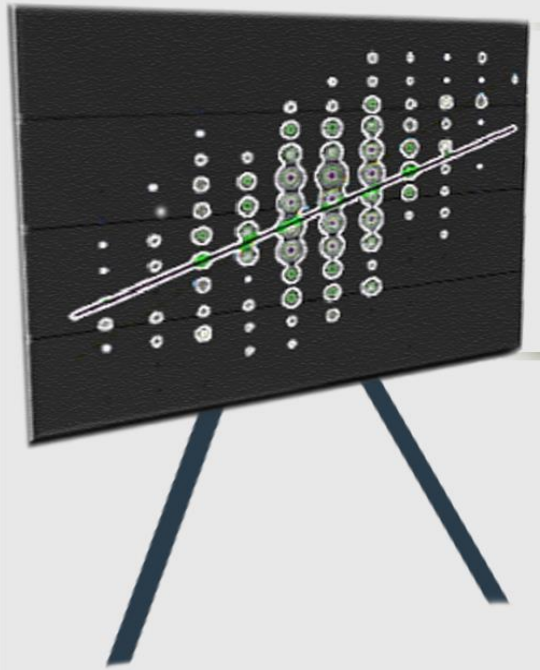
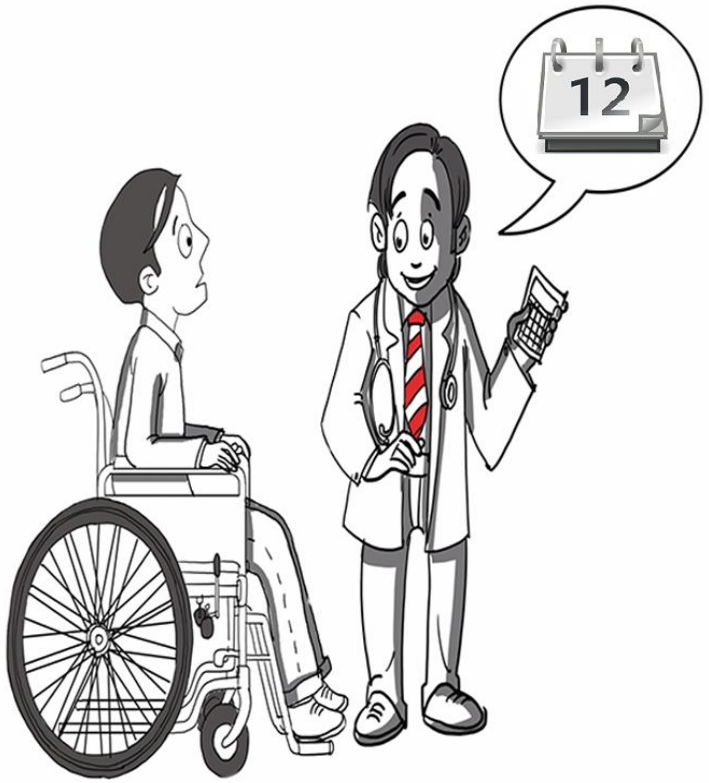


# Regression

## Introduction




IIMBX

IIMB
तेजस्वि नावधीतमस्तु

If you torture the data long enough, it will confess!

- Ronald Coase

# Interesting Hypotheses

- Good looking couples are more likely to have girl child(ren)!
- Vegetarians miss fewer flights.
- Women use camera phone more than men.
- Left handed men earn more money!
- Smokers are better sales people.
- Those who whistle at workplace are more efficient.

# What is Regression?

- Regression is a tool for finding existence of an association relationship between a dependent variable (Y) and one or more independent variables ($X_1$, $X_2$, …, $X_n$) in a study.

- The relationship can be linear or non-linear.

# Mathematical Vs Statistical Relationship

- Mathematical relationship is an exact relationship.

$$Y = \beta_0 + \beta_1 X$$

- Statistical relationship is not an exact relationship.

$$Y = \beta_0 + \beta_1 X + \epsilon$$

# Nomenclature in Regression

- A dependent variable (**response variable**) "measures an outcome of a study (also called **outcome variable**)".
- An independent variable (**explanatory variable**) "explains changes in a response variable".
- Regression often set values of explanatory variable to see how it affects response variable (predict response variable).

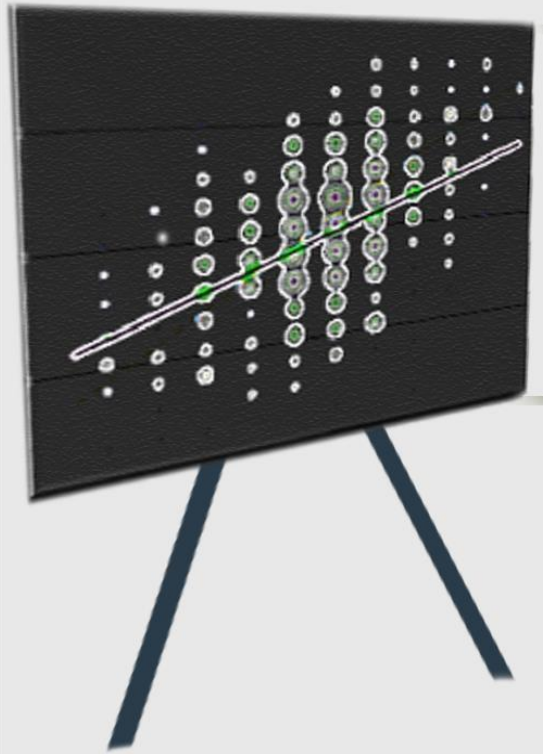
**Regression model establishes the existence of an association between two variables, but not causation.**

# Regression Nomenclature

| Dependent Variable | Independent Variable |
|---|---|
| Explained Variable | Explanatory variable |
| Regressand | Regressor |
| Predictand | Predictor |
| Endogenous Variable | Exogenous Variable |
| Controlled Variable | Control Variable |
| Target Variable | Stimulus Variable |
| Response Variable | |

# Dependent and Independent Variables

- Terms dependent and independent does not necessarily imply a causal relationship between two variables.
- Regression is **not** designed to capture **causality**.
- Purpose of regression is to predict the value of dependent variable given the value(s) of independent variable(s).



# Regression Importance


IIMBX

IIMB
तेजस्वि नावधीतमस्तु

# Why we need Regression?

- Companies would like to know about factors that have significant impact on their Key Performance Indicators (KPI).

- Regression helps to create new hypothesis that may assist the companies to improve their performance.

# Where is it used?

- **Finance:** CAPM, Non-performing assets, probability of default, chance of bankruptcy, credit risk.
- **Marketing:** Sales, market share, customer satisfaction, customer churn, customer retention, customer life time value.
- **Operations:** Inventory, productivity, efficiency.
- **HR:** Job satisfaction, attrition.

# Business Problems in Marketing/Retail

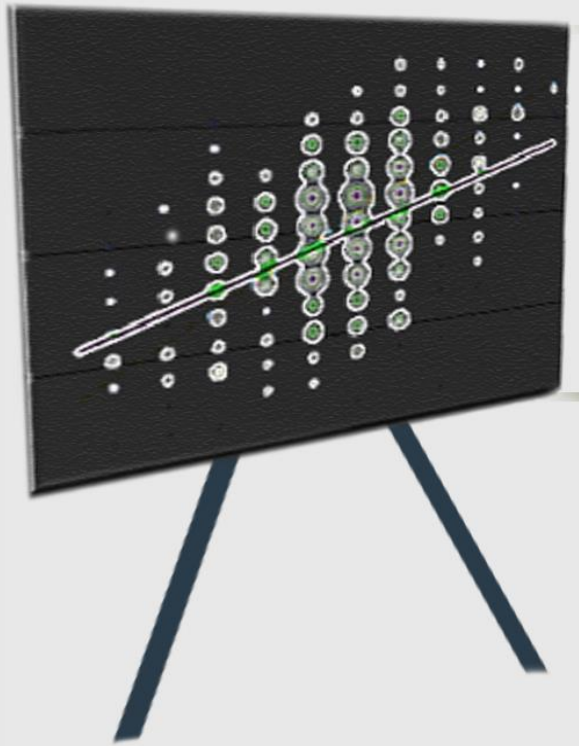


- How to improve the success probability of a new product?
- What is the impact of food label on purchase decision?
- Which promotion is more effective?

