Module 4A
Supervised Machine Learning

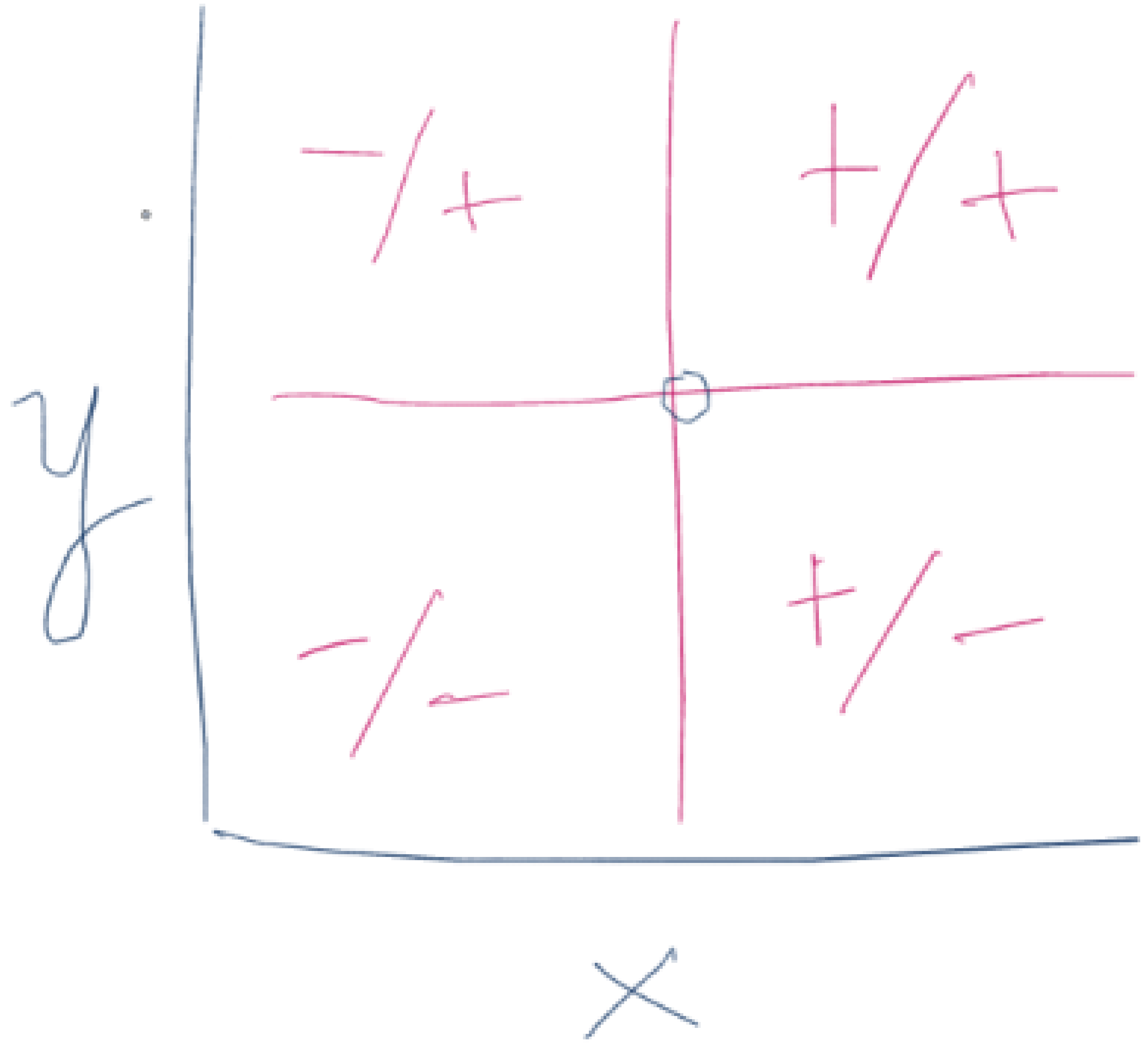
Different flavors of REGRESSION and General Linear Models

Review: Correlation:

