1. What assumptions does linear regression make?

Answer:

1. **Linearity**: The relationship between each predictor and the response is linear.

2. **Independence**: Observations (and residuals) are independent.

3. **Homoscedasticity**: Residuals have constant variance across predictor values.

4. **Normality of Errors**: Residuals (errors) are normally distributed (important for inference).

5. **No or Low Multicollinearity**: Predictors are not highly correlated with each other.

2. How do you interpret the coefficients?

Answer:

- A coefficient represents the expected change in the response variable for a one-unit increase in predictor , holding other predictors constant.
- **Sign**: Positive indicates a direct relationship; negative indicates an inverse relationship.
- **Magnitude**: Larger absolute values imply stronger effects (e.g., means each unit increase in increases the prediction by 2.5 units).

3. What is R^2 score and its significance?

Answer:

- **R^2** (coefficient of determination) measures the proportion of variance in the response explained by the model.
- Range: 0 to 1; values closer to 1 indicate a better fit.
- **Adjusted R^2** is preferred in multiple regression to account for the number of predictors.

4. When would you prefer MSE over MAE?

Answer:

- **MSE** (Mean Squared Error) penalizes larger errors more heavily (squaring amplifies outliers).
- MSE is differentiable, making it suitable for analytic solutions and gradient-based optimization.
- Use **MAE** (Mean Absolute Error) when robustness to outliers is more important.

5. How do you detect multicollinearity?

Answer:

- **Variance Inflation Factor (VIF)**: A VIF > 5–10 indicates problematic collinearity.

- **Correlation Matrix**: Look for predictor pairs with high absolute correlation ().
- **Condition Number** of the design matrix: Values > 30 signal ill-conditioning.

6. What is the difference between simple and multiple regression?

Answer:

- **Simple Regression**: One predictor ().
- **Multiple Regression**: Two or more predictors ().
- Multiple regression models more complex relationships and controls for confounders.

7. Can linear regression be used for classification?

Answer:

- Not directly, as it predicts continuous outcomes.
- Thresholding its predictions can produce a classifier (e.g., >0.5 → class 1), but this may yield values outside [0,1] and lacks probabilistic interpretation.
- **Logistic regression** is preferred for classification tasks.

8. What happens if you violate regression assumptions?

Answer:

- **Biased/Inefficient Estimates**: E.g., heteroscedasticity inflates variance of estimates.
-

Invalid Inference: P-values and confidence intervals become unreliable.

- **Poor Generalization**: Misspecified models may not predict well on new data.

- **Remedies**: Transformations, robust methods, regularization, or revisiting model specification.