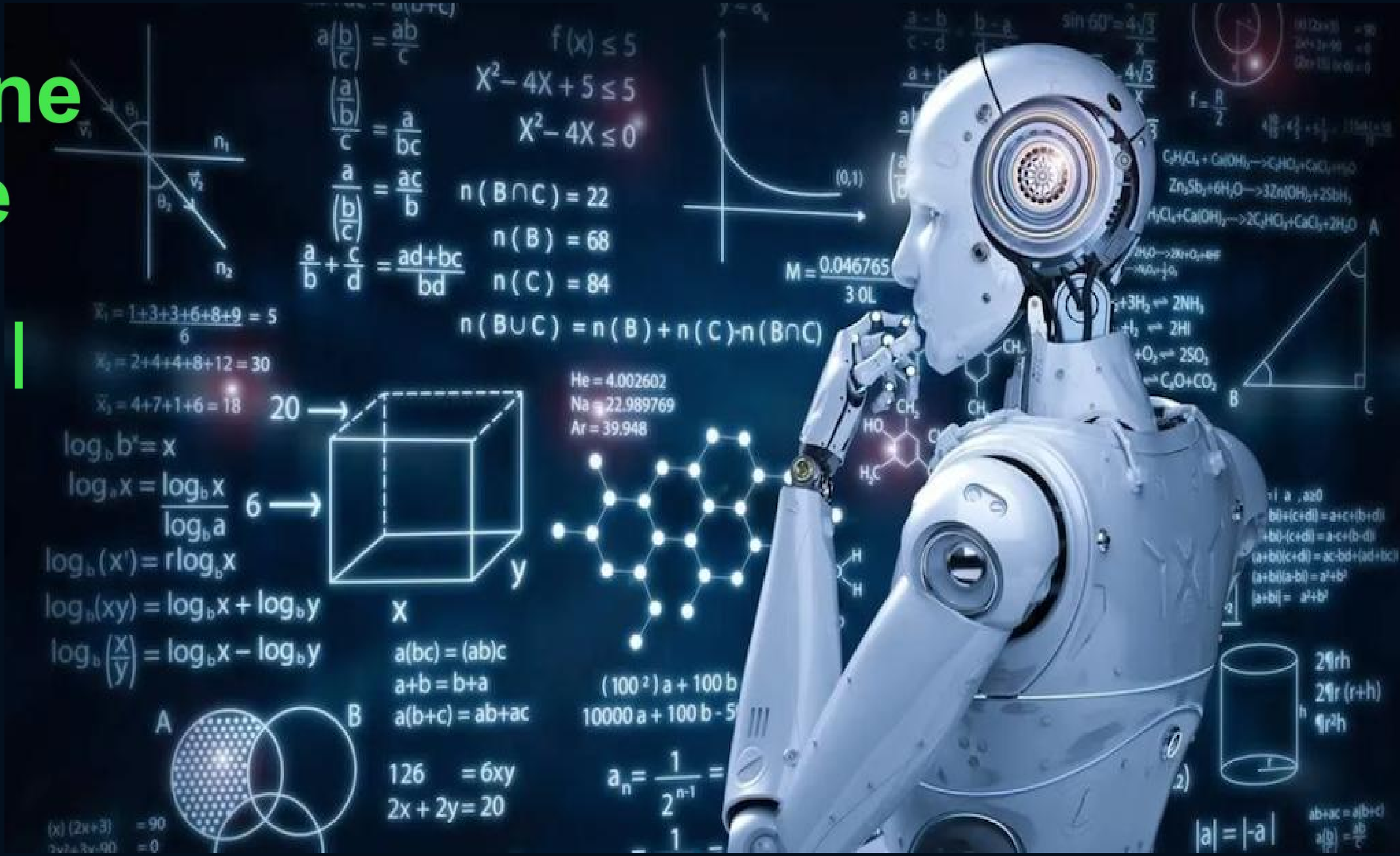
NBS Nottingham Business School

Principles of Machine Learning in Finance

2. Supervised Learning | Linear Regression

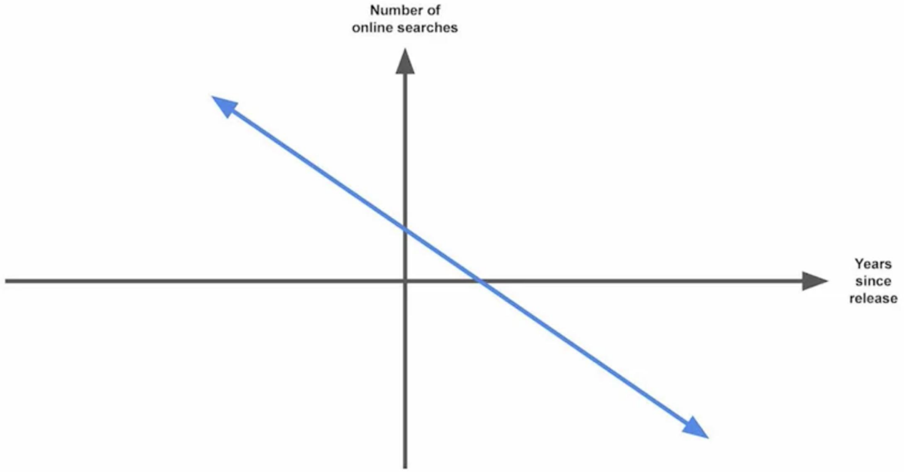


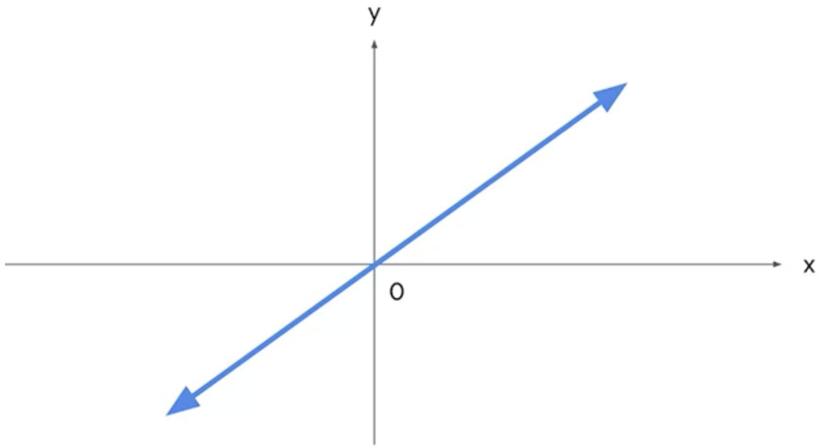
Learning Outcomes

- Regression Analysis
- Linear Regression
- Simple Regression
- Multiple Regression
- **Coding Activity 2**: Supervised ML. Linear Regression || [Regression Model For a Financial Dataset. Stock Price Prediction with Python]

Regression Analysis Overview

Regression analysis is about estimating relationships between a single dependent variable and one or more independent variables





Linear regression is a technique that estimates the linear relationship between a continuous dependent variable y and one or more independent variables x.

Example 1.Continious vs Categorical Variables

Continuous Variables	Categorical Variables
Takes on any real value between minimum and maximum value	Have a finite number of possible values
Examples: Product sales Vehicle speed Time spent on webpage	*Examples:* Types of products Educational level

Dependend and Independent Variables

- **Dependend variable (Y)**: The variable the given model estimates, also referred to as a response or outcome variable
- **Independent variable (X)**: A variable that explains trends in the dependent variable, also referred to explanatory or predictor variable

Simple Linear Regression

$$y_i = \beta_0 + \beta_1 \bullet X_i$$

where:

y_i is an i-observed value; X_i is an i-independent variable.

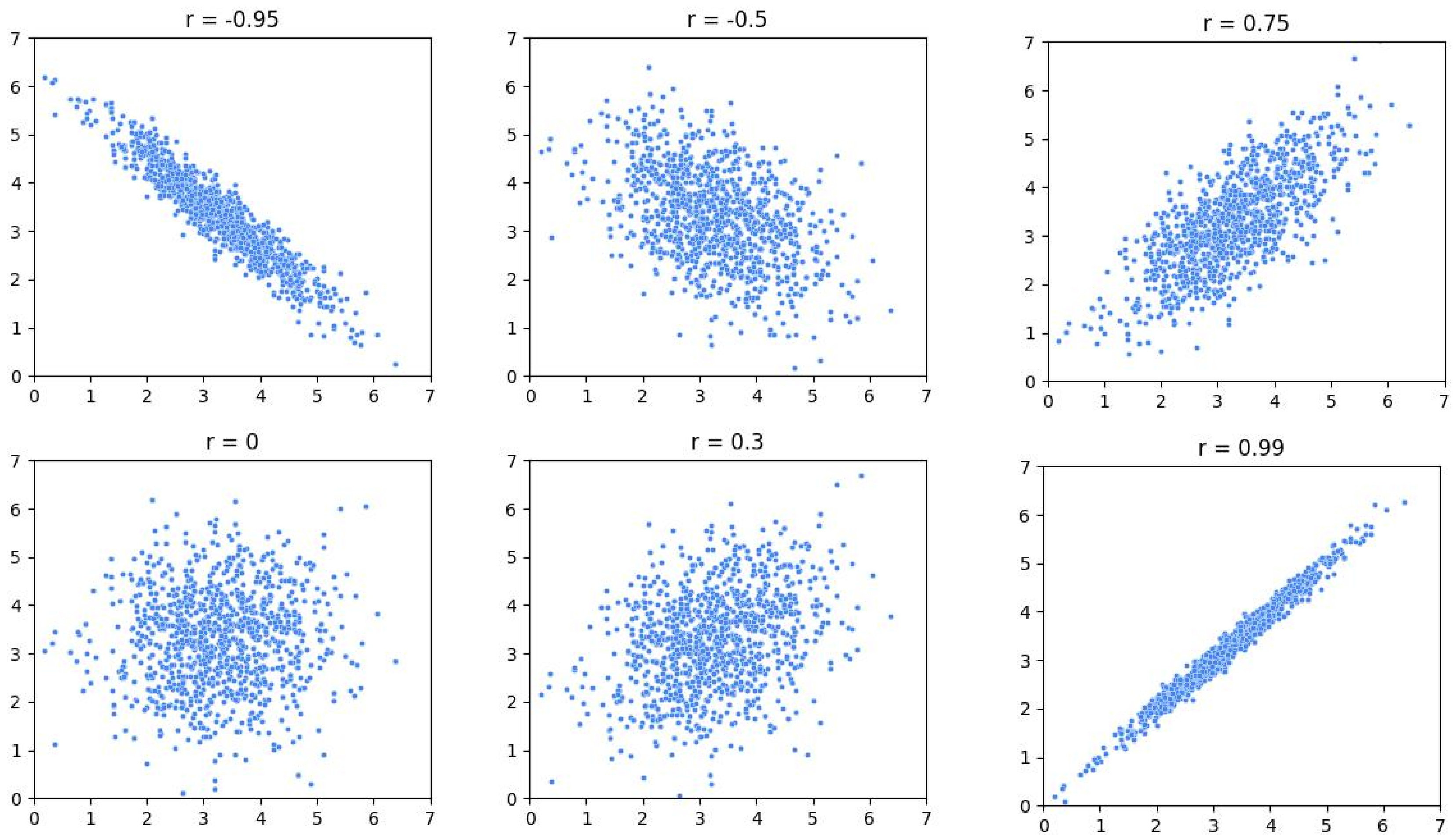
Slope is the amount that y increases or decreases per one-unit increase of X.

