Regression: Introduction

Basic idea:

Use data to identify relationships among variables and use these relationships to make predictions.

Regression

- **Regression analysis** is a set of statistical processes for estimating the relationships among variables.
- Regression analysis describes the relationship between two (or more) variables.
- It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors').

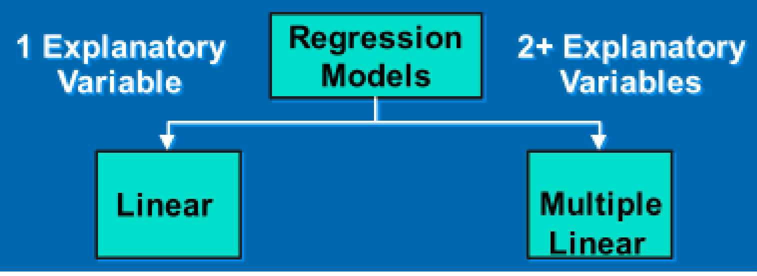
Regression Contd….

- Regression analysis helps one understand how the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed.

Regression Contd….

- Regression analysis is widely used for prediction and forecasting.
- Regression analysis is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships.
- Regression analysis can be used to infer causal relationships between the independent and dependent variables.

Types of Regression Models



Linear Regression

- Linear regression uses one independent variable to explain or predict the outcome of the dependent variable Y
- **Linear regression:** $Y = a + bX + e$
- where:
- Y = the variable that you are trying to predict (dependent variable).
- X = the variable that you are using to predict Y (independent variable).
- a = the intercept.
- b = the slope.
- e = the regression residual error

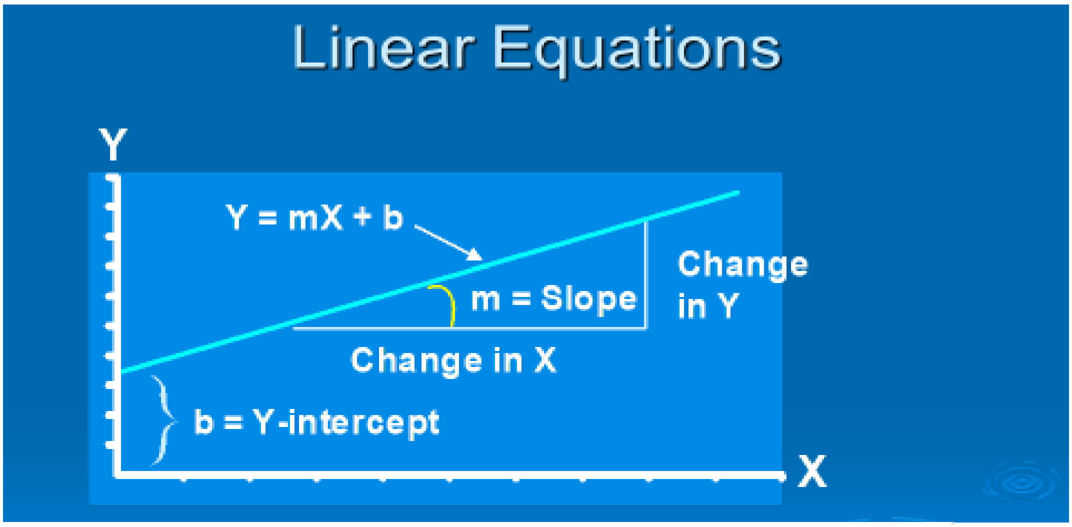
Multiple Linear Regression

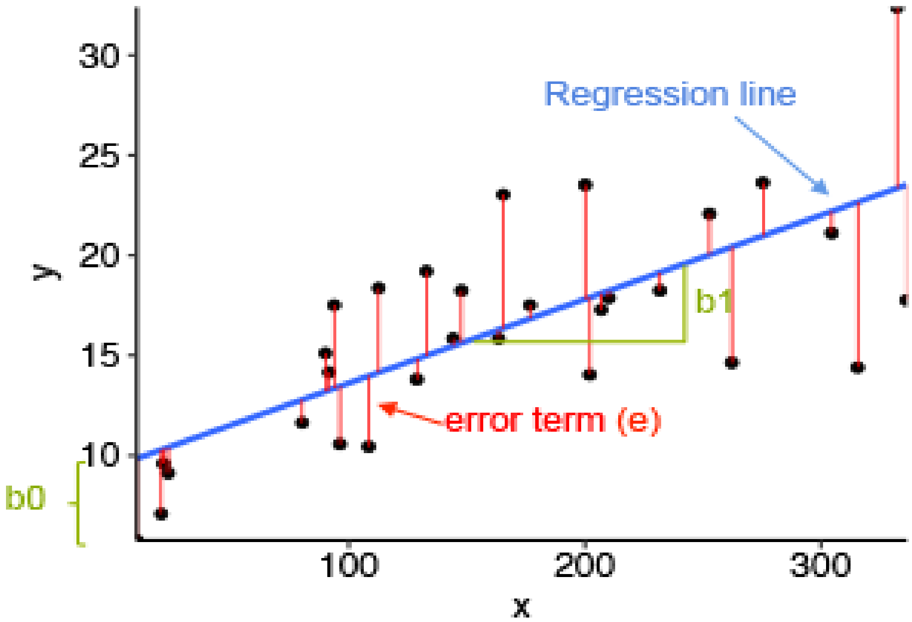
Multiple regression uses two or more independent variables to predict the outcome.

