REGRESSION AND CORRELATION METHODS

What is a Math/Stats Model?

1. Often Describe Relationship between Variables

2. Types
   - Deterministic Models (no randomness)
   - Probabilistic Models (with randomness)

Deterministic Models

1. Hypothesize Exact Relationships
2. Suitable When Prediction Error is Negligible
3. Example: Body mass index (BMI) is measure of body fat based

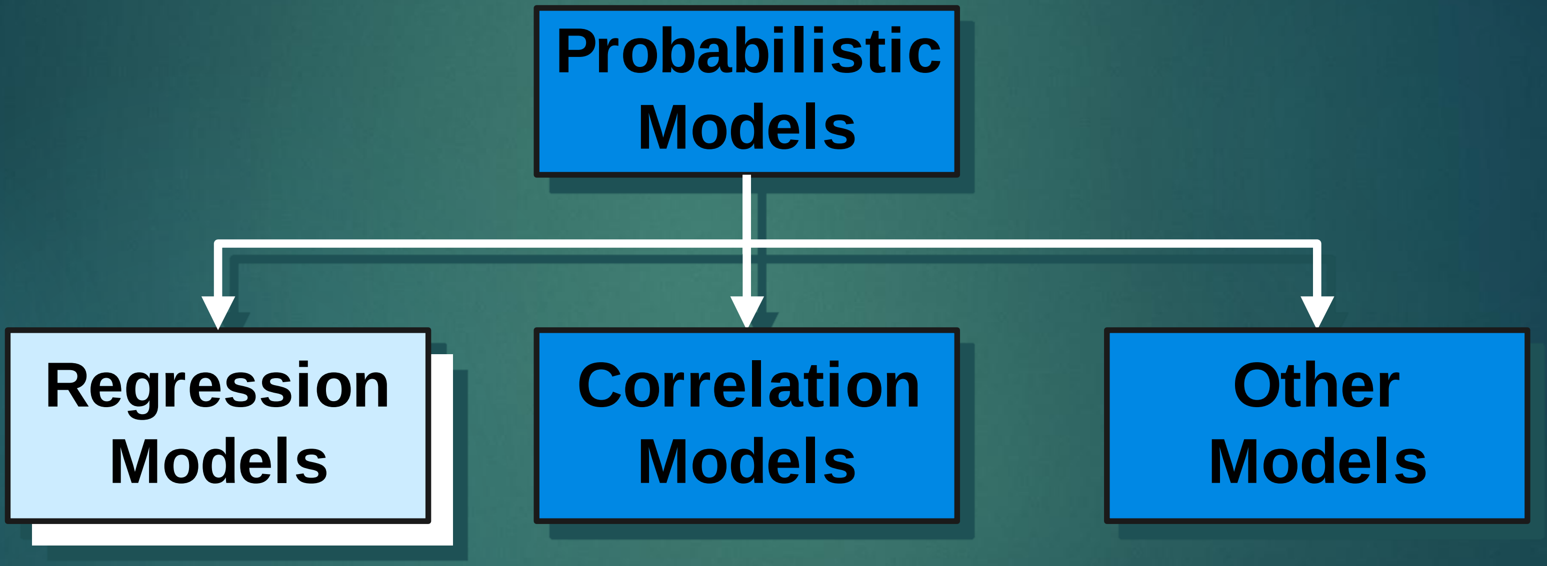
- Metric Formula: $BMI = \dfrac{\text{Weight in Kilograms}}{(\text{Height in Meters})^2}$

- Non-metric Formula: $BMI = \dfrac{\text{Weight(pounds)x703}}{(\text{Height in inches})^2}$

Probabilistic Models

1. Hypothesize 2 Components
   - Deterministic
   - Random Error
2. Example: Systolic blood pressure of newborns Is 6 Times the Age in days + Random Error
   - $SBP = 6\text{xage}(d) + \varepsilon$
   - Random Error May Be Due to Factors Other Than age in days (e.g. Birthweight)

Types of Probabilistic Models



Regression Models

Regression Models

- Relationship between one dependent variable and explanatory variable(s)
- Use equation to set up relationship
  - Numerical Dependent (Response) Variable
  - 1 or More Numerical or Categorical Independent (Explanatory) Variables
- Used Mainly for Prediction & Estimation

Regression Modeling Steps

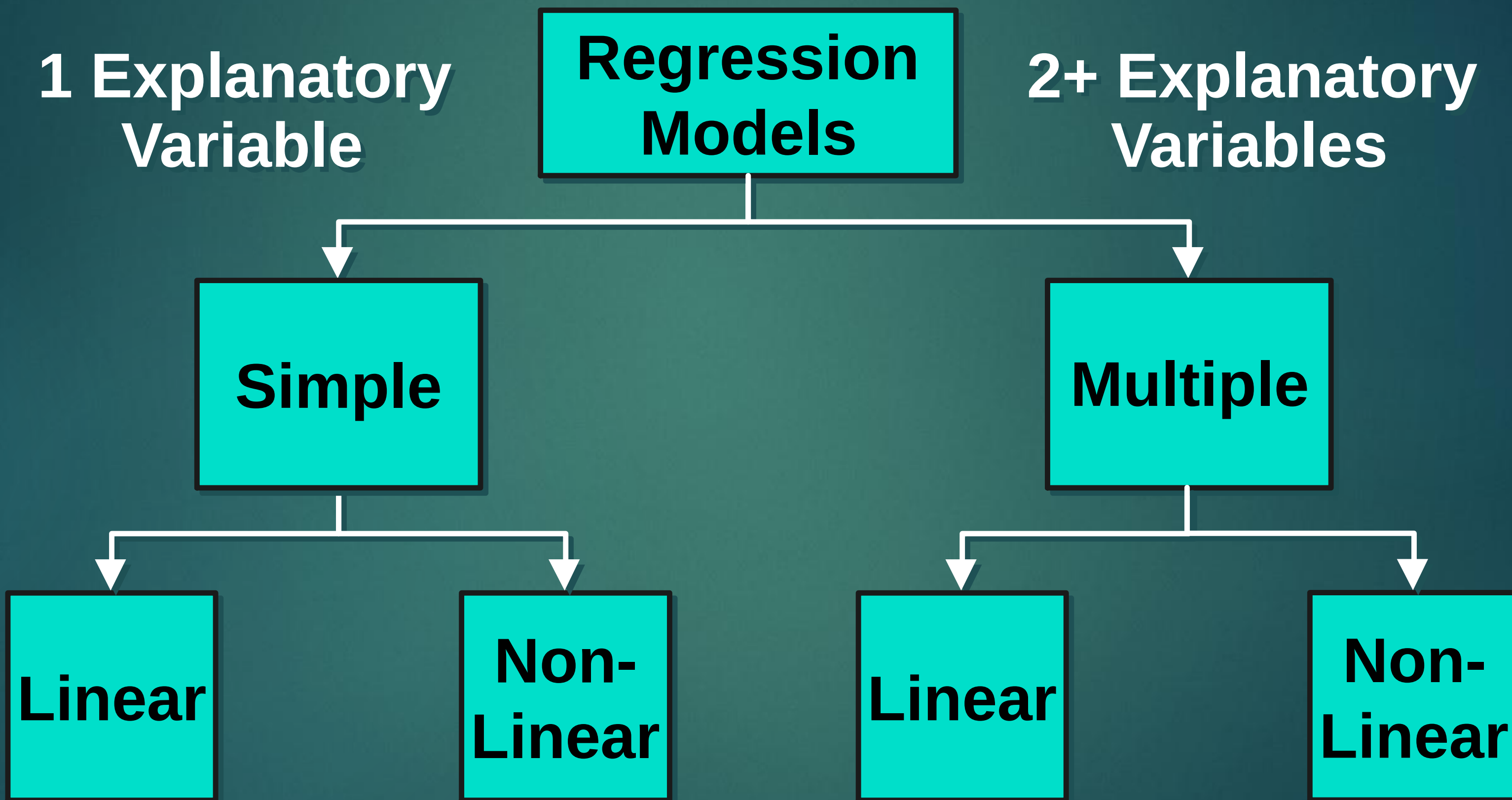
- 1. Hypothesize Deterministic Component
    - Estimate Unknown Parameters
- 2. Specify Probability Distribution of Random Error Term
    - Estimate Standard Deviation of Error
- 3. Evaluate the fitted Model
- 4. Use Model for Prediction & Estimation

Specifying the deterministic component

- 1. Define the dependent variable and independent variable

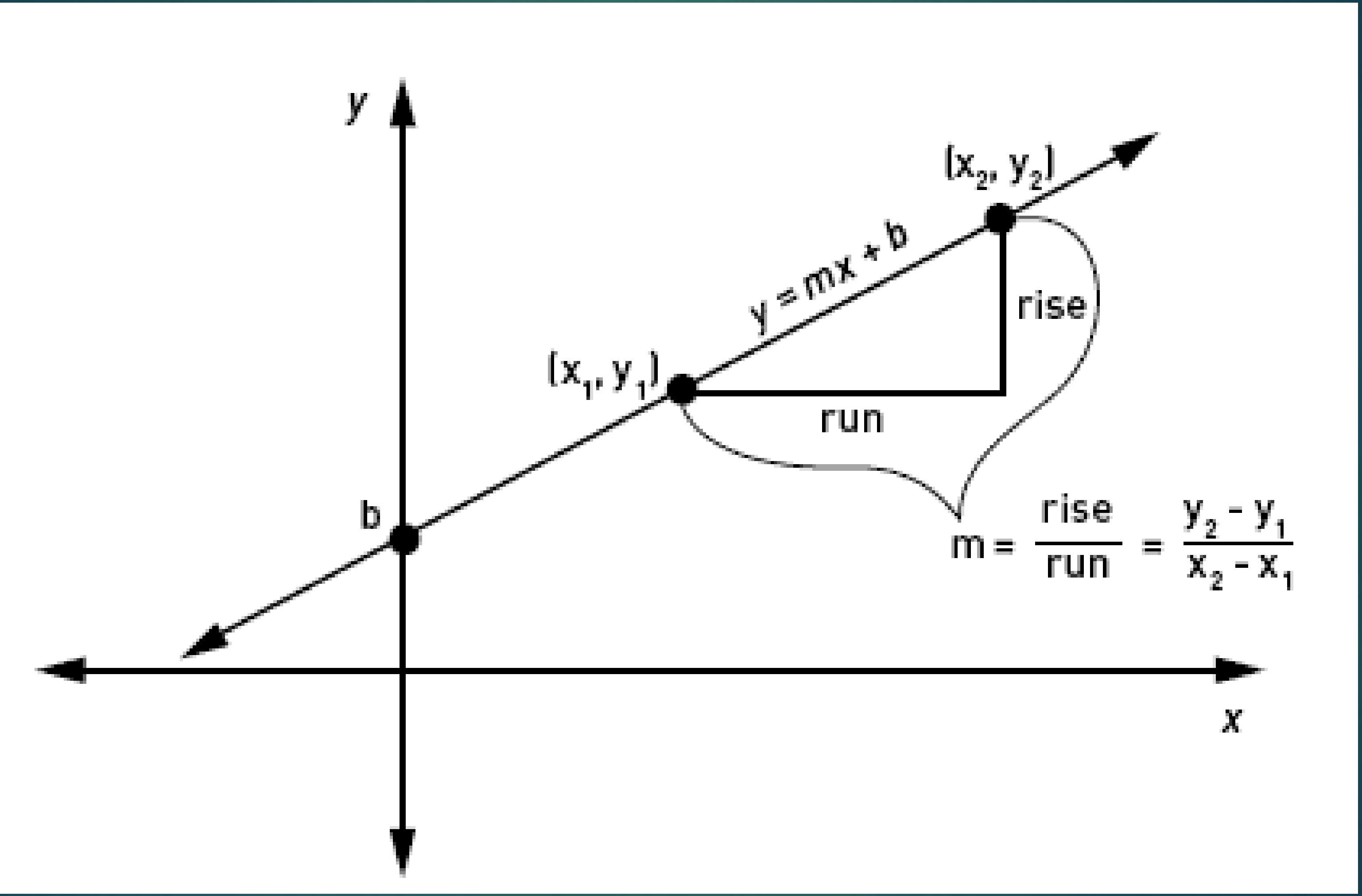
- 2. Hypothesize Nature of Relationship
  - Expected Effects (i.e., Coefficients' Signs)
  - Functional Form (Linear or Non-Linear)
  - Interactions

Types of Regression Models



Linear Regression Model

Linear Equations



Linear Regression Model

- 1. Relationship Between Variables Is a Linear Function

Population Y-Intercept

Population Slope

Random Error

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

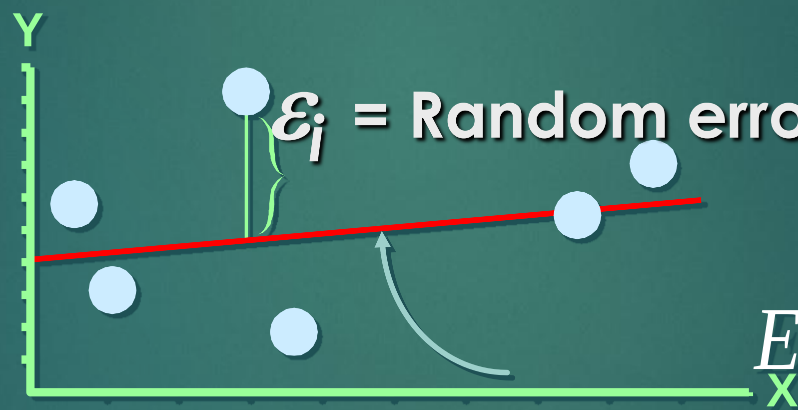
Dependent (Response) Variable (e.g., CD+ c.)

Independent (Explanatory) Variable (e.g., Years s. serocon.)

Population Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Observed value



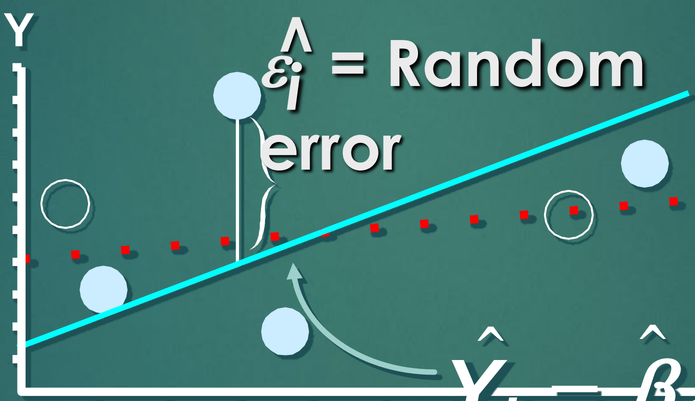
ε_i = **Random error**

$$E(Y) = \beta_0 + \beta_1 X_i$$

Observed value

Sample Linear Regression Model

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\varepsilon}_i$$



$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

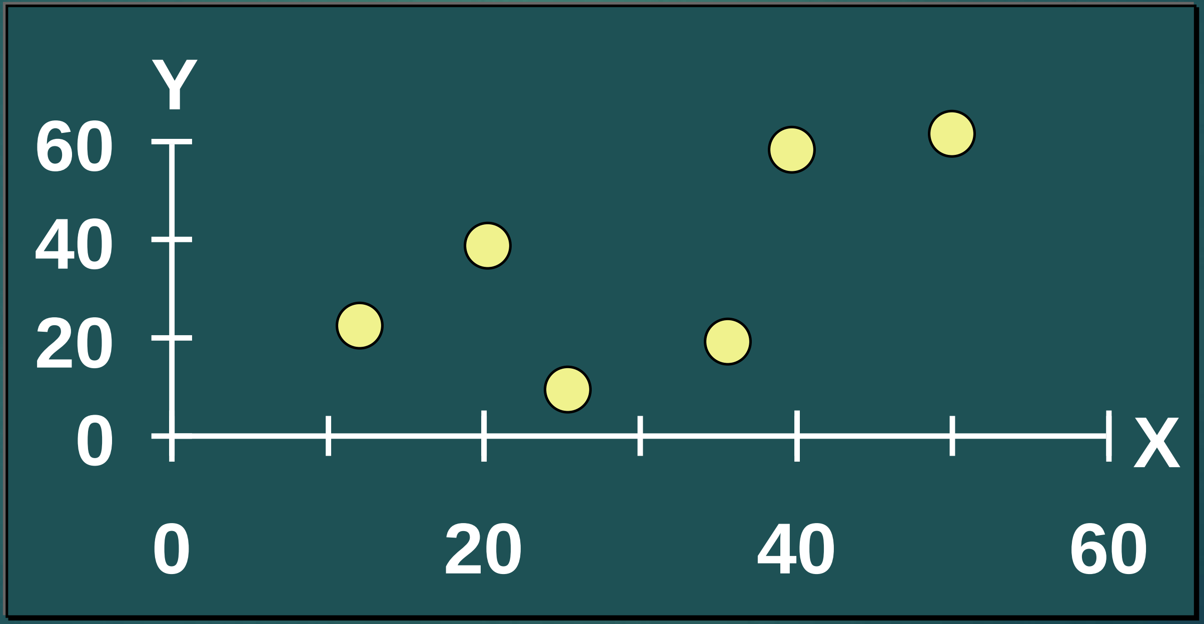
Unsampled observation

Observed value

Estimating Parameters: Least Squares Method

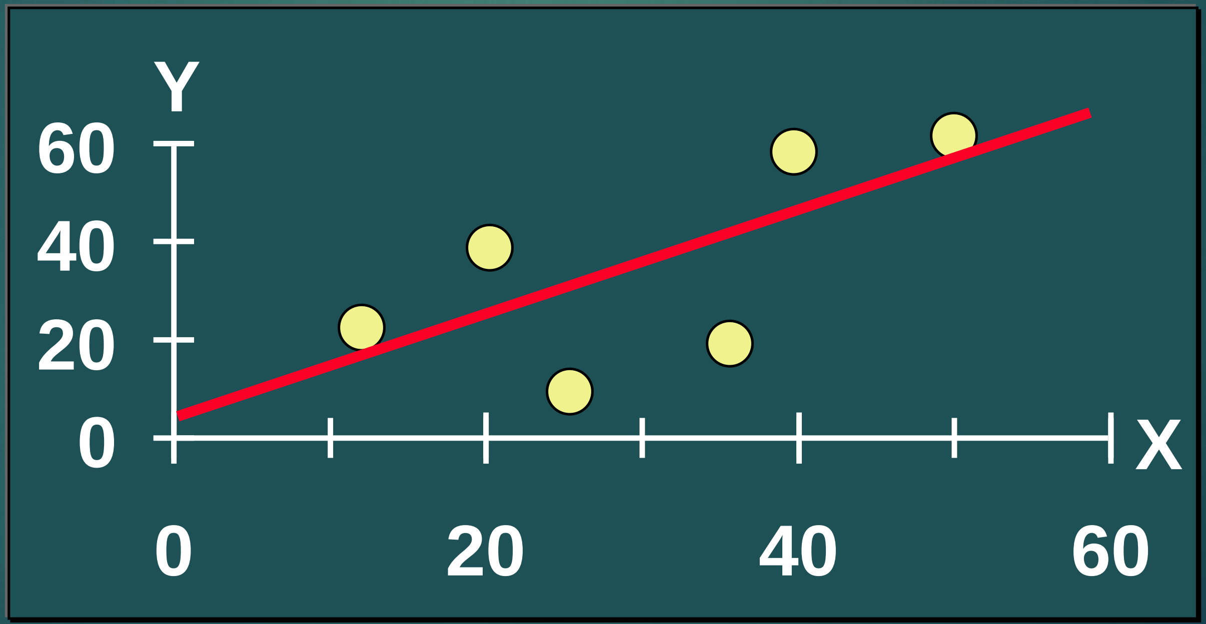
Scatter plot

- 1. Plot of All (X_i, Y_i) Pairs
- 2. Suggests How Well Model Will Fit



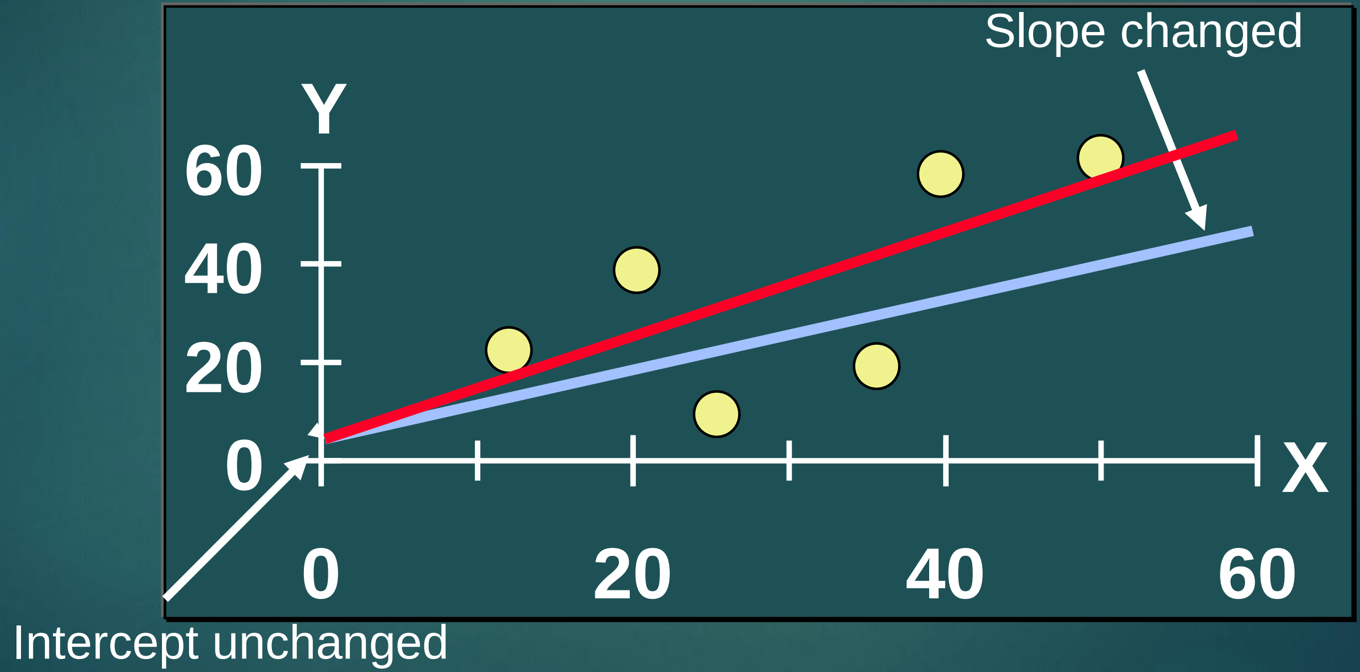
Thinking Challenge

How would you draw a line through the points? How do you determine which line 'fits best'?



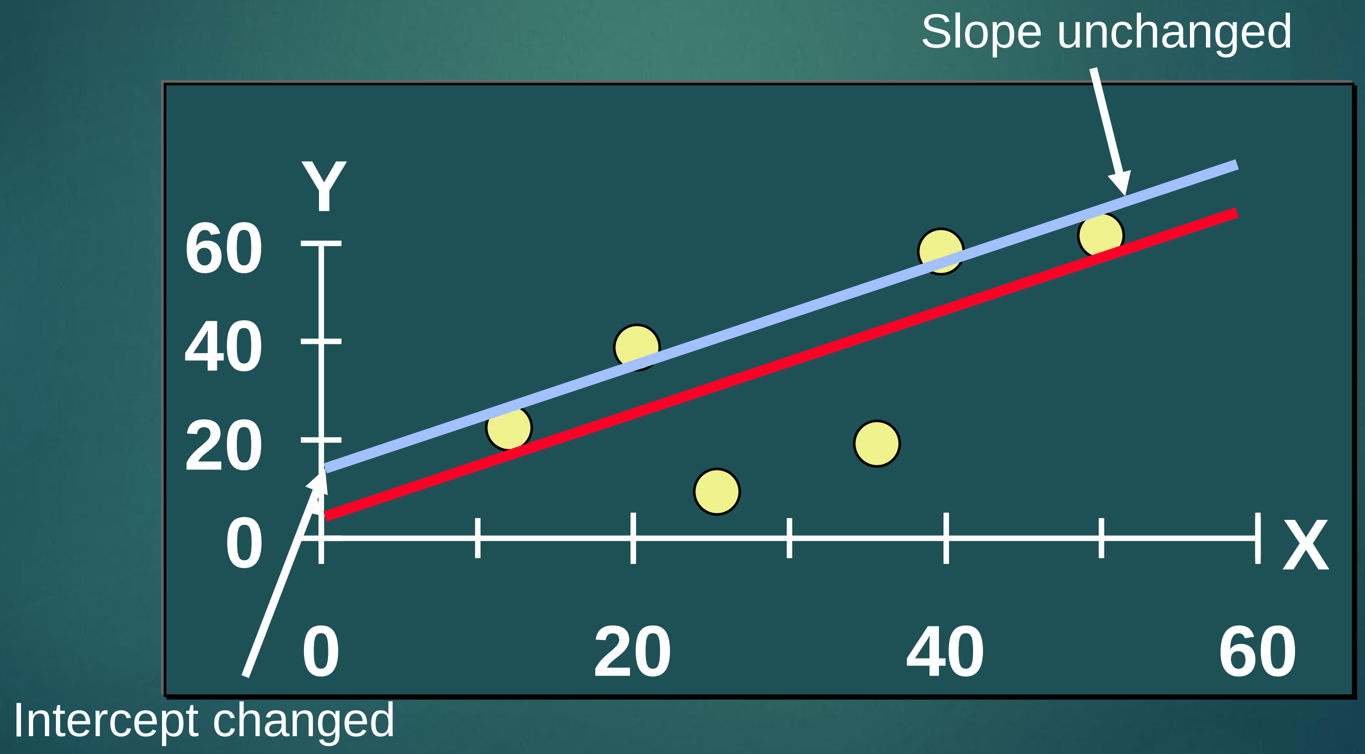
Thinking Challenge

How would you draw a line through the points? How do you determine which line 'fits best'?



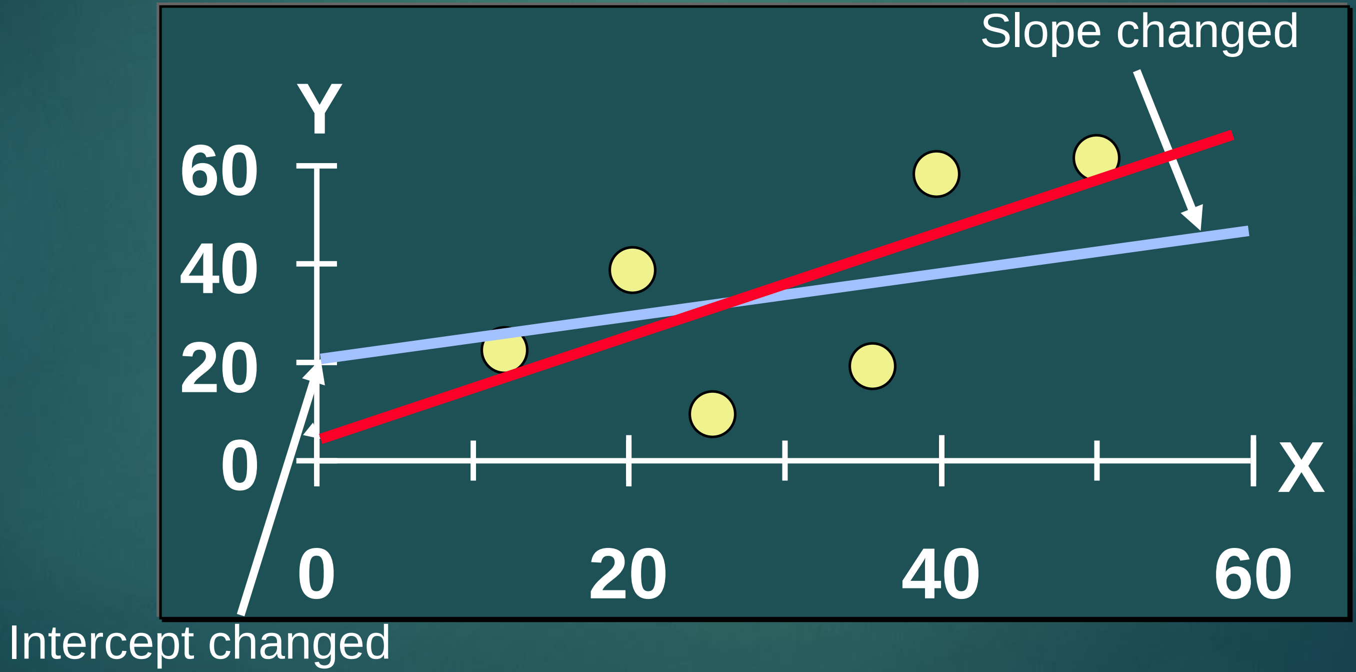
Thinking Challenge

How would you draw a line through the points? How do you determine which line 'fits best'?



Thinking Challenge

How would you draw a line through the points? How do you determine which line 'fits best'?



