# Regression

1.What is Simple Linear Regression ?

Simple Linear Regression is a statistical method used to model the relationship between a **dependent variable (Y)** and a **single independent variable (X)** using a straight line.

2.What are the key assumptions of Simple Linear Regression?

Linearity

Independence

Homoscedasticity

Normality of Errors

No Perfect Multicollinearity

3. What does the coefficient m represent in the equation Y=mX+c

It indicates the rate of change of Y with respect to X.

**Mathematically:** It tells how much **Y** changes for a one-unit increase in **X**.

**If m is positive** → Y increases as X increases (upward slope).

**If m is negative** → Y decreases as X increases (downward slope).

**If m is zero** → Y remains constant (horizontal line).

4. What does the intercept c represent in the equation Y=mX+c

**c** = Intercept (the point where the line crosses the **Y-axis**)

6. What is the purpose of the least squares method in Simple Linear Regression.

The **least squares method** is used in **Simple Linear Regression** to find the best-fitting line by minimizing the sum of the squared differences between the **actual values** and the **predicted values**.

8.What is Multiple Linear Regression ?

**Multiple Linear Regression (MLR)** is an extension of **Simple Linear Regression**, where we model the relationship between a **dependent variable (Y)** and **two or more independent variables (X₁, X₂, X₃, ...)**.

9. What is the main difference between Simple and Multiple Linear Regression?

|  |  |  |
| --- | --- | --- |
| features | Simple linear regression | Multiple linear regression |
| Number of dependent variables | one | Two or more |
| complexity | Simple and easy to interrupt | More complex |
| Use case | Used when one predictor influences the dependent variable | Used when multiple factors affect the dependent variable |

10. What are the key assumptions of Multiple Linear Regression?

Linearity

Independence

Homoscedasticity

Normality of Errors

No Perfect Multicollinearity

11. What is heteroscedasticity, and how does it affect the results of a Multiple Linear Regression model

Heteroscedasticity refers to a situation in which the **variance of the residuals (errors) is not constant** across all levels of the independent variables in a regression model.

how does it affect the results of a Multiple Linear Regression model

Biased Standard Errors

Incorrect Confidence Intervals and p-values

Inefficient Estimates

12. How can you improve a Multiple Linear Regression model with high multicollinearity?

**Check the Variance Inflation Factor (VIF):**

**Remove Highly Correlated Predictors**

**Combine Highly Correlated Variables**

**Use Regularization Techniques**

**Increase Sample Size**

**Transform Variables**

**Domain Knowledge Selection**

13. What are some common techniques for transforming categorical variables for use in regression models ?

One-Hot Encoding

Label encoding

Ordinal encoding

Frequency Encoding

14. What is the role of interaction terms in Multiple Linear Regression?

**Capture Conditional Relationships**

**Improve Model Accuracy**

**Avoid Misinterpretation**

**Model Real-World Complexity**

**Enhance Predictive Power**

15. How can the interpretation of intercept differ between Simple and Multiple Linear Regression?

|  |  |  |
| --- | --- | --- |
| Feature | Simple linear regression | Multi linear regression |
| Meaning | Baseline value of YYY when X=0X = 0X=0 | Baseline value when all Xi=0X\_i = 0Xi​=0 |
| interpretability | Often meaningful | May not be meaningful if zero is unrealistic |
| Complexity | Easier to understand | Affected by multiple predictors |

16. What is the significance of the slope in regression analysis, and how does it affect predictions?

**Measures the Effect of XXX on YYY**

**Indicates Direction**

**Determines Strength of Relationship**

**Essential for Predictions**

**Aids Decision-Making**

17. How does the intercept in a regression model provide context for the relationship between variables

**Baseline Value** – Represents the predicted value of YYY when all independent variables are zero.

**Contextual Reference** – Provides a starting point to understand how independent variables influence the dependent variable.

**Interpretability** – Meaningful if zero is a realistic value for predictors; otherwise, may be just a mathematical artifact.

**Adjustment Factor** – Ensures the regression line fits the data correctly by shifting it up or down.

**Comparison Tool** – Helps compare different models by analyzing how predictions start before applying variable effects.

18. What are the limitations of using R2 as a sole measure of model performance

**Does Not Indicate Causation**

**Ignores Model**

**Sensitive to Outliers**

**Cannot Detect Overfitting**

19. How would you interpret a large standard error for a regression coefficient

**High Variability in Estimates**

**Weak Relationship with YYY**

**Possible Multicollinearity**

**Small Sample Size Impact**

**Low Statistical Significance**

20. How can heteroscedasticity be identified in residual plots, and why is it important to address it?

Residuals form a fan-shaped or cone-shaped pattern.

The spread of residuals increases or decreases with fitted values.

Residuals show non-random patterns instead of a constant spread.

**Why It Is Important to Address**

Causes incorrect p-values and misleading statistical tests.

Reduces accuracy of regression estimates and predictions.

Violates assumptions of regression, making results unreliable.

21. What does it mean if a Multiple Linear Regression model has a high R² but low adjusted R2

T**oo Many Irrelevant Predictors**

**Overfitting**

**Multicollinearity**

**Small Sample Size**

22. Why is it important to scale variables in Multiple Linear Regression

**Improves Numerical Stability** – Prevents large differences in magnitudes from causing computational issues.

**Enhances Model Interpretability** – Helps compare the effect of different predictors on the outcome.

**Speeds Up Convergence** – Essential for optimization algorithms like Gradient Descent in large datasets.

**Reduces Multicollinearity** – Scaling can help mitigate correlation issues in polynomial and interaction terms.

23. What is polynomial regression

Polynomial regression is an extension of **linear regression** where the relationship between the independent variable (XXX) and the dependent variable (YYY) is modeled as an **n-degree polynomial** instead of a straight line.

24. How does polynomial regression differ from linear regression

|  |  |  |
| --- | --- | --- |
| feature | Linear regression | Polynomial regression |
| complexity | Simpler, easier to interpret. | More complex, can capture non-linearity. |
| Flexibility | Limited to linear trends. | Can model **quadratic, cubic, or higher-order** trends. |

24. When is polynomial regression used

**Non-Linear Relationships**

**Trend Analysis**

**Physics & Engineering**

**Stock Market & Finance**

26. Can polynomial regression be applied to multiple variable

Yes, **Polynomial Regression** can be extended to **Multiple Variables**, and it is called **Multivariable Polynomial Regression**.

27. What are the limitations of polynomial regression

**Overfitting** – Higher-degree polynomials can fit noise rather than actual trends.

**Computational Complexity** – As the degree increases, calculations become more intensive.

**Interpretability Issues** – Coefficients become harder to explain in real-world scenarios.

**Extrapolation Problems** – Predictions outside the data range can be highly inaccurate.

28. What methods can be used to evaluate model fit when selecting the degree of a polynomial

**Mean Squared Error (MSE) / Root Mean Squared Error (RMSE)**

**Cross-Validation (e.g., k-Fold CV)**

**Residual Plots**

**Akaike Information Criterion (AIC) & Bayesian Information Criterion (BIC)**

**Learning Curves**

29. Why is visualization important in polynomial regression

**Detecting Non-Linearity** – Helps identify if a polynomial model is needed instead of a simple linear regression.

**Choosing the Right Degree** – Visualizing the curve helps prevent **underfitting (too simple)** or **overfitting (too complex)**.

**Checking Residual Patterns** – Residual plots reveal if errors are randomly distributed, ensuring a good model fit.

**Comparing Different Models** – Helps compare polynomial degrees to see which one best captures the data trend.

31. How is polynomial regression implemented in Python

Polynomial regression can be implemented using **Scikit-Learn**.

**PolynomialFeatures(degree=n)** transforms input features to polynomial terms.

**LinearRegression()** fits the polynomial-transformed data