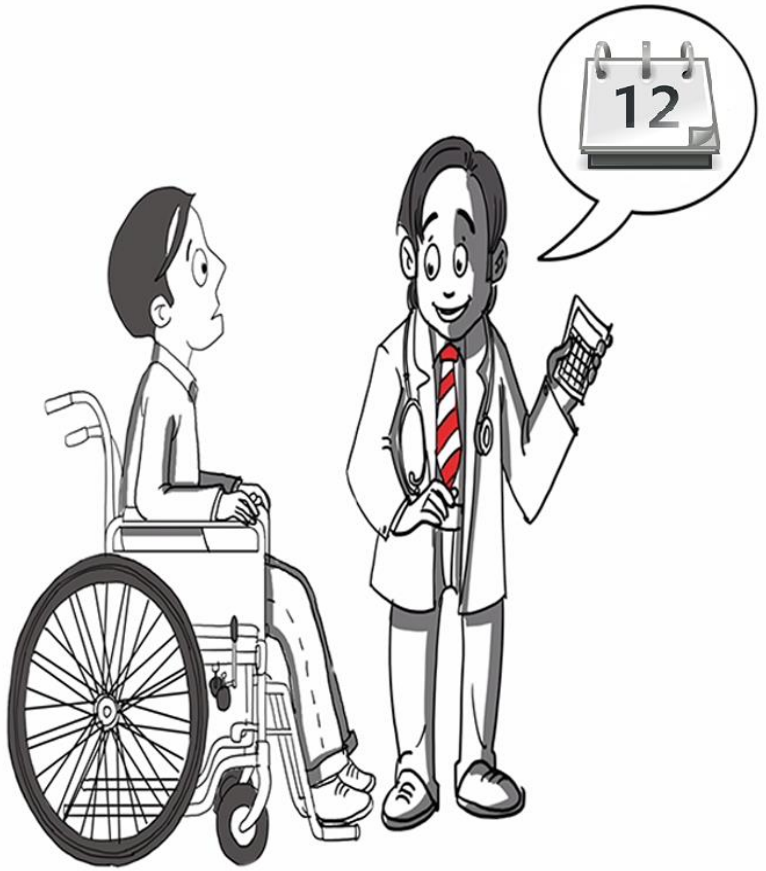
# Business Problems in Banking and Finance



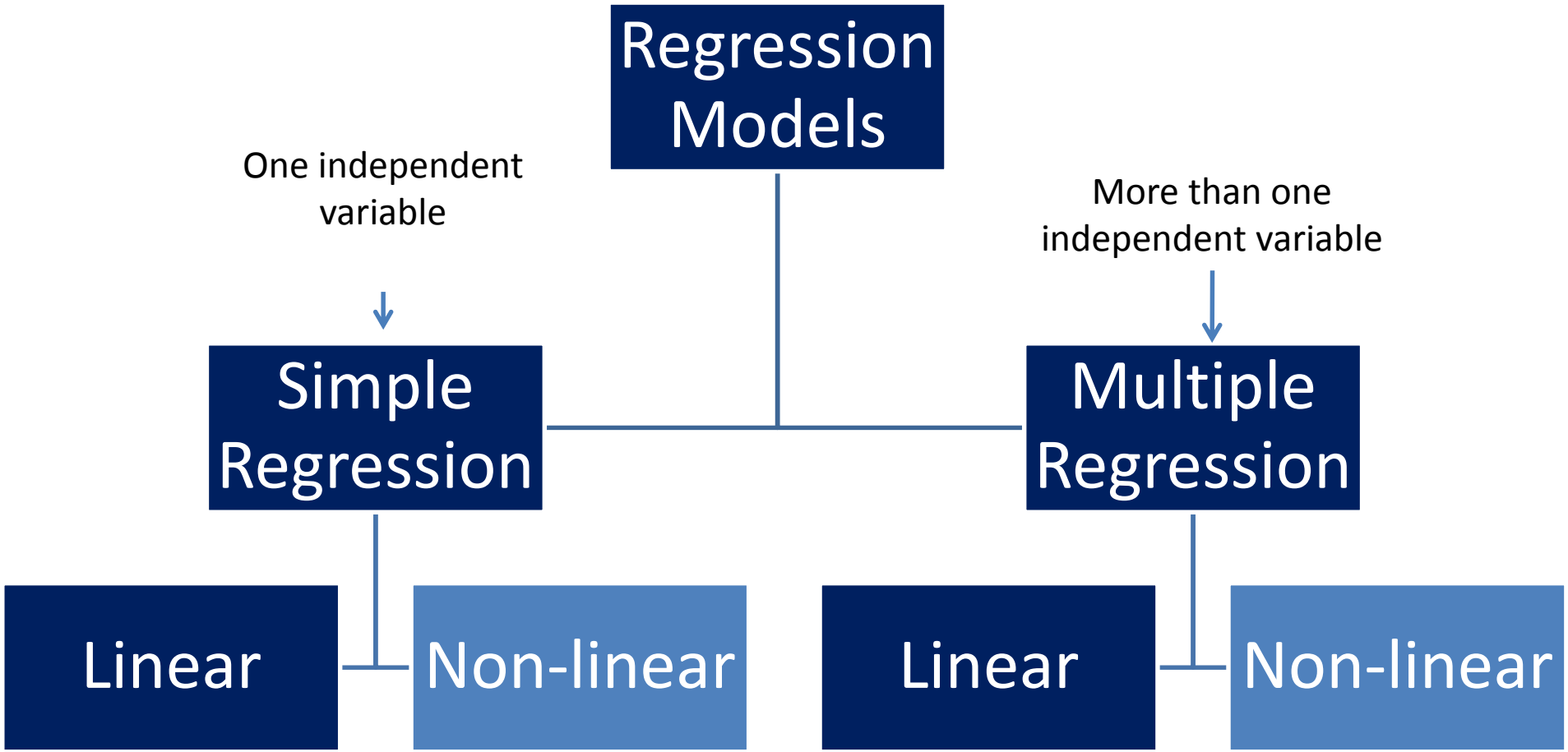
- What is the risk associated with a customer?
- Which customer is likely to default?
- What percentage of loans are likely to result in a loss?
- How to identify the most profitable customer?



# Regression Types



# Types of Regression



# Types of Regression

- **Simple** linear regression

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

- **Multiple** linear regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

- **Nonlinear** regression

$$Y = \beta_0 + \frac{1}{\beta_1 + \beta_2 X_1} + X_2^{\beta_3} + \varepsilon$$

# Multiple Linear Regression

Multiple linear regression means linear in regression parameters (beta values).

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_2^2 \ldots + \beta_k x_k + \varepsilon$$

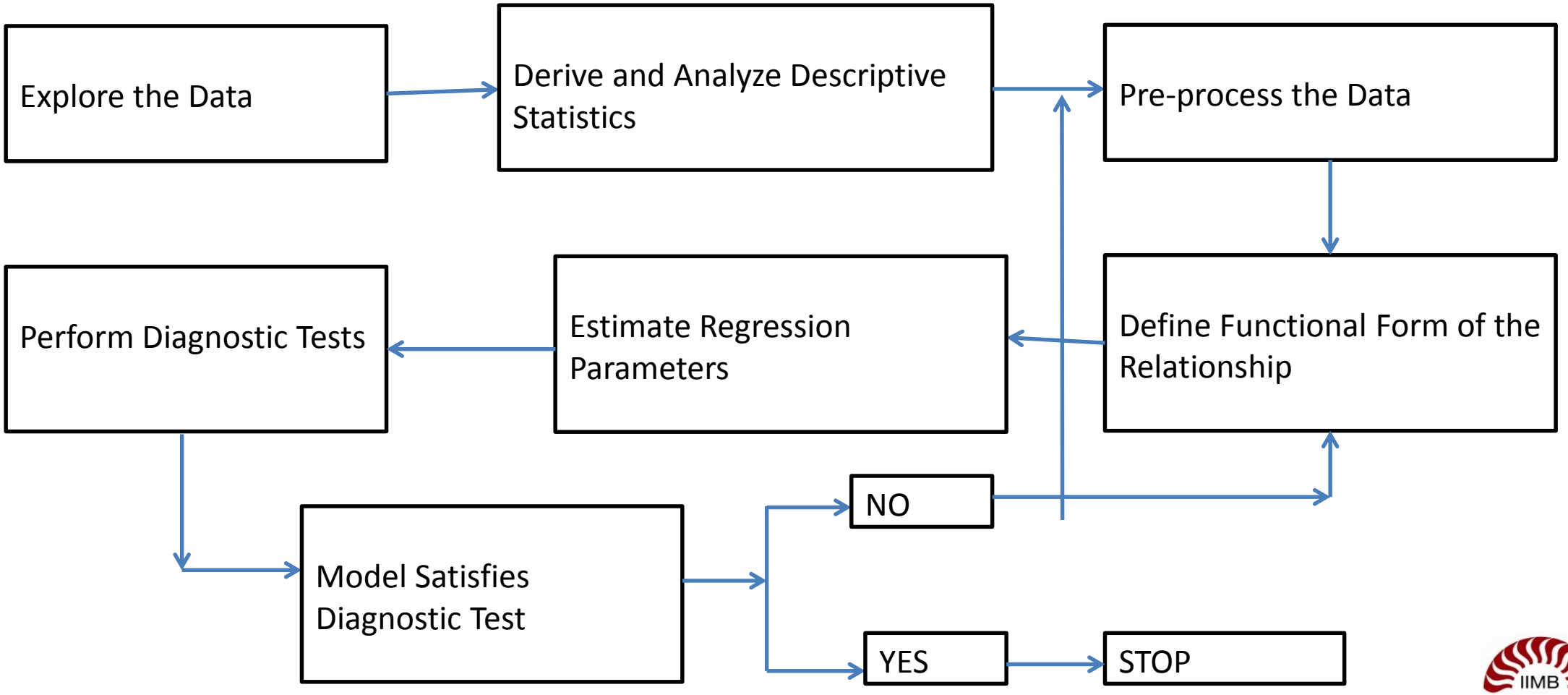
**An important task in multiple regression is to estimate the beta values ($\beta_1$, $\beta_2$, $\beta_3$ etc…)**

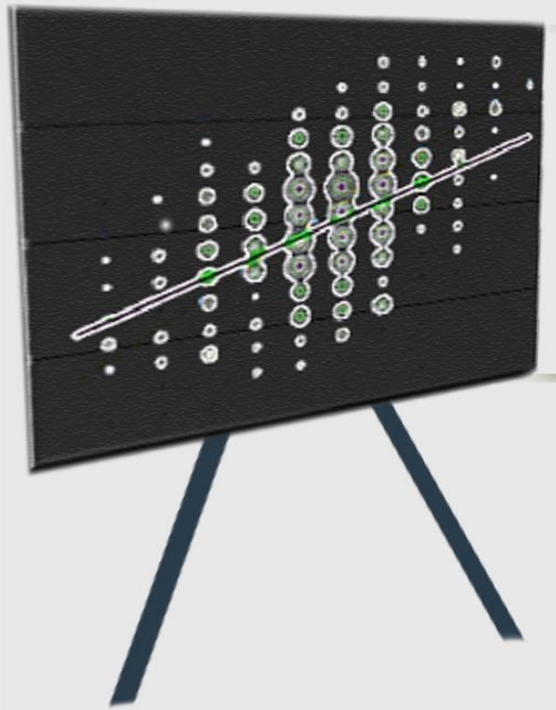
IIMBX

IIMB
तेजस्वि नावधीतमस्तु

# Regression Model Development

# Regression Model Development





# Regression

## Model Building


IIMBX

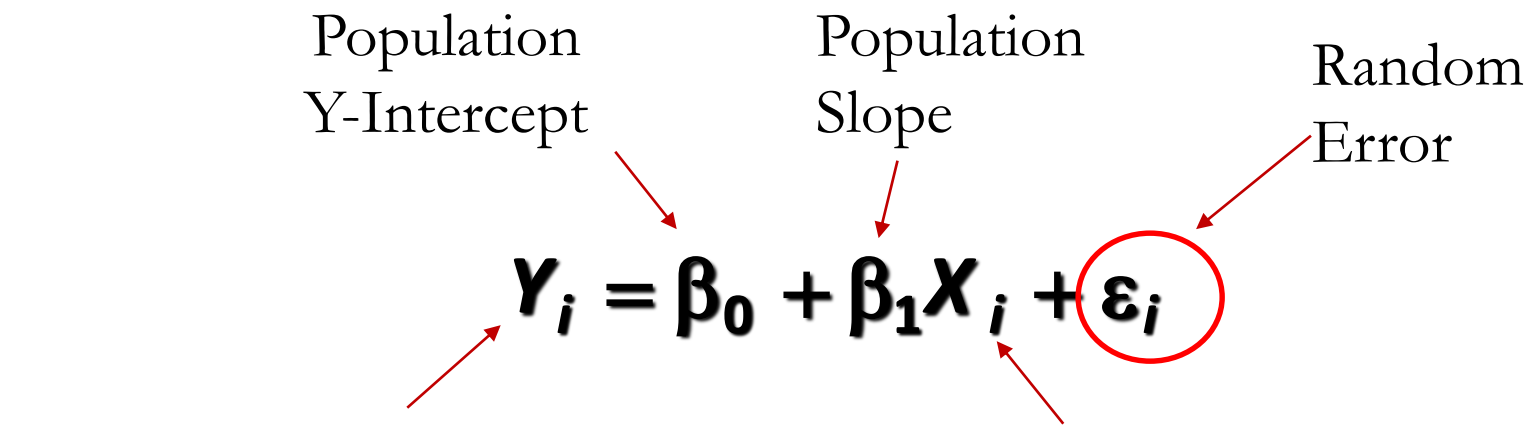
IIMB
तेजस्वि नावधीतमस्तु

# Functional Form

- Identify the explanatory variable.
- Specify the nature of relationship between dependent variable and explanatory variables.

# Linear Regression Model

Relationship between variables is a linear function.



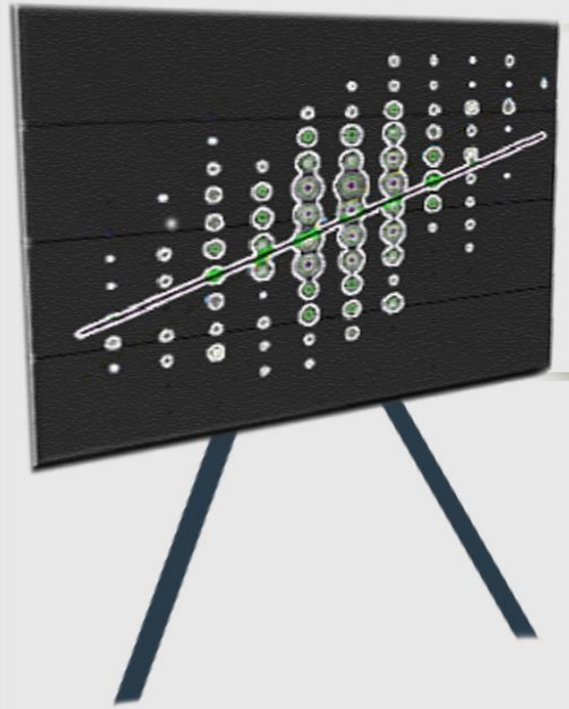
# Model Assumptions

# Linear Regression Model Assumptions

- The error term, $\varepsilon_i$, follows a normal distribution.
- For different values of X, the variance of $\varepsilon_i$ is constant (Homoscedasticity).
- There is no Multi-collinearity (no perfect linear relationship among explanatory variables).
- There is no autocorrelation between two $\varepsilon_i$ values.

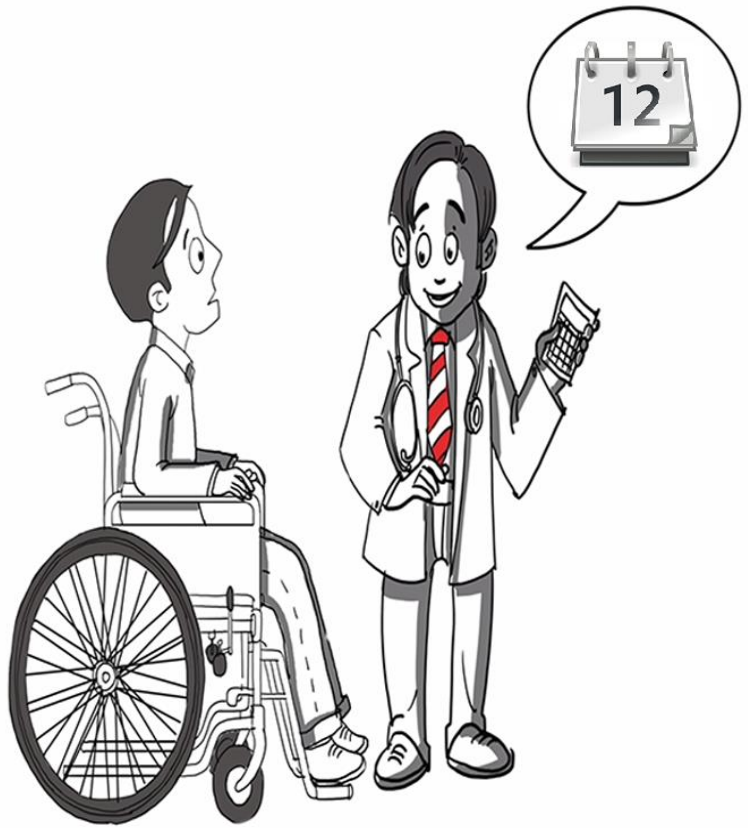
IIMBX

IIMB
तेजस्वि नावधीतमस्तु



# Regression

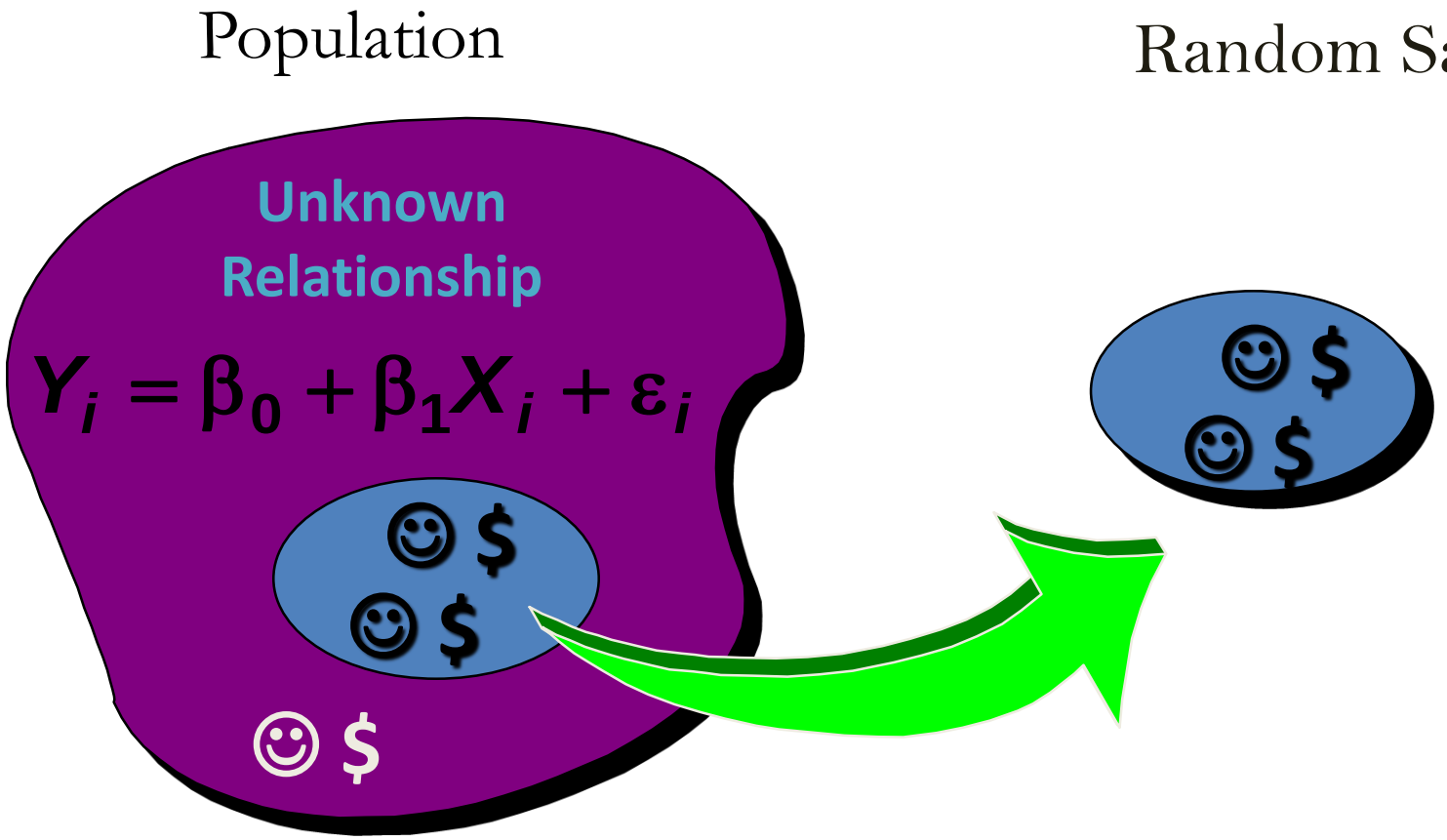
## OLS Estimation




IIMBX

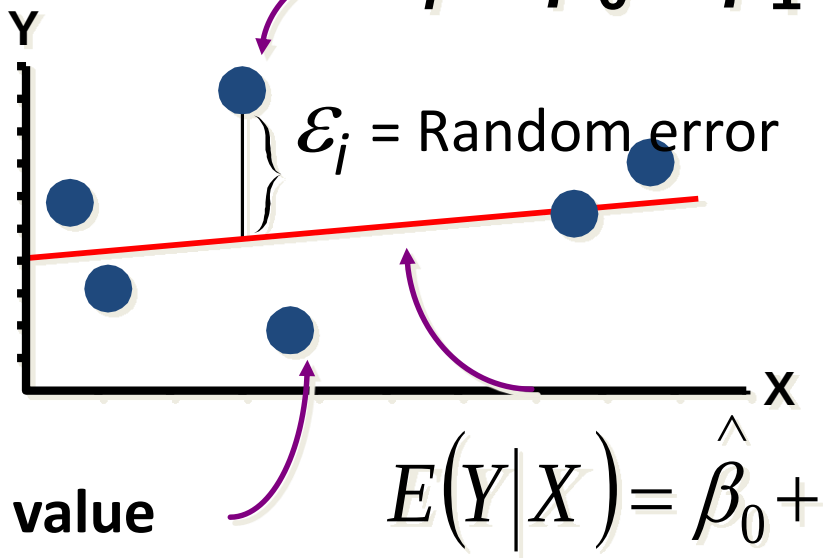
IIMB
तेजस्वि नावधीतमस्तु

# Estimation of Parameters



# Population Linear Regression Model

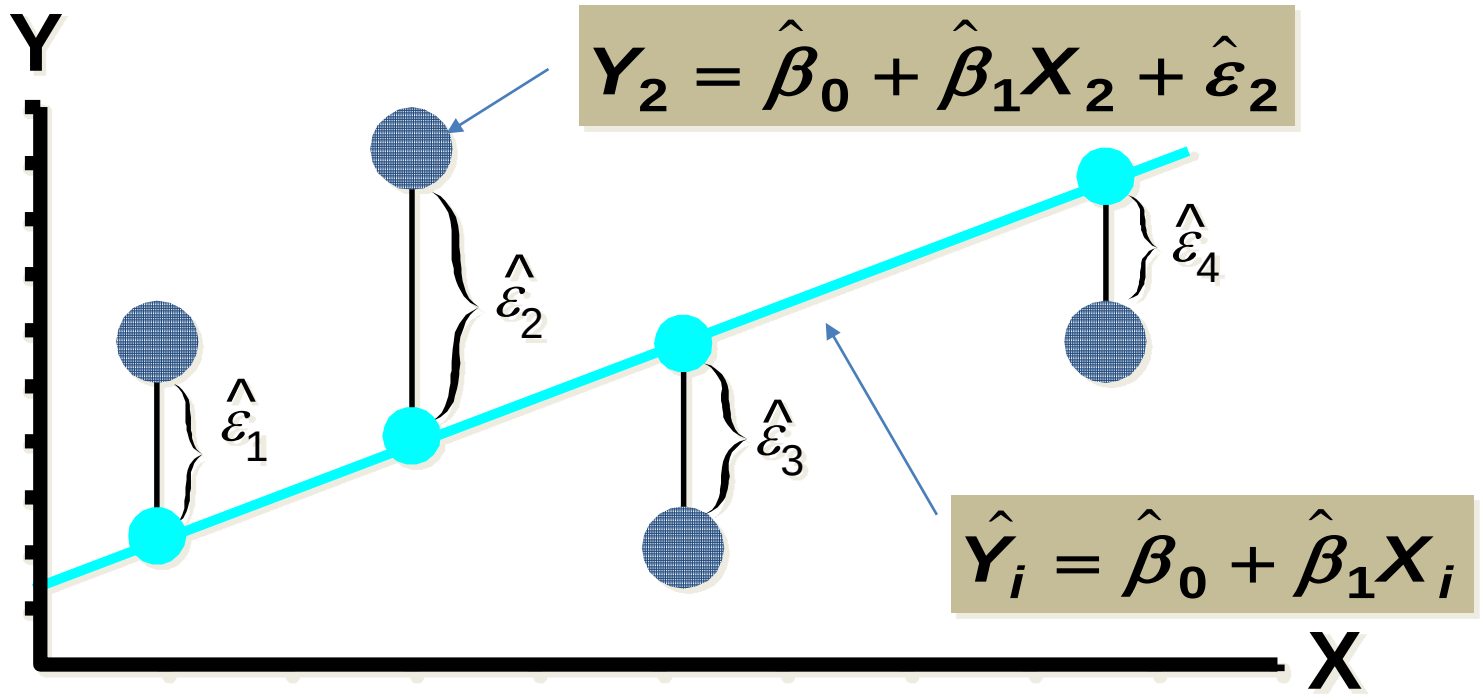
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$



# Method of Ordinary Least Squares (OLS)

# Least Squares Graphically

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



# Estimation of Parameters in Regression

The least squares function is given by:

$$SSE = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2$$

The least squares estimates must satisfy:

$$\frac{\partial SSE}{\partial \beta_0} = -2 \sum_{i=1}^{n} \left( y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) = 0$$

and

$$\frac{\partial SSE}{\partial \beta_j} = -2 \sum_{i=1}^{n} \left( y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) x_{ij} = 0 \; \forall j$$

# Coefficient Equations

**Prediction Equation:**

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

**Sample Slope:**

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

**Sample Y-intercept:**

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

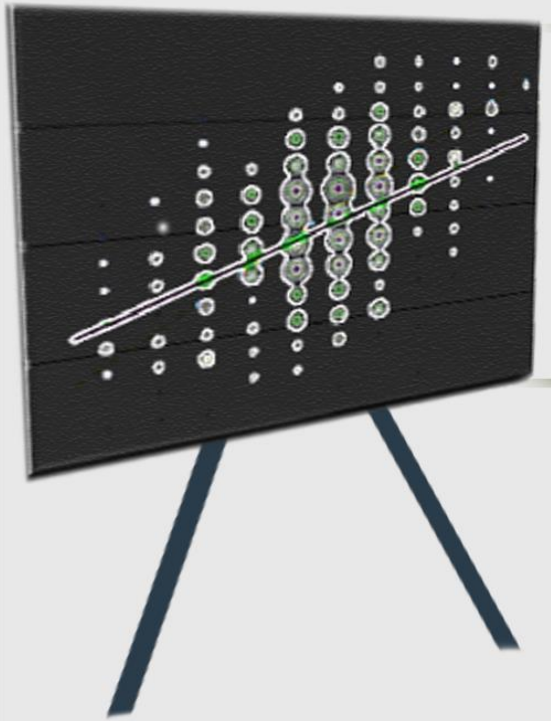
# Why Least Squares Estimate?

- OLS beta estimates are, **"Best Linear Unbiased Estimates (BLUE)"**, provided the error terms are uncorrelated (no auto regression) and have equal variance (homoscedasticity). That is,

$$E\left[\beta - \hat{\beta}\right] = 0$$

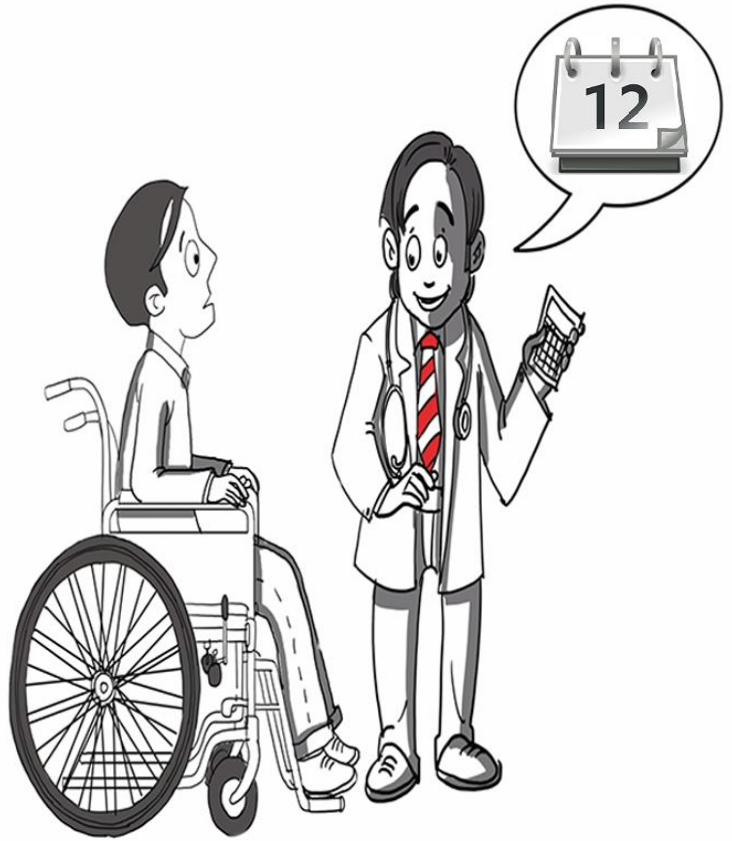
# Advantages of OLS Estimates

- They are unbiased estimates.
- They (estimates) have minimum variance.
- They have consistency, as the sample size increases, the estimate, $\hat{\beta}_i$ converges to the true population parameter value, $\beta_i$.



# Regression

## Interpret Beta




IIMBX

IIMB
तेजस्वि नावधीतमस्तु

# Interpretation of Regression Coefficients

The interpretation depends on the functional form of the relationship between the response and the explanatory variables.

# Coefficients Interpretation

- The intercept, $\beta_0$, is the mean value of the dependent variable Y, when the independent variable X = 0

- The slope, $\beta_1$, is the change in the value of the dependent variable, Y, for unit change in the independent variable X.

# Interpretation of the intercept $\beta_0$

- The intercept, $\beta_0$, is the mean value of the dependent variable Y, when the independent variable X = 0

- Y = **129110.79** + 1807.591 x Body Weight

# Interpretation of the intercept $\beta_1$

- The slope, $\beta_1$, is the change in the value of the dependent variable, Y, for unit change in the independent variable X.

