Linear and Multiple Regression

A Complete Introduction to Regression Analysis

Joerg R. Osterrieder
www.joergosterrieder.com

December 20, 2025

Course Overview

What We'll Learn:

- Understanding relationships between variables
- Predicting outcomes from data
- Building and evaluating models
- Checking assumptions
- Real-world applications

Prerequisites:

- Basic statistics (mean, variance)
- Elementary calculus
- Matrix algebra (helpful)

Course Structure:

1. Foundations
2. Simple Linear Regression
3. Multiple Linear Regression
4. Diagnostics
5. Applications

Learning Outcomes:

- Build regression models
- Interpret results
- Diagnose problems
- Make predictions

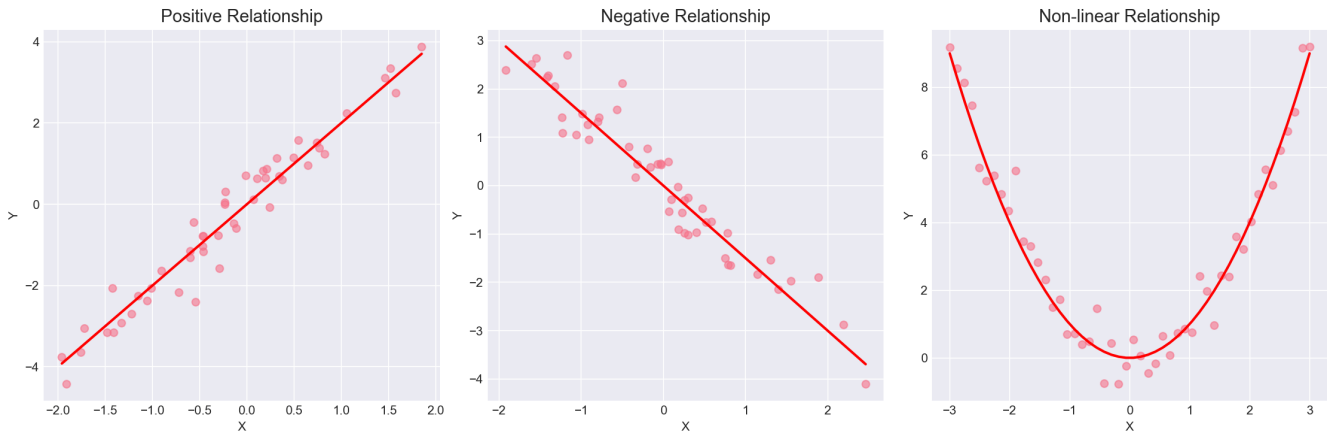
Foundation for understanding the methods covered in this section

What is Regression?

Definition:
Statistical method to model relationships between variables

Key Questions:

- How do variables relate?
- Can we predict outcomes?
- Which factors matter?
- How confident are we?

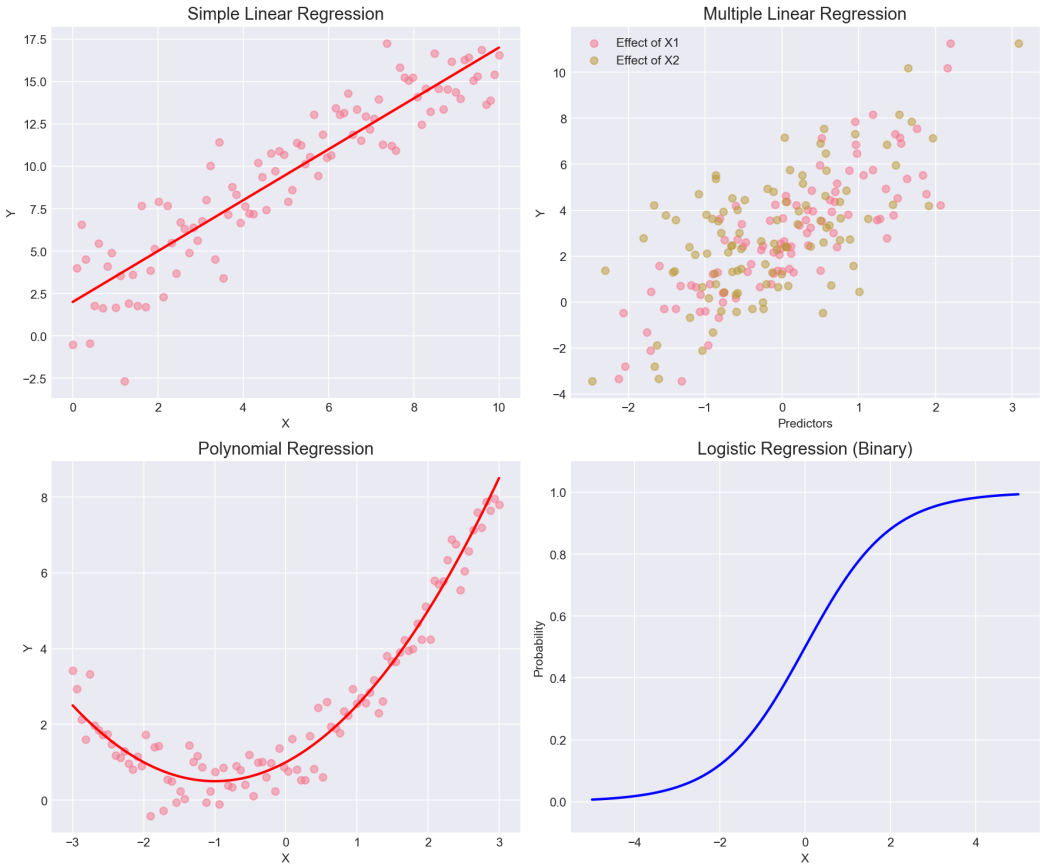


Real Examples:

- Predicting house prices from size, location, age
- Estimating exam scores from study hours
- Forecasting sales from advertising spend

Key insight from What is Regression?

Types of Regression Problems



By Predictors:
Simple (1) vs Multiple (many)
By Response:
Linear, Logistic, Poisson

Choose based on your data type

Basic Concepts

Variables:

- Dependent (Y): What we predict
- Independent (X): What we use to predict

Relationships:

- Linear vs Non-linear
- Positive vs Negative
- Strong vs Weak

Example: Student Performance

- Y = Exam Score
- X_1 = Study Hours
- X_2 = Previous GPA
- X_3 = Class Attendance

Goal: Find equation:

$$\text{Score} = f(\text{Hours, GPA, Attendance})$$

Key Insight: We want to find the best function f() that relates inputs to output

---

Key insight from Basic Concepts

Mathematical Notation

Single Observation:

- (x_i, y_i): One data point
- $i = 1, 2, ..., n$: Index
- n: Sample size

Population vs Sample:

- β: True parameter
- $\hat{\beta}$: Estimate
- Y: True value
- $\hat{Y}$: Predicted value

Summation Notation:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + ... + x_n$$

Mean:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Variance:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Key insight from Mathematical Notation

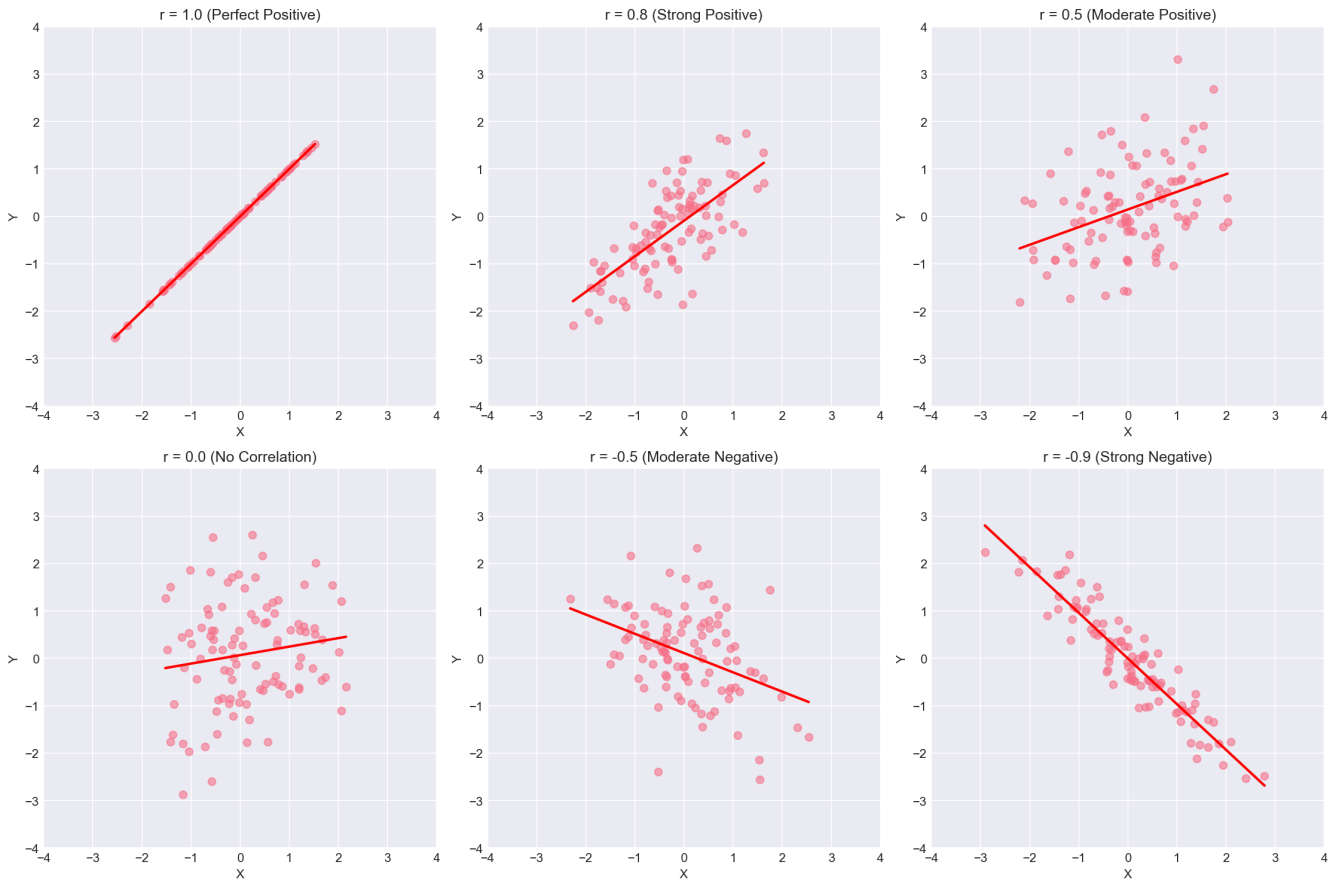
Covariance:

$$\text{Cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})($$

Correlation:

$$r = \frac{\text{Cov}(X, Y)}{s_x \cdot s_y}$$

Range: $-1 \leq r \leq 1$

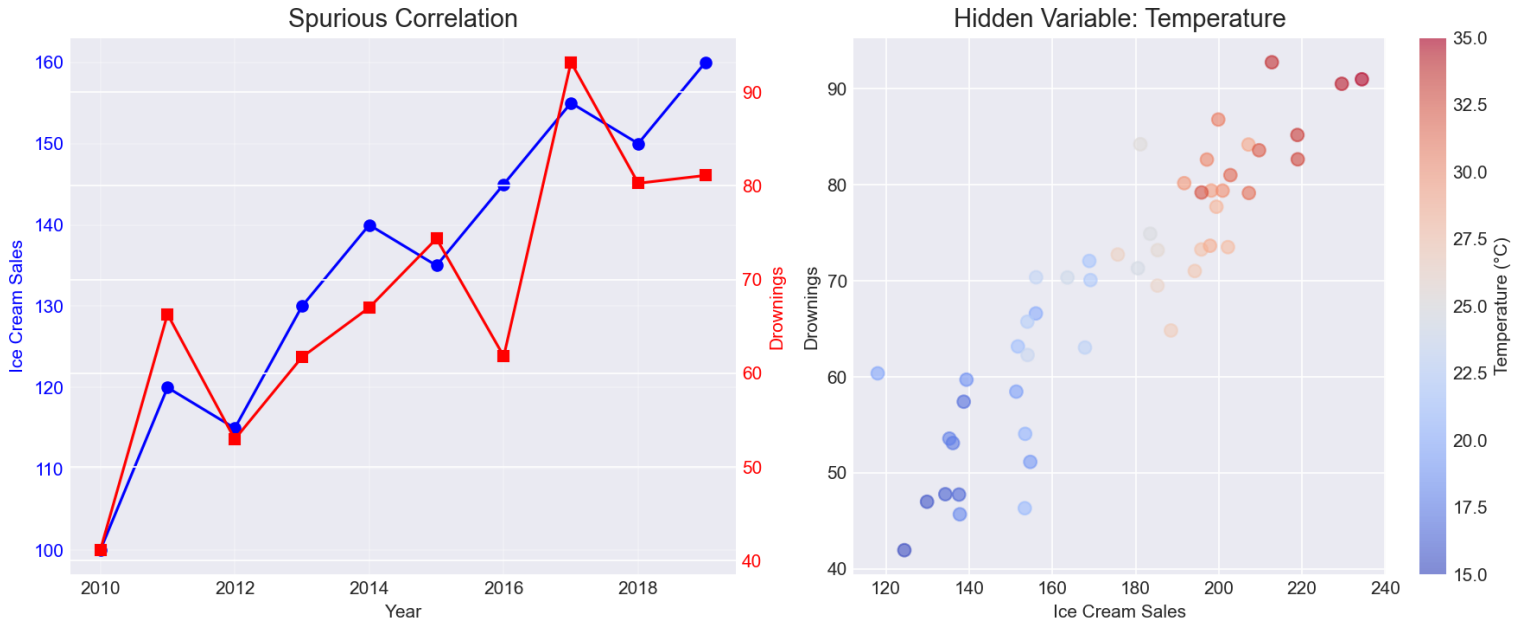


Interpretation:

- $r = 1$: Perfect positive linear relationship
- $r = -1$: Perfect negative linear relationship
- $r = 0$: No linear relationship (may still have non-linear!)

Key insight from Statistical Foundations

Correlation vs Causation



Important Distinction:

- Correlation: Variables move together
- Causation: One variable causes change in another

Classic Examples:

- Ice cream sales correlate with drownings (both increase in summer)
- Shoe size correlates with reading ability (both increase with age)

Remember: Regression finds associations, not necessarily causal relationships!

Key insight from Correlation vs Causation

Simple Linear Regression Model

The Model:

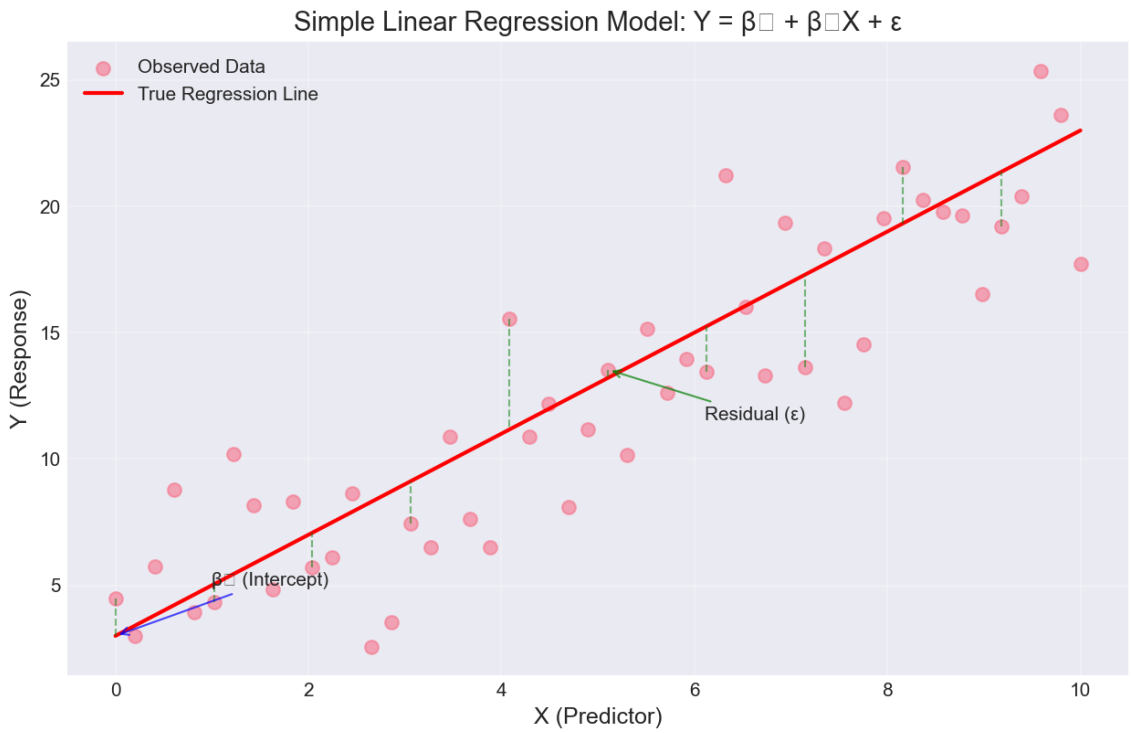
$$Y = \beta_0 + \beta_1 X + \epsilon$$

Components:

- Y: Response variable
- X: Predictor variable
- β_0: Intercept
- β_1: Slope
- ϵ: Error term

