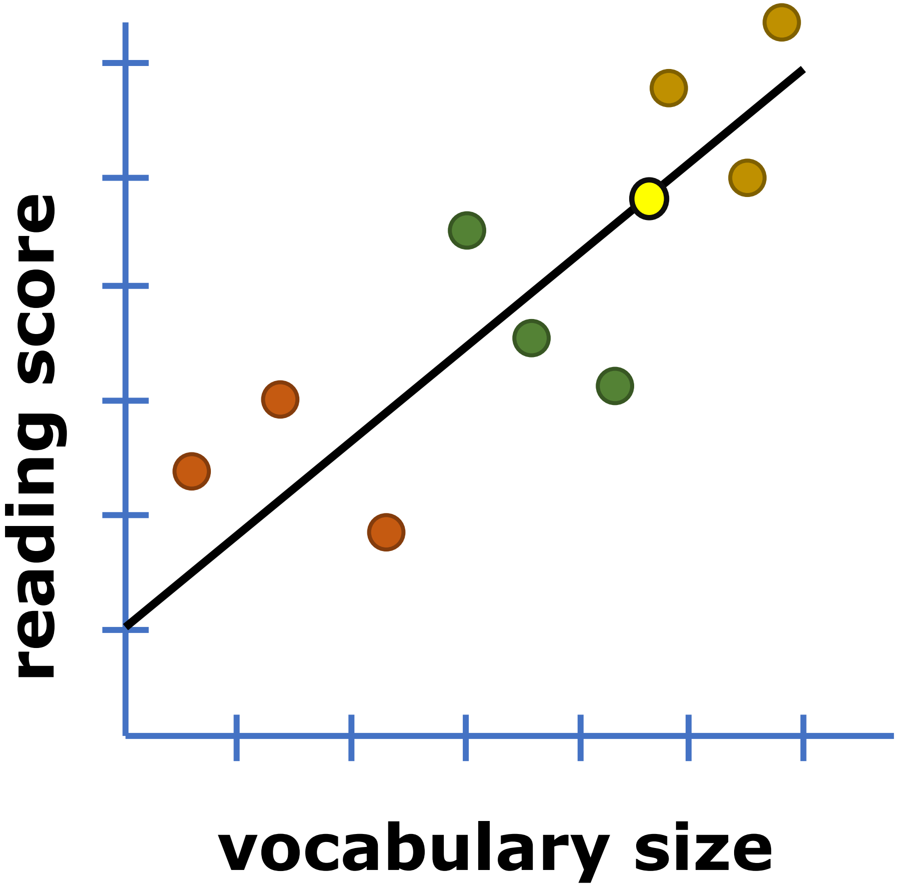
Jixing Li

Lecture 9: Correlation

ANOVA vs. Regression

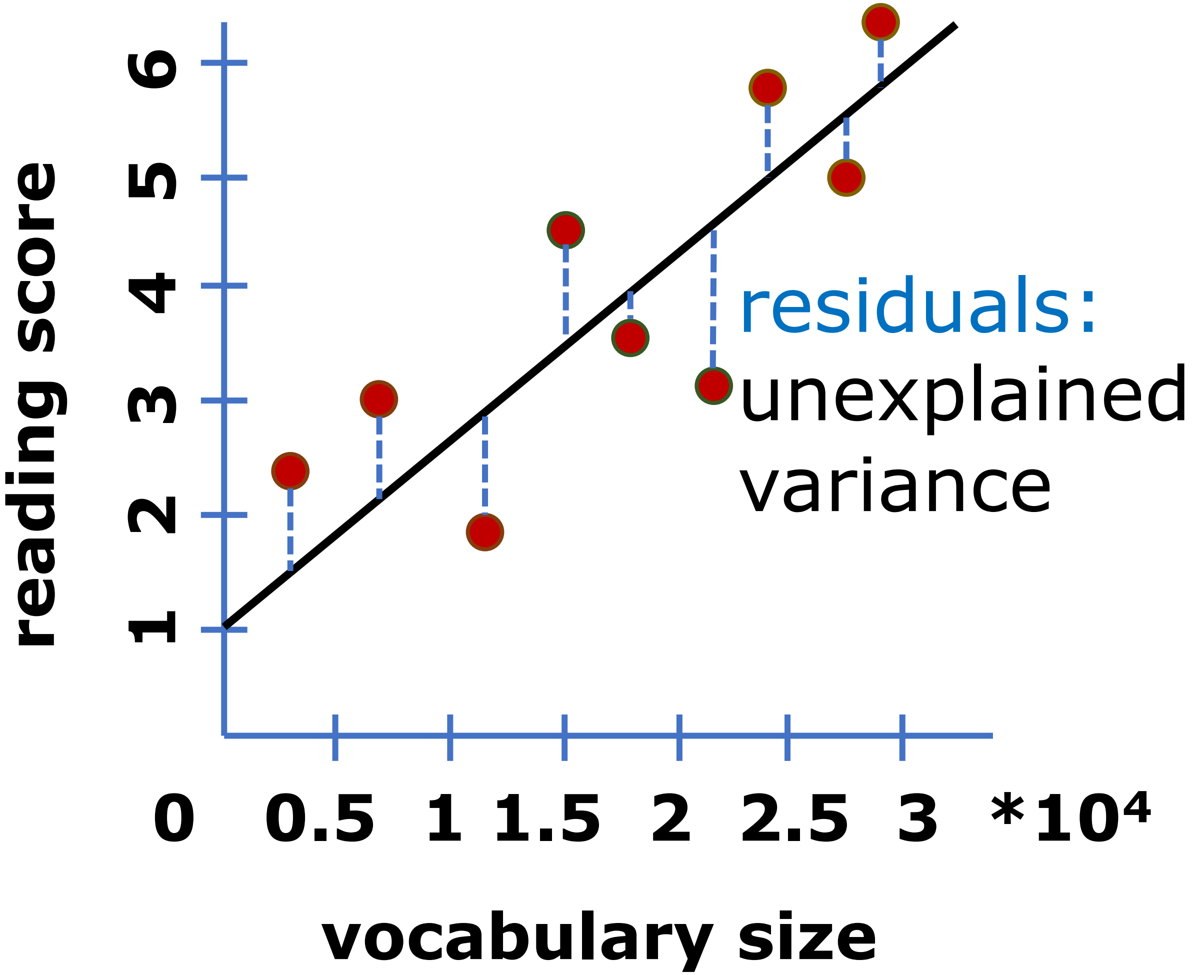


categorical
compare group mean



continuous
model relationship

Interpreting regression model



$y = 2x + 1$ $\quad$ $y = b_1x + b_0$

slope $\quad$ intercept

slope: vocabulary size increase by 10,000, reading score increases by 2 points on average.