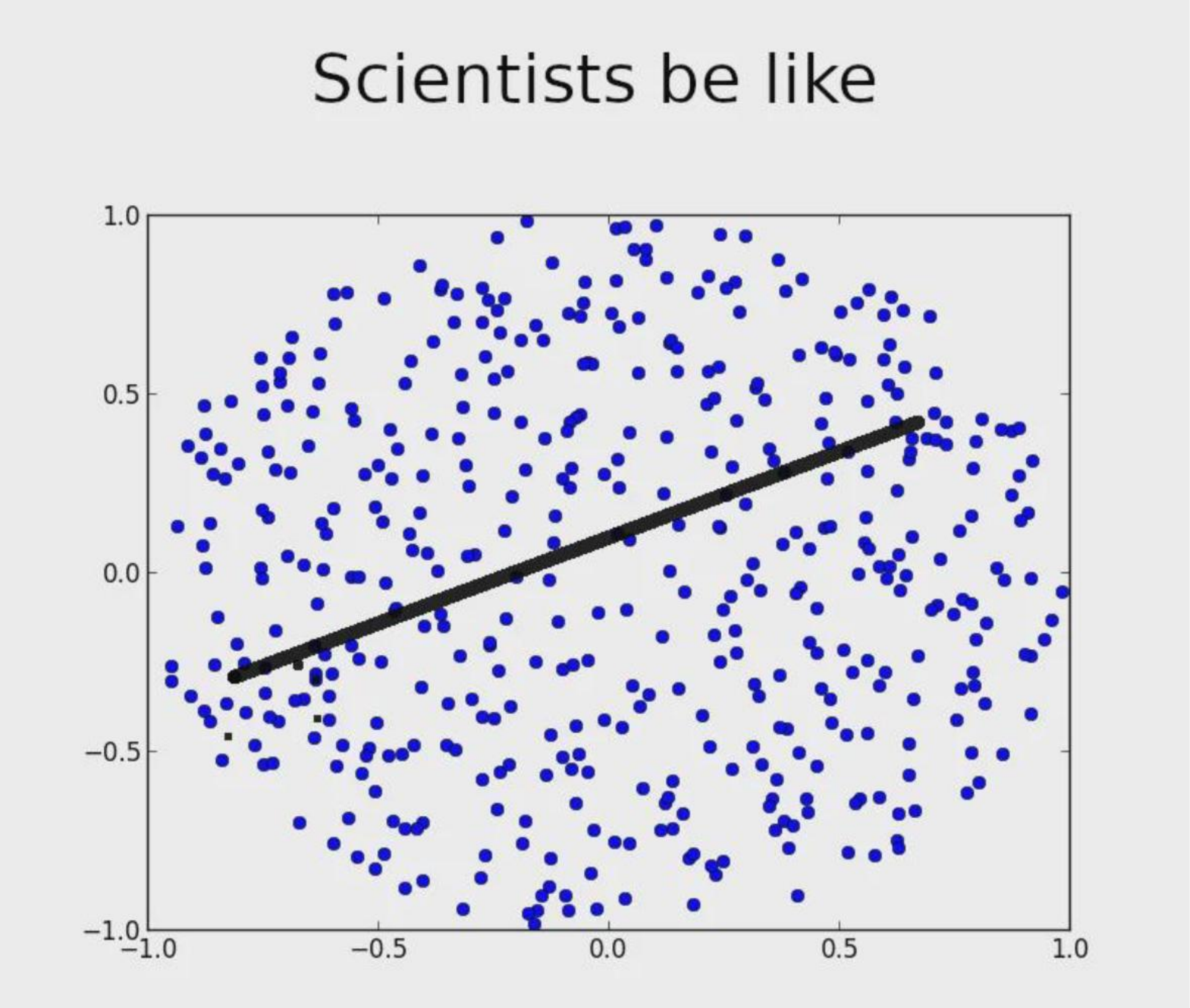
- Measures the amount/degree of linear association between two **numerical** variables

- Estimate the degree to which variables **covary**
    - With no attempt to interpret the causality of the association

Example: arm length and leg length covary together (individuals with longer arms often have longer legs) but they are influenced by other underlying variables **not** each other (longer legs do not cause longer arms)



