LR - Assumptions

Assumptions of Linear Regression

Linear regression, a fundamental statistical method for modeling the relationship between a dependent variable and one or more independent variables, relies on several key assumptions:

1. **Linearity**: The relationship between the independent variables and the dependent variable is linear. This means the change in the dependent variable due to a one-unit change in an independent variable is constant.
2. **Independence**: Observations are independent of each other. This is crucial for the reliability of the statistical inferences made from the model.
3. **Homoscedasticity**: The variance of residual errors is constant across all levels of the independent variables. If the variance changes (heteroscedasticity), it can lead to inefficient estimates.
4. **Normal Distribution of Errors**: The residuals (errors) of the model are normally distributed. This assumption facilitates the creation of confidence intervals and hypothesis tests.
5. **No or Little Multicollinearity**: Independent variables should not be too highly correlated with each other. High correlation can make it difficult to isolate the effect of each independent variable on the dependent variable.

Let's delve into each assumption of linear regression in more detail:

1. Linearity

The linearity assumption in linear regression states that there is a linear relationship between the dependent variable and each of the independent variables. This means the change in the dependent variable is expected to be directly proportional to the change in an independent variable. For a regression model $Y=\beta0+\beta1X1+\epsilon Y=\beta0+\beta1X1+\epsilon$, the effect of a one-unit change in $X1$ on Y is constant, and it's quantified by $\beta1$ (or weight W1)

Checking the assumption: You can check for linearity by plotting each predictor against the target variable using scatter plots. If plots show curvilinear patterns, the data may need transformation.

Remedying violations: Transformations such as logarithmic, square root, or reciprocal can be applied to the dependent and/or independent variables to correct non-linear relationships.

2. Independence

The independence assumption implies that the observations are independent of each other, which is crucial for the generalization of the regression results. In time-series data, where data points naturally depend on previous points, this assumption is violated.

Checking the assumption: Plotting residuals can sometimes show patterns indicative of non-independence. For time-series data, autocorrelation plots are used.

Remedying violations: For time-series data, methods like ARIMA models may be appropriate. For panel or clustered data, mixed models or generalized estimating equations can adjust for grouped data.

3. Homoscedasticity

Homoscedasticity means that the residuals (differences between observed and predicted values) should have constant variance across all levels of the independent variables. If the variance of residuals increases or decreases with the predicted values, it indicates heteroscedasticity, which can lead to inefficient estimates and affect the validity of hypothesis tests.

Checking the assumption: Residual vs. fitted value plots are typically used. Ideally, this plot shouldn’t show any pattern or funnel shape.

Remedying violations: Transformation of the dependent variable (e.g., log transformation if the variance increases with the level of impact) or using robust regression methods or weighted least squares can help (Variation of LR

and domain specific, not asked in interviews, asked only if you have used them in your project.)

4. Normal Distribution of Errors

Linear regression assumes that the residuals are normally distributed. This assumption is crucial for conducting reliable hypothesis testing including the t-tests and F-tests used in evaluating the statistical significance of the parameters.

Checking the assumption: A Q-Q (quantile-quantile) plot of the residuals can show how closely the data follow a normal distribution. The Shapiro-Wilk test is another method to test normality.

Remedying violations: Transformations like logarithmic, square root, or Box-Cox can normalize residuals. Alternatively, using non-parametric regression techniques might be necessary if transformations do not work.

5. No Multicollinearity

Multicollinearity occurs when two or more independent variables are highly correlated, making it difficult to distinguish their individual effects on the dependent variable. This can lead to inflated variances for the coefficient estimates, making them unstable and unreliable.

Checking the assumption: Checking the correlation matrix between variables can identify multicollinearity. Variance Inflation Factor (VIF) is a more quantitative measure where a VIF value greater than 10 is often considered indicative of significant multicollinearity.

Remedying violations: Removing one or more of the correlated variables can help, as can combining them into a single predictor through principal component analysis (PCA) or using regularization techniques like Ridge or Lasso regression that can shrink the coefficients of correlated predictors.