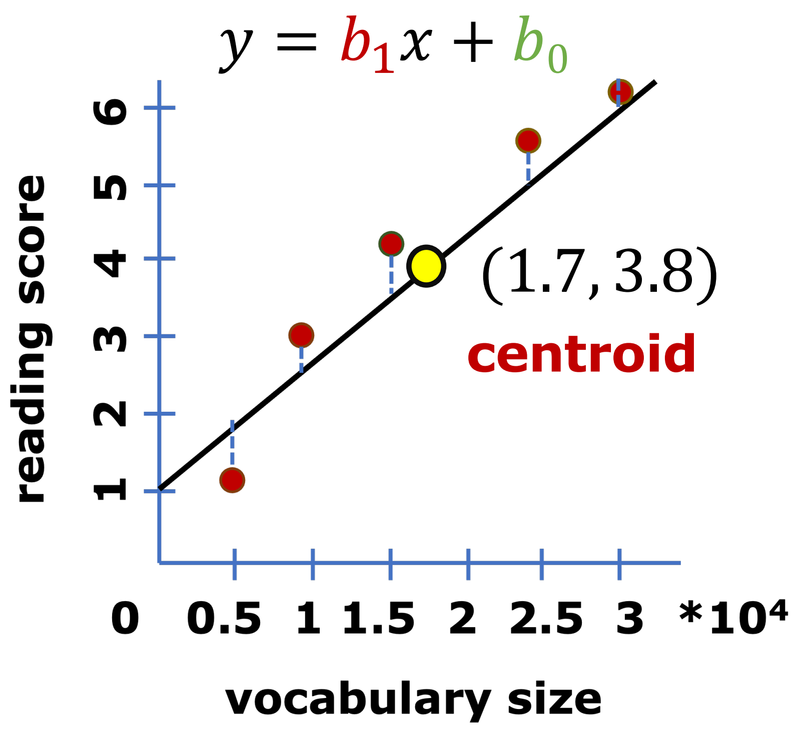
intercept: the expected y when x=0, may or may not make sense

Estimating regression coefficients

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

reading score	1,3,4,5,6 (M=3.8)
vocabulary size	0.5,1,1.5,2.5,3 (M=1.7)



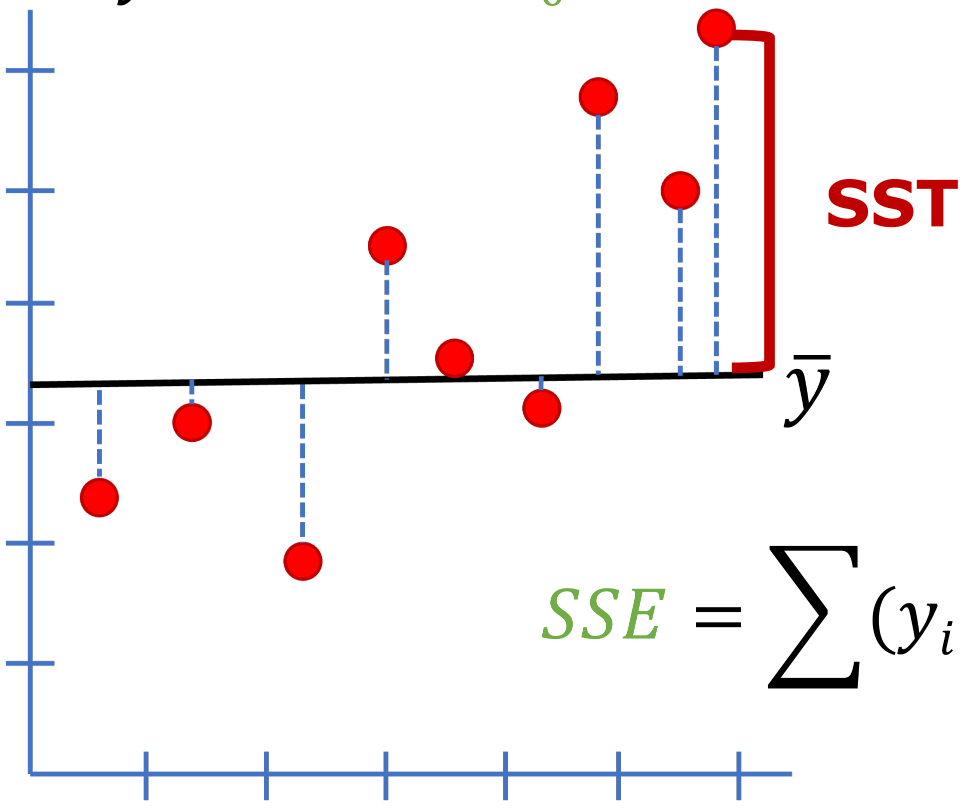
$$b_1 = \frac{(0.5-1.7)(1-3.8)+(1-1.7)(3-3.8)+(1.5-1.7)(4-3.8)+(2.5-1.7)(5-3.8)+(3-1.7)(6-3.8)}{(0.5-1.7)^2+(1-1.7)^2+(1.5-1.7)^2+(2.5-1.7)^2+(3-1.7)^2}$$

