# Fitting a Linear Model & Assessing Fit

Supervised Learning

Daniel E. Acuna
Associate Professor, University of Colorado Boulder

# Contents of This Video

In this video, we will cover:

- Ordinary Least Squares (OLS) method for fitting linear models
- How to find optimal values for $\beta_0$ and $\beta_1$
- Evaluating model fit with R-squared ($R^2$)
- Analyzing residuals to diagnose model issues
- Detecting patterns, outliers, and heteroscedasticity

# Finding the Best-Fitting Line

## The Ordinary Least Squares (OLS) Method

- Systematically determines optimal values for $\beta_0$ and $\beta_1$
- Minimizes the **sum of squared errors** between:
  - Actual values (y)
  - Predicted values ($\hat{y} = \beta_0 + \beta_1 x$)

# The Math Behind Least Squares

## Minimizing the Sum of Squared Errors (SSE):

$$\sum_{i=1}^{n}(y_i - \hat{y_i})^2 = \sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2$$

# The Least Squares Solution

## The formulas for the optimal coefficients:

$$\beta = (\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{y}$$

In simple linear regression, this gives us:

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{Cov(X, Y)}{Var(X)}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

Where $\bar{x}$ and $\bar{y}$ are the means of $x$ and $y$ values respectively.

# Evaluating Model Fit: R-squared ($R^2$)

## $R^2$ measures the proportion of variance explained by the model

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}}$$

Where:

- $SS_{residual} = \sum_{i=1}^{n} (y_i - \hat{y_i})^2$ (unexplained variance)
- $SS_{total} = \sum_{i=1}^{n} (y_i - \bar{y})^2$ (total variance)

Values range from:

- $R^2 = 1$: Perfect fit (all variance explained)
- $R^2 = 0$: Model no better than mean prediction

# Example: $R^2$ in Context

- **House Price Example**
- $R^2 \approx 0.61$ (61%)
- Interpretation:
  - 61% of the variance in house prices is explained by house size
  - 39% is due to other factors:
    - Location (neighborhood, school district)
    - House features (bedrooms, bathrooms, age)
    - Market conditions
    - Property condition
    - Local amenities

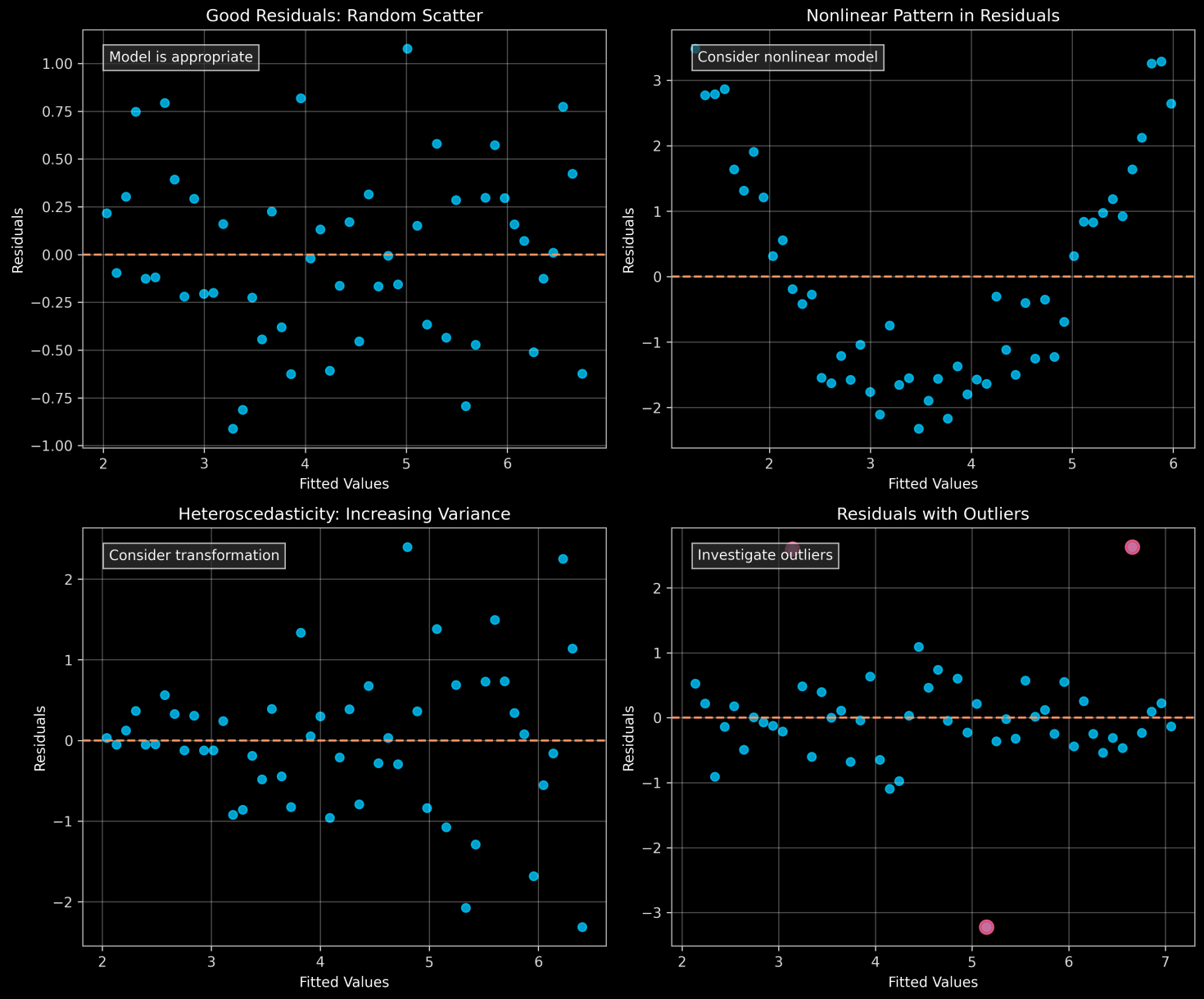
# Residual Analysis

## Residuals = Actual - Predicted

$$\text{Residual}_i = y_i - \hat{y_i}$$

- **Residual Plot**: Plot residuals vs. fitted values or vs. x
- **What to look for**:
  - Randomness: Indicates a good linear fit
  - Patterns: Suggest model inadequacies
  - Outliers: Points with unusually large residuals
  - Changing variance: Indicates heteroscedasticity

# Residual Plot Examples



# What We've Covered

In this video, we've learned:

- How the Ordinary Least Squares method fits the best line to data
- The mathematical foundation: minimizing sum of squared errors
- How to evaluate models using R-squared (variance explained)
- The importance of residual analysis for model diagnostics
- Common patterns in residuals that indicate model issues