Regression, Correlation and Geometry

stats methodologists meeting

February 2016

a problem in Correlation

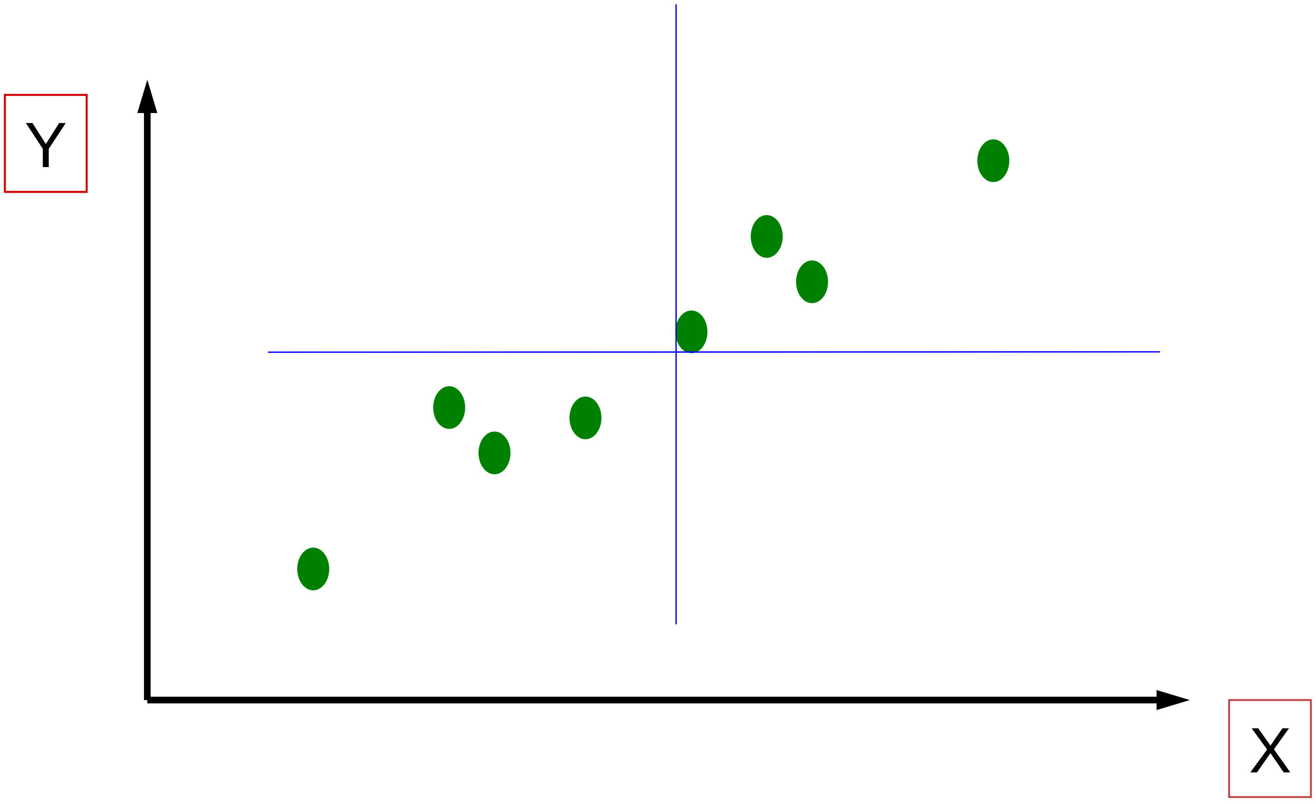
- $\rho(X,Y)$ = correlation between X and Y
- (suppose) $\rho(X,Y) = 0.7$ and $\rho(Y,Z) = 0.7$
- Q: what is the *least* possible value for

  $\rho(X,Z)$

Correlation

$$\rho = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 \sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}}$$

graphical representation of Regression



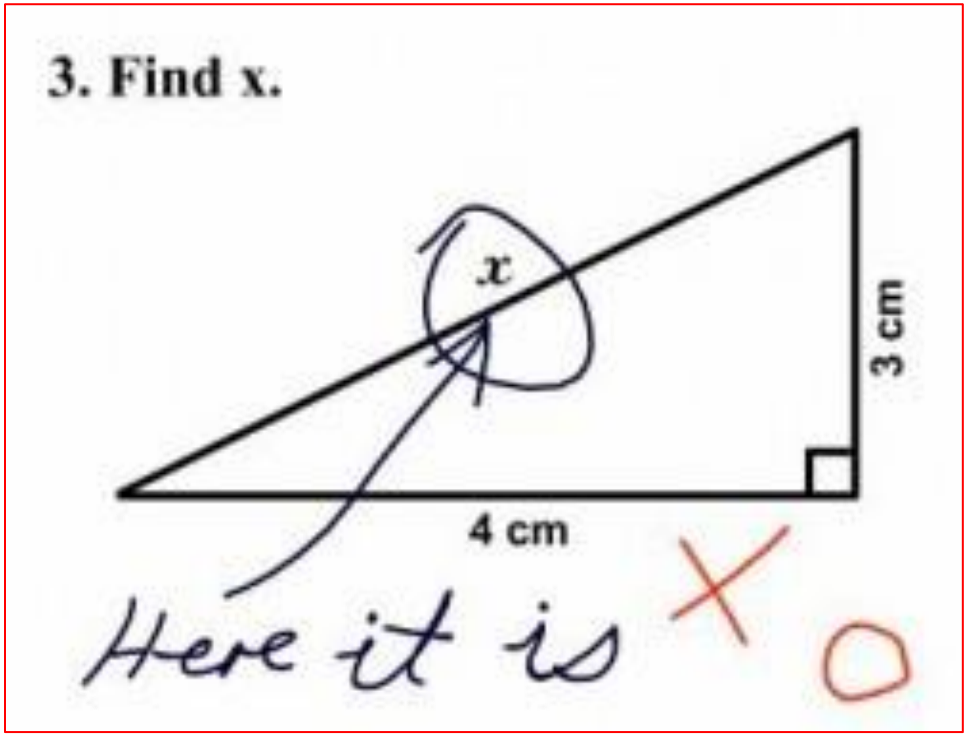
Correlation (variables centred)

$$\rho = \frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2 \sum_{i=1}^{n} Y_i^2}}$$

