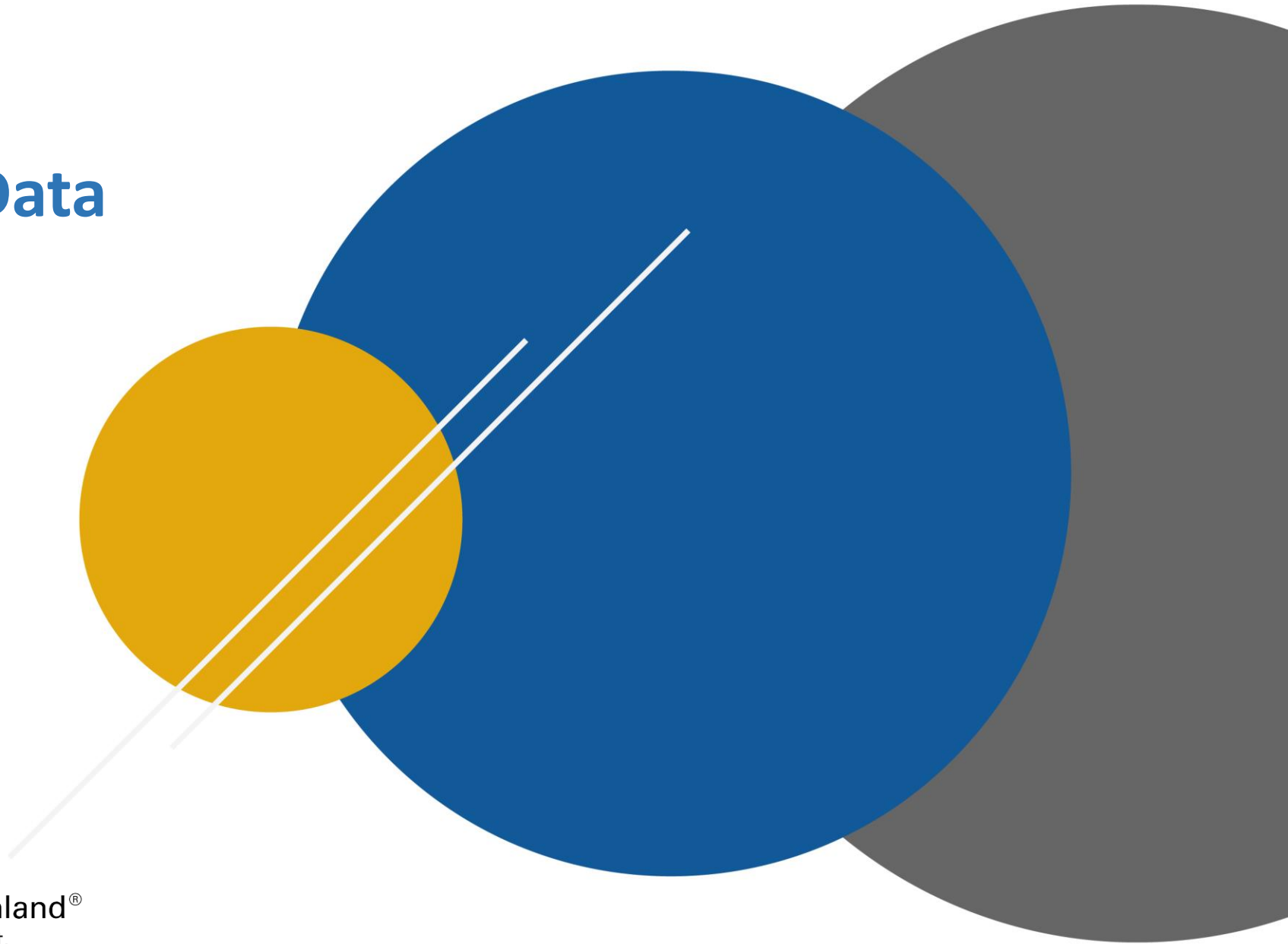


Structured Data (Data Mining Model) : Regression Model



Agenda

- **Introduction to Regression Analysis**
- Steps in Regression Analysis
- Simple Linier Regression
- Multiple Linier Regression
- Assumptions of Linier Regression Models
- Introduction to Logistic Regression Analysis

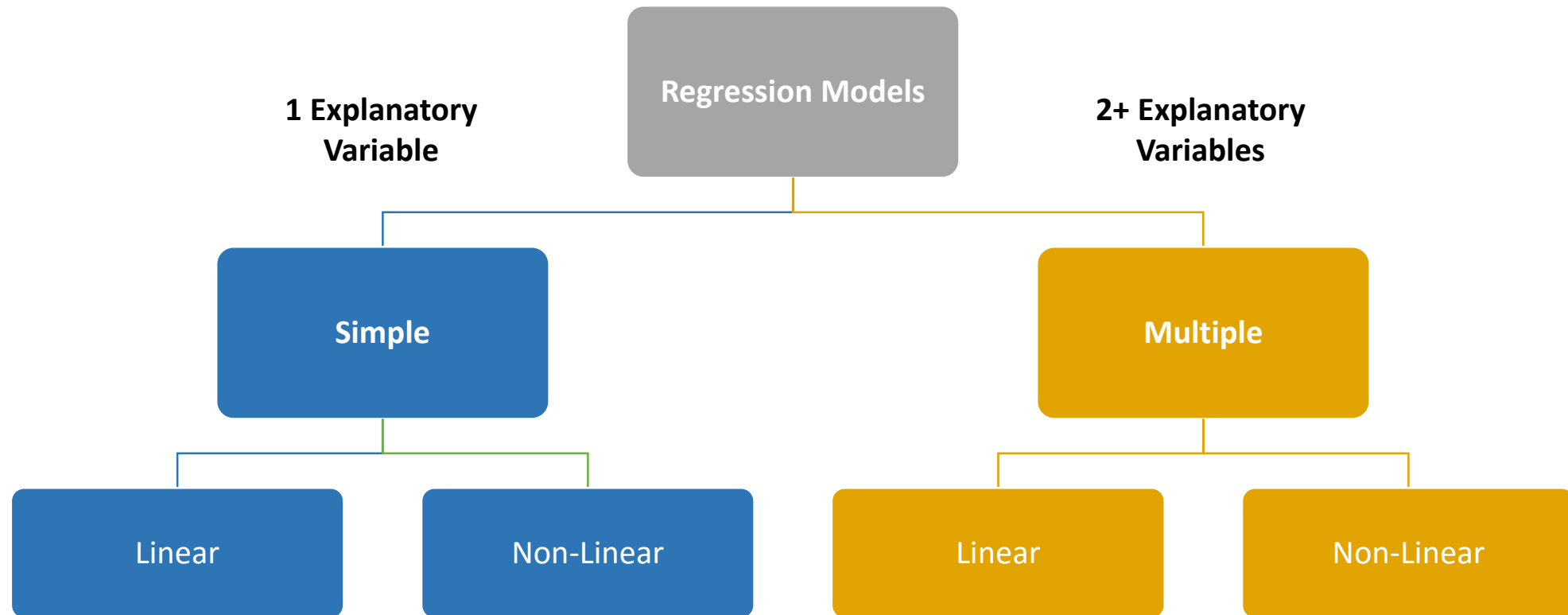


An Introduction to Regression Analysis

- Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable (s) (predictor). Regression analysis is an important tool for modelling and analyzing data.
- Regression analysis is used to:
 - Predict the value of a dependent variable based on the value of at least one independent variable
 - Explain the impact of changes in an independent variable on the dependent variable
- **Dependent variable:** the variable we wish to predict or explain
- **Independent variable:** the variable used to explain the dependent variable



Types of Regression Model



An Introduction to Regression Analysis

Regression analysis is used to model the relationship between single variables Y , we call response, output or the dependent variable with one or more predictors, inputs or explanatory variables:

- $X=1$: Simple Linear Regression Analysis
- $X>1$: Multiple Linear Regression Analysis or multivariate Regression
- $Y>1$: multivariate regression.
- Independent Variable is continuous variable
- Dependent Variable is continuous, Discrete and categorical variable



Agenda

- Introduction to Regression Analysis
- **Steps in Regression Analysis**
- Simple Linier Regression
- Multiple Linier Regression
- Assumptions of Linier Regression Models
- Introduction to Logistic Regression Analysis



Steps in Regression Analysis

A regression analysis can be broken down into 5 steps.

