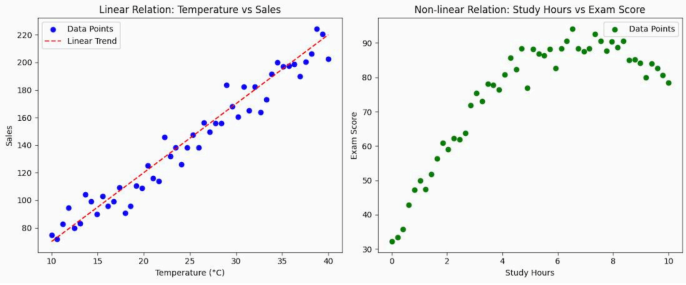
Assumptions of linear regression

Tuesday, 10 February 2026 11:38 AM

1. Linearity

The relationship between the independent and dependent variables is linear.

- The dependent variable should change proportionally with the independent variables, forming a straight-line trend.
- Curved or irregular patterns can cause underfitting and inaccurate predictions.
- When linearity fails, data transformations or non-linear models may be required.



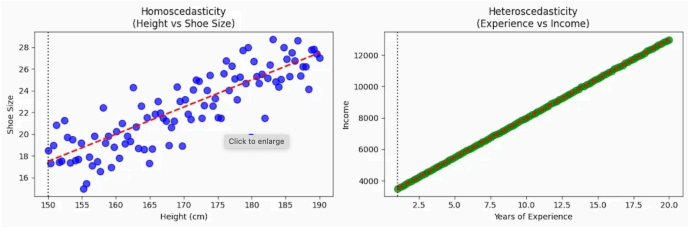
Linearity

- **Linear Relationship:** Increase in temperature results in a consistent increase in ice cream sales.
- **Non-Linear Relationship:** Increase in temperature leads to a more significant increase in ice cream sales at higher temperatures, indicating a non-linear relationship.

2. Homoscedasticity of Residuals

The variance of residuals remains constant across all levels of the independent variables.

- Residuals should appear evenly scattered, indicating uniform error spread.
- Patterns of increasing or decreasing variance lead to unreliable coefficient estimates.
- Severe heteroscedasticity may require transformations or weighted regression methods.



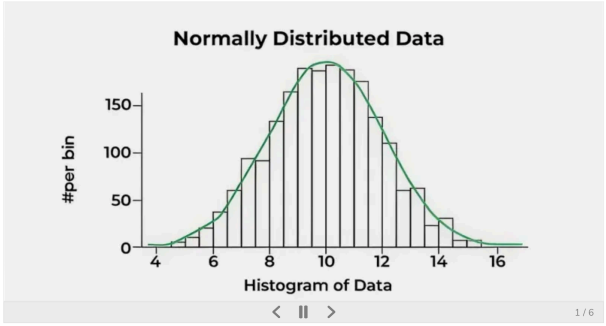
Homoscedasticity of Residuals

- **Left plot (Homoscedasticity):** The residuals are scattered evenly around the horizontal line at zero, indicating a constant variance.
- **Right plot (Heteroscedasticity):** The residuals are not evenly scattered. There is a clear pattern of increasing variance as the predicted values increase, indicating heteroscedasticity.

3. Multivariate Normality - Normal Distribution

The residuals follow a normal distribution when multiple predictors are involved.

- Normality supports valid confidence intervals, hypothesis tests and p-values.
- Skewed or peaked distributions weaken inference quality.
- Violations may be corrected through transformation or larger sample sizes.

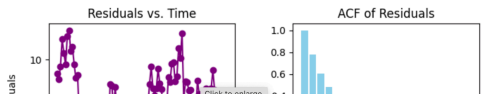


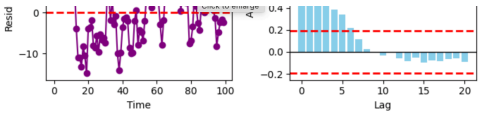
- First row shows a normally distributed dataset, as evidenced by the bell-shaped histogram and the points falling close to a straight line in the Q-Q plot.
- Second row shows a dataset that is too peaked in the middle, indicating a deviation from normality.
- Third row shows a skewed dataset, also indicating a deviation from normality.

4. Independence of Errors

Residuals must not correlate with each other across observations.

- Correlated errors suggest the model missed temporal or patterned structure.
- Autocorrelation can inflate significance and mislead conclusions.
- Time-series data often require specialized methods to resolve this.





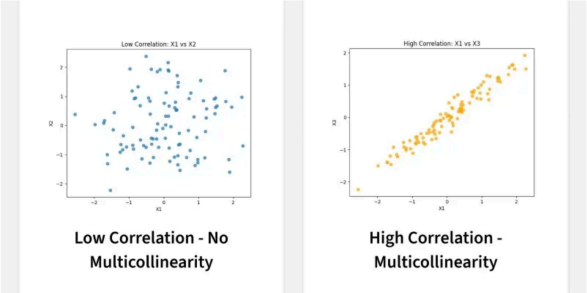
Independence of Errors

- **Residuals vs. Time plot:** Shows a random scatter of points, suggesting no clear pattern or correlation over time.
- **ACF of Residuals plot:** Shows a few spikes at low lags but they are not significant enough to indicate strong autocorrelation.

5. Lack of Multicollinearity

The independent variables are not highly correlated with each other.

- Strong collinearity inflates coefficient variance and reduces interpretability.
- It becomes difficult to assess the true contribution of each predictor.
- Feature selection or regularization helps reduce the effect.



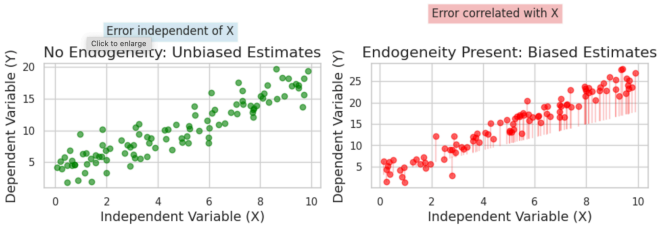
Lack of Multicollinearity

- **Left plot:** Shows scattered points with low correlation, indicating no multicollinearity.
- **Right plot:** Shows a strong linear pattern with high correlation, indicating multicollinearity.
- **Low correlation:** Features add unique information, while high correlation makes features redundant.

6. Absence of Endogeneity

Independent variables in the regression model should not be correlated with the error term.

- Endogeneity causes biased and inconsistent parameter estimates.
- Inference based on such coefficients becomes unreliable.
- Instrumental variables or additional predictors can address this issue.



Absence of Endogeneity

- **No Endogeneity (Left Side):** Residuals are independent of the input variable so the regression line fits correctly, producing unbiased estimates.
- **With Endogeneity (Right Side):** Residuals correlate with the input variable, causing the regression line to shift and produce biased, incorrect estimates.

https://www.geeksforgeeks.org/machine-learning/assumptions-of-linear-regression/