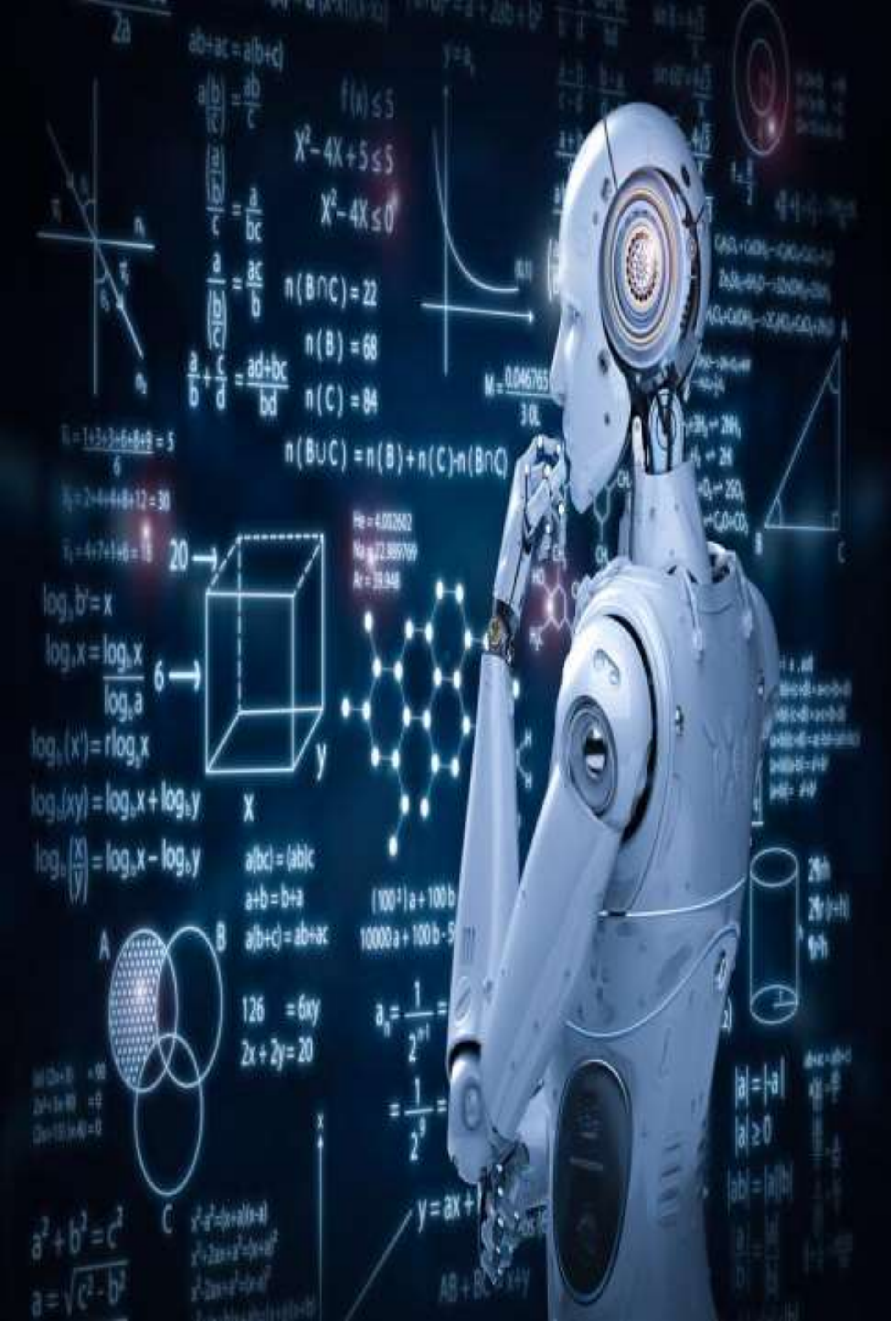


Practical Machine Learning

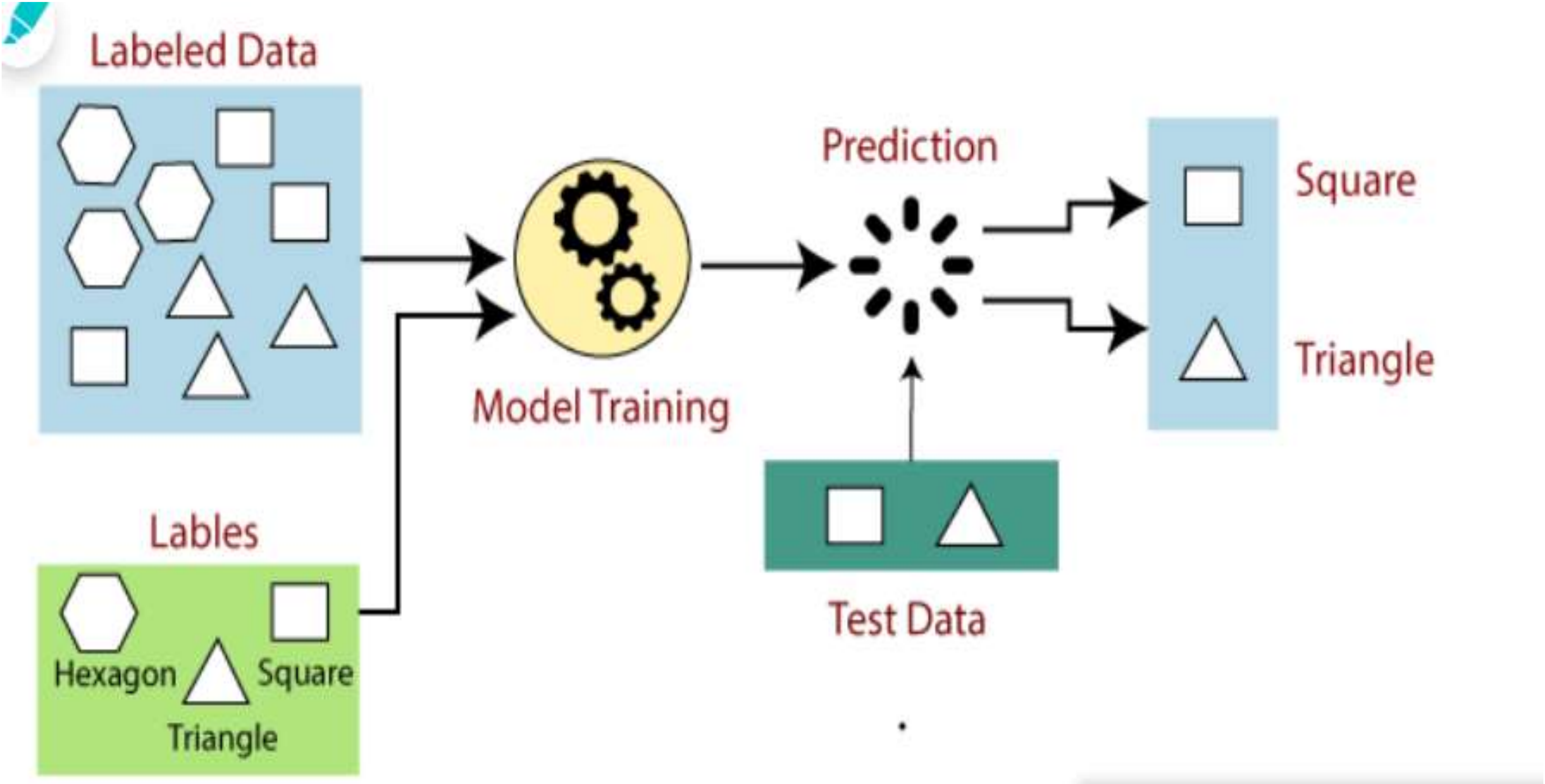
Day 5: Mar23 DBDA

Kiran Waghmare

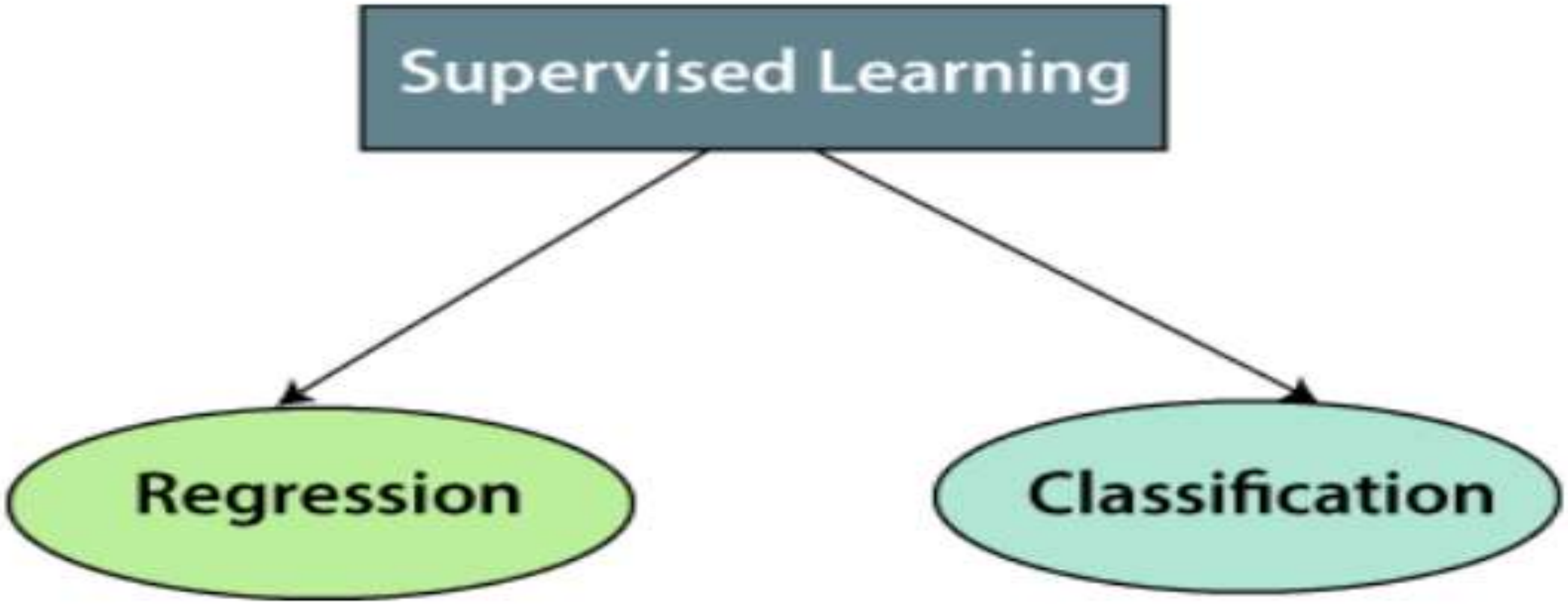
Agenda

- Regression
- Types of Regression

How Supervised Learning Works?



Types of supervised Machine learning Algorithms:



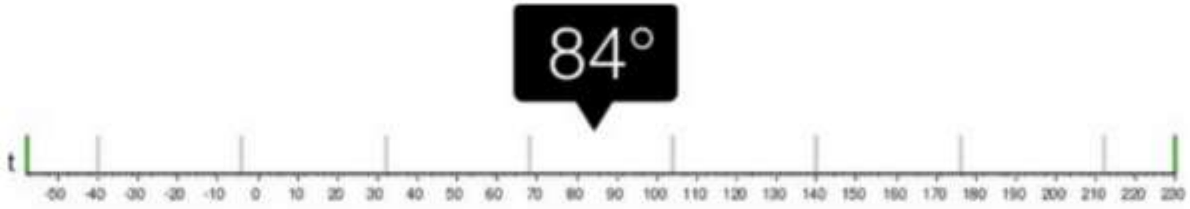
Supervised Learning

Regression

Classification

Regression

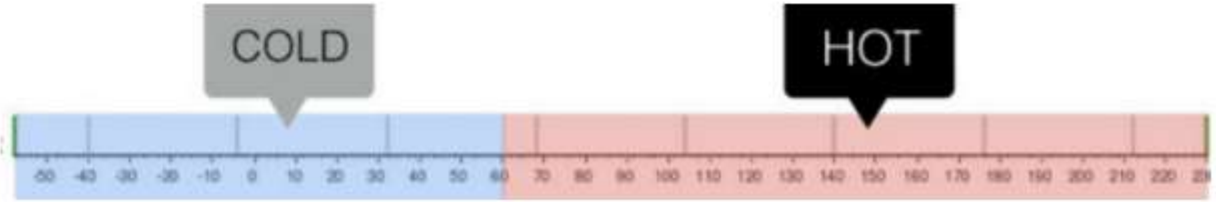
What will be the temperature tomorrow?



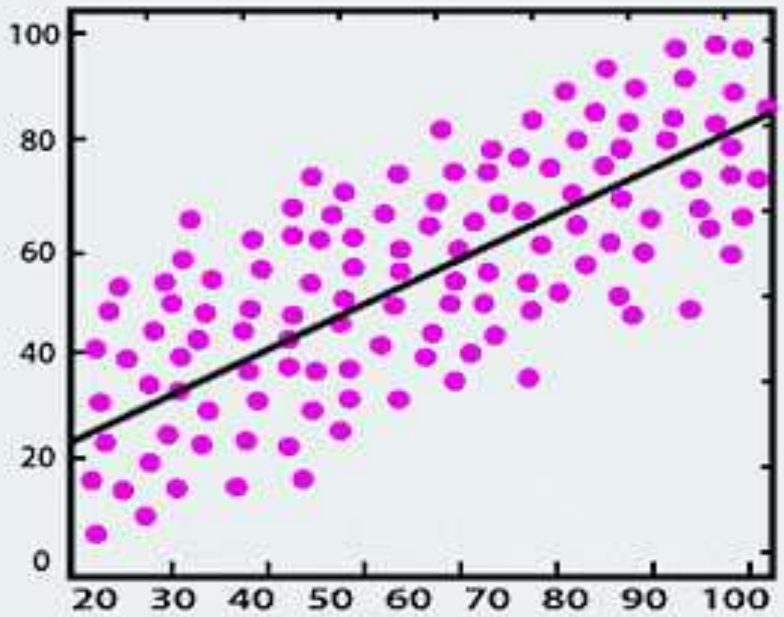
Fahrenheit

Classification

Will it be hot or cold tomorrow?

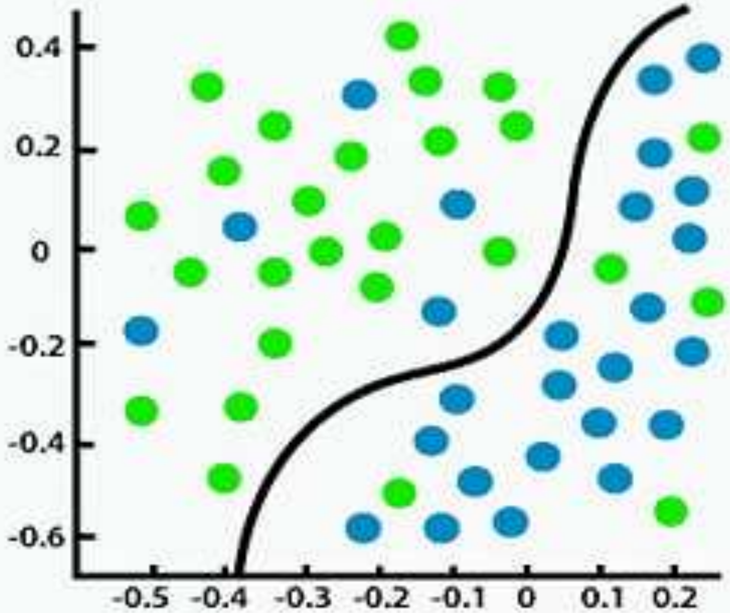


Fahrenheit



Regression

versus



Classification

Regression

- Regression is used to study the relationship between two variables.

- We can use simple regression if both the dependent variable (DV) and the independent variable (IV) are numerical.

- If the DV is numerical but the IV is categorical, it is best to use ANOVA.

Simple Linear Regression

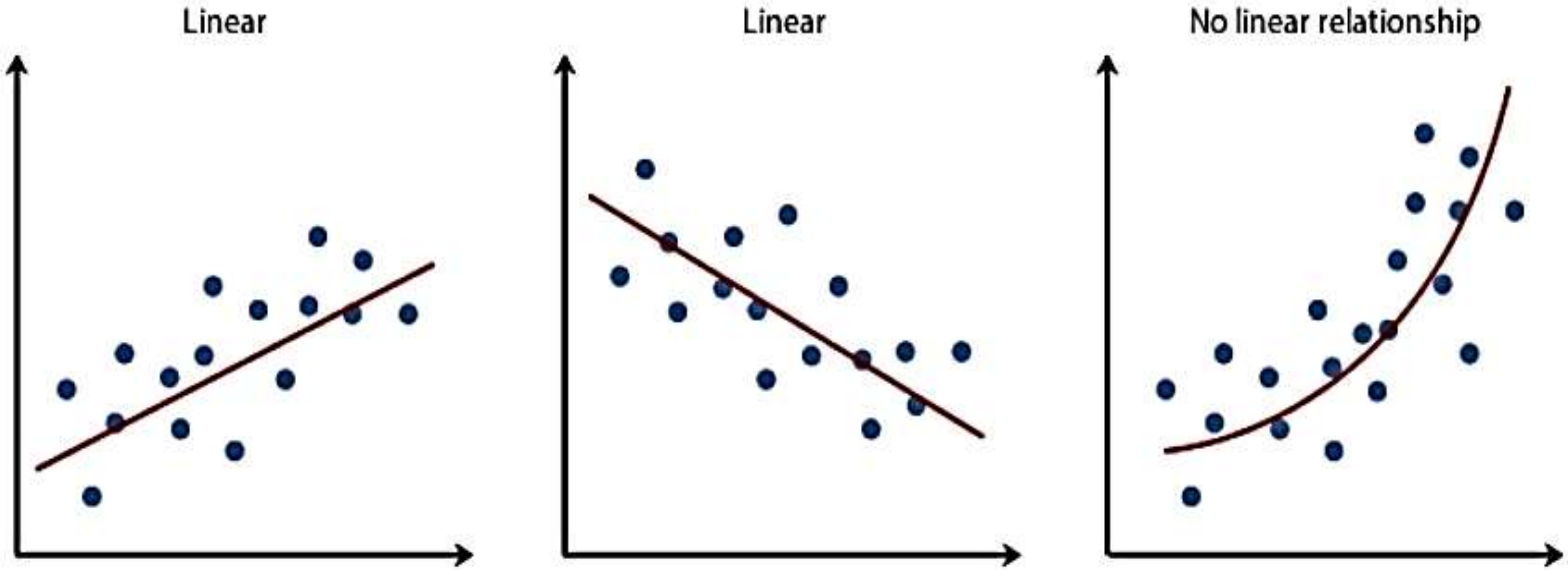
Linear Regression Key Components

- **Straight Line Equation**: y = mx + b
- **Dependent Variable (y)**: variable that is being estimated and predicted, also known as target
- **Independent Variable (x)**: input variable, also known as predictors or features
- **Coefficient**: is a numerical constant, also known as parameter
- **Slope (m)** : determines the angle of the line
- **Intercept (b)**: constant determining the value of y when x is 0

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

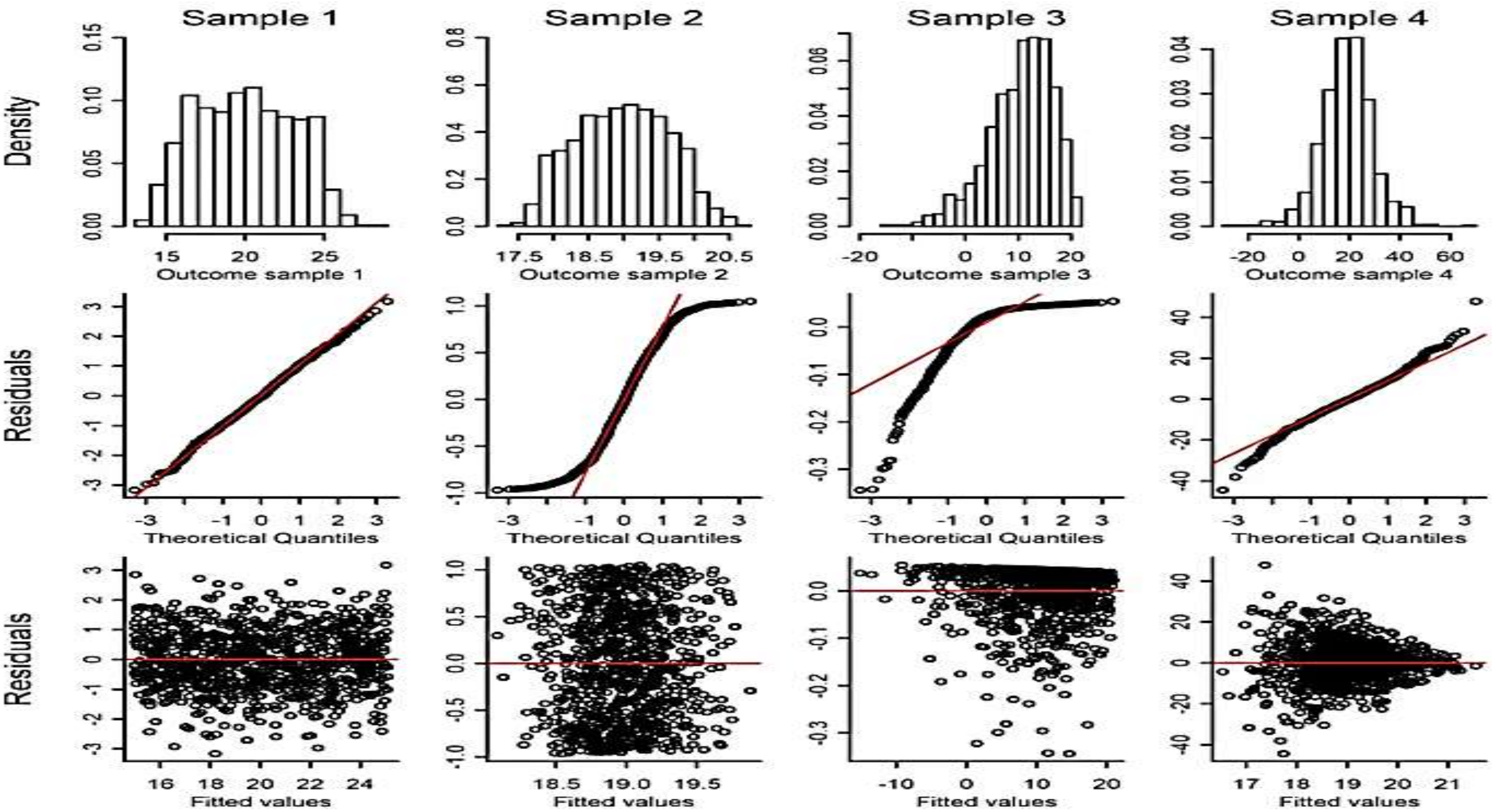
Assumptions of Linear Regression

- There are four assumptions associated with a linear regression model. If these assumptions are violated, it may lead to biased or misleading results.
- **Linearity**: relationship between independent variable(s) and dependent variable is linear
- if not respected, regression will underfit and will not accurately model the relationship between independent and dependent variables
- if there is no linear relationship, various methods can be used to make the relationship linear such as polynomial and exponential transformations for both independent and dependent variables


Linear
Linear
No linear relationship

Normality

- **Normality**: model residuals should follow a normal distribution
- if distribution is not normal, regression results will be biased and it may highlight that there are outliers or other assumptions being violated
- correct the large outliers in the data and verify if the other assumptions are not being violated

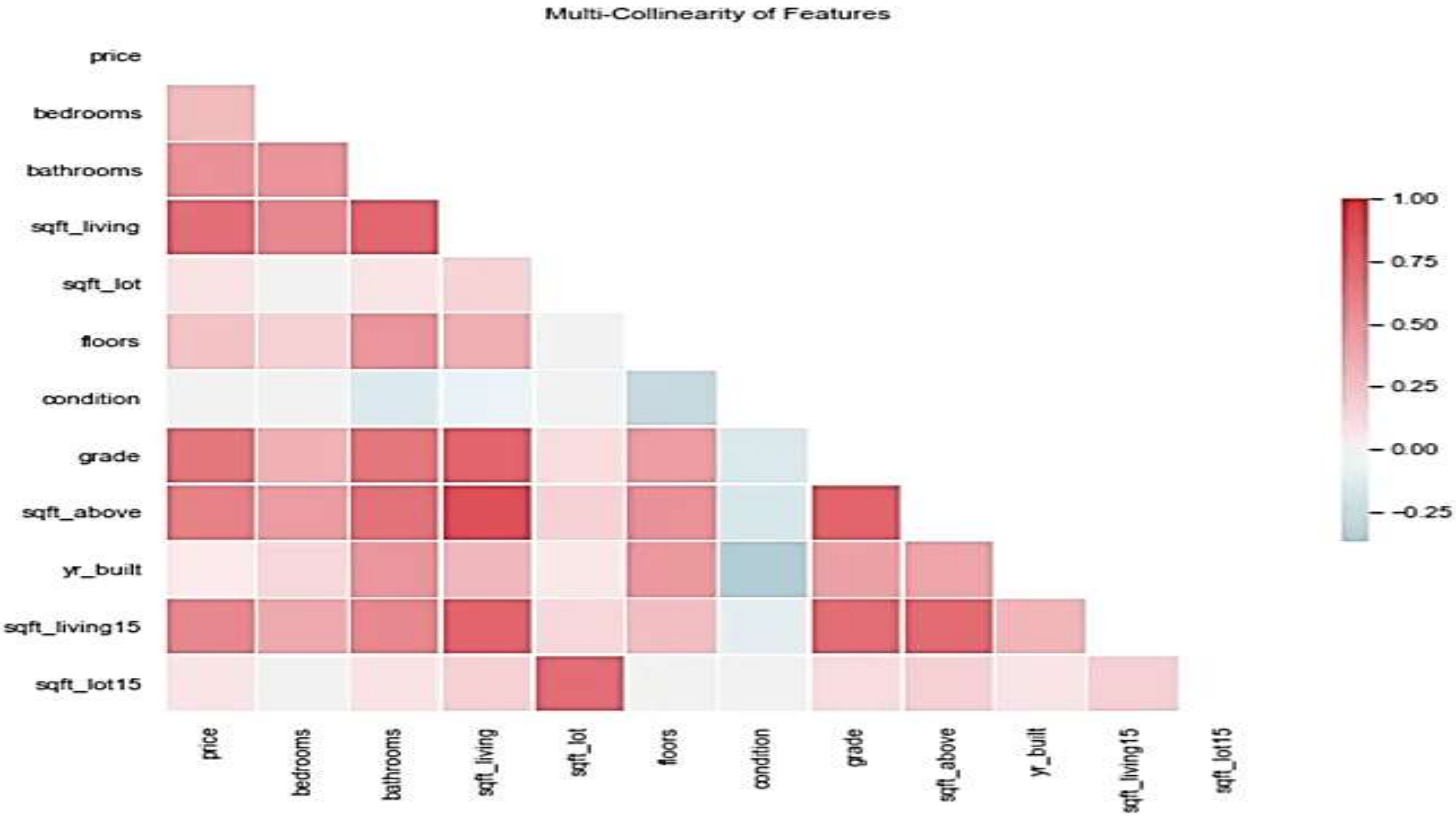


Sample 1 follows a normal distribution

Independence

- **Independence**: each independent variable should be independent from other independent variables
- multicollinearity is when independent variables are not independent from each other
- it indicates that changes in one predictor are associated with changes in another predictor
- we use heatmaps and calculate VIF (Variance Inflation Factors) scores which compares each independent variable's collinearity with other independent variables
-

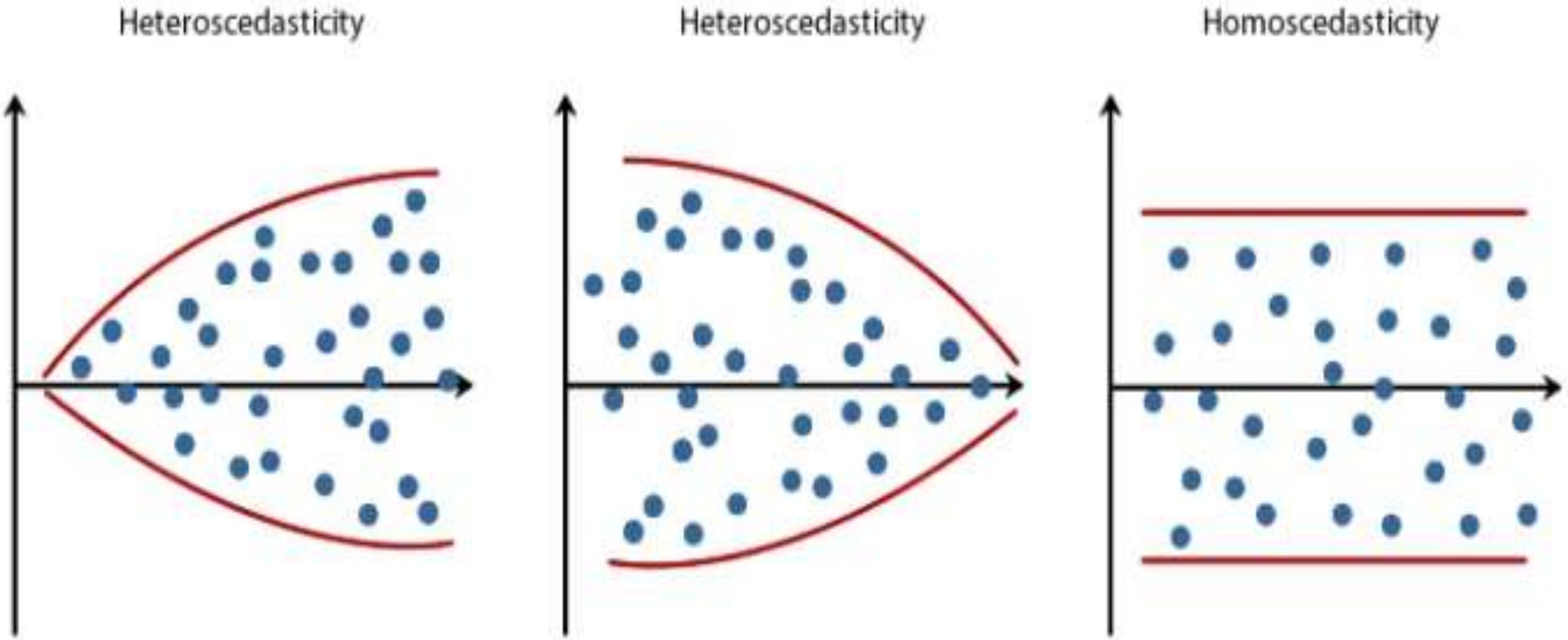
Homoscedasticity: the variance of residual is the same for any value of x



How can we validate our regression model results?

Homoscedasticity

- **Homoscedasticity**: the variance of residual is the same for any value of x, fancy word for “equal variances”
- the model does not fit all parts of the model equally which lead to biased predictions
- it can be tackled by reviewing the predictors and providing additional independent variables (and maybe even check that the linearity assumption is respected as well)

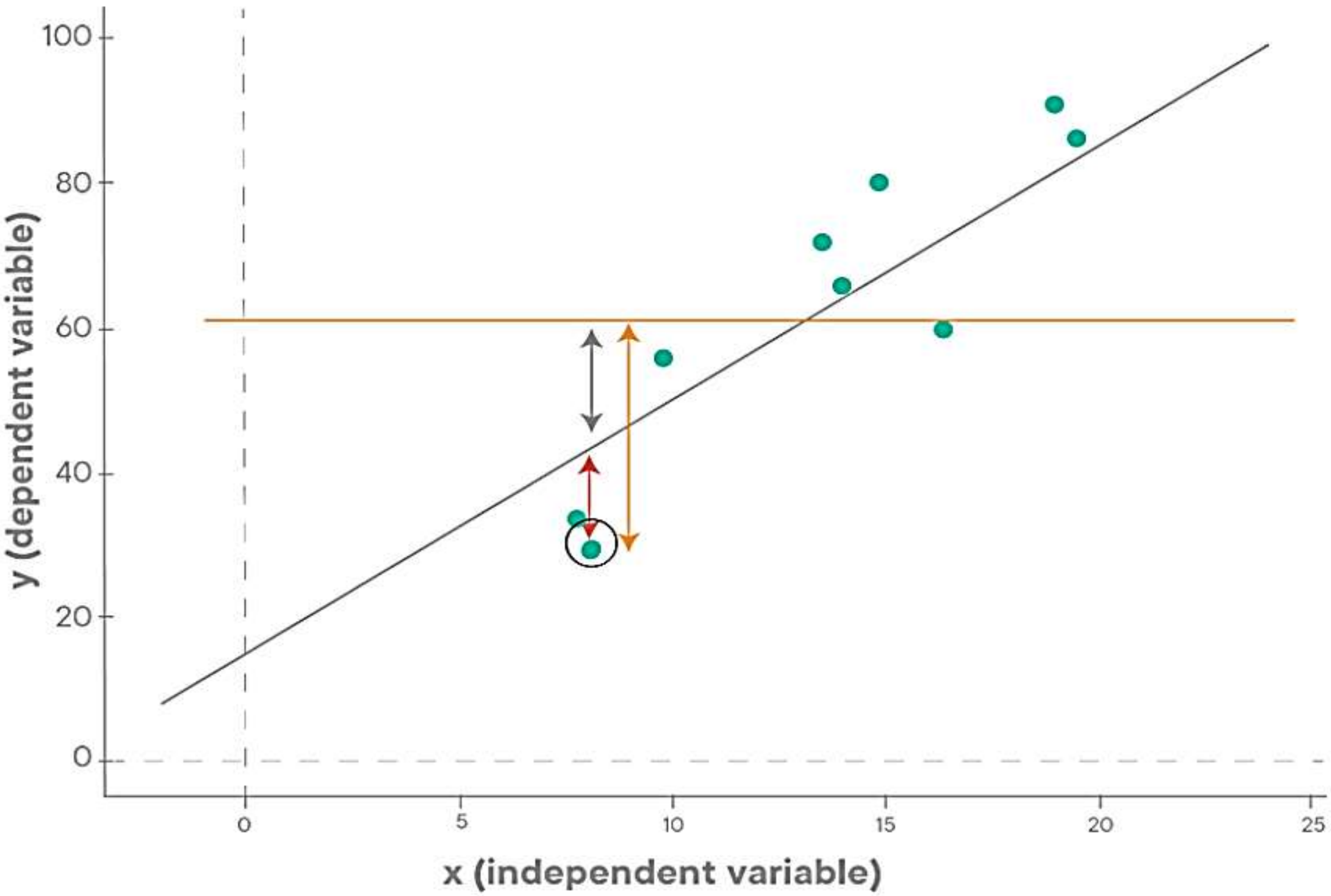

Heteroscedasticity
Heteroscedasticity
Homoscedasticity

How to determine the line of best fit?

- R-Squared (Coefficient of Determination): statistical measure that is used to assess the goodness of fit of a regression model

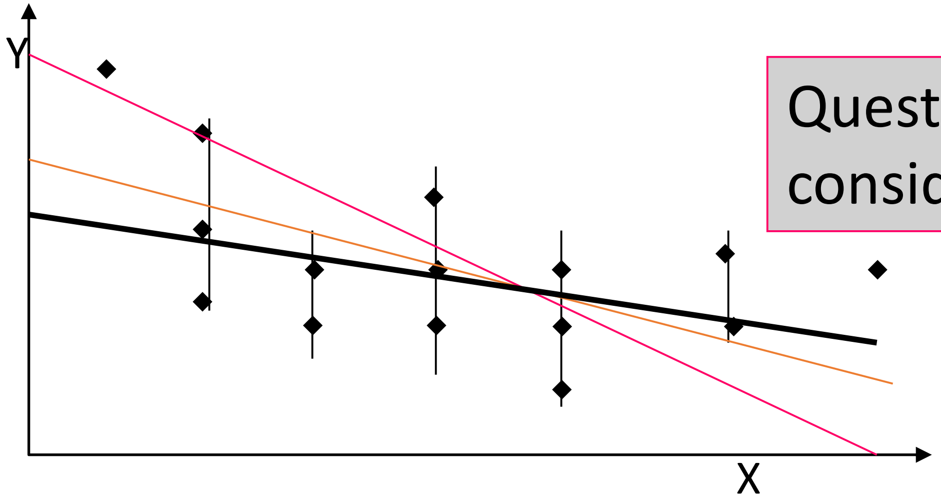
- It uses a baseline model that finds the mean of the dependent variable (y) and compares it with the regression line (yellow line below)

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$


100
80
60
40
20
0
y (dependent variable)
0
5
10
15
20
25
x (independent variable)

Estimating the Coefficients

- The estimates are determined by
  - drawing a sample from the population of interest,
  - calculating sample statistics.
  - producing a straight line that cuts into the data.



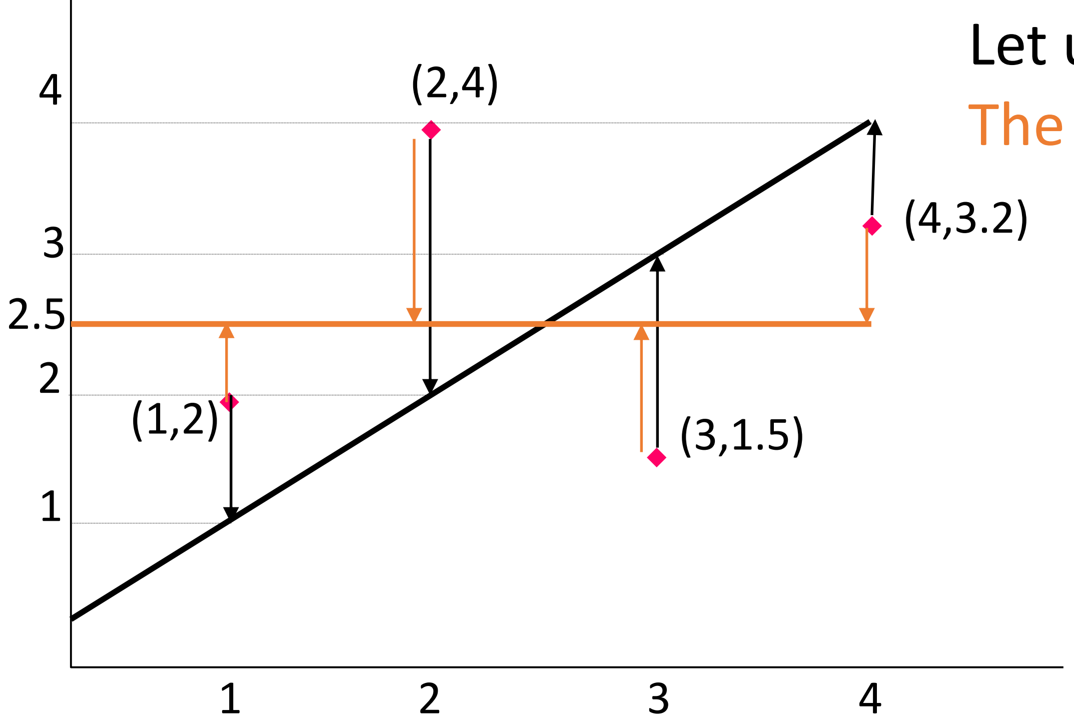
Question: What should be considered a good line?

The Least Squares (Regression) Line

A good line is one that minimizes
the sum of squared differences between the points and the line.

Sum of squared differences $=(2 - 1)^2 +(4 - 2)^2 +(1.5 - 3)^2 +(3.2 - 4)^2 = 6.89$

Sum of squared differences $=(2 -2.5)^2 +(4 - 2.5)^2 +(1.5 - 2.5)^2 +(3.2 - 2.5)^2 = 3.99$



Let us compare two lines

The second line is horizontal

The smaller the sum of squared differences the better the fit of the line to the data.