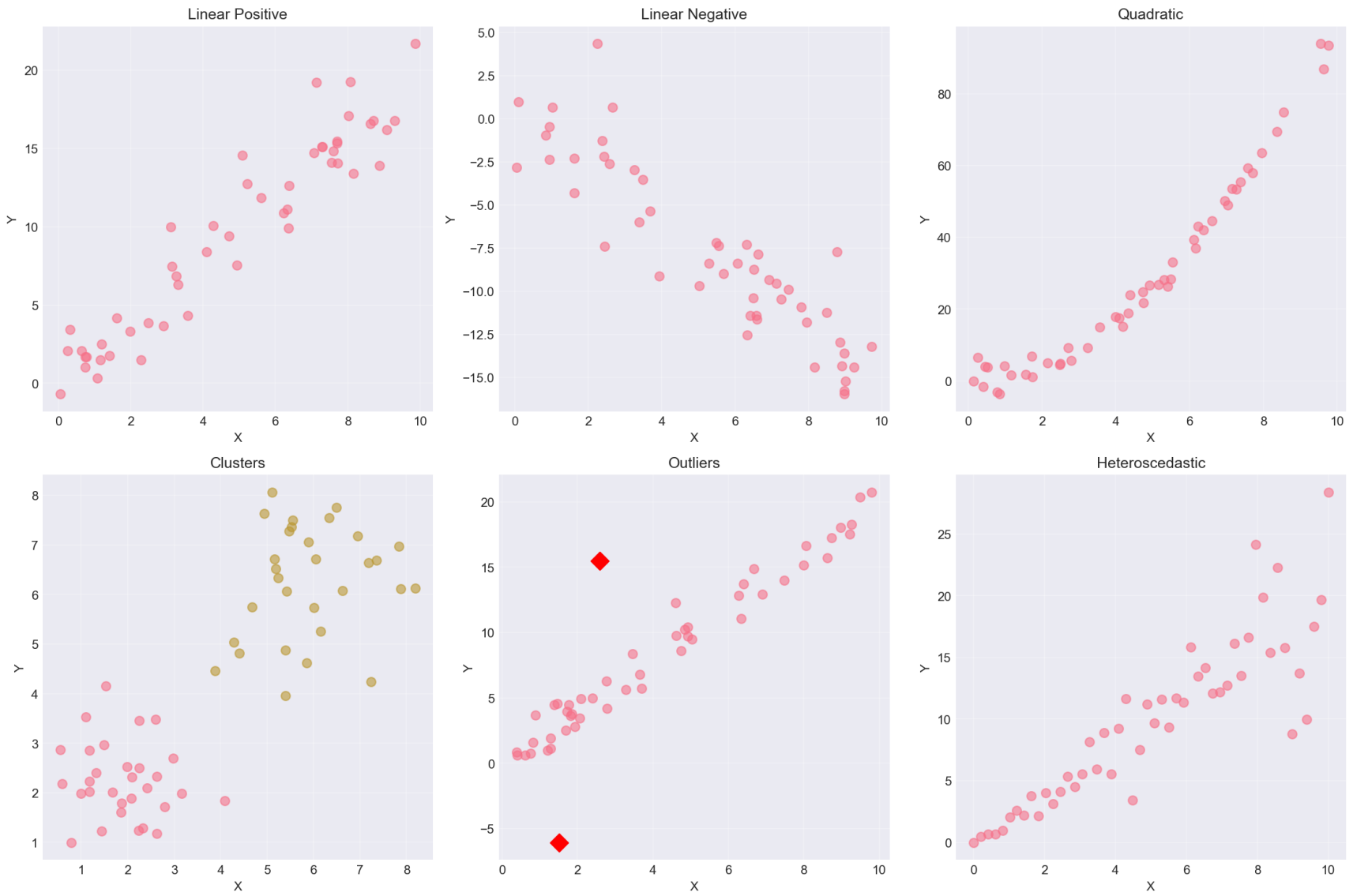
Assumptions:

- $E(\epsilon) = 0$
- $\text{Var}(\epsilon) = \sigma^2$
- $\epsilon \sim N(0, \sigma^2)$



Key insight from Simple Linear Regression Model

Scatter Plots and Visual Intuition



What to Look For:

- Direction (positive/negative)
- Clusters

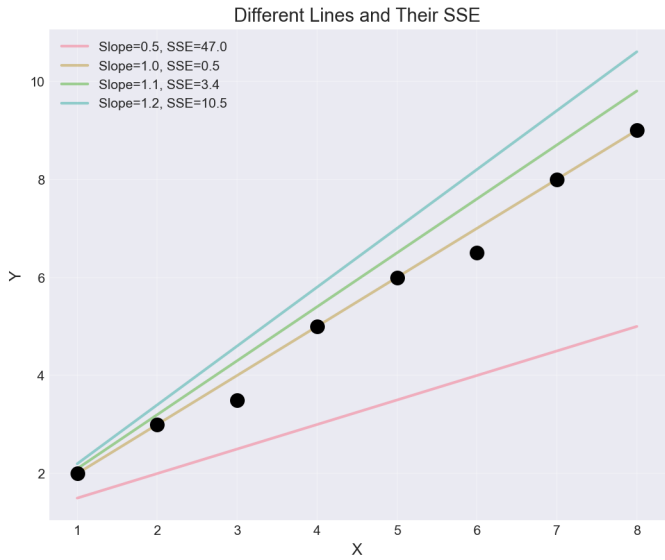
Least Squares Principle

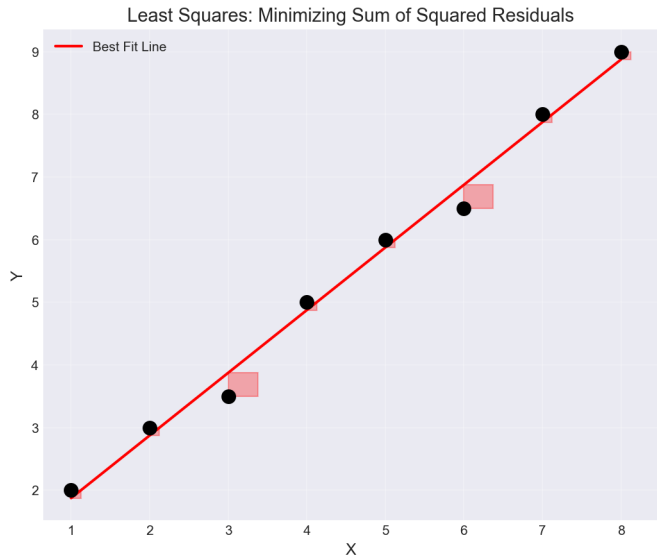
Goal: Find the "best" line
Criterion: Minimize

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Why Squared?

- Penalizes large errors
- Differentiable
- Unique solution
- Nice properties





Residual: $e_i = y_i - \hat{y}_i$ (observed - predicted)

Key insight from Least Squares Principle

Normal Equations

Minimize: $SSE = \sum(y_i - \beta_0 - \beta_1 x_i)^2$
Solutions:

$$\hat{\beta}_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Key insight: Slope = covariance/variance, intercept adjusts for means.

Minimize squared errors to find best fit

Calculating Regression Coefficients - Example

Data: Study Hours vs Score

Hours (X)	Score (Y)
1	50
2	55
3	65
4	70
5	80

Calculations:

- $\bar{x} = 3$, $\bar{y} = 64$
- $\sum(x_i - \bar{x})^2 = 10$
- $\sum(x_i - \bar{x})(y_i - \bar{y}) = 75$

Finding Coefficients:

$$\hat{\beta}_1 = \frac{75}{10} = 7.5$$

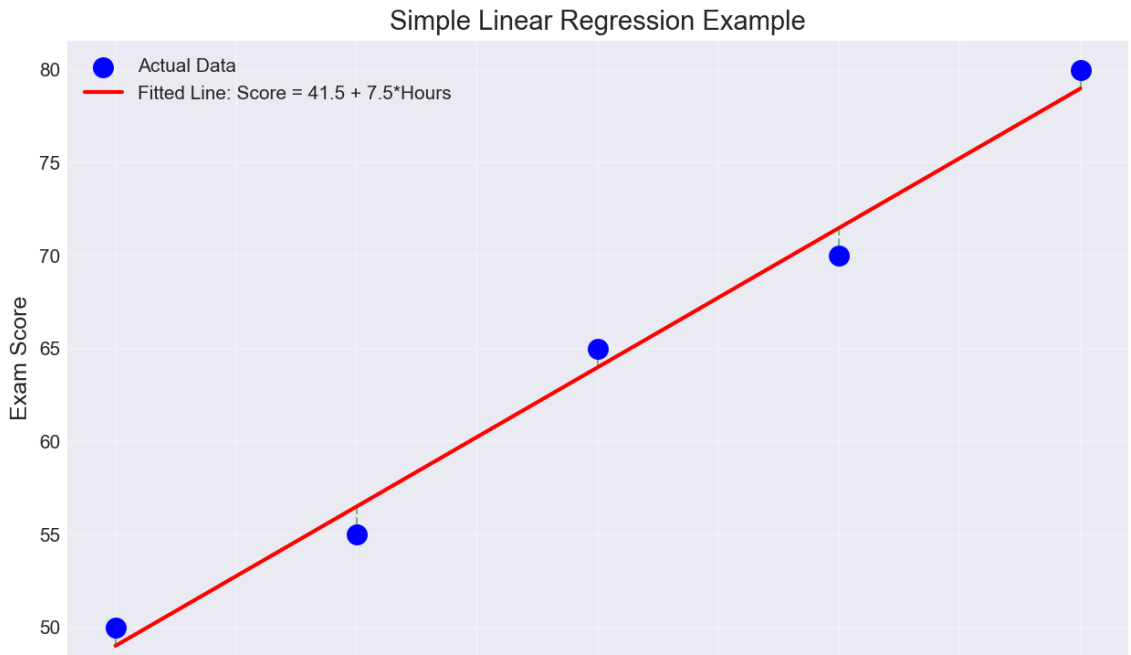
$$\hat{\beta}_0 = 64 - 7.5(3) = 41.5$$

Final Equation:

$$\hat{Y} = 41.5 + 7.5X$$

Interpretation:

- Each hour of study increases score by 7.5 points
- Base score (0 hours) is 41.5



Interpretation of Slope and Intercept

Slope (β_1):

- Change in Y for 1-unit change in X
- Direction of relationship
- Magnitude of effect

Examples:

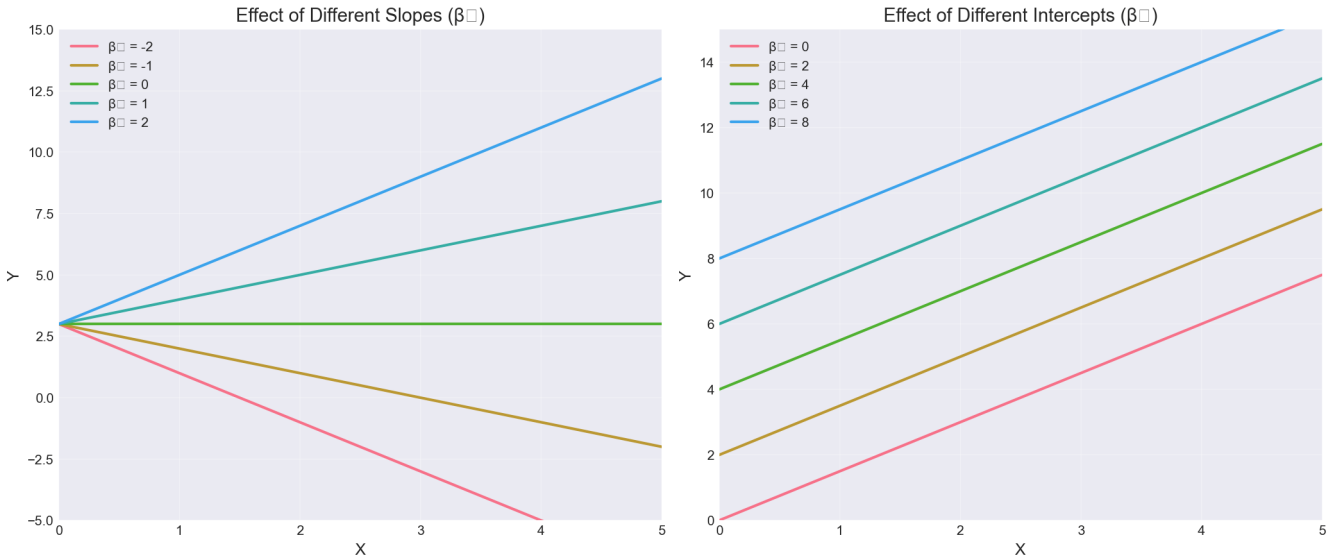
- $\beta_1 = 2.5$: Y increases by 2.5 for each unit of X
- $\beta_1 = -1.8$: Y decreases by 1.8 for each unit of X

Intercept (β_0):

- Value of Y when $X = 0$
- May not be meaningful
- Shifts line up/down

Caution:

- Extrapolation danger
- $X = 0$ may be outside data range
- Sometimes just mathematical artifact



Always interpret results in context

Residuals and Their Properties

Definition:

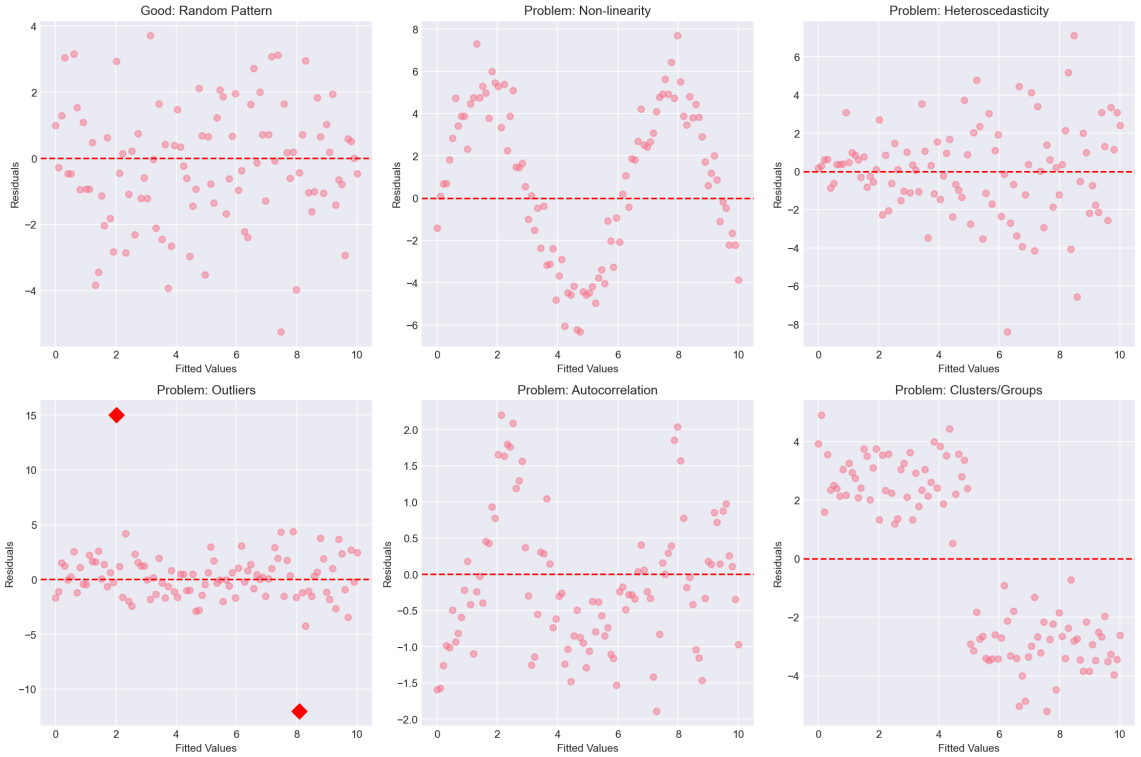
$$e_i = y_i - \hat{y}_i$$

Properties:

- $\sum e_i = 0$
- $\sum x_i e_i = 0$
- $\bar{e} = 0$
- Estimate true errors ϵ_i

Uses:

- Check assumptions
- Identify outliers
- Assess model fit



Good Residuals: Random scatter around zero

Bad Residuals: Patterns indicate model problems

Residual patterns reveal model adequacy

R-Squared and Model Fit

Total Variation:

$$SST = \sum (y_i - \bar{y})^2$$

Explained Variation:

$$SSR = \sum (\hat{y}_i - \bar{y})^2$$

Unexplained:

$$SSE = \sum (y_i - \hat{y}_i)^2$$

Relationship:

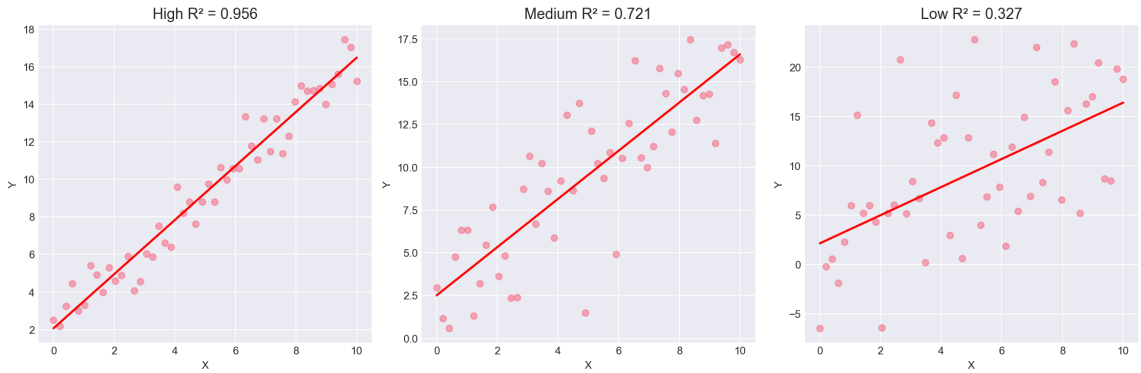
$$SST = SSR + SSE$$

Coefficient of Determination:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Interpretation:

