General Linear Regression With

Learning Objectives

1. Describe the Linear Regression Model
2. State the Regression Modeling Steps
3. Explain Ordinary Least Squares
4. Compute Regression Coefficients
5. Understand and check model assumptions
6. Predict Response Variable

Models

What is a Model?

1. Representation of Some Phenomenon

   Non-Math/Stats Model

What is a Math/Stats Model?

1. Often Describe Relationship between Variables

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

Deterministic Models

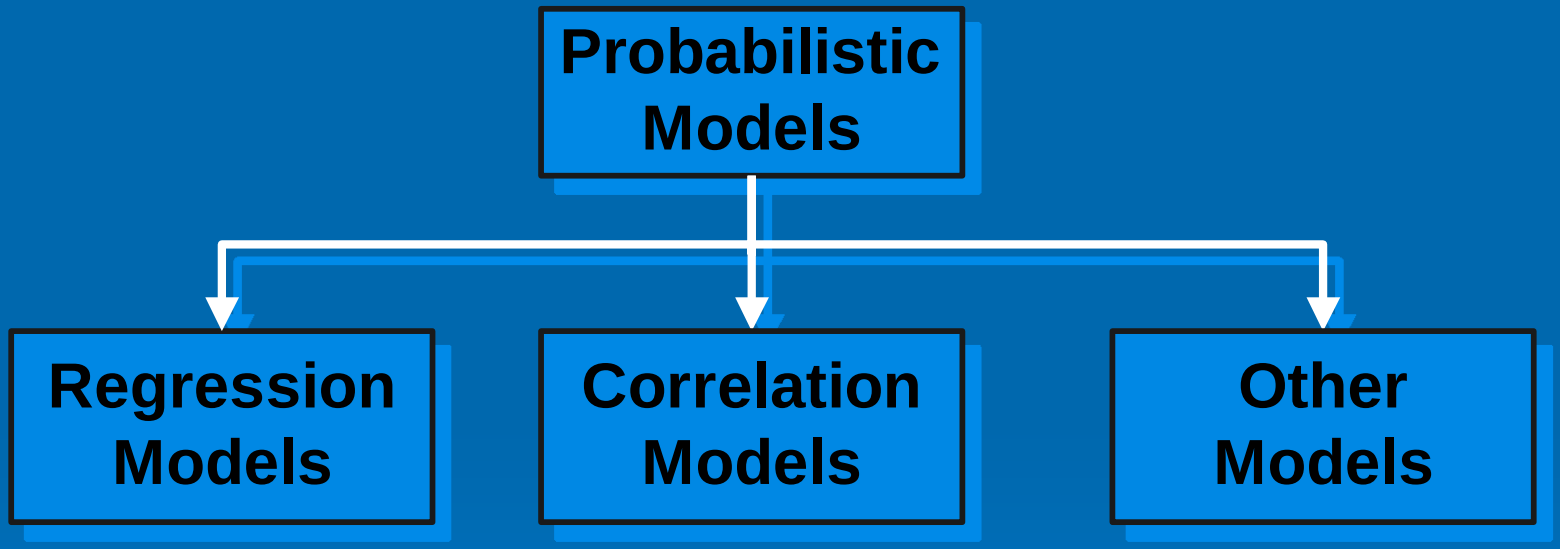
1. Hypothesize Exact Relationships
2. Suitable When Prediction Error is Negligible
3. Example: Body mass index (BMI) is measure of body fat based

- Metric Formula: $BMI = \frac{\text{Weight in Kilograms}}{(\text{Height in Meters})^2}$
- Non-metric Formula: $BMI = \frac{\text{Weight (pounds)} \times 703}{(\text{Height in inches})^2}$

Probabilistic Models

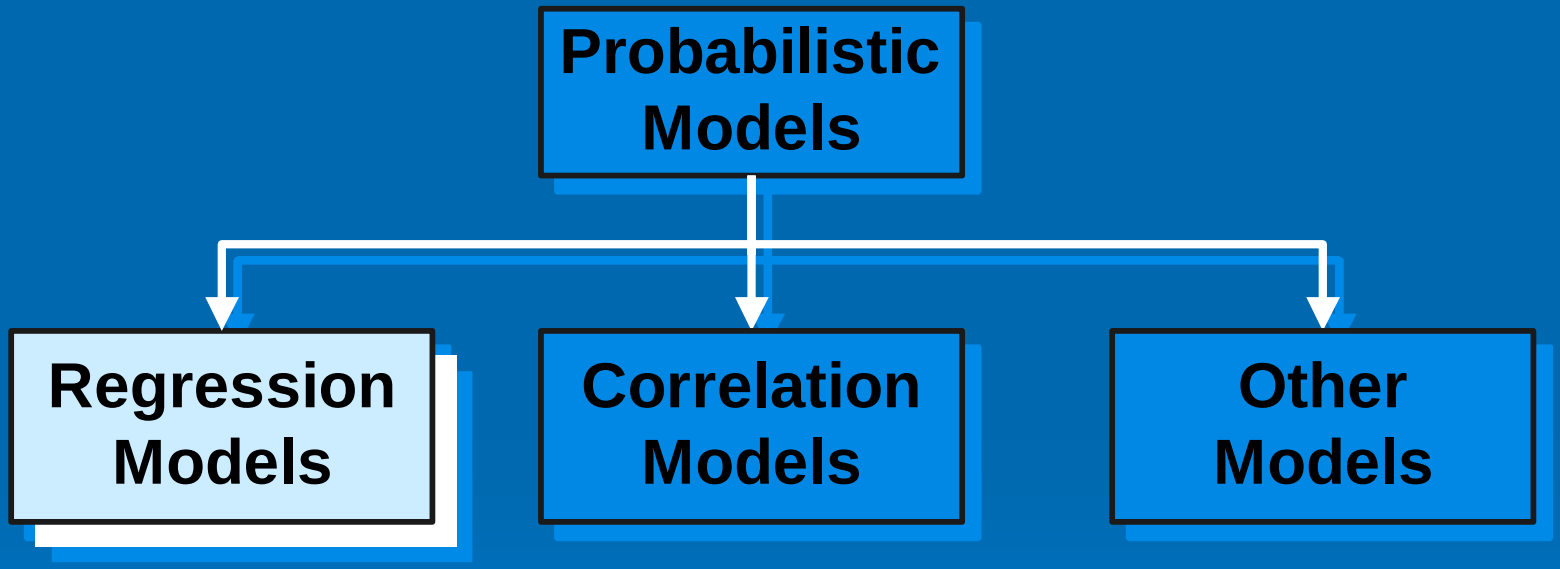
1. Hypothesize 2 Components
   - Deterministic
   - Random Error
2. Example: Systolic blood pressure of newborns Is 6 Times the Age in days + Random Error
   - $SBP = 6 \times age(d) + \varepsilon$
   - Random Error May Be Due to Factors Other Than age in days (e.g. Birthweight)

Types of Probabilistic Models



Regression Models

Types of Probabilistic 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

Model Specification

Specifying the deterministic component

- 1. Define the dependent variable and independent variable

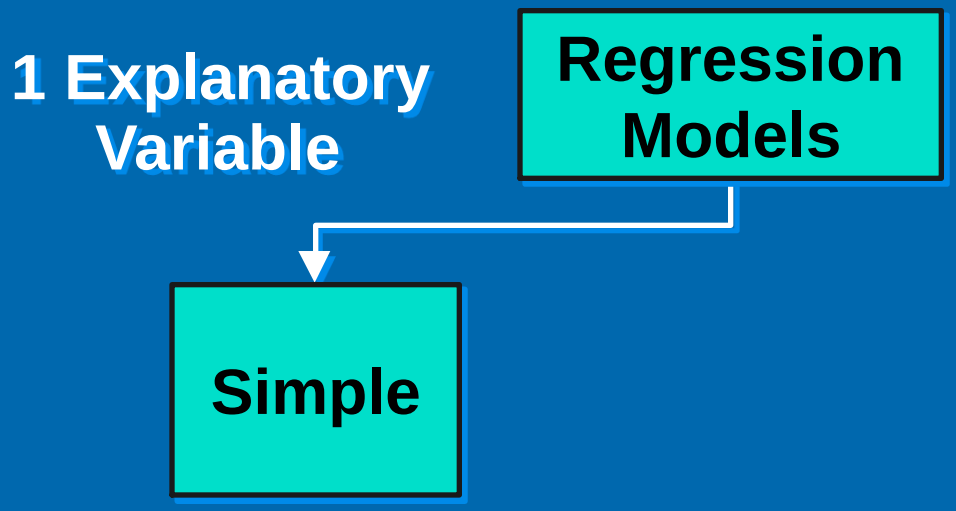
- 2. Hypothesize Nature of Relationship
  - Expected Effects (i.e., Coefficients' Signs)
  - Functional Form (Linear or Non-Linear)
  - Interactions