$$= 1.53$$

$$b_0 = 3.8 - 1.53 * 1.7 = 1.199$$

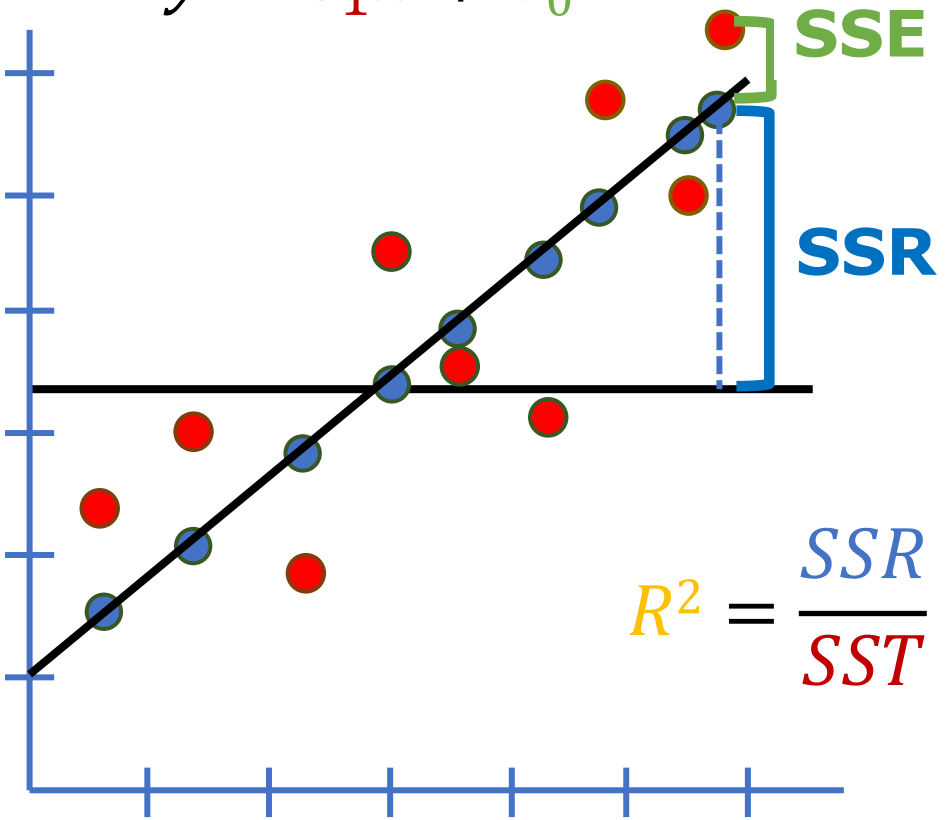
A tale of two models

$$y = 0 * x + b_0$$



$$SSE = \sum (y_i - \hat{y})^2$$

$$y = b_1 x + b_0$$



$$R^2 = \frac{SSR}{SST}$$

x has no effect on y

x has positive effect on y

$$SST = \sum (y_i - \bar{y})^2$$

$$SSR = \sum (\hat{y}_i - \bar{y})^2$$

R^2 and F ratio

$$SSR = \sum (\hat{y}_i - \bar{y})^2 \qquad MSR = \frac{SSR}{k-1} \qquad F = \frac{MSR}{MSE}$$

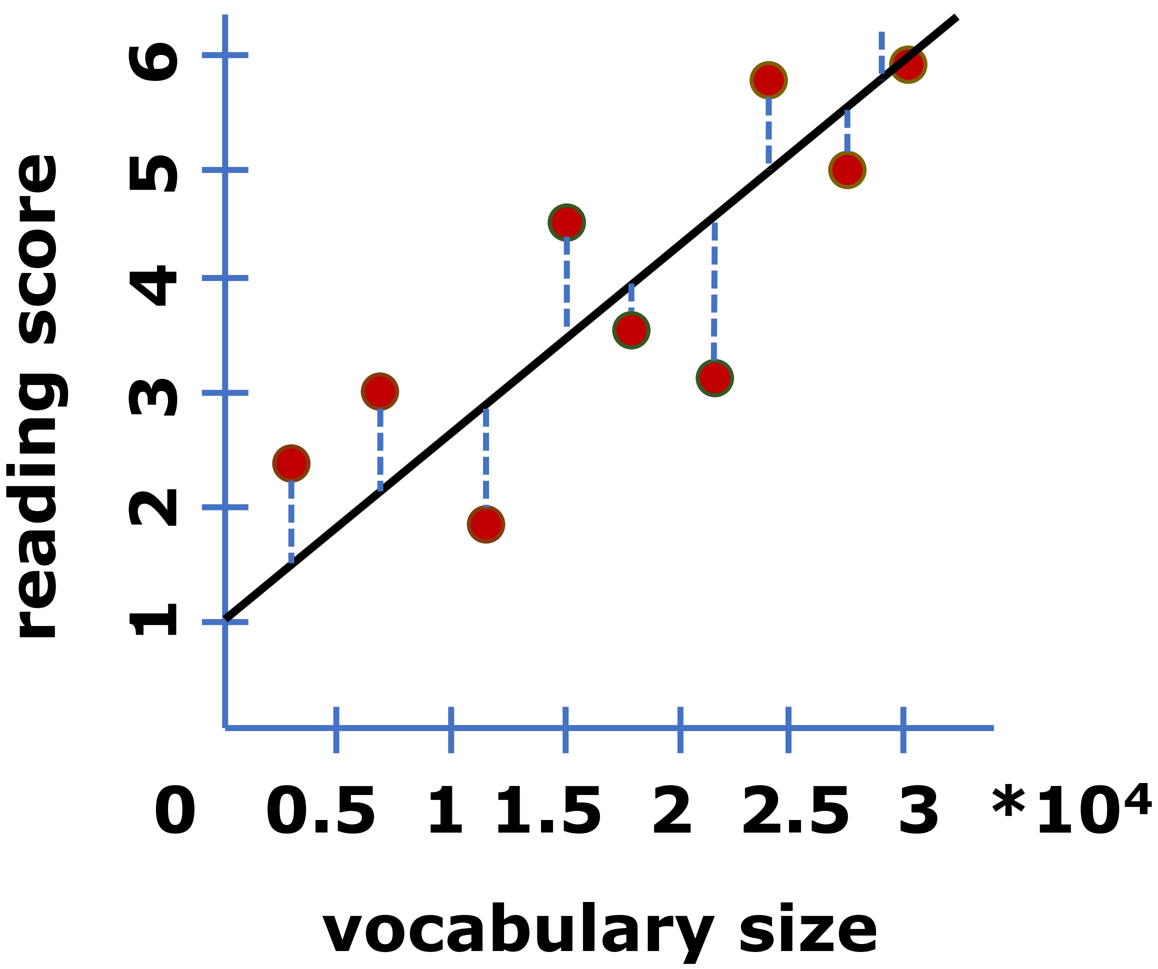
$$SSE = \sum (y_i - \hat{y})^2 \qquad MSE = \frac{SSE}{n-2}$$