- Proportion of variance explained
- Range: 0 to 1
- Higher is better (usually)



Example: $R^2 = 0.85$ means 85% of variation in Y is explained by X

R-squared alone does not indicate a good model

Standard Errors and Confidence Intervals

Standard Error of Regression:

$$s = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum e_i^2}{n-2}}$$

Standard Error of Slope:

$$SE(\hat{\beta}_1) = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}}$$

Standard Error of Intercept:

$$SE(\hat{\beta}_0) = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}$$

95% Confidence Intervals:
For slope:

$$\hat{\beta}_1 \pm t_{0.025,n-2} \cdot SE(\hat{\beta}_1)$$

For intercept:

$$\hat{\beta}_0 \pm t_{0.025,n-2} \cdot SE(\hat{\beta}_0)$$

Interpretation: We are 95% confident the true parameter lies in this interval

Confidence intervals provide more information than p-values alone

Hypothesis Testing for Coefficients

Testing Slope = 0:
$H_0 : \beta_1 = 0$ (no relationship)
$H_a : \beta_1 \neq 0$ (relationship exists)
Test Statistic:

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

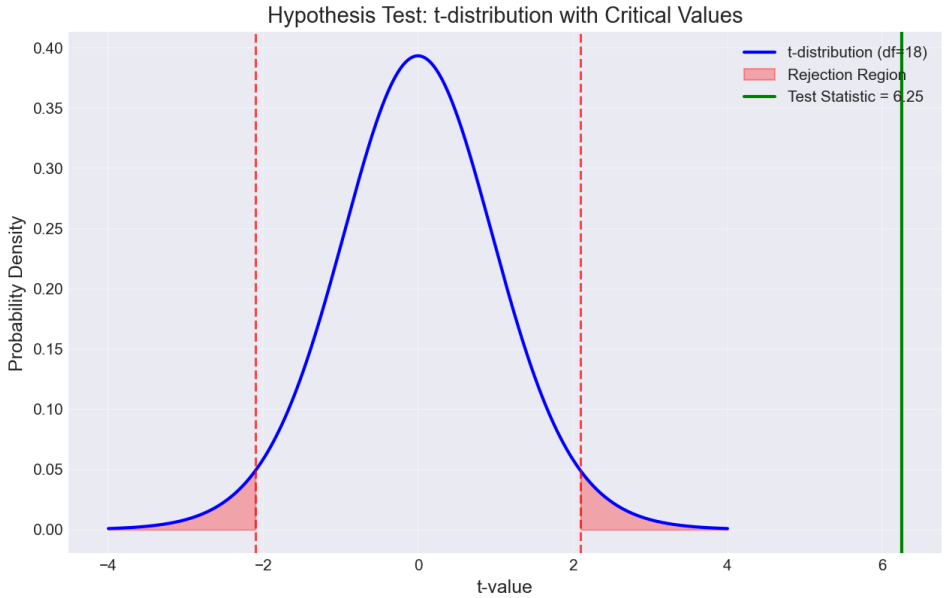
Decision Rule:

- If $|t| > t_{critical}$, reject H_0
- If p-value $< \alpha$, reject H_0

Example:

- $\hat{\beta}_1 = 7.5$
- $SE(\hat{\beta}_1) = 1.2$
- $t = 7.5/1.2 = 6.25$
- $t_{0.025,18} = 2.101$
- Since $6.25 > 2.101$, reject H_0

Conclusion: Significant relationship between X and Y



Formulate hypotheses clearly before collecting data

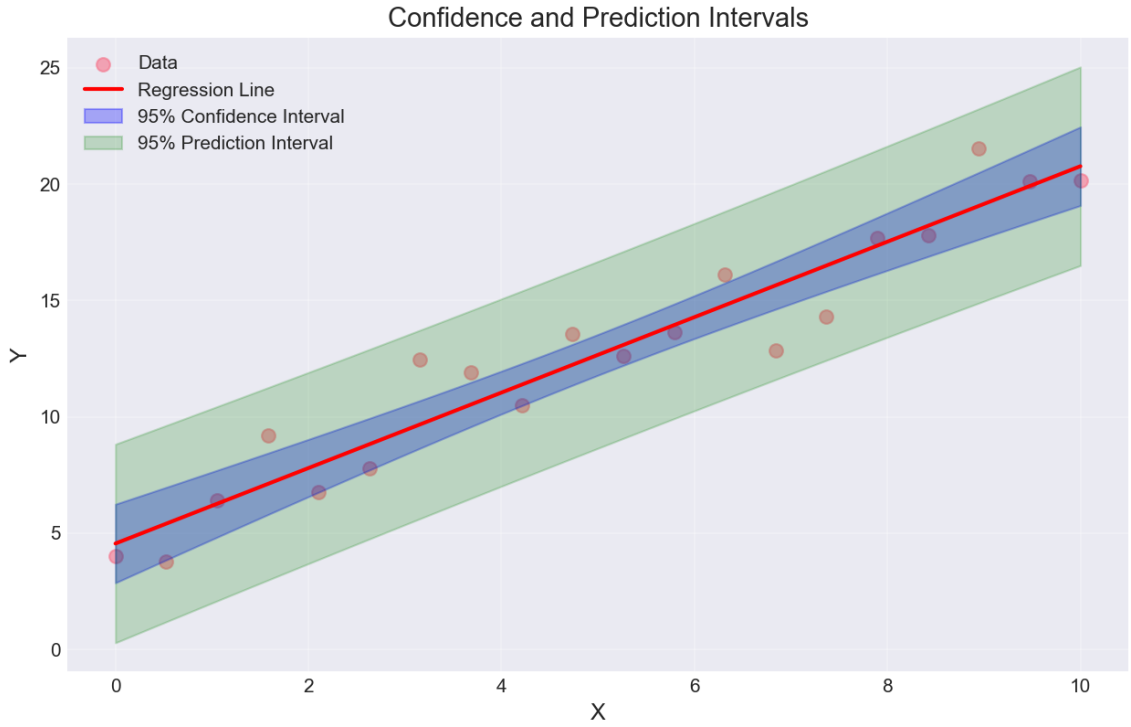
Prediction Intervals

Point Prediction:

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

Two Types of Intervals:
1. **Confidence Interval** for mean response $E(Y|X = x_0)$
2. **Prediction Interval** for individual response Y_0

PI is always wider than CI!



Formulas:

$$PI : \hat{y}_0 \pm t \cdot s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

$$CI : \hat{y}_0 \pm t \cdot s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

Key insight from Prediction Intervals

Complete Worked Example - Housing Prices

Data: House Size vs Price

Size (100 sq ft)	Price ($1000s)
12	155
15	180
18	210
20	235
22	250
25	280
28	310

Analysis Steps:

1. Plot data
2. Calculate coefficients
3. Test significance
4. Check assumptions
5. Make predictions

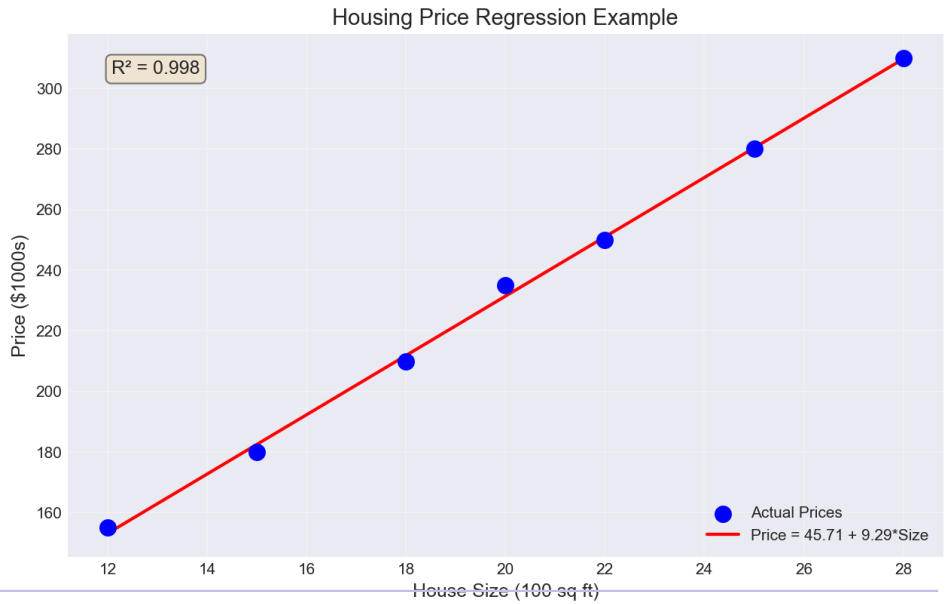
Results:

- $\hat{\beta}_0 = 45.71$
- $\hat{\beta}_1 = 9.29$
- $R^2 = 0.987$
- p-value < 0.001

Equation:

$$\text{Price} = 45.71 + 9.29 \times \text{Size}$$

Interpretation: Each 100 sq ft increases price by $9,290



This pattern applies to similar real-world scenarios

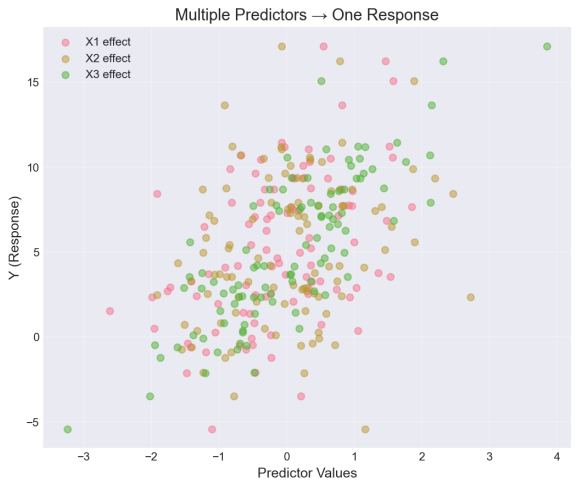
Extension to Multiple Predictors

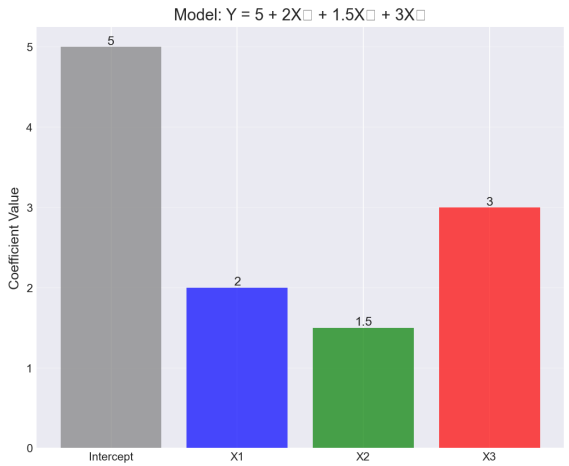
Why Multiple Regression?

- Real phenomena have multiple causes
- Better predictions
- Control for confounders
- Understand relative importance

The Model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon$$





Example: Salary Prediction

- Y: Salary
- X_1: Years of experience
- X_2: Education level
- X_3: Performance rating

Consider multicollinearity when interpreting predictors

Matrix Notation for MLR

Matrix Form:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Where:

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Least Squares Solution:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

Requirements:

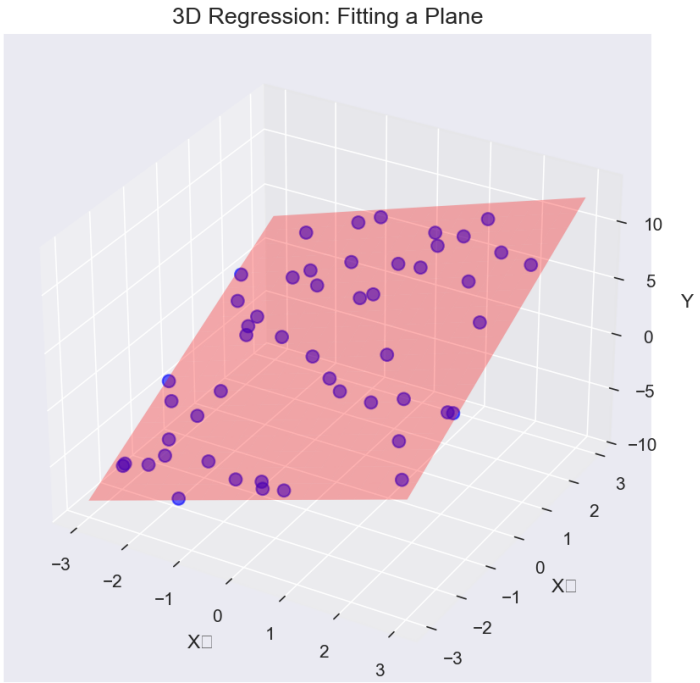
- $\mathbf{X}^T\mathbf{X}$ must be invertible
- No perfect multicollinearity
- $n > p$ (more obs than parameters)

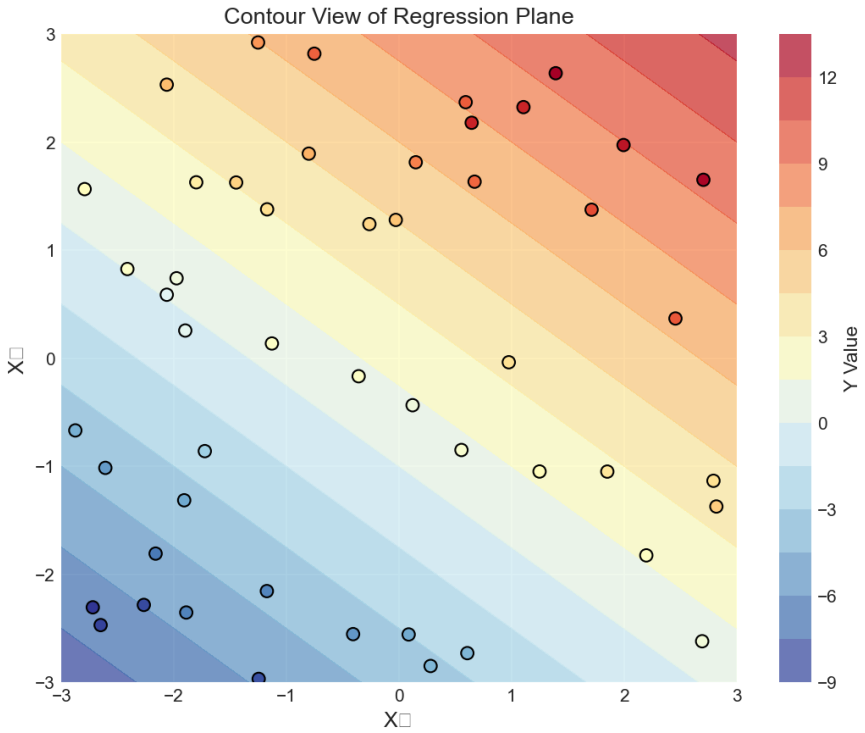
Predicted Values:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

Key insight from Matrix Notation for MLR

Geometric Interpretation





Key Concepts:

- With 2 predictors: fitting a plane in 3D space
- With p predictors: fitting a hyperplane in (p+1)-dimensional space
- Projection of Y onto column space of X
- Residuals are perpendicular to fitted surface

Always interpret results in context

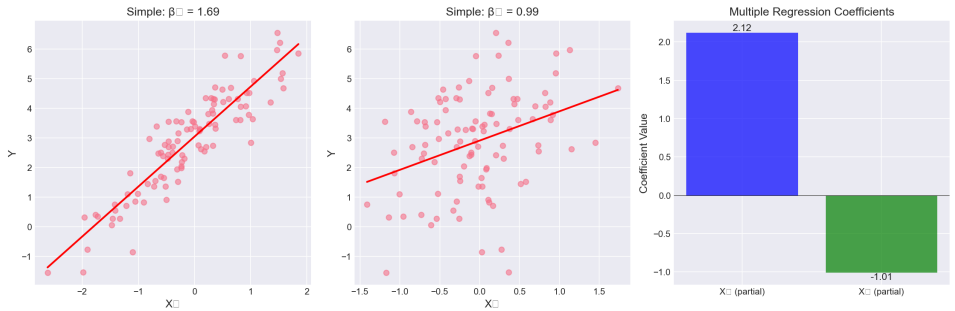
Partial Regression Coefficients

Interpretation of β_j:
Change in Y for 1-unit change in X_j, *holding all other variables constant*

Key Difference from Simple Regression:

- Controls for other variables
- "Partial" effect
- May differ from simple coefficient

Example: Ice cream sales vs temperature and advertising



Simpson's Paradox: Relationship can reverse when controlling for other variables!

Interpret coefficients in the context of your variables

Adjusted R-Squared

Problem with R^2:

- Always increases with more predictors
- Can overfit with many variables
- Not good for model comparison

Solution - Adjusted R^2:

$$R^2_{adj} = 1 - \frac{(1 - R^2)(n - 1)}{n - p - 1}$$

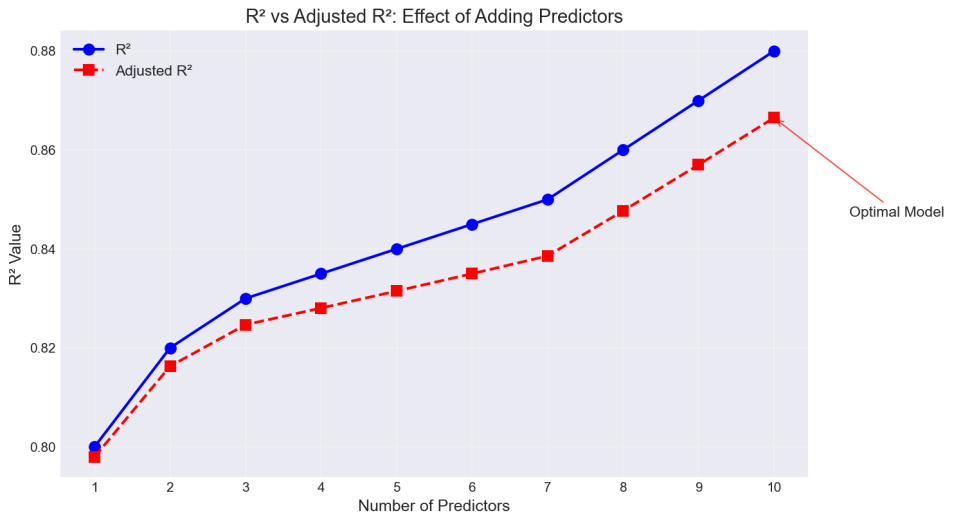
Properties:

- Penalizes for number of predictors
- Can decrease if predictor not useful
- Better for model comparison

Example Comparison:

Model	R^2	R^2_{adj}
1 predictor	0.80	0.79
2 predictors	0.82	0.80
3 predictors	0.83	0.80
10 predictors	0.88	0.75

Conclusion: Model with 2-3 predictors is best!



R-squared alone does not indicate a good model

F-Test for Overall Significance

Hypotheses:
$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$
(No predictors are significant)
H_a : At least one $\beta_j \neq 0$
(At least one predictor matters)
F-Statistic:

$$F = \frac{(SSR/p)}{(SSE/(n-p-1))} = \frac{MSR}{MSE}$$

Distribution: $F \sim F_{p,n-p-1}$ under H_0

ANOVA Table:

Source	SS	df	MS
Regression	SSR	p	MSR
Error	SSE	n-p-1	MSE
Total	SST	n-1	

Relationship to R^2:

$$F = \frac{R^2/p}{(1-R^2)/(n-p-1)}$$

Decision: Reject H_0 if $F > F_{critical}$ or p-value $< \alpha$

Statistical significance differs from practical significance

t-Tests for Individual Predictors

Testing Each Coefficient:
$H_0 : \beta_j = 0$ (given other predictors)
$H_a : \beta_j \neq 0$
Test Statistic:

$$t_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

Standard Error:

$$SE(\hat{\beta}_j) = \sqrt{MSE \cdot C_{jj}}$$

where C_{jj} is diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$

Example Output:

Variable	Coef	SE	t	p-value
Intercept	25.3	3.2	7.91	¡0.001
Age	2.1	0.5	4.20	¡0.001
Income	0.8	0.3	2.67	0.012
Education	1.5	0.9	1.67	0.103

Interpretation:

- Age and Income significant
- Education not significant (p ¿ 0.05)
- Consider removing Education

Choose the appropriate test based on your data characteristics

Multicollinearity Detection

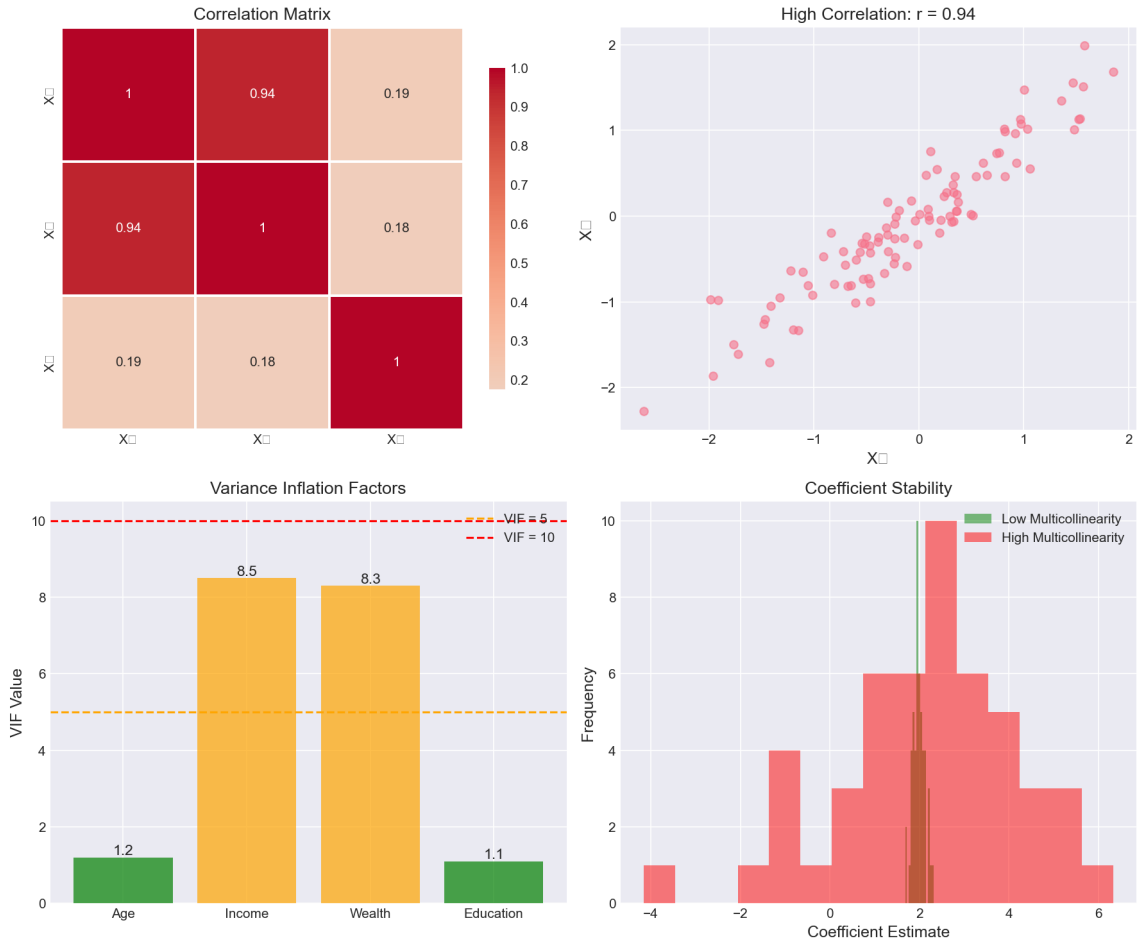
Problem: Predictors highly correlated with each other

Consequences:

- Unstable coefficients
- Large standard errors
- Difficult interpretation
- Computational issues

Detection Methods:

- Correlation matrix
- Variance Inflation Factor (VIF)
- Condition number



VIF Formula:

$$VIF_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is from regressing X_j on other predictors

Guidelines:

- VIF ¡ 5: No problem

Variable Selection Methods

1. **Forward Selection:**
   - Start with no predictors
   - Add best predictor one at a time
   - Stop when no improvement
2. **Backward Elimination:**
   - Start with all predictors
   - Remove worst predictor one at a time
   - Stop when all significant
3. **Stepwise:**
   - Combination of forward/backward
   - Can add and remove variables
4. **Best Subsets:**
   - Try all possible combinations
   - Choose best by criterion (AIC, BIC)
   - Computationally intensive

Selection Criteria:

- R^2_{adj}: Adjusted R-squared
- AIC: $n \ln(SSE/n) + 2p$
- BIC: $n \ln(SSE/n) + p \ln(n)$
- Mallows' C_p

Warning: Automated selection can miss important variables!

Model selection balances fit and complexity

Interaction Terms

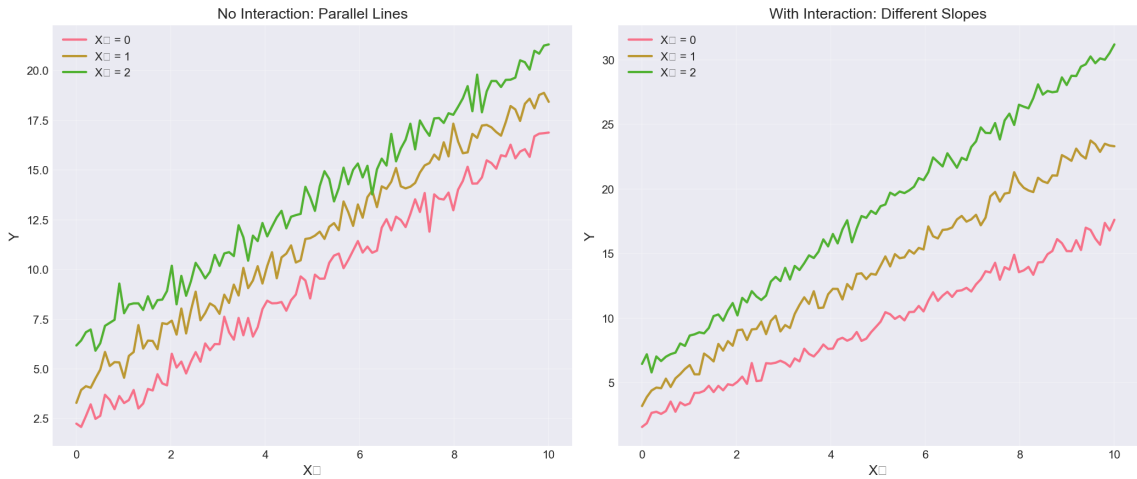
Concept: Effect of one variable depends on another

Model with Interaction:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

Interpretation:

- β_1: Effect of X_1 when $X_2 = 0$
- β_2: Effect of X_2 when $X_1 = 0$
- β_3: How effect changes



Example: Marketing

- X_1: TV advertising
- X_2: Online advertising
- Interaction: Synergy effect
- Combined effect ¿ sum of parts

Key insight from Interaction Terms

Polynomial Regression

When to Use:

- Non-linear relationships
- Curved patterns in residuals
- Theory suggests curves

Models: Quadratic:

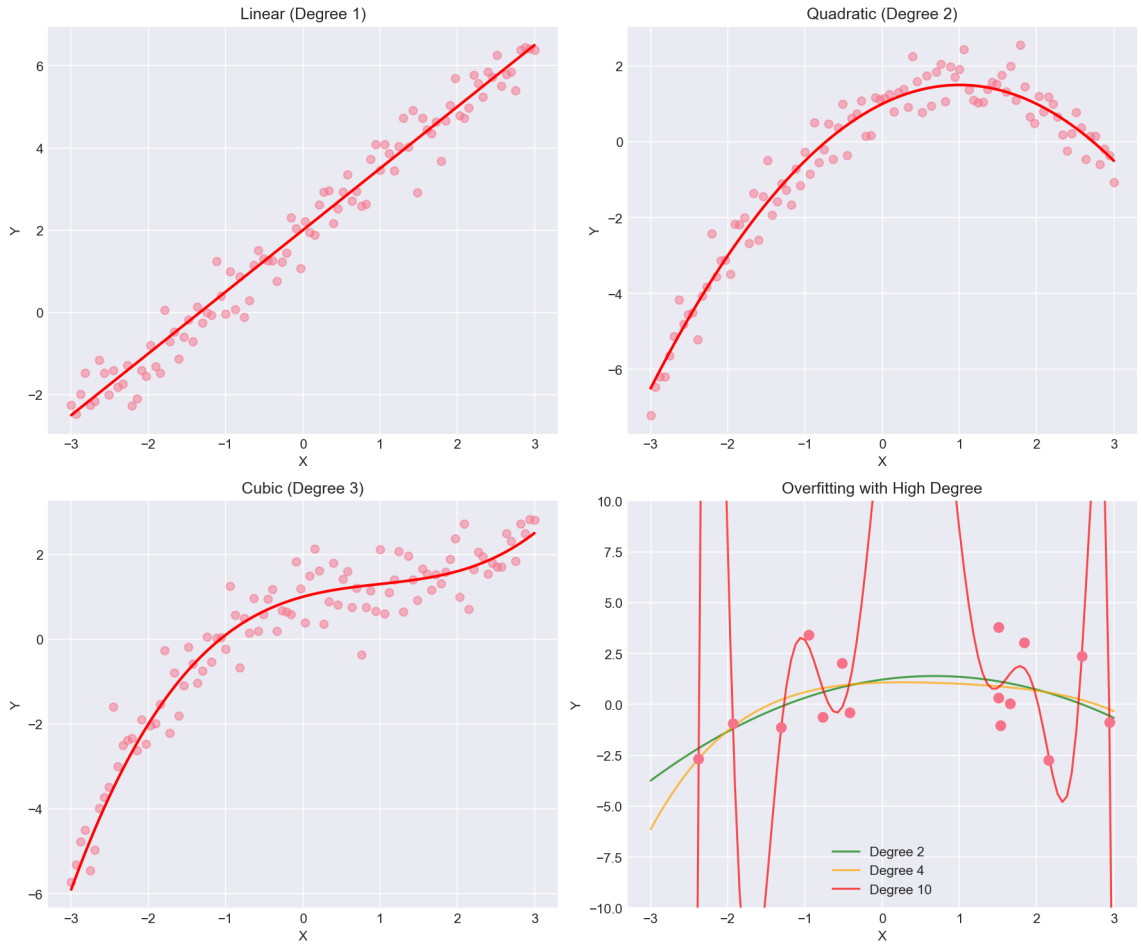
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

Cubic:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$$

Caution:

- Overfitting risk
- Extrapolation dangerous
- Multicollinearity issues



Choosing Degree:

- Start with quadratic
- Test higher order terms
- Use cross-validation
- Consider theory

Key insight from Polynomial Regression

Categorical Predictors (Dummy Variables)

Binary Variable: Gender (M/F) $\rightarrow$ One dummy

$$X = \begin{cases} 1 & \text{if Male} \\ 0 & \text{if Female} \end{cases}$$

Multiple Categories: Region (N/S/E/W) $\rightarrow$ Three dummies

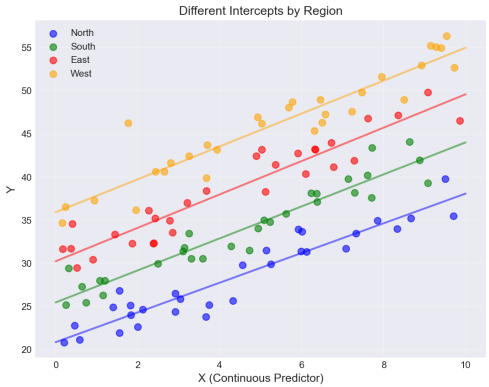
Region	D_1	D_2	D_3
North	0	0	0
South	1	0	0
East	0	1	0
West	0	0	1

Reference category: North

Interpretation:

Model: $Y = \beta_0 + \beta_1 D_{Male} + \beta_2 X_{Age}$

- β_0: Intercept for females
- $\beta_0 + \beta_1$: Intercept for males
- β_1: Gender difference



Dummy Variable Encoding (North = Reference)

Region	D_South	D_East	D_West
North	0	0	0
South	1	0	0
East	0	1	0
West	0	0	1
North	0	0	0
East	0	1	0

Rule: k categories $\rightarrow$ k-1 dummies

Consider multicollinearity when interpreting predictors

Model Building Strategy

Step 1: Understand Problem

- Define objective
- Identify outcome variable
- List potential predictors

Step 2: Explore Data

- Descriptive statistics
- Correlation matrix
- Scatter plots
- Check for outliers

Step 3: Build Initial Model

- Include theoretically important variables
- Check multicollinearity
- Assess significance

Step 4: Refine Model

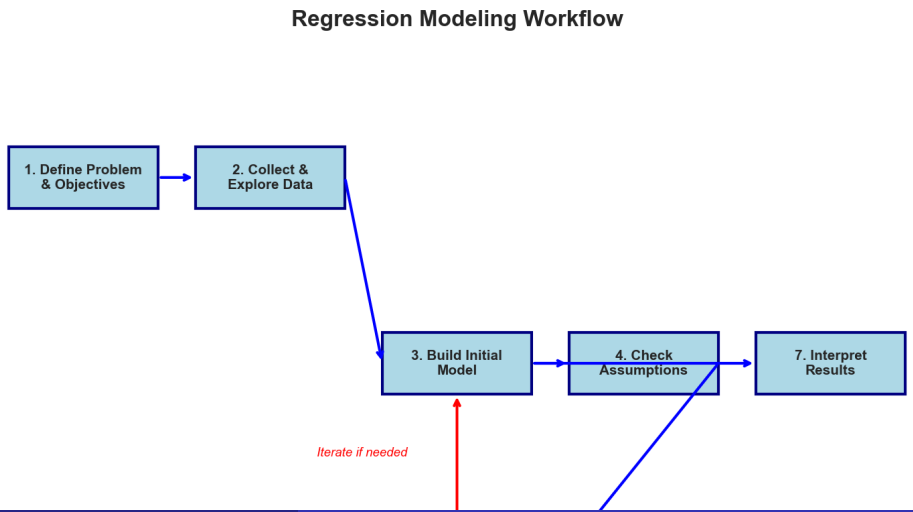
- Variable selection
- Add interactions if needed
- Consider transformations

Step 5: Validate

- Check assumptions
- Examine residuals
- Test on new data

Step 6: Interpret

- Practical significance
- Confidence intervals
- Limitations



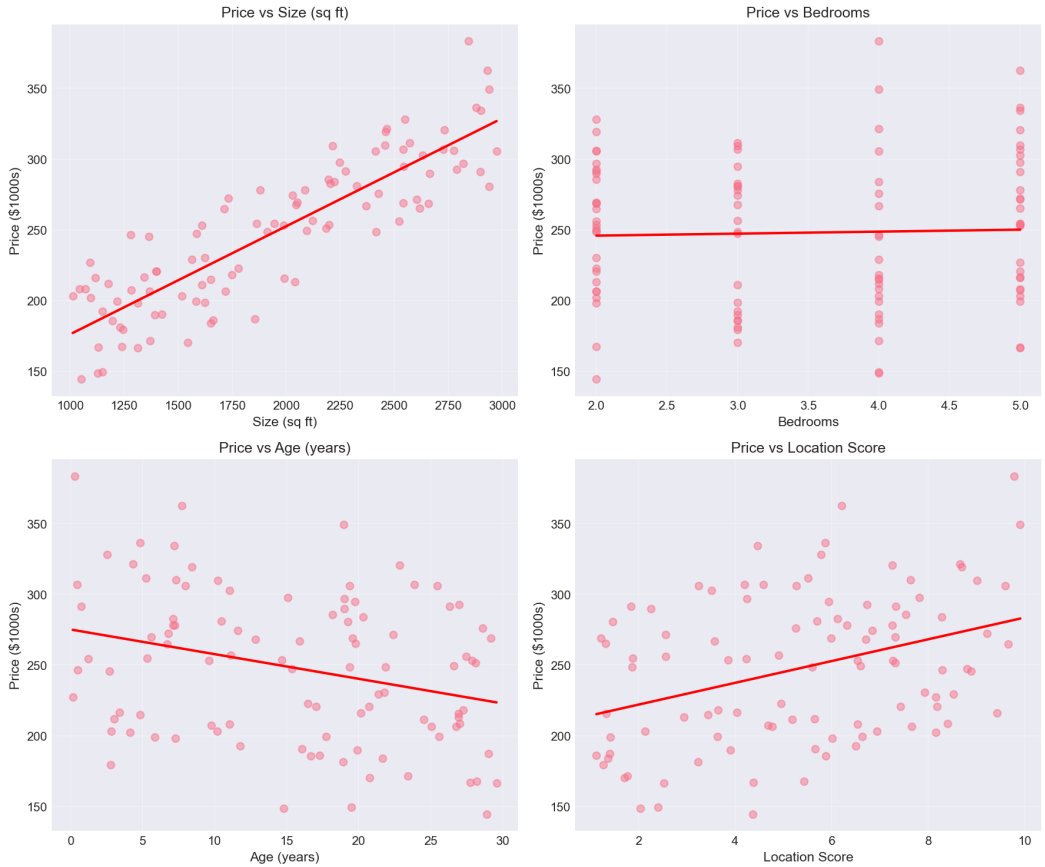
Complete MLR Example - Real Estate

Predicting House Prices:

Variable	Description
Y	Price ($1000s)
X_1	Size (sq ft)
X_2	Bedrooms
X_3	Age (years)
X_4	Garage (Y/N)
X_5	Location score

Model Results:

Var	Coef	SE	t	p
Int	50.2	8.3	6.05	¡0.001
Size	0.08	0.01	8.00	¡0.001
Beds	5.1	2.3	2.22	0.031
Age	-1.2	0.3	-4.00	¡0.001
Garage	15.3	4.1	3.73	¡0.001
Loc	8.7	1.9	4.58	¡0.001

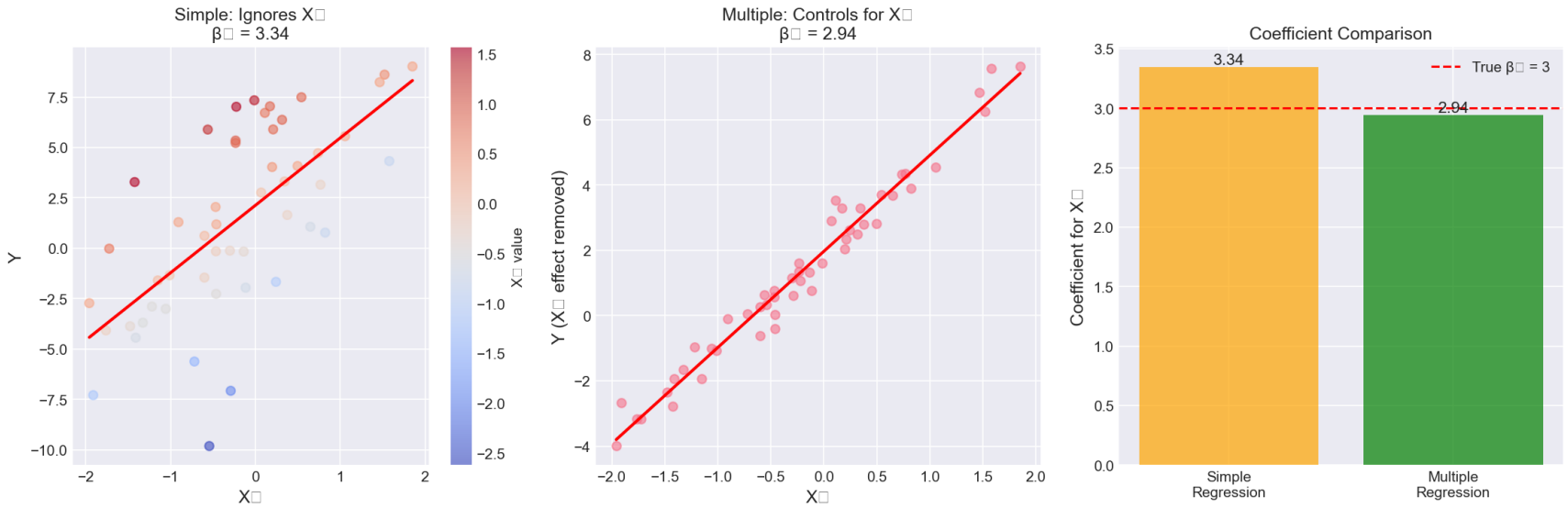


Model Performance:

- $R^2 = 0.892$
- $R^2_{adj} = 0.881$
- F-statistic = 82.4 (p ¡ 0.001)
- RMSE = 12.3

Interpretation: Each sq ft adds $80, each year reduces price by $1,200

Comparison: Simple vs Multiple Regression



Simple Linear Regression:

- One predictor
- Easy interpretation
- May miss important factors
- Lower predictive power

Multiple Linear Regression:

- Multiple predictors
- Controls for confounders
- Better predictions
- More complex interpretation

Key insight from Comparison: Simple vs Multiple Regression

Linear Regression Assumptions

L.I.N.E. Assumptions:

Linearity:

- Relationship is linear
- Check: Residual plots

Independence:

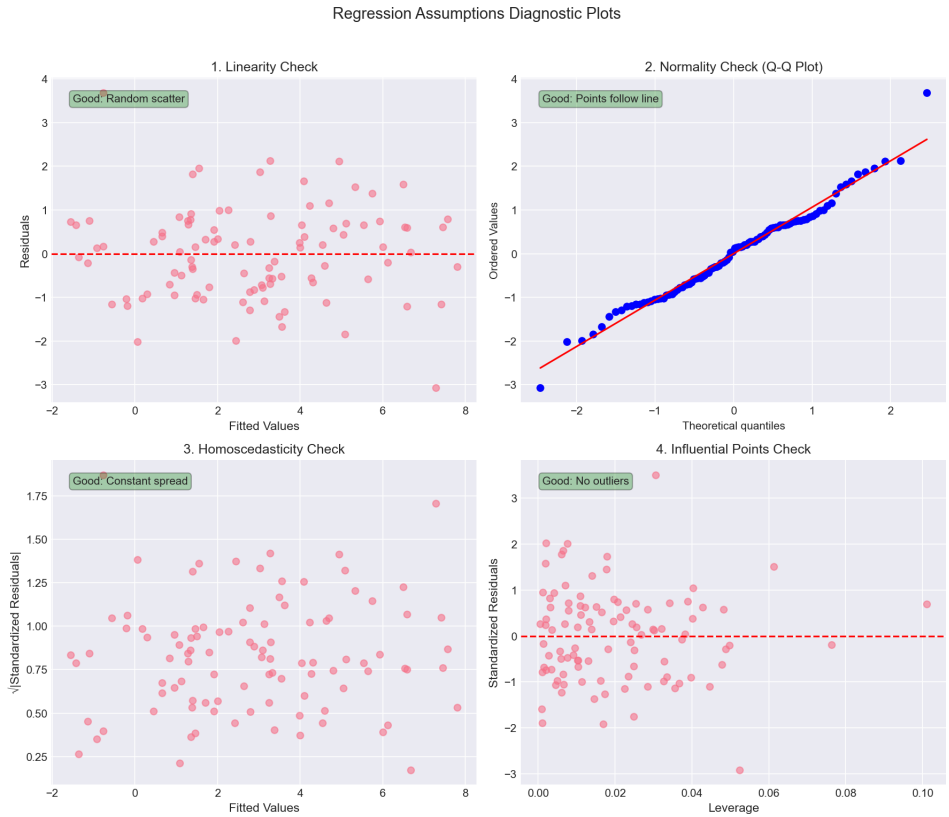
- Observations independent
- Check: Durbin-Watson test

Normality:

- Errors normally distributed
- Check: Q-Q plot

Equal variance:

- Homoscedasticity
- Check: Scale-location plot

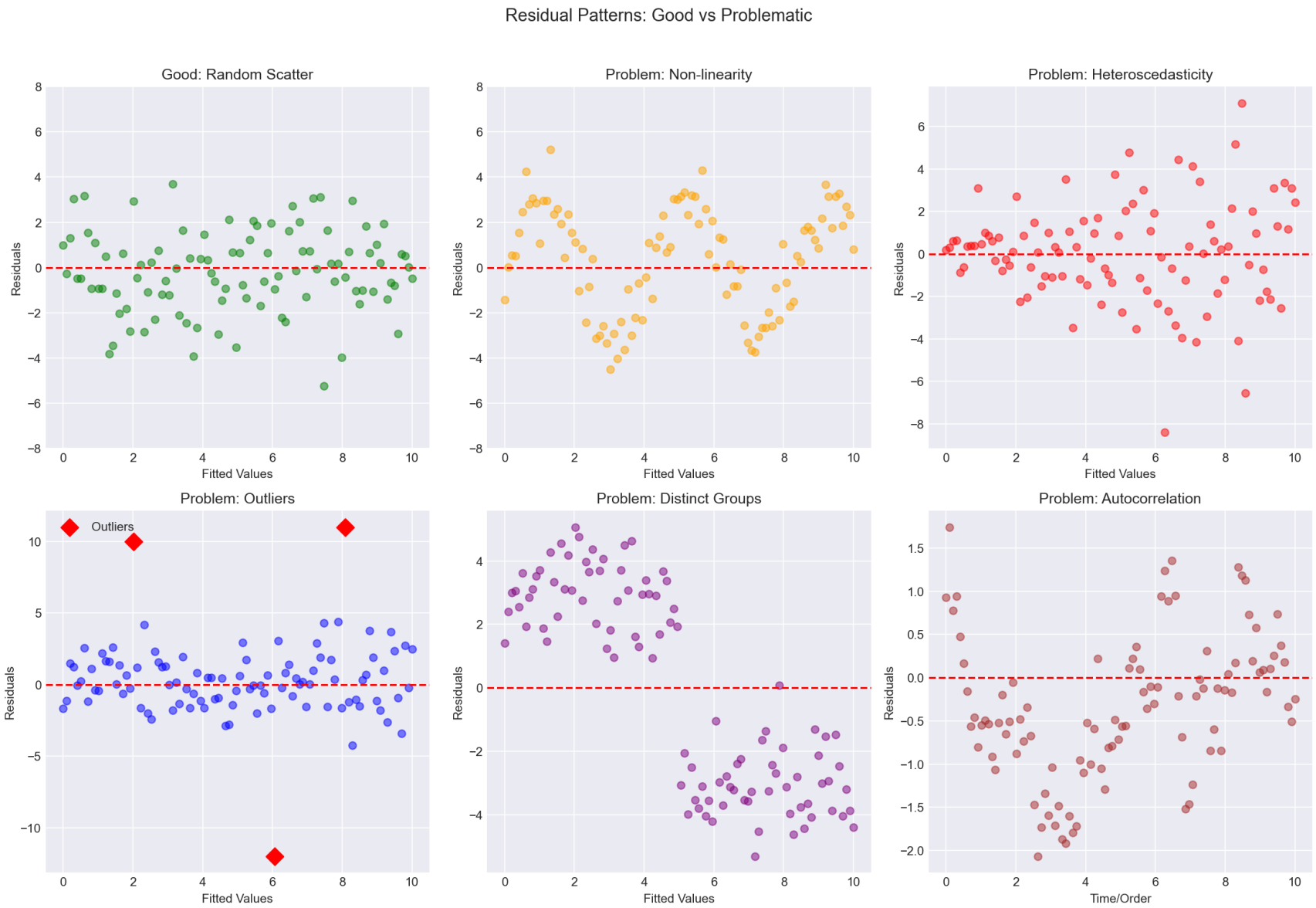


Additional Assumptions:

- No perfect multicollinearity
- More observations than parameters
- No measurement error in X

Always verify assumptions before interpreting results

Residual Plots and Patterns



Good:
Random scatter

Non-linear:
Add polynomial

Heteroscedastic:
Transform Y

Outliers:
Investigate points

Residual patterns reveal model adequacy

Normality Testing (Q-Q Plots)

Q-Q Plot:

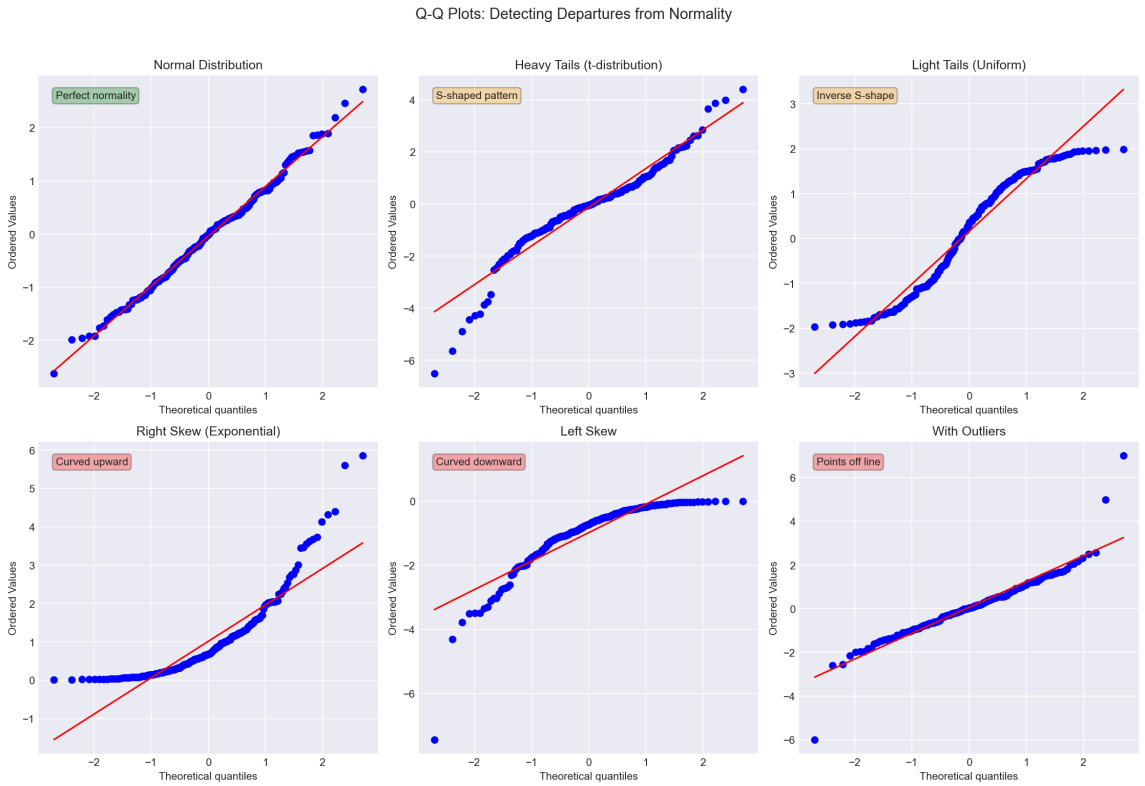
- Quantile-Quantile plot
- Compares to normal distribution
- Points should follow line

Deviations Indicate:

- S-shape: Light tails
- Inverse S: Heavy tails
- Curved: Skewness
- Outliers: Individual points off

Formal Tests:

- Shapiro-Wilk
- Kolmogorov-Smirnov
- Anderson-Darling



Note: Slight deviations acceptable for large samples (CLT)