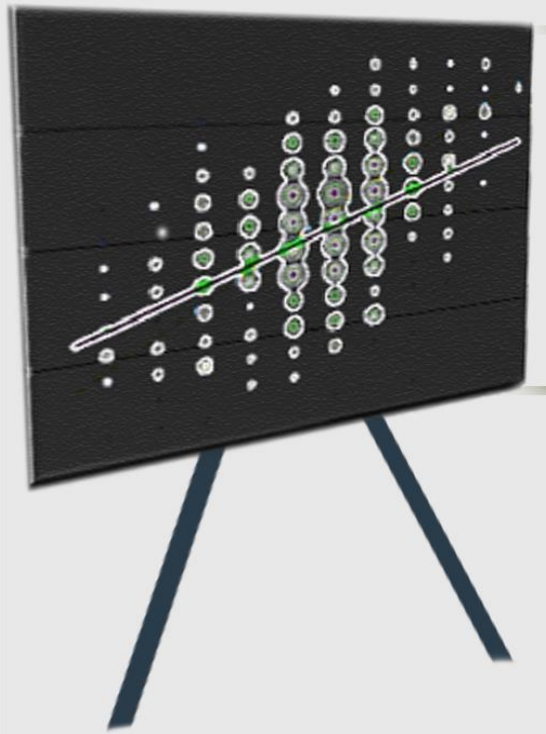
- Y = 129110.79 + **1807.591** x Body Weight

# Interpretation of $\beta_0$ and $\beta_1$ in ln(Y) = $\beta_0$ + $\beta_1$ ln(X)

Differentiating the equation with respect to X, we get:

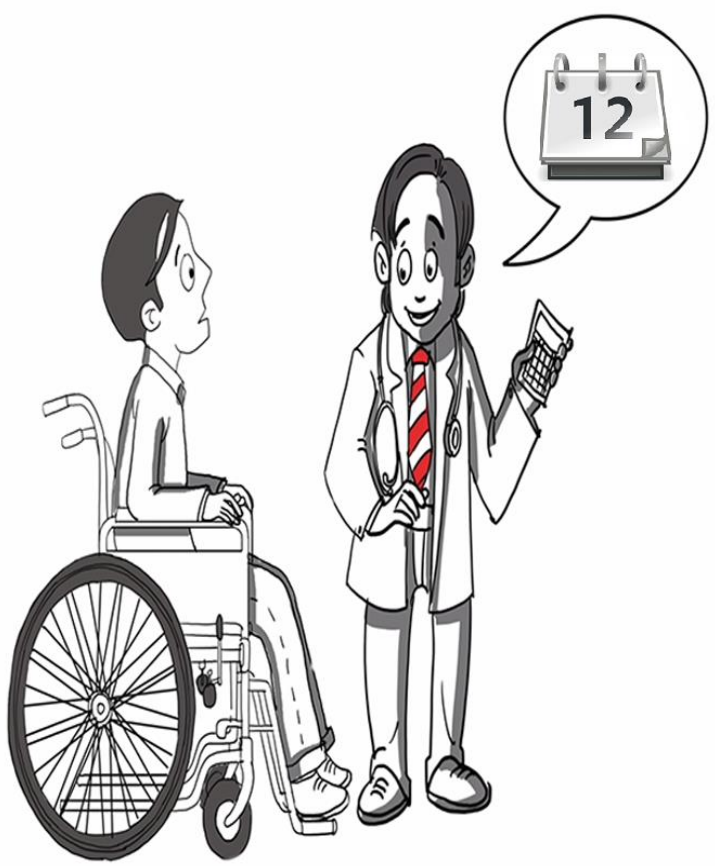
$$\frac{1}{Y}\frac{\partial Y}{\partial X} = \frac{\beta_1}{X} \Rightarrow \beta_1 = \frac{\partial Y / Y}{\partial X / X}$$

$\beta_1$ is the percentage change Y for percentage change in X.



# Regression

## Model Validation




IIMBX

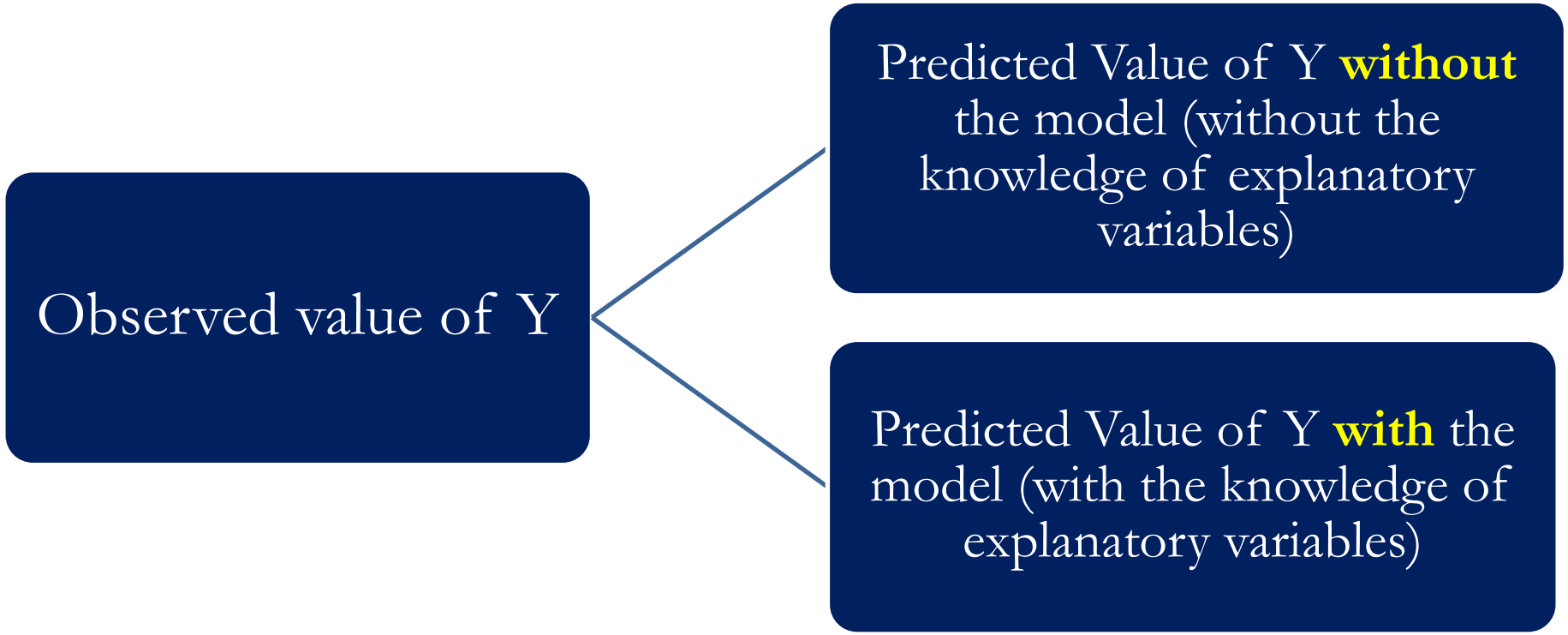
IIMB
तेजस्वि नावधीतमस्तु

# Simple Linear Regression



| | |
|---|---|
| Variable $x$ and $y$ has ***Linear*** relationship | Assumption of the world |
| $y = \beta_0 + \beta_1 x + \varepsilon$, ***Minimize SSE*** | Fitting a model |
| Is $x$ really related to $y$? ***Is $\beta_1$ statistically significant?*** | Validating the model |
| ***Predict*** $y$ for a given $x$. | Using a model |

# Objective Model Validation



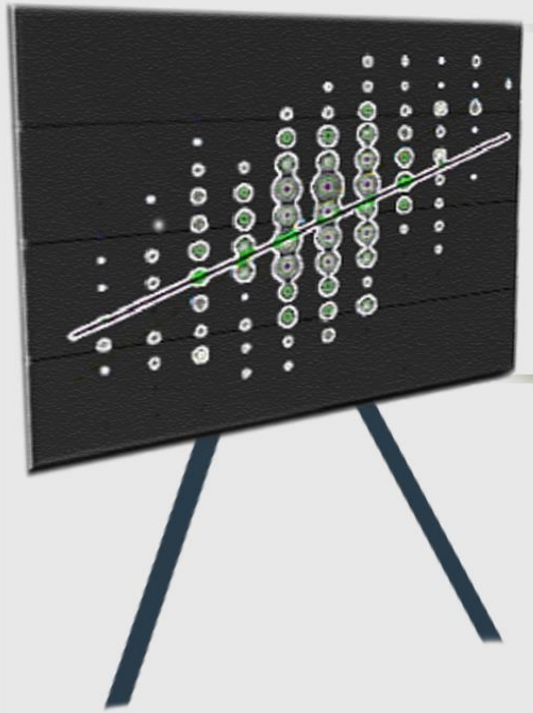

IIMBX

IIMB
तेजस्वि नावधीतमस्तु

# Model Validation



- Use of co-efficient of determination to check the goodness of fit of regression.
- Analysis of Variance (ANOVA) and F test to check the overall fitness of the regression model.
- t-test to validate relationship dependent and individual independent variable.
- Residual analysis to check the model adequacies.



