Fundamentals of Statistics for Language Sciences LT2206

Jixing Li

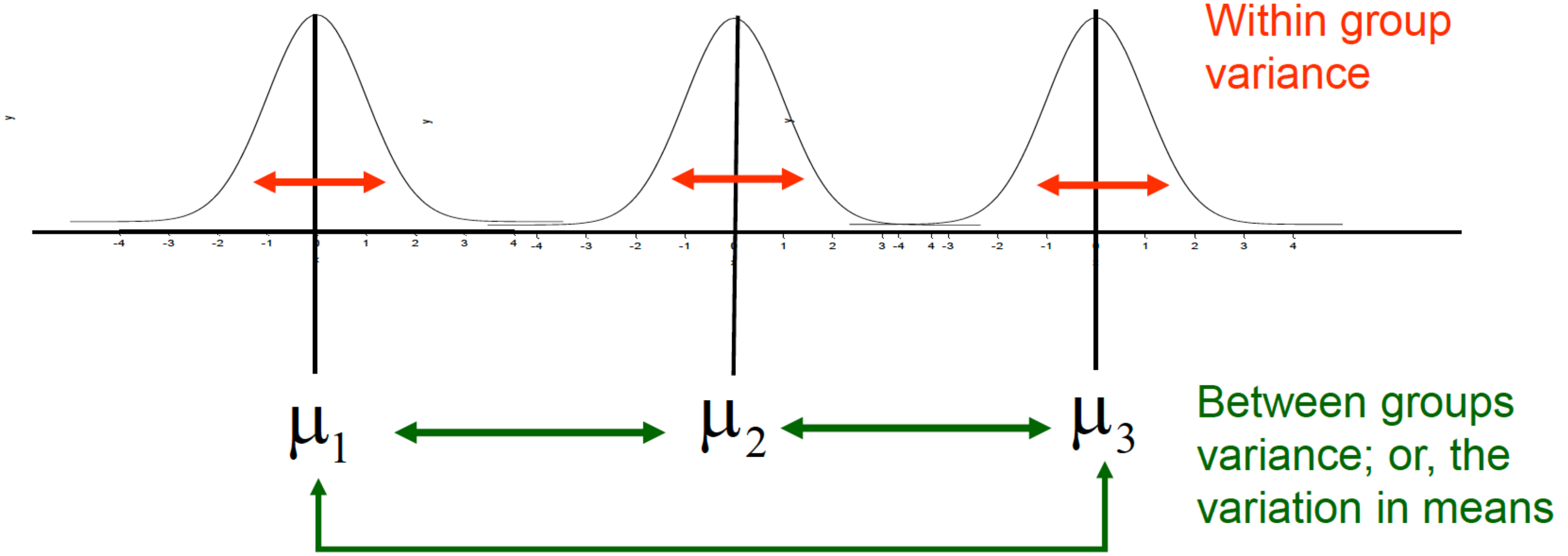
Lecture 8: Simple Linear Regression

One-way ANOVA

1. **Compute sum of squares between group (SSB):**

$$\sum_{1}^{k} n_k(\overline{X_k} - \overline{X})^2$$

2. **Compute sum of squares within group / error (SSE):**
   sum of the squared differences between each individual observation and the group mean of that observation.
3. **Compute sum of squares total (SST): SSB+SSE**



The ANOVA table

Source of variation	Degrees of freedom	Sum of squares	Mean squares	*F* statistic
Between	k-1	SSB	MSB =SSB/(k-1)	MSB/MSE
Within (Error)	n-k	SSE	MSE =SSE/(n-k)	
Total	n-1	SST		

k: number of groups
n: total number of samples

Under H_0, F should tend to be close to 1. Under H_a, F should exceed 1, by an amount depending on both n and k.

Example: The cocktail party experiment

group	mixed	mean	grand mean
hearing-impaired	2,2,3	2.33	2.78
normal	2,4,4	3.33	
children	2,3,3	2.67	

SSB = $3*(2.33-2.78)^2 + 3*(3.33-2.78)^2 + 3*(2.67-2.78)^2 = 2.77$

SSE = $(2-2.33)^2 + (2-2.33)^2 + (3-2.33)^2$
$+ (2-3.33)^2 + (4-3.33)^2 + (4-3.33)^2$
$+ (2-2.67)^2 + (3-2.67)^2 + (3-2.67)^2 = 4$