Least Squares

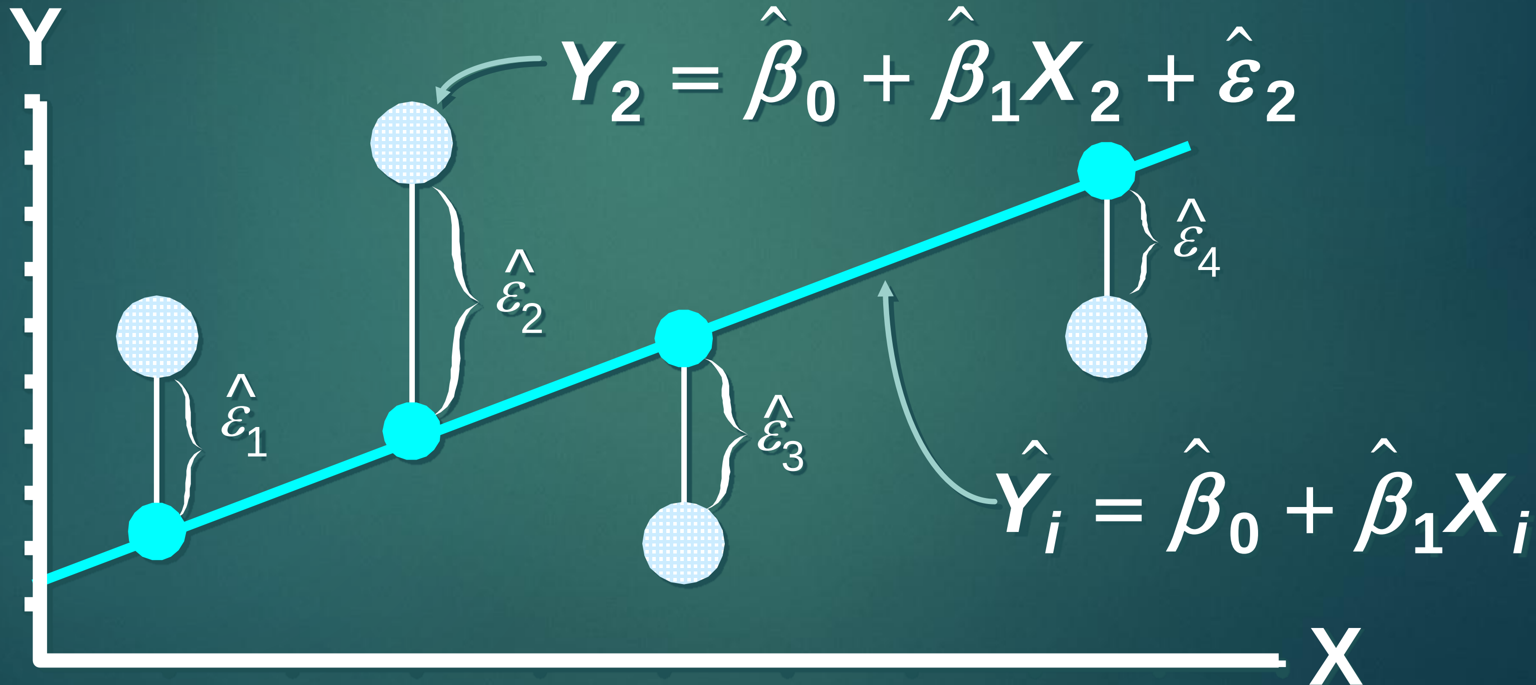
- 1. 'Best Fit' Means Difference Between Actual Y Values & Predicted Y Values Are a Minimum. *But* Positive Differences Off-Set Negative. So square errors!

$$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

- 2. LS Minimizes the Sum of the Squared Differences (errors) (SSE)

Least Squares Graphically

LS minimizes $\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \hat{\varepsilon}_1^2 + \hat{\varepsilon}_2^2 + \hat{\varepsilon}_3^2 + \hat{\varepsilon}_4^2$



Computation Table

X_i	Y_i	X_i^2	Y_i^2	X_iY_i
X_1	Y_1	X_1^2	Y_1^2	X_1Y_1
X_2	Y_2	X_2^2	Y_2^2	X_2Y_2
:	:	:	:	:
X_n	Y_n	X_n^2	Y_n^2	X_nY_n
ΣX_i	ΣY_i	ΣX_i^2	ΣY_i^2	ΣX_iY_i

Interpretation of Coefficients

- 1. Slope (β_1)
  - Estimated *Y* Changes by β_1 for Each 1 Unit Increase in *X*
    - If $\beta_1 = 2$, then *Y* Is Expected to Increase by 2 for Each 1 Unit Increase in *X*
- 2. Y-Intercept (β_0)
  - Average Value of *Y* When $X = 0$
    - If $\beta_0 = 4$, then Average *Y* Is Expected to Be 4 When *X* Is 0