**Multiple Linear Regression**

1. **What is Multiple Linear Regression?**
   * **Answer:** Multiple Linear Regression is a statistical technique that models the relationship between one dependent variable and two or more independent variables by fitting a linear equation to the observed data.
2. **How does Multiple Linear Regression differ from Simple Linear Regression?**
   * **Answer:** Simple Linear Regression involves one independent variable, while Multiple Linear Regression involves two or more independent variables.
3. **What assumptions are made in Multiple Linear Regression?**
   * **Answer:** The main assumptions are linearity, independence, homoscedasticity, normality of errors, and no multicollinearity among predictors.
4. **What is multicollinearity, and how can it be detected?**
   * **Answer:** Multicollinearity occurs when independent variables are highly correlated. It can be detected using Variance Inflation Factor (VIF) or correlation matrices.
5. **How can multicollinearity be addressed in a model?**
   * **Answer:** It can be addressed by removing highly correlated variables, combining them, or using regularization techniques like Ridge or Lasso regression.
6. **Explain the purpose of the intercept in a Multiple Linear Regression model.**
   * **Answer:** The intercept represents the expected value of the dependent variable when all independent variables are zero.
7. **What is the coefficient of determination (R²) in Multiple Linear Regression?**
   * **Answer:** R² measures the proportion of the variance in the dependent variable that is predictable from the independent variables.
8. **How would you interpret a negative coefficient in a Multiple Linear Regression model?**
   * **Answer:** A negative coefficient indicates an inverse relationship between the independent variable and the dependent variable.
9. **Can you include categorical variables in Multiple Linear Regression? How?**
   * **Answer:** Yes, categorical variables can be included by converting them into dummy or indicator variables.
10. **What is the difference between Adjusted R² and R²?**
    * **Answer:** Adjusted R² adjusts for the number of predictors in the model, penalizing the inclusion of irrelevant variables, whereas R² simply measures the goodness of fit.
11. **Why might you choose a Multiple Linear Regression model over a more complex model?**
    * **Answer:** Multiple Linear Regression is simpler, easier to interpret, and requires fewer assumptions, making it preferable if it provides a good fit.
12. **What are residuals in Multiple Linear Regression?**
    * **Answer:** Residuals are the differences between the observed and predicted values of the dependent variable.
13. **Explain the purpose of feature scaling in the context of Multiple Linear Regression.**
    * **Answer:** Feature scaling helps in aligning all variables on a similar scale, which can improve the accuracy of the model and reduce computational complexity.
14. **What are interaction terms, and when should they be included in a Multiple Linear Regression model?**
    * **Answer:** Interaction terms capture the combined effect of two or more independent variables on the dependent variable and should be included when there is reason to believe that the effect of one variable depends on another.
15. **What is stepwise regression?**
    * **Answer:** Stepwise regression is a method of fitting regression models in which the choice of predictive variables is carried out by an automatic procedure, usually based on statistical significance.

**Model Evaluation**

1. **What are some common metrics for evaluating a Multiple Linear Regression model?**
   * **Answer:** Common metrics include Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R².
2. **How do you interpret the p-value of a coefficient in Multiple Linear Regression?**
   * **Answer:** The p-value indicates the probability that the observed effect is due to chance. A low p-value (typically < 0.05) suggests that the coefficient is statistically significant.
3. **What is overfitting, and how can it be avoided in regression models?**
   * **Answer:** Overfitting occurs when the model captures noise instead of the underlying pattern. It can be avoided using techniques like cross-validation, regularization, or reducing the number of predictors.
4. **What is cross-validation, and why is it important?**
   * **Answer:** Cross-validation is a technique for assessing how well a model generalizes to an independent dataset. It helps in preventing overfitting and gives an estimate of model performance.
5. **Explain the difference between train-test split and cross-validation.**
   * **Answer:** A train-test split divides the data into two sets, while cross-validation divides it into multiple subsets to ensure every observation has a chance to be in the training and test set.
6. **What is the purpose of regularization in regression models?**
   * **Answer:** Regularization penalizes large coefficients to prevent overfitting and improve model generalization.
7. **What is the difference between Ridge and Lasso regression?**
   * **Answer:** Ridge regression adds a penalty equal to the sum of the squared coefficients, while Lasso adds a penalty equal to the sum of the absolute values of the coefficients. Lasso can shrink some coefficients to zero, effectively performing variable selection.
8. **How would you evaluate the performance of a regression model on unseen data?**
   * **Answer:** The performance can be evaluated using metrics like MAE, MSE, RMSE, and R² on the test dataset, as well as cross-validation scores.
9. **What is the bias-variance tradeoff?**
   * **Answer:** The bias-variance tradeoff is the balance between the error introduced by bias (error due to overly simplistic models) and variance (error due to overly complex models).
10. **What are some common pitfalls in model evaluation?**
    * **Answer:** Common pitfalls include data leakage, overfitting, underfitting, ignoring assumptions, and using inappropriate evaluation metrics.

**Logical Questions**

1. **Given two models with the same R² but different coefficients, how would you decide which one to use?**
   * **Answer:** I would examine the p-values, interpretability, multicollinearity, and domain knowledge to decide which model better represents the underlying relationships.
2. **If a model has a high R² but poor performance on new data, what might be the problem?**
   * **Answer:** The model may be overfitting, capturing noise instead of the underlying pattern, or there might be a data leakage issue.
3. **Why might you choose to use a simpler model over a complex one, even if the complex model has better performance?**
   * **Answer:** A simpler model is often preferred for its interpretability, ease of implementation, and lower risk of overfitting, especially if the performance gain from the complex model is marginal.
4. **How would you handle missing data in a dataset before applying Multiple Linear Regression?**
   * **Answer:** I would consider options like imputation (mean, median, mode), using models to predict missing values, or removing rows/columns with missing data depending on the extent and importance of the missing values.
5. **If adding a new feature to your model decreases R², what does that indicate?**
   * **Answer:** It indicates that the new feature may not be contributing useful information and could be introducing noise, leading to a worse fit. This might also suggest multicollinearity or overfitting.