$$SST = \sum (y_i - \bar{y})^2$$

$$R^2 = \frac{SSR}{SST}$$

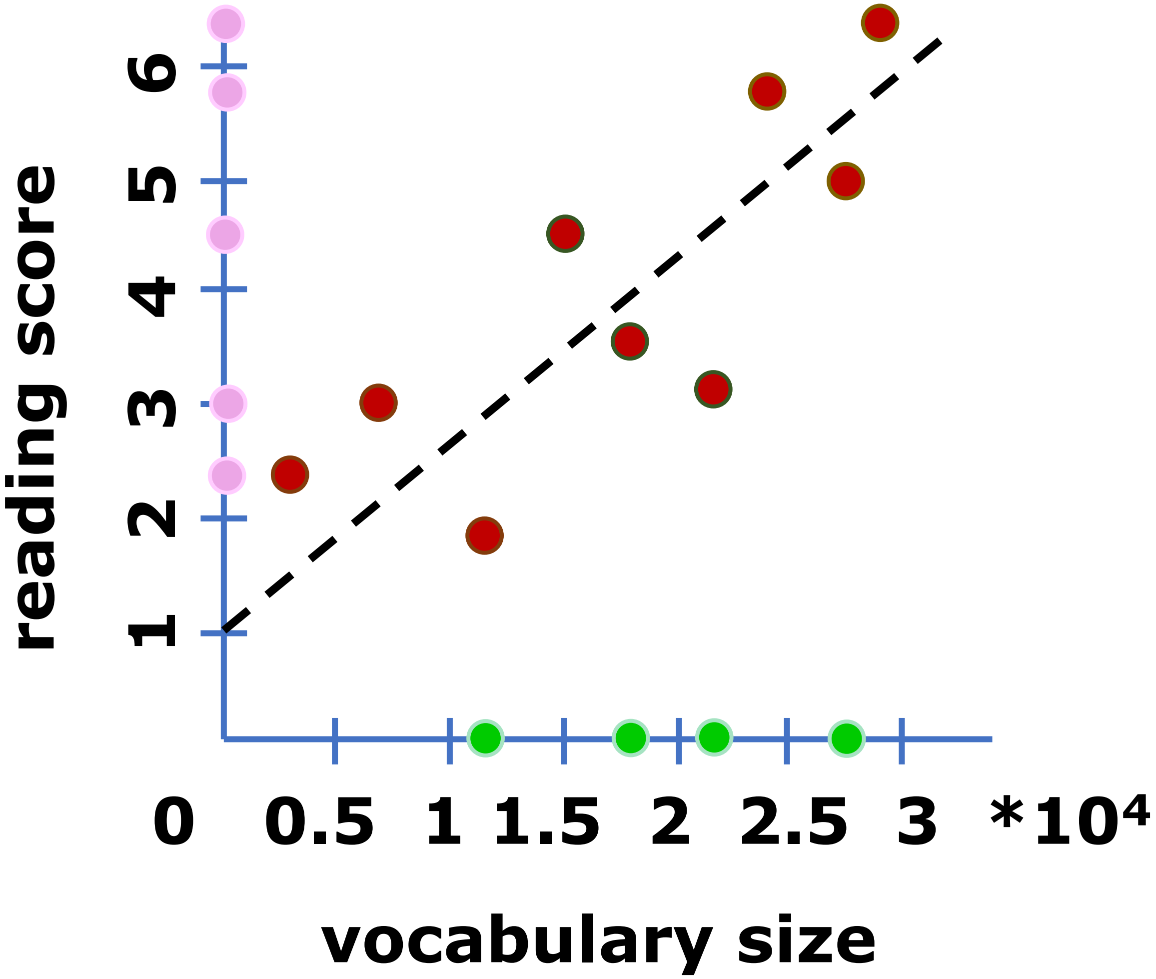
k: number of model parameters (slope, intercept)
n: number of data points

