1. **What does R-squared represent in a regression model?**
R-squared (R²) is a statistical measure that represents the proportion of the variance in the dependent variable that is predictable from the independent variables. It ranges from 0 to 1, where:

- 0 means the model explains none of the variance.
- 1 means the model explains all the variance. A higher R-squared indicates a better fit of the model to the data.

2. **What are the assumptions of linear regression?**
Linear regression relies on the following assumptions:

- **Linearity**: The relationship between independent and dependent variables is linear.
- **Independence**: The residuals (errors) are independent.
- **Homoscedasticity**: The residuals have constant variance.
- **Normality**: The residuals are normally distributed.
- **No multicollinearity**: Independent variables are not highly correlated.

3. **What is the difference between R-squared and Adjusted R-squared?**

- **R-squared**: Measures the proportion of variance explained by the model but can increase with the addition of more predictors, even if they are irrelevant.
- **Adjusted R-squared**: Adjusts for the number of predictors and penalizes the addition of irrelevant predictors. It provides a more accurate measure of model performance.

4. **Why do we use Mean Squared Error (MSE)?**
MSE measures the average of the squared differences between observed and predicted values. It penalizes larger errors more than smaller ones, making it sensitive to outliers. It provides a way to quantify model accuracy.

5. **What does an Adjusted R-squared value of 0.85 indicate?**
An Adjusted R-squared of 0.85 indicates that 85% of the variance in the dependent variable is explained by the model, accounting for the number of predictors. This suggests a strong model fit.

6. **How do we check for normality of residuals in linear regression?**

- **Histogram of residuals**
- **Q-Q plot (Quantile-Quantile plot)**
- **Shapiro-Wilk test**
- **Kolmogorov-Smirnov test**

- **Anderson-Darling test** If residuals deviate significantly from normal, it may indicate model misspecification.

7. **What is multicollinearity, and how does it impact regression?**
   Multicollinearity occurs when independent variables are highly correlated. This can lead to:
- Unstable coefficients
- Reduced interpretability
- Inflated standard errors
- Difficulty in determining the effect of individual predictors

8. **What is Mean Absolute Error (MAE)?**
   MAE measures the average absolute difference between observed and predicted values. It is less sensitive to outliers than MSE and provides a straightforward measure of prediction error.

9. **What are the benefits of using an ML pipeline?**
- Automates data preprocessing, model training, and evaluation
- Ensures consistency and reproducibility
- Streamlines workflows
- Facilitates model deployment
- Reduces errors and improves efficiency

10. **Why is RMSE considered more interpretable than MSE?**
    RMSE (Root Mean Squared Error) is the square root of MSE, expressed in the same units as the target variable, making it easier to interpret and compare to actual data.

11. **What is pickling in Python, and how is it useful in ML?**
    Pickling serializes Python objects (e.g., models, datasets) into byte streams for saving and reloading. This is useful for preserving trained models and avoiding retraining.

12. **What does a high R-squared value mean?**
    A high R-squared indicates that a large proportion of the variance in the dependent variable is explained by the model, suggesting a good fit.

13. **What happens if linear regression assumptions are violated?**

- **Linearity violation**: Model underperformance.
- **Independence violation**: Overfitting or autocorrelation.
- **Homoscedasticity violation**: Biased standard errors.
- **Normality violation**: Inaccurate confidence intervals.
- **Multicollinearity**: Unstable coefficients.

14. **How can we address multicollinearity in regression?**

- **Remove highly correlated predictors**
- **Use dimensionality reduction (e.g., PCA)**
- **Combine correlated variables**
- **Regularization techniques (Lasso, Ridge)**

15. **Why do we use pipelines in machine learning?**
Pipelines ensure that data preprocessing, feature engineering, and modeling steps are performed consistently and sequentially. This improves reproducibility and reduces errors.

16. **How is Adjusted R-squared calculated?**
Adjusted R-squared is calculated as: Where:

- **n** = number of observations
- **k** = number of predictors
- **R²** = R-squared value

17. **Why is MSE sensitive to outliers?**
MSE squares the errors, amplifying larger deviations. This makes it sensitive to outliers, which disproportionately affect the overall error.

18. **What is the role of homoscedasticity in linear regression?**
Homoscedasticity ensures that residuals have constant variance across all levels of independent variables. It stabilizes coefficient estimates and maintains the accuracy of statistical tests.

19. **What is Root Mean Squared Error (RMSE)?**
RMSE is the square root of MSE, providing a measure of the average magnitude of prediction errors in the same units as the target variable.

20. **Why is pickling considered risky?**

- Security risks (malicious code execution during unpickling)
- Compatibility issues between Python versions
- Corruption or loss of data integrity

21. **What alternatives exist to pickling for saving ML models?**

- **Joblib** (more efficient for large objects)
- **ONNX** (cross-platform, interoperable models)
- **HDF5** (hierarchical data format)
- **TensorFlow/Keras save_model**

22. **What is heteroscedasticity, and why is it a problem?**

Heteroscedasticity means residual variance changes across levels of an independent variable. It can:

- Lead to biased coefficient estimates
- Reduce the reliability of hypothesis tests

23. **How does adding irrelevant predictors affect R-squared and Adjusted R-squared?**

- **R-squared**: Increases regardless of predictor relevance.
- **Adjusted R-squared**: Decreases if the added predictors do not improve model performance.