- Step 1: state the hypothesis.
- Step 2: test the hypothesis (estimate the relationship).
- Step 3: interpret the test results. This step would enable us to answer the following questions,
- Step 4: check for and correct common problems of regression analysis.
- Step 5: evaluate the test results.

Let us explain each step one by one.



Agenda

- Introduction to Regression Analysis
- Steps in Regression Analysis?
- **Simple Linier Regression**
- Multiple Linier Regression
- Assumptions of Linier Regression Models
- Introduction to Logistic Regression Analysis



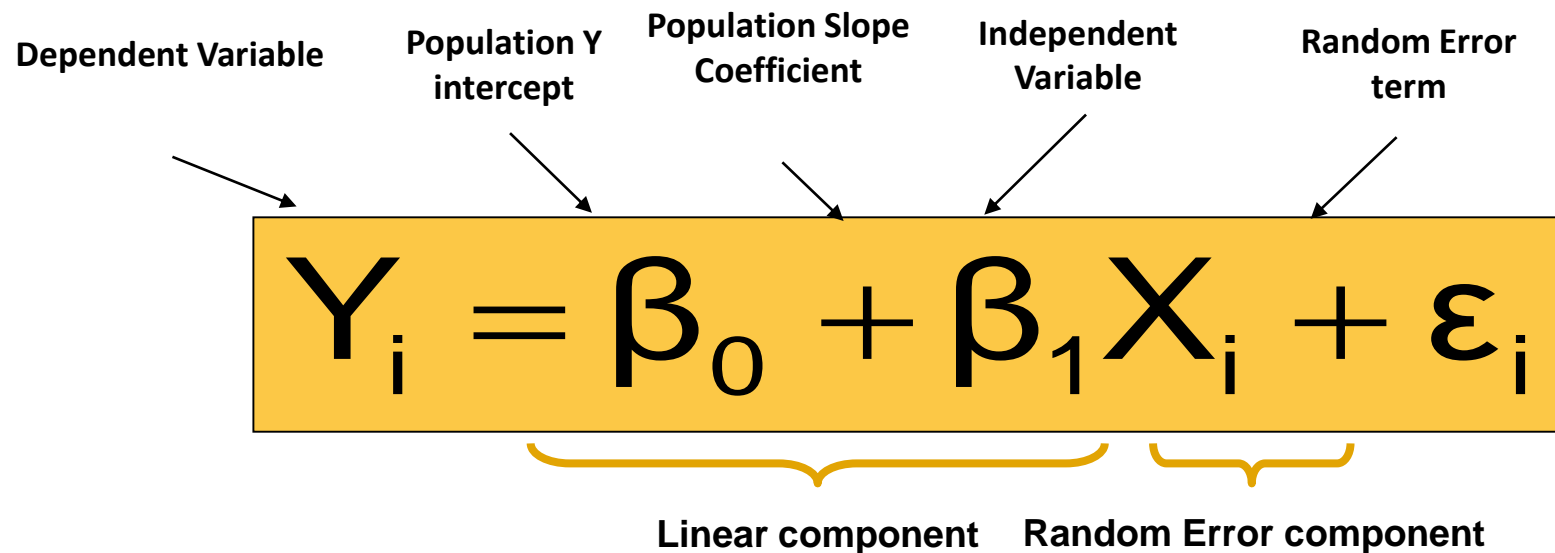
Simple Linear Regression Model

- Only one independent variable, X Relationship between X and Y is described by a linear function.
- Changes in Y are assumed to be caused by changes in X? Relationship between variables is a linear function

Dependent Variable Population Y intercept Population Slope Coefficient Independent Variable Random Error term