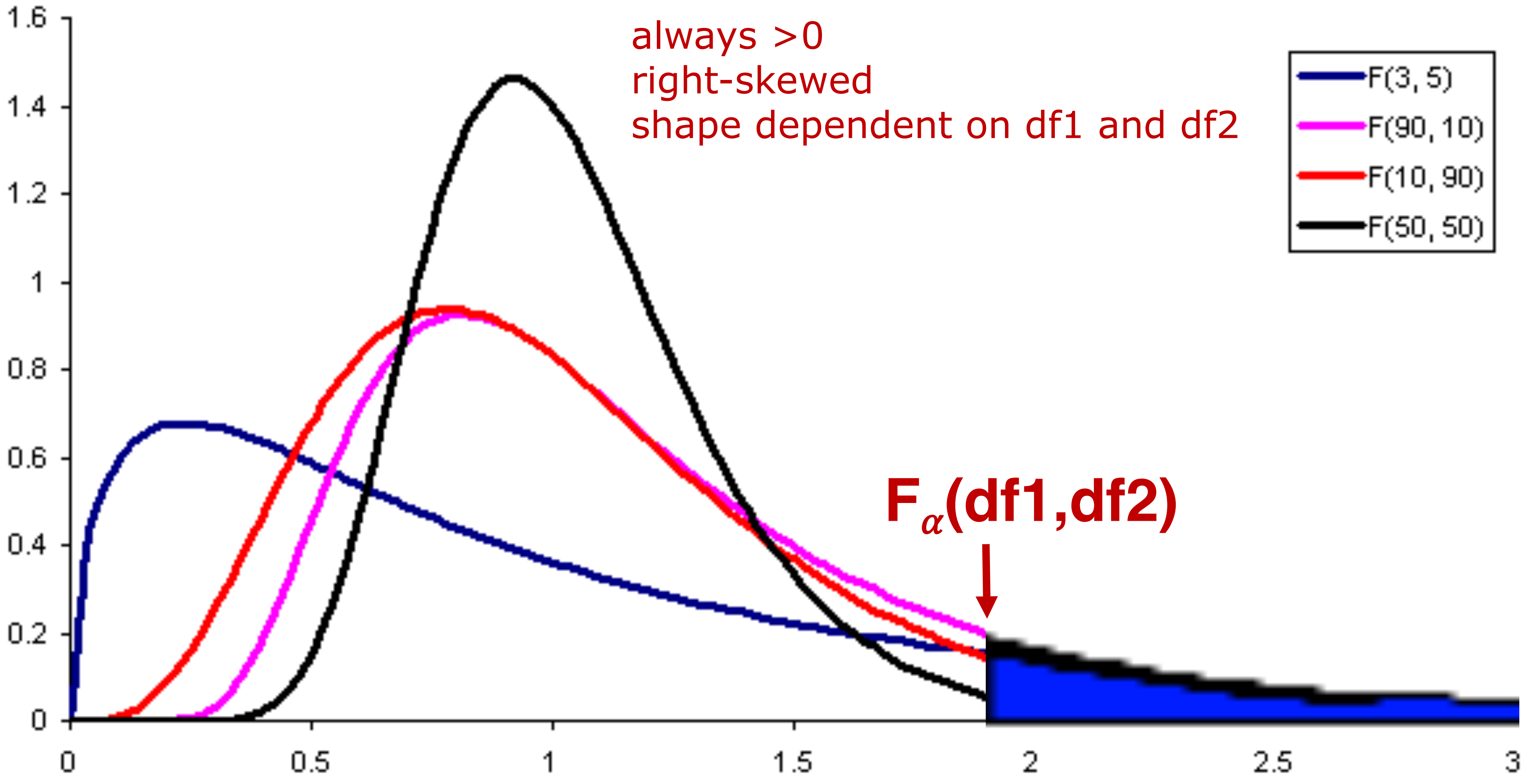
k = 3, **n** = 9

MSB= SSB / (k-1) = 2.77/2 = 1.385

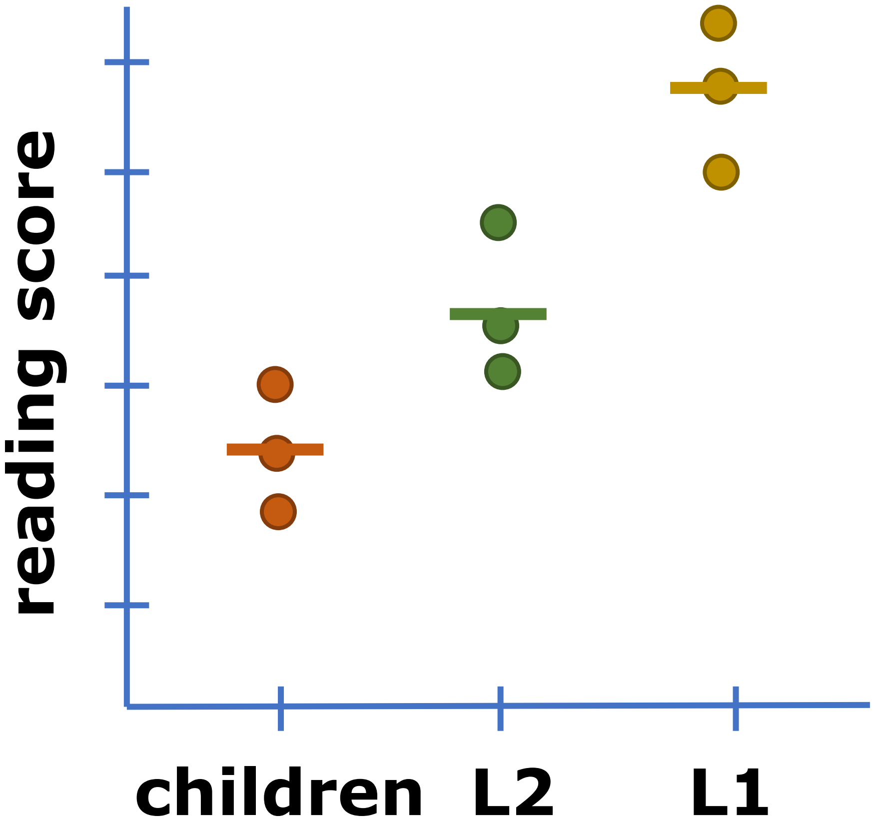
MSE = SSE / (n-k) = 0.67

F = MSB/MSE = 2.07

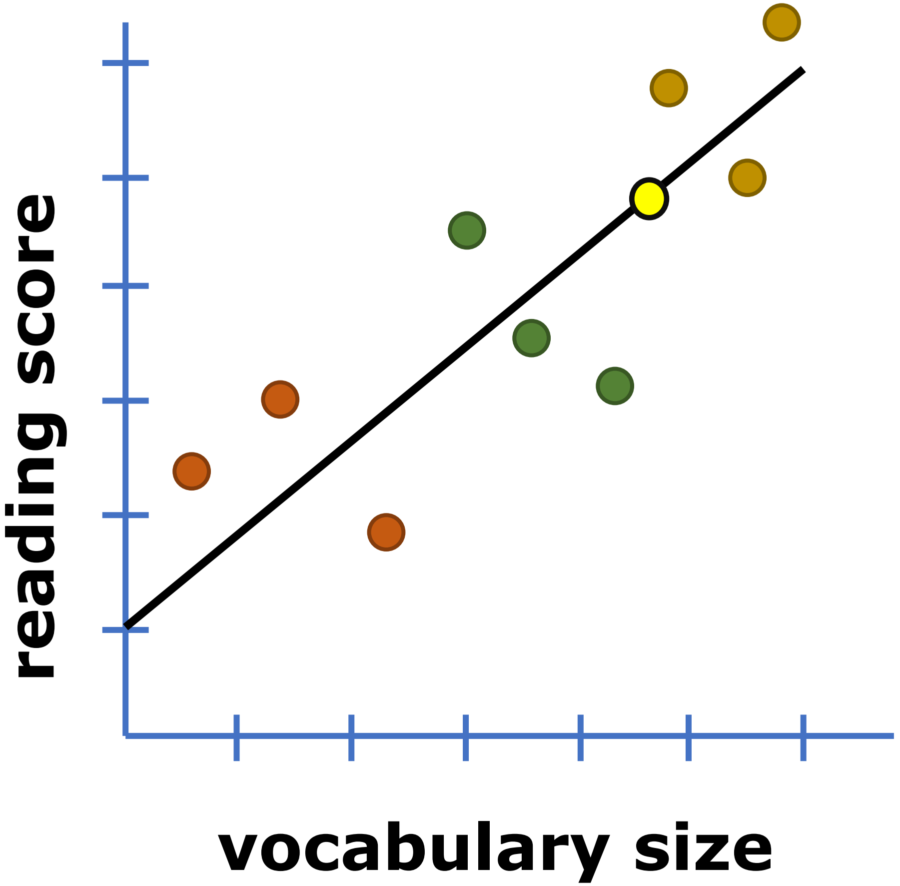
F-distribution



ANOVA vs. Regression

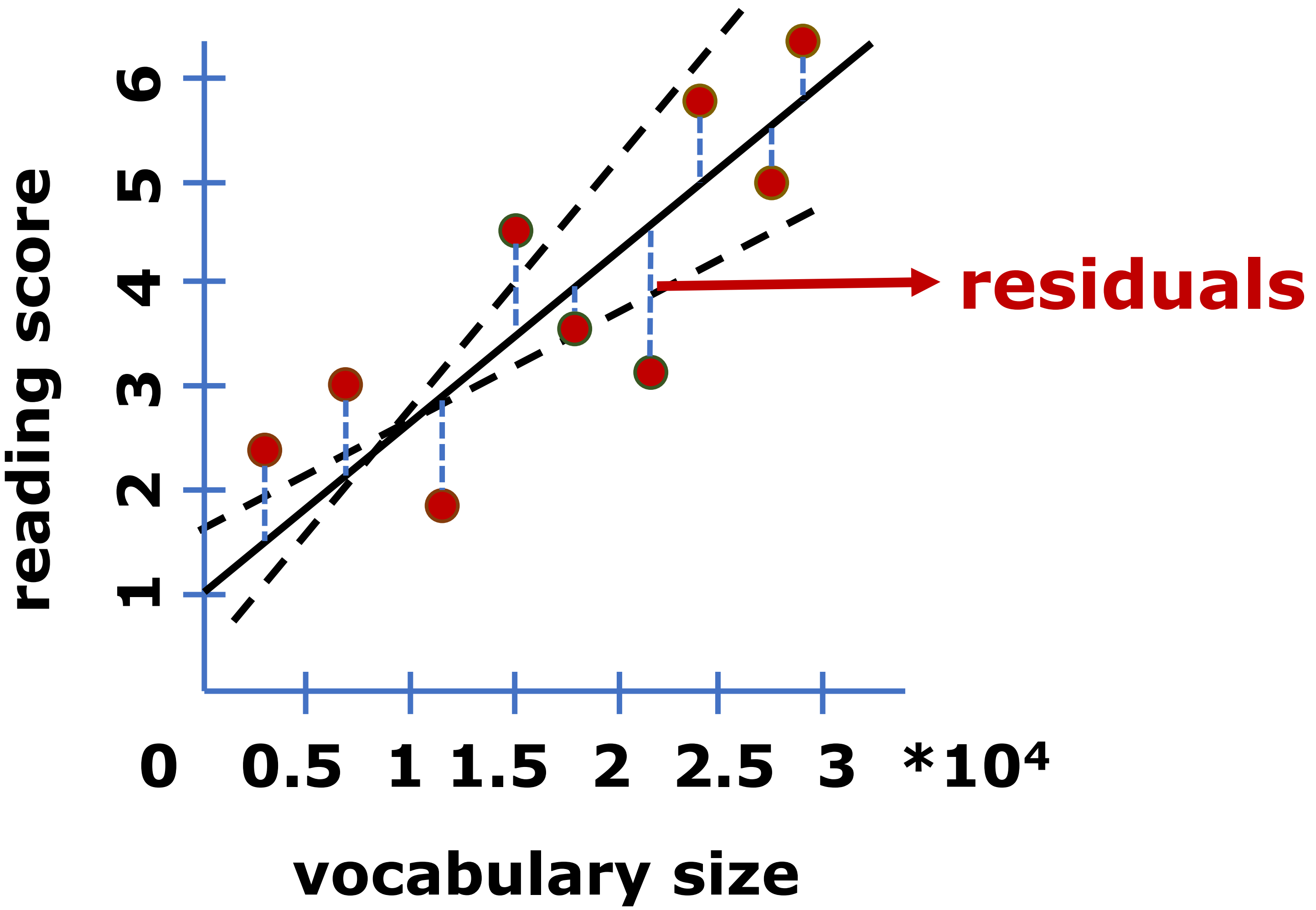


categorical