-/+
+/+
0
-/-
+/-
y
x

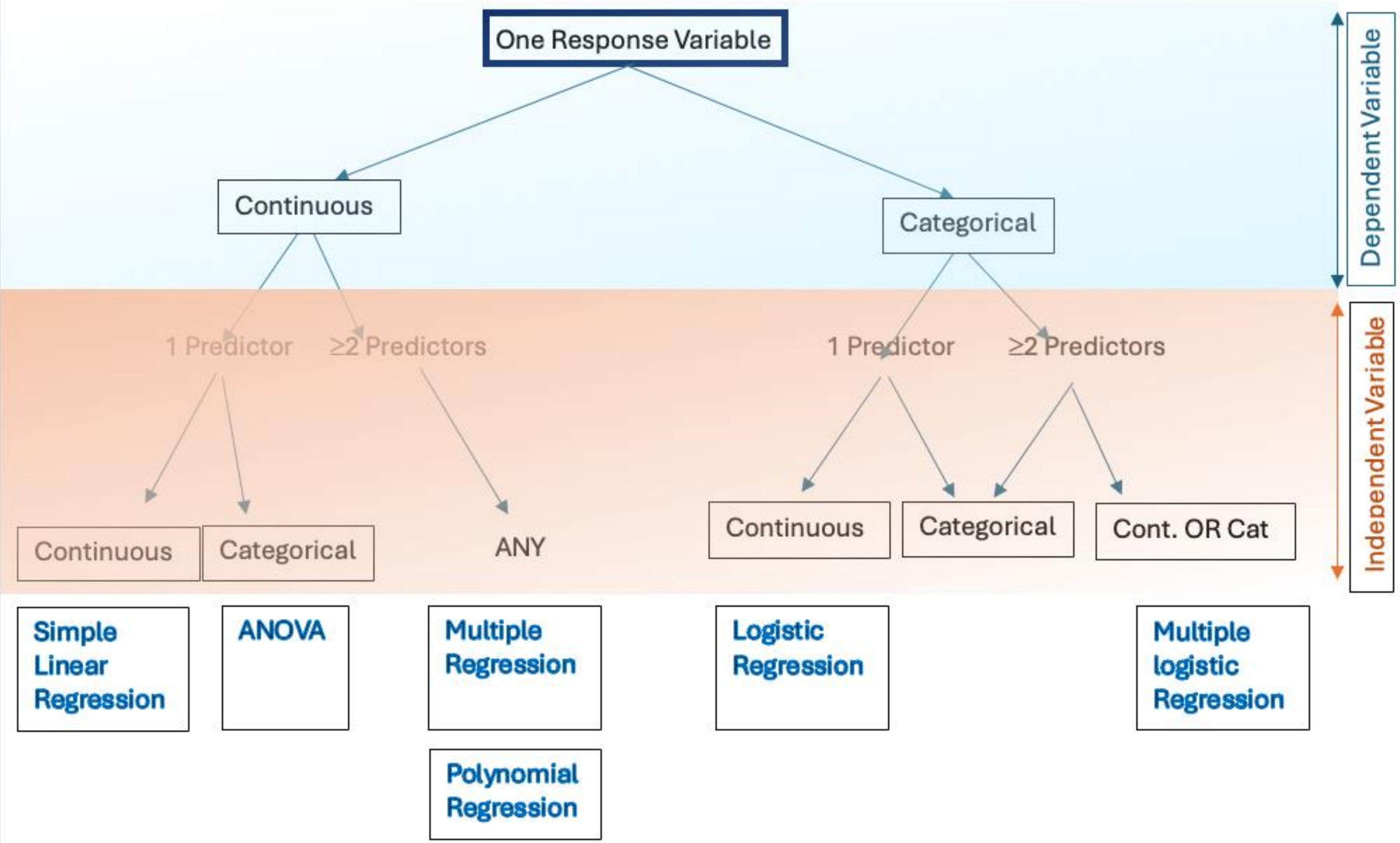

Scientists be like
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0

Regression:

- Statistics is about prediction
- Used to **predict** value of one numerical variable from the value of another
  - predicting dependent/response variable, Y from independent/predictor X
- Linear regression assumes that the relationship between X and Y can be described by a line
  - Fits a straight line to a (messy) scatterplot

Example: ambient temperature may impact growth rate of a plant species, but the reverse is probably not true

A common point of confusion: Pearson's correlation is the slope of the best-fit line ***after standardizing*** **both** variables. The regression slope is the **unstandardized** version that tells you how much Y changes per unit X.

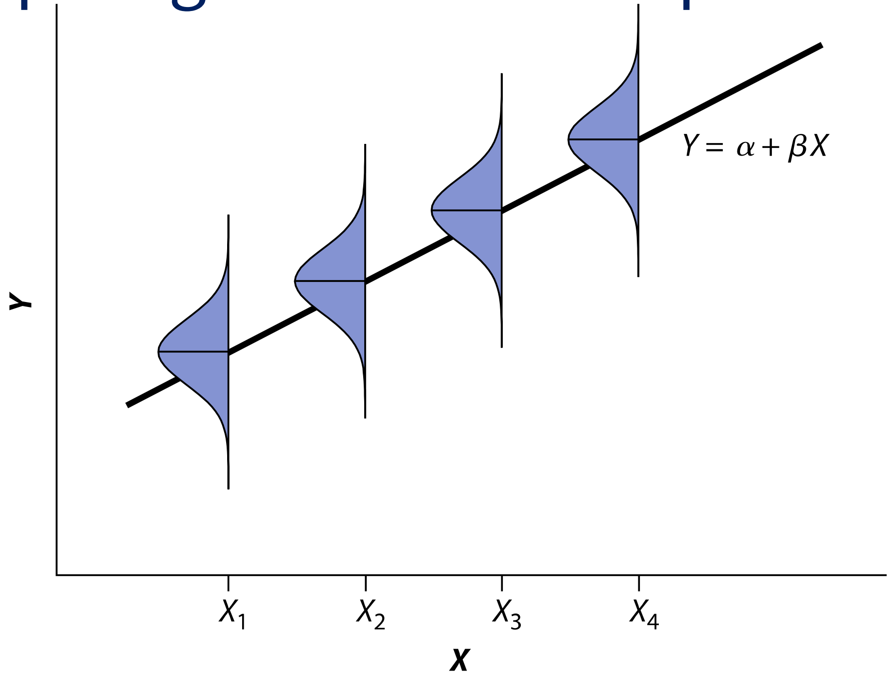


- Other kinds:
    - Lasso
        - Variable selection (weighting a predictor variable by 0)
    - Ridge
        - Allows analysis even in the face of **collinearity**