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

Linear component Random Error component

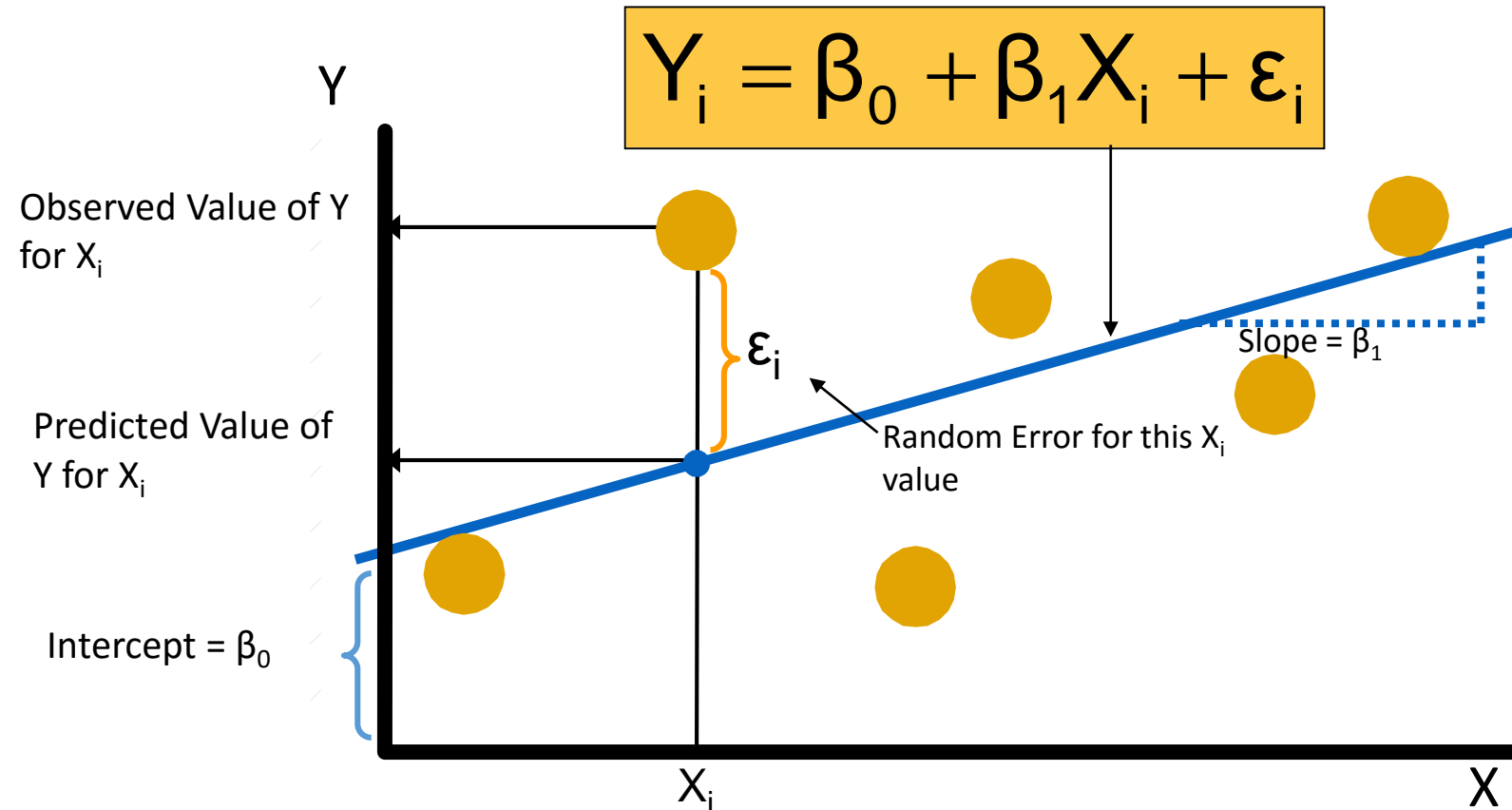


The diagram illustrates the Simple Linear Regression Model equation: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$. The equation is presented in a yellow box. Above the box, five labels with arrows point to their respective parts: 'Dependent Variable' points to Y_i , 'Population Y intercept' points to β_0 , 'Population Slope Coefficient' points to β_1 , 'Independent Variable' points to X_i , and 'Random Error term' points to ε_i . Below the box, two curly braces group the terms: the first brace under $\beta_0 + \beta_1 X_i$ is labeled 'Linear component', and the second brace under ε_i is labeled 'Random Error component'.



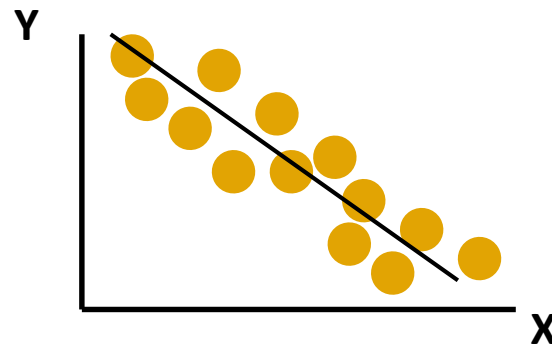
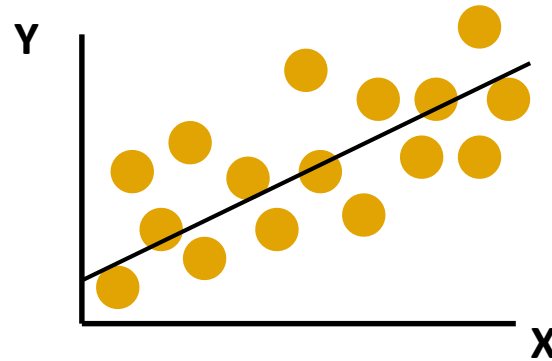
Simple Linear Regression Model

(continued)

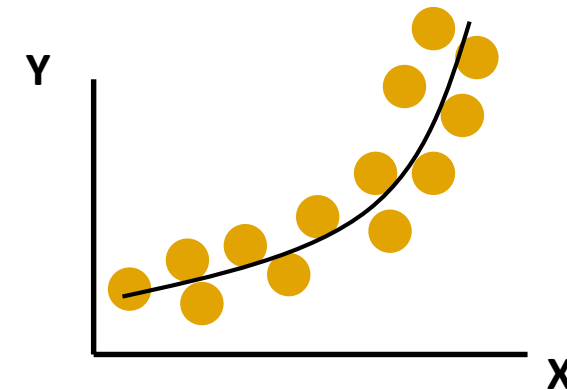
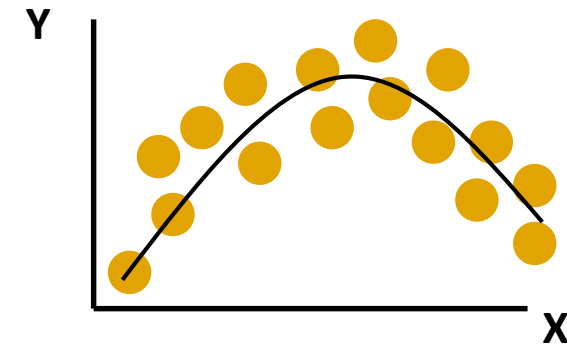


Types of Relationship

Linear relationships



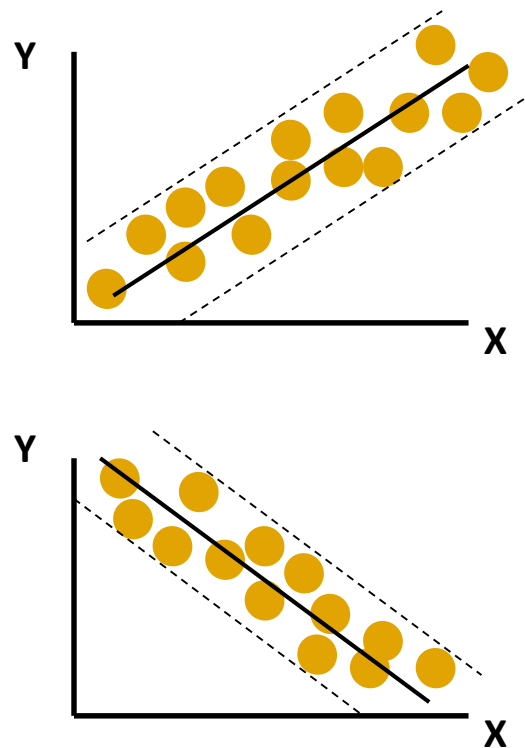
Curvilinear relationships



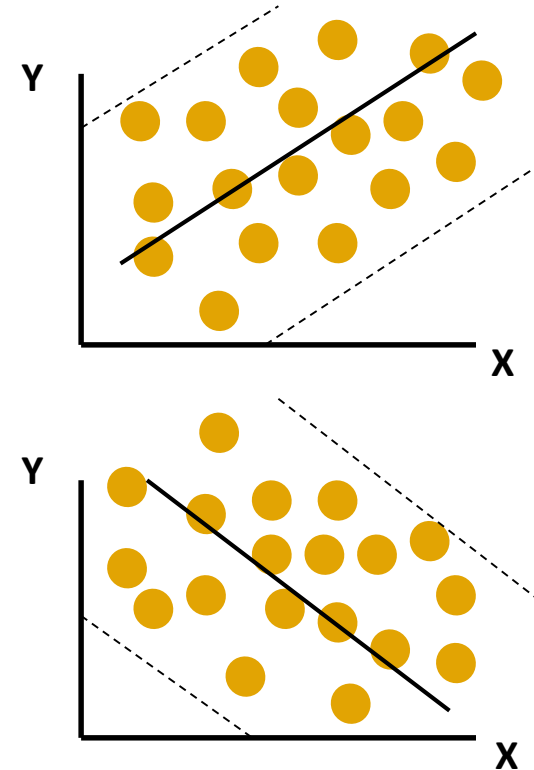
Types of Relationship

(continued)