compare group mean



continuous
model relationship

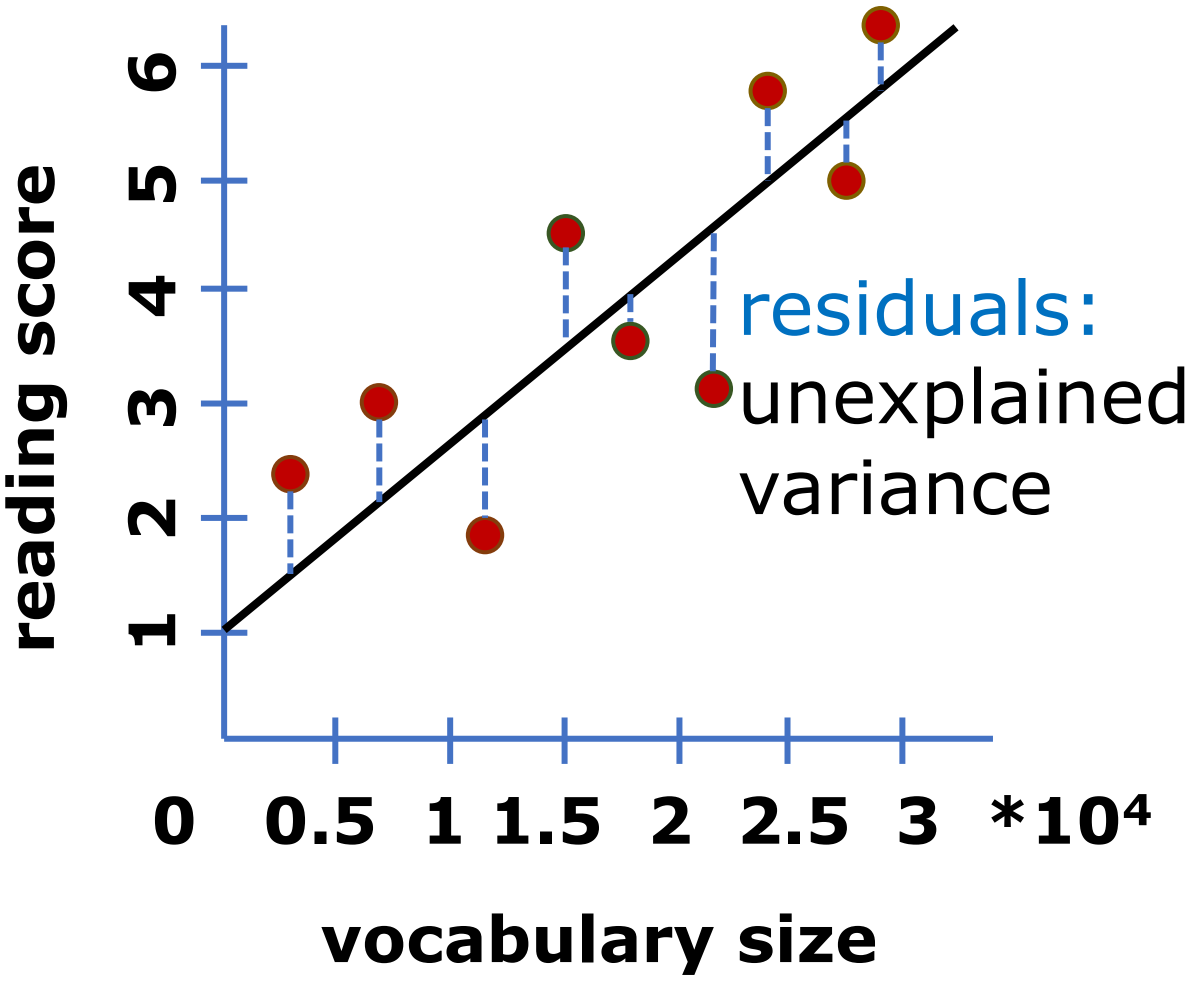
Fit the regression line



The best fit regression line minimizes the sum of squared residuals

→ least-squares regression line

Interpreting regression model



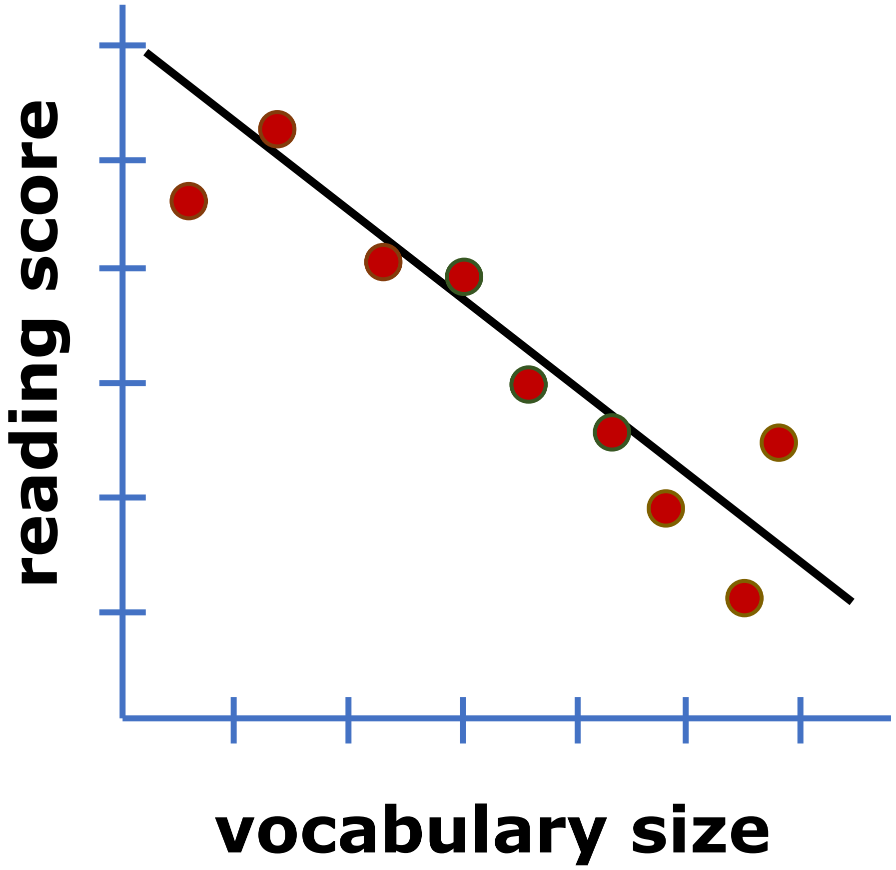
$y = 2x + 1$ $y = b_1x + b_0$

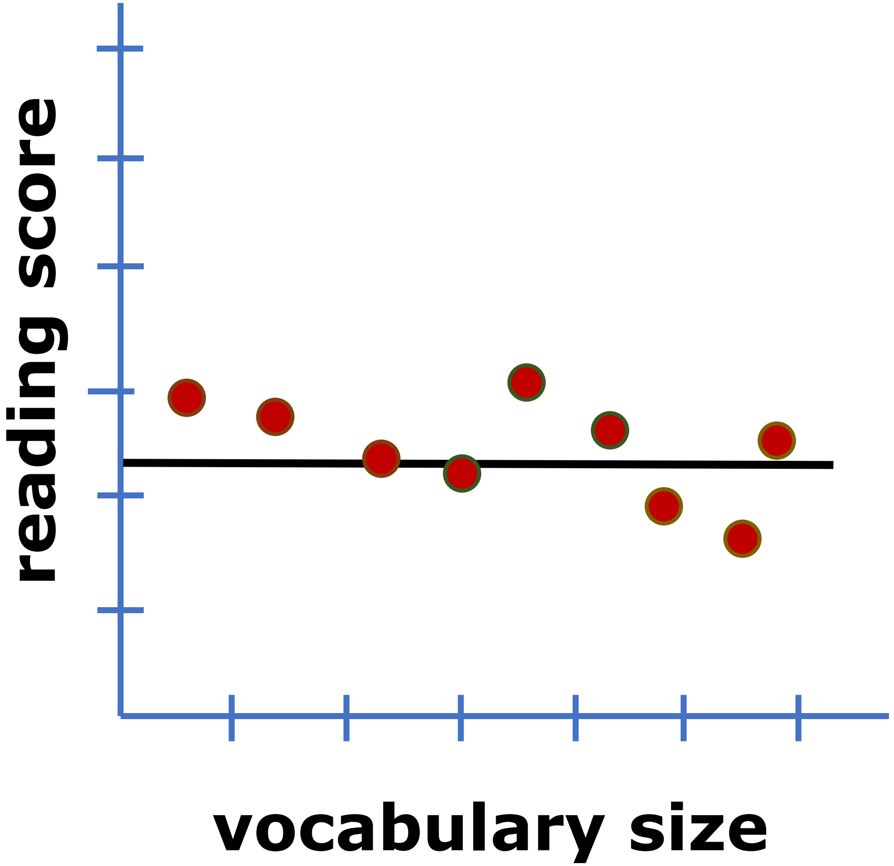
slope intercept

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

Interpreting regression model





negative b1: vocabulary size increases, reading score decreases

b1 close to 0: vocabulary size does not influence reading score

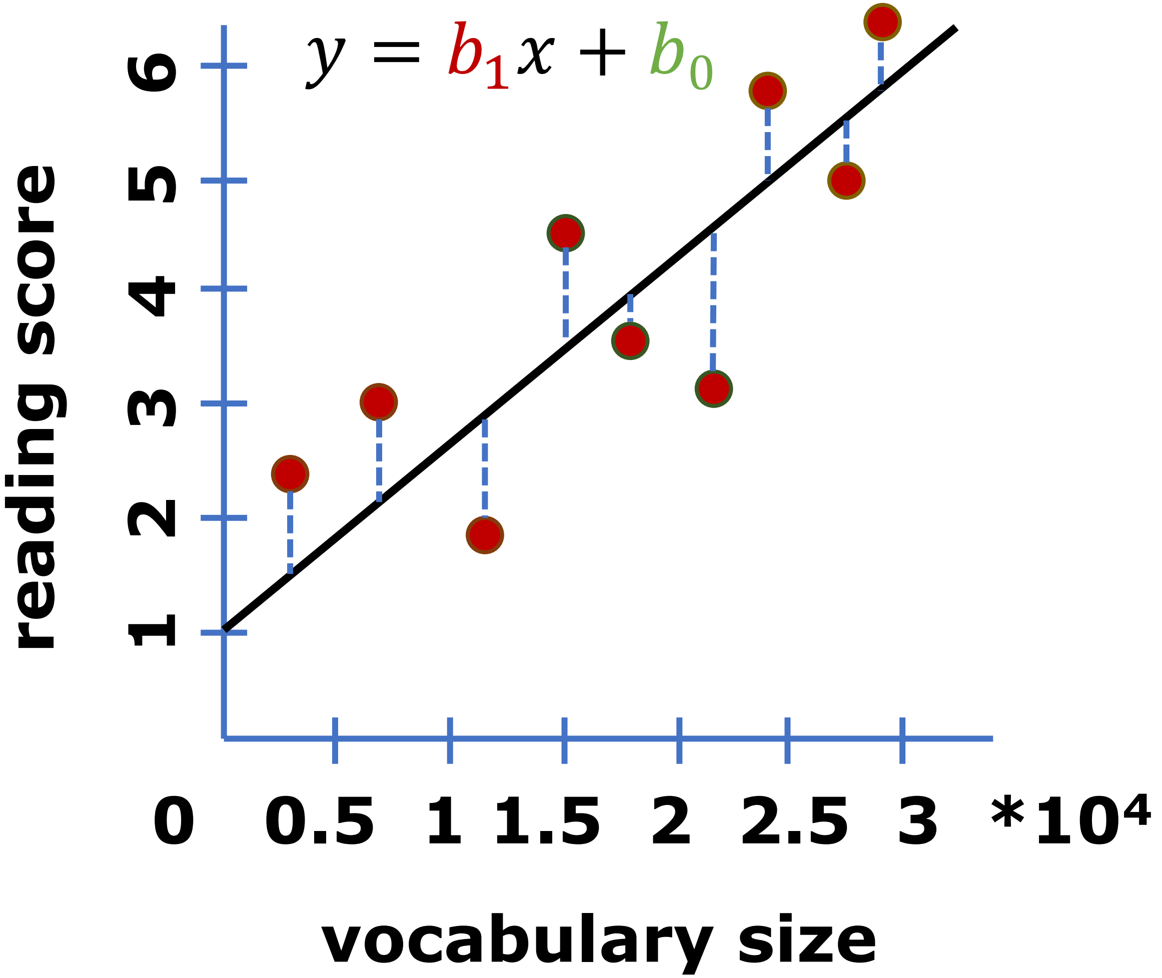
Estimating regression coefficients

Solution (calculus):

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

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

x_i = value of independent variable
y_i = value of dependent variable
$\bar{x}$ = mean of the independent variable
$\bar{y}$ = mean of the dependent variable

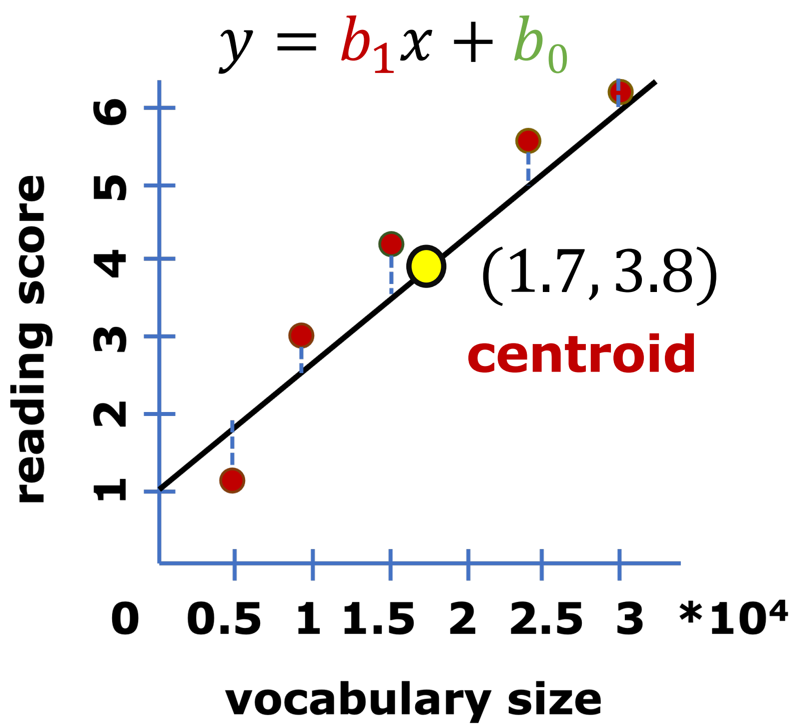


Estimating regression coefficients

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

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

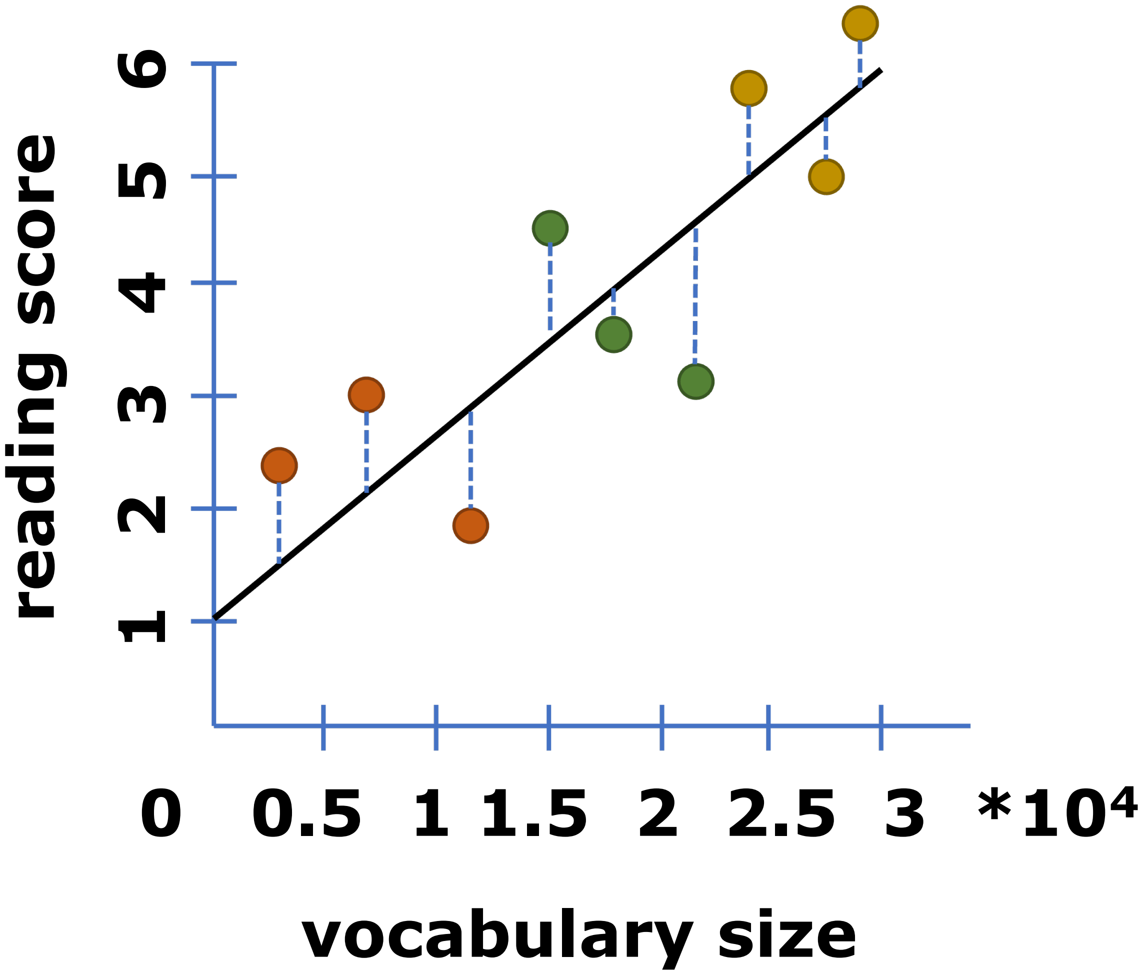
reading score	1,3,4,5,6 (M=3.8)
vocabulary size	0.5,1,1.5,2.5,3 (M=1.7)

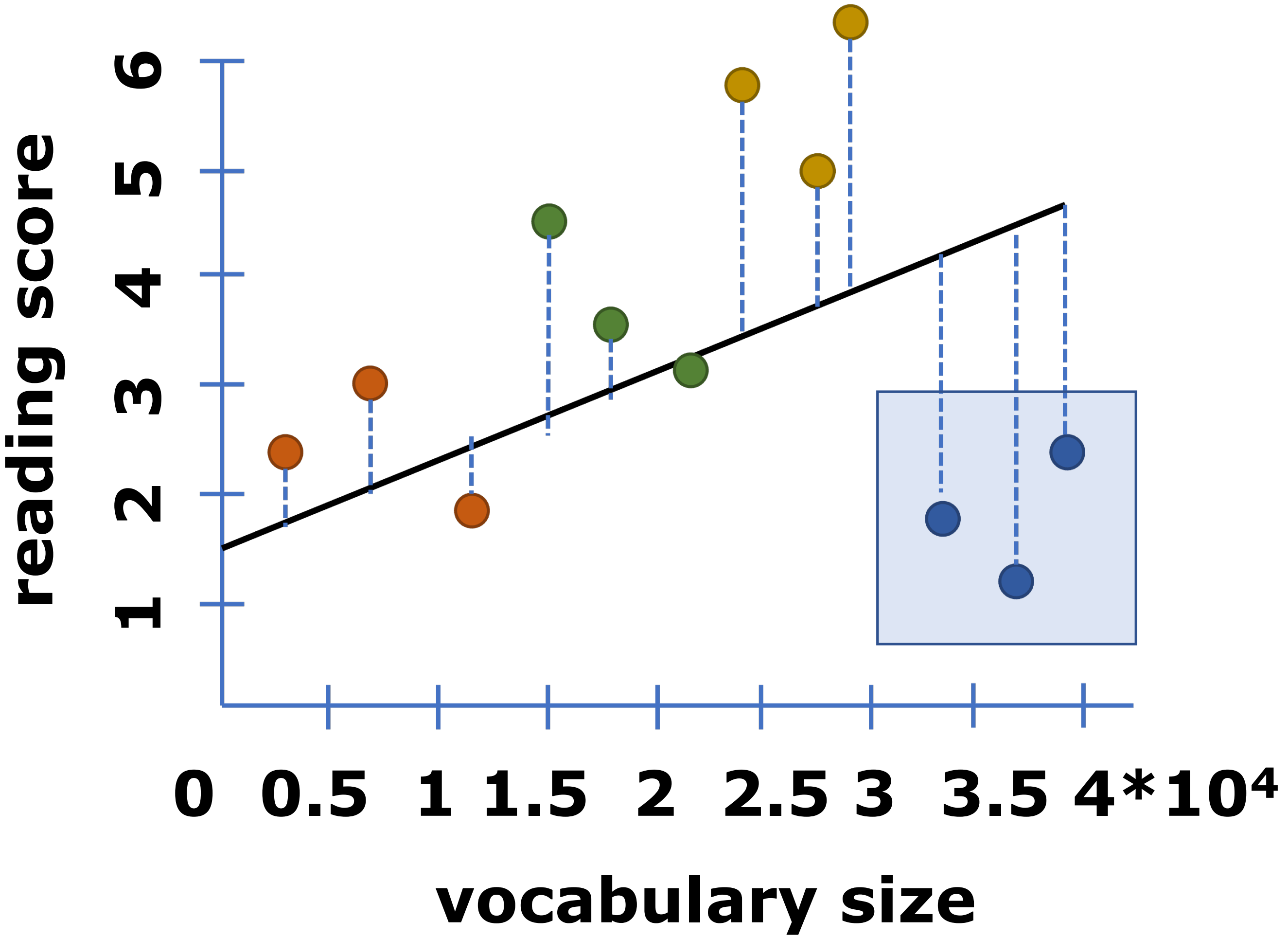


$$b_1 = \frac{(0.5-1.7)(1-3.8)+(1-1.7)(3-3.8)+(1.5-1.7)(4-3.8)+(2.5-1.7)(5-3.8)+(3-1.7)(6-3.8)}{(0.5-1.7)^2+(1-1.7)^2+(1.5-1.7)^2+(2.5-1.7)^2+(3-1.7)^2}$$
$$= 1.53$$

