Regression

1) What is Simple Linear Regression ?

Simple linear regression attempts to determine the strength and characteristics of the relationship between one independent and one dependent variable let say y .

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

(i) The relationship between the dependent and independent variable is linear.

(ii) For all lines we want error = Yactual – Y pred to be least .

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

Here m , in Y = mX + c represents the slope of the line .

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

Here c , in Y =mX + c represents the intercept of the line on y-axis.

5) How do we calculate the slope m in Simple Linear Regression ?

We can calculate the slope of a line in Simple Linear Regression by this, change is y / change in x (y2-y1/x2-x1).

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

In Simple Linear Regression ,when we add the errors , these sometimes cancel outs each other and is not true that error is zero. So, that's why we squared the difference of Y actual & Y pred and taken the least squared values .

7) How is the coefficient of determination (R^2) interpreted in Simple Linear Regression ?

It measures how well the independent variable(x) explains the variability in the dependent in the dependent variable (y).

R**2 = Explained variation by x in y / Total variation in Y

R**2 = 1 (Perfect fit)

R**2 = 0 (No explanatory power)

It always lies between 0 and 1.

8) What is Multiple Linear Regression ?

Multiple linear Regression is nothing ,just the extension of Simple Linear Regression that models the relationship between one dependent(y) and two or more independent variable (x1,x2,x3,…….xn).

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

Main Differences between Simple and Multiple Linear Regression are :

In Simple Linear Regression the number of independent variable is only one and simple and easy to use , whereas in Multiple Linear Regression the number of independent variables are two or more than two and these are more complex than S.L.R.

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

(i) The relationship between independent variables and dependent variable is linear.
(ii) Independent variables should not be highly correlated with each other.

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

Heteroscedasticity occurs when the variance of residuals (errors) is not constant across all levels of the independent variables.

It affects in many ways :

(i) Violates Regression Assumptions
(ii) Affects Standard Errors (t-tests, F- tests becomes invalid).\

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

Multicollinearity occurs when two or more independent variables in Multiple Linear Regression (MLR) model are highly correlated . This makes it difficult to determine the individual effect of each predictor on the dependent variable.

Ways to improve :

(i) Remove highly correlated variables
(ii) Combine correlated variables (avg).
(iii)Use Principal Component Analysis (PCA).

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

(i) One Hot Encoding (O.H.E).

(ii) Label Encoding

(iii)Ordinal Encoding

(iv)Target Encoding

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

Interaction term in MLR helps in capturing the combined effect of two or more independent variables on the dependent variable .They allow the relationship between an independent variable and the dependent variable to change depending on the value of another independent variable.

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

In Simple Linear Regression intercept represents the baseline of Y when there is no influence from X ,whereas in Multiple Linear Regression the intercept is the point where hyperplane intersect the vertical axis , it is often meaningless because 'all predictors = 0' may not be realistic scenario.

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

The slope in regression analysis represents the rate of change in the dependent variable (Y) for a one-unit increase in the independent variable(X) ,assuming all other factors remain constant.

Affects in prediction by the following :

(i) A higher absolute value of the slope means the predictor variable has a stronger impact on Y.

(ii) If the slope is close to zero , the independent variable has little to no effect.

(iii) In MLR, slope help prioritize important predictors for decision making.

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

The intercept in regression models represents the expected value of the dependent variable(Y) when all independent variables(X) are zero.

18) What are the limitations of using R^2 as a sole measure of model performance?

(i) A high R**2 does not imply causation.

(ii) A low R**2 does not mean the model is useless -there may be non-linear relationship.

(iii)Adding two or more variables (in MLR) artificially increase R**2 . So, adjusted R**2 is often used.

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

It measures the variability or uncertainty in the estimated coefficient . A large standard error suggests that the coefficient estimate is unstable , meaning it may vary with different samples.

(i) Address Multicollinearity

(ii) Increase Sampe Size.

(iii)Remove Unnecessary Variables

(iv)Use Feature engineering

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

(i) **Identification:** Heteroscedasticity occurs when residuals exhibit a non-constant variance, often forming a funnel shape in residual plots.

(ii) **Importance:** It violates the assumption of homoscedasticity in regression models, leading to inefficient estimates and unreliable hypothesis tests. Corrective measures include transforming the dependent variable or using robust standard errors.

21) What does it mean if a Multiple Linear Regression model has a high R^2 but low adjusted R^2?

(i) A high R2R^2R2 suggests the model explains a large portion of the variance in the dependent variable.

(ii) A low adjusted R2R^2R2 indicates that adding independent variables is not significantly improving the model and may introduce unnecessary complexity.

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

(i) Prevents some variables from dominating others due to different units or magnitudes.

(ii)Improves numerical stability and convergence in optimization algorithms.

(iii)Helps in better interpretation of coefficients, especially when using regularization techniques like Ridge or Lasso regression.

23) What is polynomial regression ?

Polynomial regression is a type of regression analysis where the relationship between the independent variable(s) and the dependent variable is modeled as an nth-degree polynomial. It is an extension of linear regression that allows for capturing nonlinear relationships.

