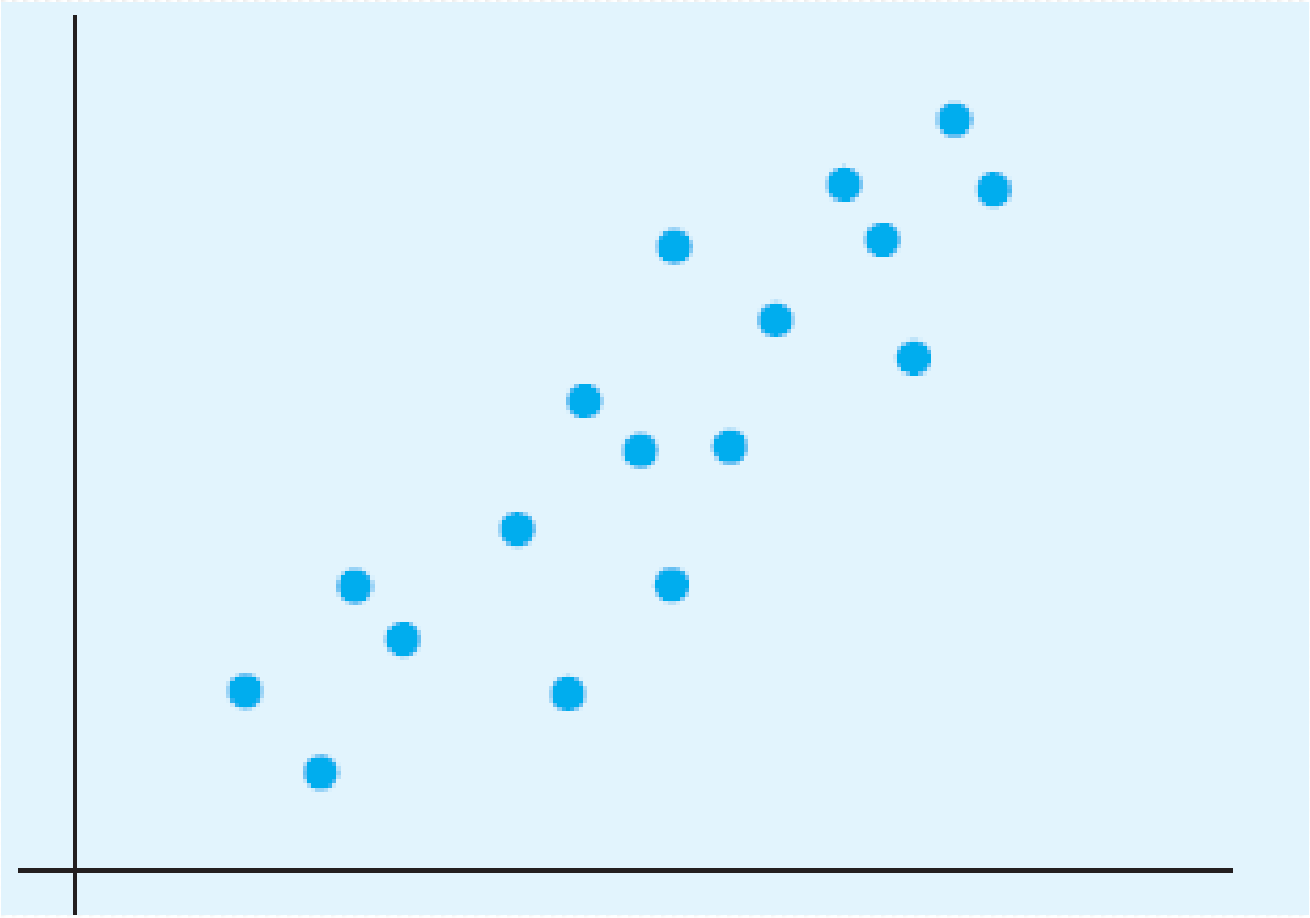
Regression

Dr Parag Chavda
Asst Prof
Community Medicine

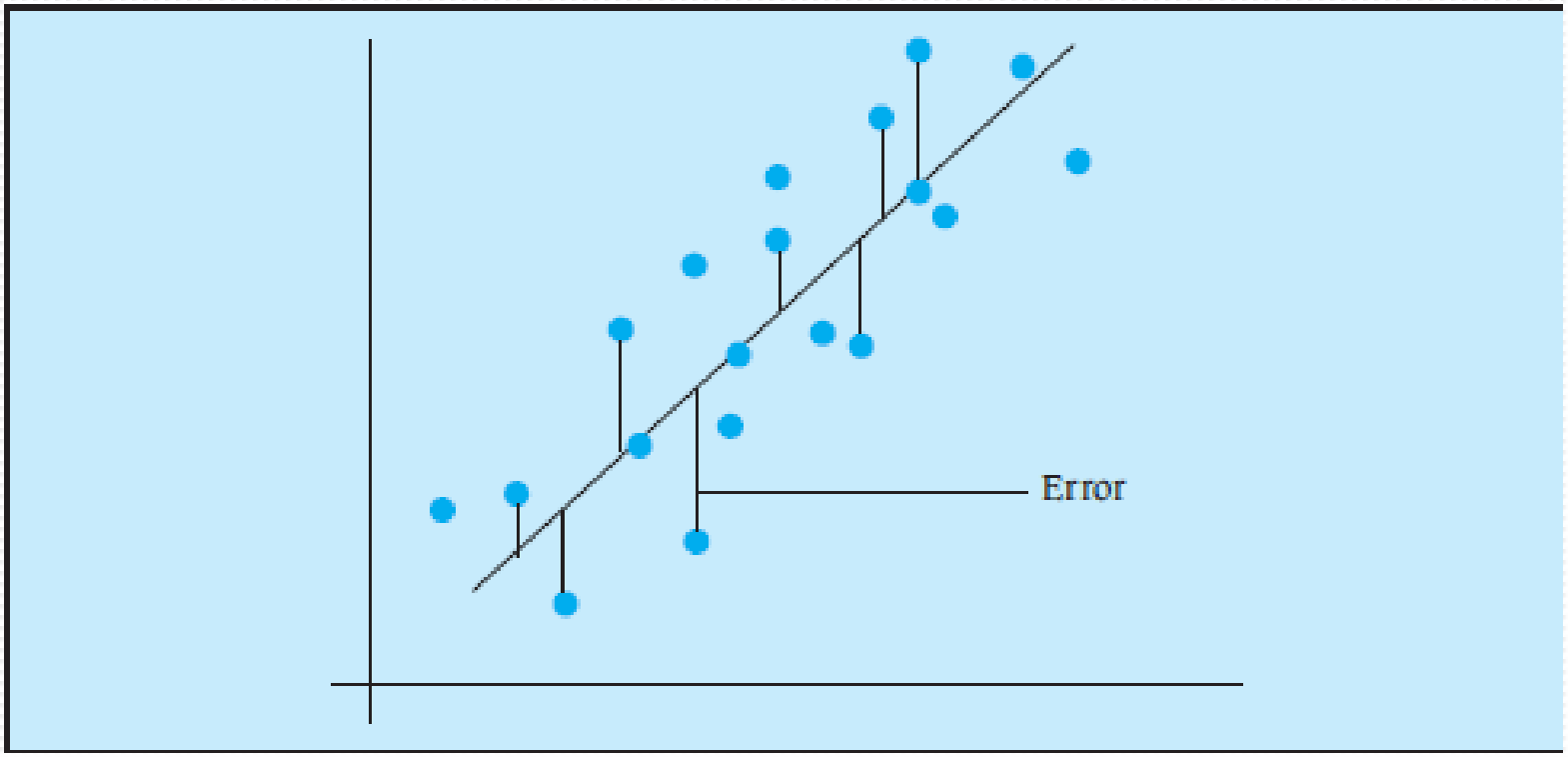
Correlation and Regression

- ‘Correlation’ indicates the relationship between two quantitative variables in which,
  - with changes in the values of one variable, the values in the other variable also changes.
- When the objective is to determine the strength of relationship between two such variables,
  - we use correlation coefficient (r).
- If the objective is to describe the existing relationship with a view of prediction,