The Estimated Coefficients

To calculate the estimates of the line coefficients, that minimize the differences between the data points and the line, use the formulas:

$$b_1 = \frac{\operatorname{cov}(X,Y)}{s_X^2} \left(= \frac{s_{XY}}{s_X^2} \right)$$

$$b_0 = \overline{Y} - b_1 \overline{X}$$

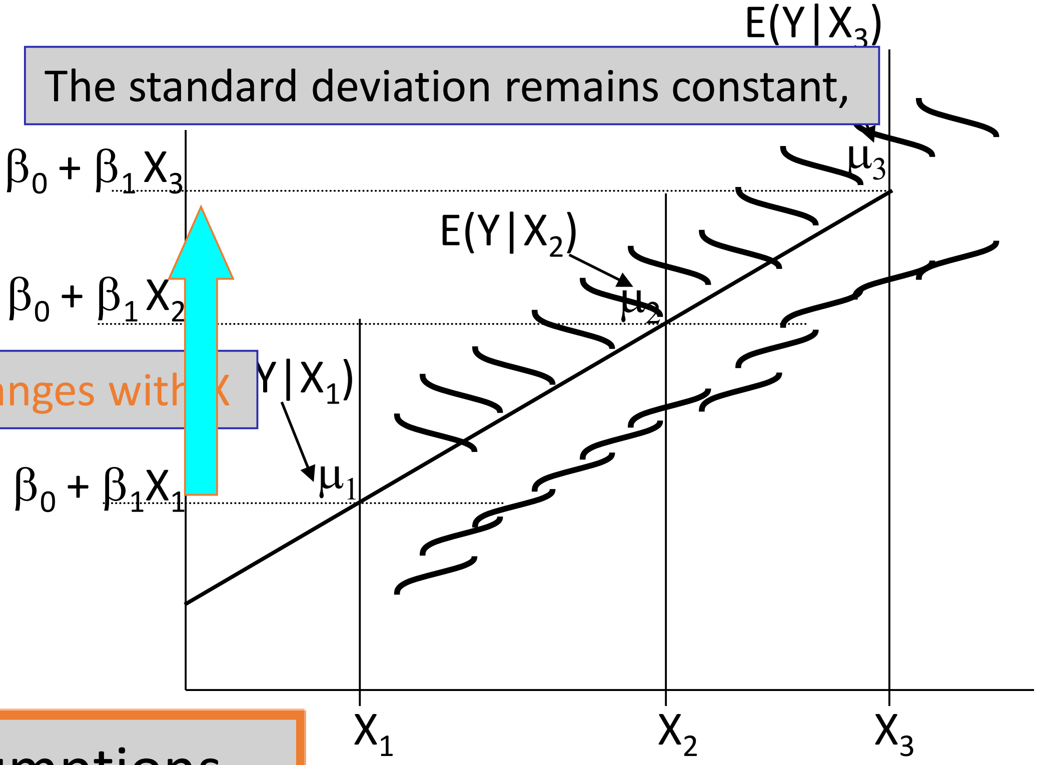
The regression equation that estimates the equation of the first order linear model is:

$$\hat{Y} = b_0 + b_1 X$$

Error Variable: Required Conditions

- The error ε is a critical part of the regression model.
- Four requirements involving the distribution of ε must be satisfied.
  - The probability distribution of ε is normal.
  - The mean of ε is zero: $E(\varepsilon) = 0$.
  - The standard deviation of ε is σ_ε for all values of X.
  - The set of errors associated with different values of Y are all independent.

The Normality of ε



The standard deviation remains constant,

but the mean value changes with X

From the first three assumptions we
have: Y is normally distributed
with mean $E(Y) = \beta_0 + \beta_1 X$, and a
constant standard deviation σ_ε

Sum of Squares for Errors

- This is the sum of differences between the points and the regression line.
- It can serve as a measure of how well the line fits the data. SSE is defined by

$$\boxed{\mathrm{SSE} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 .}$$

  - A shortcut formula

$$\boxed{\mathrm{SSE} = (n-1)s_Y^2 - \frac{[\mathrm{cov}(X,Y)]^2}{s_X^2}}$$

Standard Error of Estimate

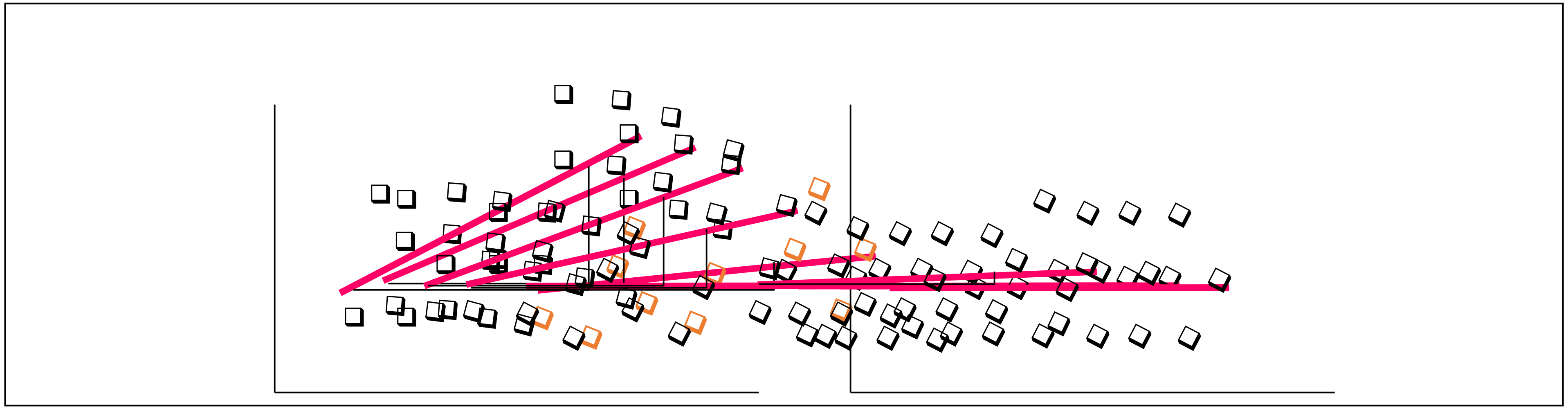
- The mean error is equal to zero.
- If σ_ε is small the errors tend to be close to zero (close to the mean error). Then, the model fits the data well.
- Therefore, we can, use σ_ε as a measure of the suitability of using a linear model.
- An estimator of σ_ε is given by s_ε

$$S\tan dard\ \ Error\ \ of\ \ Estimate$$

$$s_\varepsilon = \sqrt{\frac{\mathrm{SSE}}{n-2}}$$

Testing the Slope

- When no linear relationship exists between two variables, the regression line should be horizontal.



Linear relationship.
Different inputs (X) yield different outputs (Y).

The slope is not equal to zero

No linear relationship.
Different inputs (X) yield the same output (Y).

The slope is equal to zero

- We can draw inference about β_1 from b_1 by testing

  H_0: $\beta_1 = 0$
  H_1: $\beta_1 \neq 0$ (or < 0,or > 0)

  - The test statistic is

$$t = \frac{b_1 - \beta_1}{s_{b_1}}$$

where

$$s_{b_1} = \frac{s_\varepsilon}{\sqrt{(n-1)s_X^2}}$$

The standard error of b_1.

  - If the error variable is normally distributed, the statistic has Student t distribution with d.f. = n-2.

Coefficient of Determination

- To measure the strength of the linear relationship we use the coefficient of determination:

$$R^2 = \frac{[\mathrm{cov}(X,Y)]^2}{s_X^2 s_Y^2} \quad \left(\text{or, } = r_{XY}^2\right);$$

$$\text{or, } R^2 = 1 - \frac{\text{SSE}}{\sum (Y_i - \bar{Y})^2} \qquad \text{(see p. 18 above)}$$

- R^2 measures the proportion of the variation in Y that is explained by the variation in X.

$$R^2 = 1 - \frac{\text{SSE}}{\sum (Y_i - \bar{Y})^2} = \frac{\sum (Y_i - \bar{Y})^2 - \text{SSE}}{\sum (Y_i - \bar{Y})^2} = \frac{\text{SSR}}{\sum (Y_i - \bar{Y})^2}$$

- R^2 takes on any value between zero and one.