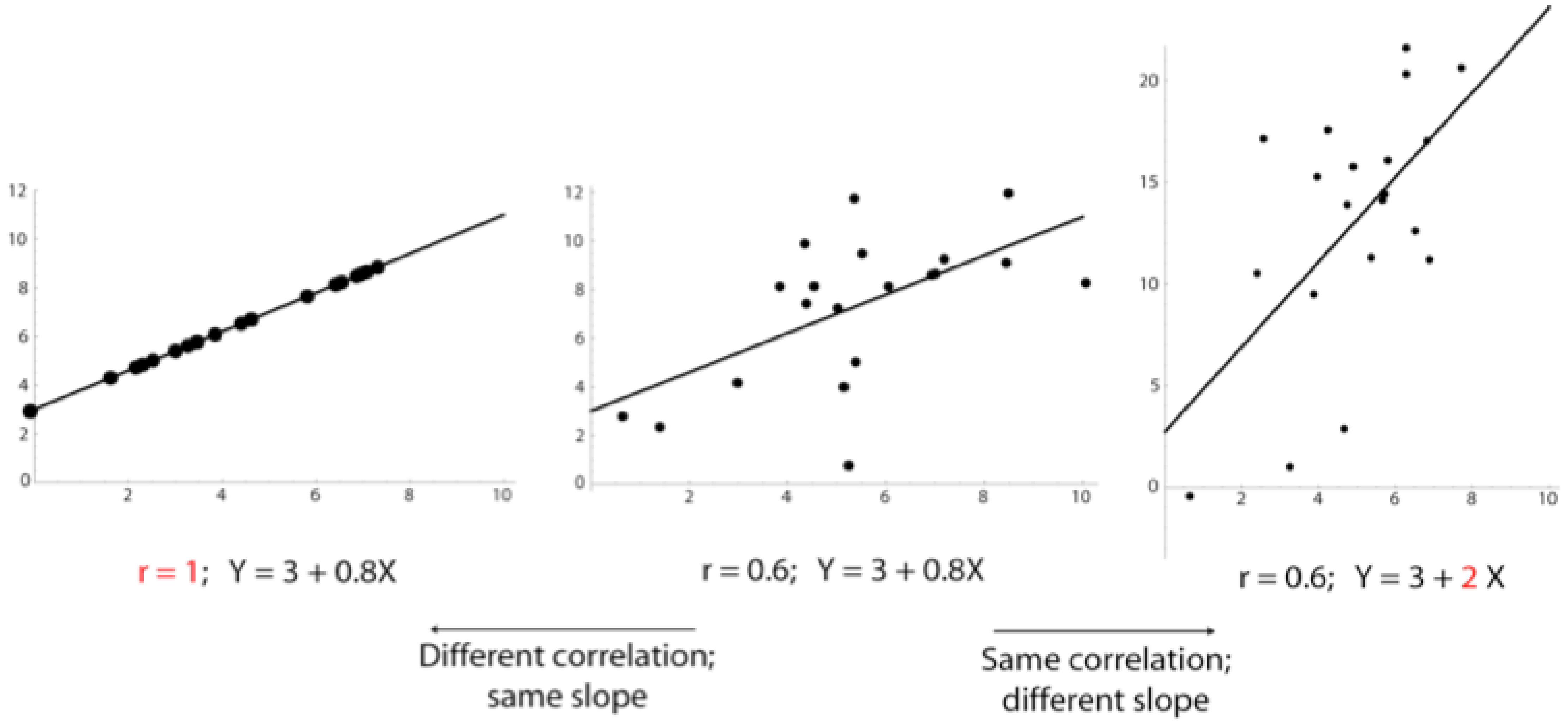
Regression:

- Linear regression assumes that the relationship between X and Y can be described by a line
  - Fits a straight line to a (messy) scatterplot
- Homoscedasticity: Y is normally distributed with equal variance for all values of X

Example: ambient temperature may impact growth rate of a plant species, but the reverse is probably not true



correlation vs regression



The parameters of linear regression

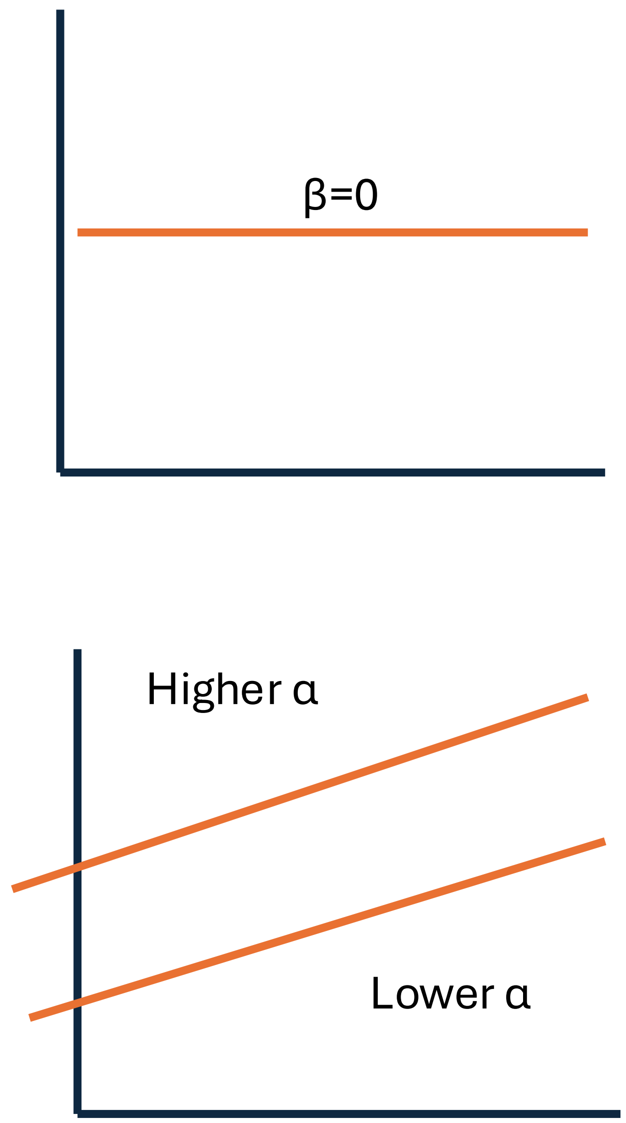
$$Y = \boxed{\alpha} + \beta X + \varepsilon_1$$

intercept

Slope*

* Very similar to Pearson's correlation (but normalized only wrt to X)

Regression Overview



Estimating a regression line

Y = a + bX + ε_1

Quick Review:

$$Z = \frac{\bar{Y} - \mu}{\sigma_{\bar{Y}}} = \frac{\bar{Y} - \mu}{\sigma / \sqrt{n}}$$

$$t = \frac{\bar{Y} - \mu}{SE_{\bar{Y}}} = \frac{\bar{Y} - \mu}{s / \sqrt{n}}$$

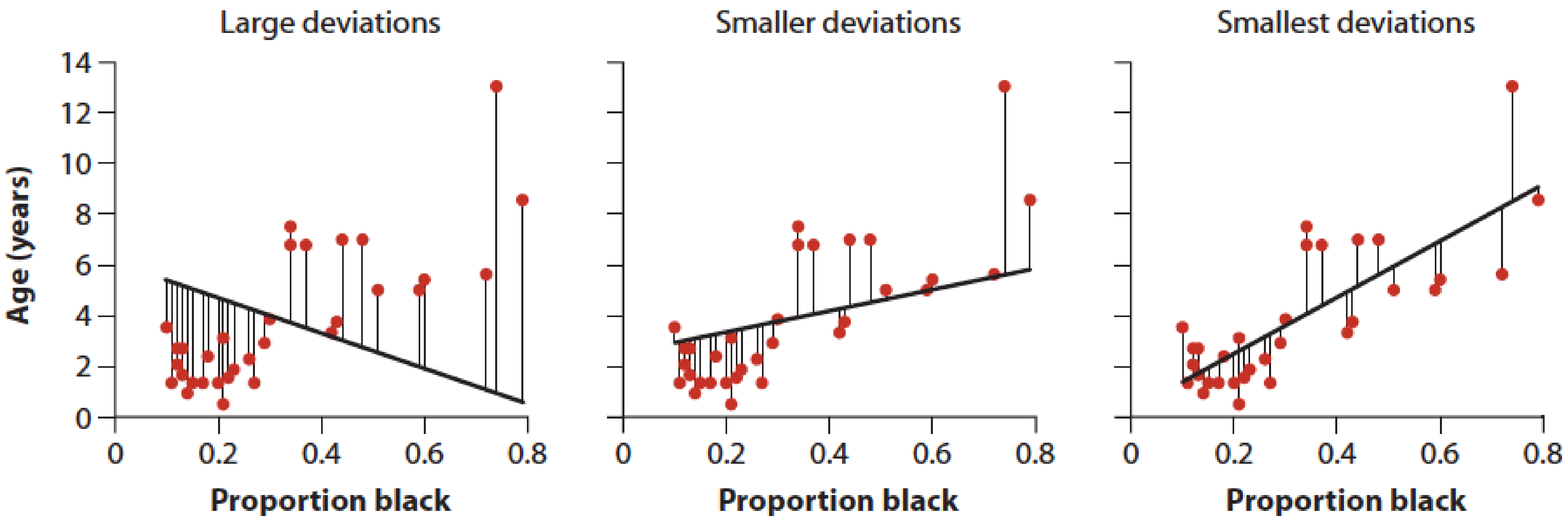
F-value = $\dfrac{\text{MS}_{\text{group}}}{\text{MS}_{\text{error}}}$ SIGNAL / NOISE

$$r = \frac{\square (X - \bar{X})(Y - \bar{Y})}{\sqrt{\square (X - \bar{X})^2} \sqrt{\square (Y - \bar{Y})^2}} = \frac{Co\,var\,iance(X,Y)}{s_x s_y}$$

(Ordinary) Least Squares:

- Best fitting line through a scatterplot
  - Line that minimized spread of y values

- Minimize $SS_{residuals}$
  - Measurement of how much the line's predicted y_i deviate from actual data values



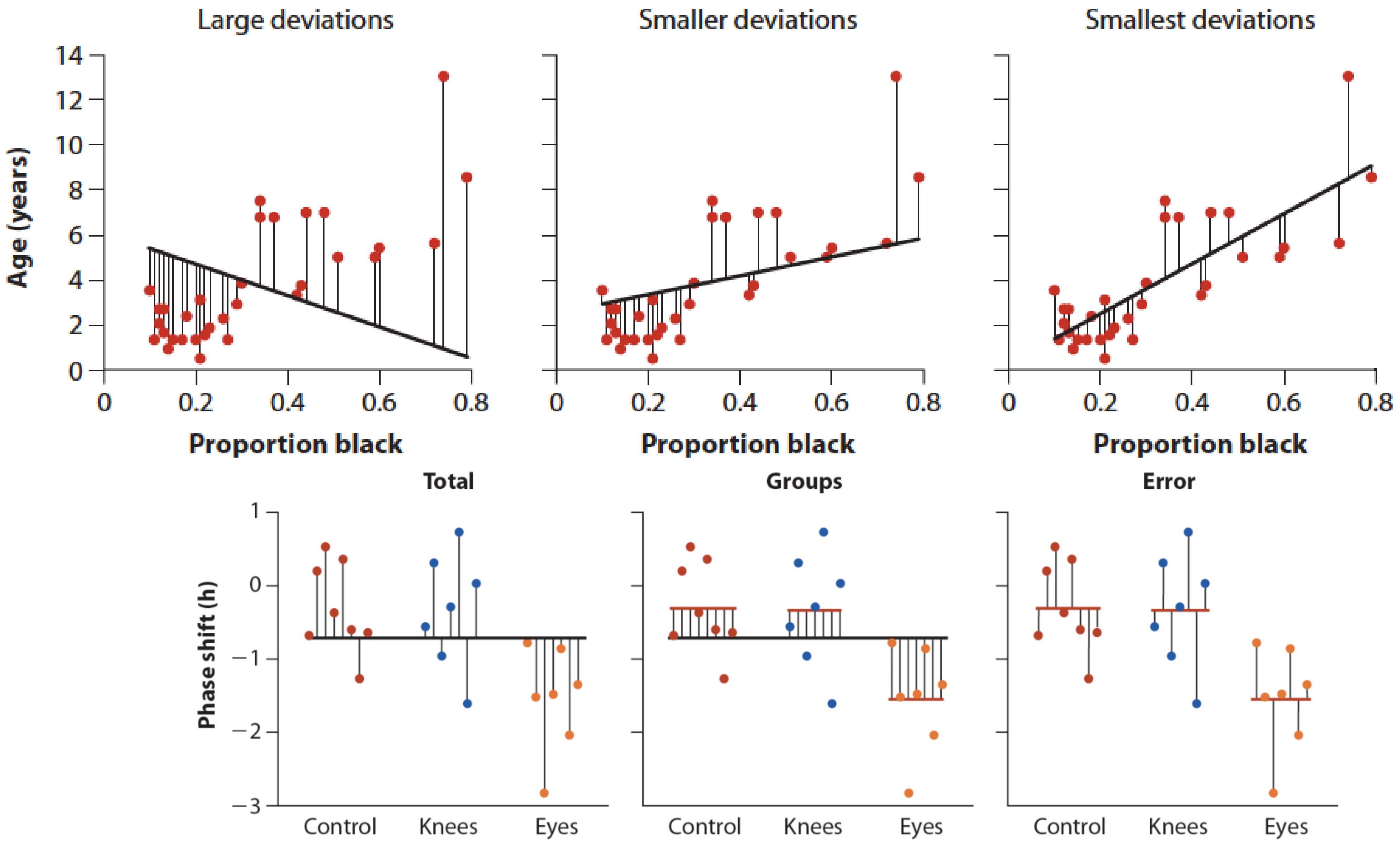


Figure 20.1: Whitock and Schluter, Fig 15.1.2 – Illustrating the partitioning of sum of squares into MS_{group} and MS_{error} components.