Correlation



Regression line



$$y = ax + b$$

- Generally
  - ‘y’ is taken as dependent (outcome) variable and
  - ‘x’ as independent variable(predictor, explanatory)
- Regression analysis determine the form of the relationship by a line which best fits the data - called as ‘Regression Equation’.
- Equation of the straight line: Y=a+bx
  - ‘a’ is the intercept and
  - ‘b’ is the slope of the line which measures the amount of increase or change in y for unit change in x

- ‘a’ and ‘b’ can be estimated using the following formula

$$a = \bar{y} - b\bar{x} \text{ and } b = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

- Many times the researcher may come across situations where he is dealing with more than one independent variables.
- For example, the researcher may be interested in finding out as to how much change is likely to occur in serum cholesterol level (outcome or dependent variable) following changes in age, body weight, alcohol intake, calories consumed per day, and calories spent per day in physical exercise.

- Dependant variable
  - serum cholesterol level
- Independent variables are
  - age,
  - body weight
  - alcohol intake
  - calories consumed per day
  - physical exercise (calories spent per day)
- Thus, we have a set of 5 independent variables, all of which are likely to affect the outcome variable by themselves.

- $y = a + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_n x_n$

- Such situations where more than one independent variable and one dependent variable are available and the aim is to predict dependent variable on the basis of many independent variables, the technique used for analysis is called as '*Multivariate Regression Analysis'*

Types of multiple regression analysis

- Multiple linear regression
- Multiple logistic regression
- Cox regression model

Multiple linear regression

- One dependant variable
- Multiple independent variables

- Why linear?
- outcome (dependent) variable which is measured on a “Numerical continuous” or “Numerical discrete” Scale

Multiple linear regression – Our Example

- One dependant variable – serum cholesterol
- Multiple independent variables – 5 variables

- Why linear?
- outcome (dependent) variable which is measured on a “Numerical continuous” or “Numerical discrete” Scale
- Serum cholesterol – measured in mg/dl is

Multiple linear regression

- Analysis usually done on computers
- ß coefficient indicates the change in the dependent variable for the unit change in independent variable.
- For example, if ß for weight is 2.1, it means that
  - for every 1 kilogram increase in weight,
  - the serum cholesterol will increase by 2.1 mg/dl.

Multiple linear regression

- Computer analysis provides
  - ß coefficient
  - 95% confidence interval,
  - significance value for each variable and
  - regression equation.

Multiple Logistic Regression Model

- Applied when the outcome variable is measured on a "dichotomous scale" of
  - "either having a disease (cases) or not having that disease (controls)"
  - "either survived or died"
- the independent (predictor or exposure) variables, which are at least two or more, are measured either on numerical continuous/ discrete or categorical dichotomous scale
- Logistic regression is widely used in medical science

- The beta coefficients (ß) in multiple logistic regression represent the “natural logarithm of the risk of the outcome due to that particular independent variable, following one unit change in the independent variable, or else as compared to the baseline category”
- This estimate is adjusted for the confounders
- Exponential of this beta coefficient is

- Example: if the beta coefficient for age is 0.037; then the odds ratio is 1.04.
- The interpretation of OR is that it represents the independent, adjusted estimate of risk of the outcome, due to “particular level” of the independent variable, as compared to its baseline category

Cox Regression Model

- The outcome variable
  - Continuous or discrete – multiple linear regression
  - Dichotomous – multiple logistic
  - Survival data - Cox Regression Model