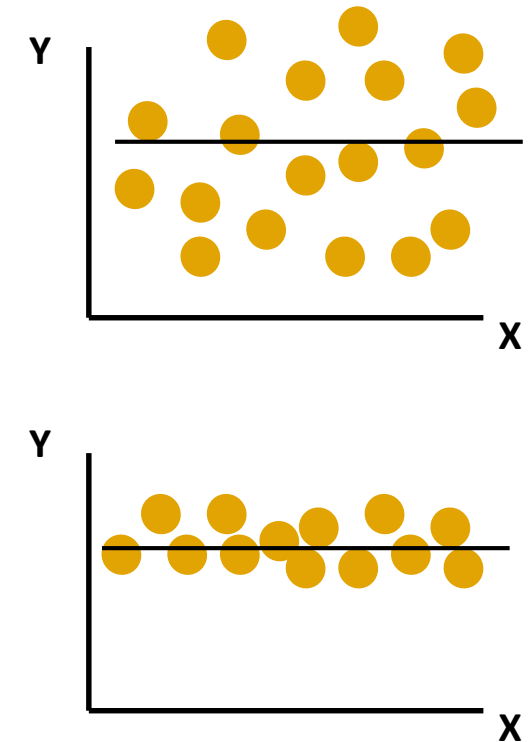
Strong relationships



Weak relationships



No relationship



Coefficient Equations

Prediction Equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Slope

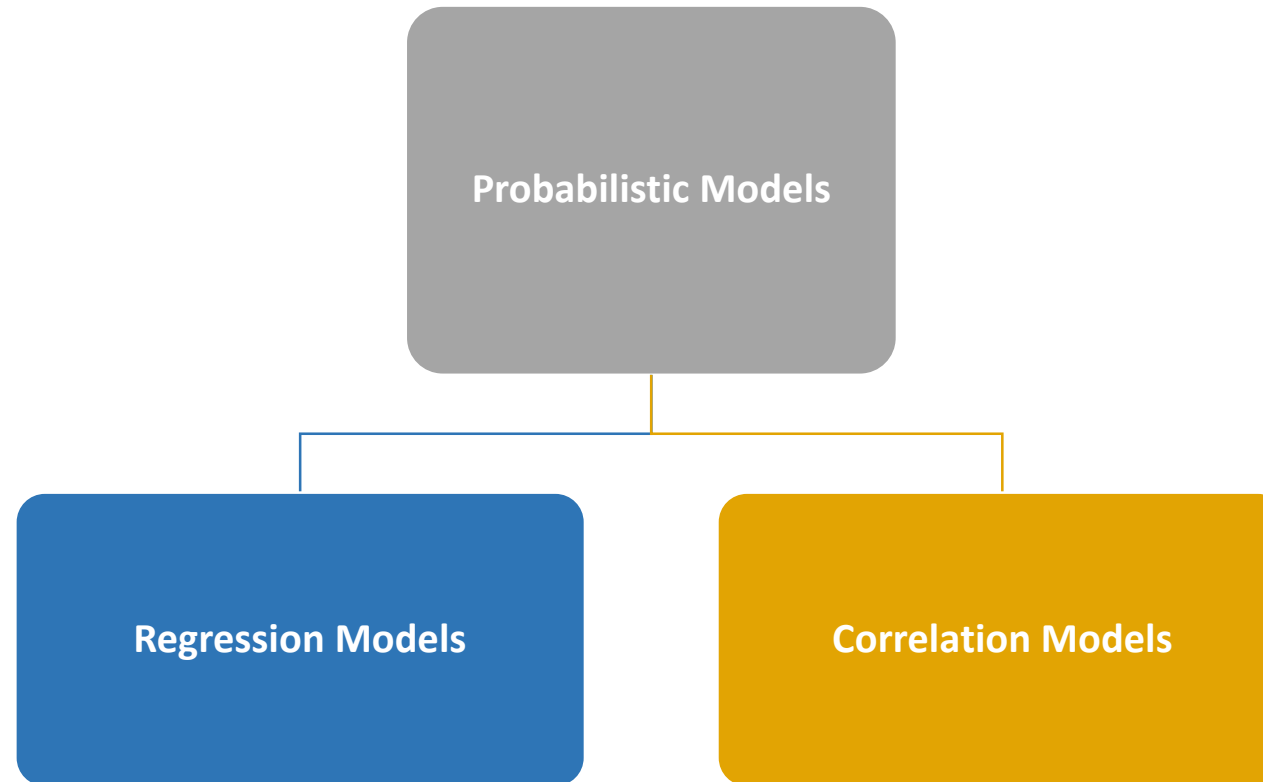
$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n}}{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}}$$

y-intercept

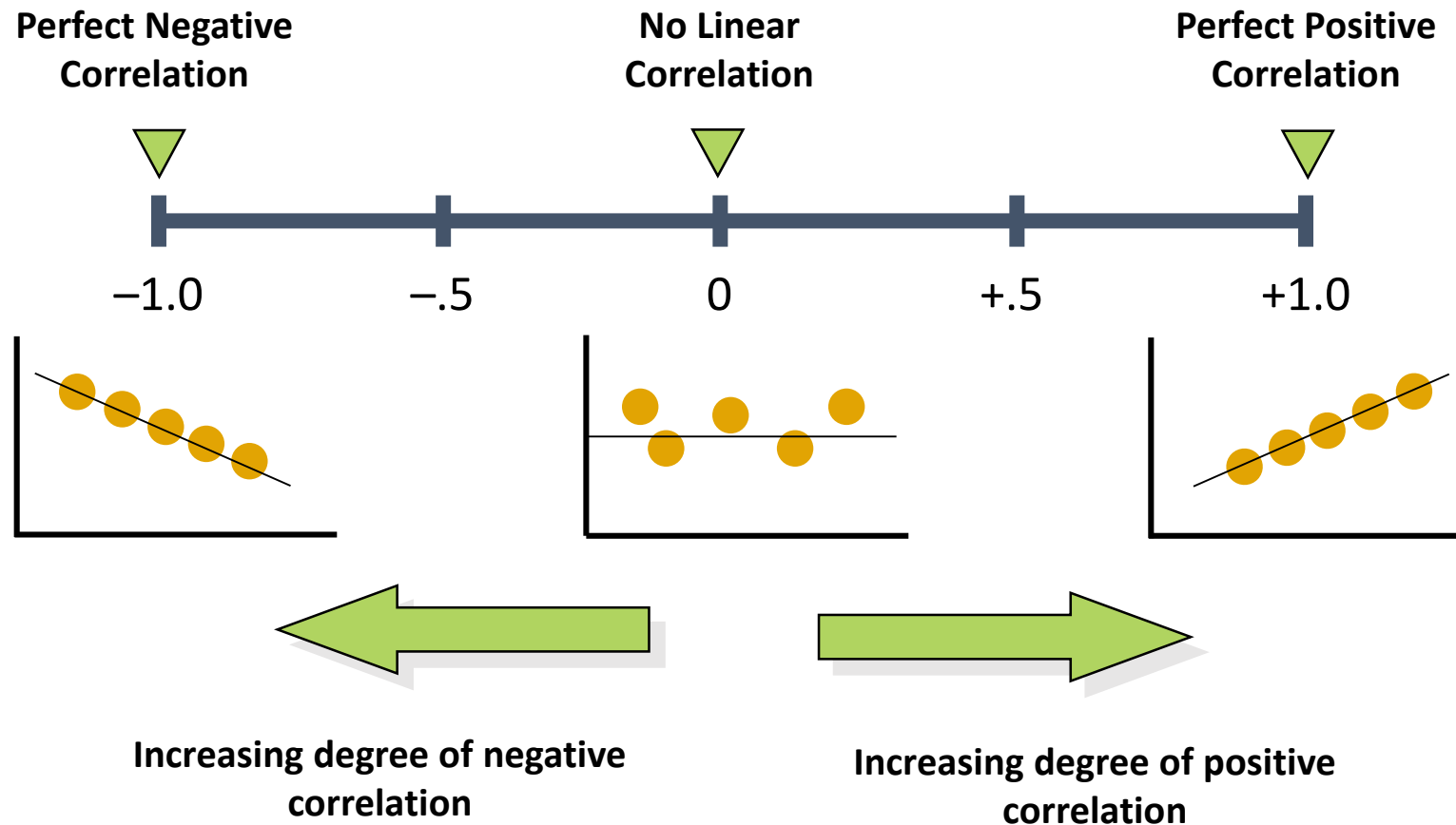
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$



Types of Probabilistic Model



Coefficient of Correlation Values



Agenda

- Introduction to Regression Analysis
- Steps in Regression Analysis?
- Simple Linier Regression
- **Multiple Linier Regression**
- Assumptions of Linier Regression Models
- Introduction to Logistic Regression Analysis



Multiple Linier Regression Model

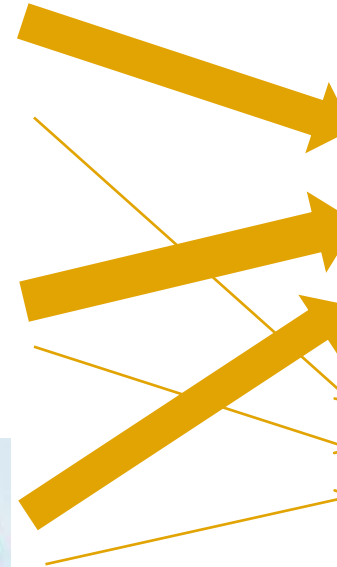
- Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. Every value of the independent variable x is associated with a value of the dependent variable
- Multiple Regression is a statistical method for estimating the relationship between a dependent variable and two or more independent (or predictor) variables.
- Simply, MLR is a method for studying the relationship between a dependent variable and two or more independent variables.
- Purposes:
 - Prediction
 - Explanation
 - Theory building



Multiple Linier Regression Model

Example:

The relationship between an adult's health and his/her daily eating amount of wheat, vegetable and meat.



Multiple Linier Regression Model

Idea: Examine the linear relationship between 1 dependent (y) & 2 or more independent variables (x_i)

Population model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon$$

Diagram labels for the population model equation:

- β_0 : Y-intercept
- $\beta_1, \beta_2, \dots, \beta_k$: Population slopes
- ε : Random Error

Estimated multiple regression model:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_k x_k$$

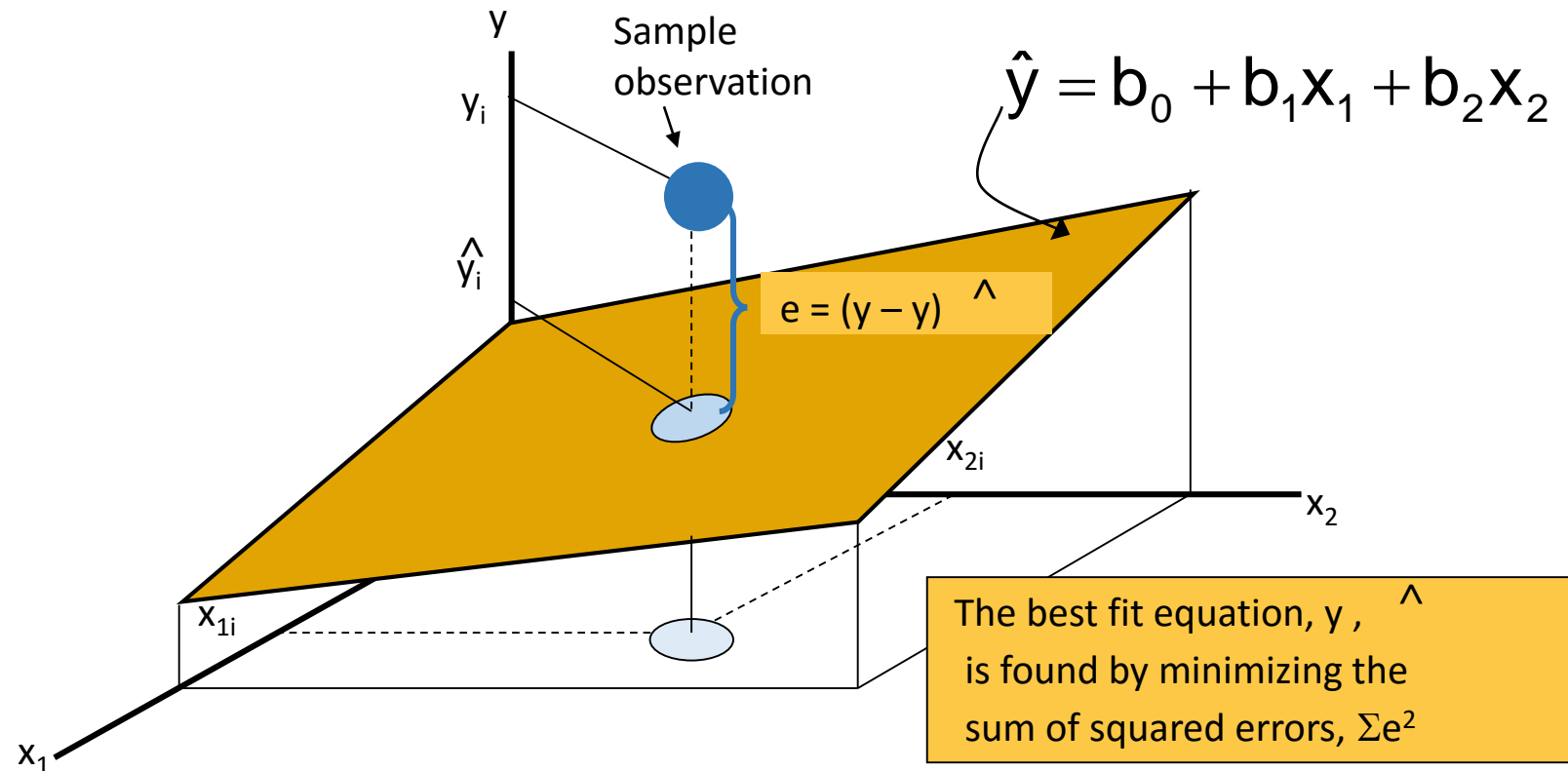
Diagram labels for the estimated multiple regression model equation:

- \hat{y} : Estimated (or predicted) value of y
- b_0 : Estimated intercept
- b_1, b_2, \dots, b_k : Estimated slope coefficients



Multiple Linear Regression Model

Two variable model



Agenda

- Introduction to Regression Analysis
- Steps in Regression Analysis?
- Simple Linier Regression
- Multiple Linier Regression
- **Assumptions of Linier Regression Models**
- Introduction to Logistic Regression Analysis



The Assumptions

1. The distribution of residuals is normal (at each value of the dependent variable).
2. The variance of the residuals for every set of values for the independent variable is equal.
 - violation is called heteroscedasticity.
3. At every value of the dependent variable the expected (mean) value of the residuals is zero
4. No non-linear relationships. The expected correlation between residuals, for any two cases, is 0.
 - The independence assumption (lack of autocorrelation)
5. No independent variables are a perfect linear function of other independent variables (no perfect multicollinearity)