# Regression

## R- Square


IIMBX

IIMB
तेजस्वि नावधीतमस्तु

# What is coefficient of determination?

- The coefficient of determination ($R^2$ square) is a measure of how well the regression line fits the data.
- The value of $R^2$ lies between 0 and 1 and is the percentage of variation explained by the regression model.
- $R^2$ is a rough indicator of the worth of the regression model.
- $R^2$ is the square of the correlation coefficient r ($R^2 = r^2$).

# Variation in Y

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

| Variation in $Y_i$ | = | Systemic Variation | + | Random Variation |
|---|---|---|---|---|

or

| Variation in $Y_i$ | = | Explained Variation | + | Unexplained Variation |
|---|---|---|---|---|

# Variation in Y

$$\underbrace{Y_i - \bar{Y}}_{\text{Total variation}} = \underbrace{\hat{Y}_i - \bar{Y}}_{\text{Explained variation}} + \underbrace{Y_i - \hat{Y}_i}_{\text{Unexplained variation}}$$

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{SST} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{SSR} + \underbrace{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}_{SSE}$$

- TOTAL SUM OF SQUARES (**SST**):
  - $SST = \Sigma(Y_i - \bar{Y})^2$
  - How much error is there in predicting Y without the knowledge of X?
- SUM OF SQUARES ERROR (**SSE**):
  - $SSE = \Sigma\ (Y_i - \hat{Y}_i)^2$
  - How much error is there in predicting Y with the knowledge of X?

- SUM OF SQUARES REGRESSION (**SSR**):

$$SSR = \Sigma\ (\hat{Y}_i - \bar{Y})^2$$ (Amount of variation explained by the model).

Mathematically, **SST = SSR + SSE**

# Coefficient of determination

Coefficient of determination is the ratio sum of squares due to regression to the total sum of squares.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

## SUMMARY OUTPUT

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.351 |
| **R Square** | 0.123 |
| Adjusted R Square | 0.120 |
| Standard Error | 111886.046 |
| Observations | 247.000 |

ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 1.000 | 4.30902E+11 | 4.30902E+11 | 34.4212328 | 1.4321E-08 |
| Residual | 245.000 | 3.06703E+12 | 12518487400 | | |
| Total | 246.000 | 3.49793E+12 | | | |

| | | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
|---|---|---|---|---|---|---|---|---|---|
| Intercept | | **129110.795** | **13564.94009** | 9.517977506 | 1.76841E-18 | 102392.0146 | 155829.5747 | 102392.0146 | 155829.5747 |
| | 49.2 | **1807.591** | **308.0966635** | 5.866961121 | 1.4321E-08 | 1200.735021 | 2414.447272 | 1200.735021 | 2414.447272 |

IIMBX

# Spurious Regression

One of the major problems with coefficient of determination is that two sets of data without any relationship can have a very high coefficient of determination value.

# The data shows the number of Facebook users (in millions) and the number of people who died of Helium poisoning in UK between 2004 and 2012

| Year | Number of Facebook users in millions | Number of people who died of Helium Poisoning |
|---|---|---|
| 2004 | 1 | 2 |
| 2005 | 6 | 2 |
| 2006 | 12 | 2 |
| 2007 | 58 | 2 |
| 2008 | 145 | 11 |
| 2009 | 360 | 21 |
| 2010 | 608 | 31 |
| 2011 | 845 | 40 |
| 2012 | 1056 | 51 |