Correlation vs. Regression



causal relationship, make prediction

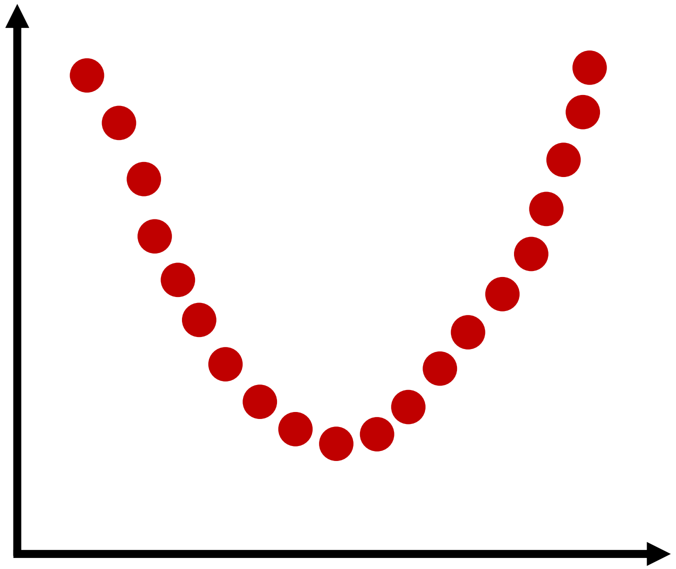
x ⟶ y



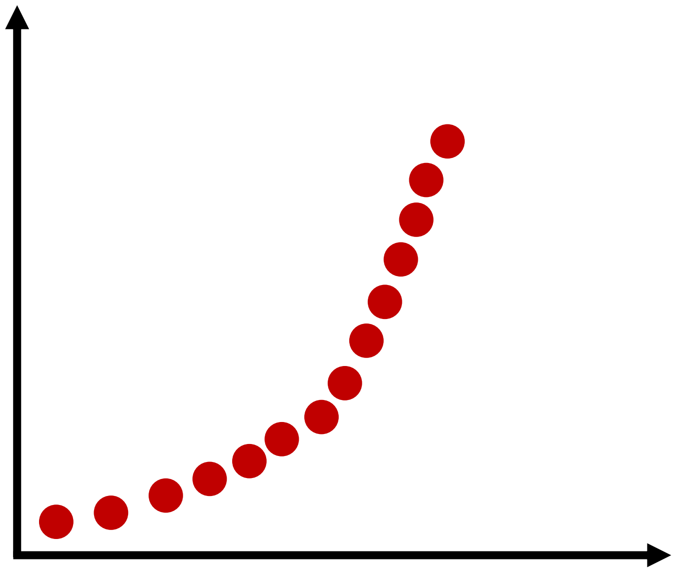
linear relationship, interchangeable

x ⟷ y

Non-linear relationships



Quadratic

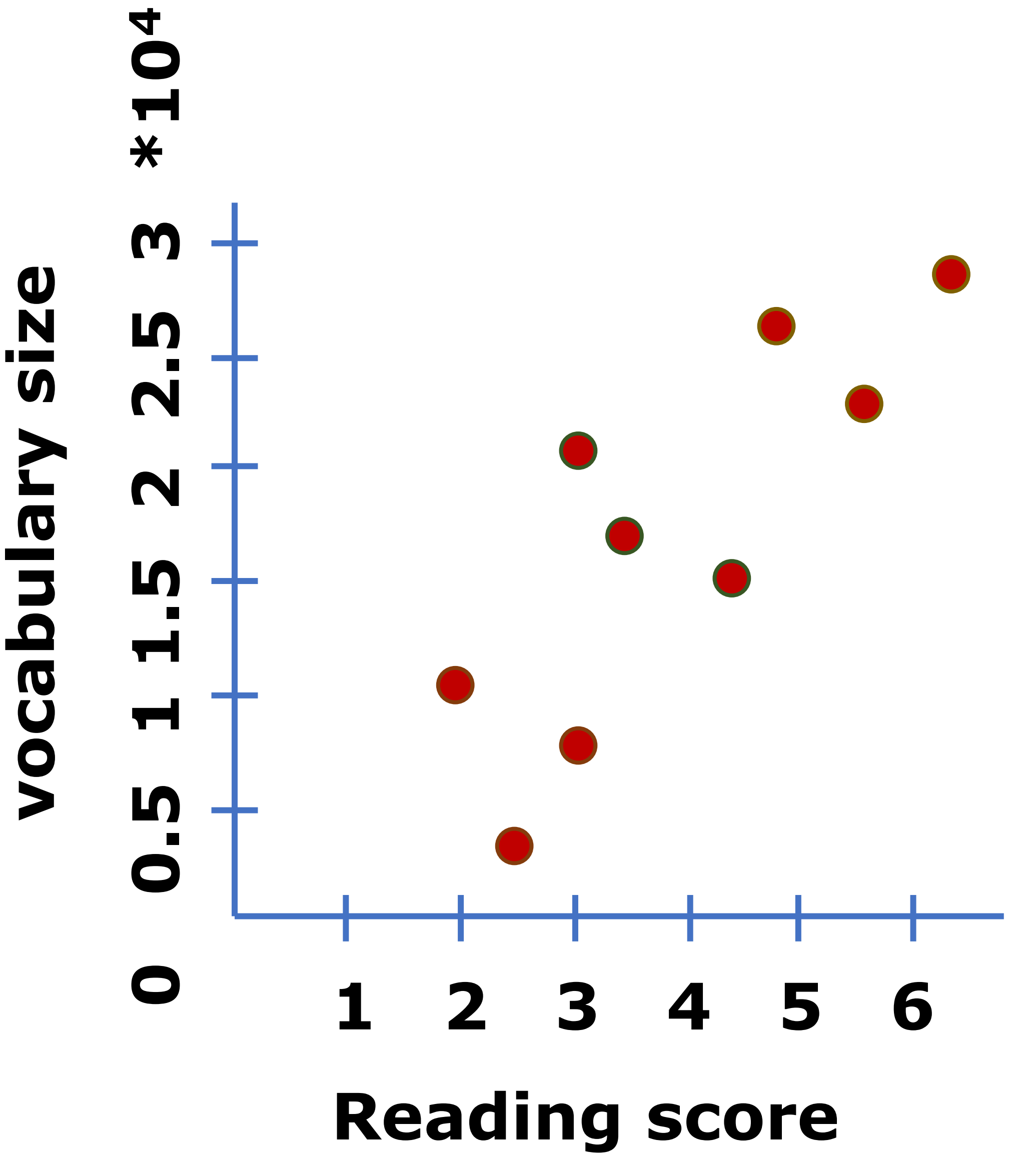