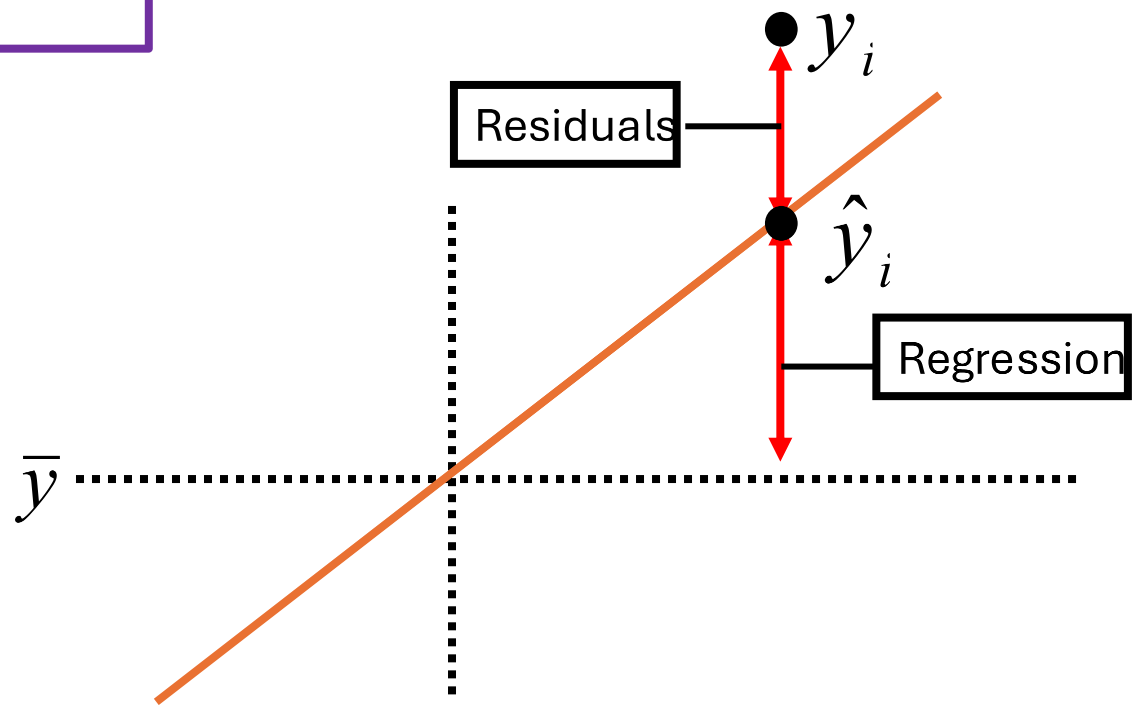
Regression Overview

Least Squares:

- What are the elements of this equation?

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

$$\hat{y}_i = a + bx_i$$

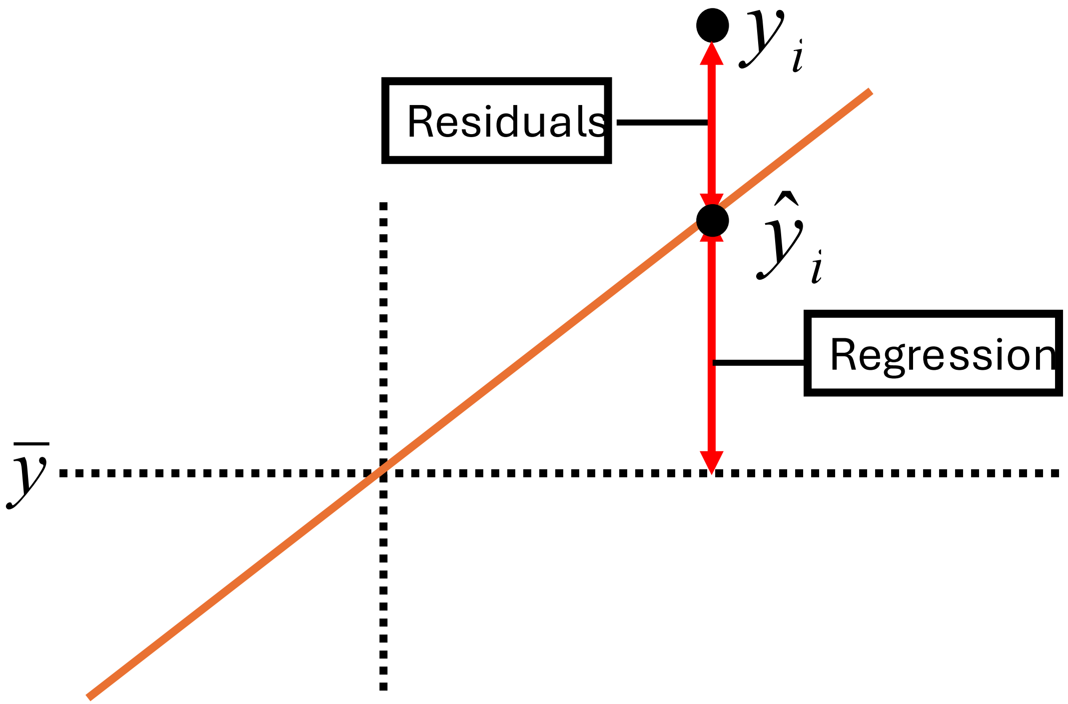


- **Residuals:**
  - Residuals measure the scatter of points above and below the least squares regression line
  - $MS_{residual}$ is the variance of the residuals, $residual = Y_i - \hat{Y}_i$