- **Multiple regression:** $Y = a + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_tX_t + e$
- Where:
- Y = the variable that you are trying to predict (dependent variable).
- X = the variable that you are using to predict Y (independent variable).
- a = the intercept.
- b = the slope.
- e = the regression residual error

Linear regression

- Linear dependence: constant rate of increase of one variable with respect to another (as opposed to, e.g., diminishing returns).
- Examples:
  - Income and educational level
  - Demand for electricity and the weather
  - Home sales and interest rates
- Our focus:
  - Gain some understanding of the mechanics.
    - the regression line
    - regression error
  - Learn how to interpret and use the results.
  - Learn how to setup a regression analysis.


Linear Equations
Y
Y = mX + b
m = Slope
Change in Y
Change in X
b = Y-intercept
X


Regression line
b1
error term (e)
b0
y
x
30
25
20
15
10
100
200
300

Linear Regression Model

- 1. Relationship Between Variables Is a Linear Function

Population Y-Intercept

Population Slope

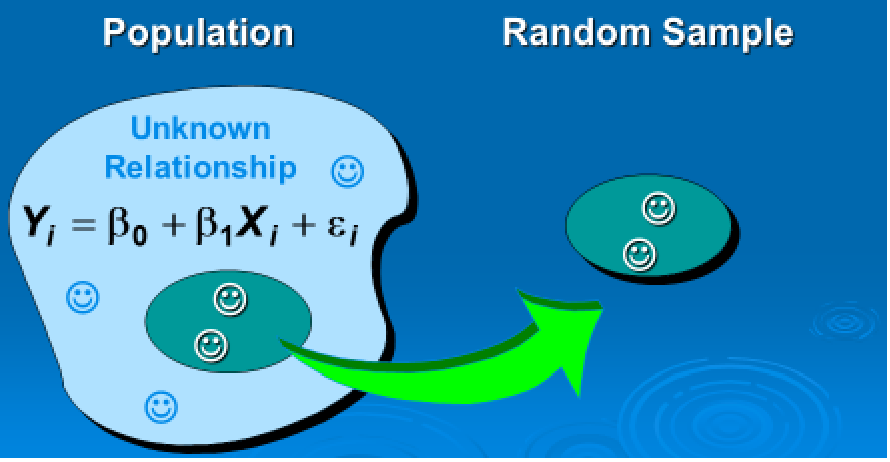
Random Error

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

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

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

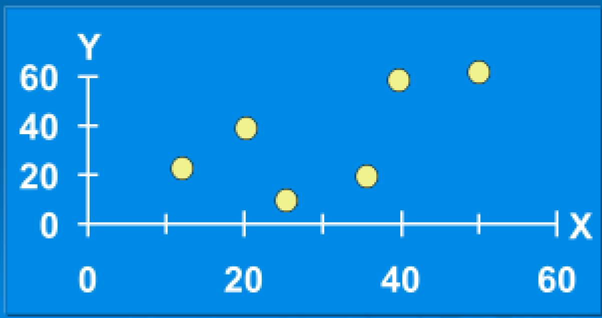
Population & Sample Regression Models



Estimating Parameters: Least Squares Method

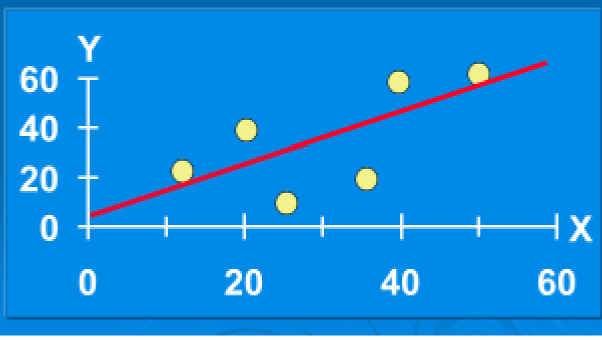
Scatter plot

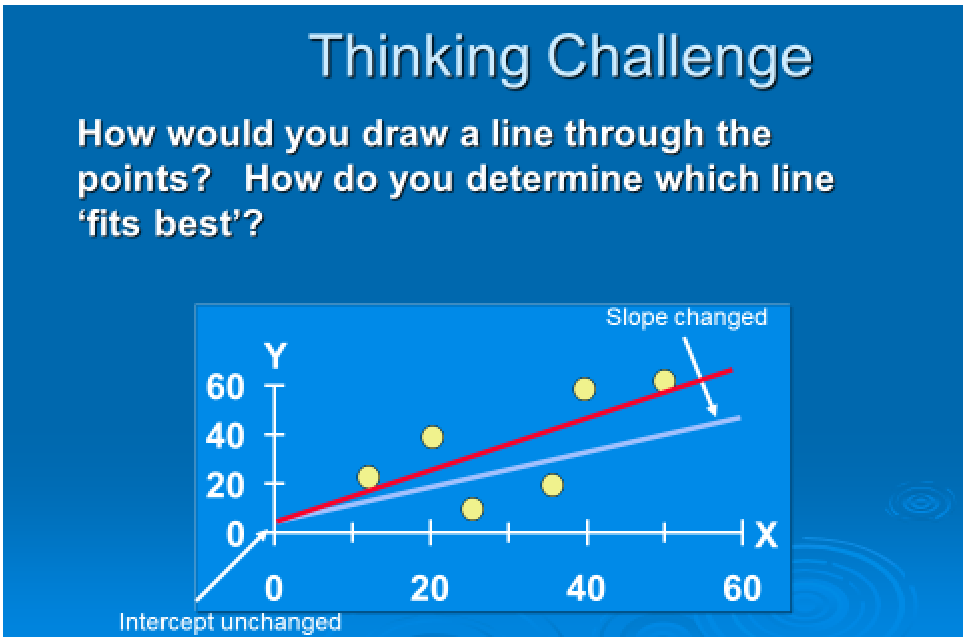
- 1. Plot of All (X_i, Y_i) Pairs
- 2. Suggests How Well Model Will Fit



