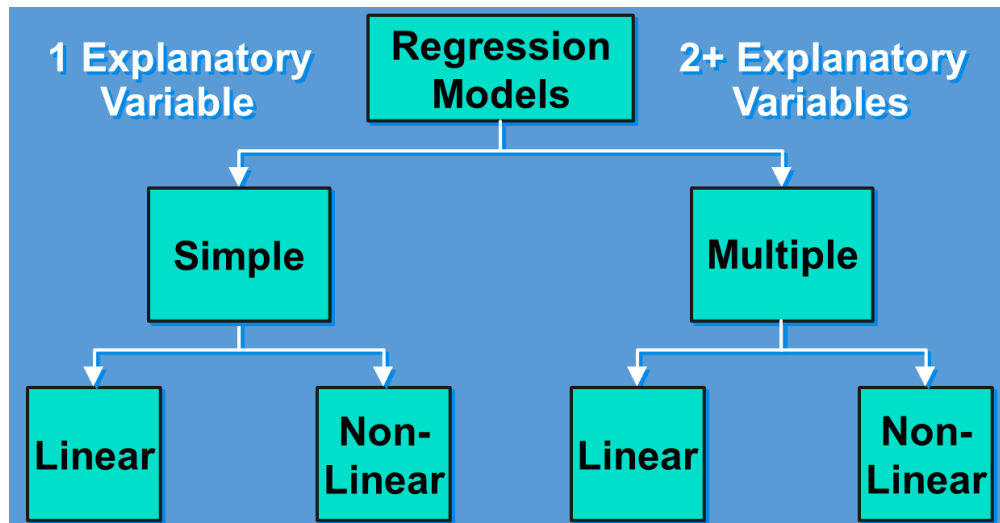


Types of Regression Models



Simple Regression

- Simple regression gives us the ability to estimate the mathematical relationship between a dependent variable (usually called y) and an independent variable (usually called x).
- The dependent variable is the variable for which we want to make a prediction
- Regression Analysis was first developed by Sir Francis Galton, who studied the relation between heights of sons and fathers.

Simple Linear Regression

Relationship between Variables is a Linear Function

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

Population Y-Intercept points to β_0

Population Slope points to β_1

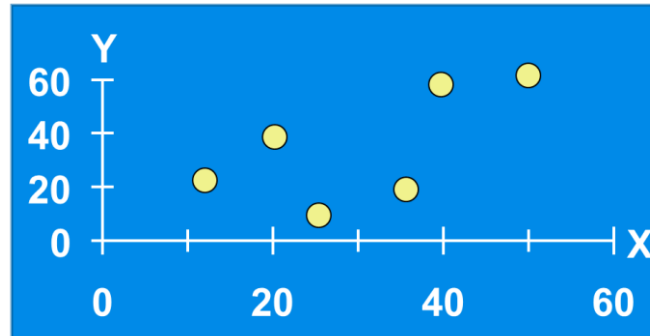
Random Error points to ε_i

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

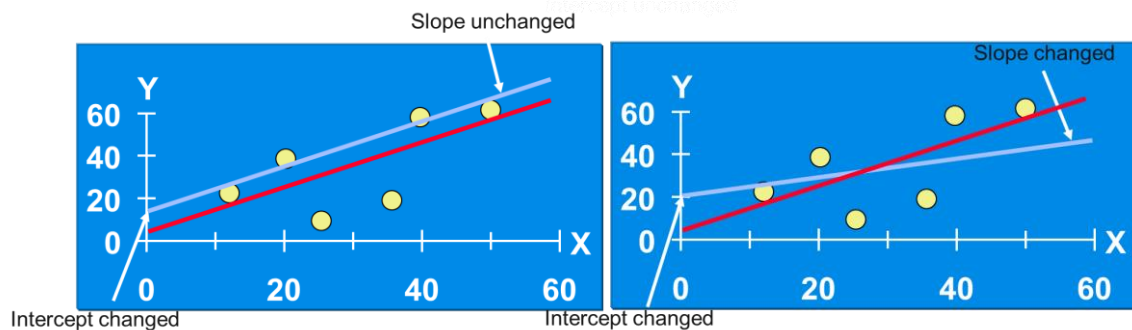
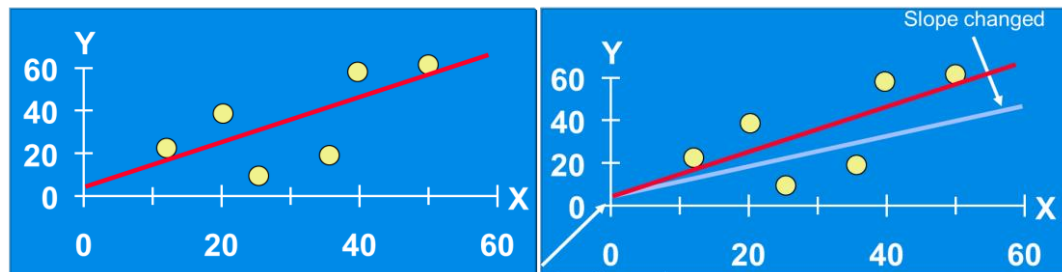
Independent (Explanatory) Variable (e.g., Years s. serocon.) points to X_i

Parameter Estimation of Simple Linear Regression Model using Least Squares Method