Types of Regression Models

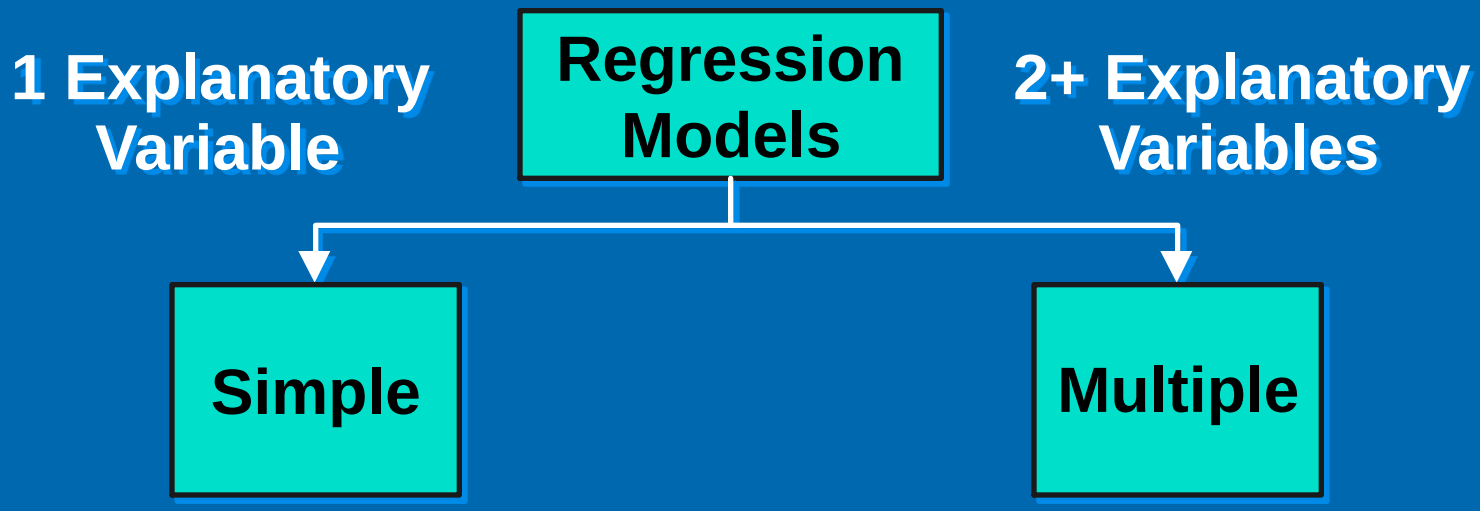
Types of Regression Models

Regression Models

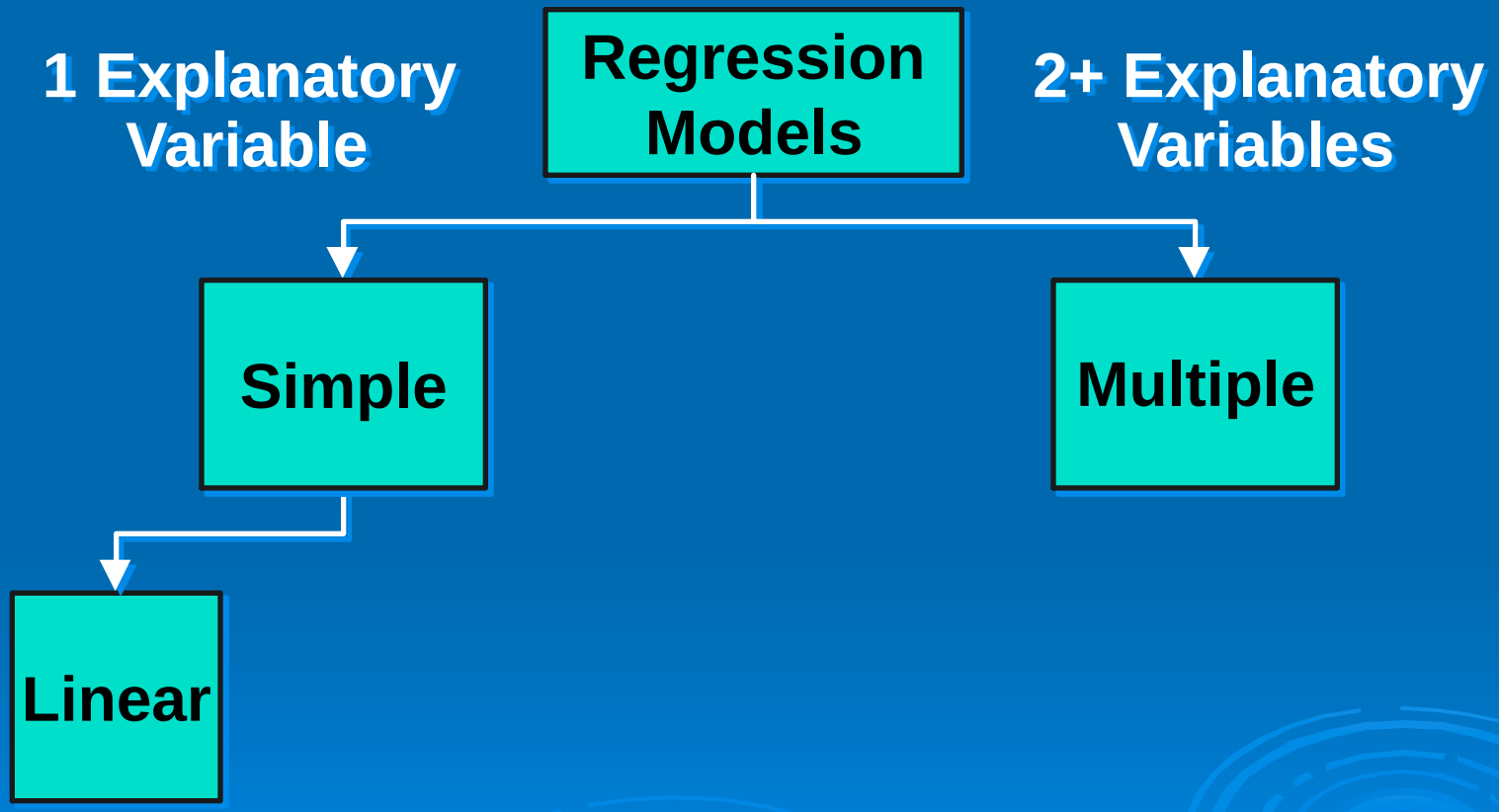
Types of Regression Models



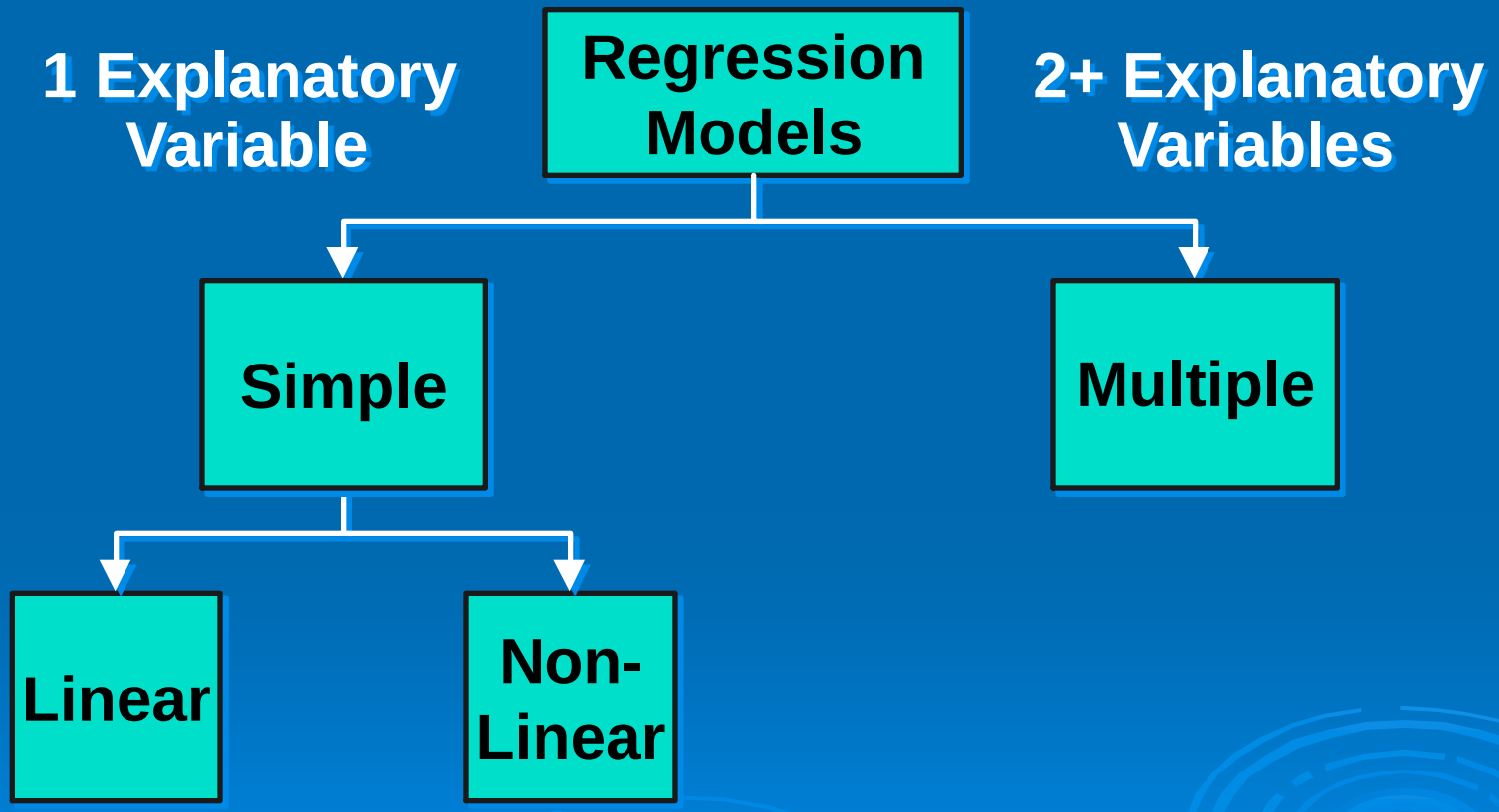
Types of Regression Models



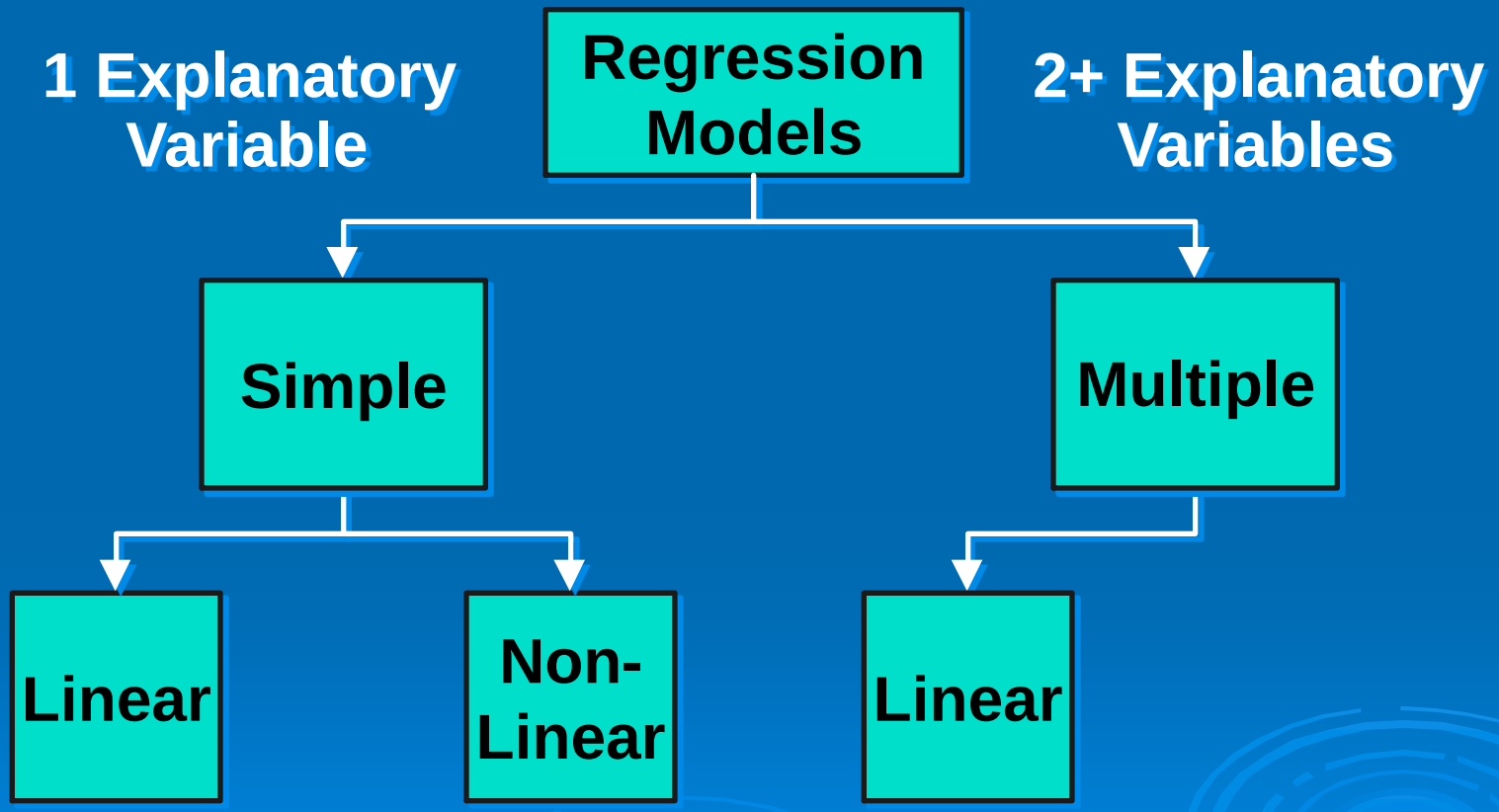
Types of Regression Models



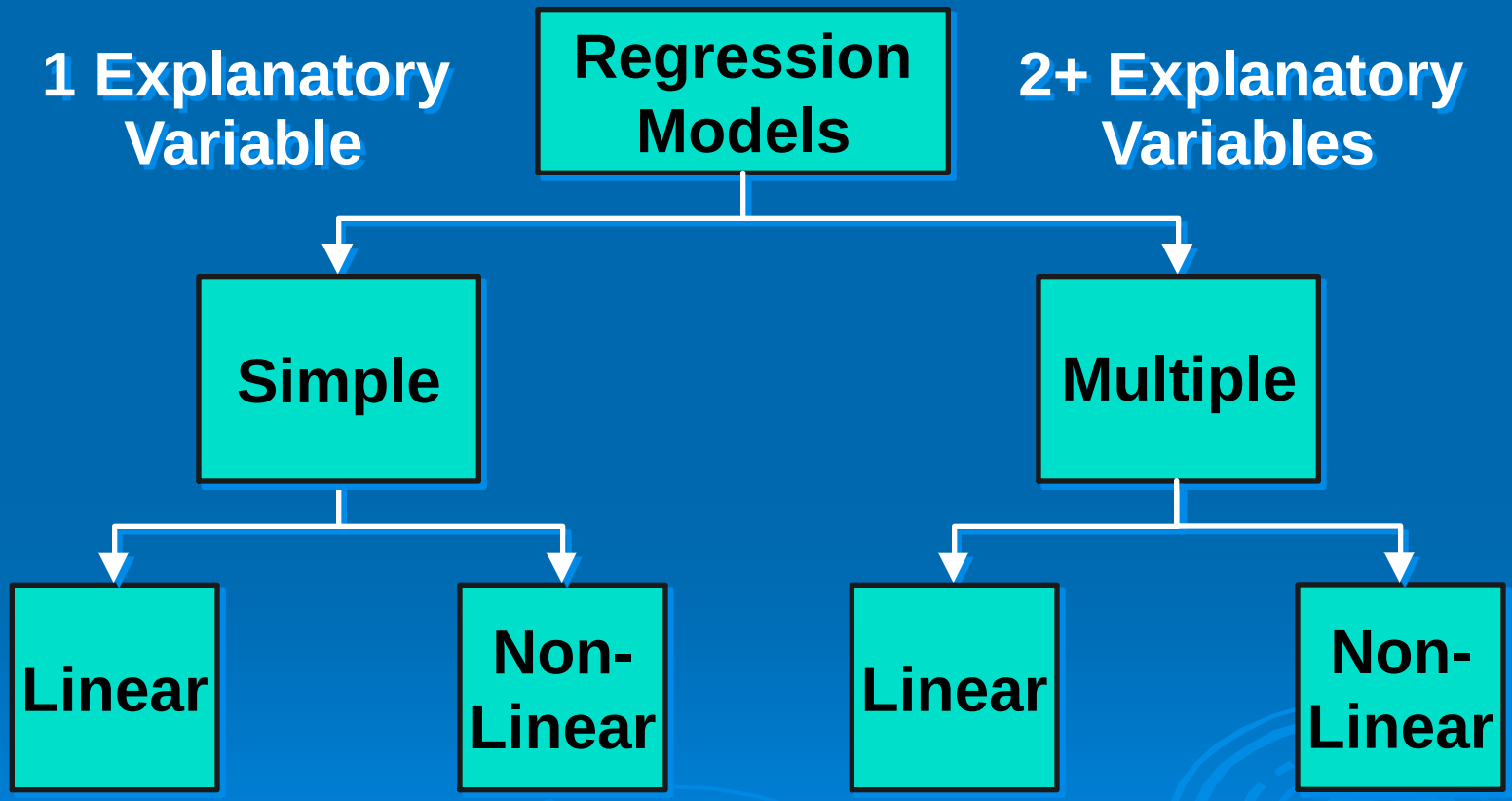
Types of Regression Models



Types of Regression Models

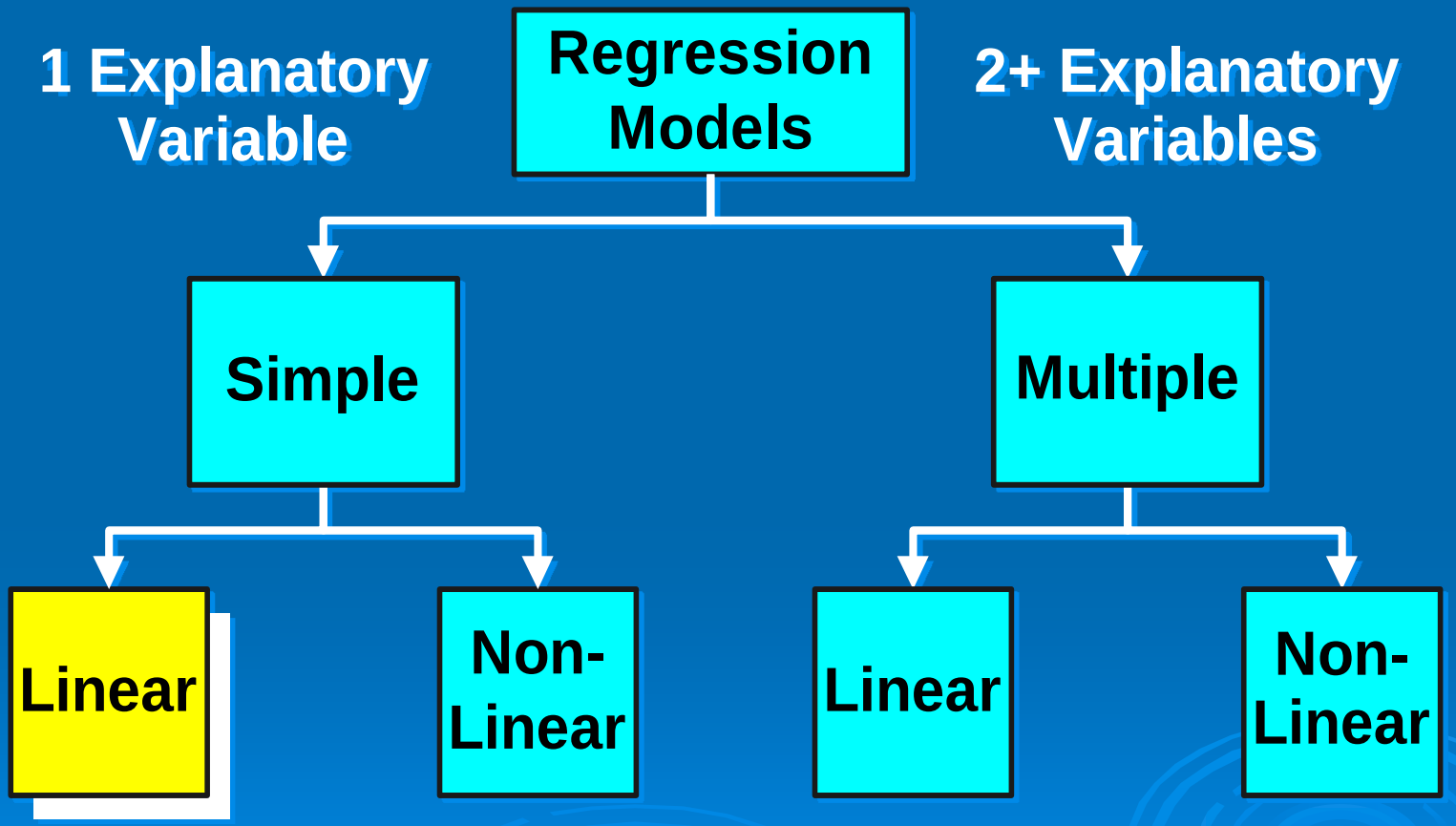


Types of Regression Models



Linear Regression Model

Types of Regression Models



The Explicit Assumptions

These assumptions are explicitly stated by the model:

1. The residuals are independent
2. The residuals are normally distributed
3. The residuals have a mean of 0 at all values of X
4. The residuals have constant variance

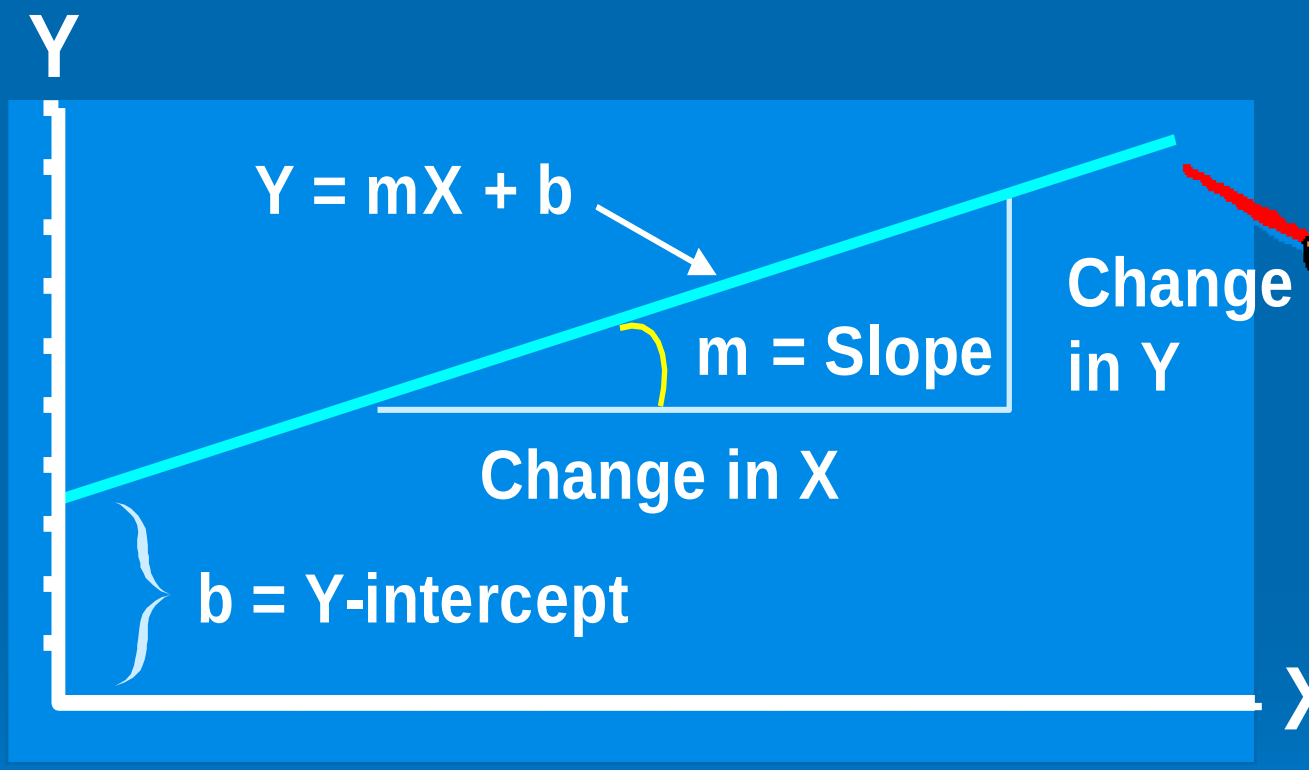
The Implicit Assumptions

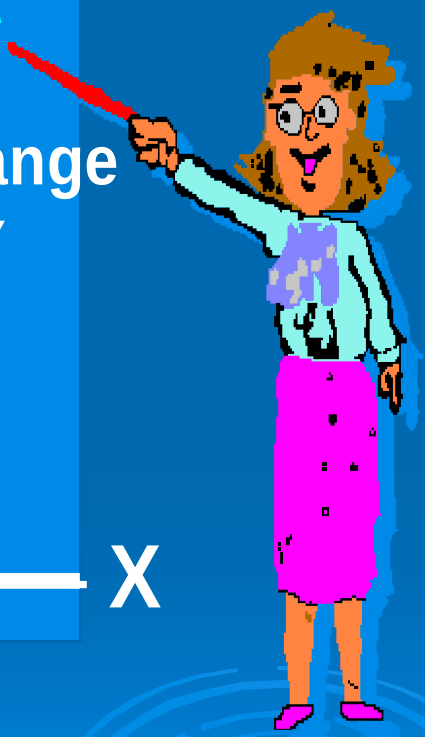
These assumptions aren't, but the specification of the model implies them. This is the way I've summarized them-they can be written with different terminology, of course.

1. All X are fixed and are measured without error
2. The model is linear in the parameters
3. The predictors and response are specified correctly
4. There is a single source of unmeasured random variance

If there is an assumption you've heard not on this list, chances are it is a logical extension of one of these core assumptions.

Linear Equations





Linear Regression Model

- 1. Relationship Between Variables Is a Linear Function

Population Y-Intercept

Population Slope

Random Error

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

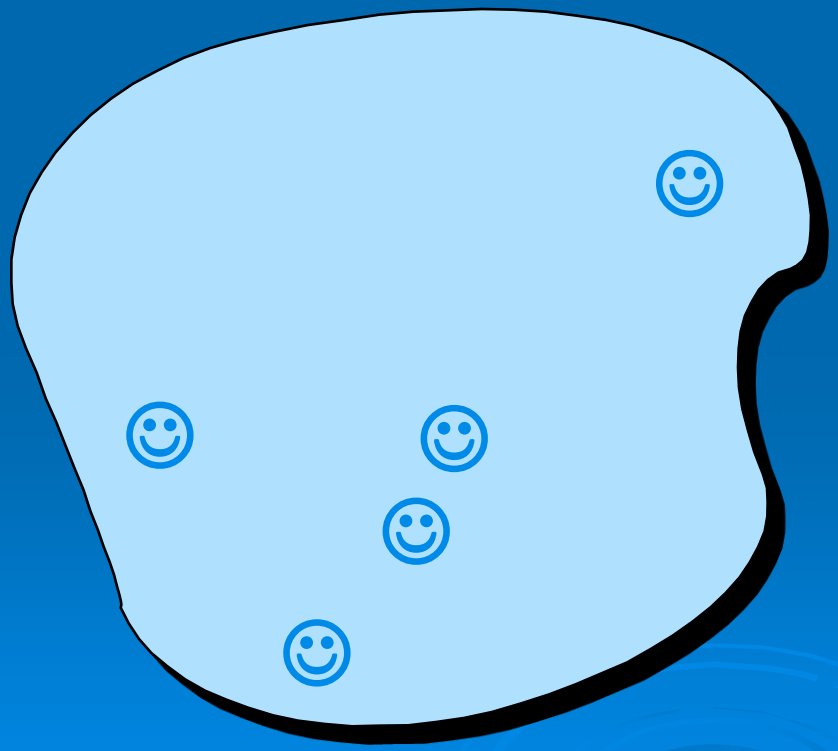
Dependent (Response) Variable (e.g., CD+ c.)

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

Population & Sample Regression Models

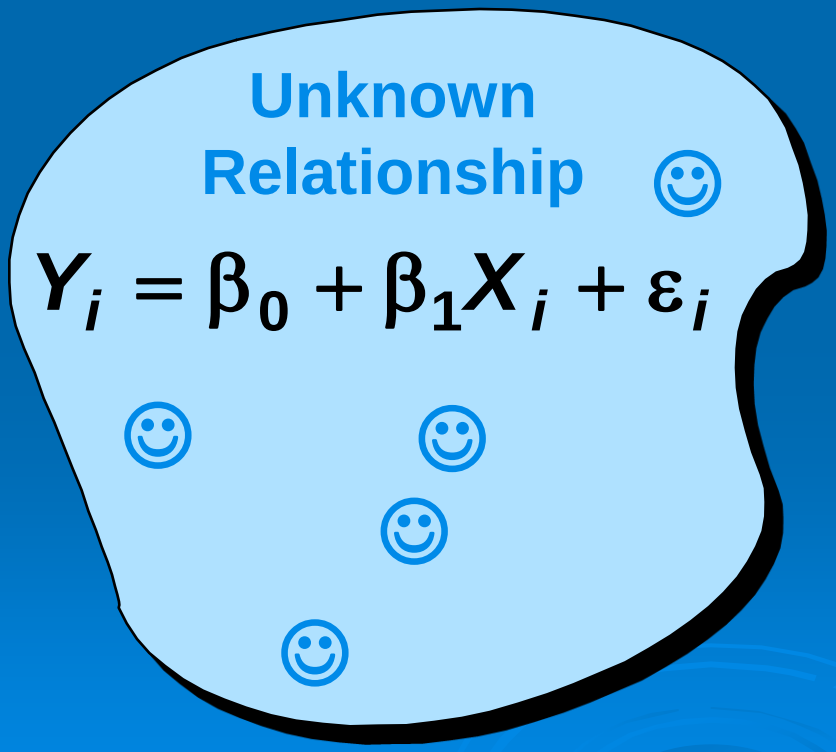
Population & Sample Regression Models

Population



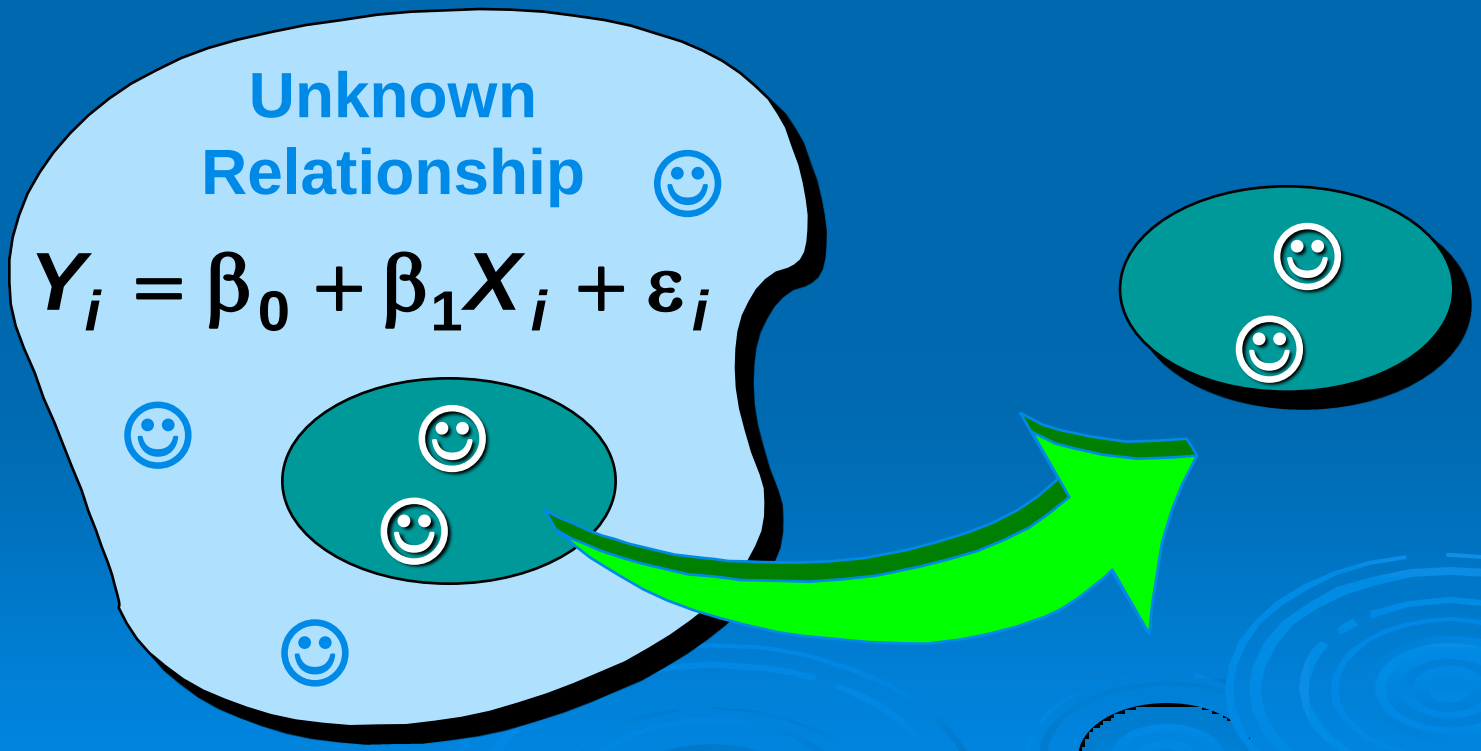
Population & Sample Regression Models

Population

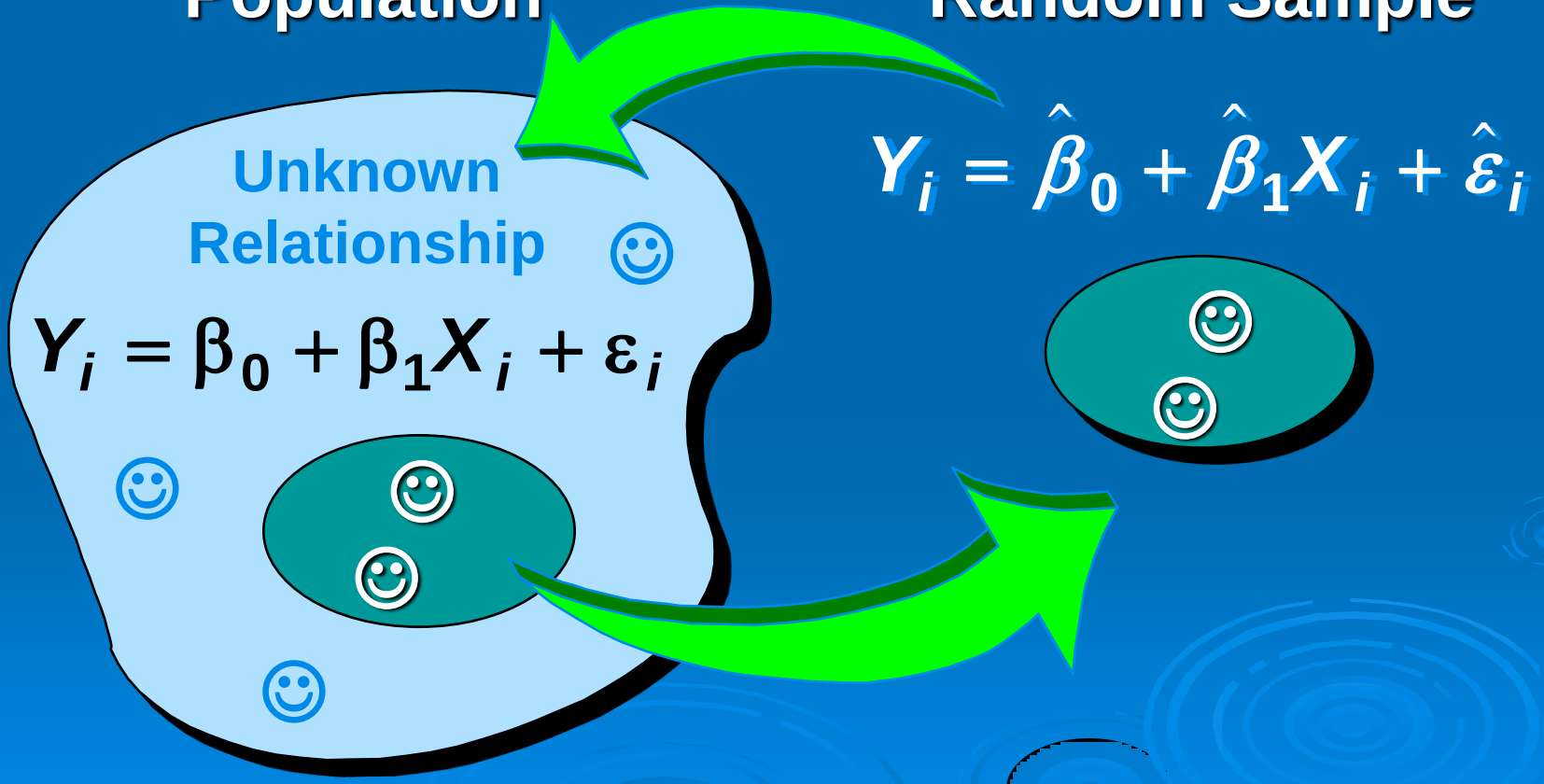


Population & Sample Regression Models

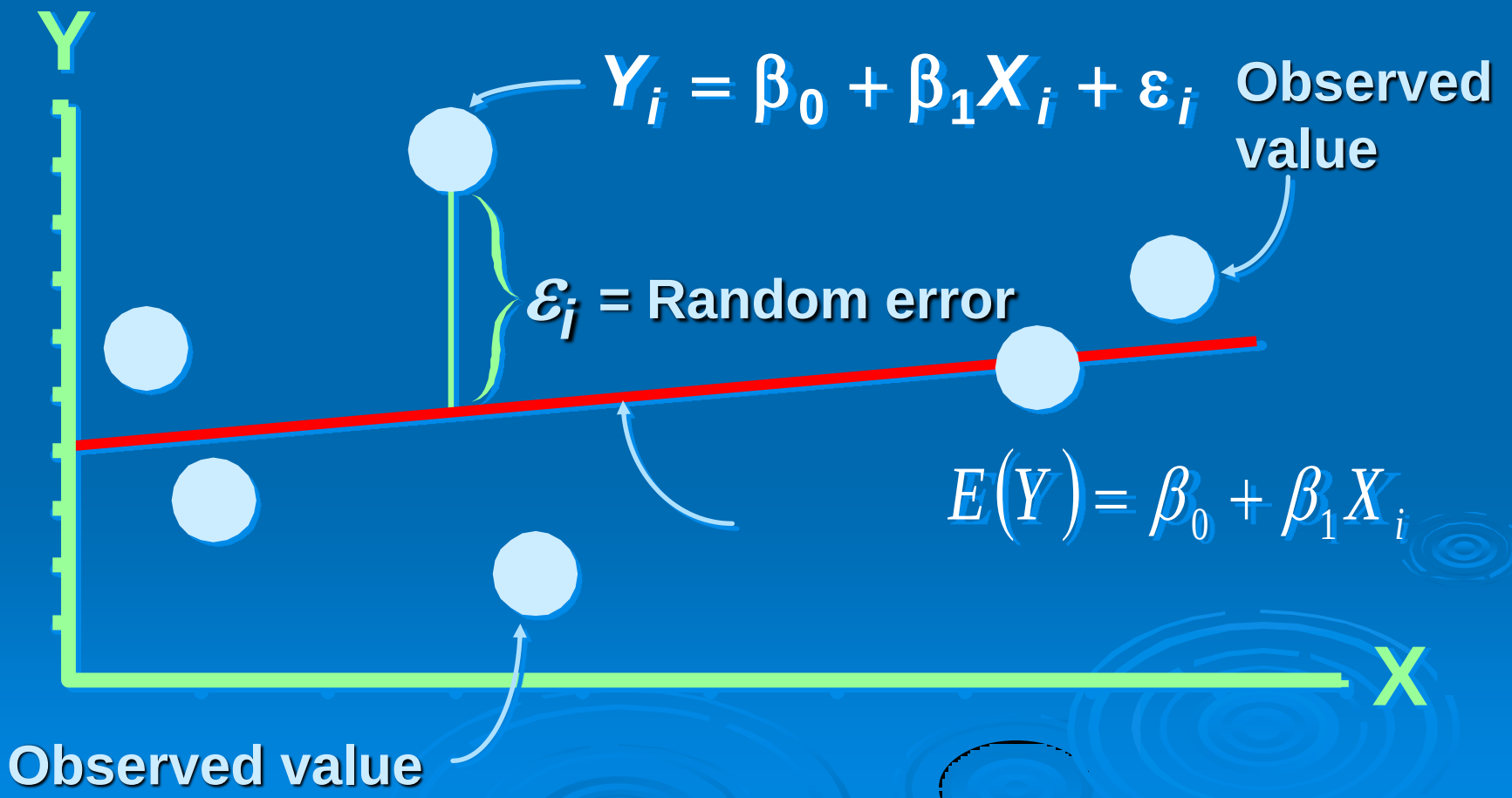
Population **Random Sample**



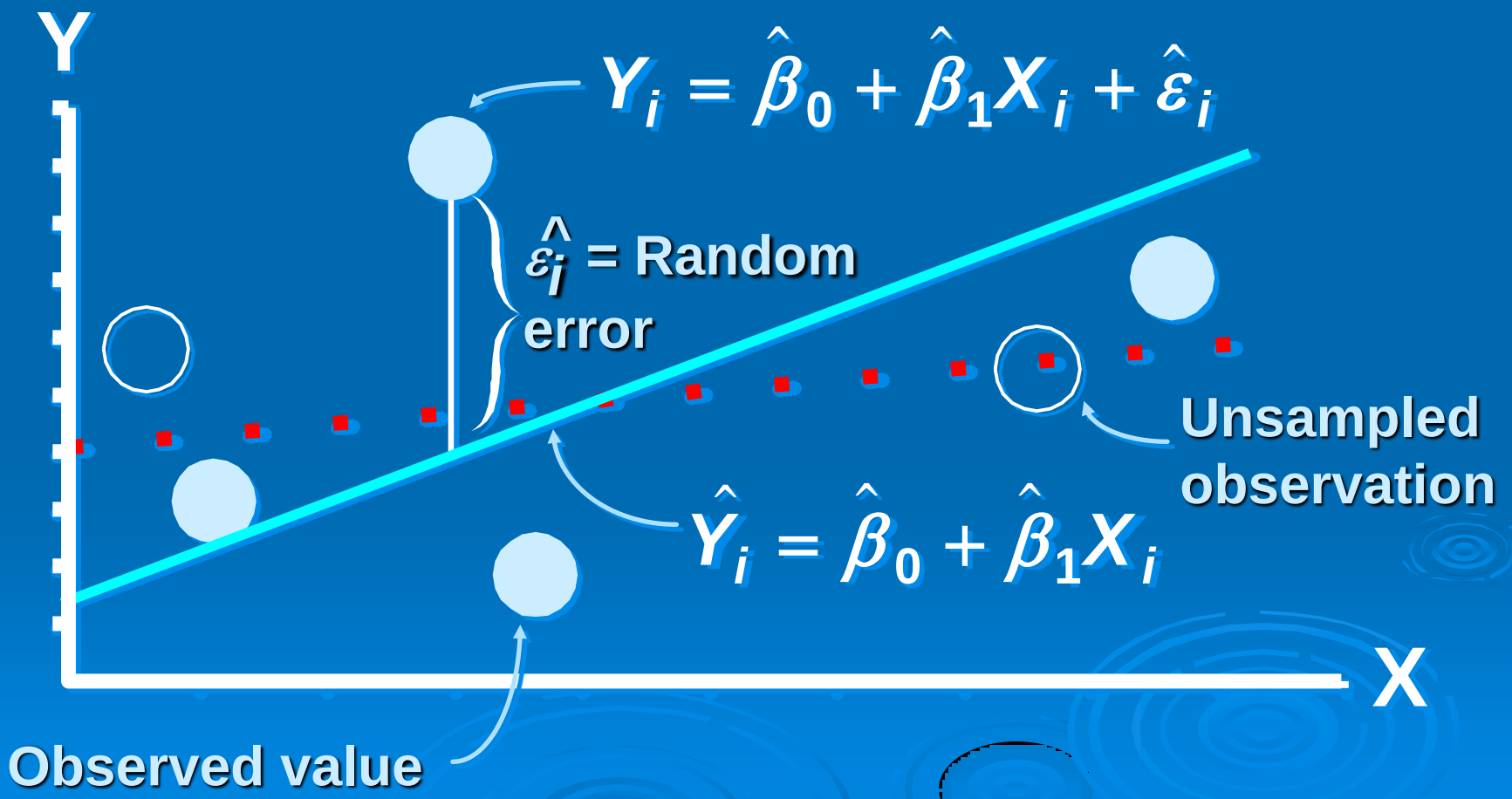
Population & Sample Regression Models



Population Linear Regression Model



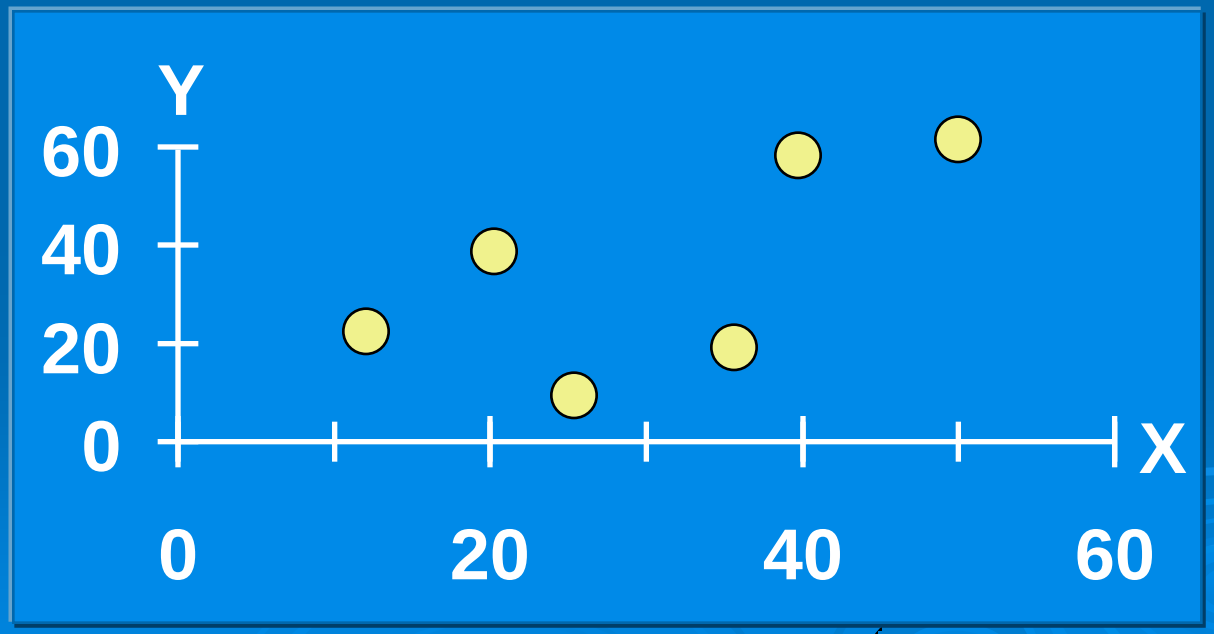
Sample Linear Regression Model



Estimating Parameters: Least Squares Method

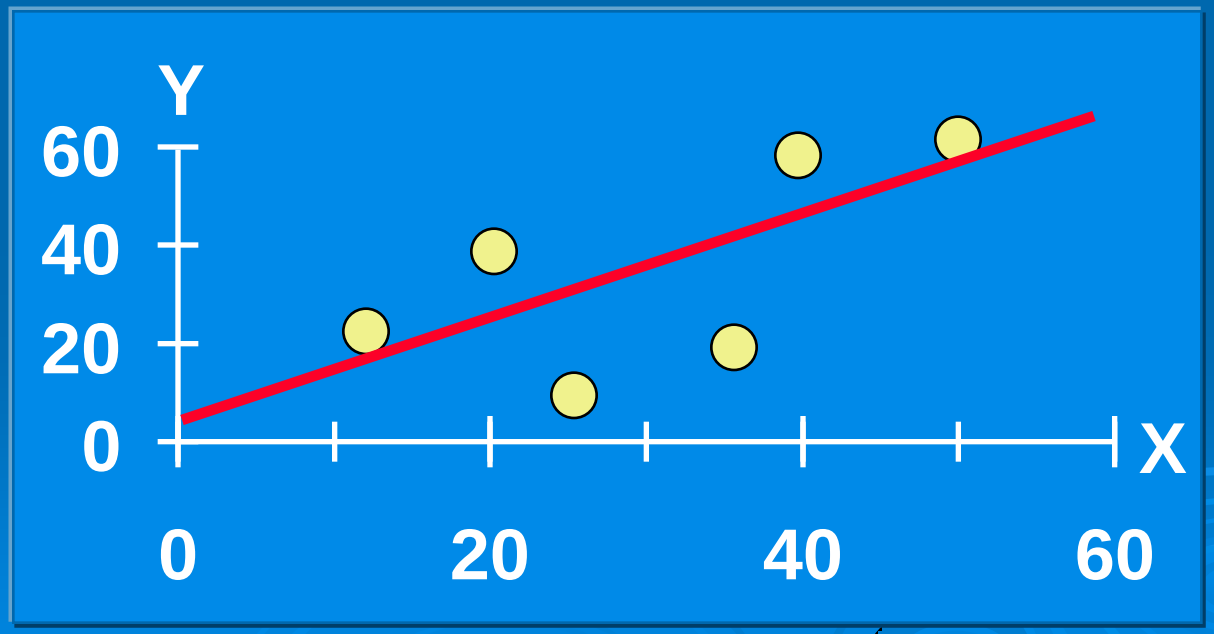
Scatter plot

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



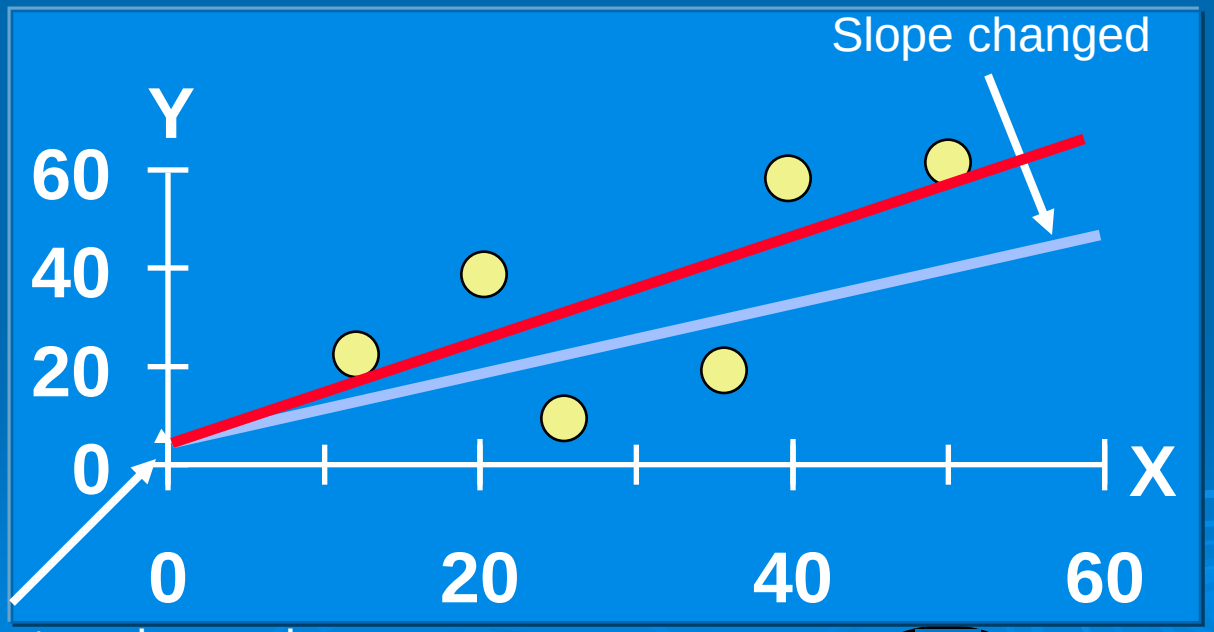
Thinking Challenge

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



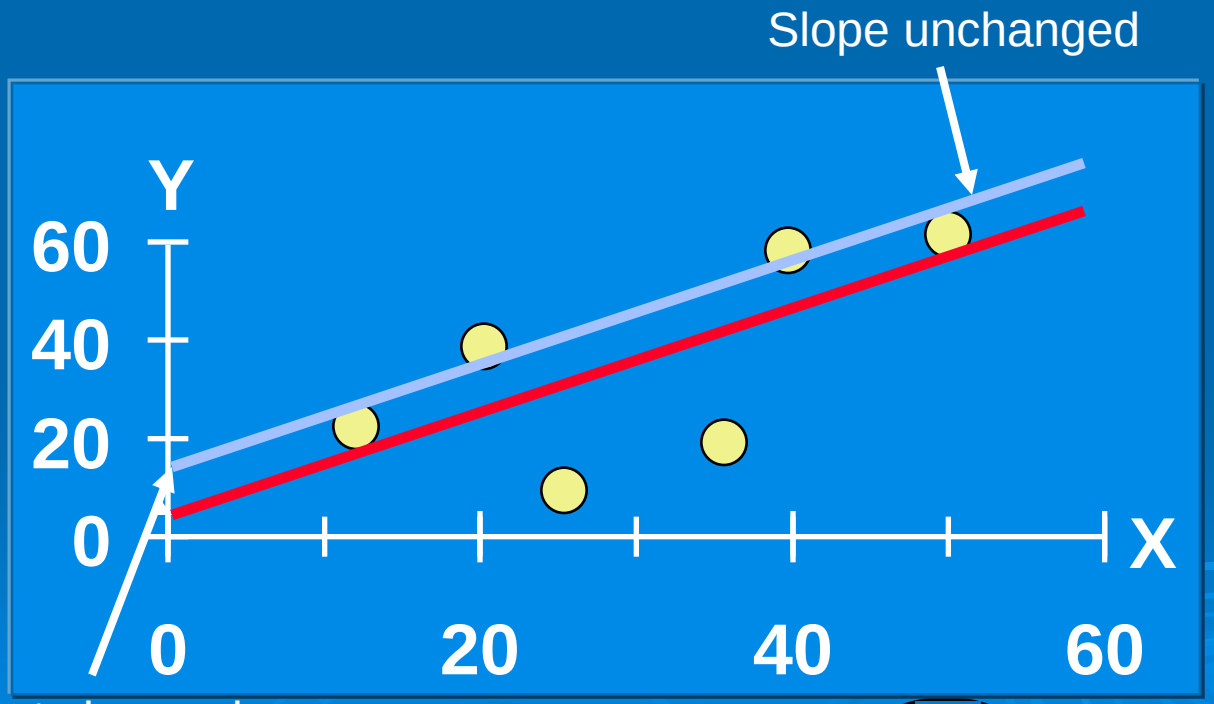
Thinking Challenge

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



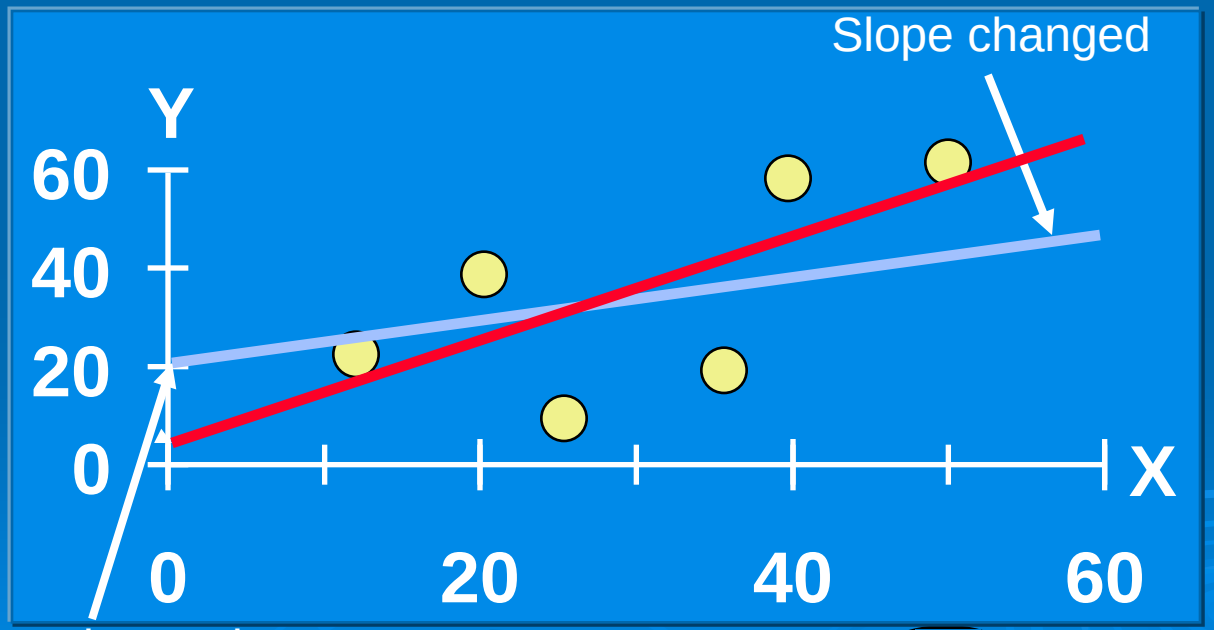
Thinking Challenge

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 ones

Least Squares

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

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

Least Squares

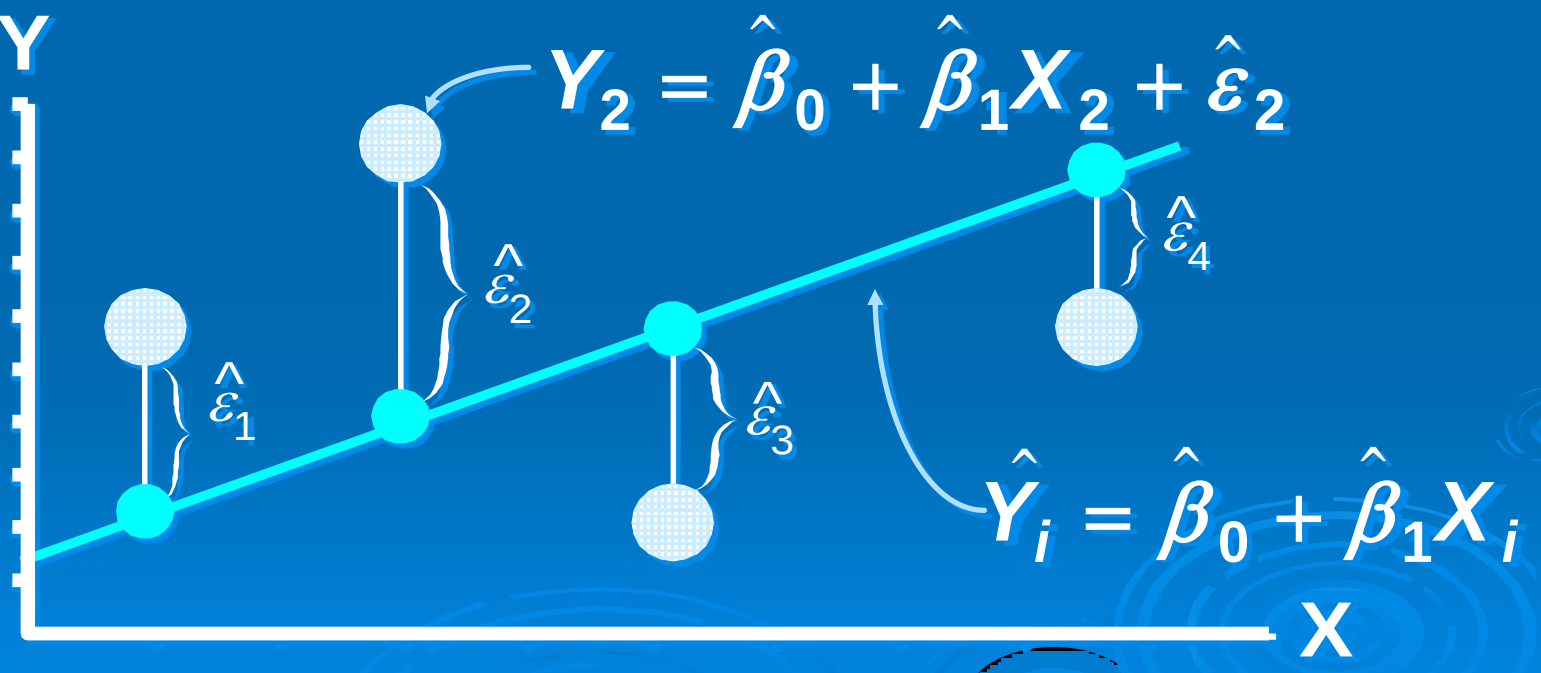
- 1. ‘Best Fit’ Means Difference Between Actual Y Values & Predicted Y Values Are a Minimum. *But* Positive Differences Off-Set Negative. So square errors!

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

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

Least Squares Graphically

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



Coefficient Equations

- Prediction equation

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- Sample slope

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

- Sample Y - intercept

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Derivation of Parameters (1)

- Least Squares (L-S):

  Minimize squared error

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i \right)^2$$

$$0 = \frac{\partial \sum \varepsilon_i^2}{\partial \beta_0} = \frac{\partial \sum \left(y_i - \beta_0 - \beta_1 x_i \right)^2}{\partial \beta_0}$$

$$= -2\left(n\bar{y} - n\beta_0 - n\beta_1 \bar{x} \right)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Derivation of Parameters (1)

- Least Squares (L-S):

  Minimize squared error

$$0 = \frac{\partial \sum \varepsilon_i^2}{\partial \beta_1} = \frac{\partial \sum (y_i - \beta_0 - \beta_1 x_i)^2}{\partial \beta_1}$$

$$= -2 \sum x_i (y_i - \beta_0 - \beta_1 x_i)$$

$$= -2 \sum x_i (y_i - \bar{y} + \beta_1 \bar{x} - \beta_1 x_i)$$

$$\beta_1 \sum x_i (x_i - \bar{x}) = \sum x_i (y_i - \bar{y})$$

$$\beta_1 \sum (x_i - \bar{x})(x_i - \bar{x}) = \sum (x_i - \bar{x})(y_i - \bar{y})$$

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$$

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

Interpretation of Coefficients

- 1. Slope ($\hat{\beta}_1$)
  - Estimated *Y* Changes by $\hat{\beta}_1$ for Each 1 Unit Increase in *X*
    - If $\hat{\beta}_1 = 2$, then *Y* Is Expected to Increase by 2 for Each 1 Unit Increase in *X*

Interpretation of Coefficients

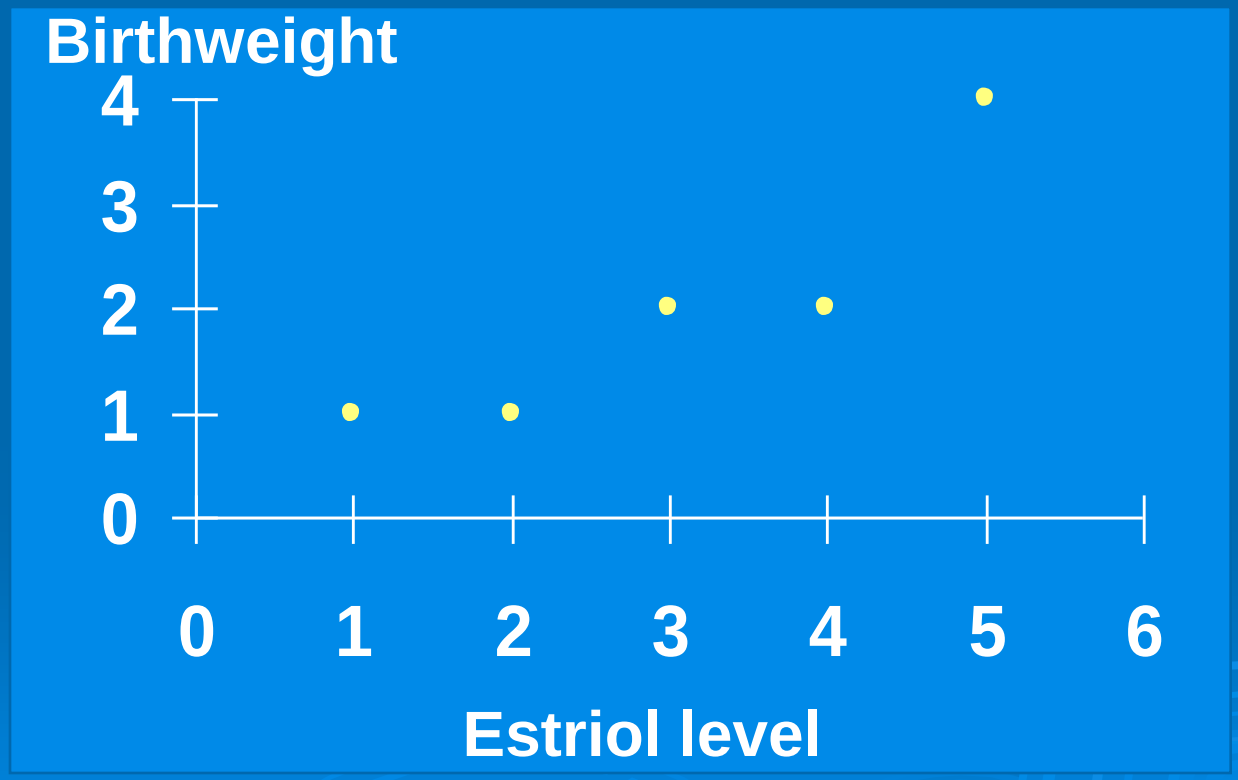
- 1. Slope ($\hat{\beta}_1$)
  - Estimated *Y* Changes by $\hat{\beta}_1$ for Each 1 Unit Increase in *X*
    - If $\hat{\beta}_1 = 2$, then *Y* Is Expected to Increase by 2 for Each 1 Unit Increase in *X*
- 2. Y-Intercept ($\hat{\beta}_0$)
  - Average Value of *Y* When *X* = 0
    - If $\hat{\beta}_0 = 4$, then Average *Y* Is Expected to Be 4 When *X* Is 0

Parameter Estimation Example

- Obstetrics: What is the **relationship** between Mother's Estriol level & Birthweight using the following data?

Estriol (mg/24h)	Birthweight (g/1000)
1	1
2	1
3	2
4	2
5	4

Scatterplot
Birthweight vs. Estriol level



Parameter Estimation Solution Table

X_i	Y_i	X_i^2	Y_i^2	X_iY_i
1	1	1	1	1
2	1	4	1	2
3	2	9	4	6
4	2	16	4	8
5	4	25	16	20
15	10	55	26	37

Parameter Estimation Solution

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n}}{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}} = \frac{37 - \frac{(15)(10)}{5}}{55 - \frac{(15)^2}{5}} = 0.70$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 2 - (0.70)(3) = -0.10$$

Coefficient Interpretation Solution

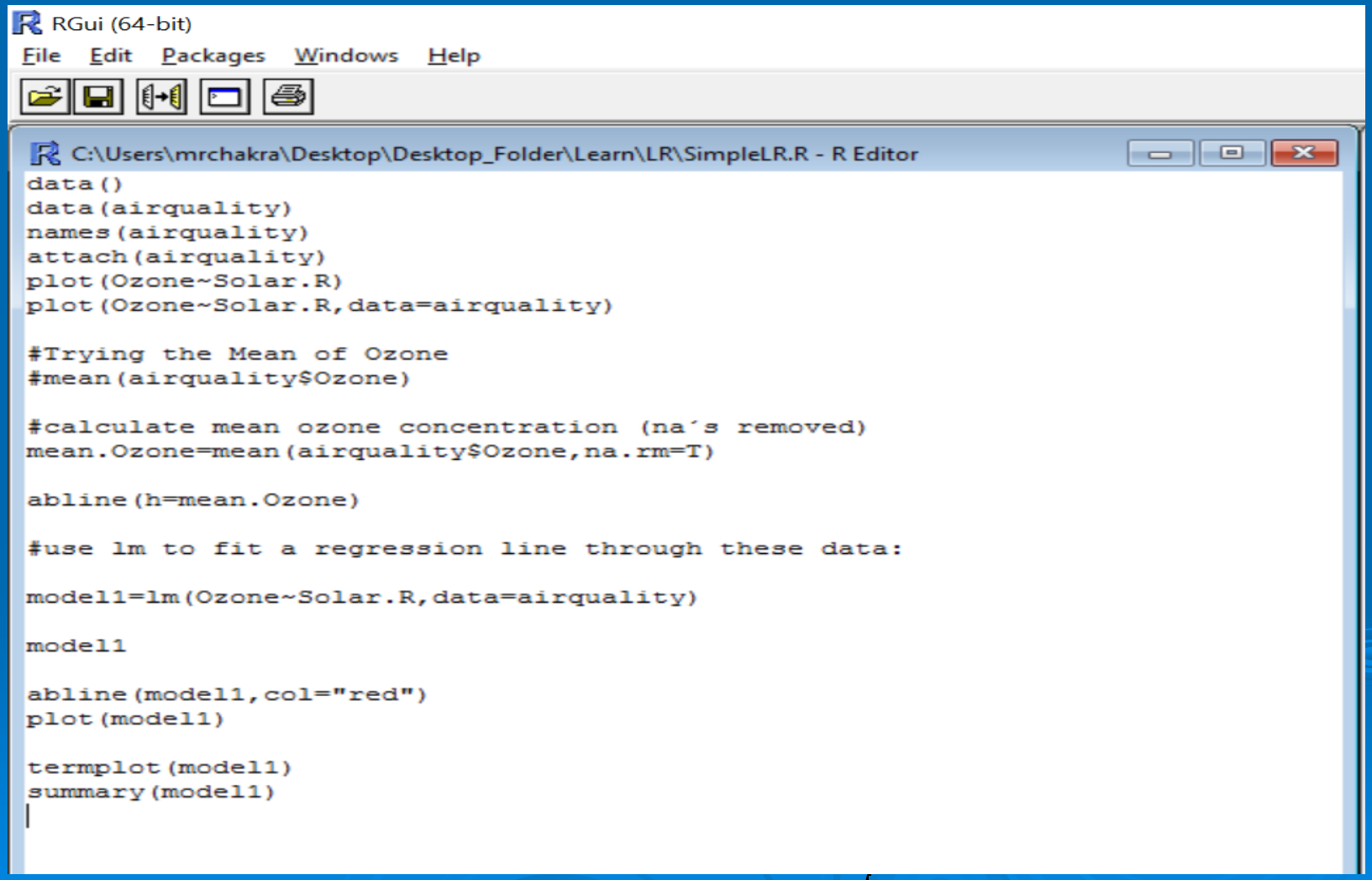
Coefficient Interpretation Solution

- 1. Slope ($\hat{\beta}_1$)
  - Birthweight (Y) Is Expected to Increase by .7 Units for Each 1 unit Increase in Estriol (X)

Coefficient Interpretation Solution

- 1. Slope ($\hat{\beta}_1$)
  - Birthweight (Y) Is Expected to Increase by .7 Units for Each 1 unit Increase in Estriol (X)
- 2. Intercept ($\hat{\beta}_0$)
  - Average Birthweight (Y) Is -.10 Units When Estriol level (X) Is 0
    - Difficult to explain
    - The birthweight should always be positive

Regression with R



Questions?