Thinking Challenge

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

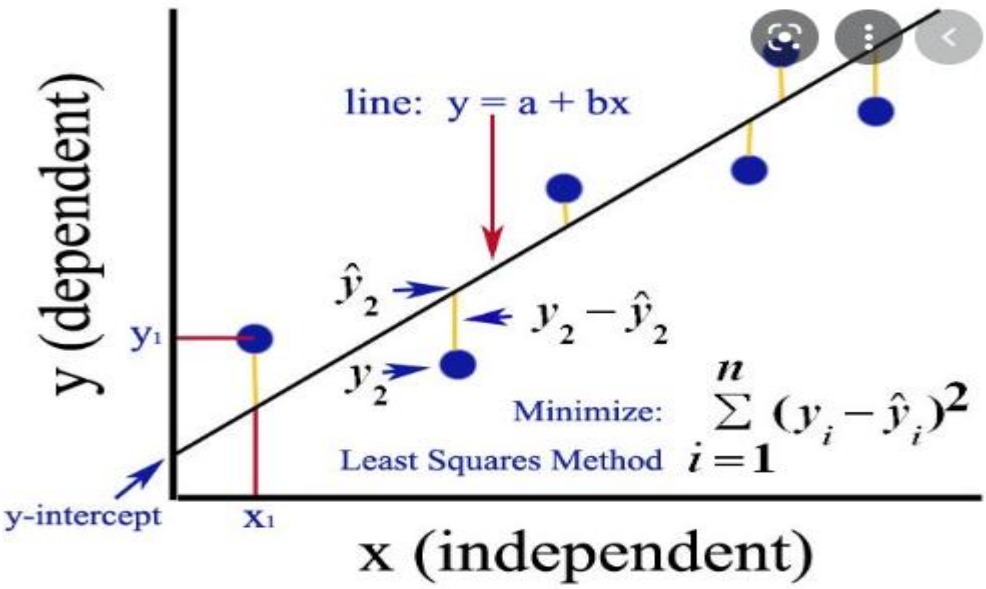


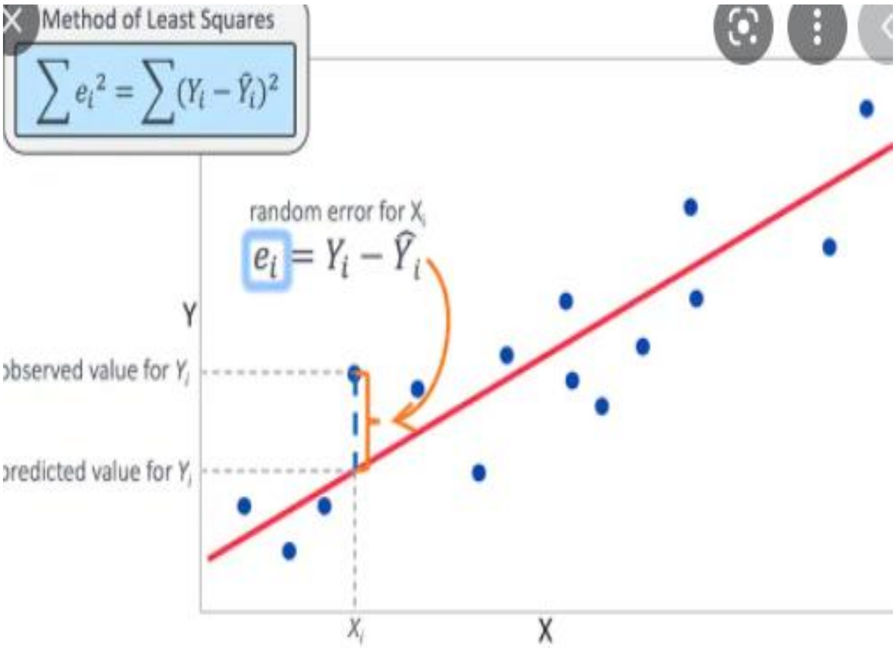


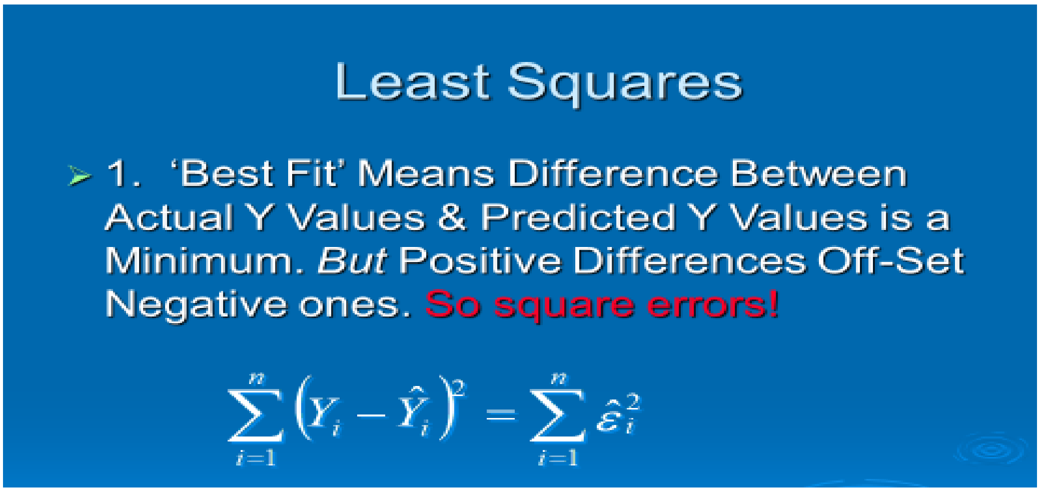
Ordinary Least Squares

Least Squares

- 1. 'Best Fit' Means Difference Between Actual Y Values & Predicted Y Values Are a Minimum.




Method of Least Squares
$\sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2$
random error for X_i
$e_i = Y_i - \hat{Y}_i$
Y
bserved value for Y_i
redicted value for Y_i
X_i
X

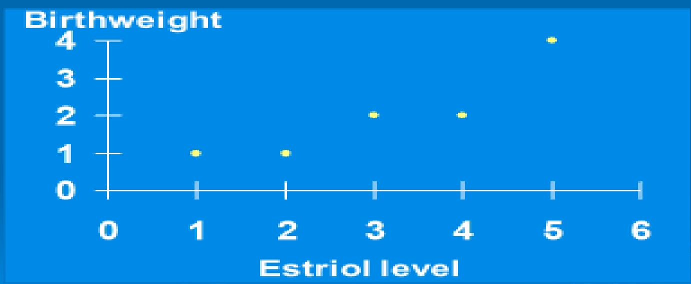

Least Squares
1. 'Best Fit' Means Difference Between Actual Y Values & Predicted Y Values is a Minimum. *But* Positive Differences Off-Set Negative ones. So square errors!
$\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2$

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

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

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