$$\boxed{MS_{residual} = \frac{\square\,(Y_i - \hat{Y}_i)^2}{n-2}}$$

  - $MS_{regression}$ is the variance of the regression

$$MS_{regression} = \frac{\sum(\hat{Y}_i - \bar{Y}\ )^2}{n-2}$$



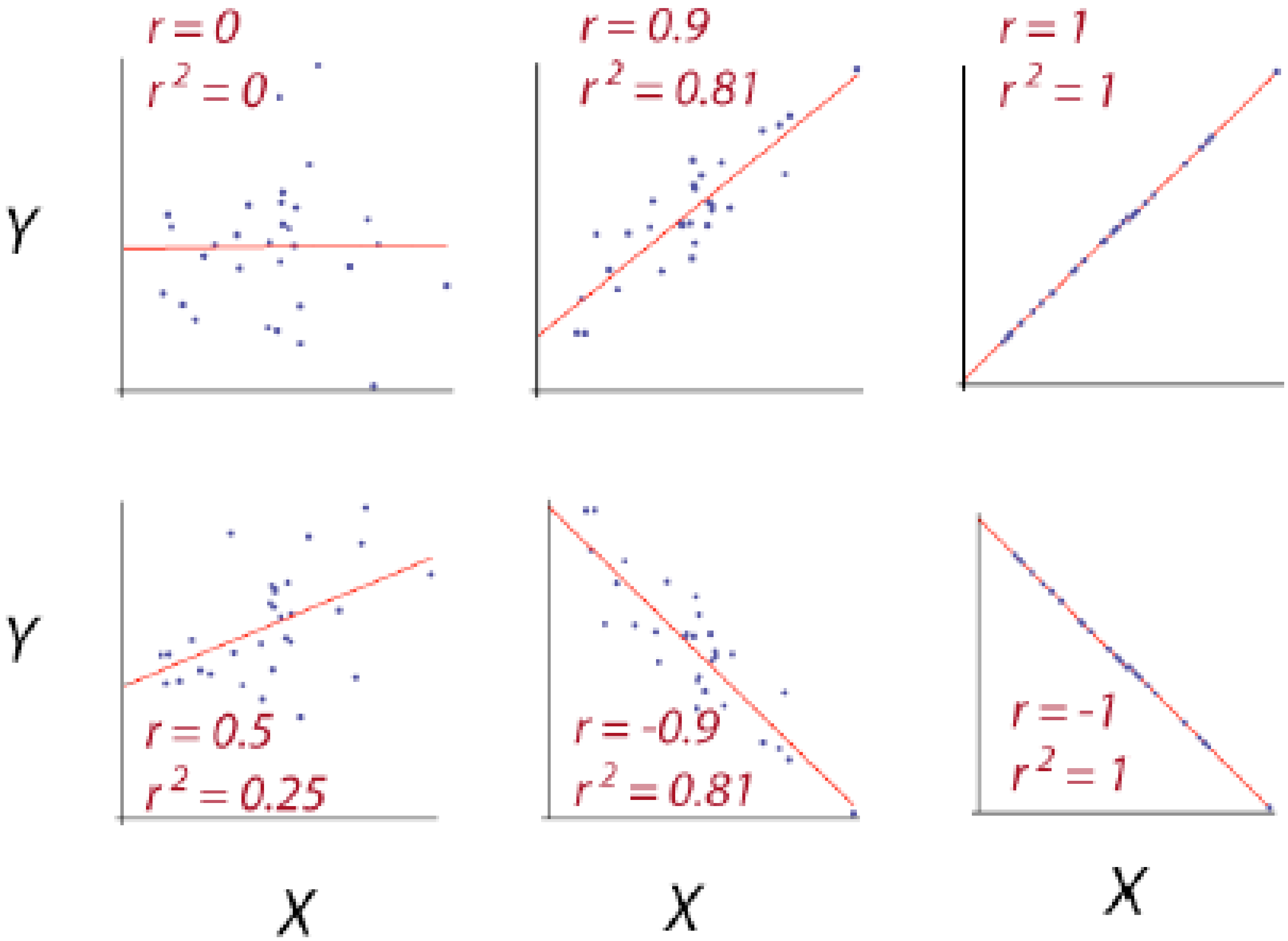
  - Coefficient of determination (r^2) = SSR/SST

R^2 predicts the amount of variance in Y explained by the regression line

We saw this in ANOVA where R^2 gave 'precision' of model (i.e. Ability of the model to explain variation)

- The coefficient of determination
- Sometimes written as r^2
- Square of the correlation coefficient, r

$$R^2 = \frac{SS_{regression}}{SS_{Total}}$$



r = 0
r 2 = 0
r = 0.9
r 2 = 0.81
r = 1
r 2 = 1
Y
r = 0.5
r 2 = 0.25
r = -0.9
r 2 = 0.81
r = -1
r 2 = 1
Y
X
X
X

Best estimate of slope:

b = Sum of cross products / Sum of squares of X

$$b = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

Best estimate of slope:

$$b = \frac{\text{Sum of cross products}}{\text{Sum of squares of X}}$$

$$b = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2} \sqrt{\sum (Y - \bar{Y})^2}} = \frac{Covariance(X,Y)}{s_x s_y}$$

Denominator normalizes based on **both** X and Y variables

Denominator ONLY normalizes based on X (Independent/Explanatory variable)

Finding **a**:

$$\boxed{\overline{Y} = a + b\overline{X}}$$

OR

$$\boxed{a = \overline{Y} - b\overline{X}}$$