Choose the appropriate test based on your data characteristics

Homoscedasticity Checking

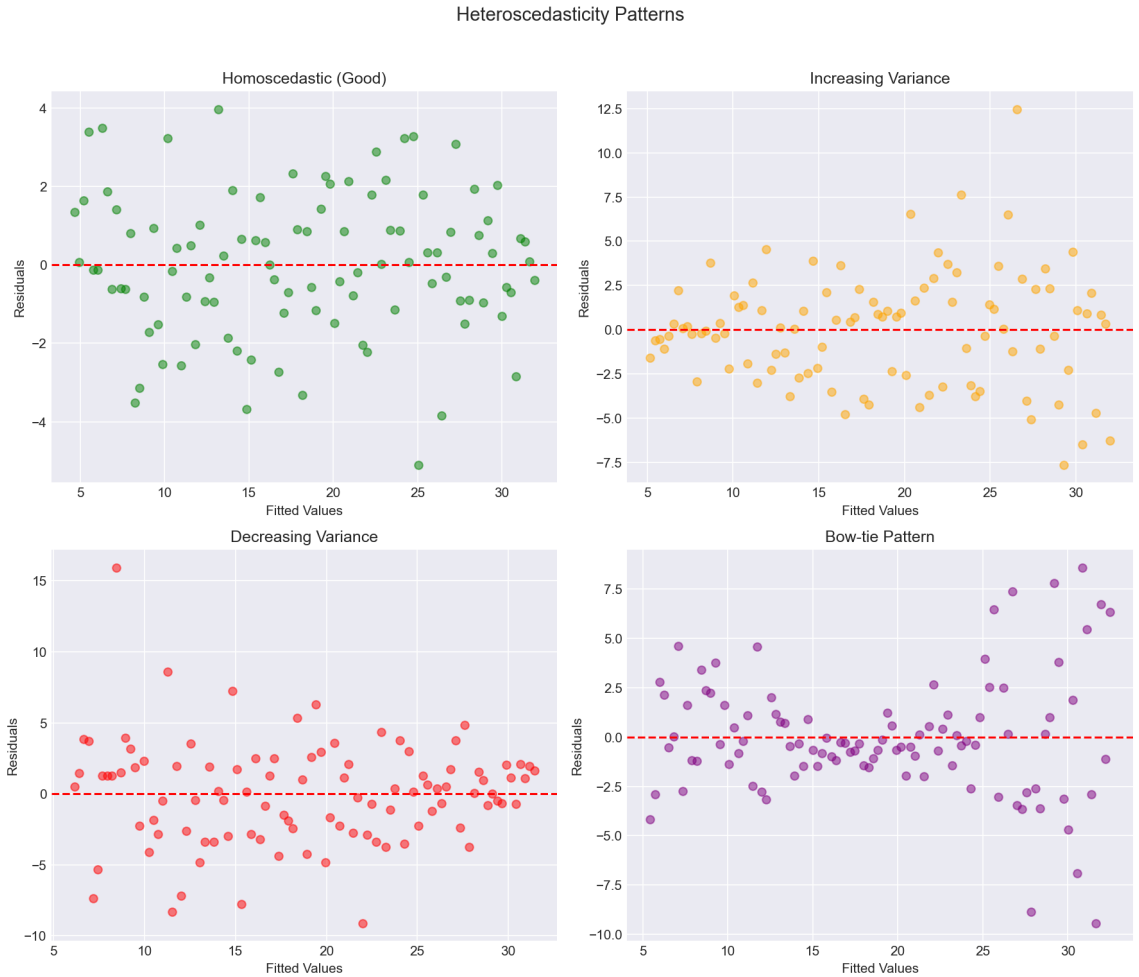
Equal Variance Assumption: Variance of errors constant across all X values

Detection:

- Scale-Location plot
- Residuals vs Fitted
- Breusch-Pagan test
- White test

Consequences of Violation:

- Inefficient estimates
- Wrong standard errors
- Invalid inference



Solutions:

- Transform Y (log, sqrt)
- Weighted least squares
- Robust standard errors

Key insight from Homoscedasticity Checking

Independence Testing

Durbin-Watson Test:

$$DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

Range: 0 to 4

- DW $\approx$ 2: No autocorrelation
- DW ¡ 2: Positive autocorrelation
- DW ¿ 2: Negative autocorrelation

Rule of Thumb:

- 1.5 ¡ DW ¡ 2.5: Acceptable
- Otherwise: Problem

Time Series Context:

Other Contexts:

- Spatial data
- Clustered data
- Repeated measures

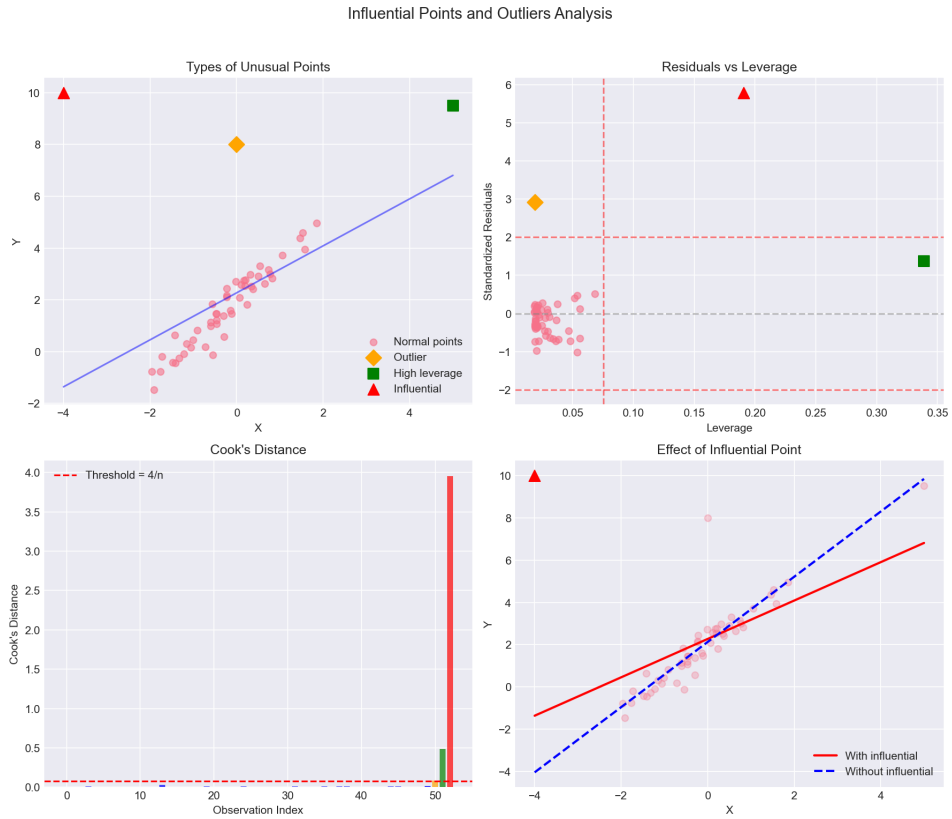
Solutions:

- Add lagged variables
- Use time series methods

Influential Points and Outliers

Types of Unusual Points:

1. **Outlier:**
   - Large residual
   - Far from fitted line
   - Check standardized residuals
2. **High Leverage:**
   - Unusual X values
   - Far from $\bar{X}$
   - Hat values: $h_{ii} > 2p/n$
3. **Influential:**
   - Changes regression line
   - High leverage + outlier
   - Cook's D ¿ 1



Detection Measures:

- Standardized residuals: $|r_i| > 3$
- Leverage: $h_{ii} > 2(p+1)/n$
- Cook's D: $D_i > 4/n$
- DFFITS: $|DFFITS_i| > 2\sqrt{p/n}$

Key insight from Influential Points and Outliers

Leverage and Cook's Distance

Leverage:

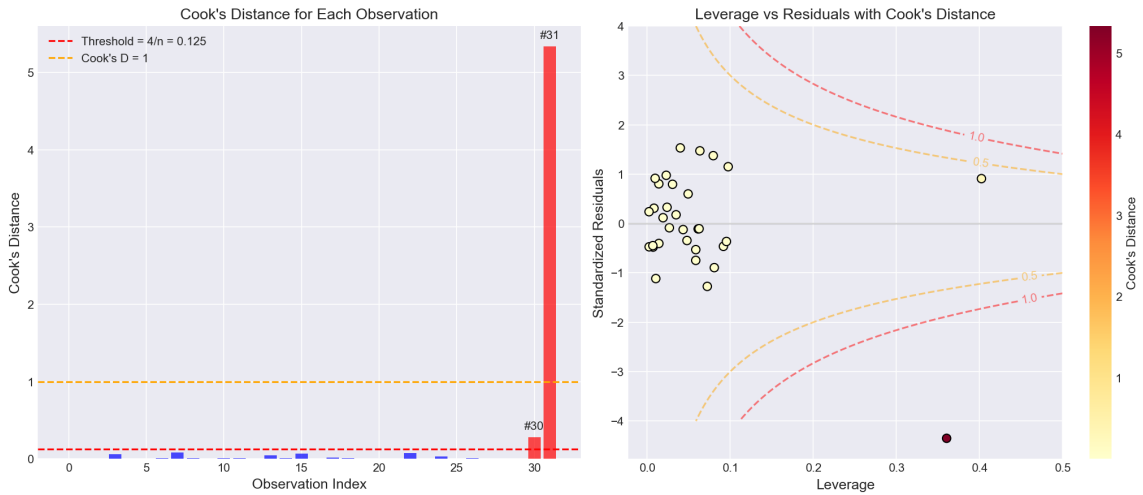
$$h_{ii} = \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i$$

Measures how far x_i is from center of X-space

Cook's Distance:

$$D_i = \frac{r_i^2}{p} \cdot \frac{h_{ii}}{1 - h_{ii}}$$

Combines residual and leverage



Action Steps:

- Investigate points with high Cook's D
- Check for data entry errors
- Consider separate analysis
- Report sensitivity

Distance metric choice affects clustering results

Remedial Measures

For Non-linearity:

- Add polynomial terms
- Transform variables
- Use splines
- Consider different model

For Non-normality:

- Transform Y (Box-Cox)
- Use robust regression
- Bootstrap confidence intervals
- Large sample $\rightarrow$ rely on CLT

For Heteroscedasticity:

- Transform Y (log, sqrt)
- Weighted least squares
- Robust standard errors
- Generalized least squares

For Autocorrelation:

- Add time lags
- First differences
- ARIMA errors
- Panel data methods

General Principle: Understand the problem before applying remedy

---

Key insight from Remedial Measures

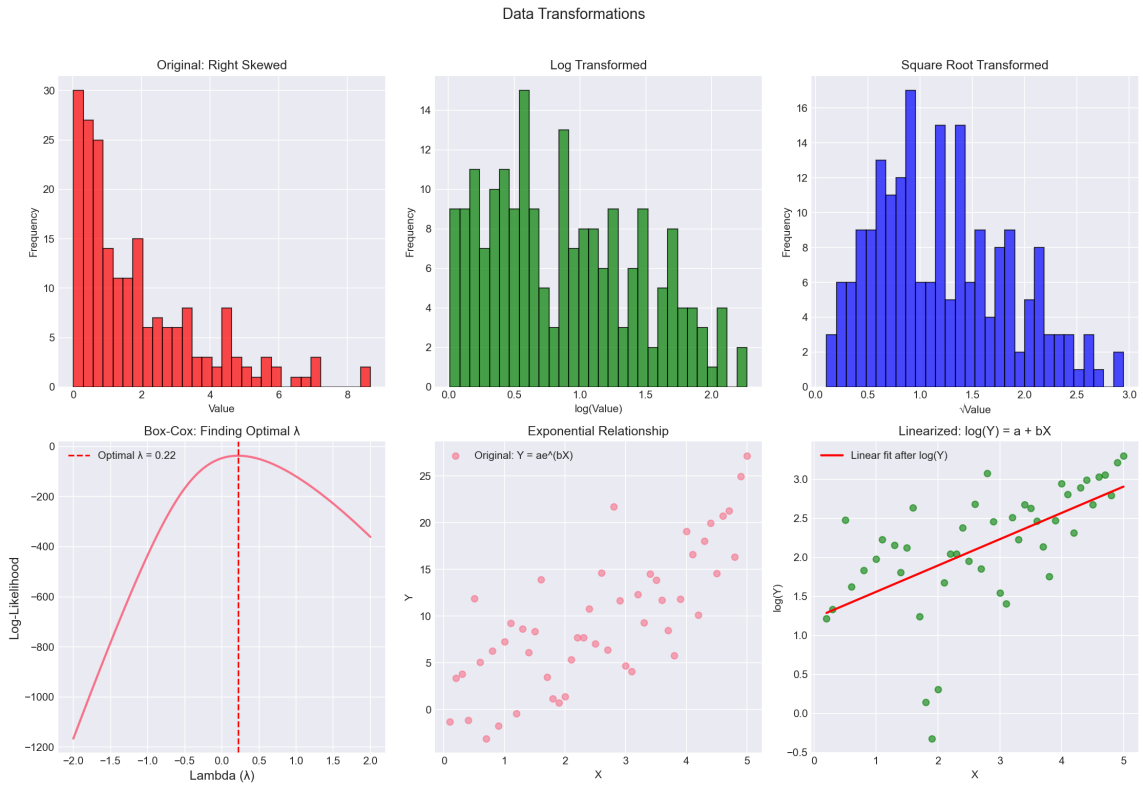
Transformations

Common Transformations:
For Y:

- Log: Right skew, multiplicative
- Square root: Count data
- Reciprocal: Extreme skew
- Box-Cox: Optimal power

For X:

- Log: Diminishing returns
- Polynomial: Curves
- Centering: Reduce multicollinearity



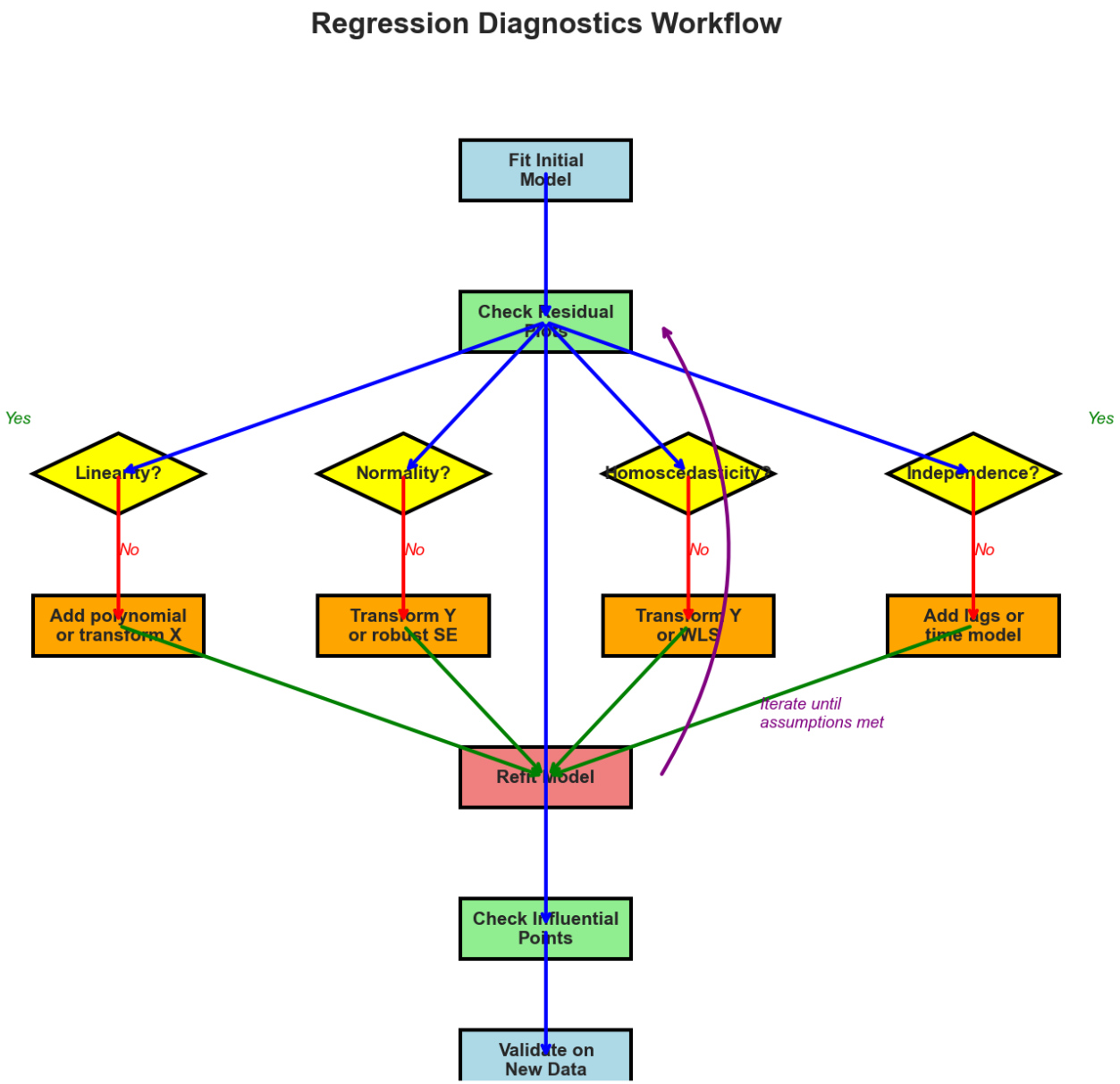
Box-Cox Transformation:

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log(y) & \lambda = 0 \end{cases}$$

Choose λ to maximize normality

Key insight from Transformations

Diagnostic Workflow



Systematic Approach:

Real Dataset Analysis - Boston Housing

Dataset: 506 Boston suburbs
Variables:

- MEDV: Median home value
- CRIM: Crime rate
- RM: Average rooms
- DIS: Distance to employment
- NOX: Nitrogen oxide

Objective: Predict home values

Final Model:

$$\text{MEDV} = 36.5 - 0.11 \cdot \text{CRIM} + 3.8 \cdot \text{RM}$$

$$-17.8 \cdot \text{NOX} + 0.69 \cdot \text{DIS} - 0.52 \cdot \text{LSTAT}$$

$R^2 = 0.74$, RMSE $= 4.75$

Key insight from Real Dataset Analysis - Boston Housing

Model Comparison Techniques

Information Criteria:
AIC (Akaike):

$$AIC = n \ln(SSE/n) + 2p$$

- Lower is better
- Balances fit and complexity

BIC (Bayesian):

$$BIC = n \ln(SSE/n) + p \ln(n)$$

- Stronger penalty for parameters
- More conservative

Cross-Validation:

- Split data: train/test
- k-fold CV
- Leave-one-out
- Best for prediction focus

Key insight from Model Comparison Techniques

Cross-Validation

k-Fold CV Process:

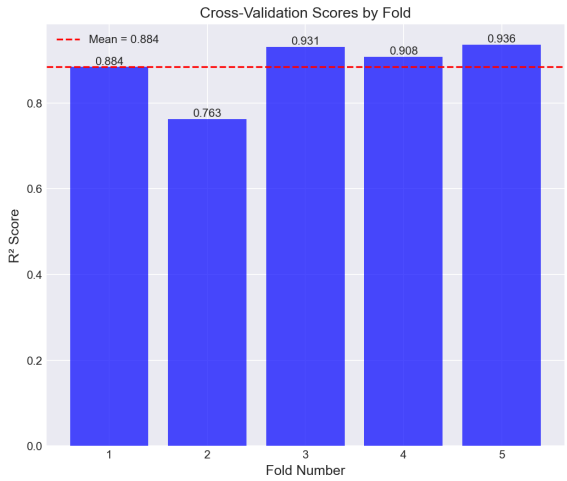
1. Split data into k folds
2. Train on k-1 folds
3. Test on remaining fold
4. Repeat k times
5. Average performance

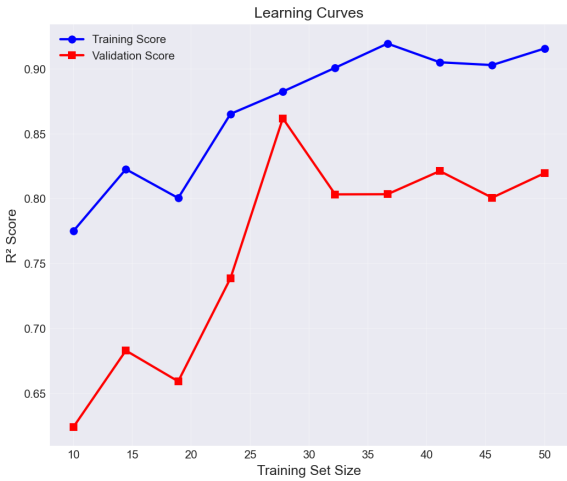
Common Choices:

- k = 5 or 10
- Leave-one-out: k = n

Metrics:

- RMSE
- MAE
- R^2





Benefits:

- Avoid overfitting
- Better generalization estimate
- Model selection
- Hyperparameter tuning

Key insight from Cross-Validation

Common Pitfalls

Data Issues:

- Missing not at random
- Measurement error
- Selection bias
- Survivorship bias

Model Issues:

- Overfitting with many predictors
- Ignoring non-linearity
- Extrapolation beyond data
- Ignoring interactions

Inference Issues:

- p-hacking
- Multiple testing
- Post-hoc theorizing

Interpretation Issues:

- Confusing correlation/causation
- Ignoring practical significance
- Over-interpreting R^2
- Simpson's paradox

Validation Issues:

- No holdout test set
- Data leakage
- Temporal issues
- Different populations

Best Practices:

- Pre-register analysis plan
- Use holdout data
- Report all models tried
- Consider practical significance

Avoid these common mistakes in your analysis

Summary and Key Takeaways

Simple Linear Regression:

- One predictor $\rightarrow$ one outcome
- $Y = \beta_0 + \beta_1 X + \epsilon$
- Least squares minimization
- Clear interpretation

Multiple Linear Regression:

- Multiple predictors
- Partial effects
- Control for confounders
- Better predictions

Key Concepts:

- Coefficient interpretation
- Hypothesis testing
- R^2 and adjusted R^2
- Prediction intervals

Diagnostics:

- Check all assumptions
- Residual analysis crucial
- Identify influential points
- Apply appropriate remedies

Model Building:

- Start simple
- Add complexity carefully
- Validate thoroughly
- Interpret cautiously

Remember:

- "All models are wrong, some are useful" - Box
- Focus on practical significance
- Always validate on new data
- Document your process

Review these key points before moving to the next section

Further Resources

Textbooks:

- Kutner et al. - Applied Linear Regression Models
- James et al. - Introduction to Statistical Learning
- Faraway - Linear Models with R
- Montgomery - Introduction to Linear Regression

Software:

- R: lm(), glm()
- Python: statsmodels, scikit-learn
- SPSS, SAS, Stata

Online Resources:

- ISLR videos (Stanford)
- Andrew Ng's course
- Cross Validated (Stack Exchange)
- R-bloggers

Practice Datasets:

- Boston Housing
- mtcars
- Advertising
- California Housing
- Ames Housing

Next Topics: Logistic Regression, GLMs, Time Series, Machine Learning

---

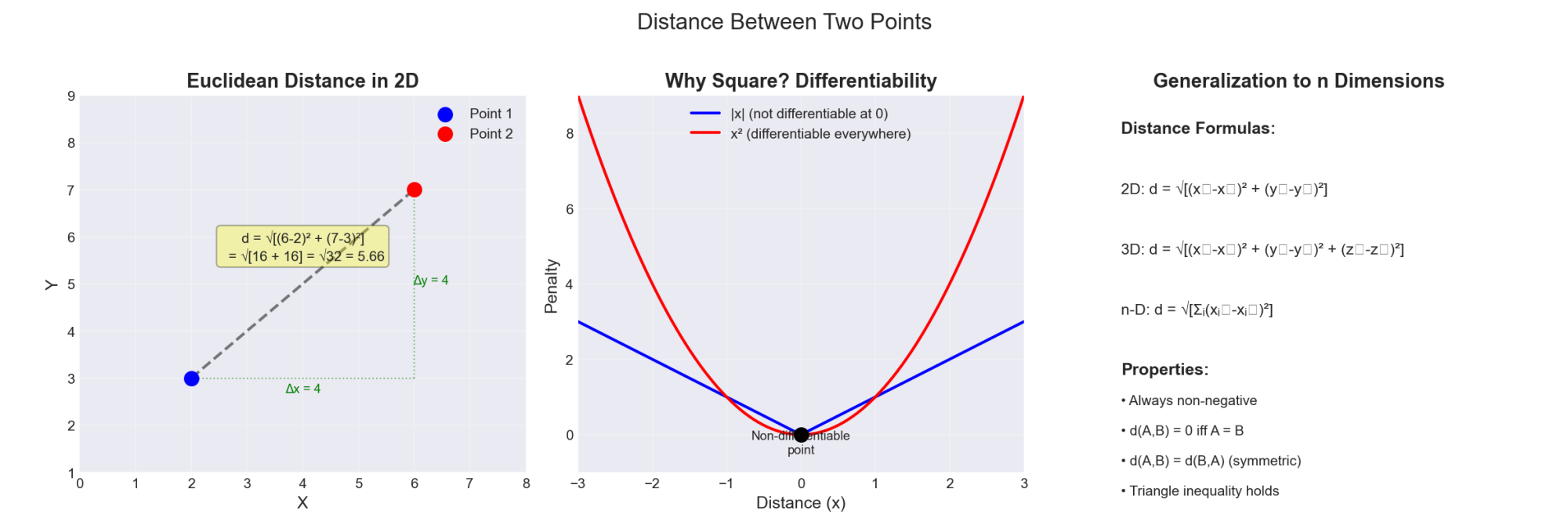
Use these resources to deepen your understanding

Appendix

Mathematical Foundations

Building Concepts from First Principles

Distance Between Two Points

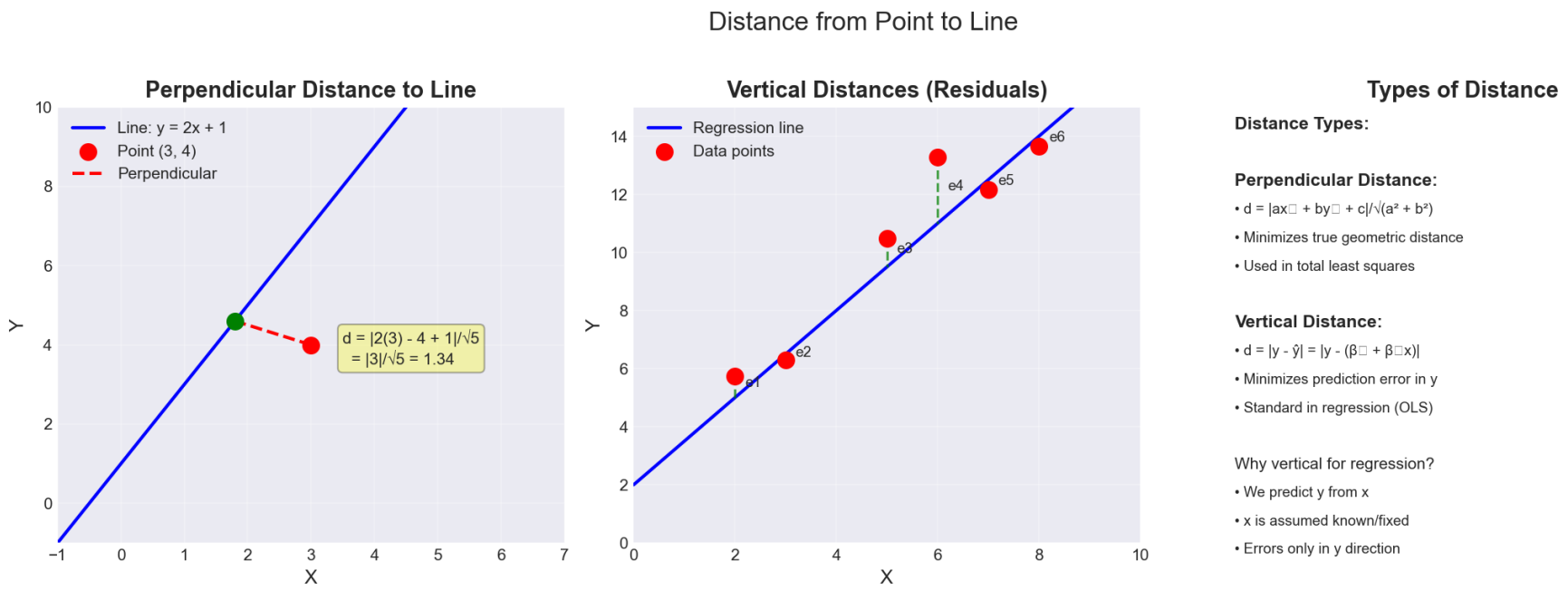


Key Concepts:

- Euclidean distance: $d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$
- Why square? Differentiability and mathematical convenience
- Generalizes to n dimensions: $d = \sqrt{\sum_i (x_{i2} - x_{i1})^2}$

Key insight from center

Distance from Point to Line

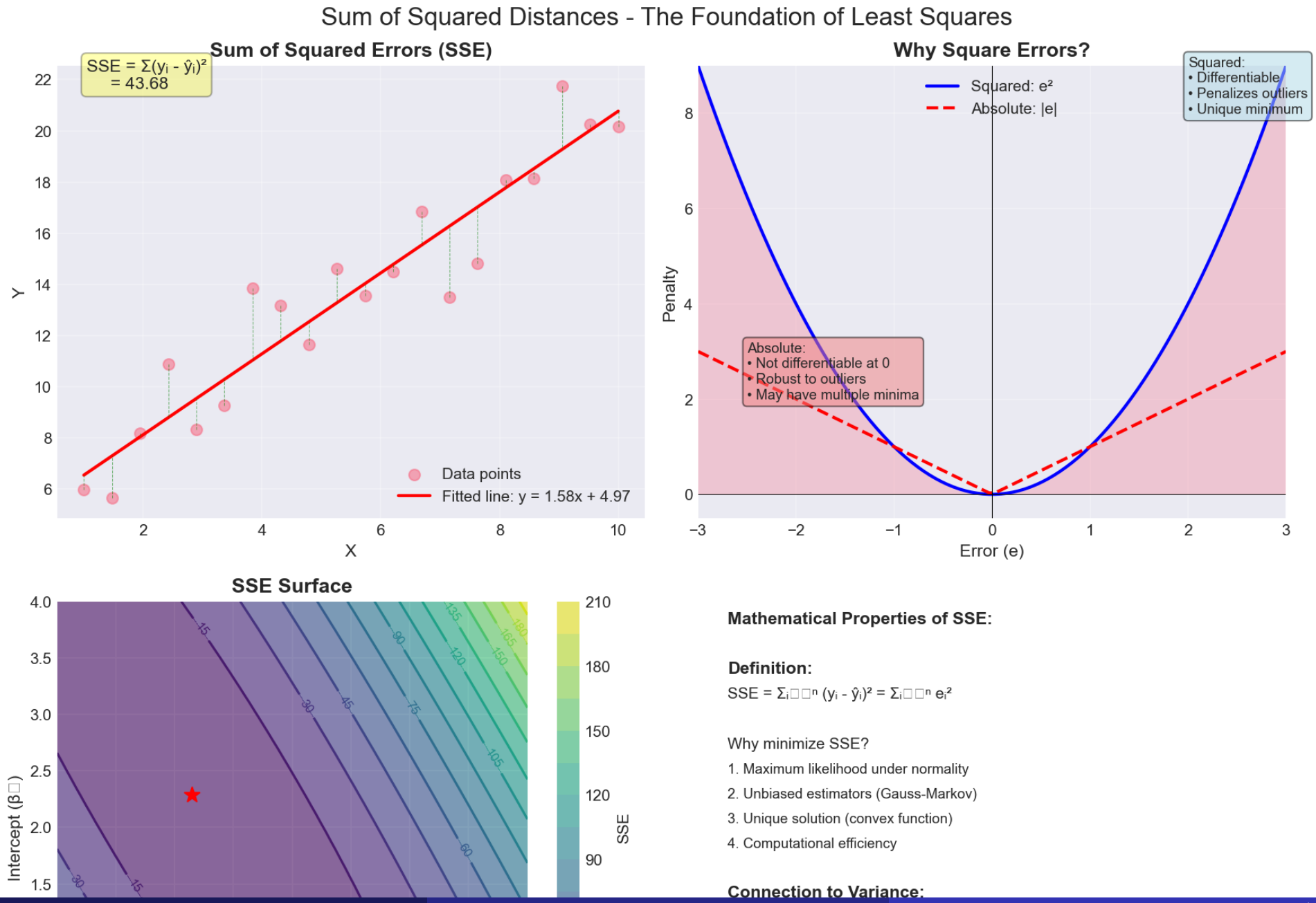


Two Types of Distance:

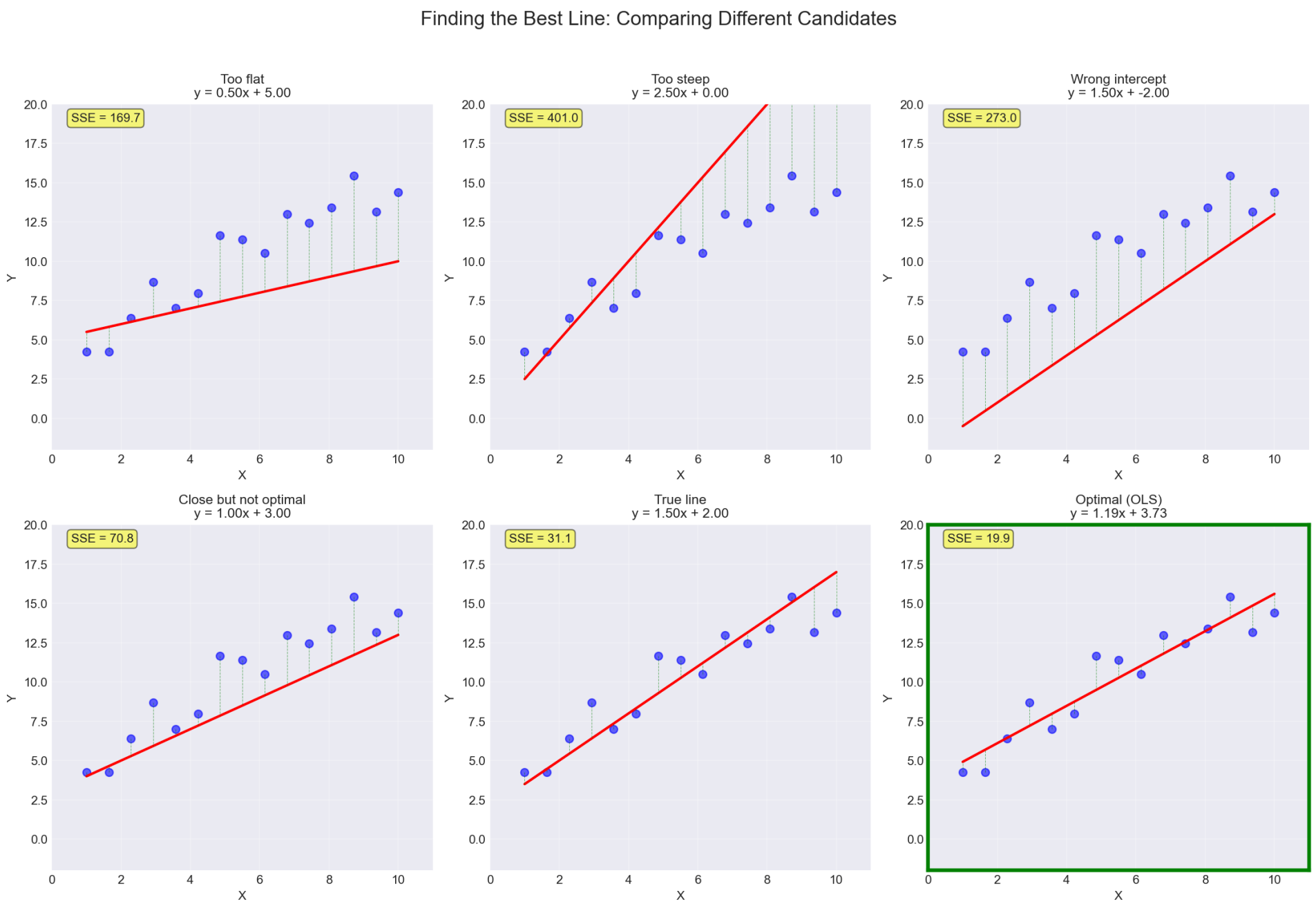
- **Perpendicular**: True geometric distance (used in total least squares)
- **Vertical**: Distance in y-direction $= |y - \hat{y}|$ (used in OLS)
- Why vertical for regression? We predict y from x, assuming x is known