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

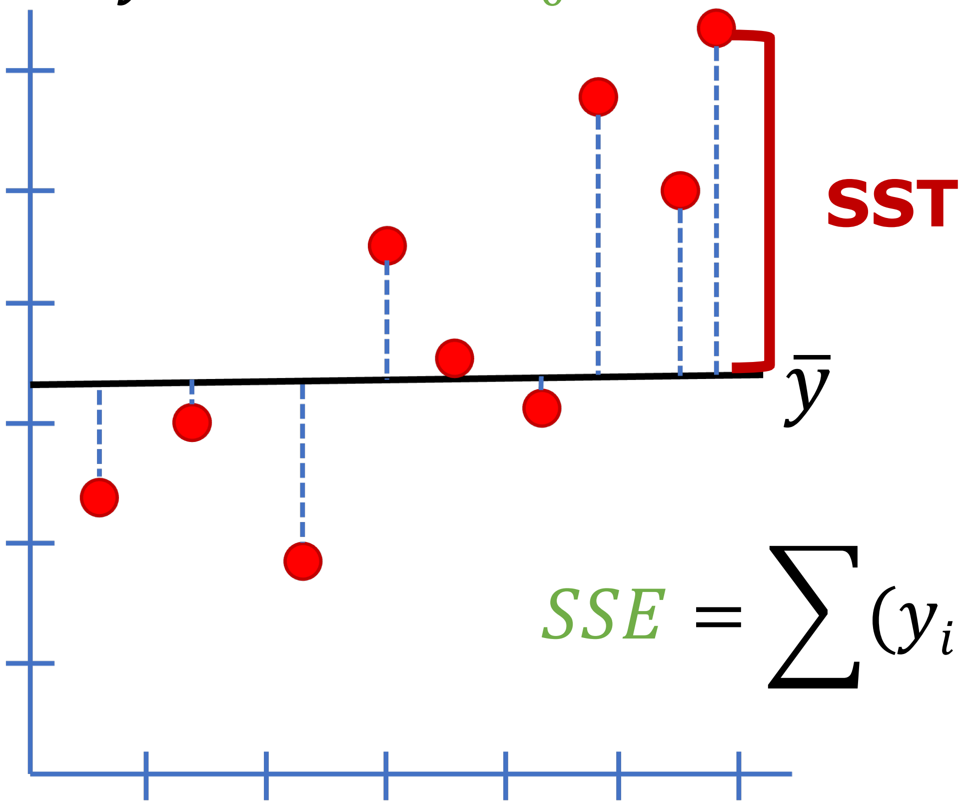
Goodness of fit





A tale of two models

$y = 0 * x + b_0$

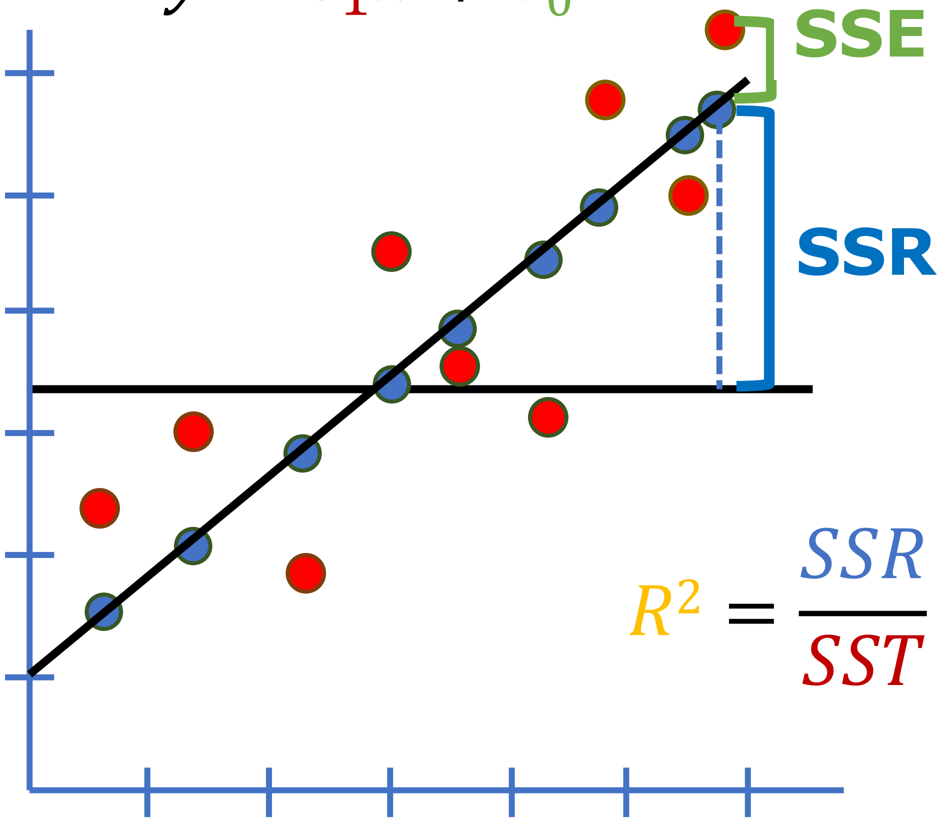


$SSE = \sum (y_i - \hat{y})^2$

x has no effect on y

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

$y = b_1 x + b_0$



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

x has positive effect on y

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

R^2 and F ratio

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

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

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

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

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

Example: Alice dataset

Participants listened to the first chapter of *Alice in Wonderland* in the fMRI scanner.

Question: How is the frequency of each word in the story affect the brain activity from 4 brain regions?