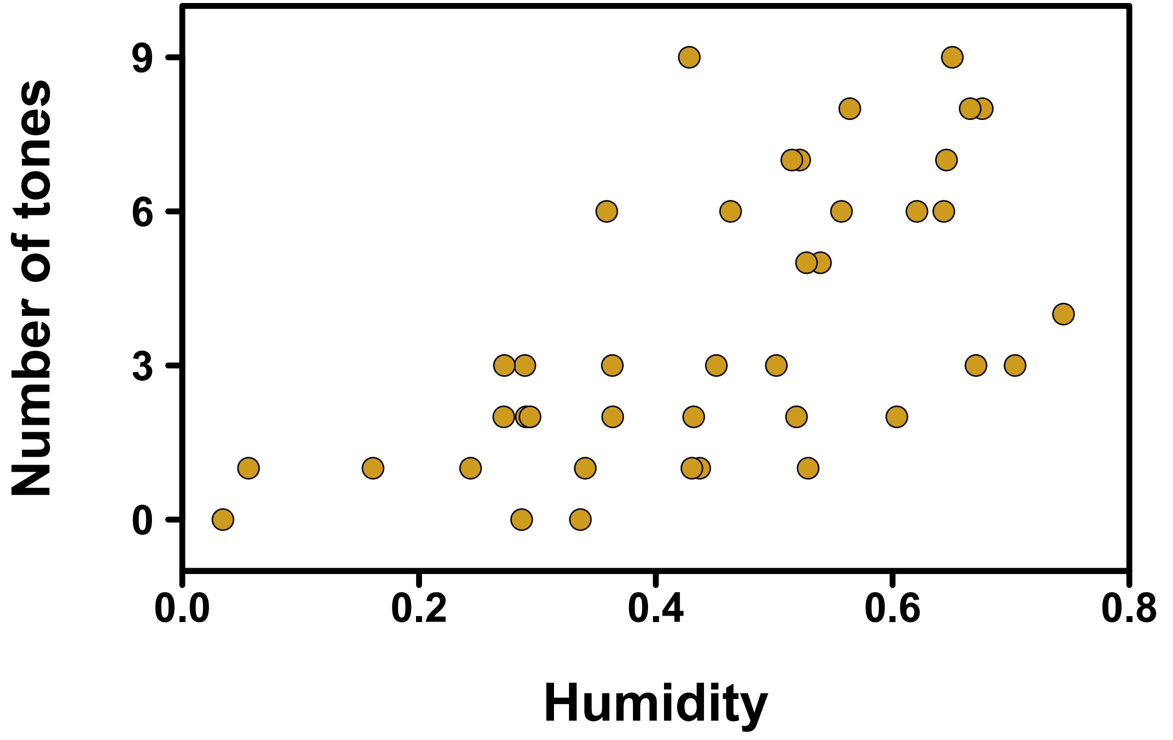
Exponential

e.g., energy and temperature

e.g., pandemic outbreak

Correlation is not causation



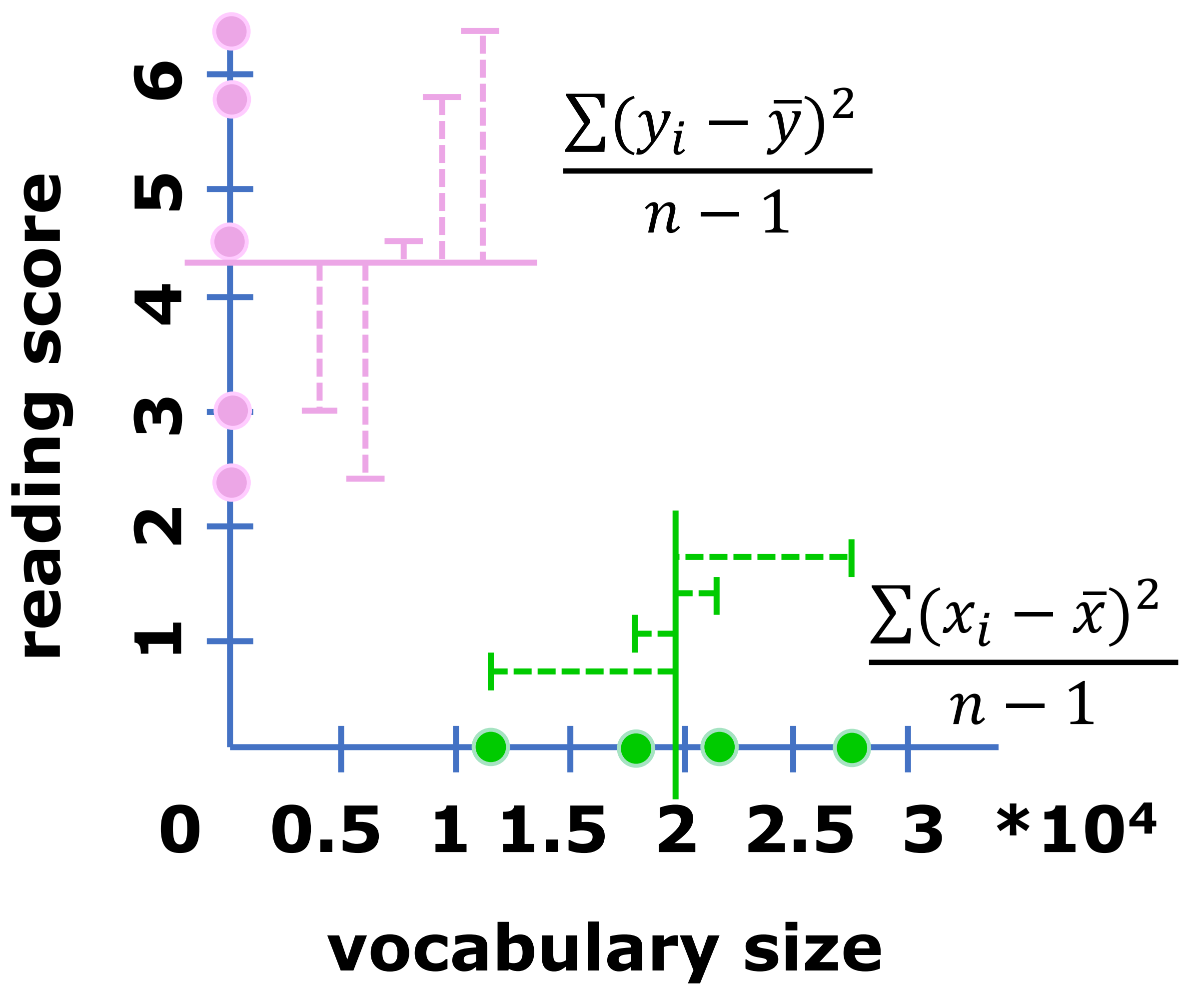


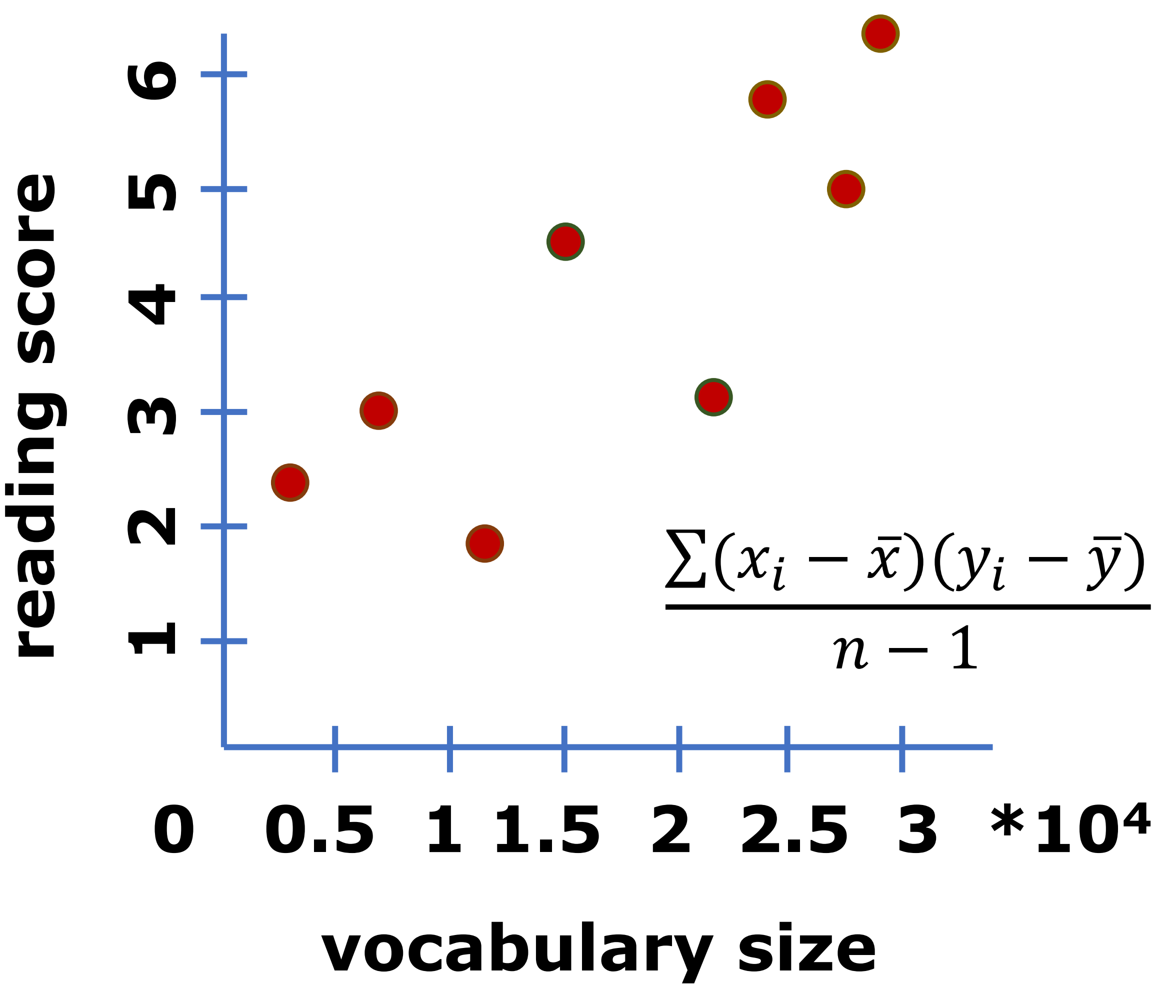
Correlation to causation:

The correlation is strong.