—● Poin 1 2 3 4 for simple linier regression

—● Poin 1 2 3 4 5 for multipel linier regression



Look at Normal Distributions

A normal distribution : symmetrical, bell-shaped (so they say)

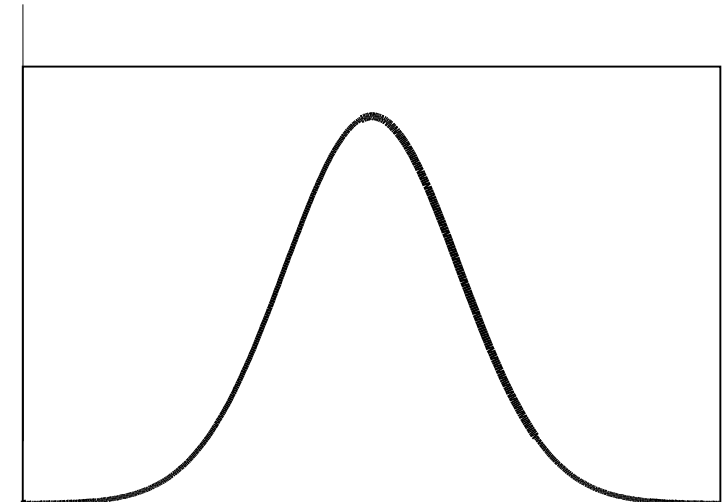
Decriptive:

What can go wrong?

- Skew
 - non-symmetricality
 - one tail longer than the other

- Kurtosis
 - too flat or too peaked
 - Kurtosed

- Outliers
 - Individual cases which are far from the distribution
- Analytic : Kolmogorv Smirnov ($N > 50$) and Shapiro Wilks Test ($N < 50$)

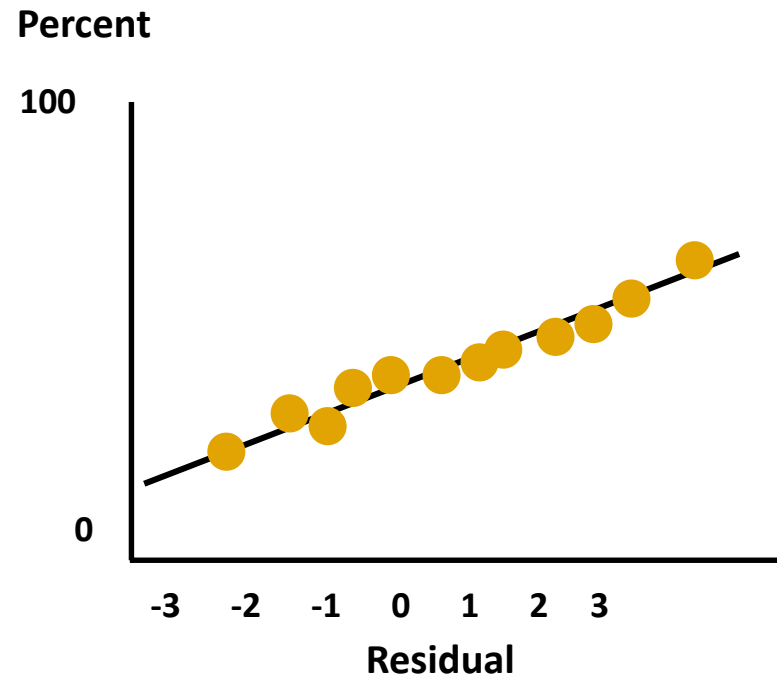


Checking for Normality

- Examine the Stem-and-Leaf Display of the Residuals
- Examine the Boxplot of the Residuals
- Examine the Histogram of the Residuals
- Construct a Normal Probability Plot of the Residuals



Residual Analysis for Normality



- When using a normal probability plot, normal errors will approximately display in a straight line



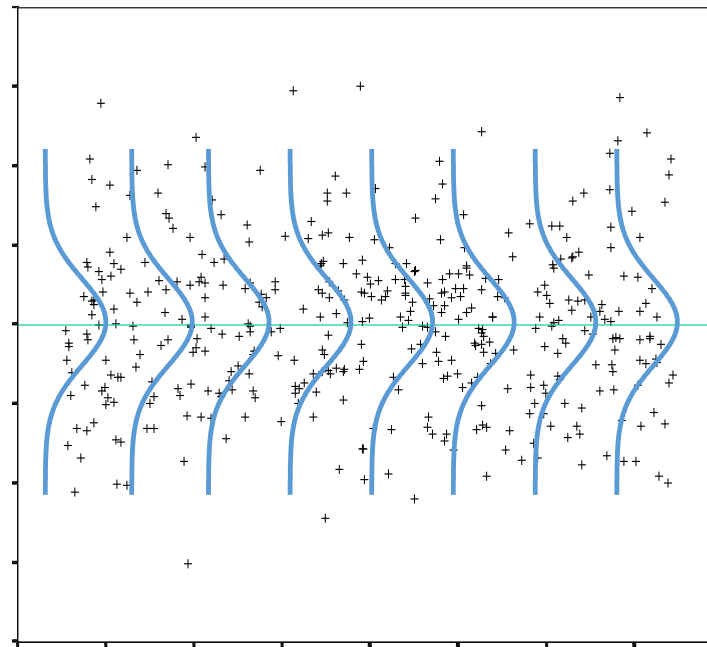
Heteroscedasticity

- The variance of the residuals for every set of values for the independent variable is equal
- This assumption is about heteroscedasticity of the residuals
 - Hetero=different
 - Scedastic = scattered
- We don't want heteroscedasticity
 - we want our data to be homoscedastic
- Draw a scatterplot to investigate



