Regression Assignment Q&A

1. What is Simple Linear Regression?
Simple Linear Regression is a statistical method used to model the relationship between a single independent variable (X) and a dependent variable (Y) by fitting a straight line through the data points. This line, called the regression line, is used to predict the value of Y based on X. The relationship is assumed to be linear, meaning changes in X result in proportional changes in Y.

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

- Linearity: The relationship between X and Y must be linear.
- Independence: Observations should be independent of each other.
- Homoscedasticity: The variance of residuals (errors) should remain constant across all values of X.
- Normality of residuals: The residuals should be normally distributed.
  These assumptions ensure the reliability of coefficient estimates and model predictions.

3. What does the coefficient m represent in the equation Y = mX + c?
The coefficient m is the slope of the regression line. It quantifies how much the dependent variable Y is expected to change for a one-unit increase in the independent variable X. A positive slope indicates a direct relationship, while a negative slope indicates an inverse relationship.

4. What does the intercept c represent in the equation Y = mX + c?
The intercept c is the predicted value of Y when X is zero. It represents the point at which the regression line crosses the Y-axis and serves as the baseline level of Y before considering the effect of X.

5. How do we calculate the slope m in Simple Linear Regression?
The slope m is calculated using the least squares method:

m=∑(Xi−X¯)(Yi−Y¯)∑(Xi−X¯)2m = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}m=∑(Xi−X¯)2∑(Xi−X¯)(Yi−Y¯)

This formula measures how X and Y co-vary relative to the variance in X.

6. What is the purpose of the least squares method in Simple Linear Regression?
The least squares method minimizes the sum of the squared differences between actual values (Y) and predicted values (Ŷ). This method ensures the line is the "best fit" by reducing the total error in prediction.

7. How is the coefficient of determination (R^2) interpreted in Simple Linear Regression?
R^2 measures the proportion of variance in the dependent variable that is explained by the independent variable. An R^2 of 0.85 means 85% of the variability in Y is explained by X. It ranges from 0 to 1, with values closer to 1 indicating a better fit.

8. What is Multiple Linear Regression?
Multiple Linear Regression (MLR) extends Simple Linear Regression by using two or more independent variables to predict the dependent variable. The model is:

Y=β0+β1X1+β2X2+⋯+βnXn+ϵY = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilonY=β0+β1X1+β2X2+⋯+βnXn+ϵ

MLR helps model more complex relationships and captures the combined effect of multiple predictors on the target variable.

9. What is the main difference between Simple and Multiple Linear Regression?
Simple Linear Regression uses one independent variable, while Multiple Linear Regression uses two or more. MLR can model more complex and realistic scenarios with multiple factors influencing the outcome.

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

- Linearity between each predictor and the target
- Independence of errors
- Homoscedasticity (constant error variance)
- Normality of residuals

- No multicollinearity (independent variables should not be highly correlated)

11. What is heteroscedasticity, and how does it affect the results of a Multiple Linear Regression model?
Heteroscedasticity occurs when the residuals have non-constant variance. This violates one of the key assumptions and can lead to inefficient estimates and incorrect conclusions from hypothesis tests (e.g., t-tests for coefficients).

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

- Remove or combine highly correlated predictors
- Use Principal Component Analysis (PCA)
- Apply regularization techniques like Ridge or Lasso regression
  Multicollinearity inflates variance of coefficients and can make them unstable.

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

- One-hot encoding: Converts categories into binary columns.
- Label encoding: Assigns a unique number to each category.
- Ordinal encoding: Used when categories have a meaningful order.

14. What is the role of interaction terms in Multiple Linear Regression?
Interaction terms model the combined effect of two variables that might not be captured by individual effects. For example, the effect of X1 * X2 shows how X1's impact on Y changes at different levels of X2.

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

- In Simple Regression, it represents the expected Y when X = 0.

- In Multiple Regression, it's the expected Y when all predictors are 0 — which might not be a meaningful real-world situation.

16. What is the significance of the slope in regression analysis, and how does it affect predictions?
The slope represents the change in Y for a one-unit change in X, assuming other variables are held constant. It determines the direction and strength of the relationship between predictors and the target.

17. How does the intercept in a regression model provide context for the relationship between variables?
The intercept serves as the baseline value of Y before any effect from the independent variables. It helps interpret the model, especially when X = 0 has practical meaning.

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

- R^2 always increases with more variables, even if they're irrelevant.
- It doesn't account for overfitting or complexity.
- It doesn't indicate if a model is appropriate or significant.
  Use Adjusted R^2, AIC, BIC, and residual analysis along with R^2.

19. How would you interpret a large standard error for a regression coefficient?
A large standard error indicates that the coefficient estimate is unstable and could vary greatly with new data. It may mean the variable isn't statistically significant.

20. How can heteroscedasticity be identified in residual plots, and why is it important to address?
If residuals fan out or form patterns in a residual vs. predicted plot, it signals heteroscedasticity. Addressing it improves the accuracy of standard errors and significance tests.

21. What does it mean if a Multiple Linear Regression model has a high R^2 but low adjusted R^2?
It suggests that irrelevant variables have been added, inflating R^2 but not improving the model's real explanatory power.

22. Why is it important to scale variables in Multiple Linear Regression?
Scaling ensures that variables with large ranges don't dominate the model. It also improves convergence for algorithms like Ridge, Lasso, or Gradient Descent.

23. What is polynomial regression?
Polynomial regression models non-linear relationships by adding polynomial terms (e.g., X2,X3X^2, X^3X2,X3) to a linear model. It captures curvature in data.

24. How does polynomial regression differ from linear regression?
Linear regression fits a straight line, while polynomial regression fits a curved line by including higher powers of the independent variable.

25. When is polynomial regression used?
When the data shows non-linear trends that can't be modeled accurately by a straight line.

26. What is the general equation for polynomial regression?

Y=β0+β1X+β2X2+⋯+βnXn+ϵY = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_n X^n + \epsilonY=β0+β1X+β2X2+⋯+βnXn+ϵ

27. Can polynomial regression be applied to multiple variables?
Yes. You can create interaction and polynomial terms for multiple variables, but the model becomes more complex and computationally expensive.

28. What are the limitations of polynomial regression?

- Risk of overfitting
- Poor extrapolation beyond data range

- Can become unstable and hard to interpret at high degrees

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

- Cross-validation
- Adjusted R^2
- AIC/BIC (penalize complexity)
- Residual analysis

30. Why is visualization important in polynomial regression?
It helps you see the fit, identify overfitting or underfitting, and communicate model behavior effectively.

31. How is polynomial regression implemented in Python?

```
from sklearn.preprocessing import PolynomialFeatures
from sklearn.linear_model import LinearRegression
from sklearn.pipeline import make_pipeline

# Create and fit model
model = make_pipeline(PolynomialFeatures(degree=3), LinearRegression())
model.fit(X, y)
y_pred = model.predict(X)
```