24) How does polynomial regression differ from linear regression?

In linear regression the line is straight line and less flexible and risk of overfitting is very less , whereas in Polynomial Regression is non-linear (curved) and more flexible and overfitting risk is higher ,especially for high-degree polynomials.

25) When is polynomial regression used?

i. The data exhibits a nonlinear relationship between independent and dependent variables.
ii. A simple linear model does not fit the data well.
iii. The relationship between variables follows a curved pattern rather than a straight line.

26) What is the general equation for polynomial regression?

For a single independent variable xxx, the polynomial regression equation of degree n is:

$y=\beta0+\beta1x+\beta2x^2+...+\beta nx^n+\epsilon y$

where:

- y is the dependent variable.
- x is the independent variable.
- $\beta0,\beta1,...,\beta n$ are the regression coefficients.
- ϵ is the error term.

27) Can polynomial regression be applied to multiple variables ?

Yes, polynomial regression can be extended to multiple independent variables (Multivariate Polynomial Regression). The equation becomes:

$y=\beta0+\beta1x1+\beta2x2+\beta3x12+\beta4x22+\beta5x1x2+...+\epsilon$

where interaction terms (e.g., x1,x2) can also be included.

28) What are the limitations of polynomial regression ?

i. **Overfitting:** Higher-degree polynomials can fit training data too closely and fail to generalize.
ii. **Extrapolation Issues:** Predictions outside the observed data range may be highly inaccurate.
iii. **Multicollinearity:** High-degree polynomials can cause correlated predictors, leading to unstable coefficients.

29) What methods can be used to evaluate model fit when selecting the degree of a polynomial ?

i. **R^2 and Adjusted R^2:** Measure how well the model explains variance, but adjusted R^2 penalizes unnecessary complexity.
ii. **Cross-Validation:** Ensures the model generalizes well to unseen data.
iii. **Residual Plots:** Helps detect underfitting or overfitting.

30) Why is visualization important in polynomial regression ?

i. Helps determine if a polynomial fit is necessary by revealing nonlinear patterns.
ii. Allows assessment of overfitting by comparing training and test set fits.
iii. Residual plots help diagnose model performance.

31) How is polynomial regression implemented in Python?

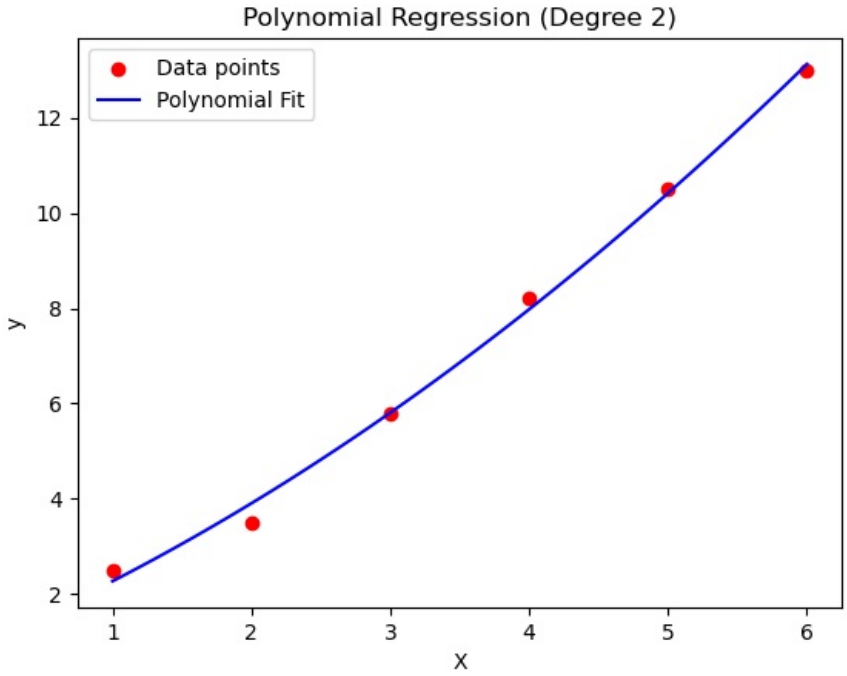
In [1]:
```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.preprocessing import PolynomialFeatures
from sklearn.linear_model import LinearRegression
from sklearn.pipeline import make_pipeline

X = np.array([1, 2, 3, 4, 5, 6]).reshape(-1, 1)
y = np.array([2.5, 3.5, 5.8, 8.2, 10.5, 13])

degree = 2
model = make_pipeline(PolynomialFeatures(degree), LinearRegression())
model.fit(X, y)

X_pred = np.linspace(1, 6, 100).reshape(-1, 1)
y_pred = model.predict(X_pred)

plt.scatter(X, y, color='red', label='Data points')
plt.plot(X_pred, y_pred, color='blue', label='Polynomial Fit')
plt.xlabel("X")
plt.ylabel("y")
plt.legend()
plt.title("Polynomial Regression (Degree 2)")
plt.show()
```



In [ ]: