

Evaluation Metrics and Regression Implementation

Q1. What does R-squared represent in a regression model?

R-squared (R^2) represents the proportion of variance in the dependent variable that is explained by the independent variables in a regression model. Its value lies between 0 and 1. A higher R^2 value indicates a better model fit.

Q2. What are the assumptions of linear regression?

The main assumptions of linear regression are linearity, independence of observations, homoscedasticity (constant variance of residuals), normality of residuals, and absence of multicollinearity among predictors.

Q3. What is the difference between R-squared and Adjusted R-squared?

R-squared increases whenever a new predictor is added, even if it is not useful. Adjusted R-squared penalizes unnecessary predictors and increases only when a new variable improves the model, making it more reliable.

Q4. Why do we use Mean Squared Error (MSE)?

Mean Squared Error measures the average squared difference between actual and predicted values. It penalizes larger errors more strongly and is widely used as a loss function in regression models.

Q5. What does an Adjusted R-squared value of 0.85 indicate?

An Adjusted R-squared value of 0.85 means that 85% of the variation in the dependent variable is explained by the model after accounting for the number of predictors, indicating a strong model.

Q6. How do we check for normality of residuals in linear regression?

Normality of residuals can be checked using histograms, Q–Q plots, or statistical tests such as the Shapiro–Wilk test. Residuals should approximately follow a normal distribution.

Q7. What is multicollinearity, and how does it impact regression?

Multicollinearity occurs when independent variables are highly correlated. It makes coefficient estimates unstable and difficult to interpret, reducing the reliability of regression results.

Q8. What is Mean Absolute Error (MAE)?

Mean Absolute Error is the average of the absolute differences between actual and predicted values. It is easy to interpret because it is expressed in the same unit as the target variable.

Q9. What are the benefits of using an ML pipeline?

ML pipelines prevent data leakage, ensure consistent preprocessing, improve reproducibility, and simplify model

deployment by combining preprocessing and modeling steps.

Q10. Why is RMSE considered more interpretable than MSE?

RMSE is the square root of MSE and is expressed in the same unit as the dependent variable, making it easier to understand and interpret prediction errors.

Q11. What is pickling in Python, and how is it useful in ML?

Pickling is the process of serializing Python objects. In machine learning, it is used to save trained models so they can be reused later without retraining.

Q12. What does a high R-squared value mean?

A high R-squared value indicates that the model explains a large portion of the variance in the dependent variable, though it does not guarantee that the model is correct or unbiased.

Q13. What happens if linear regression assumptions are violated?

Violation of assumptions can lead to biased coefficients, incorrect standard errors, and unreliable hypothesis testing, reducing the validity of the model.

Q14. How can we address multicollinearity in regression?

Multicollinearity can be reduced by removing correlated variables, using VIF analysis, applying Ridge or Lasso regression, or using dimensionality reduction techniques like PCA.

Q15. How can feature selection improve model performance in regression analysis?

Feature selection removes irrelevant or redundant variables, reduces overfitting, improves interpretability, and enhances model generalization.

Q16. How is Adjusted R-squared calculated?

Adjusted R-squared is calculated as: $\text{Adjusted } R^2 = 1 - [(1 - R^2)(n - 1)/(n - k - 1)]$, where n is the number of observations and k is the number of predictors.

Q17. Why is MSE sensitive to outliers?

MSE squares the errors, so large errors caused by outliers have a much greater impact on the final value, making it highly sensitive to extreme observations.

Q18. What is the role of homoscedasticity in linear regression?

Homoscedasticity ensures constant variance of residuals across all levels of predictors, which is essential for reliable statistical inference and confidence intervals.

Q19. What is Root Mean Squared Error (RMSE)?

RMSE is the square root of Mean Squared Error and measures the average magnitude of prediction errors in the same unit as the dependent variable.

Q20. Why is pickling considered risky?

Pickling is risky because loading pickle files from untrusted sources can execute malicious code, and pickled files may break across Python versions.

Q21. What alternatives exist to pickling for saving ML models?

Alternatives include Joblib, ONNX, PMML, and saving model parameters manually. Joblib is commonly preferred for scikit-learn models.

Q22. What is heteroscedasticity, and why is it a problem?

Heteroscedasticity occurs when residual variance is not constant. It leads to inefficient estimates and unreliable hypothesis testing.

Q23. How can interaction terms enhance a regression model's predictive power?

Interaction terms allow the effect of one predictor to depend on another, enabling the model to capture more complex relationships and improve prediction accuracy.