Regression Assumptions & Diagnostics

Supervised Learning

Daniel E. Acuna

Associate Professor, University of Colorado Boulder

Contents of This Video

In this video, we will cover:

- The four key assumptions of linear regression
- Why these assumptions matter for prediction
- Diagnostic tools to check assumptions
- Residual plots and how to interpret them
- Dealing with outliers and high leverage points
- What to do when assumptions are violated

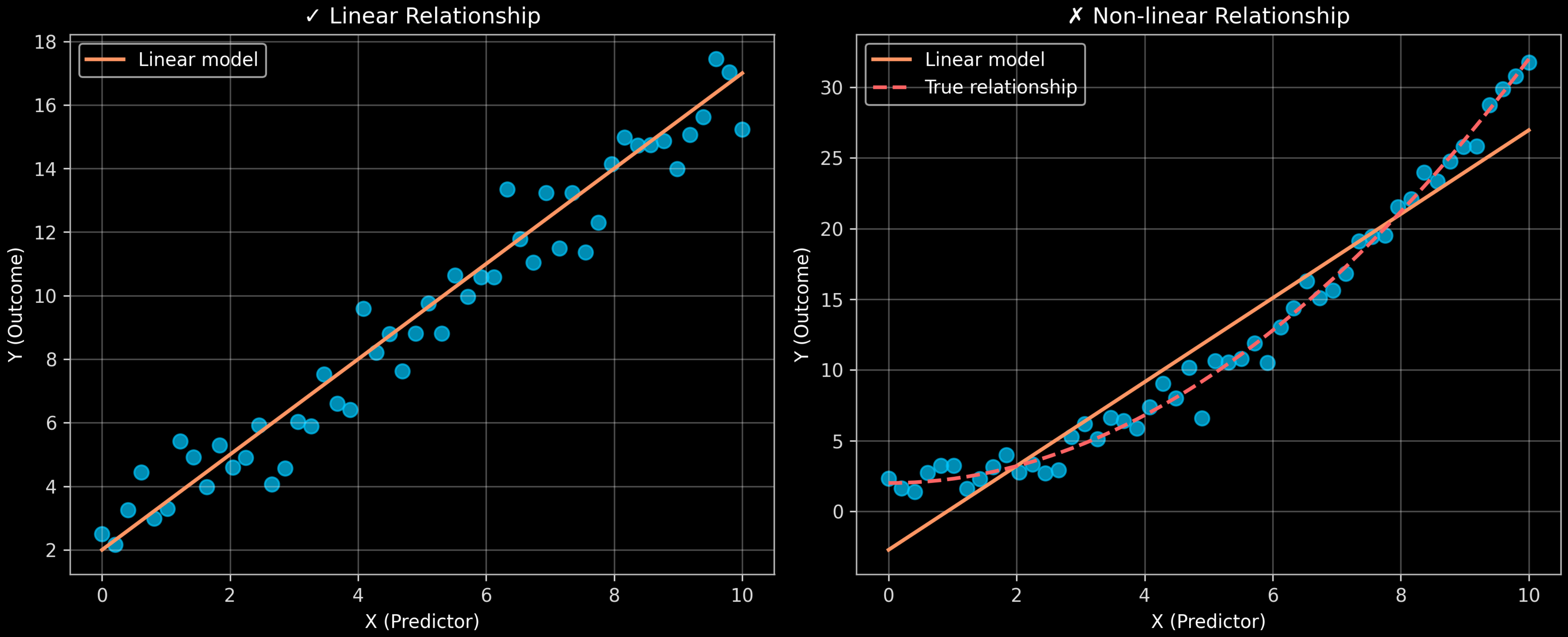
The Four Key Assumptions

The main assumptions of classical linear regression:

- **Linearity:** The relationship between predictors and outcome is linear
- **Independent errors:** Residuals are independent of each other
- **Constant variance (Homoscedasticity):** Residuals have constant variance
- **Normality of errors:** Residuals are approximately normally distributed