Intercept is the value of y, the dependent variable when x, the independent variable equals 0.

Correlation

$$\rho_{x,y} = \frac{Cov(x,\ y)}{\sigma_x \bullet \sigma_y}$$

$$\rho_{x,y} \in [-1;\ 1]$$



Causation

Positive correlation is a relationship between two variables that tend to increase of decrease together:

$$\rho^{+} \in (0;\ 1]$$

Negative correlation is an inverse relationship between two variables, where one variable increases, the other tends to decrease and vice versa:

$$\rho^{-} \in [-1;\ 0)$$

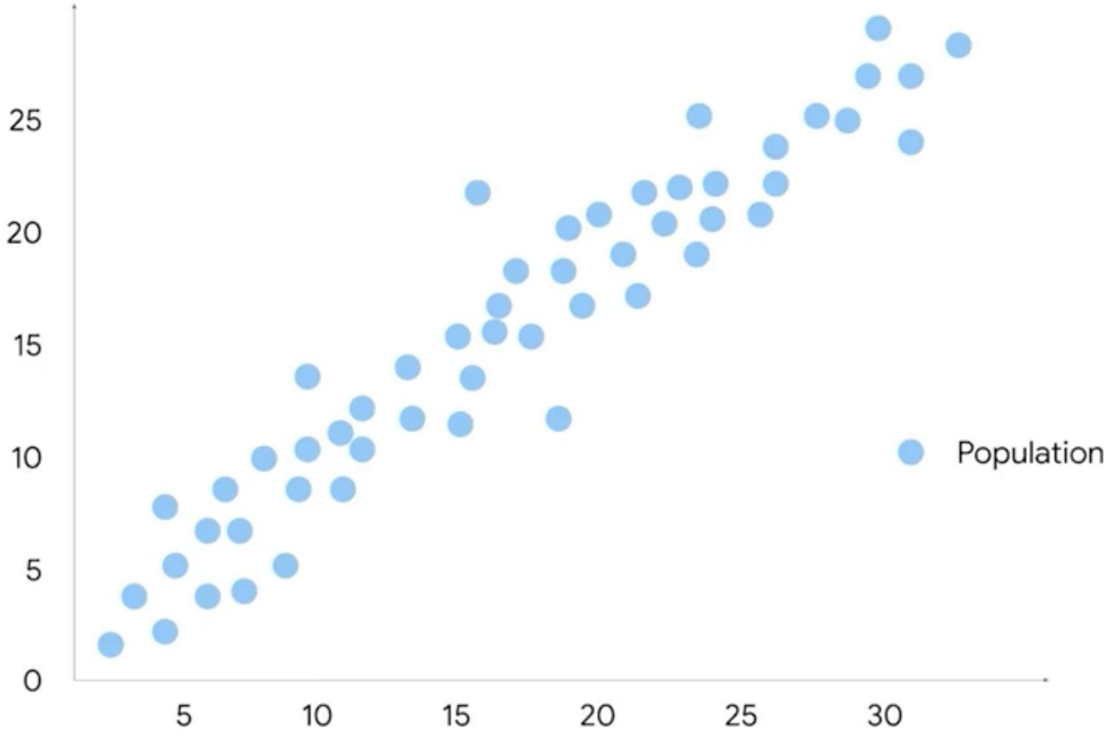
Causation is a cause-and-effect relationship where one variable directly causes the other to change in a particular

Linear Regression: Overview

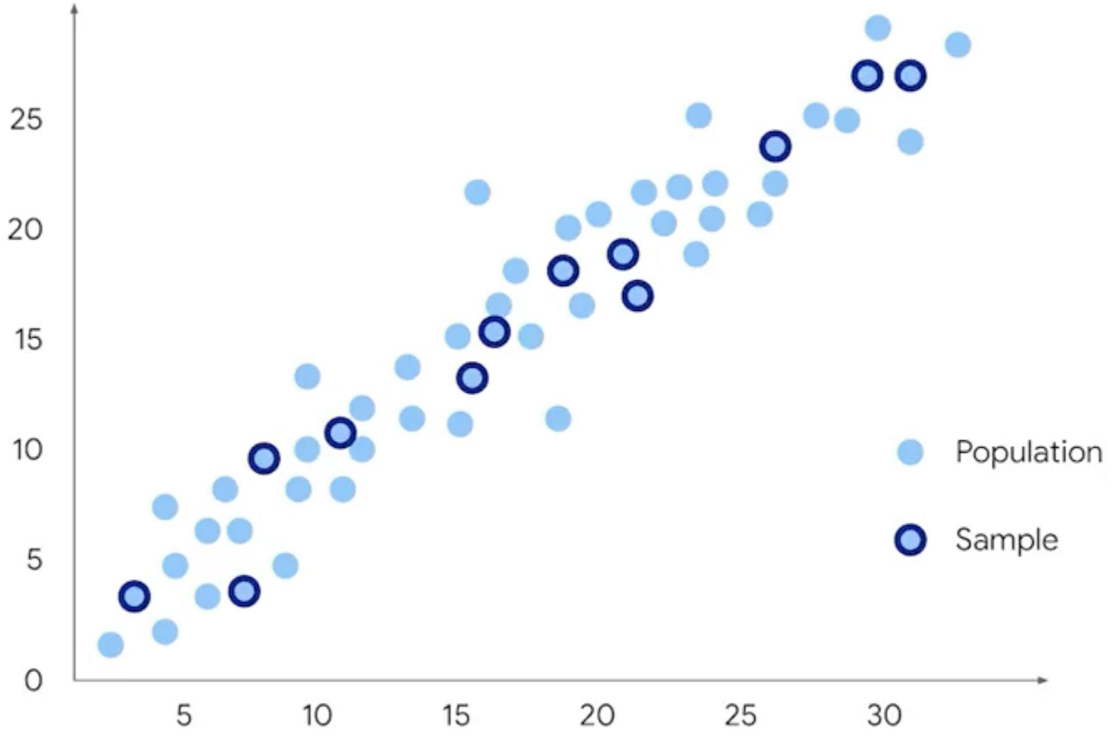
- Linear regression is a way to model linear relationships
- Dependent variables vary according to independent variables
- The slope identifies how much the dependent variable changes per one-unit change in the independent variable
- Correlation describes linear relationships between variables
- **Correlation is not causation**

Example 2. Sample vs Population

A) Population



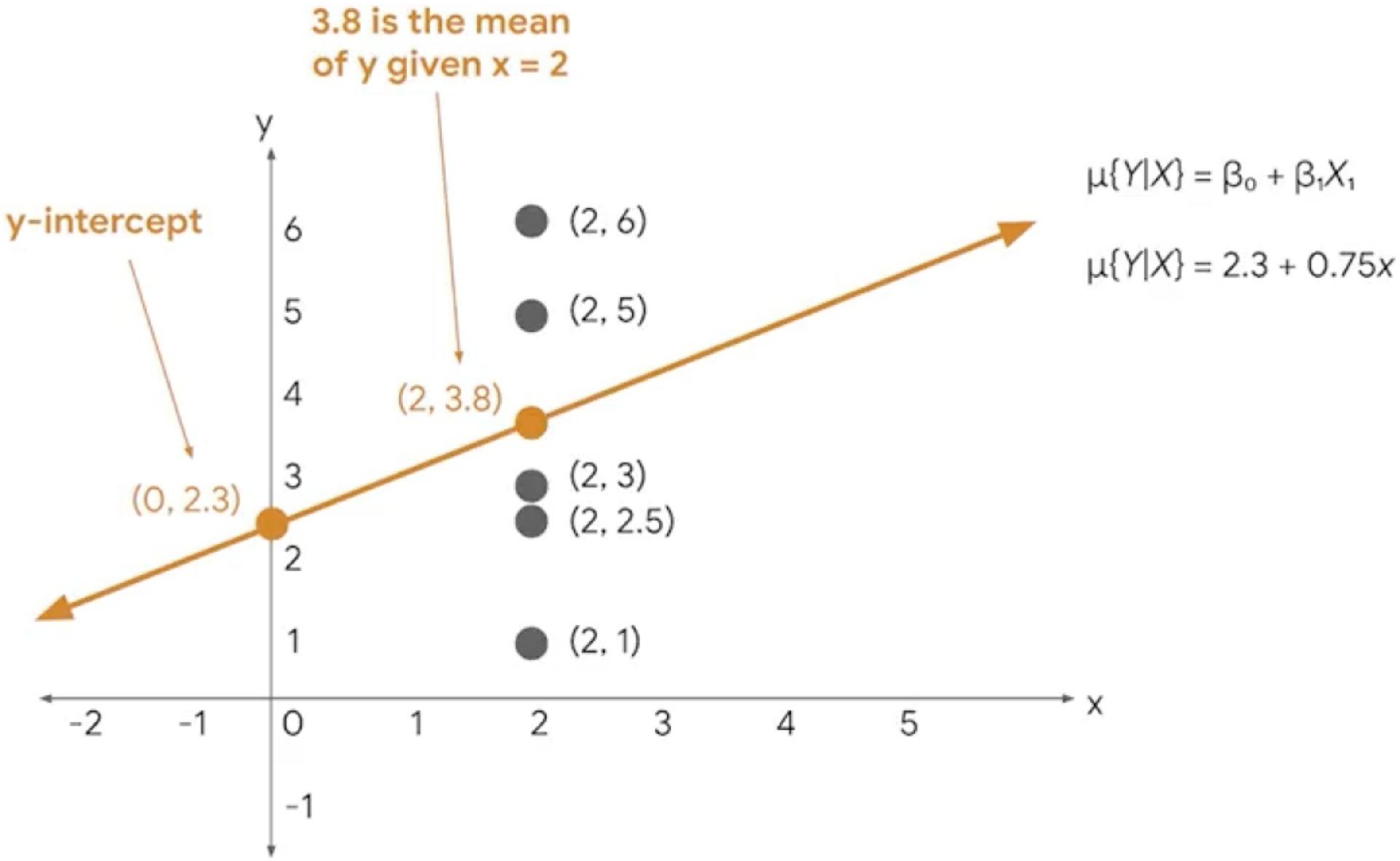
B) Sample vs Population



Data: Sample and Population

- **Sample** is a selection (subset) of data from a larger group of data (**Population**)
- **Observed values (Actual values)** are the existing sample of data
- **Each data point** in the sample is represented by an observed value of the dependent variable and an observed of independent variable

Example 3. Regression Analysis



Linear Regression Equation

$$\mu(Y|X) = \beta_0 + \beta_1 \bullet X$$

Slope is the amount that y increases or decreases per one-unit increase of X.

Intercept is the value of y, the dependent variable when X, the independent variable equals 0.

Betas ($\boldsymbol{\beta_i}$) are parameters.

Linear Regression Estimation

$$\hat{\mu}(Y|X) = \hat{\beta_0} + \hat{\beta_1} \bullet X$$

$$y = \hat{\beta_0} + \hat{\beta_1} \bullet X$$

Regression coefficients are the estimated betas in a regression model, represented as $\hat{\beta_i}$.

Example 4. Linear Regression Estimation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

$$= -1 + 5X$$

X	$\hat{y}$
0	-1
1	4
2	9
3	14

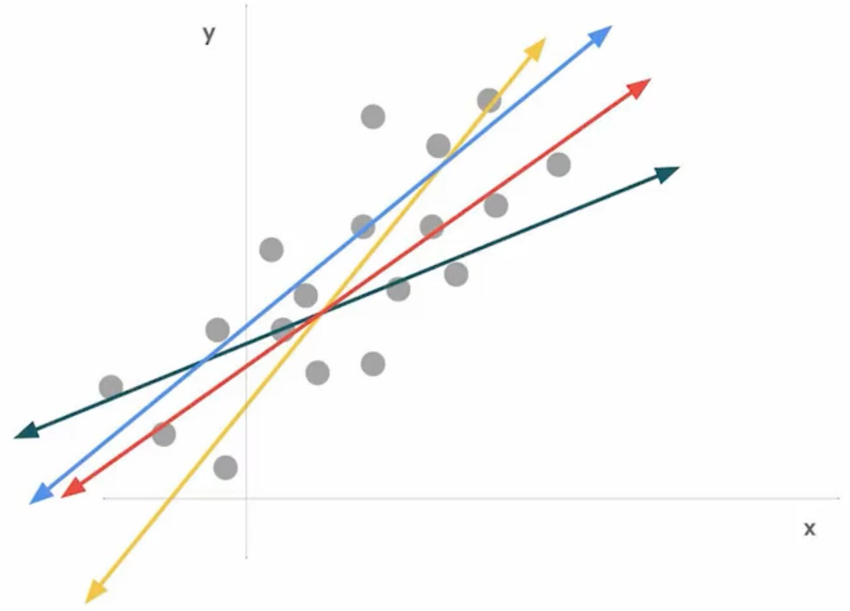
For every one-unit increase in X, we get a 5-unit increase in Y

Ordinary Least Squares (OLS)

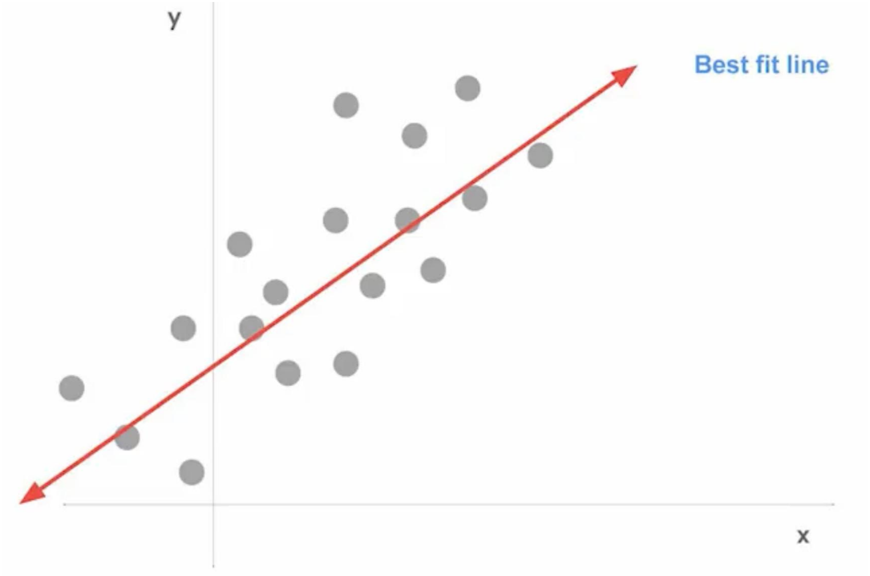
- **OLS** is a method that minimizes the sum of squared residuals to estimate parameters in a linear regression model
- **Loss function** is a function that measures the distance between the observed values and the model's estimated values

Simple Linear Regression

A)



B)



- **Best fit line** is the line that fits the data best by minimizing some loss function or error
- **Predicted values** are the etimated Y for each X calculated by a model

Residuals

Residual is the difference between observed or actual values and the predicted values of the regression line

$$\varepsilon_i = y_i - \hat{y_i}$$

The sum of the residuals is always equal to zero for OLS estimators

Sum of squared residuals (SSP)

Sum of squared residuals is the sum of squared differences between each observed value and its associated predicted value

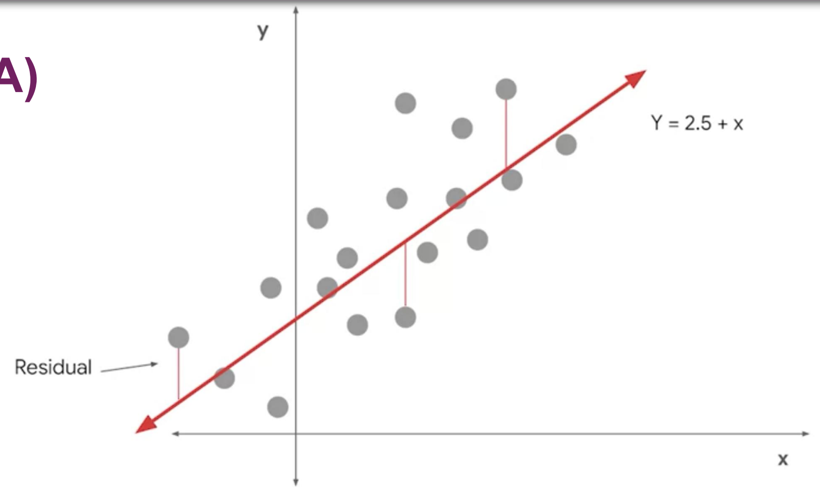
$$SSR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

where:

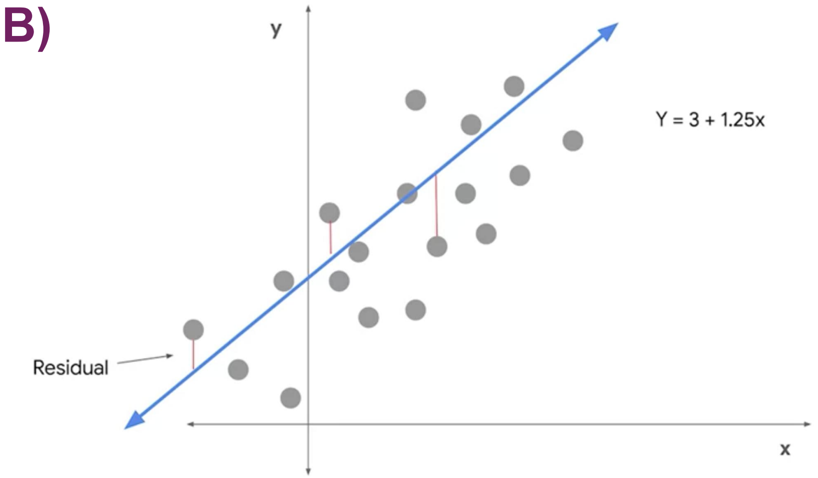
y_i *is an i - observed value; and* $\hat{y}_i$ *is an i - predicted value.*

Example 5. Simple Linear Regression

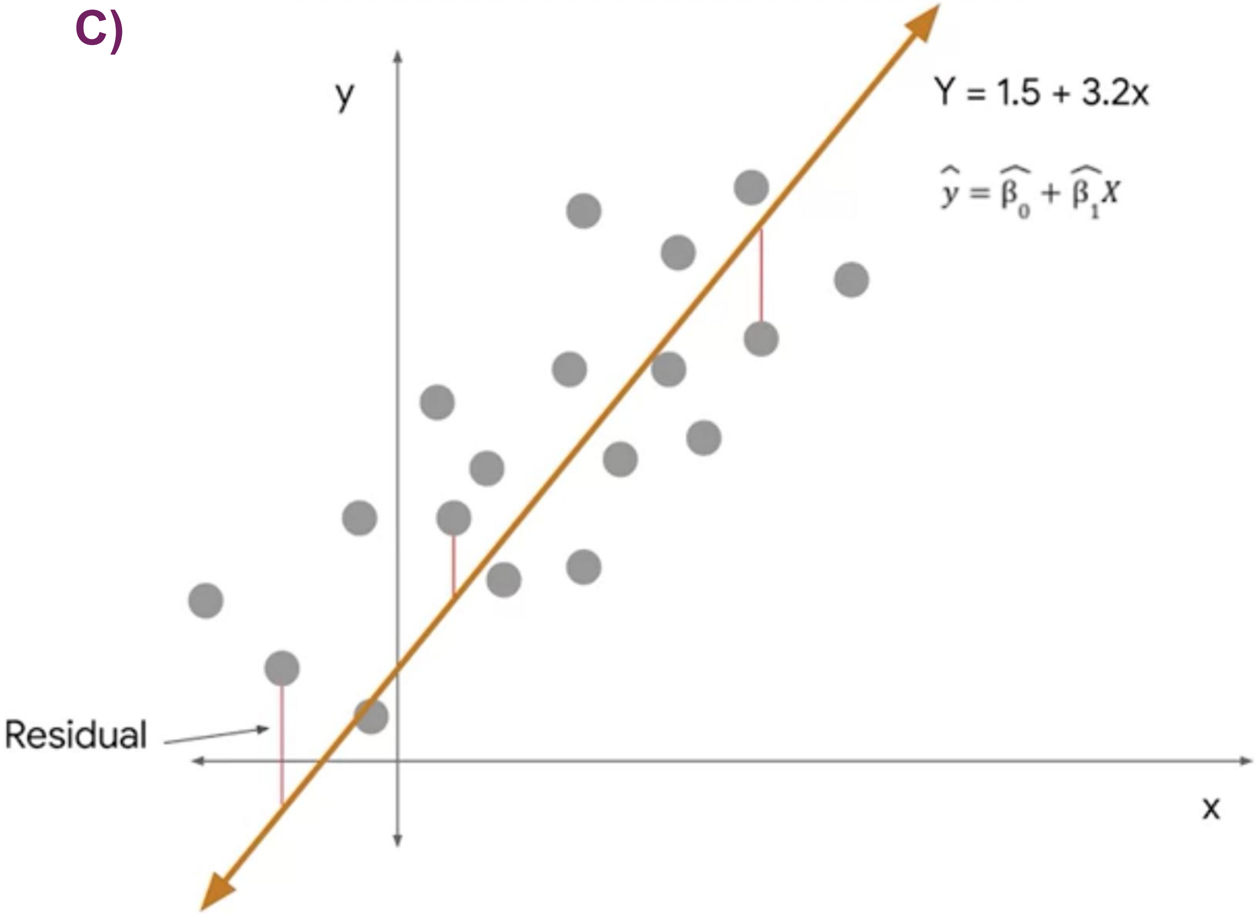
A)



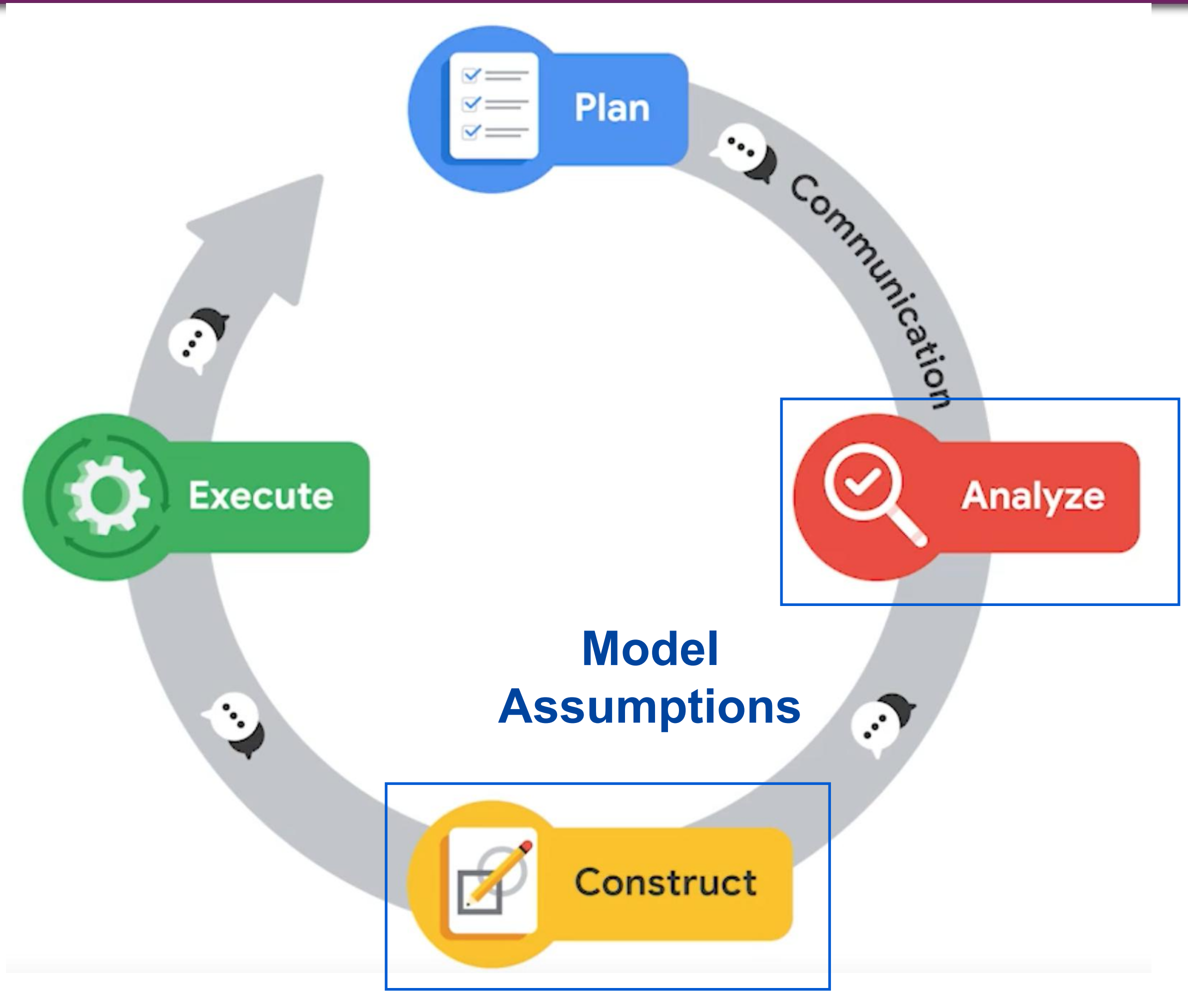
B)



C)



Model Assumptions



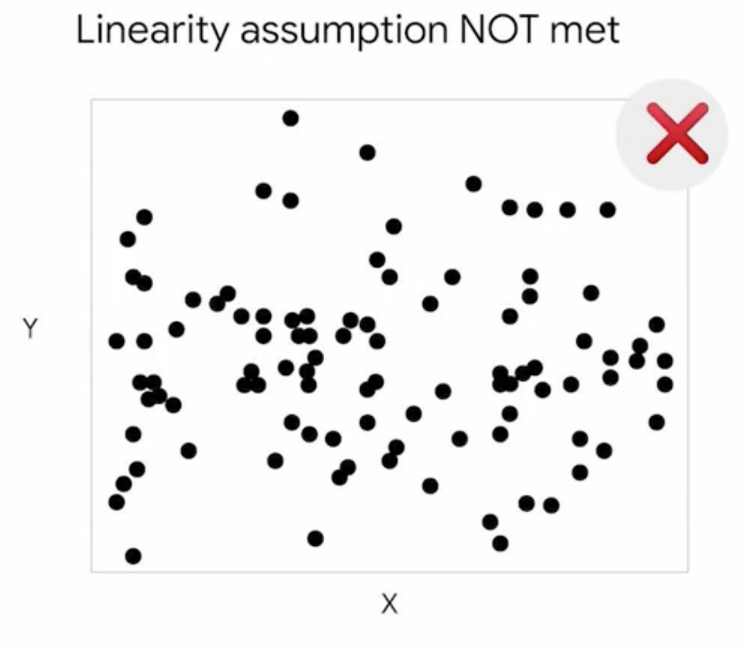
Model Assumptions are statements about the data that must be true in order to justify the use of a particular modelling technique

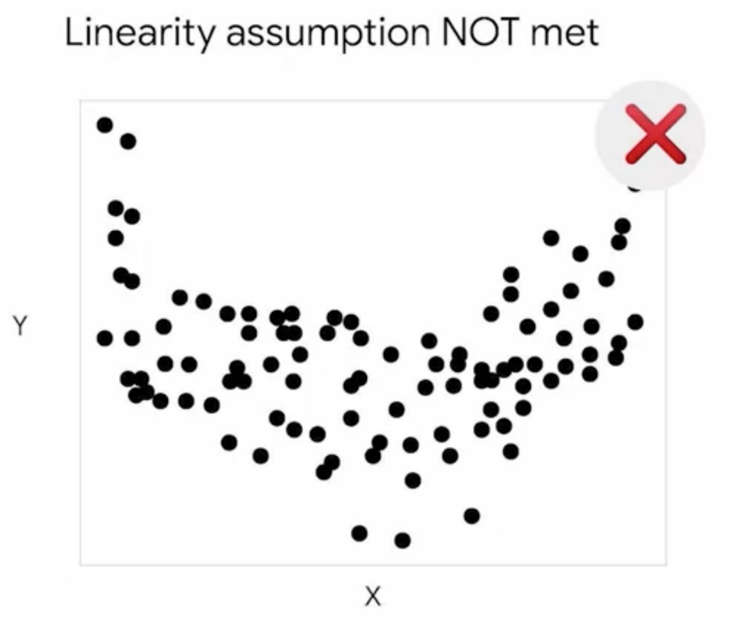
Simple Linear Regression: Assumptions

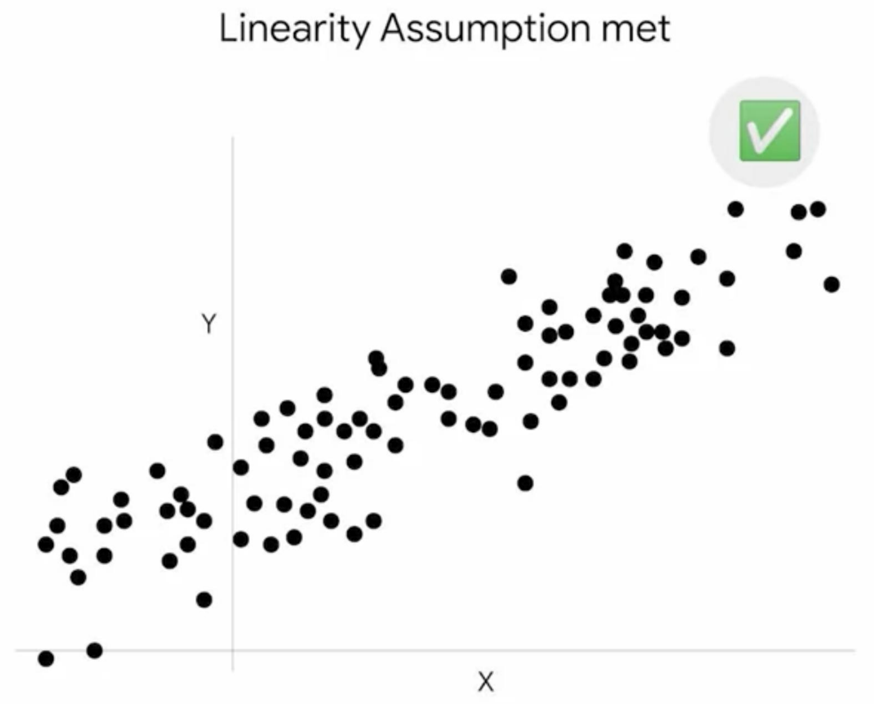
- **Linearity**: Each predictor variable (Xi) is linearly related to the outcome variable (Y)
- **Normality**: The errors are normally distributed.*
- **Independent Observations**: Each observation in the dataset is independent.
- **Homoscedasticity**: The variance of the errors is constant or similar across the model.*

Linearity Assumption

Linearity Assumption: Each predictor variable (Xi) is linearly related to the outcome variable (Y).





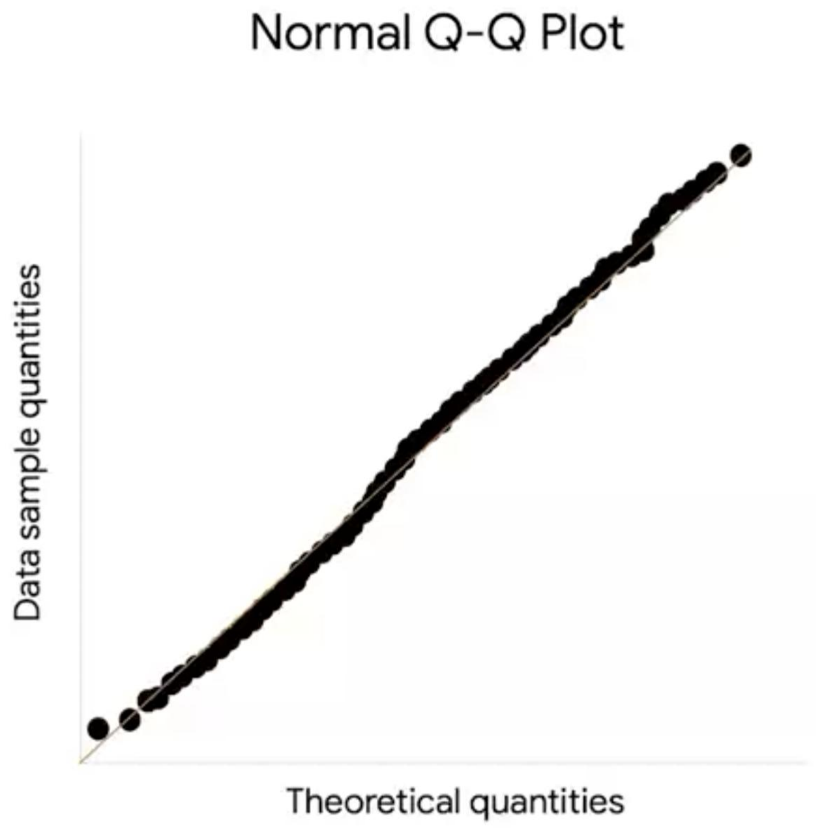


Normality Assumption

Normality Assumption: The residuals or errors are normally distributed.

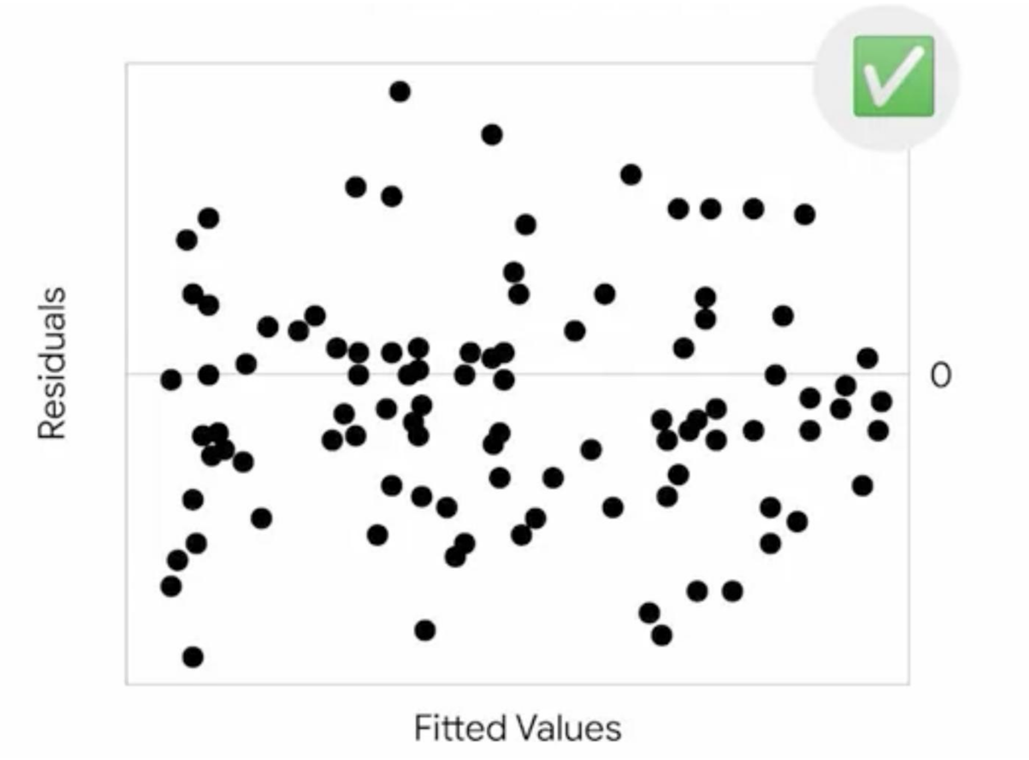
Note:

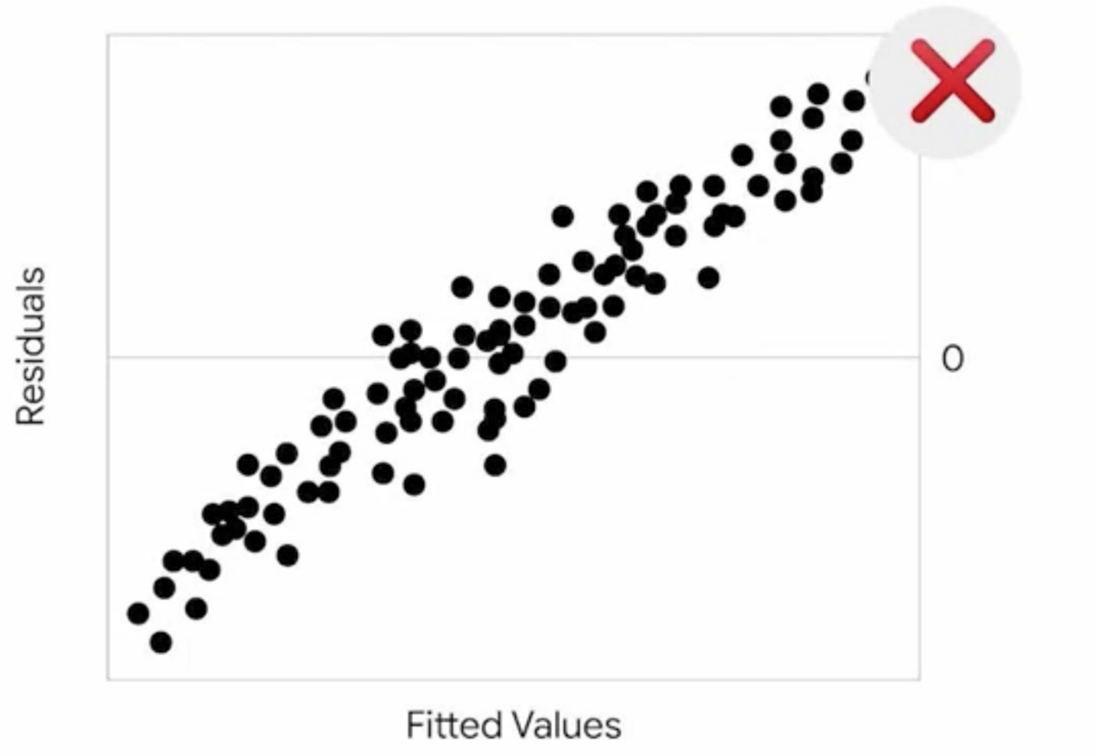
- You can not check the assumption until after the model is buillt;
- Use a specific plot called a quantile-quantile or QQ plot of the residuals.



Independent Observation Assumption

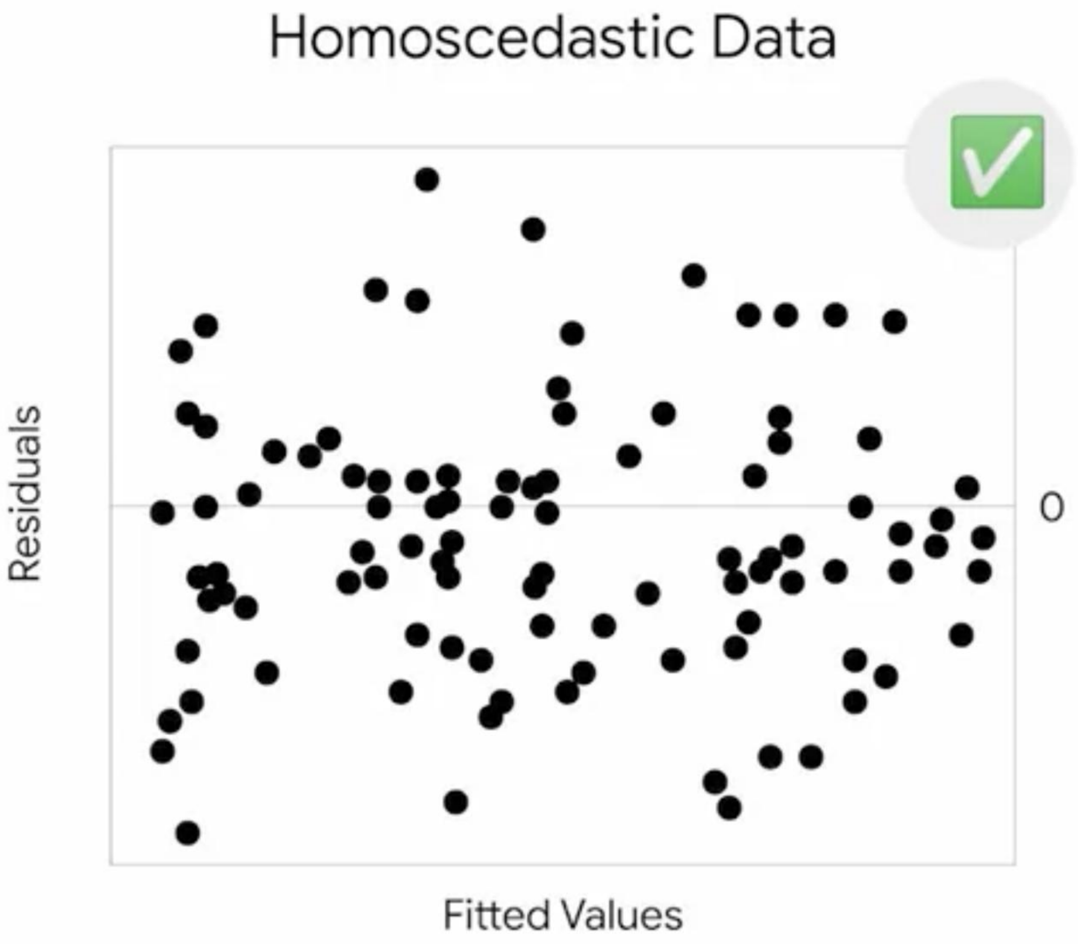
Independent Observation Assumption: Each observation in the dataset is independent





Homoscedasticity Assumption

Homoscedasticity Assumption: The variation of the residuals (errors) is constant or similar across the model

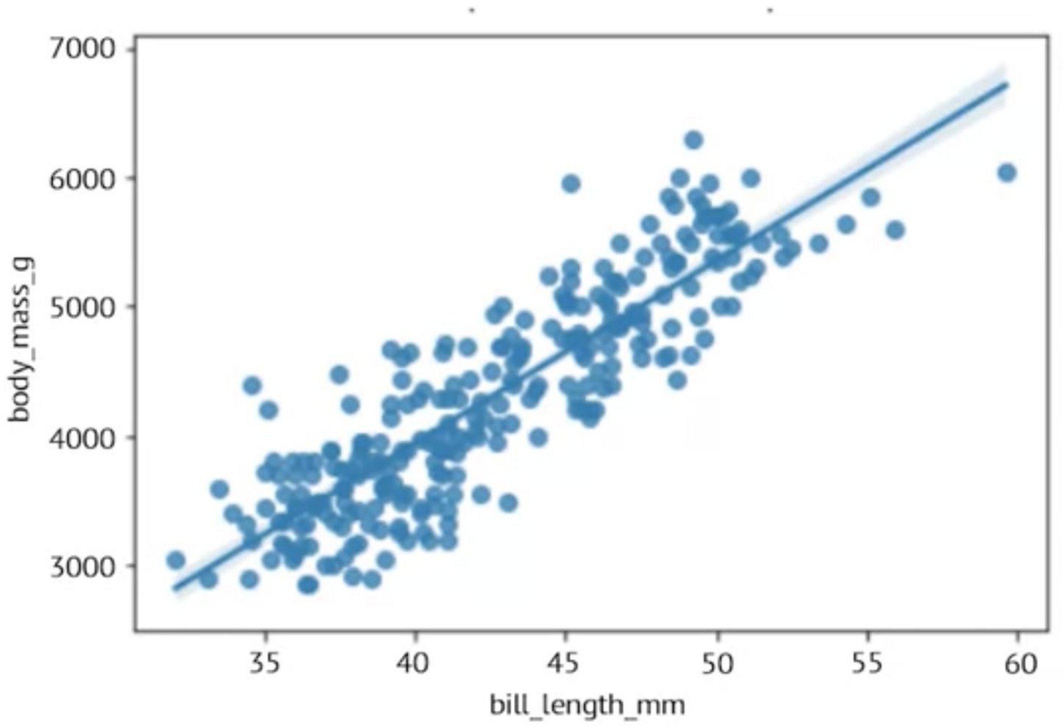


Confidence Interval and Confidence Band

Confidence interval is a range of values that describes the uncertainty surrounding an estimate

Confidence band is an area surrounding the line that describes the uncertainty around the predicted outcome at every value of X

```
sns.regplot(x = "bill_length_mm", y = "body_mass_g", data = ols_data)
```



Linear Regression: Evaluation Metrics

- **Coefficient of determination (R^2)** measures the proportion of variation in the independent variable, Y, explained by the independent variable(s), X:

$$R^2 \in [0;\ 1]$$

- Mean Squared Error (MSE)
- Mean Absolute Error (MAE)

Hold-out Sample

Hold-out sample is a random sample of observed data that is not used to fit the model

Multiple Linear Regression

Multiple linear regression or multiple regression is a technique that estimates the relationship between one continious dependent variable and two or more independent variables

$$y = \beta_0 + \beta_1 \bullet X_1 + \beta_2 \bullet X_2 + \ldots + \beta_n \bullet X_n$$

or

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i \bullet X_i$$

where: y is an observed value; Xi is an i-independent variable.

One-hot encording and Interaction term

- **One-hot encording** is a data transformation technique that turns one categorical variable into several binary variables

- **Iteraction term** is a term that represents how the relationship between two independent variables is associated with changes in the mean of the dependent variable

Example 6. Website Clicks and Advertisements

Categorical Variable 1	Categorical Variable 2	Categorical Variable 3
Ad Color	**Call to Action**	**Streaming Service**
Black-and-white Color	Call to action No call to action	Service A Service B Service C

Example 6. Website Clicks and Advertisements (2)

$$X_{Action} = \begin{cases} 1, \; if \;\; A \; has \; a \; call \; to \; action \\ 0, \; if \;\; A \; doesn't \; have \; a \; call \; to \; action \end{cases}$$

$$y = \beta_0 + \beta_1 \bullet X_1 + \beta_2 \bullet X_2 + \beta_{Action} \bullet X_{Action}$$

Example 6. Website Clicks and Advertisements (3)

$X_{\text{service A}}$	Service A	Service B	Service C
$\boxed{1}$	Ad plays on service A	Ad does NOT play on service B	Ad does NOT play on service C
0	Ad does NOT play on service A	Ad plays on EITHER service B OR C	

Example 6. Website Clicks and Advertisements (4)

# of categories	# of binary variables
2	1
3	2

Example 6. Website Clicks and Advertisements (5)

$X_{\text{service A}}$	$X_{\text{service B}}$	Service A	Service B	Service C
1	0	Plays on service A	Does not play on service B	Does not play on service C
0	1	Does not play on service A	Plays on service B	Does not play on service C
0	0	Does not play on service A	Does not play on service B	Plays on service C

Example 6. Website Clicks and Advertisements (6)

$$y = \beta_0 + \beta_1 \bullet X_1 + \beta_2 \bullet X_2 + \overbrace{\beta_{Service\ A} \bullet X_{Service\ A}}^{\beta_3 \bullet X_3} + \\ + \underbrace{\beta_{Service\ B} \bullet X_{Service\ B}}_{\beta_4 \bullet X_4}$$

where: X_1 is a number of people in the advertisement;

X_2 is the length of the advertisement

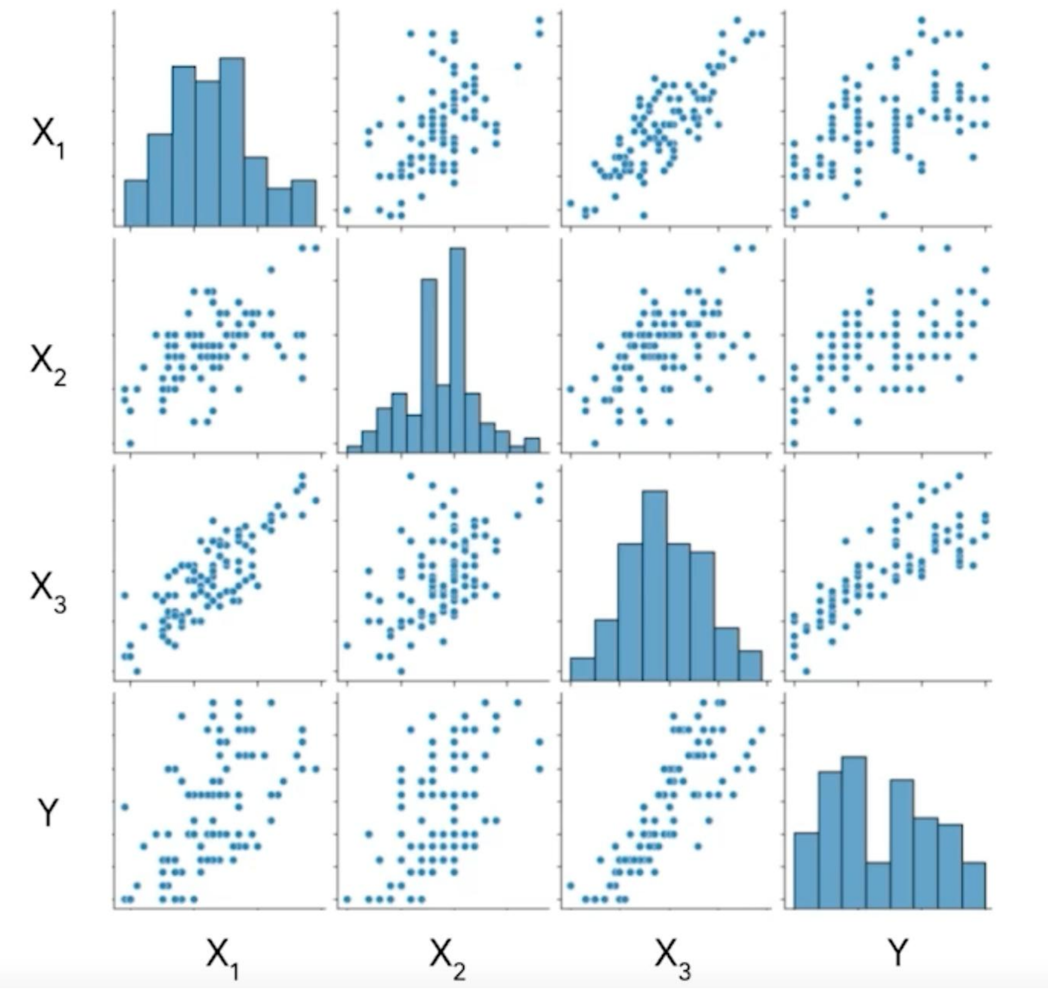
Multiple Regression Assumptions

- **Linearity**: Each predictor variable (Xi) is linearly related to the outcome variable (Y)
- **Normality**: The errors are normally distributed.
- **Independent Observations**: Each observation in the dataset is independent.
- **Homoscedasticity**: The variance of the errors is constant or similar across the model.
- **No multicollinearity**: No two independent variables (X_i and X_j) can be highly correlated with each other

No Multicollinearity Assumption

No multicollinearity: No two independent variables (X_i and X_j) can be highly correlated with each other.

So, X_i and X_j can not be linear related to each other.



Variance Inflation Factors (VIF)

Variance Inflation Factor (VIF) quantifies how correlated each independent variable is with all of the other independent variables

$$VIF \in [1; +\infty)$$

Example 7. Multiple Regression

1) $Sales = -38 + 4 \bullet Temperature$

2) $Sales = \beta_0 + \beta_{Temperature} \bullet X_{Temperature} + \beta_{Ad} \bullet X_{Ad}$

3) $Sales = \beta_0 + \beta_{Temperature} \bullet 15 + \beta_{Ad} \bullet 1$

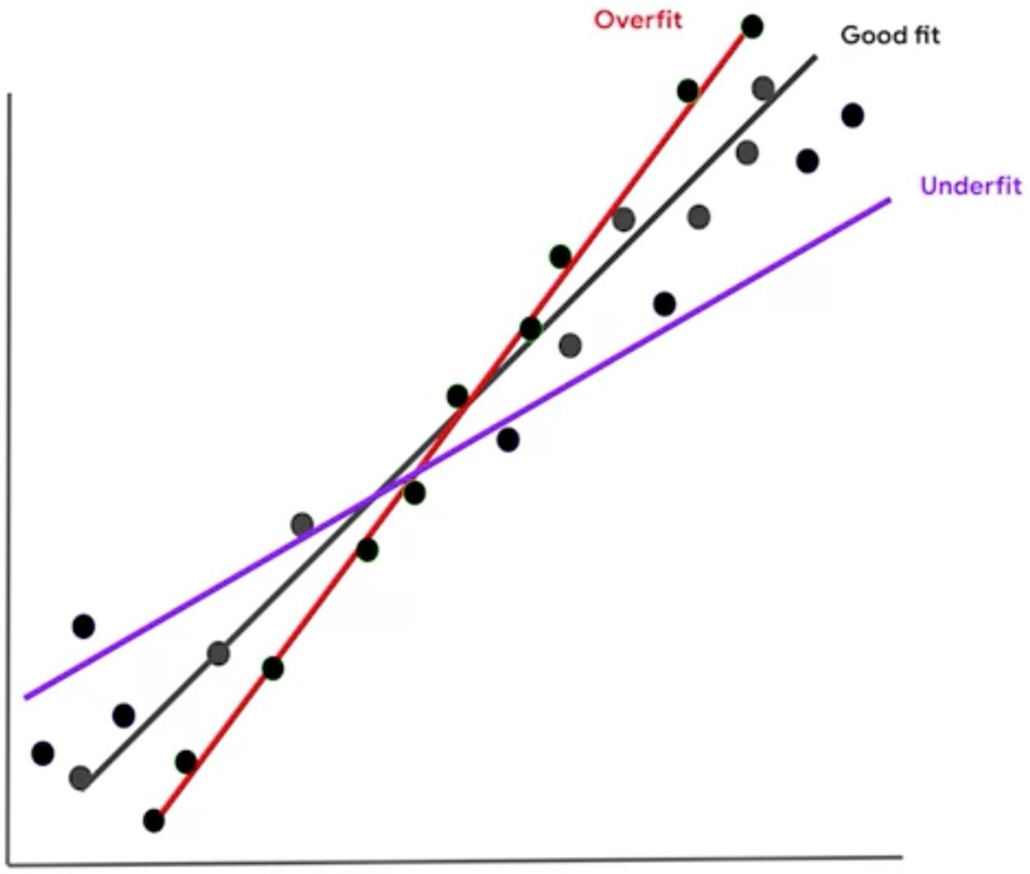
4) $Sales = \beta_0 + \beta_{Temperature} \bullet 15 + \beta_{Ad} \bullet 0 = \beta_0 + \beta_{Temperature} \bullet 15$

Example 7. Multiple Regression (2)

5) $$Sales = \beta_0 + \beta_{Temperature} \bullet X_{Temperature} + \\ + \beta_{Transportation} \bullet X_{Transportation}$$

6) $$Sales = \beta_0 + \beta_{Temperature} \bullet X_{Temperature} + \\ + \beta_{Transportation} \bullet X_{Transportation} + \\ + \beta_{Interaction} \bullet (X_{Temperature} \bullet X_{Transportation})$$

Overfitting



Overfitting

When a model fits the observed or training data too specifically, and is unable to generate suitable estimates for the general population

Adjusted R^2

Adjusted R^2 is a variation of the R^2 regression evaluation metric that penalizes unnecessary explanatory variables

$$Adj.\, R^2 \in \; [0;\; 1]$$

Adjusted R^2 vs R^2

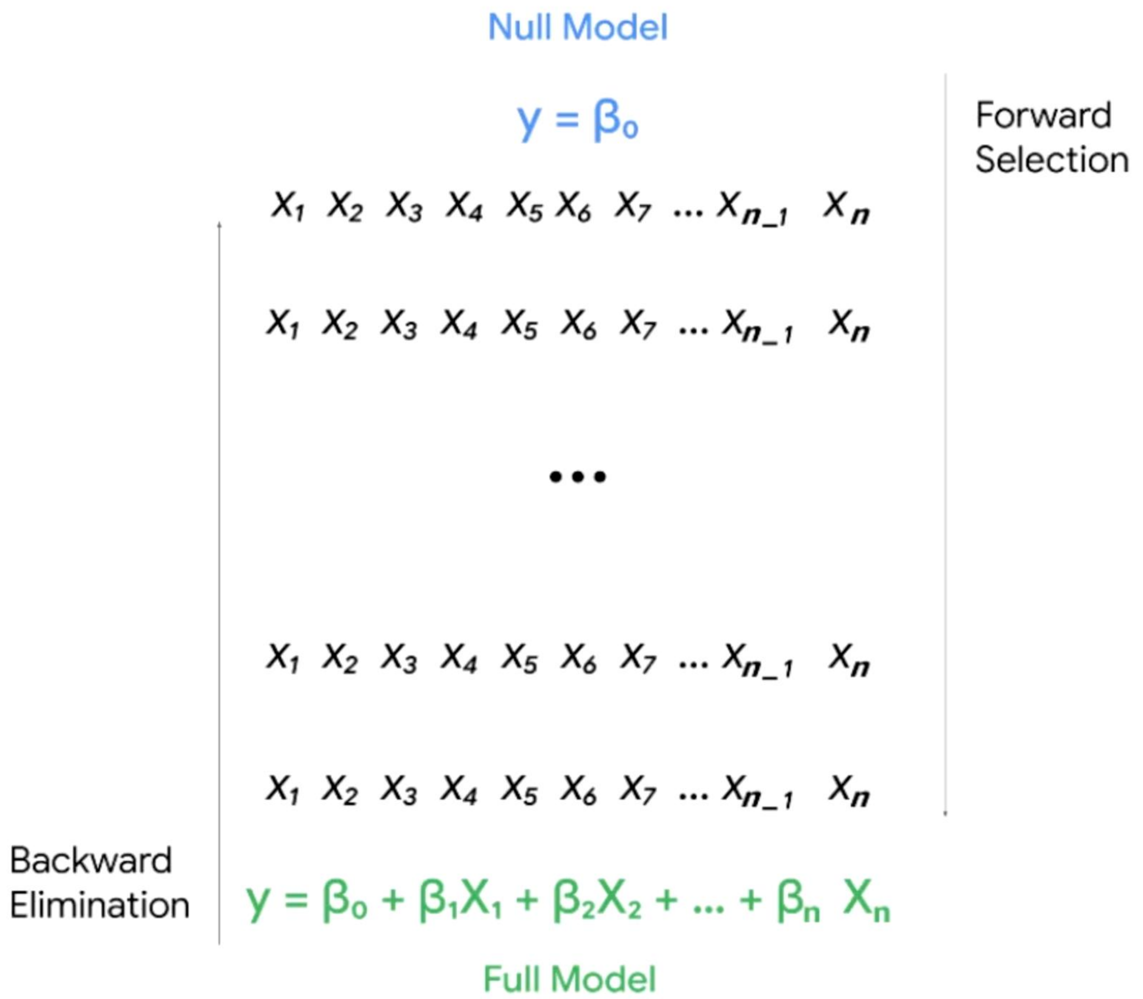
Adjusted R^2 is used to compare models of varying complexity:

- determine is you should add another variable or not

R^2 is more easily interpretable:

- determine how much variation in the dependent variable is explained by the model

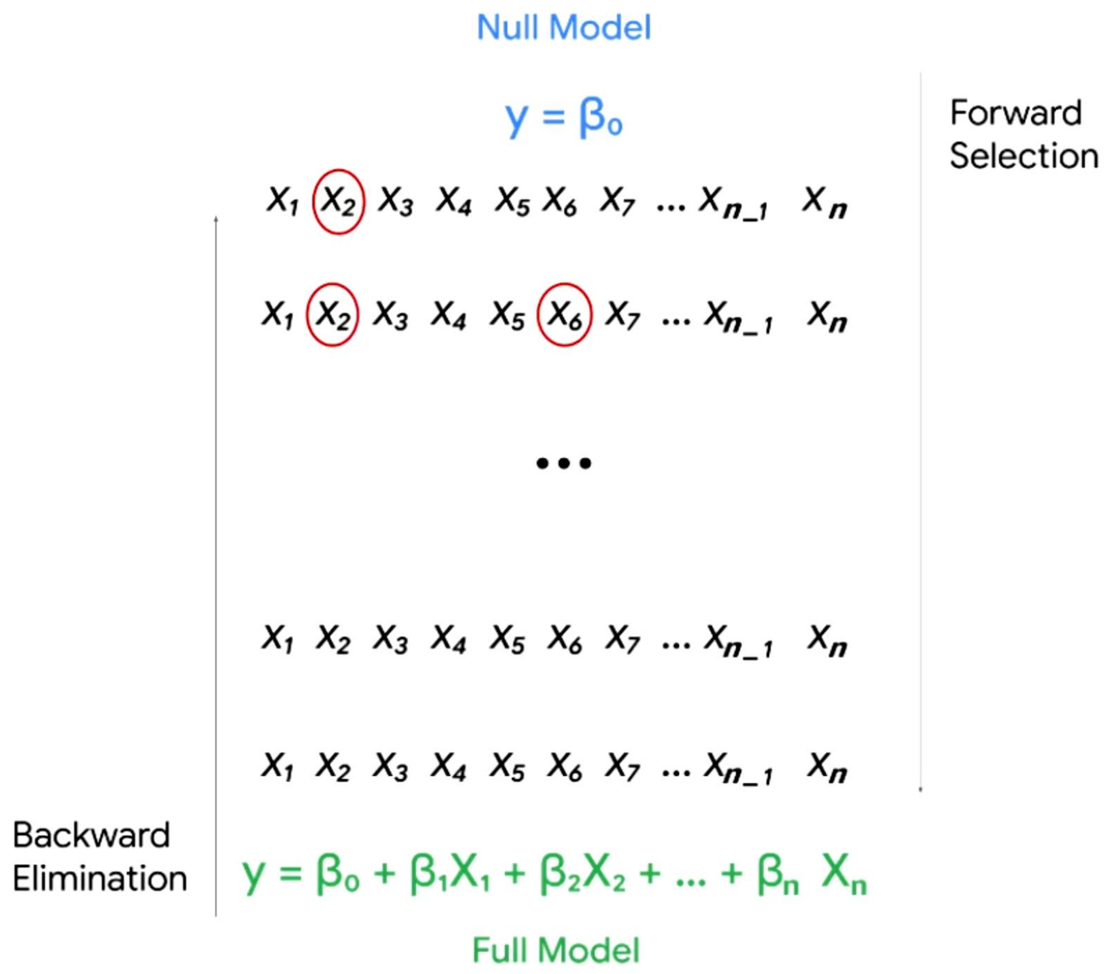
Forward Selection and Backward Elimination



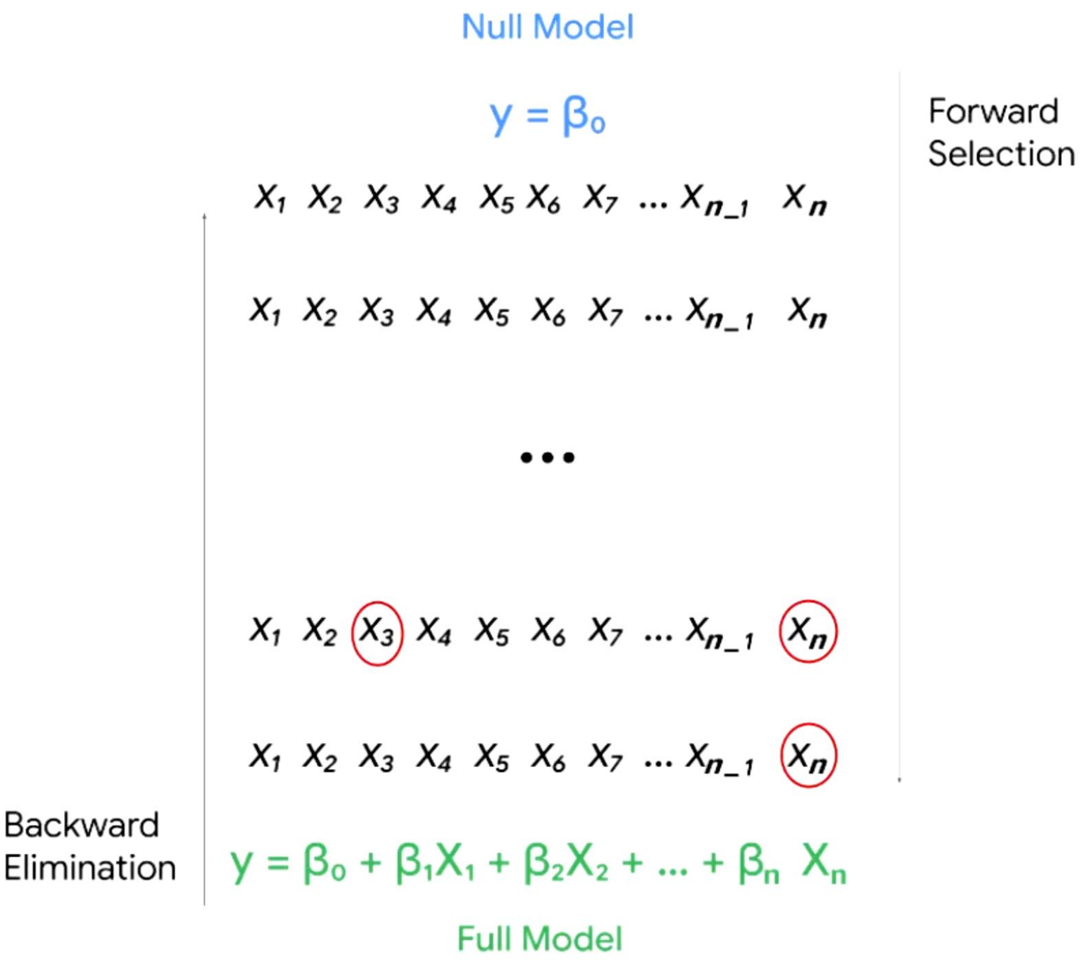
Variable selection or feature selection is the process of determining which variables or features to include in a given model

Forward Selection

Forward selection is a stepwise variable selection process that begins with the null model, with 0 independent variables, considers all posible variables to add. It incorporates the independent variable that contributes the most explanatory power to the model.



Backward Elimination



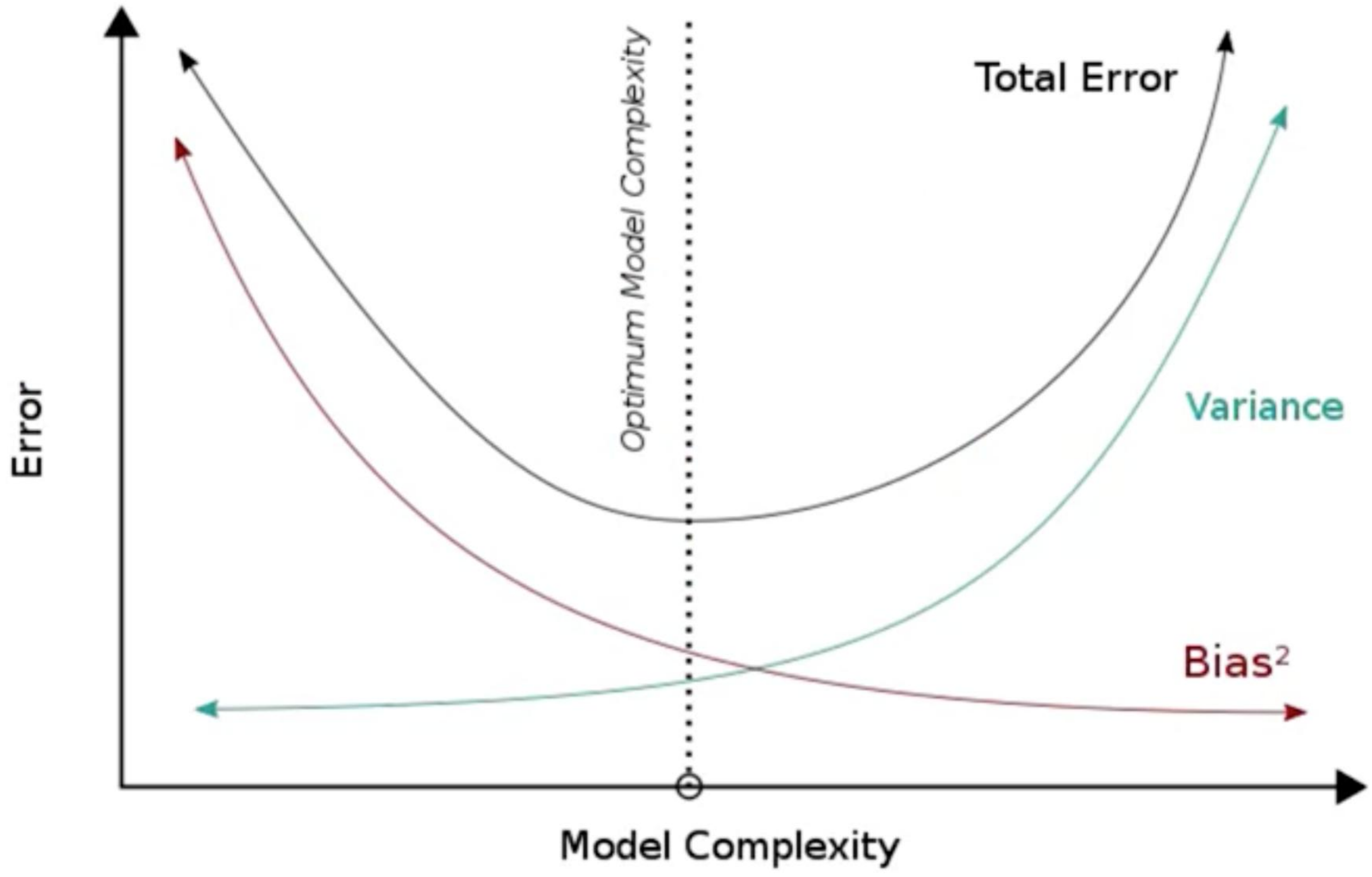
Backward elemination is a stepwise variable selection process that begins with the full model, with all possible independent variables, and removes the independent variable that adds the least explanatory power to the model

Extra-sum-of-squares F-test

Extra-sum-of-squares F-test quantifies the difference between the amount of variance that is left unexplained by a reduced model that is explained by the full model

Bias-Variance Tradeoff

Bias-Variance Tradeoff is a balance between two model qualities, bias and variance, to minimize overall error for unobserved data



Bias

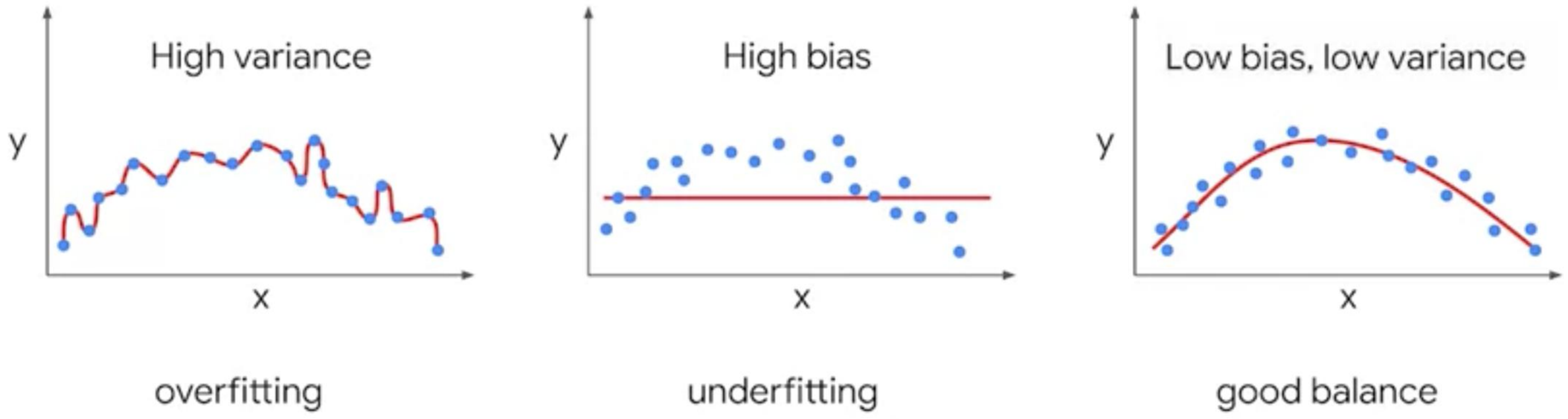
Bias simplifies the model predictions by making assumptions about the variable relationships.

A **highly biased model** may:

- oversimplify the relationship
- underfitting to the observed data
- generating inacurate estimates

Variance

Variance allows for a model flexibility and complexity, so the model learn from existing data. A model with high variance can overfit the observed data and generate inaccurate estimates for unseen data.



Regularization

Regulatization is a set of regression techniques that shrinks regression coefficient estimates toward zero, adding in bias, to reduce variance

Regularized regression:

- Lasso regression
- Ridge regression
- Elastic-net regression

Chi-squared (χ^2)

Chi-squared (χ^2) tests will help you determine if two categorical variables are associated with one another, and whether categorical variable follows an expected distribution

Coding Activity 2. Supervised ML. Regression

Lab 2. Supervised Machine Learning. Linear Regression ||
Regression Model for a Financial Dataset.
Stock Price Prediction with Python

Steps to follow:

1. Upload the following files from the module learning room:
   – Jupiter notebook "Lab2_Stock_Price_Prediction_with_Python.ipynb"
   – Csv-dataset file "data-appl_regression.csv"
2. Follow along in the Jupiter notebook

Q&A

Thank you!