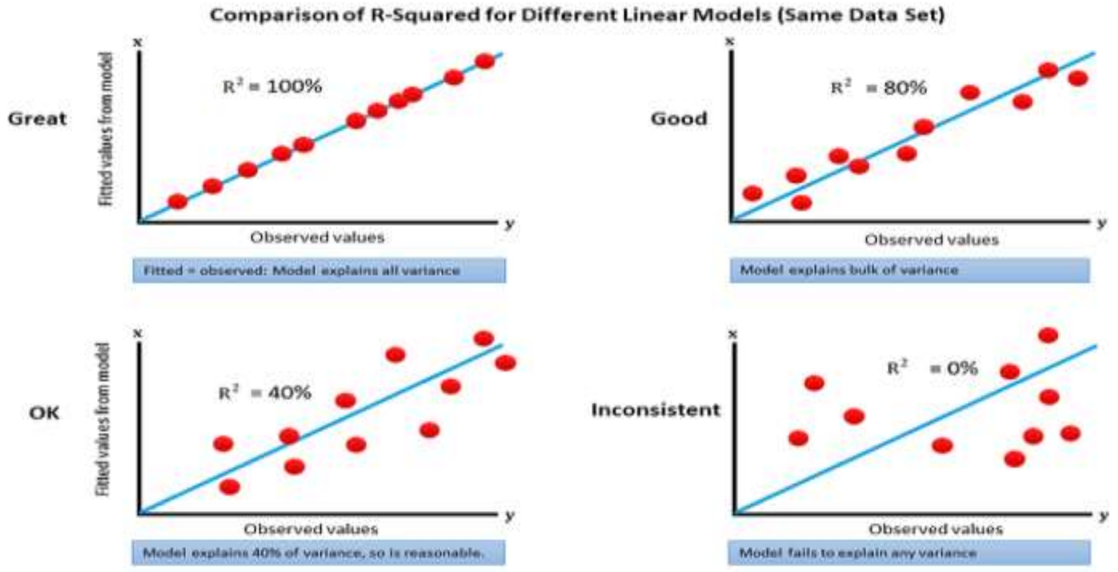
  R^2 = 1: Perfect match between the line and the data points.

  R^2 = 0: There are no linear relationship between X and Y.

- **Residual Sum of Squared Errors (RES) :** also known as SSE and RSS, is the sum of squared difference between y and predicted y (red arrow)

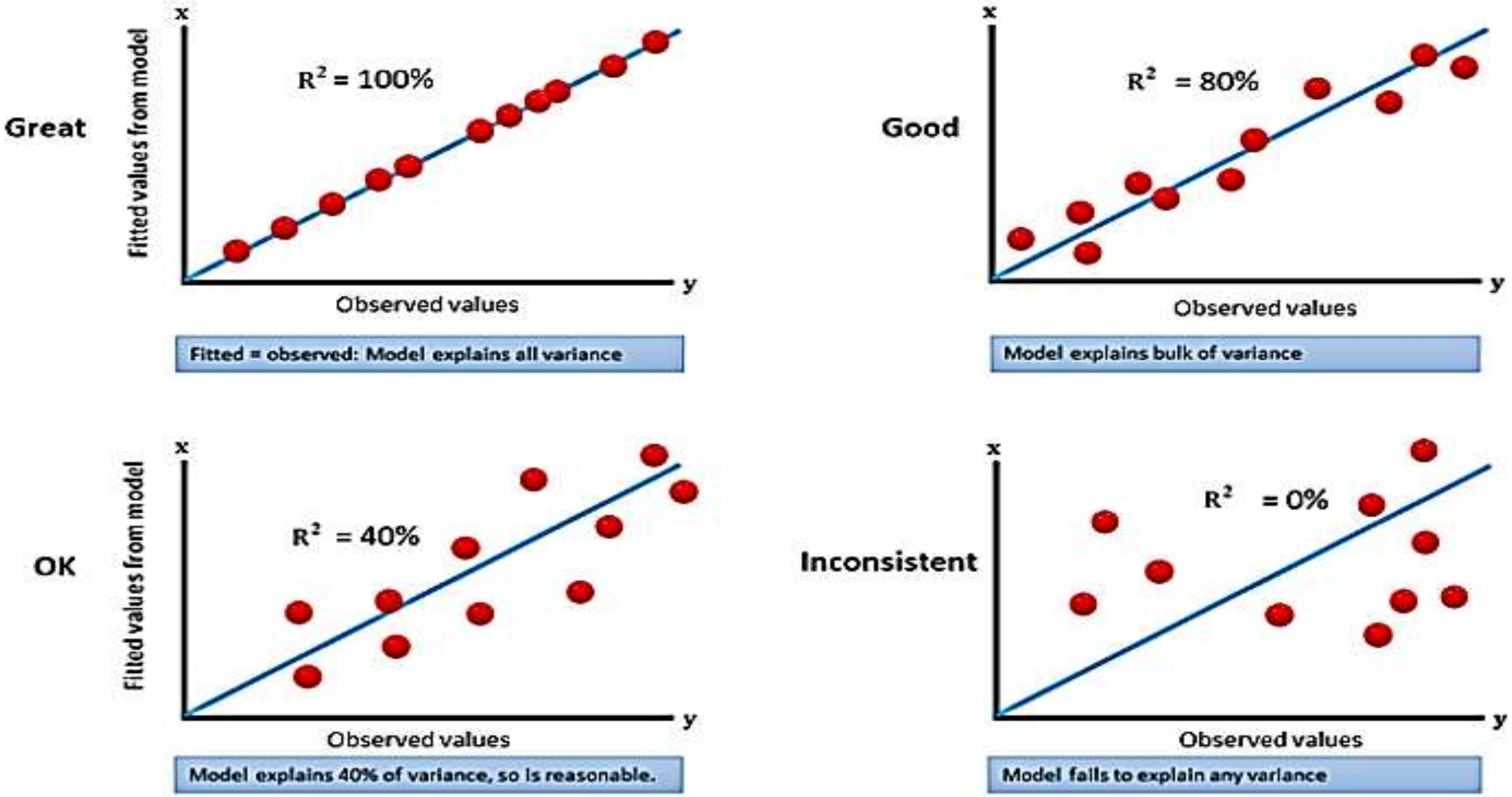
- **Total Sum of Squared Errors (TOT):** also known as TSS, is the sum of squared difference between y and predicted y (orange arrow)

- R-Squared can take a value between 0 and 1 where values closer to 0 represents a poor fit and values closer to 1 represent an (almost) perfect fit



85% of the variations in dependent variable y are explained by the

Comparison of R-Squared for Different Linear Models (Same Data Set)

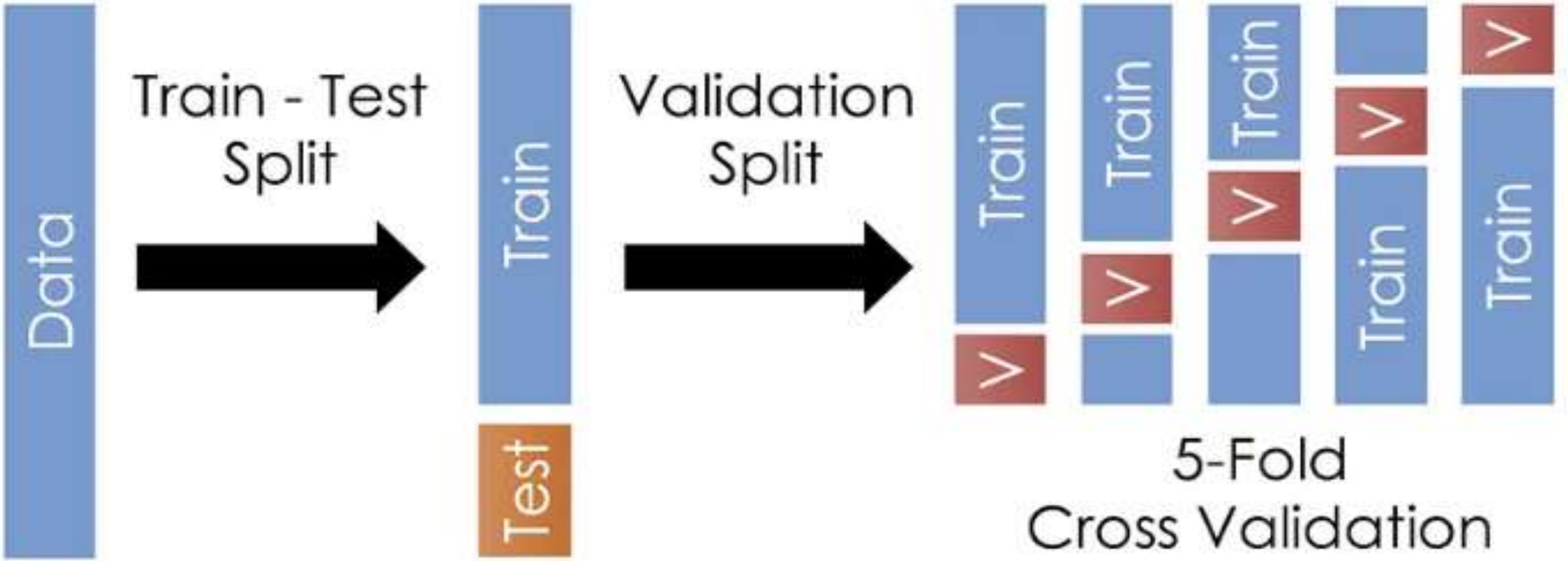


Fitted = observed: Model explains all variance

Model explains bulk of variance

Model explains 40% of variance, so is reasonable.