$$X_i = \left(x_i - \overline{x}\right)$$

$$Y_i = \left(y_i - \overline{y}\right)$$

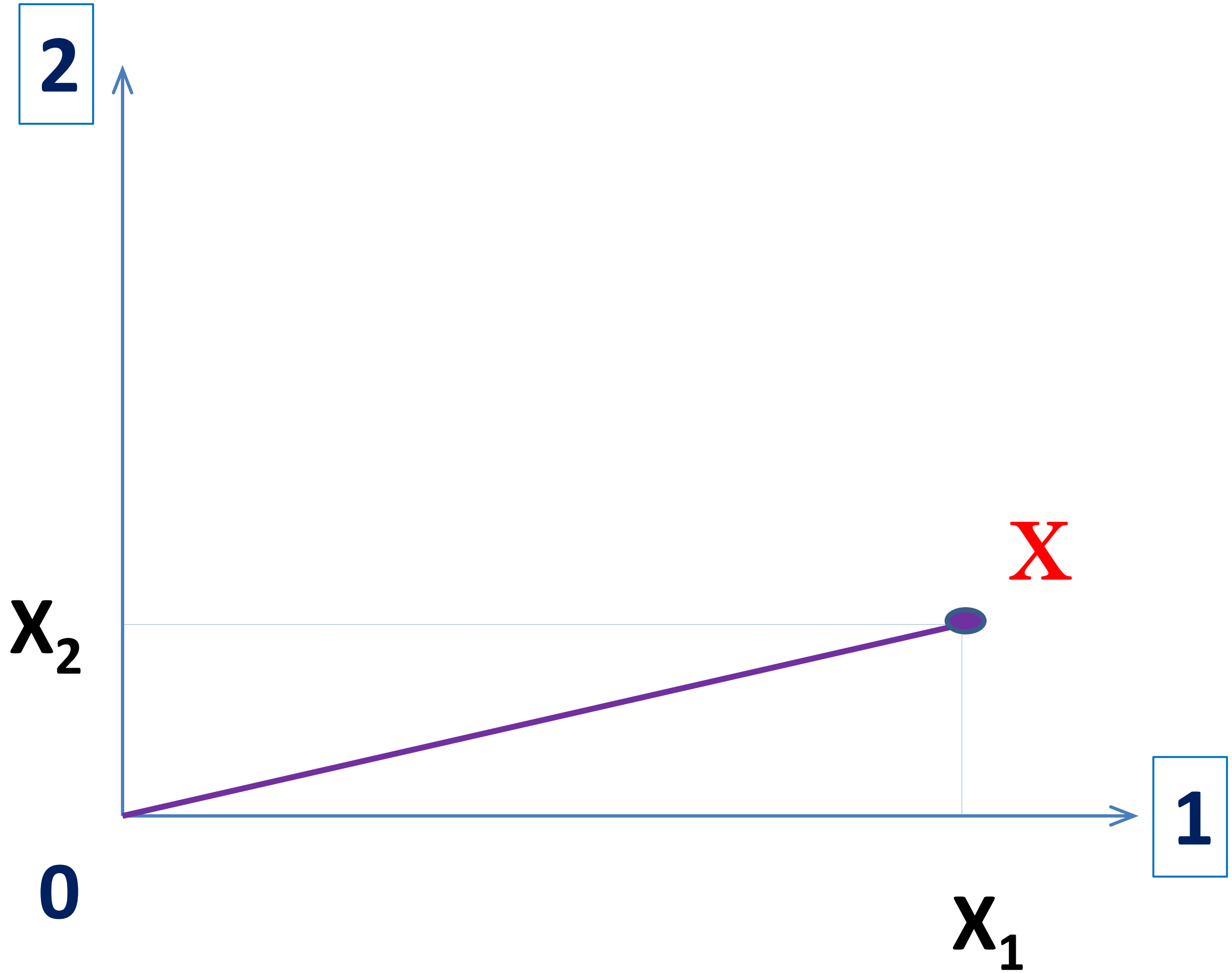
a problem in Geometry



representation of (all) the values of a (centred) variable

$$X_i = \left(x_i - \bar{x}\right)$$

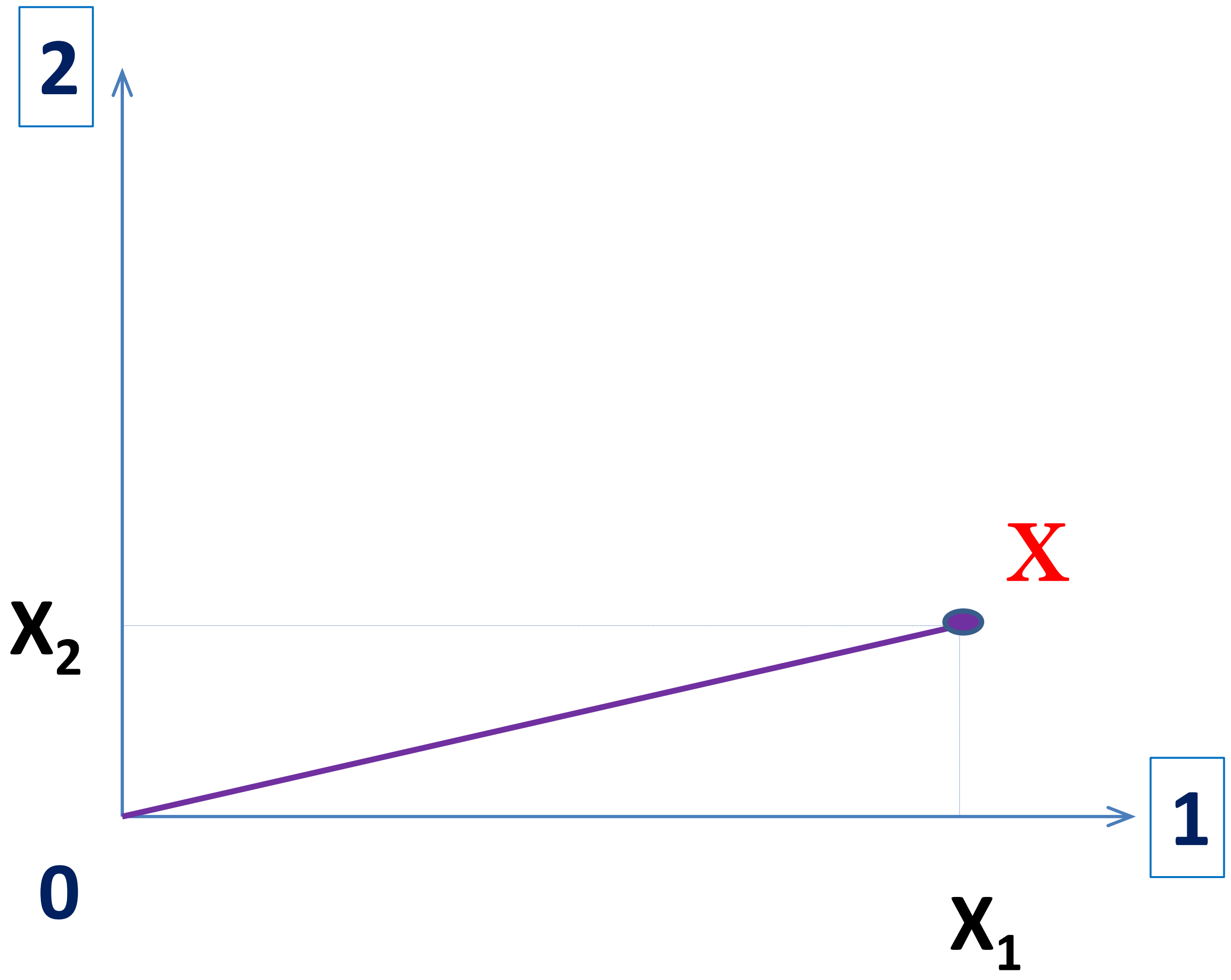
$$i = 1 \ldots n$$



representation of X with a regression coefficient: b.X fitted values

$$X_i = \left(x_i - \bar{x}\right)$$

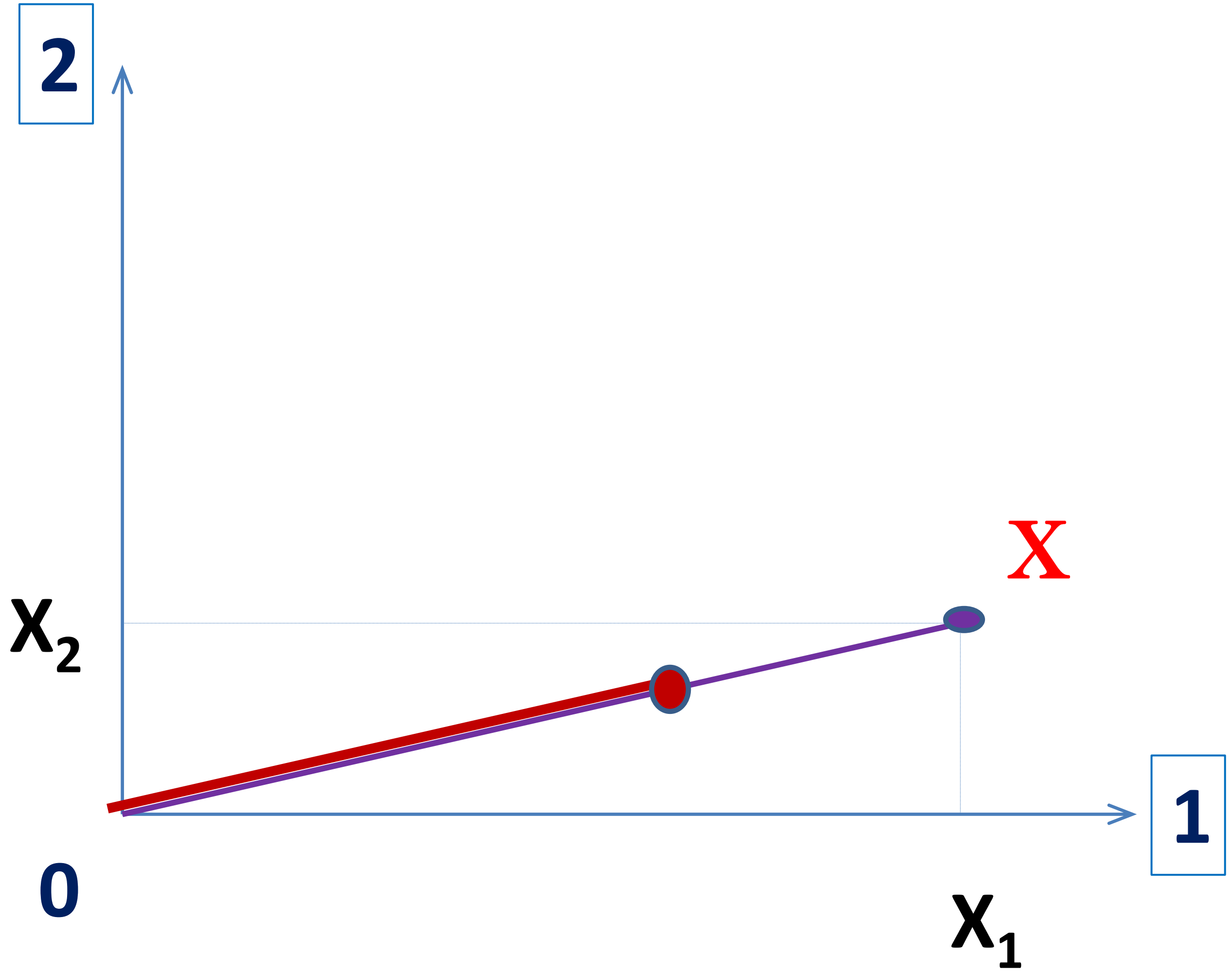
$$i = 1 \ldots n$$



representation of X with a regression coefficient: b.X b<1

$$X_i = \left(x_i - \bar{x}\right)$$

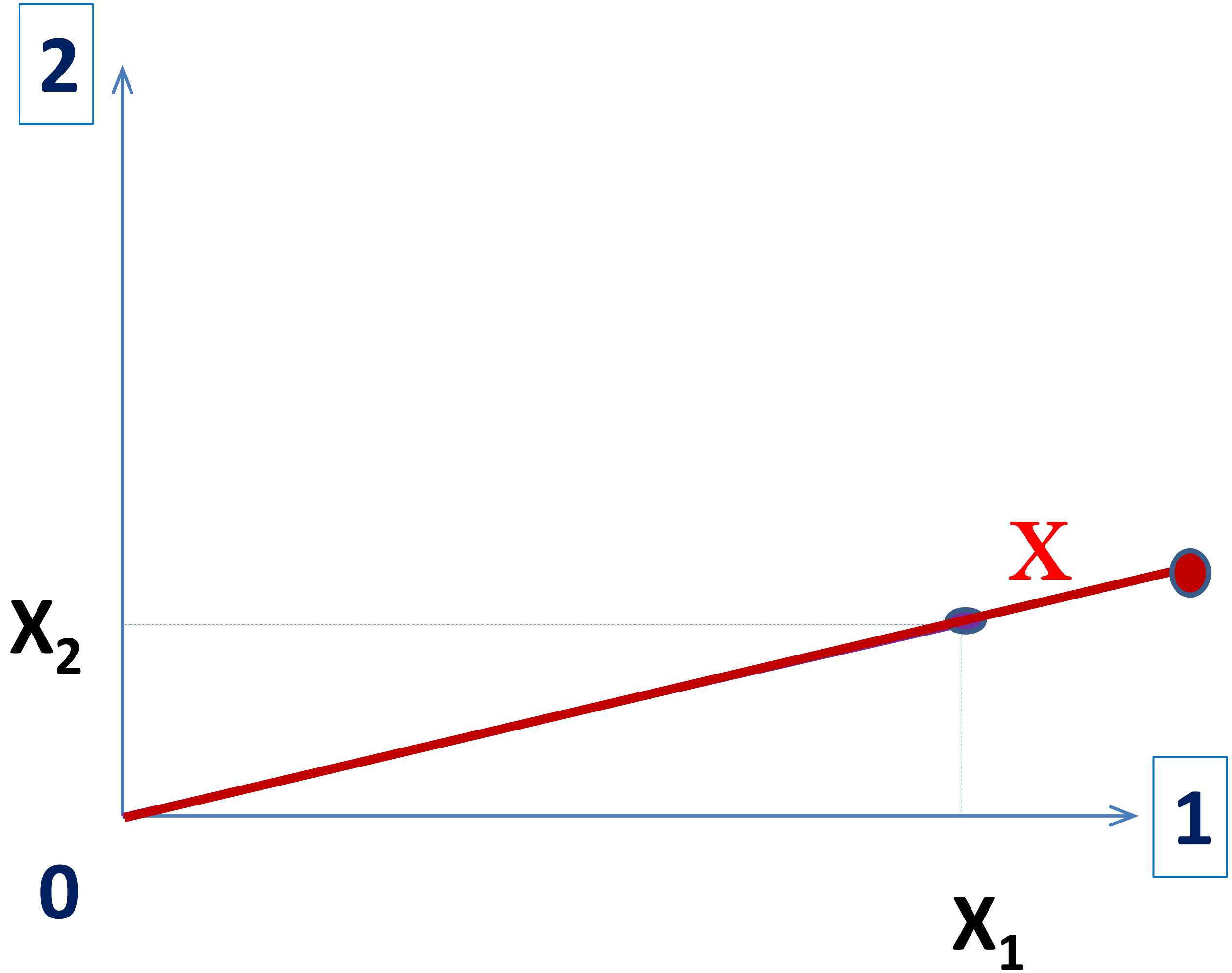
$$i = 1 \ldots n$$