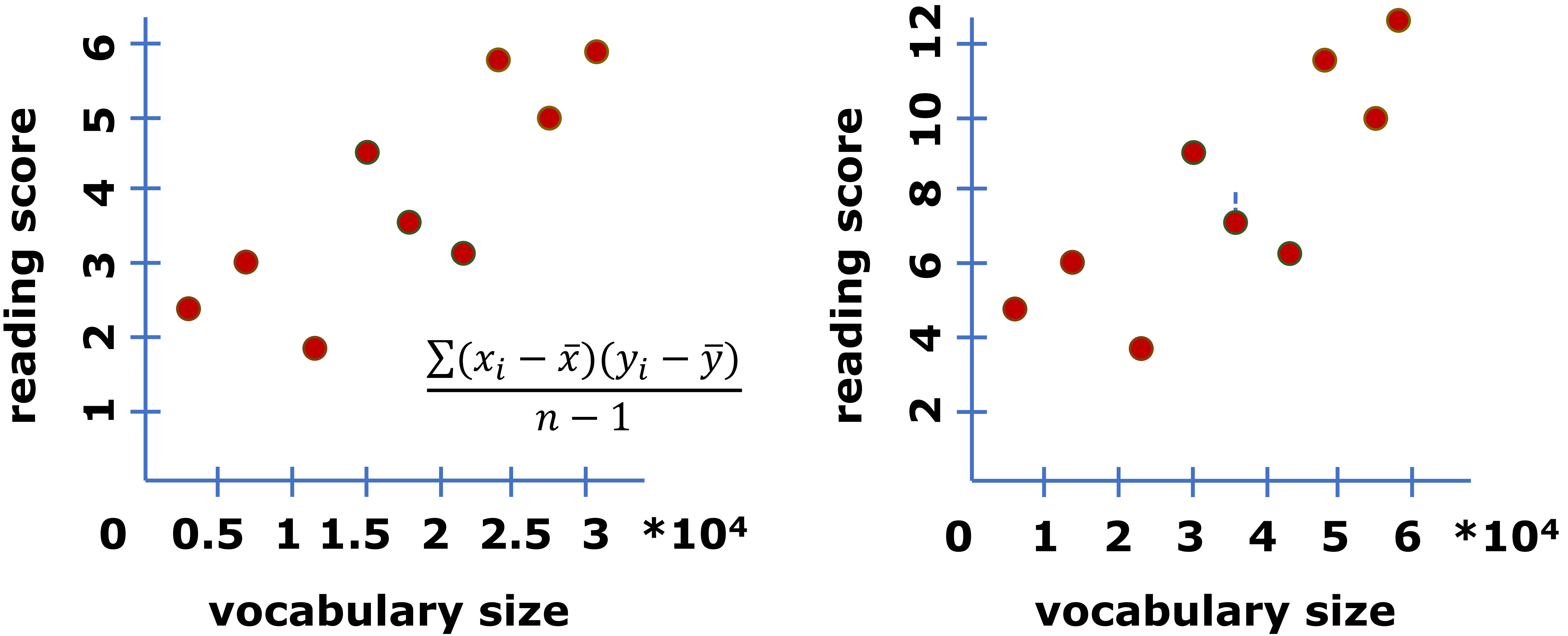
The causal effect is plausible.

Variance vs. Covariance





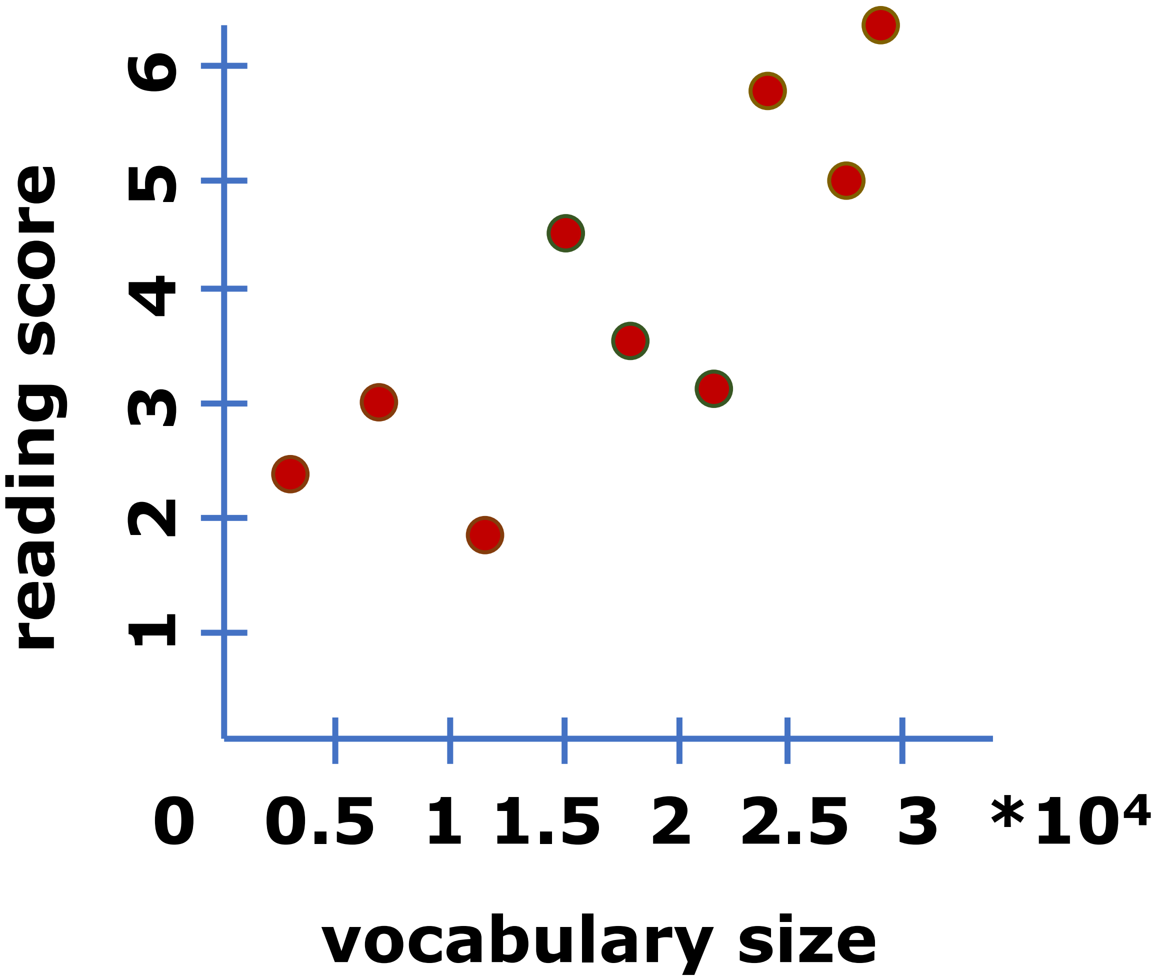
Covariance: Direction of the relationship