Linearity Assumption

The relationship follows the linear form our model assumes



If violated: Model systematically misses patterns, hurting prediction accuracy

Independence Assumption

Each observation’s error is unrelated to others

Usually reasonable when:

- Data points are individual and unconnected
- Examples: different households, separate market regions

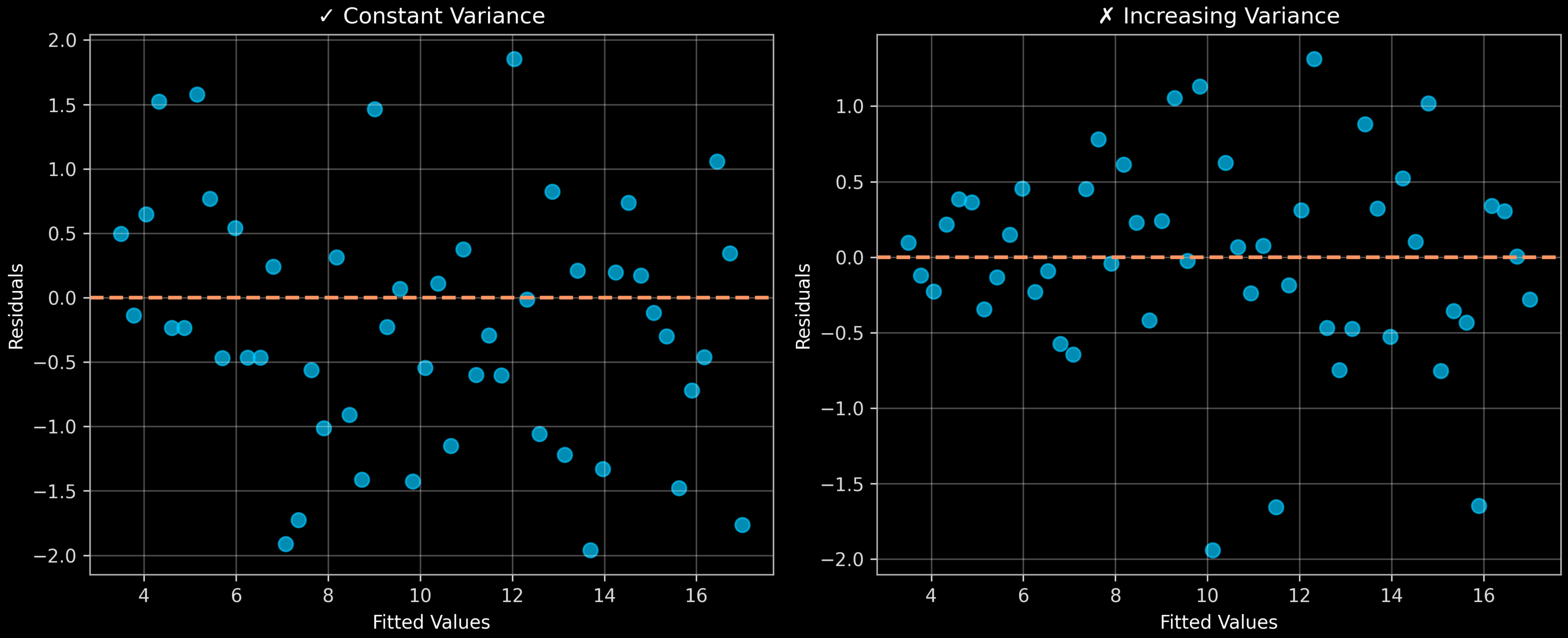
Violated when:

- Repeated measurements of same person/entity
- Time series data with autocorrelation
- Spatial clustering in geographic data

Consequence: Can lead to overfitting and poor generalization to new data

Constant Variance (Homoscedasticity)

Residuals should have equal spread across all prediction levels

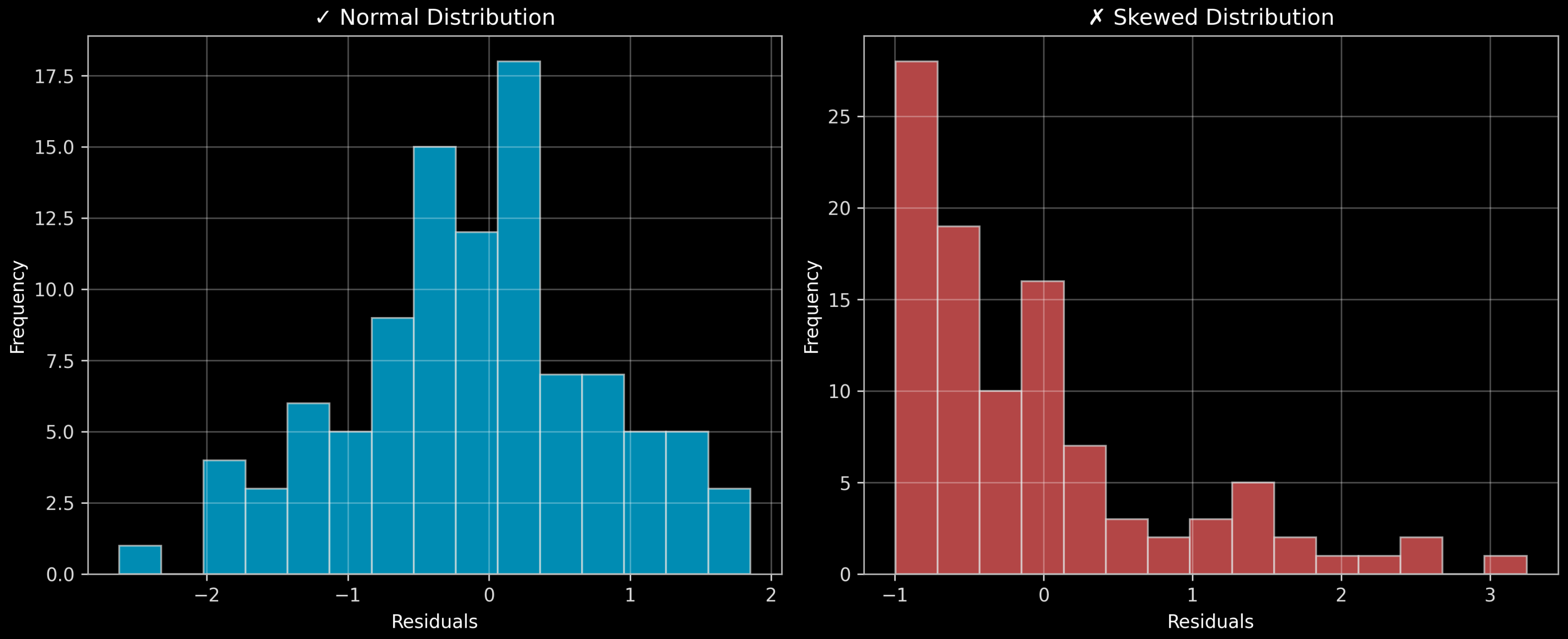


Example: Income prediction - errors small for low income, large for high income

Normality of Errors

Residuals should be approximately normally distributed

- **Not critical for prediction** - model can still make good predictions
- **Less important than other assumptions** for ML applications
- **Less crucial with large samples**
- **May indicate need for data transformation**

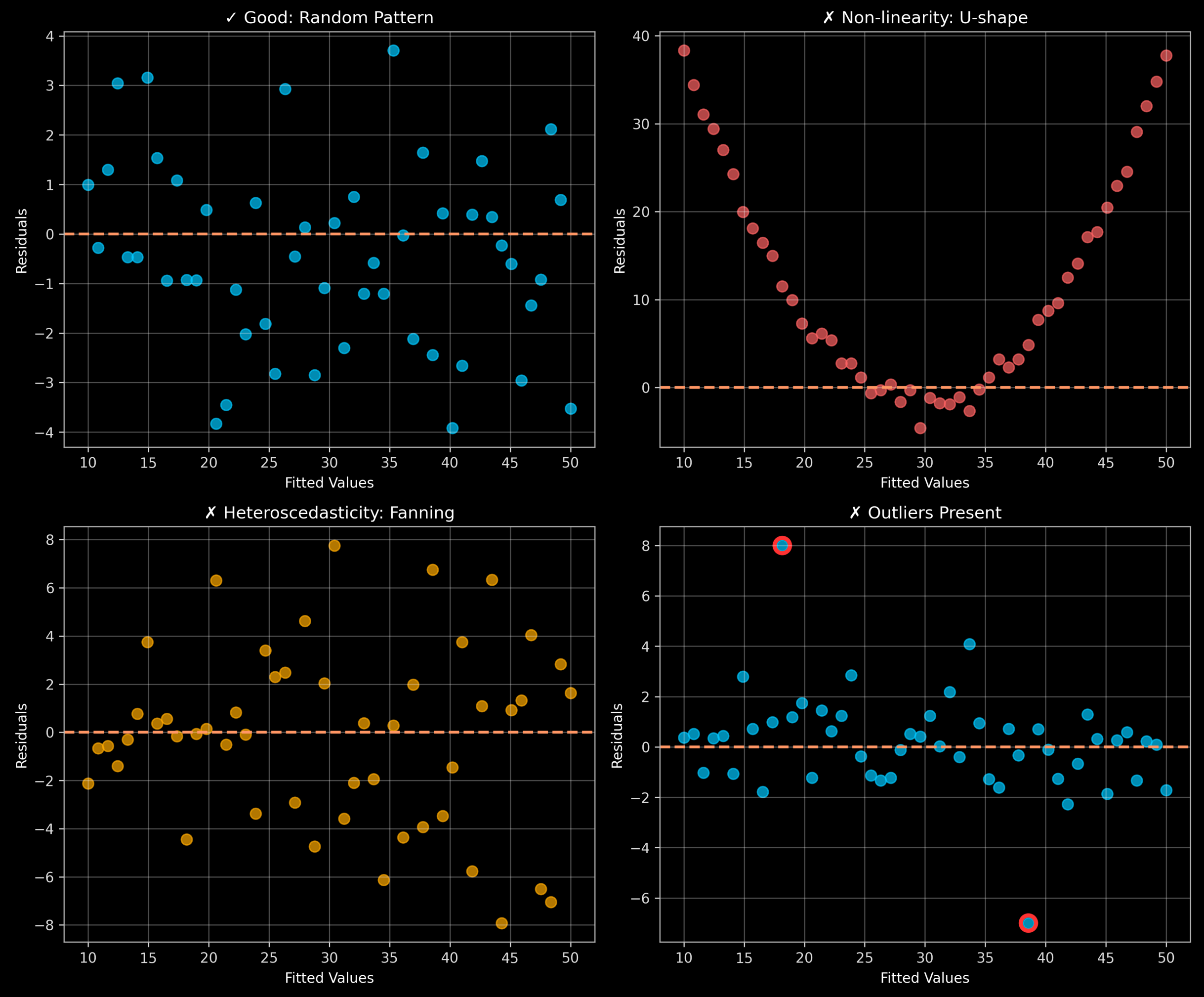


✓ Normal Distribution

✗ Skewed Distribution

Residual Plots: Primary Diagnostic Tool

Plot residuals vs fitted values to check assumptions

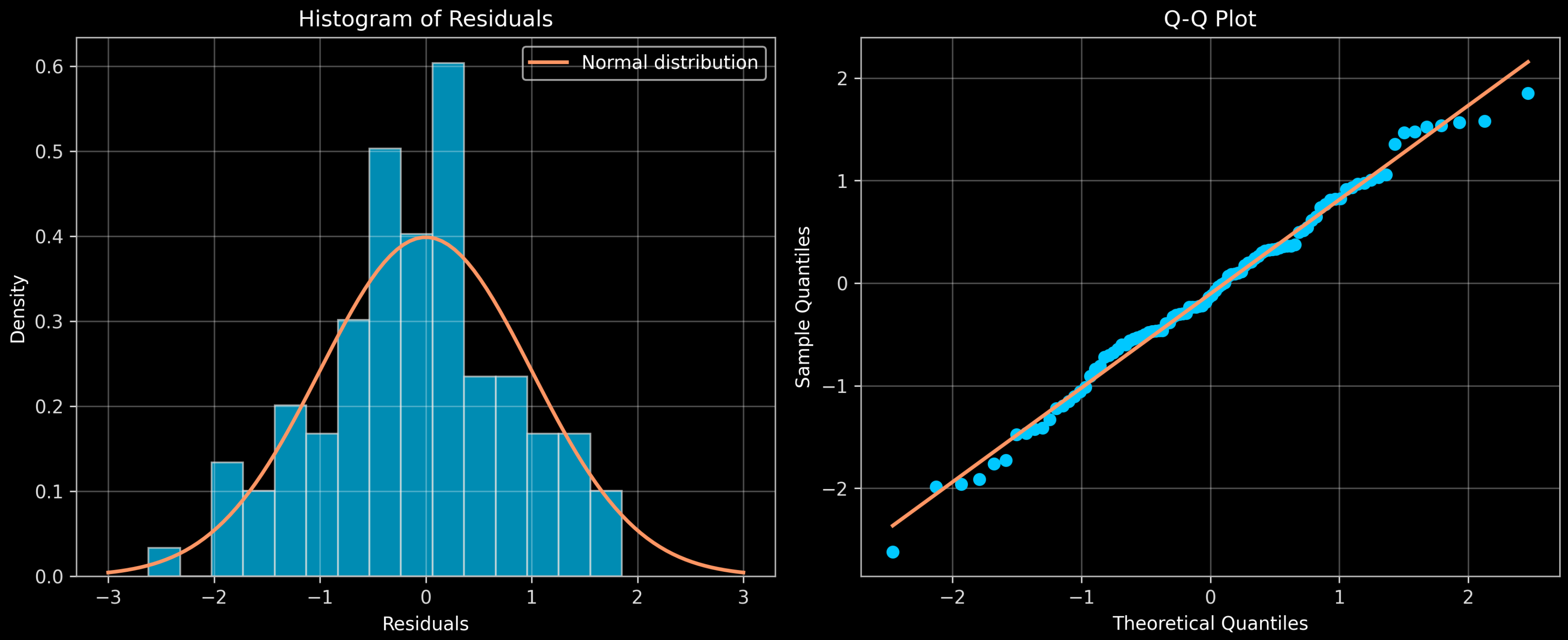


Ideal: Random cloud with no pattern **Problems:** Curves, fanning, outliers

Checking Normality of Residuals

Tools for assessing normality:

- **Histogram** of residuals
- **Q-Q plot** (quantile-quantile plot)



Q-Q plot interpretation: Points on diagonal line = normal residuals

Detecting Outliers and Leverage Points

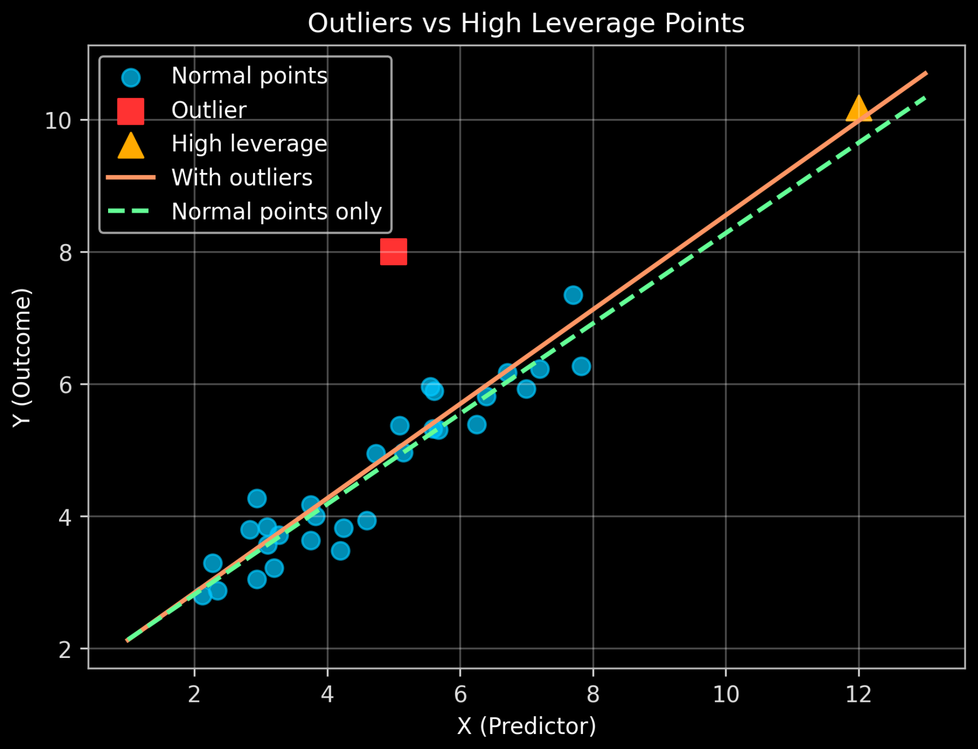
Two types of problematic observations:

Outliers

- Large residuals (unusual Y values)
- Far from regression line
- Can pull the line toward them

High Leverage Points

- Extreme predictor values (unusual X values)
- Can have disproportionate influence
- May or may not be outliers



When Assumptions Are Violated

No dataset perfectly meets all assumptions - focus on major problems

Common solutions:

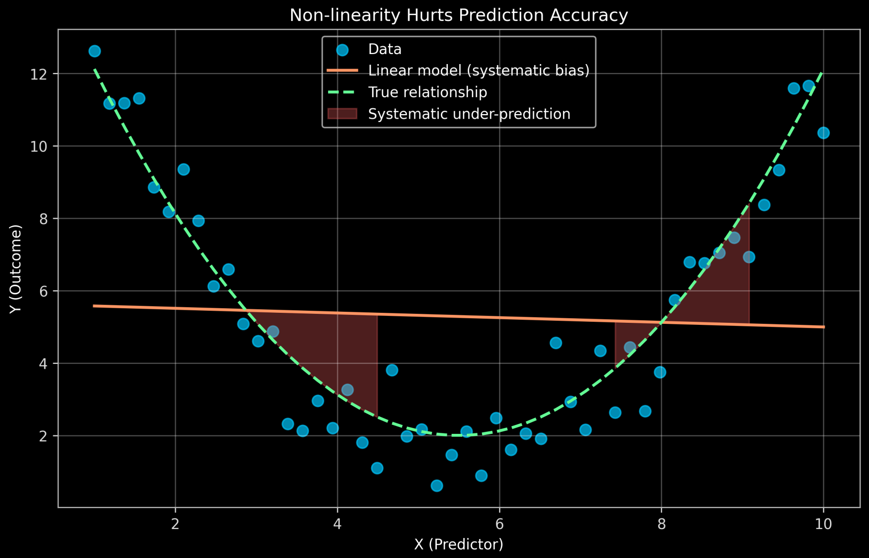
- **Linearity violation:** Add polynomial terms, use flexible models
- **Non-constant variance:** Transform Y (e.g., log transformation)
- **Normality violation:** Often not critical for prediction (large samples)
- **Independence violation:** Use specialized models (time series, clustering)

Key insight: Linearity is most critical for prediction accuracy

Focus on Predictive Accuracy

Since our focus is prediction, some assumptions are less crucial:

- **Normality:** Not critical for predictive ability
- **Independence:** Still important but often assumed
- **Linearity:** Most critical - non-linear relationships hurt prediction
- **Constant variance:** Can affect prediction reliability



Gross violations of linearity will hurt model accuracy

What We’ve Covered

In this video, we’ve learned:

- Four key assumptions: linearity, independence, constant variance, normality
- Diagnostic tools: residual plots, histograms, Q-Q plots
- How to identify outliers and high leverage points
- Solutions when assumptions are violated
- Why linearity is most critical for prediction accuracy
- Practical approach: focus on major violations, not perfection

Diagnostics help us improve model performance and reliability