representation of X with a regression coefficient: b.X b>1

$$X_i = \left(x_i - \bar{x}\right)$$

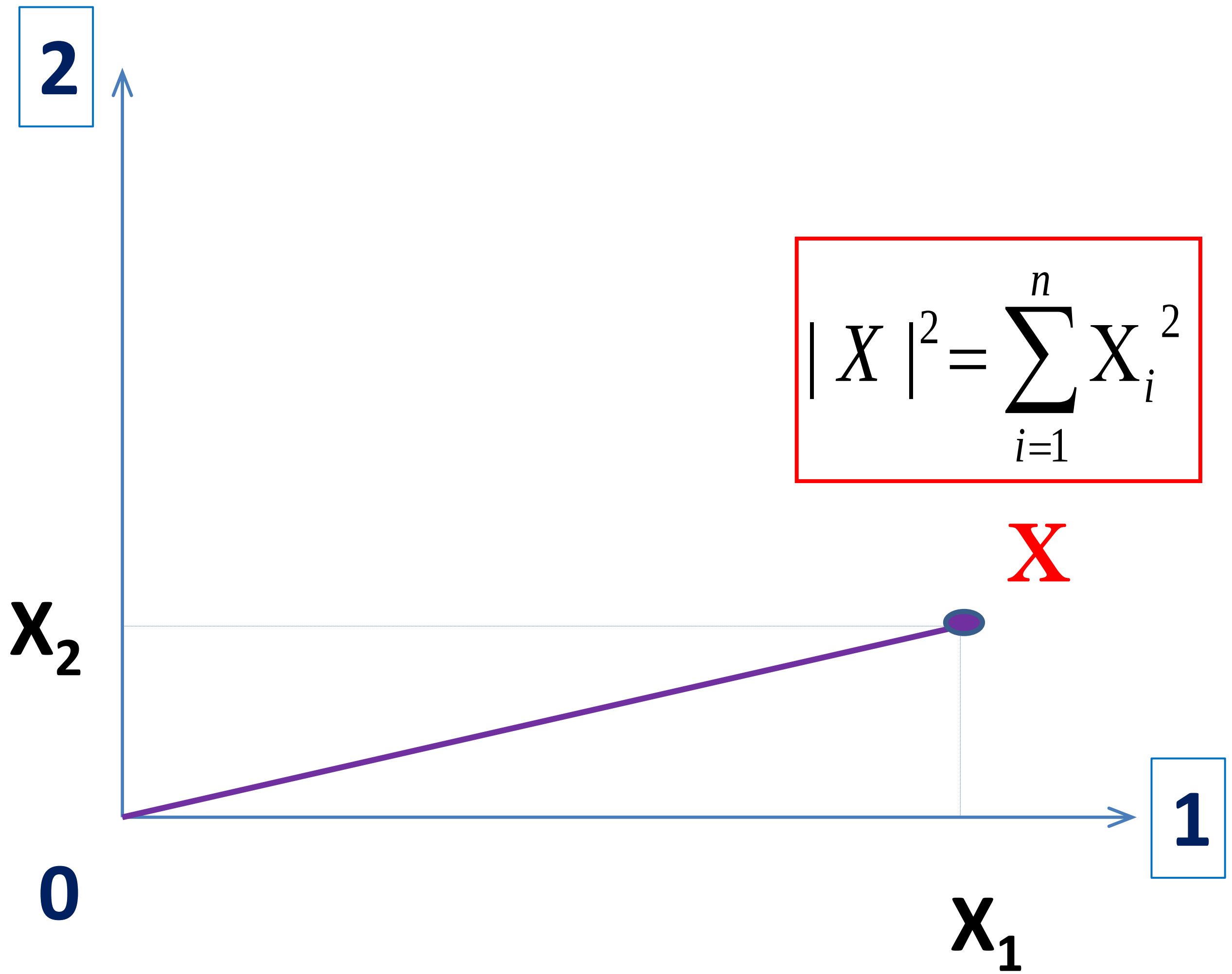
$$i = 1 \ldots n$$



Sum of Squares (variance) = Length^2

$$X_i = \left(x_i - \bar{x}\right)$$

$$i = 1 \ldots n$$



$$|X|^2 = \sum_{i=1}^{n} X_i^{\,2}$$

X and Y together…

$$X_i = \left(x_i - \bar{x}\right)$$

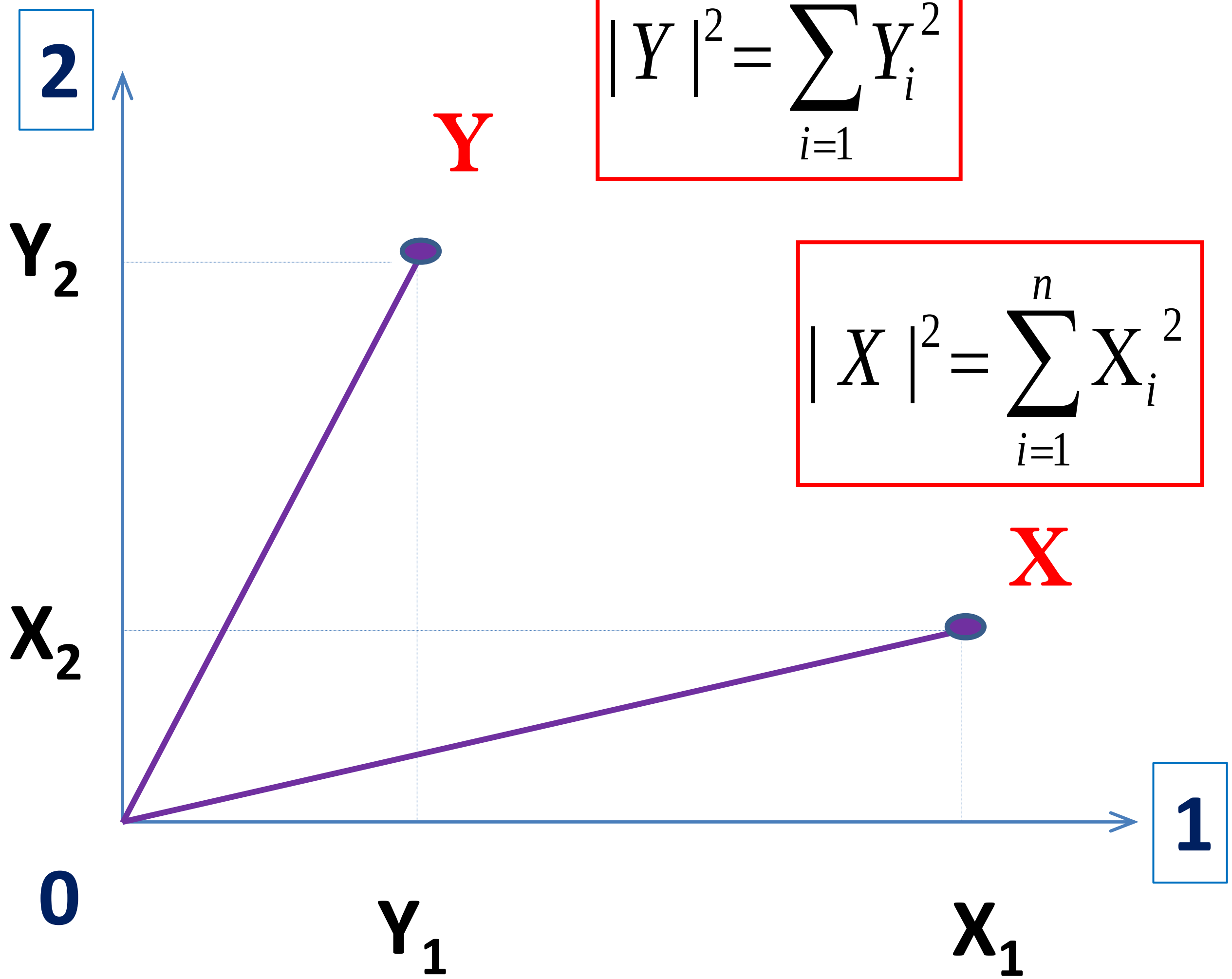
$$i = 1 \ldots n$$

$$Y_i = \left(y_i - \bar{y}\right)$$

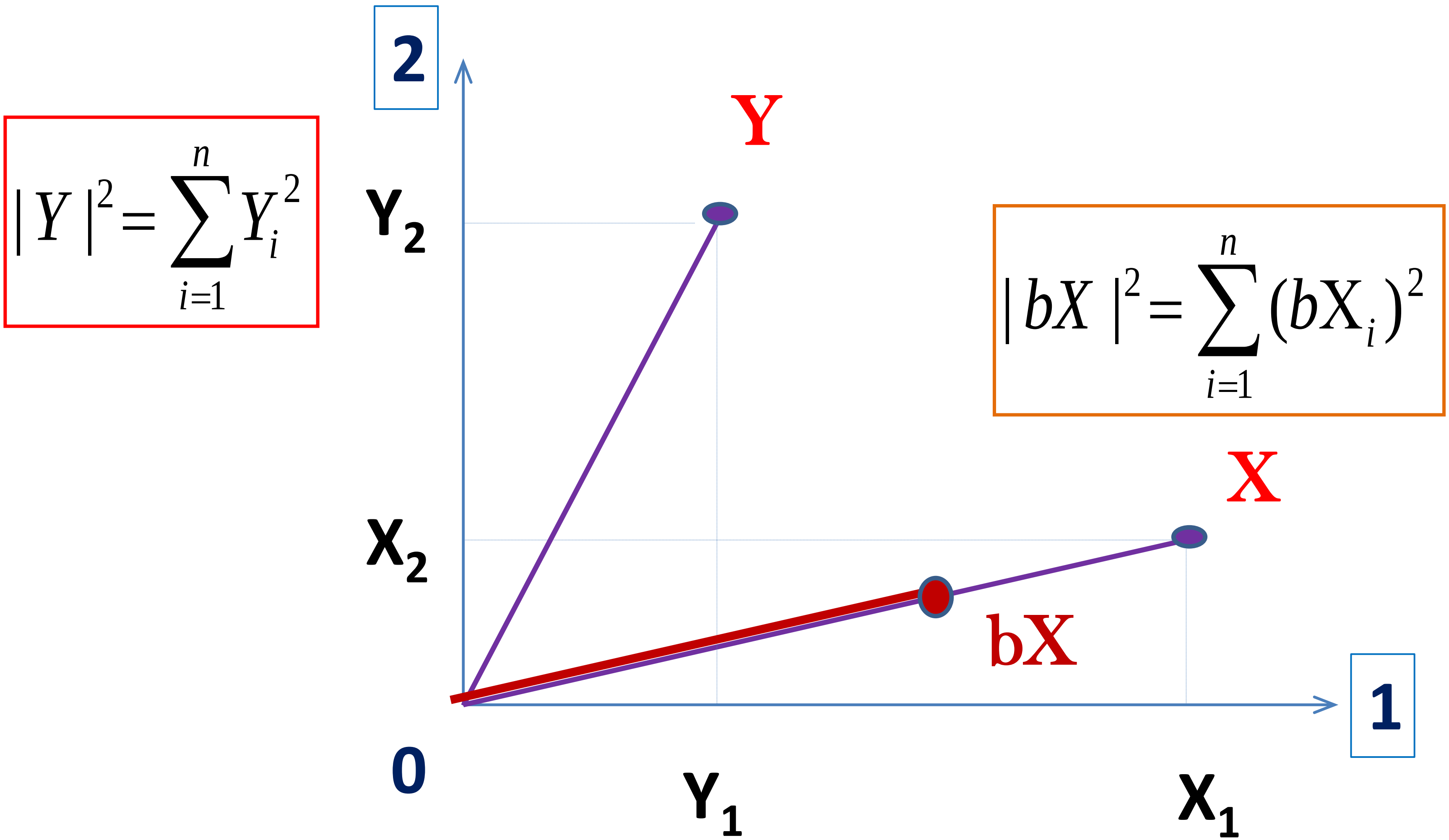
$$i = 1 \ldots n$$

$$|Y|^2 = \sum_{i=1}^{n} Y_i^2$$

$$|X|^2 = \sum_{i=1}^{n} X_i^2$$



X and Y together, with fitted values…

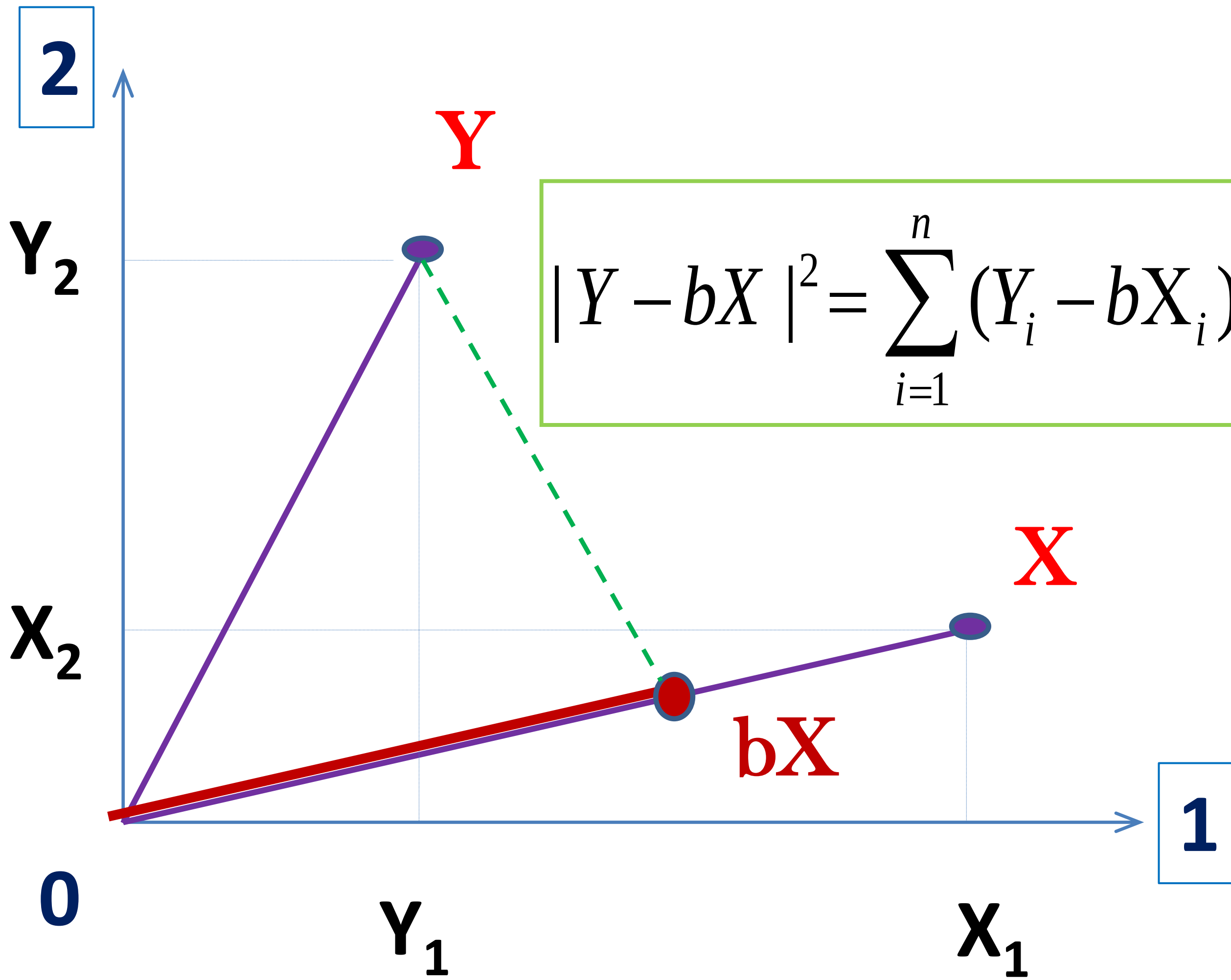