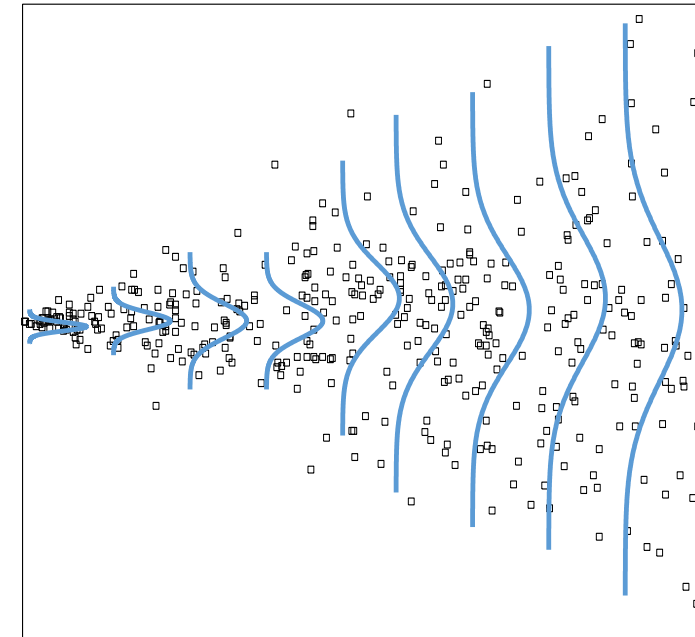
Heteroscedasticity

Good – no



Predicted Value

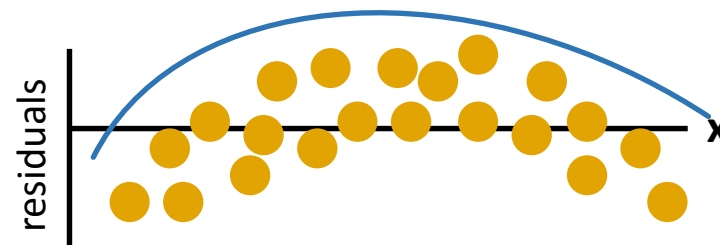
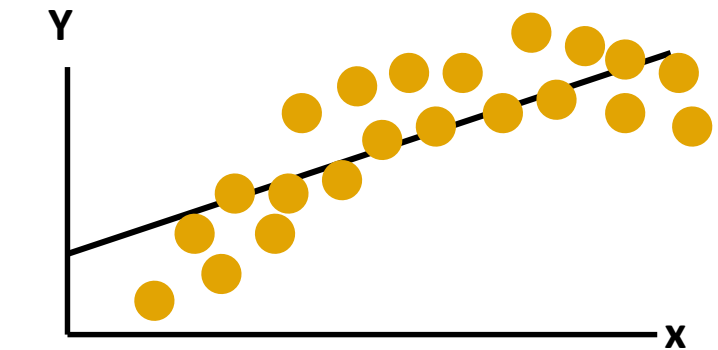
Bad – heteroscedasticity



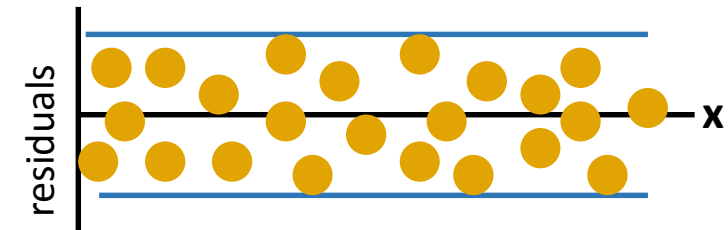
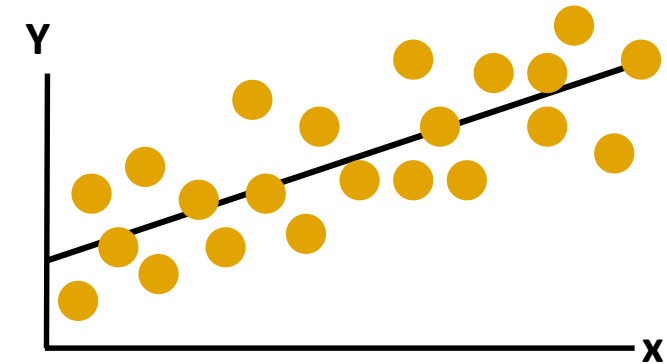
Predicted Value



Residual Analysis for Linearity



Not Linear

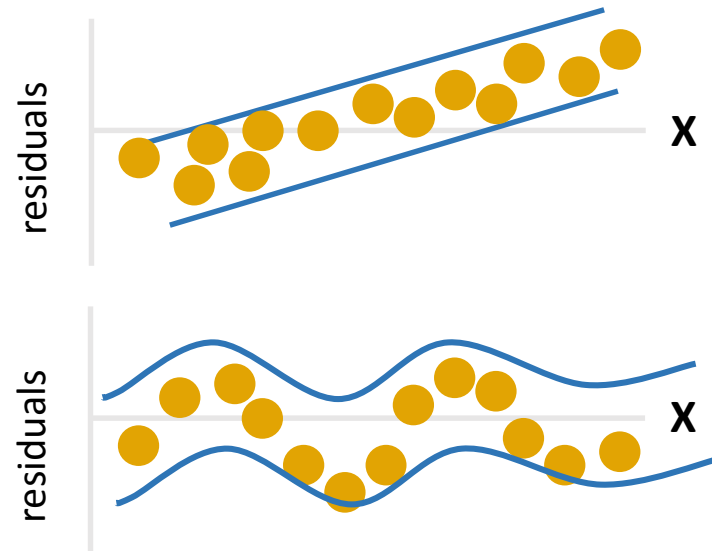


Linear

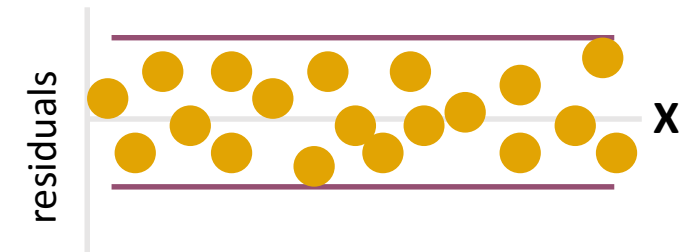


Residual Analysis for Independence

Not Independent



Independent



Independence Assumption

Measuring Autocorrelation:

The Durbin-Watson Statistic

- Used when data are collected over time to detect if autocorrelation is present
- Autocorrelation exists if residuals in one time period are related to residuals in another period




Multicollinearity

- Multicollinearity: High correlation exists between two independent variables
- This means the two variables contribute redundant information to the multiple regression model
- Including two highly correlated independent variables can adversely affect the regression results
 - No new information provided
 - Can lead to unstable coefficients (large standard error and low t-values)
 - Coefficient signs may not match prior expectations




Detect Collinearity

(Variance Inflationary Factor)

-  VIF_j is used to measure collinearity:

$$VIF_j = \frac{1}{1 - R_j^2}$$

-  R_j^2 is the coefficient of determination when the j^{th} independent variable is regressed against the remaining $k - 1$ independent variables

If $VIF_j > 5$, x_j is highly correlated with the other explanatory variables



Agenda

- Introduction to Regression Analysis
- Steps in Regression Analysis?
- Simple Linier Regression
- Multiple Linier Regression
- Assumptions of Linier Regression Models
- **Introduction to Logistic Regression Analysis**



An Introduction to Logistic Regression

- Why use logistic regression?
- Estimation by maximum likelihood
- Interpreting coefficients
- Hypothesis testing
- Evaluating the performance of the model



An Introduction to Logistic Regression

- Need to extend to scientific questions of higher dimension.
- When the number of potential covariates increases, traditional methods of contingency table analysis become limited
- One alternative approach to stratified analyses is the development of regression models that incorporate covariates and interactions among variables. Logistic regression is a form of regression analysis in which the outcome variable is binary or dichotomous
- General theory: analysis of variance (ANOVA) and logistic regression all are special cases of **General Linear Model (GLM)**