Model fails to explain any variance


Data
Train - Test Split
Train
Test
Validation Split
Train
V
Train
V
Train
V
V
Train
V
Train
5-Fold Cross Validation

- To calculate the results for both train and test data, a popular metric is the Root Mean Squared Error (RMSE).

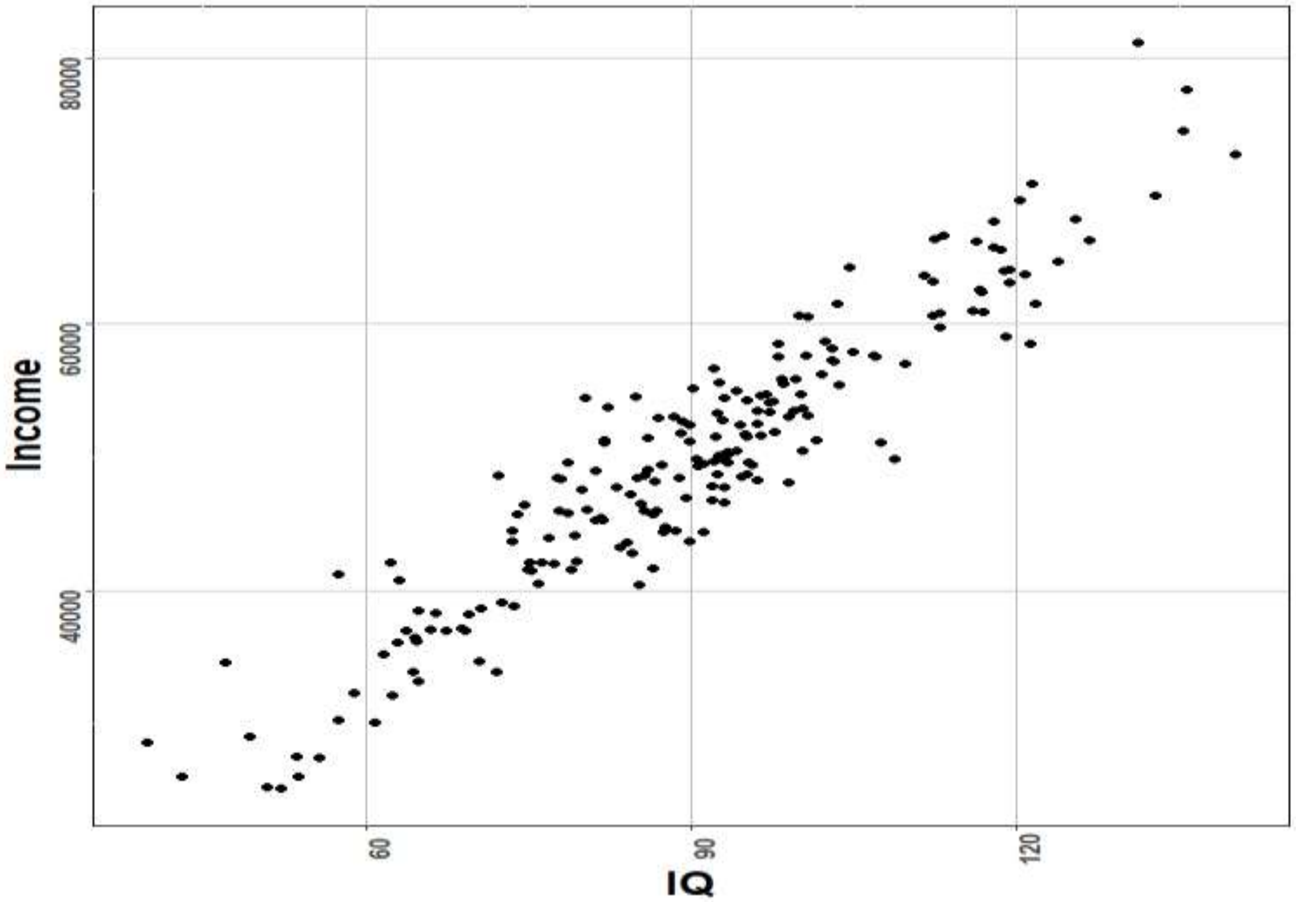
- **Mean Squared Error (MSE)**: average squared difference between the estimated values and the actual value
- the smaller the MSE, the closer the fit is to the data

- **Root Mean Squared Error (RMSE)**: square root of MSE
- easier to interpret since it is the same units as the quantity plotted on the x axis
- the RMSE is the distance on average of a data point from the fitted line, measured along a vertical line

- **5-Fold Cross Validation**
- Running a model with different Train-Test Split will lead to different results. This is where **5-Fold Cross Validation** comes in where we split the data into *"k"* equal sections of data, with each linear model using a different section of data as the test data and all other sections combined as the training set.

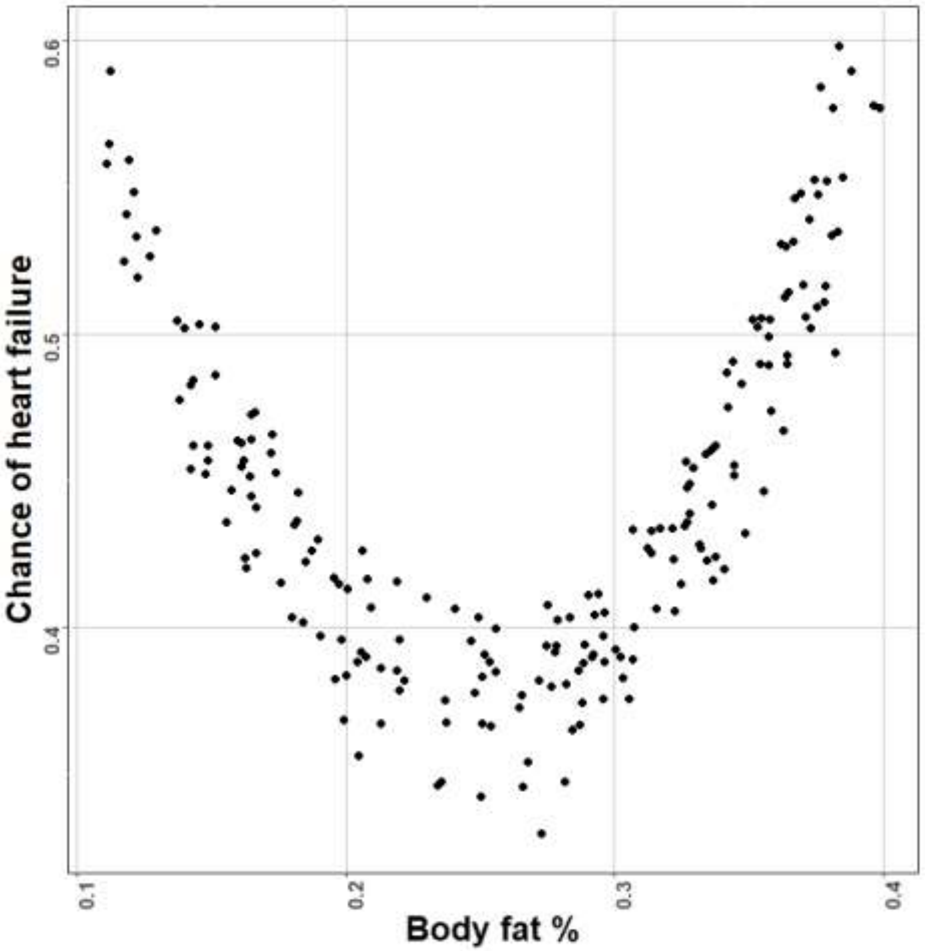
Displaying the data

When both the DV and IV are numerical, we can represent data in the form of a scatterplot.



Displaying the data

It is important to perform a scatterplot because it helps us to see if the relationship is linear.



In this example, the relationship between body fat % and chance of heart failure is not linear and hence it is not sensible to use linear regression.