X and Y together, with fitted values, and residuals…

$$|Y|^2 = \sum_{i=1}^{n} Y_i^2$$

$$|bX|^2 = \sum_{i=1}^{n} (bX_i)^2$$

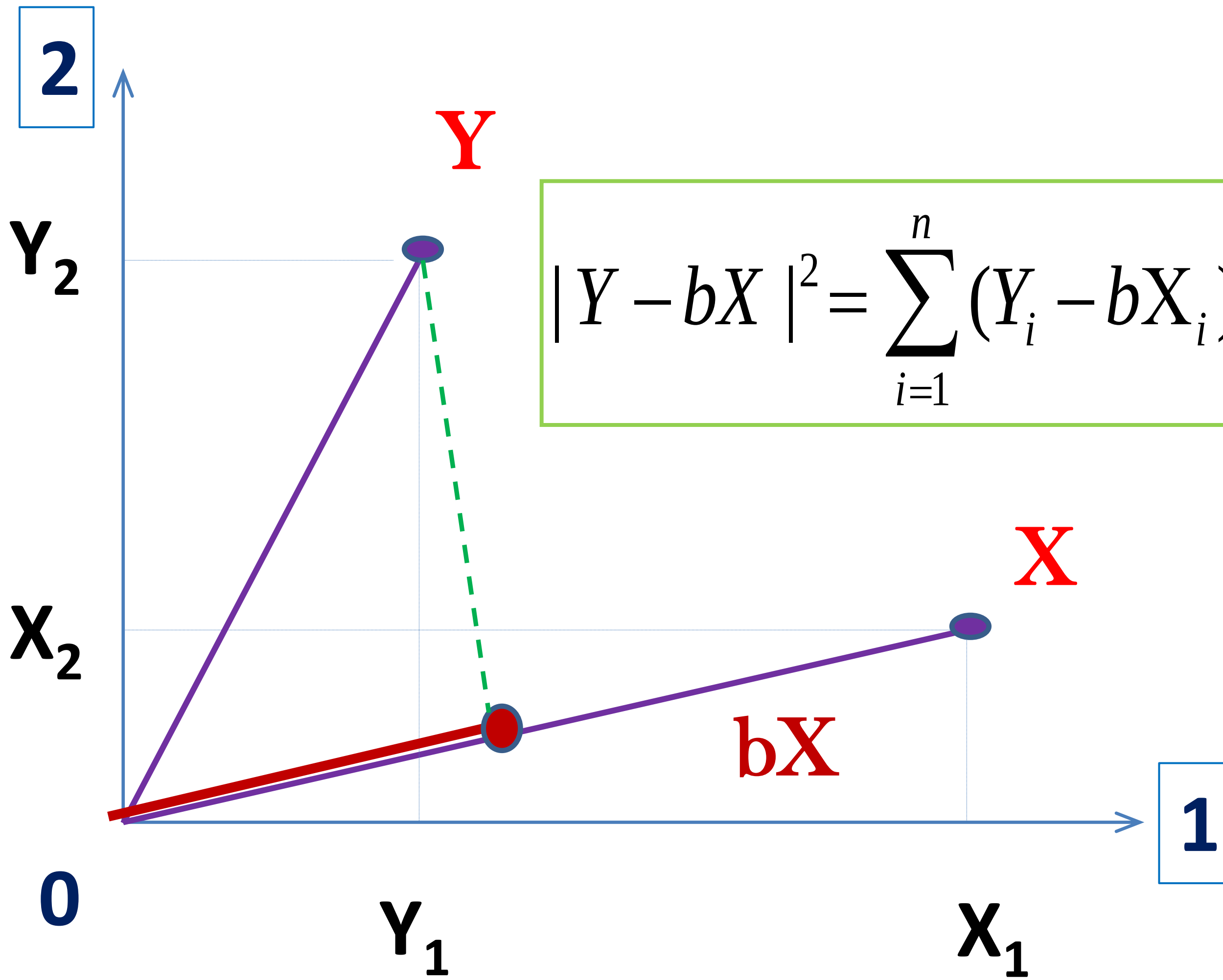
$$|Y - bX|^2 = \sum_{i=1}^{n} (Y_i - bX_i)^2$$



X and Y together, with fitted values, and residuals...

$$|Y|^2 = \sum_{i=1}^{n} Y_i^2$$

$$|bX|^2 = \sum_{i=1}^{n} (bX_i)^2$$

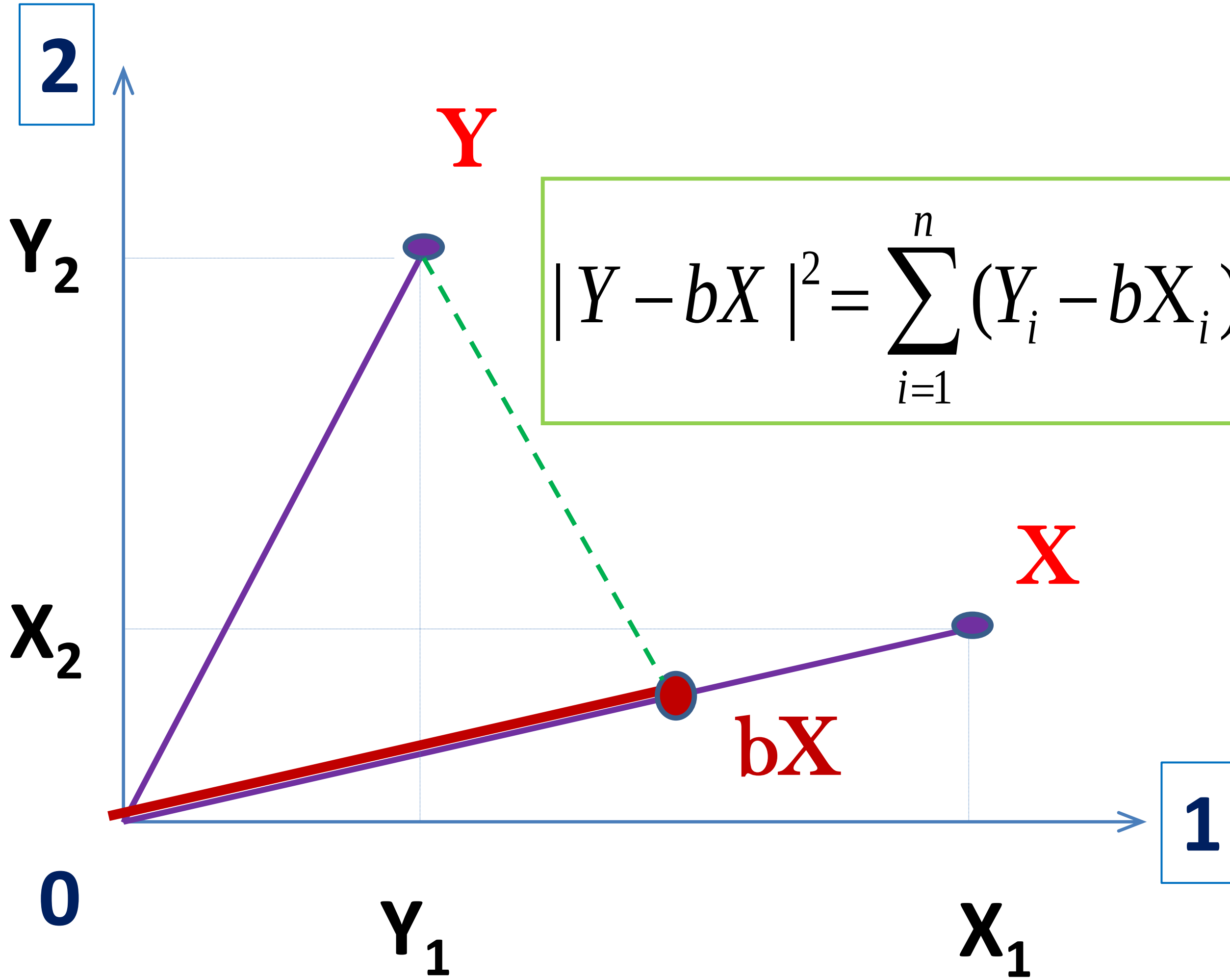


$$|Y - bX|^2 = \sum_{i=1}^{n} (Y_i - bX_i)^2$$

X and Y together, with fitted values, and residuals…

$$|Y|^2 = \sum_{i=1}^{n} Y_i^2$$

$$|bX|^2 = \sum_{i=1}^{n} (bX_i)^2$$



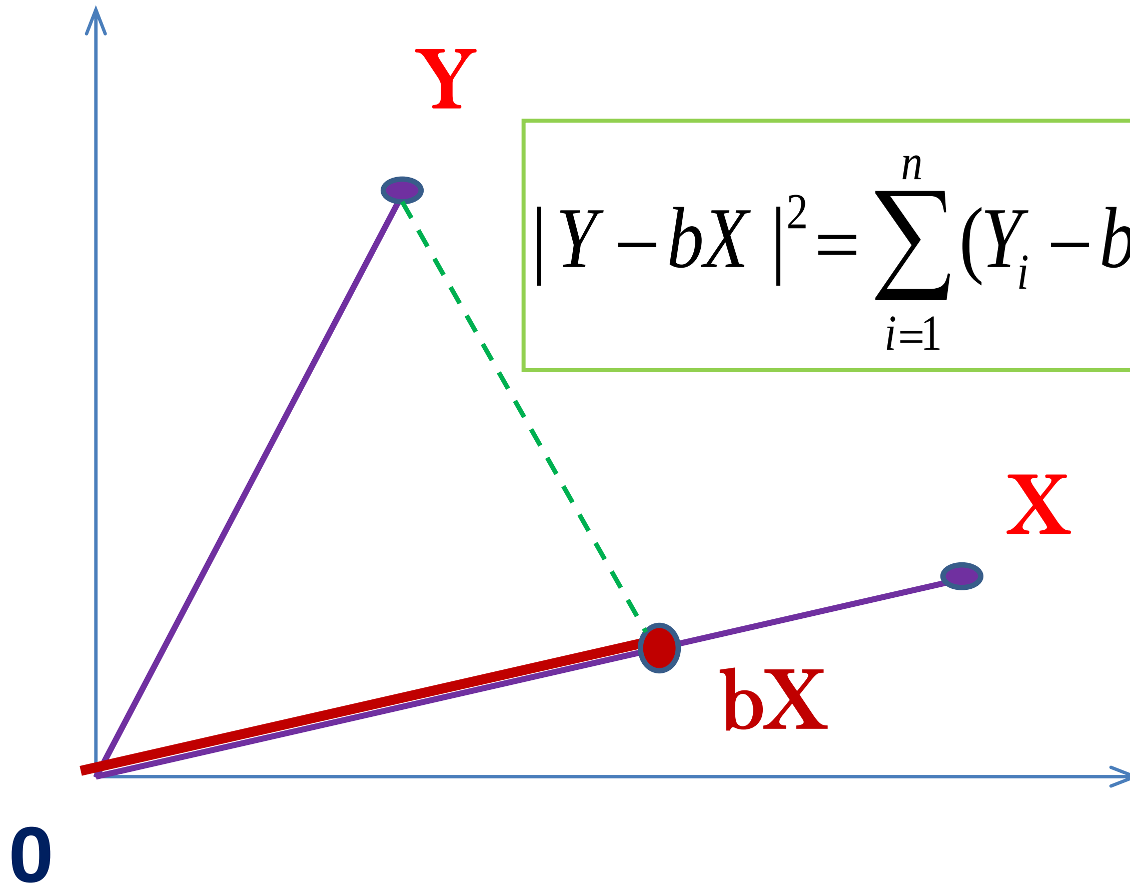
$$|Y - bX|^2 = \sum_{i=1}^{n} (Y_i - bX_i)^2$$

…on their own 2-dimensional 'slice'

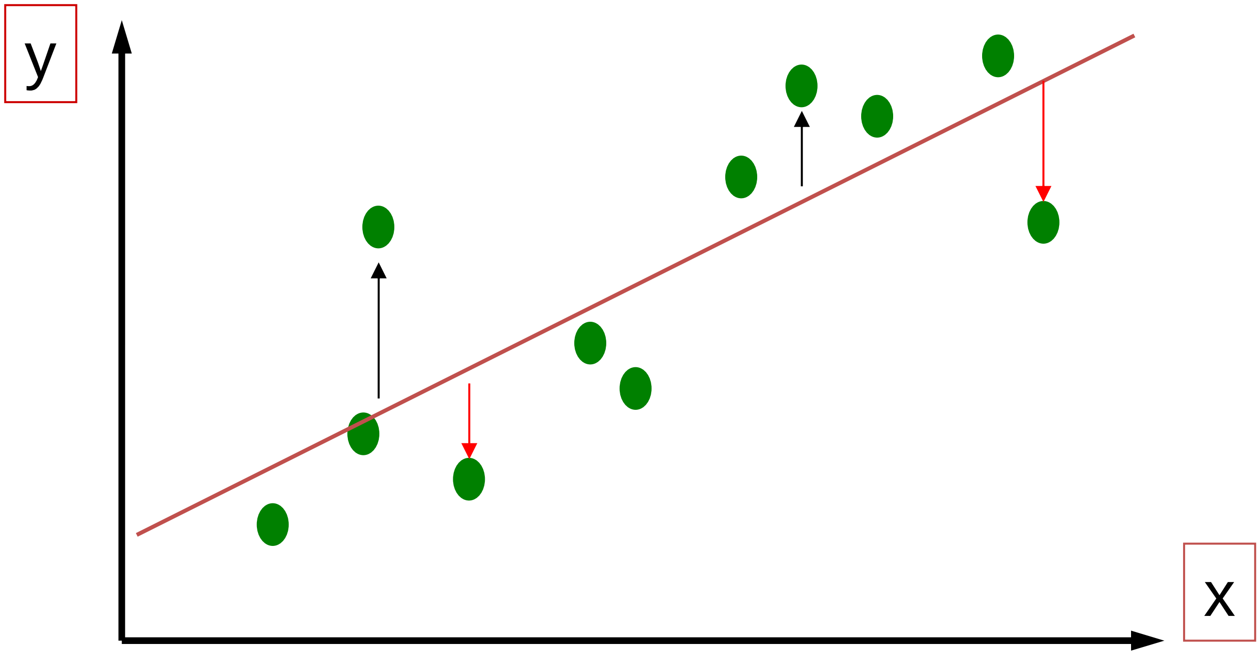
$$|Y|^2 = \sum_{i=1}^{n} Y_i^2$$

$$|bX|^2 = \sum_{i=1}^{n} (bX_i)^2$$

$$|Y - bX|^2 = \sum_{i=1}^{n} (Y_i - bX_i)^2$$



least-squares linear regression



regression line placed to minimize

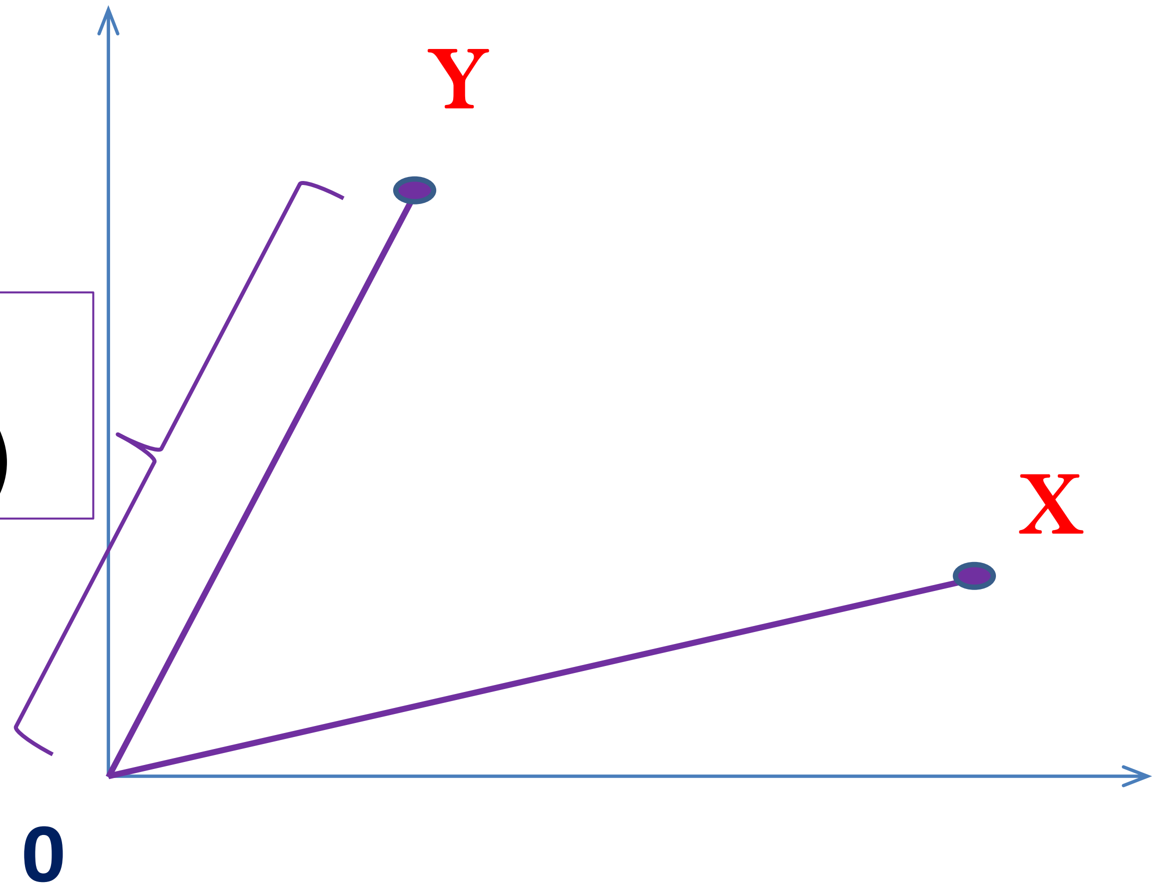
sum of squares of the ↑↓ differences

between y and the fitted values

least-squares fitting

$$|Y|^2 = \sum_{i=1}^{n} Y_i^2$$

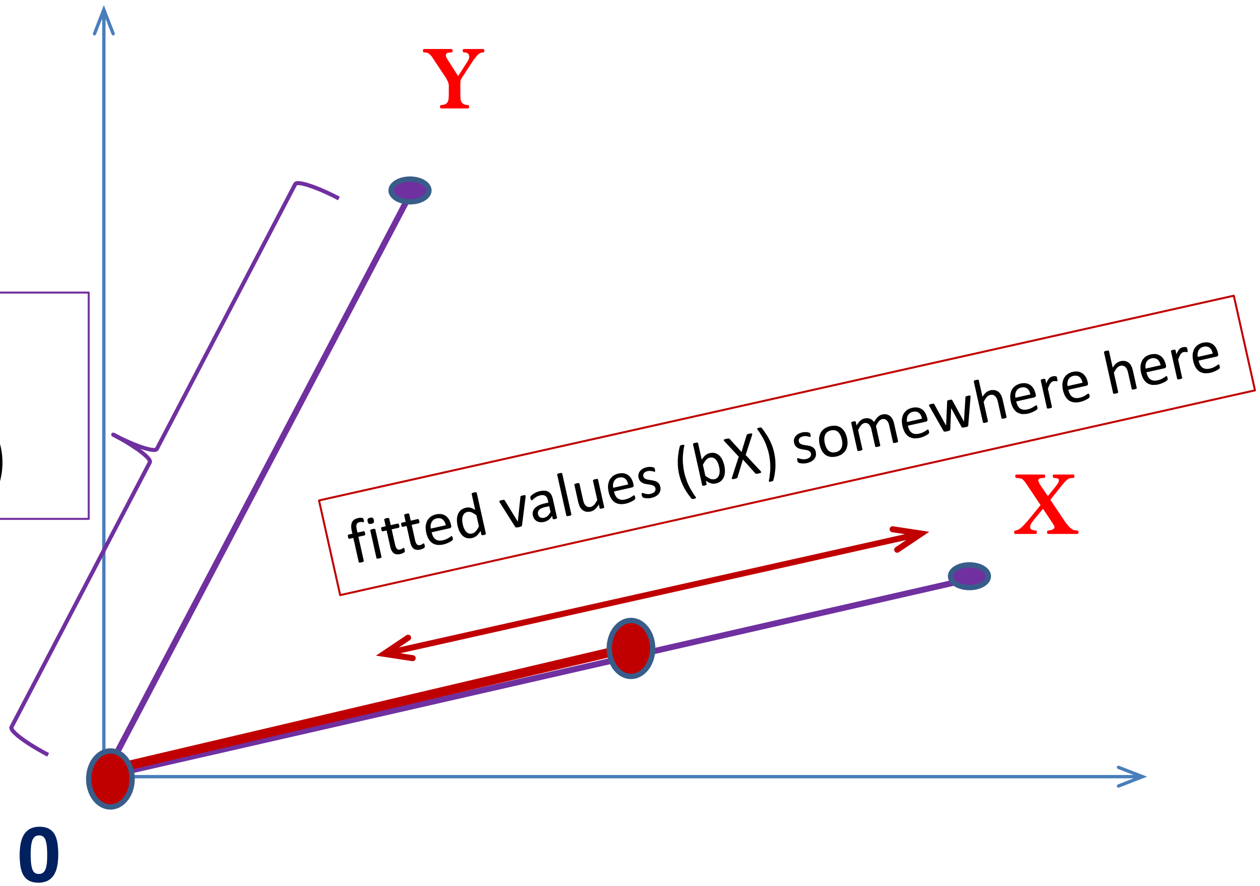
length^2 =
total SS (for Y)



least-squares fitting

$$|Y|^2 = \sum_{i=1}^{n} Y_i^2$$

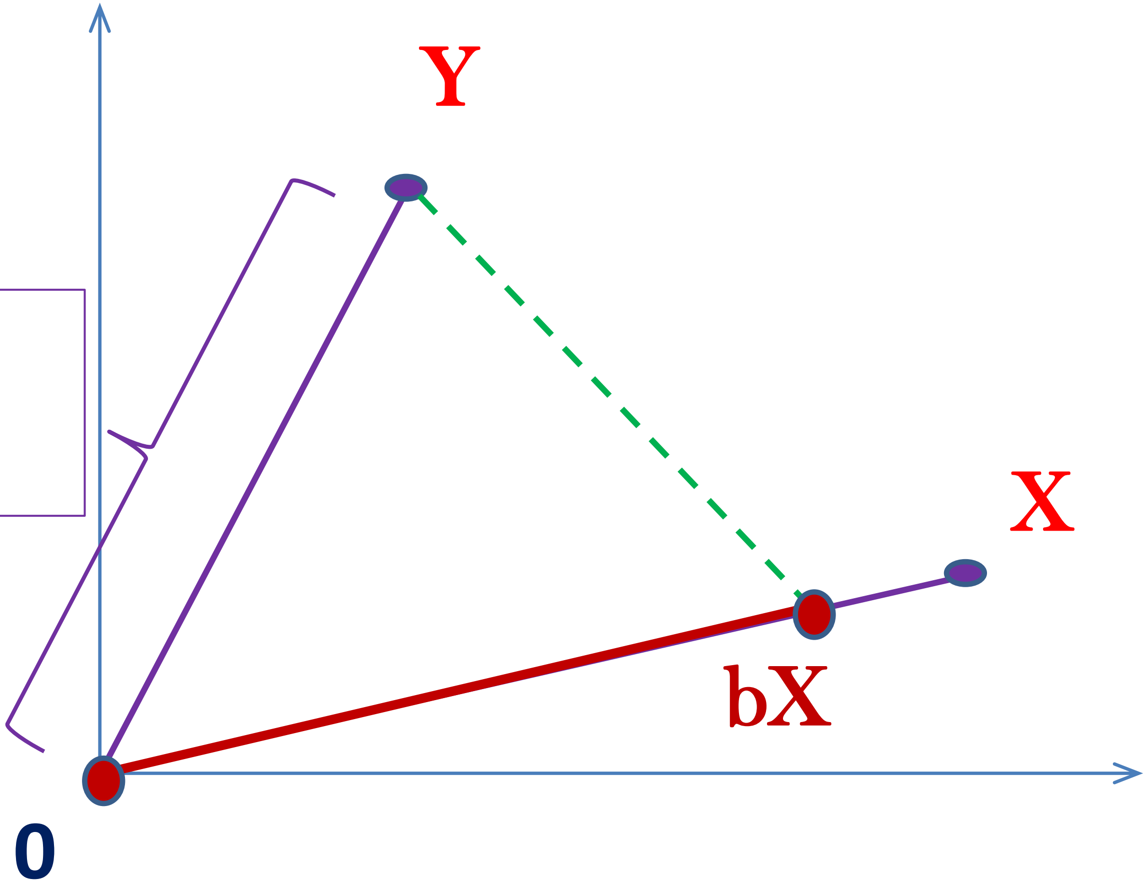
length^2 =
total SS (for Y)



least-squares fitting

$$|Y|^2 = \sum_{i=1}^{n} Y_i^2$$

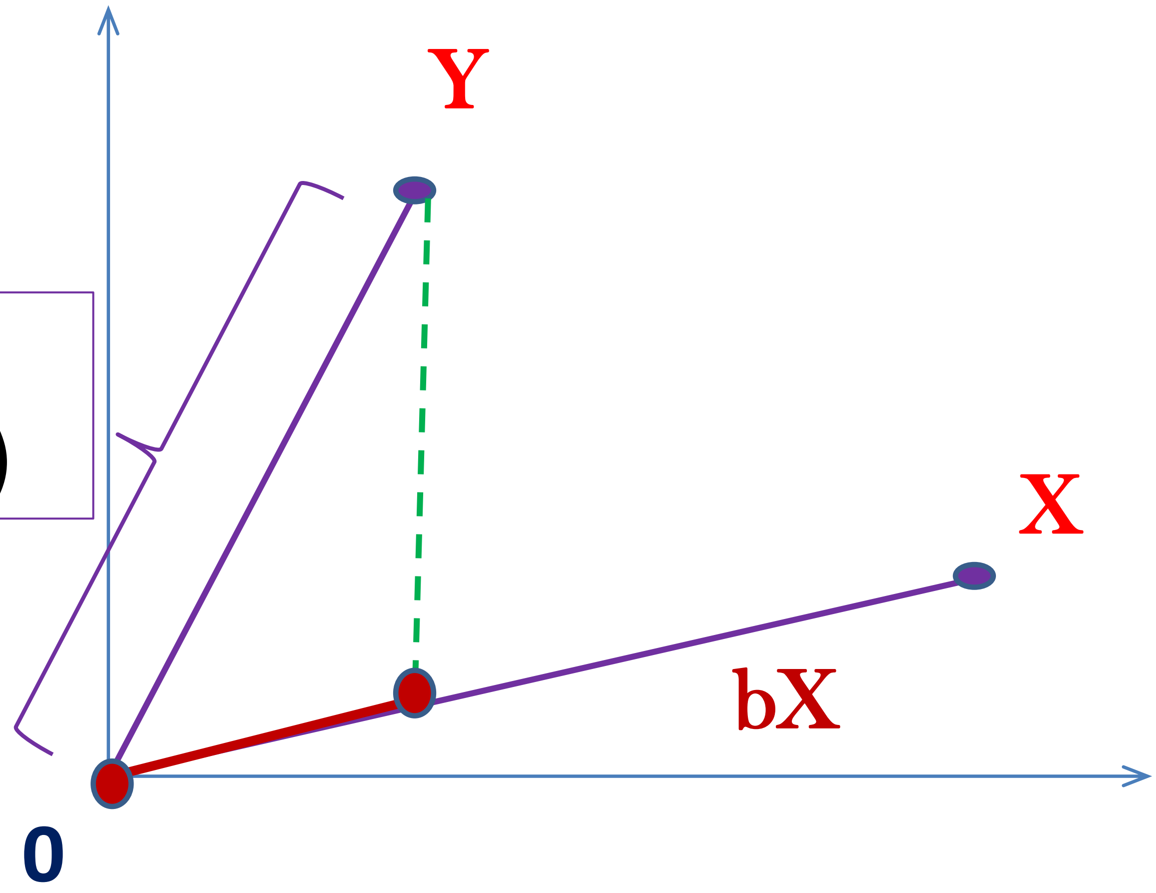
length^2 =
total SS (for Y)



least-squares fitting

$$|Y|^2 = \sum_{i=1}^{n} Y_i^2$$

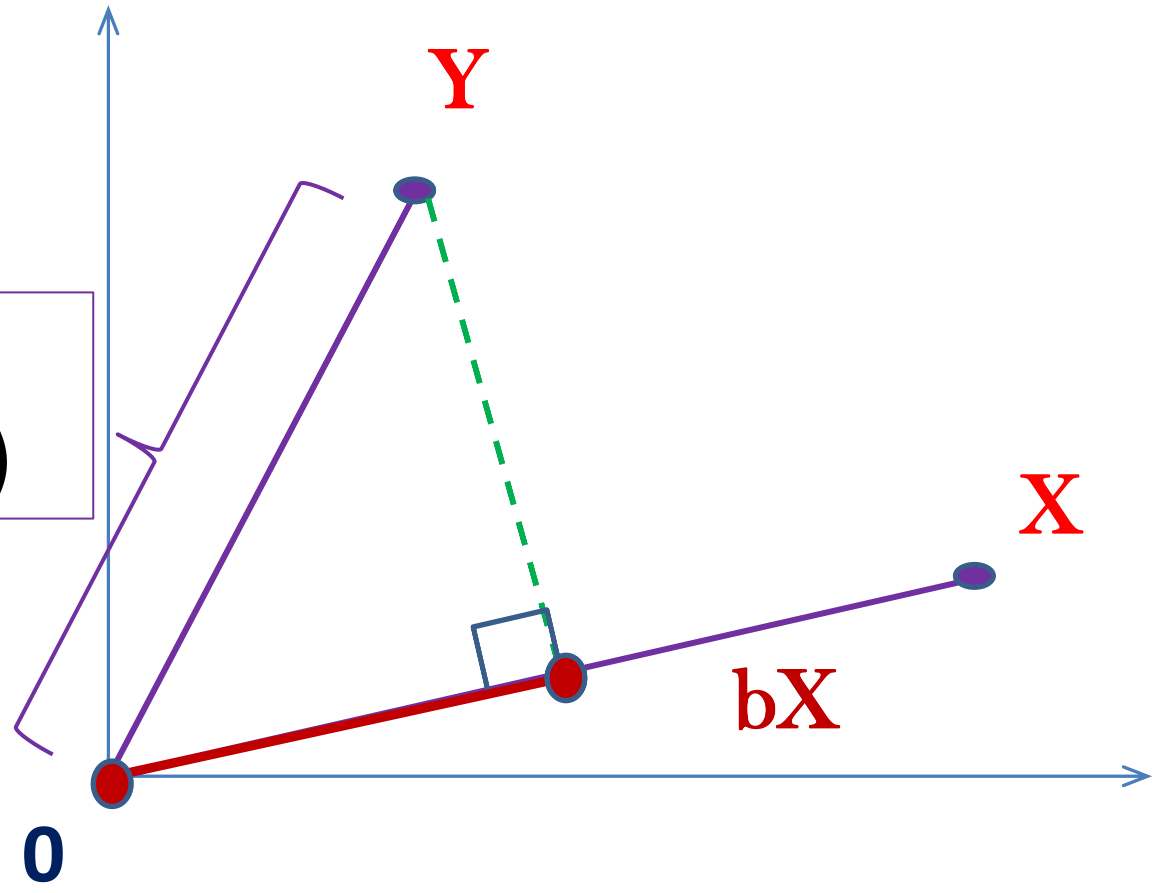
length^2 =
total SS (for Y)



least-squares fitting

$$|Y|^2 = \sum_{i=1}^{n} Y_i^2$$

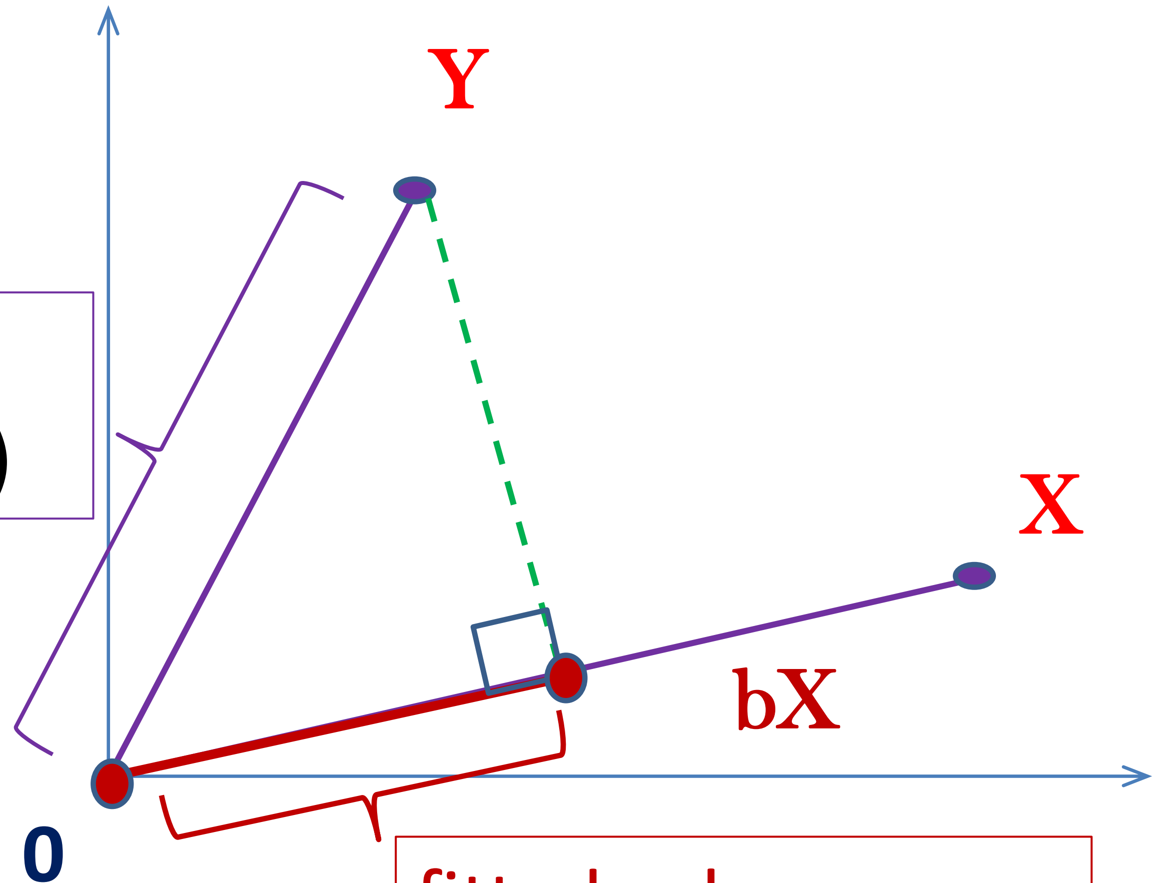
length² =
total SS (for Y)