The Logistic Regression Model

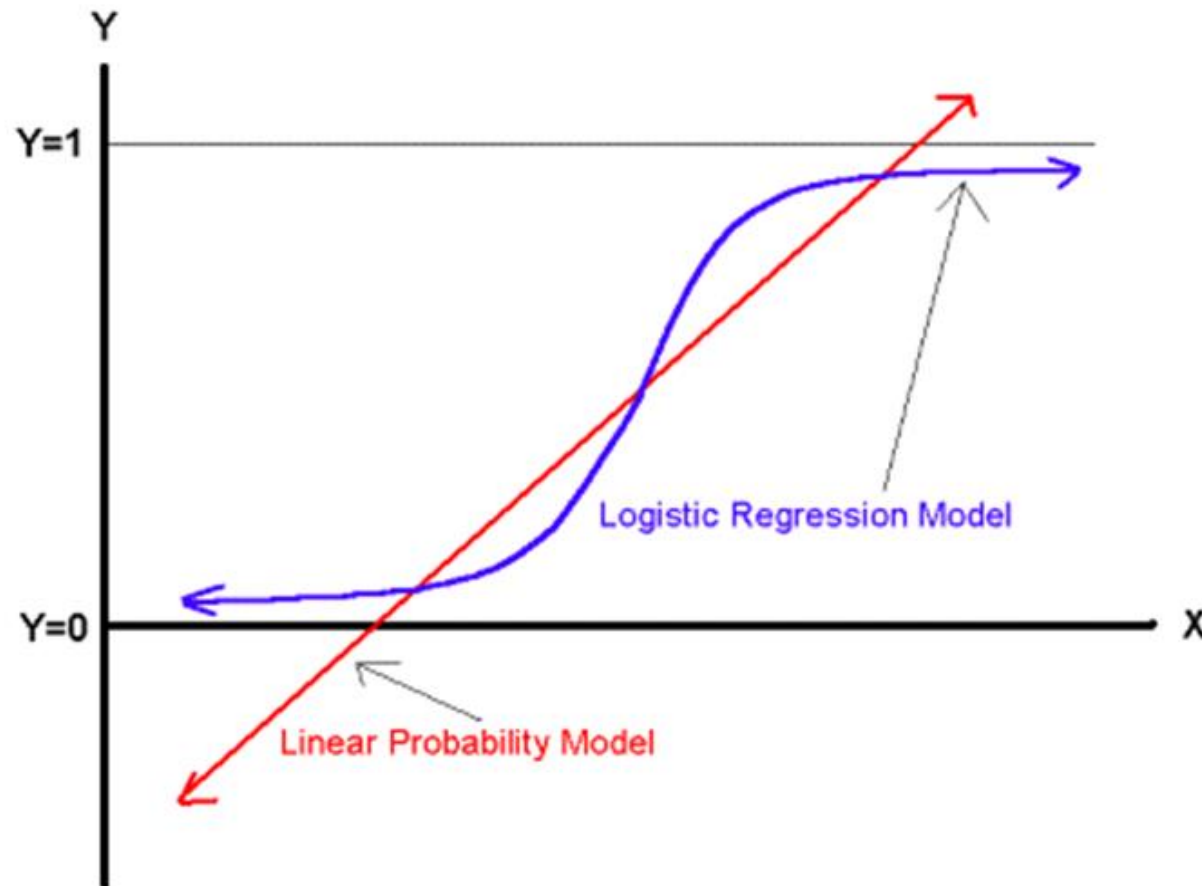
The "logit" model solves these problems:

$$\ln[p/(1-p)] = \alpha + \beta X + e$$

- p is the probability that the event Y occurs, $p(Y=1)$
- $p/(1-p)$ is the "odds ratio"
- $\ln[p/(1-p)]$ is the log odds ratio, or "logit" The logistic distribution constrains the estimated probabilities to lie between 0 and 1.
- The estimated probability is: $p = 1/[1 + \exp(-\alpha - \beta X)]$
- if you let $\alpha + \beta X = 0$, then $p = .50$
- as $\alpha + \beta X$ gets really big, p approaches 1
- as $\alpha + \beta X$ gets really small, p approaches 0

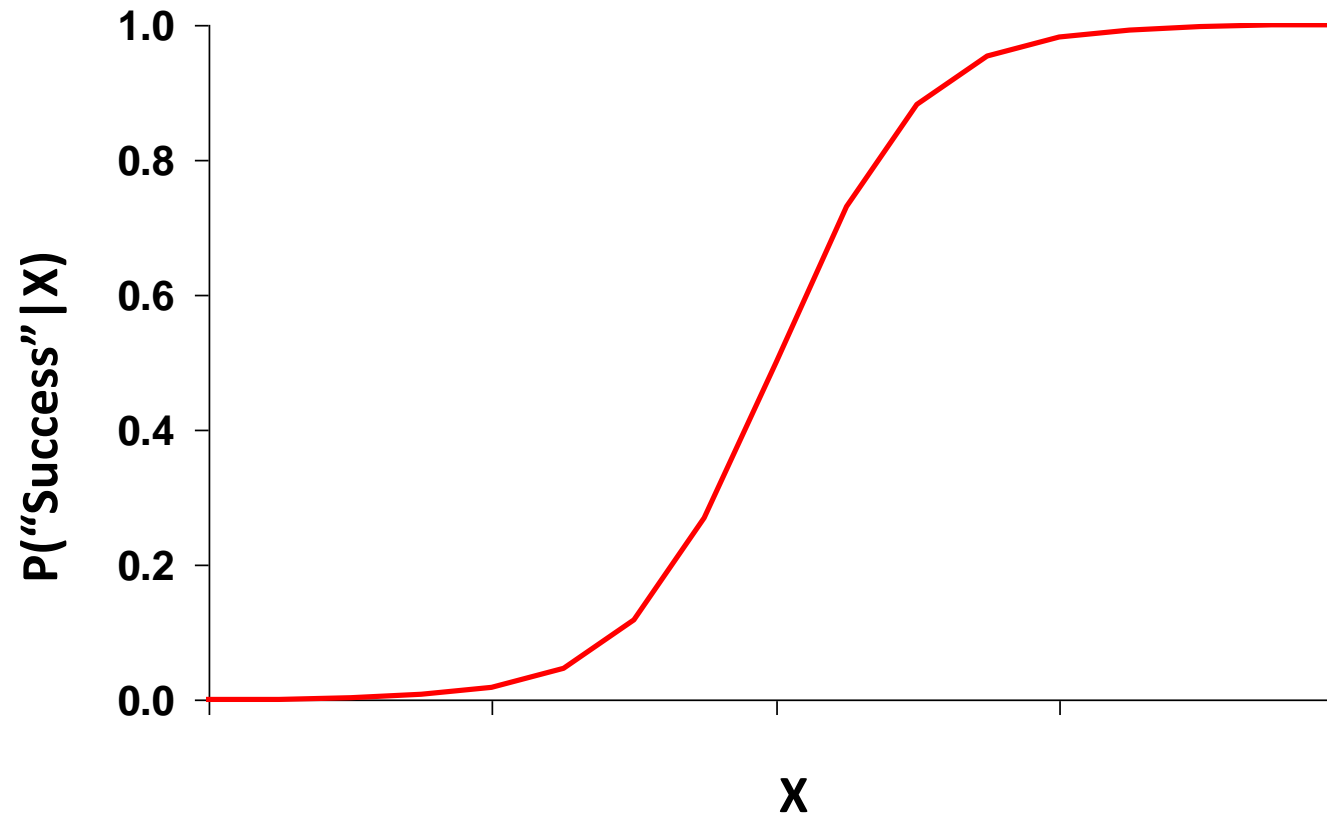


Comparing the LP and Logit Models



Logistic Function

$$P(\text{"Success"} | X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$



Why Use Logistic Regression?

Introduction

- A statistical method used to model dichotomous or binary outcomes (but not limited to) using predictor variables. Used when the research method is focused on whether or not an event occurred, rather than when it occurred (time course information is not used).

What is the “Logistic” component?

Instead of modeling the outcome, Y , directly, the method models the log odds(Y) using the logistic function. Why use logsitic regression?

- There are many important research topics for which the dependent variable is "limited."
- For example: voting, morbidity or mortality, and participation data is not continuous or distributed normally.
- Binary logistic regression is a type of regression analysis where the dependent variable is a dummy variable: coded 0 (did not vote) or 1(did vote)



Evaluating the Performance of the Model

There are several statistics which can be used for comparing alternative models or evaluating the performance of a single model:

- Model Chi-Square
- Percent Correct Predictions
- Pseudo-R²