Plot of All (X_i, Y_i) Pairs



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

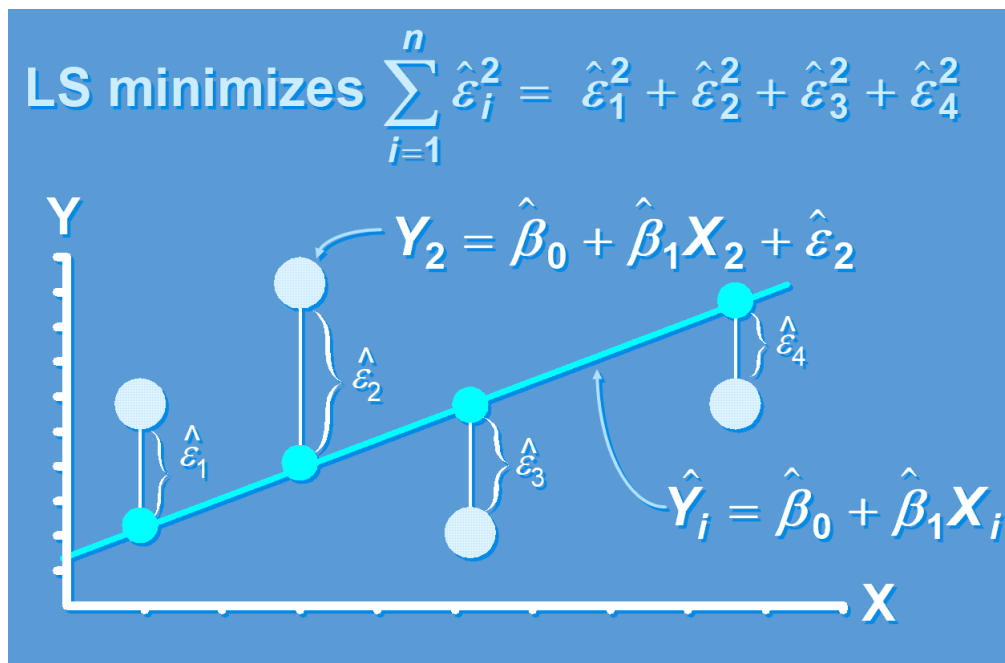


Intercept changed

Intercept changed

'Best Fit' Means Difference Between Actual Y Values & Predicted Y Values Are a Minimum. *But* Positive Differences Off-Set Negative ones. Hence, LS Minimizes the Sum of the Squared Differences (errors) (SSE)

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



➤ 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

Multiple Linear Regression

Relationship between 1 dependent & 2 or more independent variables is a linear function

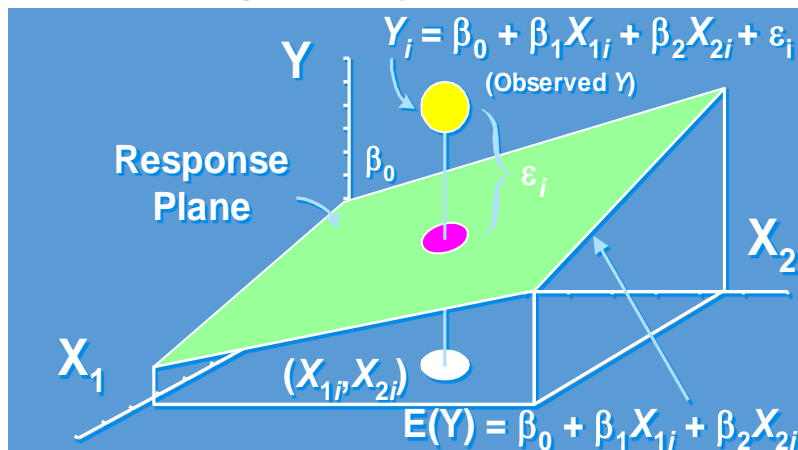
- Decide what you want to do and select the dependent variable
- List all potential independent variables for your model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \varepsilon_i$$

Diagram illustrating the components of the Multiple Linear Regression equation:

- Population Y-intercept:** β_0
- Population slopes:** $\beta_1, \beta_2, \dots, \beta_k$
- Random error:** ε_i
- Dependent (response) variable:** Y_i
- Independent (explanatory) variables:** $X_{1i}, X_{2i}, \dots, X_{ki}$

Parameter Estimation using Least Squares



1. Slope ($\hat{\beta}_k$)

- Estimated Y Changes by $\hat{\beta}_k$ for Each 1 Unit Increase in X_k **Holding All Other Variables Constant**
 - Example from textbook: If $\hat{\beta}_1 = 0.13$, then the systolic blood pressure (Y) Is Expected to Increase by 0.13 for Each 1 Unit Increase in birthweight (X_1) Given fixed age (X_2)

2. Y-Intercept ($\hat{\beta}_0$), predicted average value of Y When all X_k 's are set 0

- Proportion of Variation in Y 'Explained' by All X Variables **Taken Together**

$$R^2 = \frac{\text{Explained variation}}{\text{Total variation}} = \frac{SS_{yy} - SSE}{SS_{yy}} = 1 - \frac{SSE}{SS_{yy}}$$

SSE = Residual Error
SS_{yy} = Variance

- R^2 Never Decreases When New X Variable Is Added to Model (Disadvantage When Comparing Models)
- Solution: Adjusted R^2
 - Each additional variable reduces adjusted R^2 , unless SSE reduces enough to compensate

$$R_a^2 = 1 - \left[\frac{n-1}{n-(k+1)} \right] \frac{SSE}{SS_{yy}} \leq 1 - \frac{SSE}{SS_{yy}} = R^2$$