least-squares fitting

$$|Y|^2 = \sum_{i=1}^{n} Y_i^2$$

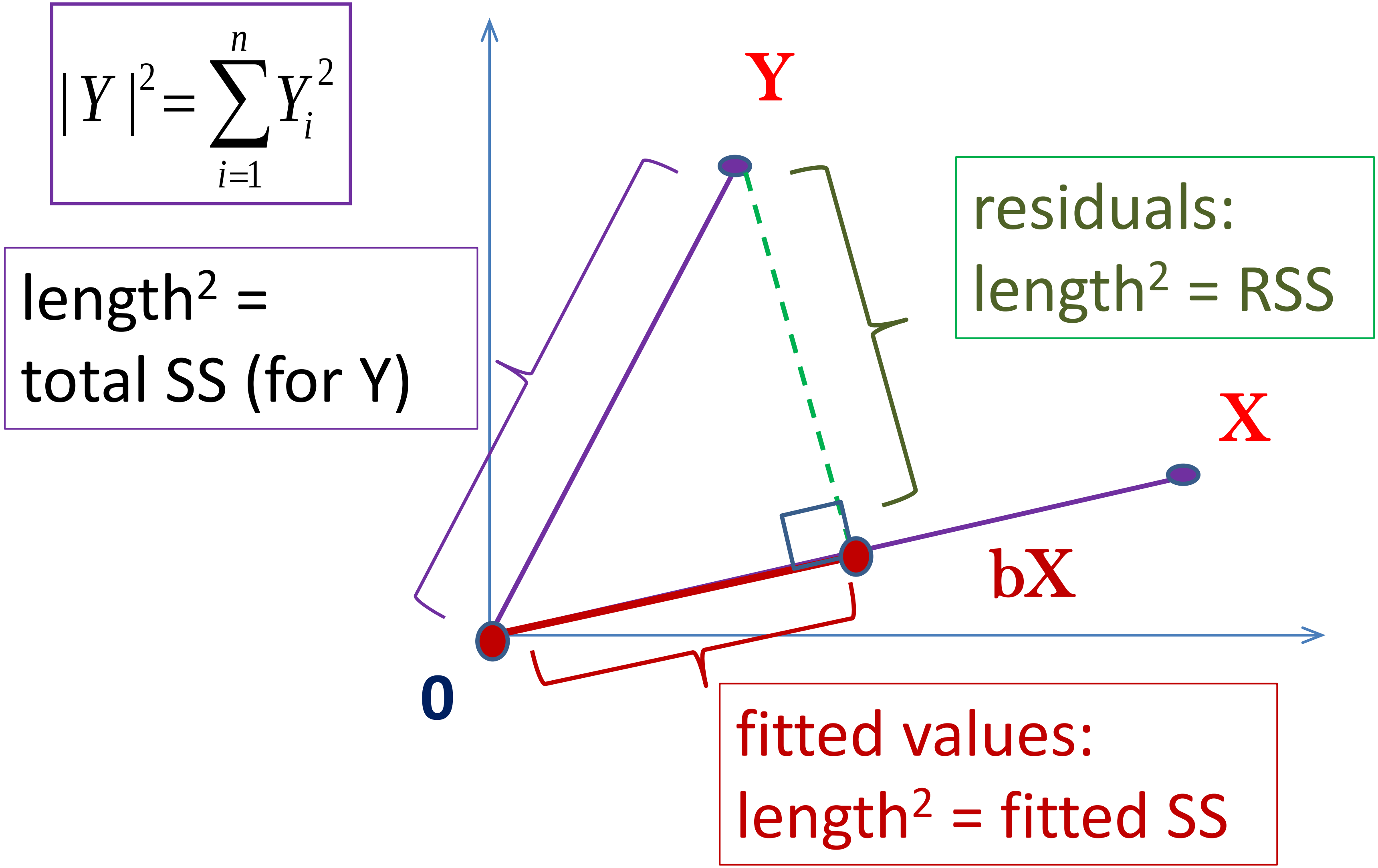
length^2 =
total SS (for Y)



fitted values:
length^2 = fitted SS

least-squares fitting

$$|Y|^2 = \sum_{i=1}^{n} Y_i^2$$



least-squares fitting

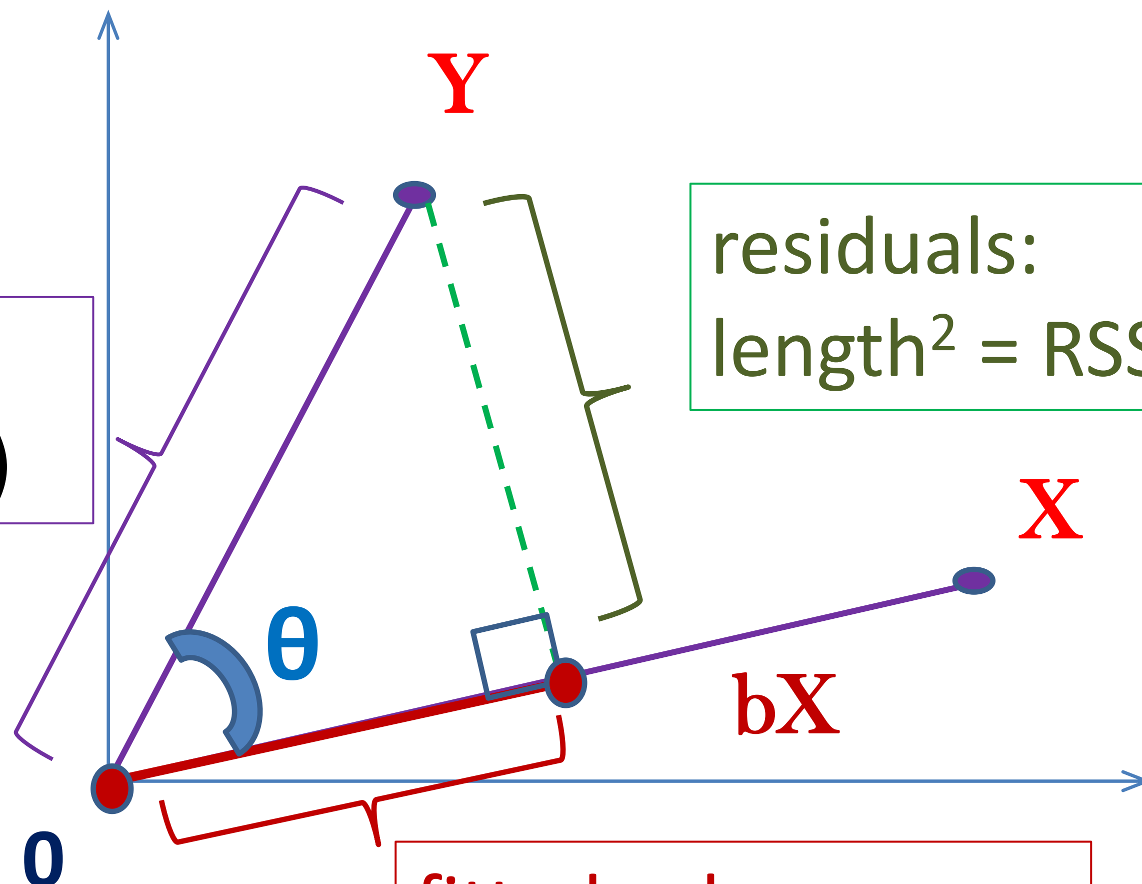
$$|Y|^2 = \sum_{i=1}^{n} Y_i^2$$

length^2 =
total SS (for Y)

" %var. acct. for " = ρ^2

= FSS/TSS

= $\cos^2(\theta)$

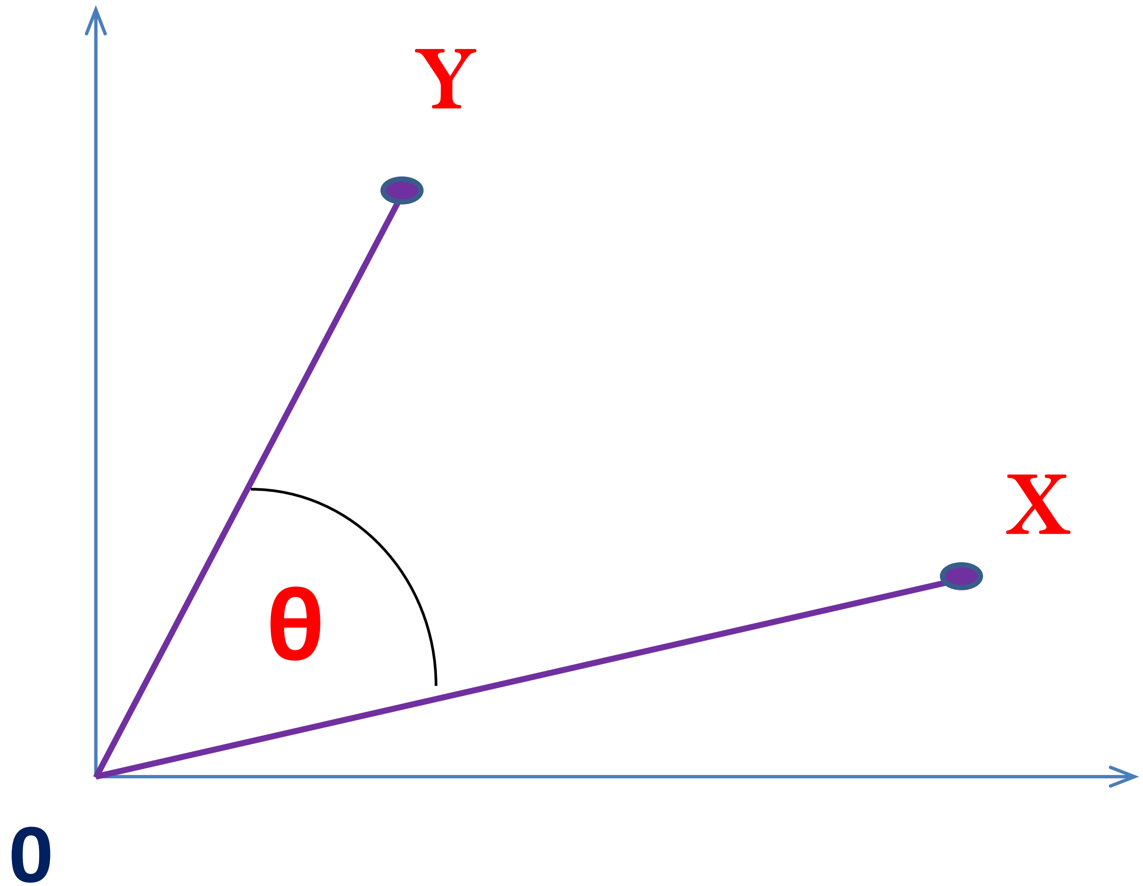


residuals:
length^2 = RSS

fitted values:
length^2 = fitted SS

Correlation and Angle

$\rho(X,Y) = \cos(\theta)$



Correlation (variables centred)

$$\rho = \frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2 \sum_{i=1}^{n} Y_i^2}}$$