$$\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

covariance is influenced by scale: can only tell the direction of the relationship.

Correlation: Direction and strength

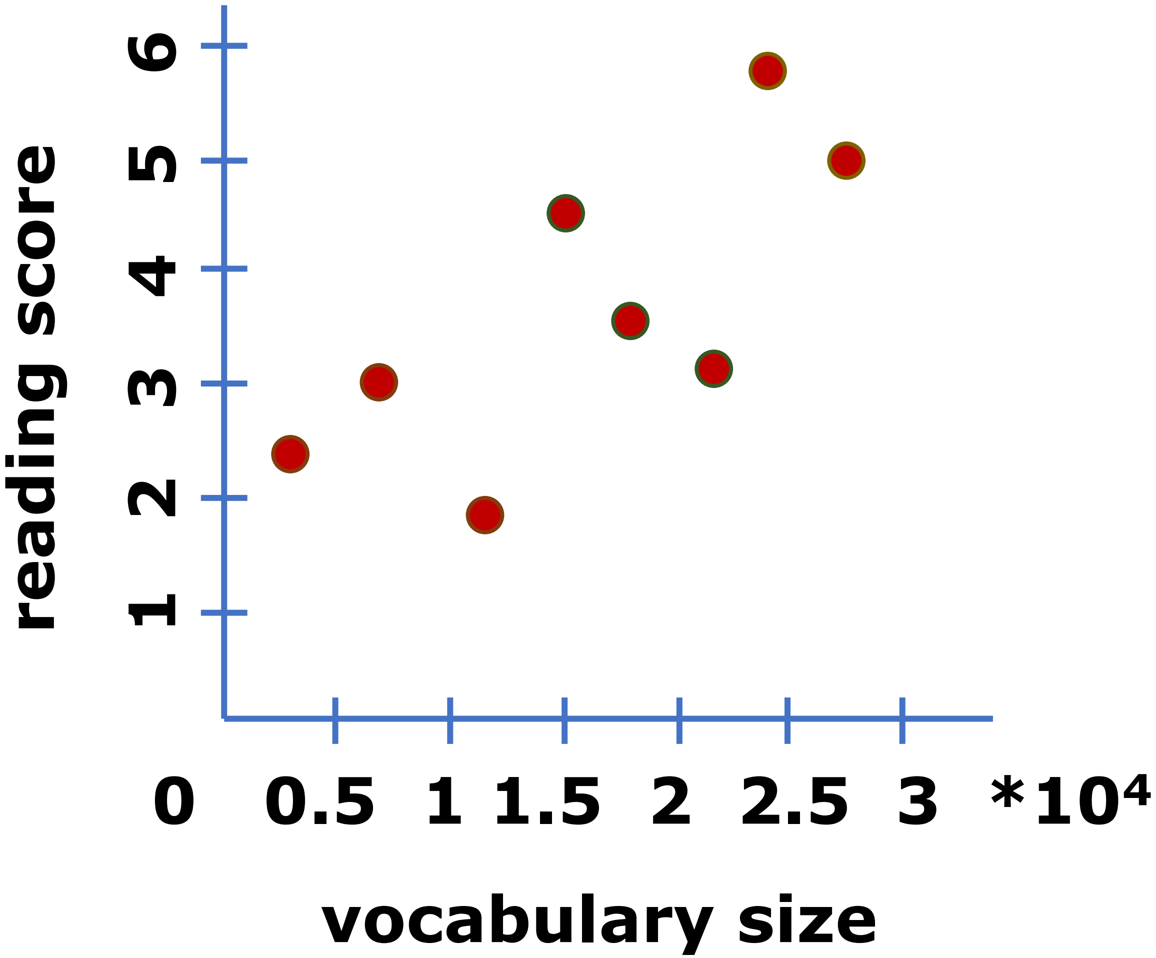


$$Cov = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$r = \frac{Cov(x,y)}{std(x) * std(y)}$$

***r* (correlation coefficient):**
a number between -1 and 1

Significance of correlation coefficient



1. Determine H_0 and H_a:

H_0: $\rho = 0$

H_a: $\rho \neq 0$ → population level

2. Calculate the test statistic

$$t = \frac{r-\rho}{se(r)} \qquad se(r) = \frac{\sqrt{1-r^2}}{\sqrt{n-2}}$$

$$= \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad \sim T_{n-2}$$

3. Compare *p* with α

Interpret an r value in the context of the scientific question

r	**Interpretation**
0.00-0.10	No correlation
0.10-0.39	Weak correlation
0.40-0.69	Moderate correlation
0.70-0.89	Strong correlation
0.90-1.00	Very strong correlation

Interpret *r* value in the context of the scientific question

Neural dynamics of semantic composition

Bingjiang Lyu[a], Hun S. Choi[a], William D. Marslen-Wilson[a], Alex Clarke[a], Billi Randall[a], and Lorraine K. Tyler[a,1]

[a]Centre for Speech, Language and the Brain, Department of Psychology, University of Cambridge, CB2 3EB Cambridge, United Kingdom

PNAS RESEARCH ARTICLE | NEUROSCIENCE PSYCHOLOGICAL AND COGNITIVE SCIENCES

A hierarchy of linguistic predictions during natural language comprehension

Micha Heilbron[a,b,1], Kristijan Armeni[a], Jan-Mathijs Schoffelen[a], Peter Hagoort[a,b], and Floris P. de Lange[a]