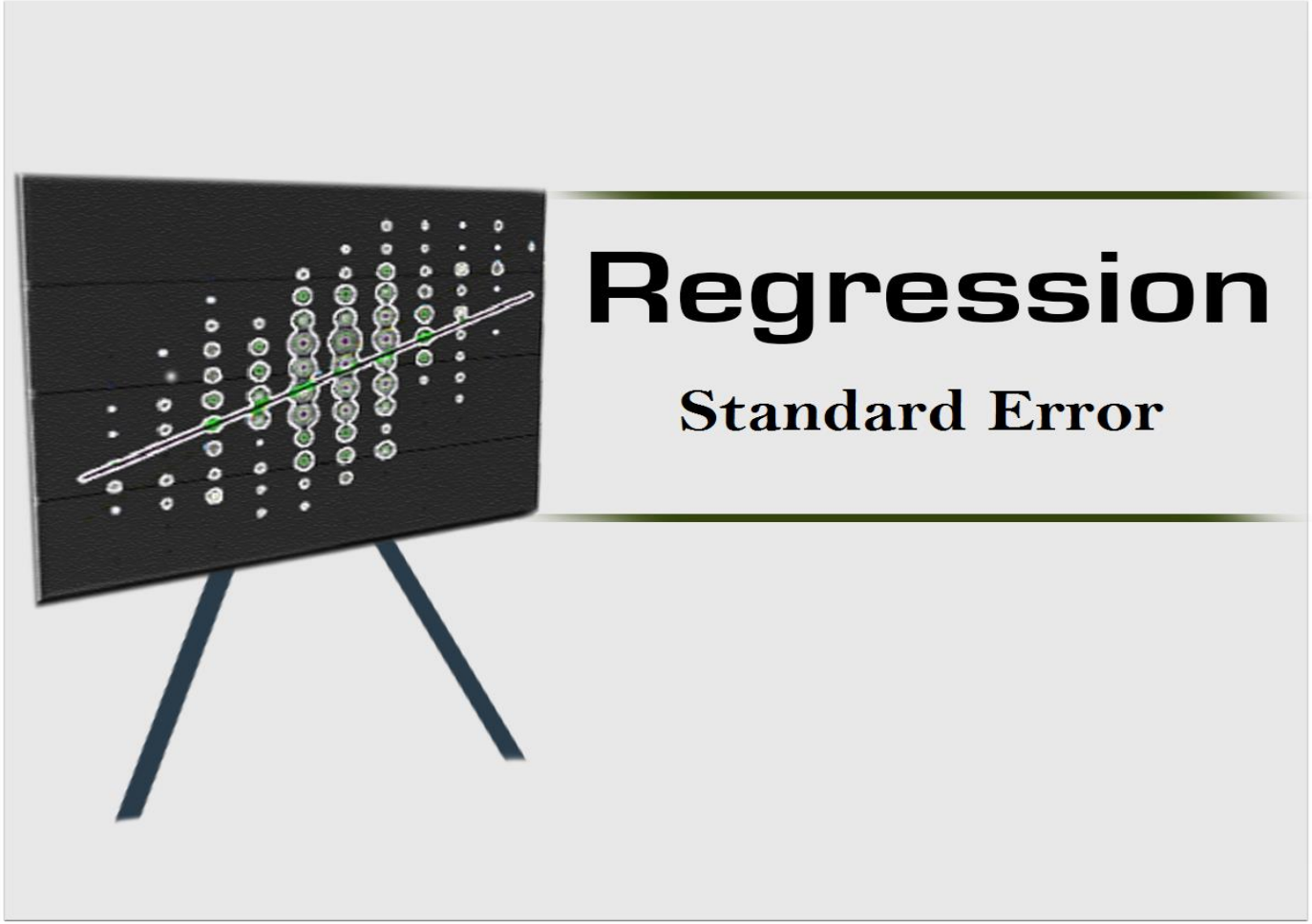
SUMMARY OUTPUT

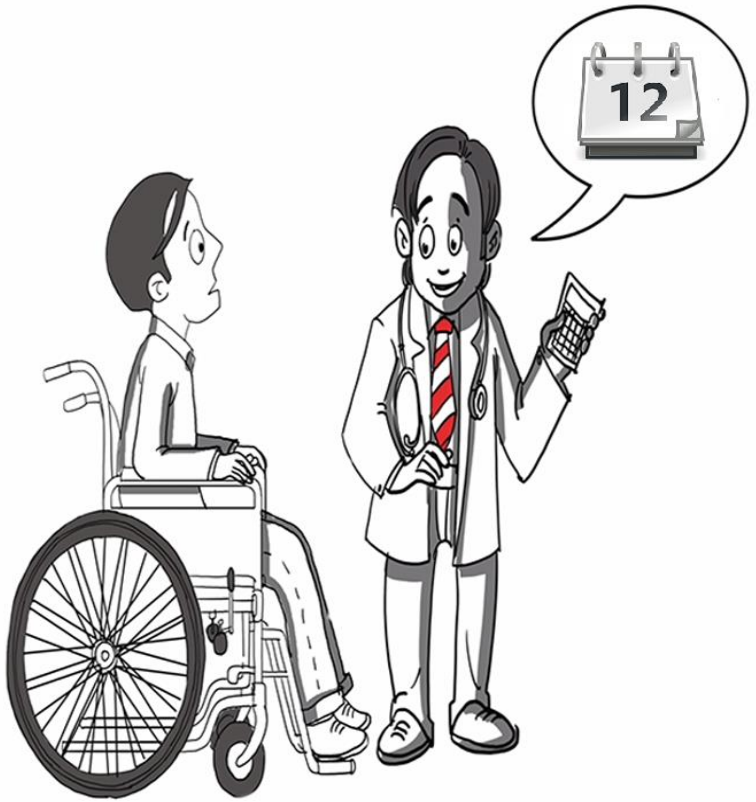
| *Regression Statistics* | |
|---|---|
| Multiple R | 0.996442 |
| R Square | 0.992896 |
| Standard Error | 1.69286 |
| Observations | 9 |

ANOVA

| | *Df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 1 | 2803.94 | 2803.94 | 978.4229 | 8.82E-09 |
| Residual | 7 | 20.06042 | 2.865775 | | |
| Total | 8 | 2824 | | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
|---|---|---|---|---|---|---|
| Intercept | 1.996718 | 0.76169 | 2.62143 | 0.034338 | 0.195607 | 3.79783 |
| FB | 0.046596 | 0.00149 | 31.27975 | 8.82E-09 | 0.043074 | 0.050119 |



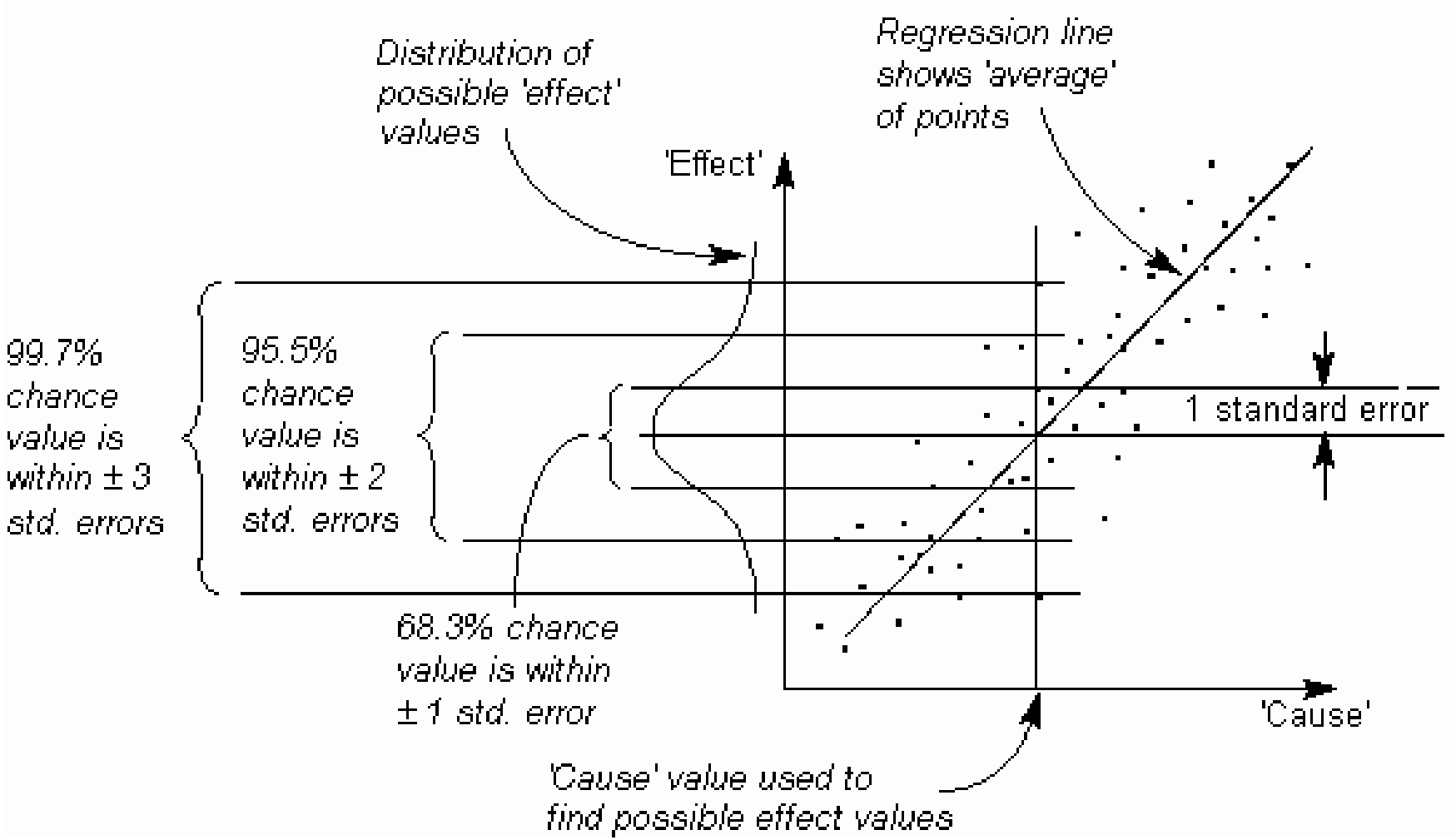



# Standard Error of Estimate

- Standard error is the estimate of the standard deviation of the regression errors.

- Standard error of estimate, $S_e$, measures the variability or scatter of the observed values around the regression line.

$$S_e = \sqrt{\frac{\sum \left( Y_i - \hat{Y}_i \right)^2}{n-2}}$$



# Interpreting the Standard Error of Estimate

- A smaller standard error of estimate indicates better fit.
- The larger the standard error of estimate, the greater the scattering of points around the regression line.
- If $S_e = 0$, then we can expect a perfect fit.

**SUMMARY OUTPUT**

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.351 |
| **R Square** | 0.123 |
| Adjusted R Square | 0.120 |
| Standard Error | 111886.046 |
| Observations | 247.000 |

ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 1.000 | 4.30902E+11 | 4.30902E+11 | 34.4212328 | 1.4321E-08 |
| Residual | 245.000 | 3.06703E+12 | 12518487400 | | |
| Total | 246.000 | 3.49793E+12 | | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
|---|---|---|---|---|---|---|---|---|
| Intercept | **129110.795** | **13564.94009** | 9.517977506 | 1.76841E-18 | 102392.0146 | 155829.5747 | 102392.0146 | 155829.5747 |
| 49.2 | **1807.591** | **308.0966635** | 5.866961121 | 1.4321E-08 | 1200.735021 | 2414.447272 | 1200.735021 | 2414.447272 |

IIMBX

# Standard Error of Estimate for Regression Coefficients

Standard error of estimate for regression coefficient measures the amount of sampling error in a regression coefficient.

# Standard error of $\beta_0$ and $\beta_1$

Standard error of $\beta_0$ and $\beta_1$ is given by:

$$S(\beta_0) = \frac{S_e \times \sqrt{\sum x^2}}{\sqrt{nSS_x}}$$

$$S(\beta_1) = \frac{S_e}{\sqrt{SS_x}}$$

$$SS_x = \sum_i (X_i - \bar{X})^2$$