$$X_i = \left(x_i - \bar{x}\right)$$

$$Y_i = \left(y_i - \bar{y}\right)$$

Correlation and Angle

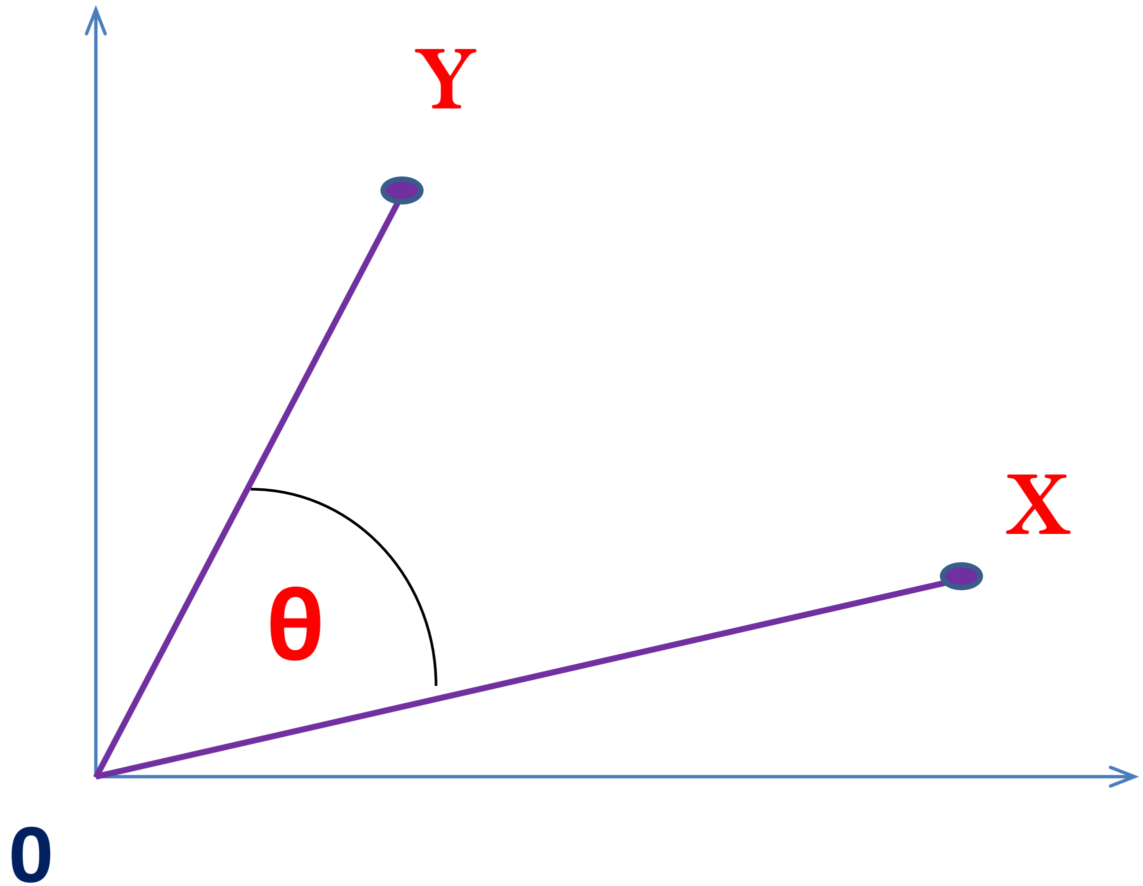
$\rho(X,Y) = \cos(\theta)$

$\theta = 0$

$\cos(\theta) = 1$

$\theta = 90^o$

$\cos(\theta) = 0$

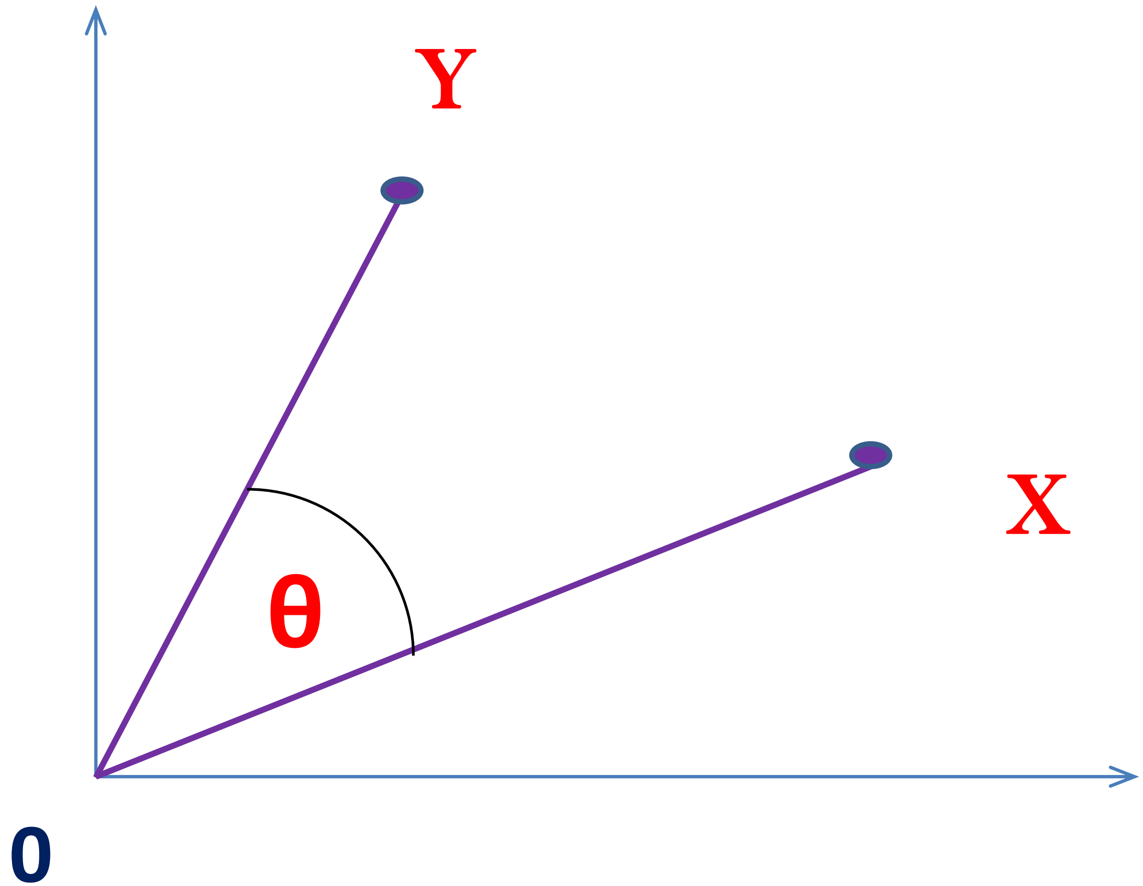


a Problem in Correlation

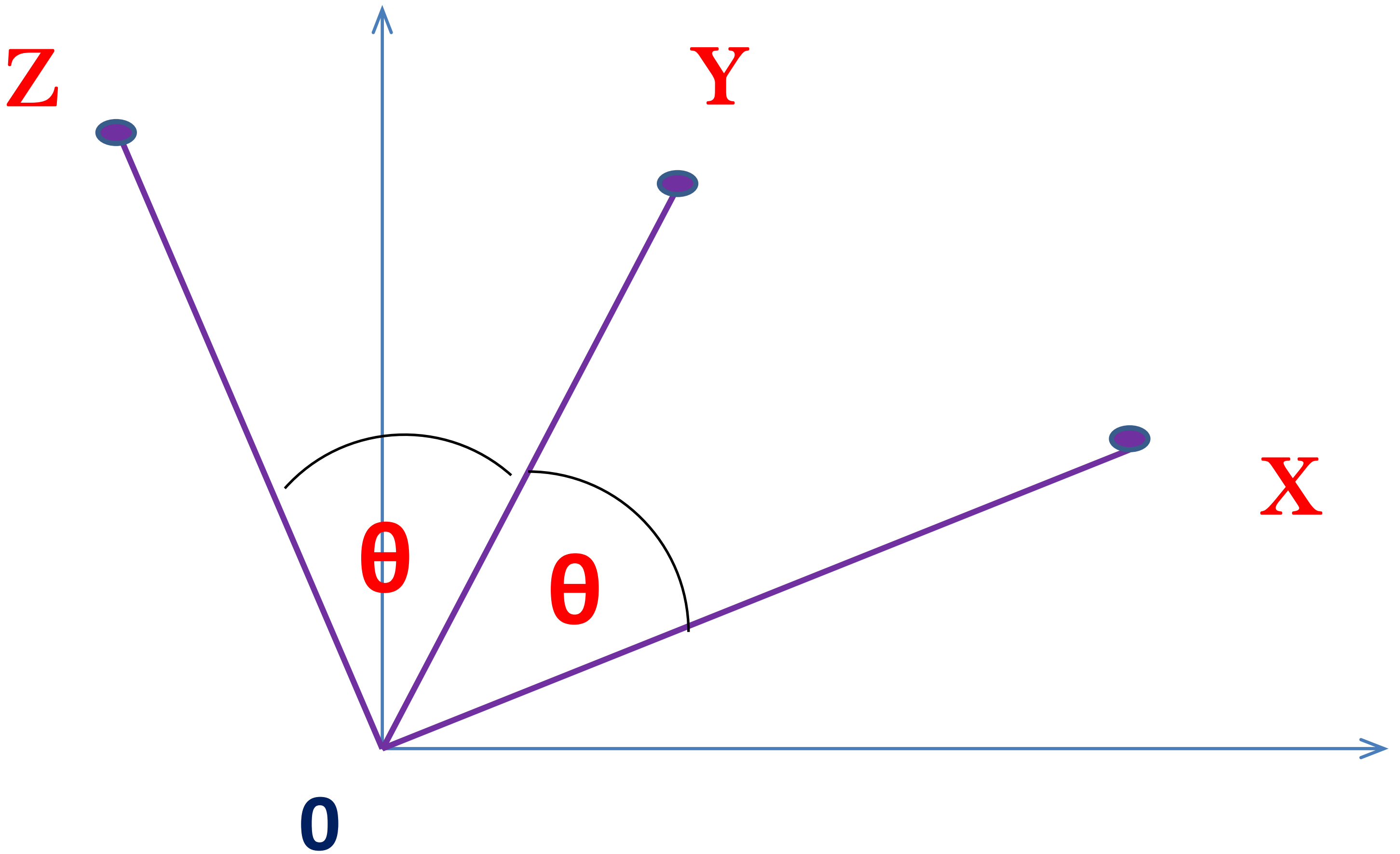
- $\rho(X,Y)$ = correlation between X and Y
- (suppose) $\rho(X,Y) = 0.7$ and $\rho(Y,Z) = 0.7$
- Q: what is the *least* possible value for

$$\rho(X,Z)$$

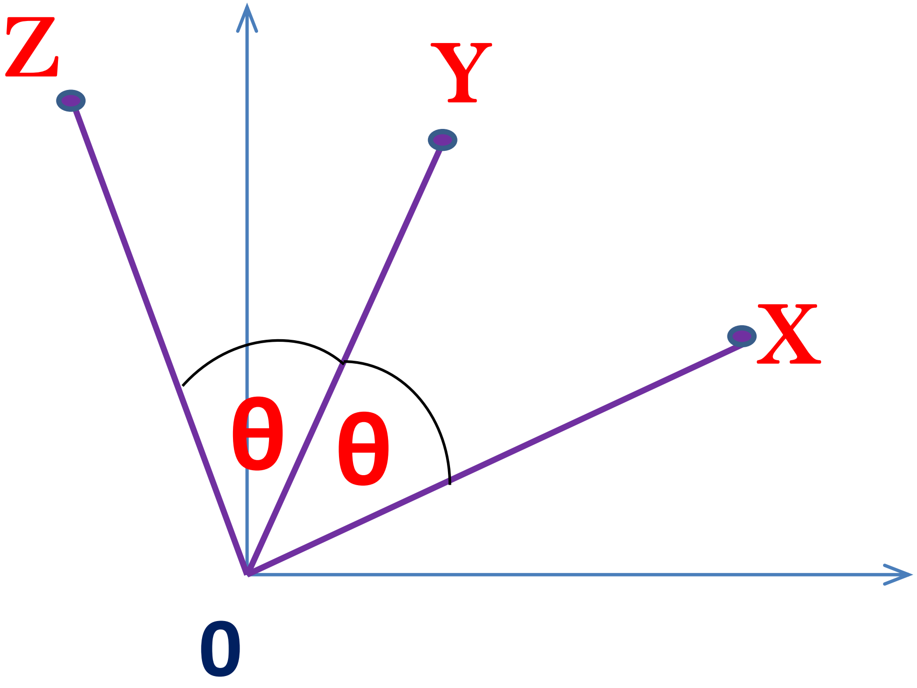
Representation



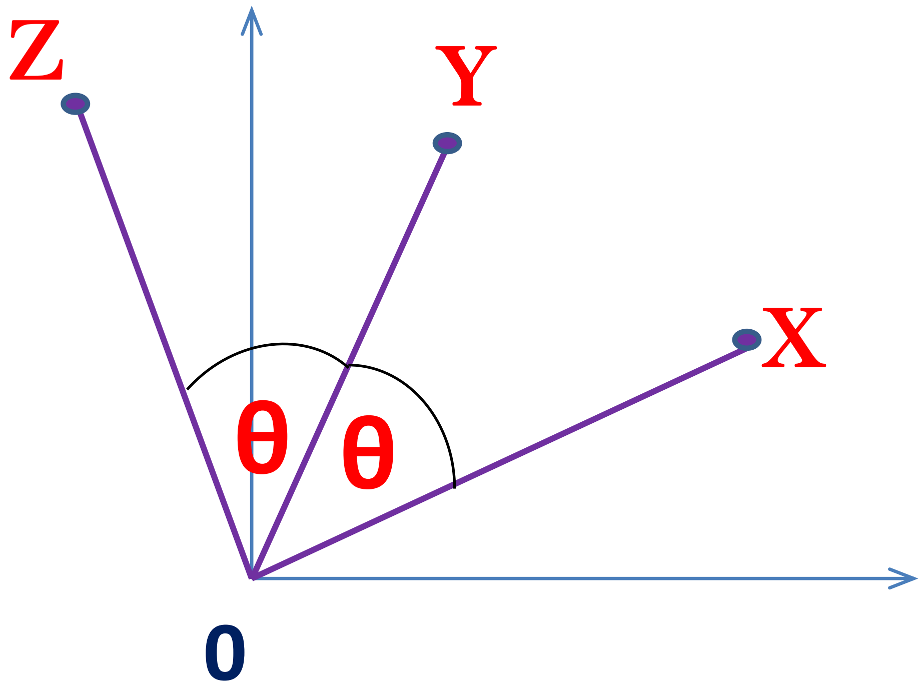
Geometric Solution



Geometric Solution



Formulae



$\rho(X,Y) = \cos(\theta) = 0.7$

$\rho(Y,Z) = \cos(\theta) = 0.7$

$\rho(X,Z) = \cos(2\theta)$

$\cos(2\theta) = 2\cos^2(\theta) - 1$

$\min[\rho(X,Z)] = 2\rho^2 - 1$