Distance metric choice affects clustering results

Sum of Squared Distances



Finding the Best Line - Optimization



Normal Equations - Complete Derivation

Normal Equations: From Calculus to Solution

Normal Equations: Complete Derivation

Objective: Minimize $SSE = \Sigma_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$

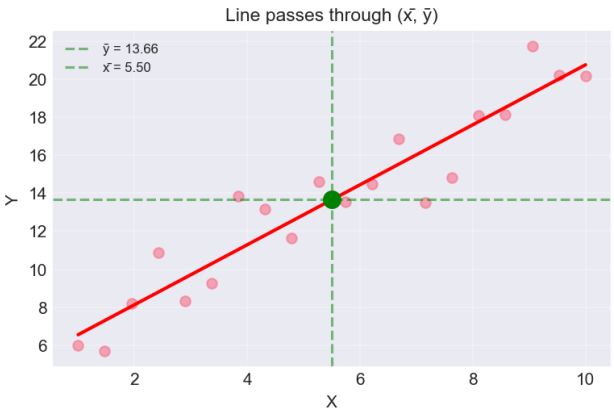
Step 1: Expand the squared term
$SSE = \Sigma_{i=1}^{n} (y_i^2 - 2y_i\beta_0 - 2y_i\beta_1 x_i + \beta_0^2 + 2\beta_0\beta_1 x_i + \beta_1^2 x_i^2)$

Step 2: Take partial derivatives
$\partial SSE/\partial\beta_0 = \Sigma_{i=1}^{n} (-2y_i + 2\beta_0 + 2\beta_1 x_i) = -2\Sigma y_i + 2n\beta_0 + 2\beta_1 \Sigma x_i$
$\partial SSE/\partial\beta_1 = \Sigma_{i=1}^{n} (-2y_i x_i + 2\beta_0 x_i + 2\beta_1 x_i^2) = -2\Sigma x_i y_i + 2\beta_0 \Sigma x_i + 2\beta_1 \Sigma x_i^2$

Step 3: Set derivatives equal to zero and simplify
$-2\Sigma y_i + 2n\beta_0 + 2\beta_1 \Sigma x_i = 0 \rightarrow n\beta_0 + \beta_1 \Sigma x_i = \Sigma y_i$
$-2\Sigma x_i y_i + 2\beta_0 \Sigma x_i + 2\beta_1 \Sigma x_i^2 = 0 \rightarrow \beta_0 \Sigma x_i + \beta_1 \Sigma x_i^2 = \Sigma x_i y_i$

Step 4: Solve the system (Normal Equations)
From first equation: $\beta_0 = (\Sigma y_i - \beta_1 \Sigma x_i)/n = \bar{y} - \beta_1 \bar{x}$

Substitute into second equation:
$(\bar{y} - \beta_1 \bar{x})\Sigma x_i + \beta_1 \Sigma x_i^2 = \Sigma x_i y_i$
$\beta_1(\Sigma x_i^2 - \bar{x}\Sigma x_i) = \Sigma x_i y_i - \bar{y}\Sigma x_i$
$\beta_1 = (\Sigma x_i y_i - n\bar{x}\bar{y})/(\Sigma x_i^2 - n\bar{x}^2)$



Final Formulas:

$\hat{\beta}_1 = \Sigma(x_i - \bar{x})(y_i - \bar{y})/\Sigma(x_i - \bar{x})^2$

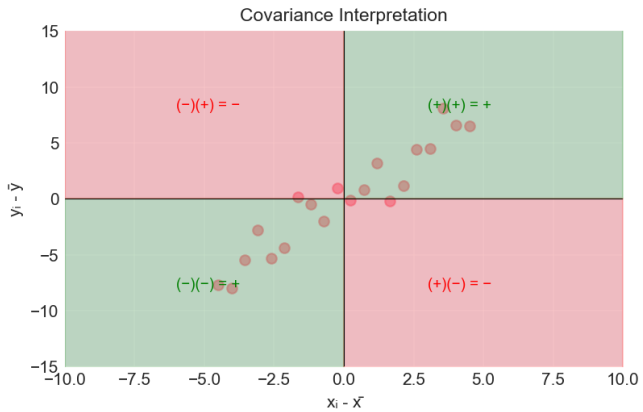
$= Cov(X,Y)/Var(X)$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

Alternative form:

$\hat{\beta}_1 = (n\Sigma x_i y_i - \Sigma x_i \Sigma y_i)/$

$(n\Sigma x_i^2 - (\Sigma x_i)^2)$



From Calculus to Solution:

- Minimize: $SSE = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$

Matrix Form Derivation

Matrix Form of Linear Regression

Matrix Form Setup

Data: n observations (x_i, y_i)

Design Matrix X:

$$\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

Response Vector Y:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Parameter Vector β:

$$\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

Prediction Ŷ:

$\hat{Y} = X\beta$

Error Vector e:

$e = Y - \hat{Y} = Y - X\beta$

Key Matrix Products

$$X^TX = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \Sigma x_i \\ \Sigma x_i & \Sigma x_i^2 \end{bmatrix}$$

$$X^TY = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \Sigma y_i \\ \Sigma x_i y_i \end{bmatrix}$$

$SSE = e^Te = (Y-X\beta)^T(Y-X\beta)$

Minimization

Minimize: $SSE = (Y-X\beta)^T(Y-X\beta)$

$= Y^TY - 2\beta^TX^TY + \beta^TX^TX\beta$

Take derivative w.r.t. β:

$\partial SSE/\partial\beta = -2X^TY + 2X^TX\beta$

Set to zero:

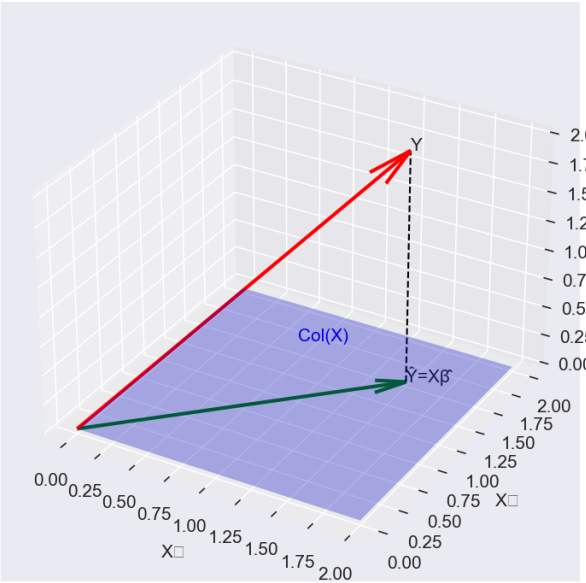
$X^TX\beta = X^TY$

Solution (if X^TX is invertible):

$\hat{\beta} = (X^TX)^{-1}X^TY$

This is the OLS estimator!

Projection onto Column Space



When is $(X^TX)^{-1}$ Exists?

(X^TX) is invertible when:

1. X has full column rank
   - No perfect multicollinearity
   - Columns linearly independent
2. $n \geq p$ (observations ≥ parameters)
   - More equations than unknowns
   - System is not underdetermined
3. Variance in predictors
   - Not all x_i are the same
   - $\Sigma(x_i-\bar{x})^2 > 0$

Counter-examples:

- All x values identical → singular
- Perfect correlation between predictors
- $n < p$ (more parameters than data)

Connection to Normal Equations

Matrix form: $X^TX\beta = X^TY$

Expand for simple regression:

$$\begin{bmatrix} n & \Sigma x_i \\ \Sigma x_i & \Sigma x_i^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \Sigma y_i \\ \Sigma x_i y_i \end{bmatrix}$$

This gives two equations:

$n\beta_0 + \beta_1\Sigma x_i = \Sigma y_i$

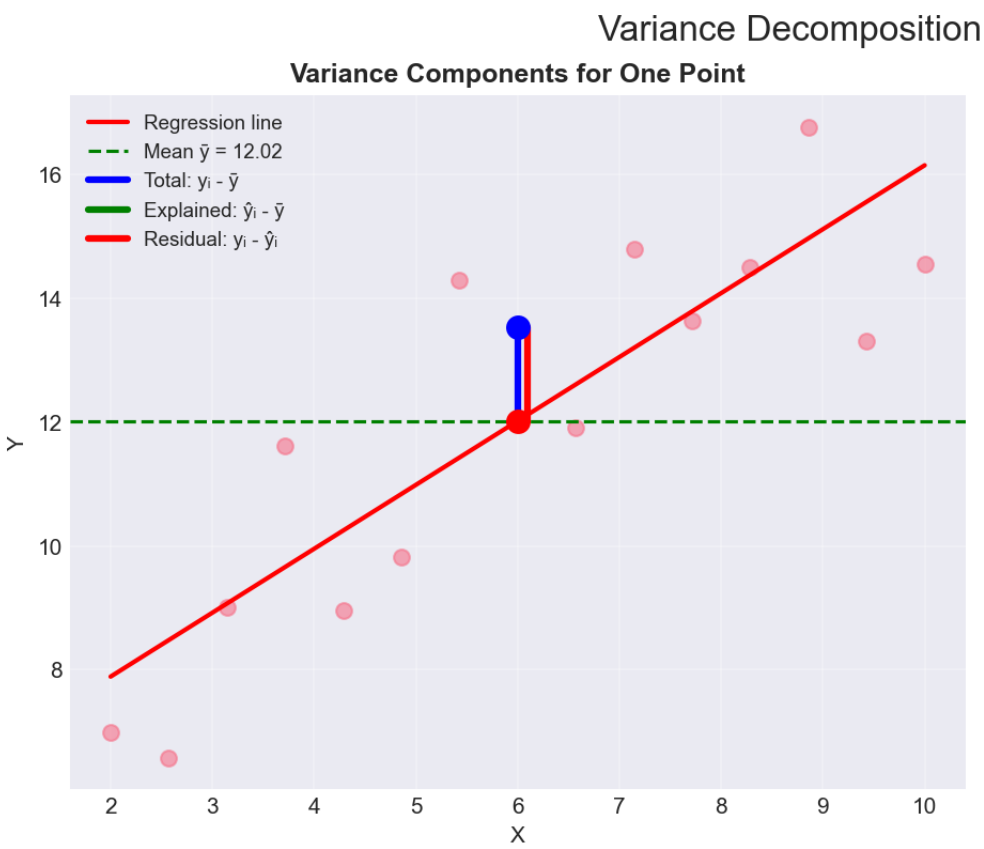
$\beta_0\Sigma x_i + \beta_1\Sigma x_i^2 = \Sigma x_i y_i$

These are exactly the Normal Equations!

Matrix form is more general:

- Works for multiple regression
- Compact notation
- Computational efficiency

Variance Decomposition: Why SST = SSR + SSE



Algebraic Proof

Start with: $y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$

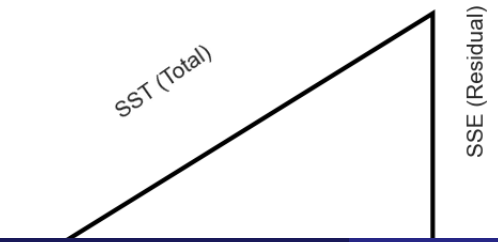
Square both sides and sum:
$\Sigma(y_i - \bar{y})^2 = \Sigma[(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2$

Expand the square:
$= \Sigma(y_i - \hat{y}_i)^2 + \Sigma(\hat{y}_i - \bar{y})^2 + 2\Sigma(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$

Key insight: The cross-product term = 0
Because $\Sigma e_i(\hat{y}_i - \bar{y}) = 0$ (orthogonality)

Therefore:
SST = SSE + SSR
Total = Unexplained + Explained

Pythagorean Theorem in n-dimensions



Numerical Example

From our data (n = 15):

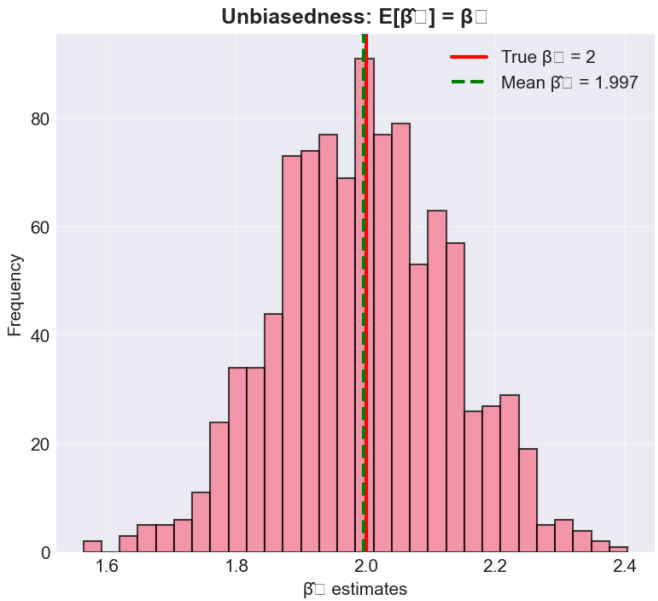
$\bar{y} = 12.02$

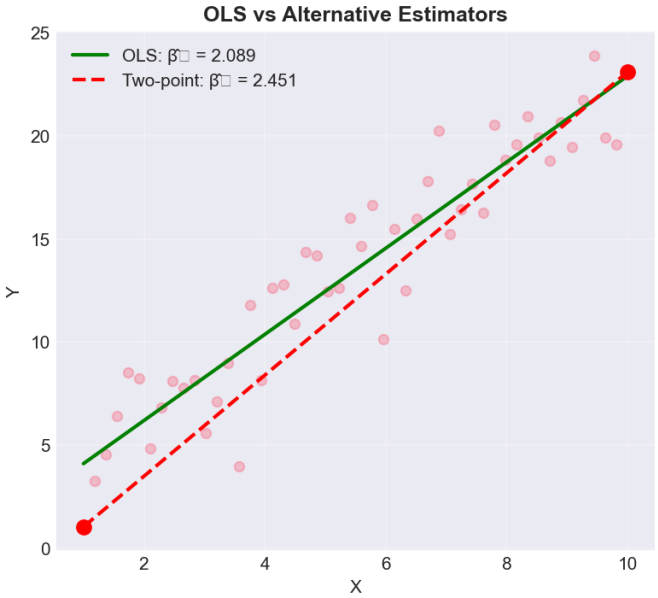
$SST = \Sigma(y_i - \bar{y})^2$
$= 132.94$

$SSR = \Sigma(\hat{y}_i - \bar{y})^2$
$= 97.54$

$SSE = \Sigma(y_i - \hat{y}_i)^2$

Properties of Least Squares Estimators





Properties of Least Squares Estimators

Variance Formulas

For simple regression:

$$\text{Var}(\hat{\beta}_1) = \sigma^2/\Sigma(x_i - \bar{x})^2 = \sigma^2/S_{xx}$$

$$\text{Var}(\hat{\beta}_0) = \sigma^2[1/n + \bar{x}^2/\Sigma(x_i - \bar{x})^2] = \sigma^2[1/n + \bar{x}^2/S_{xx}]$$

where $\sigma^2 = \text{Var}(\varepsilon_i)$

Standard Errors (estimated):

$$SE(\hat{\beta}_1) = s/\sqrt{S_{xx}}$$

$$SE(\hat{\beta}_0) = s\sqrt{[1/n + \bar{x}^2/S_{xx}]}$$

where $s^2 = MSE = SSE/(n-2)$

Gauss-Markov Theorem

Under assumptions:
1. $E[\varepsilon_i] = 0$ (zero mean errors)
2. $\text{Var}(\varepsilon_i) = \sigma^2$ (homoscedasticity)
3. $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ (independence)
4. X is non-stochastic

OLS estimators are BLUE:
Best - minimum variance
Linear - linear in Y
Unbiased - $E[\hat{\beta}] = \beta$
Estimators

Among all linear unbiased estimators, OLS has the smallest variance!

Covariance Matrix

For $\hat{\beta} = [\hat{\beta}_0, \hat{\beta}_1]^T$:

$$\text{Cov}(\hat{\beta}) = \sigma^2(X^TX)^{-1}$$

For simple regression:

$$\text{Cov}(\hat{\beta}) = \begin{bmatrix} \text{Var}(\hat{\beta}_0) & \text{Cov}(\hat{\beta}_0,\hat{\beta}_1) \\ \text{Cov}(\hat{\beta}_0,\hat{\beta}_1) & \text{Var}(\hat{\beta}_1) \end{bmatrix}$$

where:

$$\text{Cov}(\hat{\beta}_0,\hat{\beta}_1) = -\sigma^2\bar{x}/S_{xx}$$

Note: $\hat{\beta}_0$ and $\hat{\beta}_1$ are correlated! (unless $\bar{x} = 0$, i.e., centered)

Key Properties Summary

□ **Unbiased:**

$E[\hat{\beta}] = \beta$

□ **Consistent:**

$\hat{\beta} \to^p \beta$ as $n \to \infty$

□ **Efficient:**

Minimum variance (BLUE)

□ **Normally distributed:**

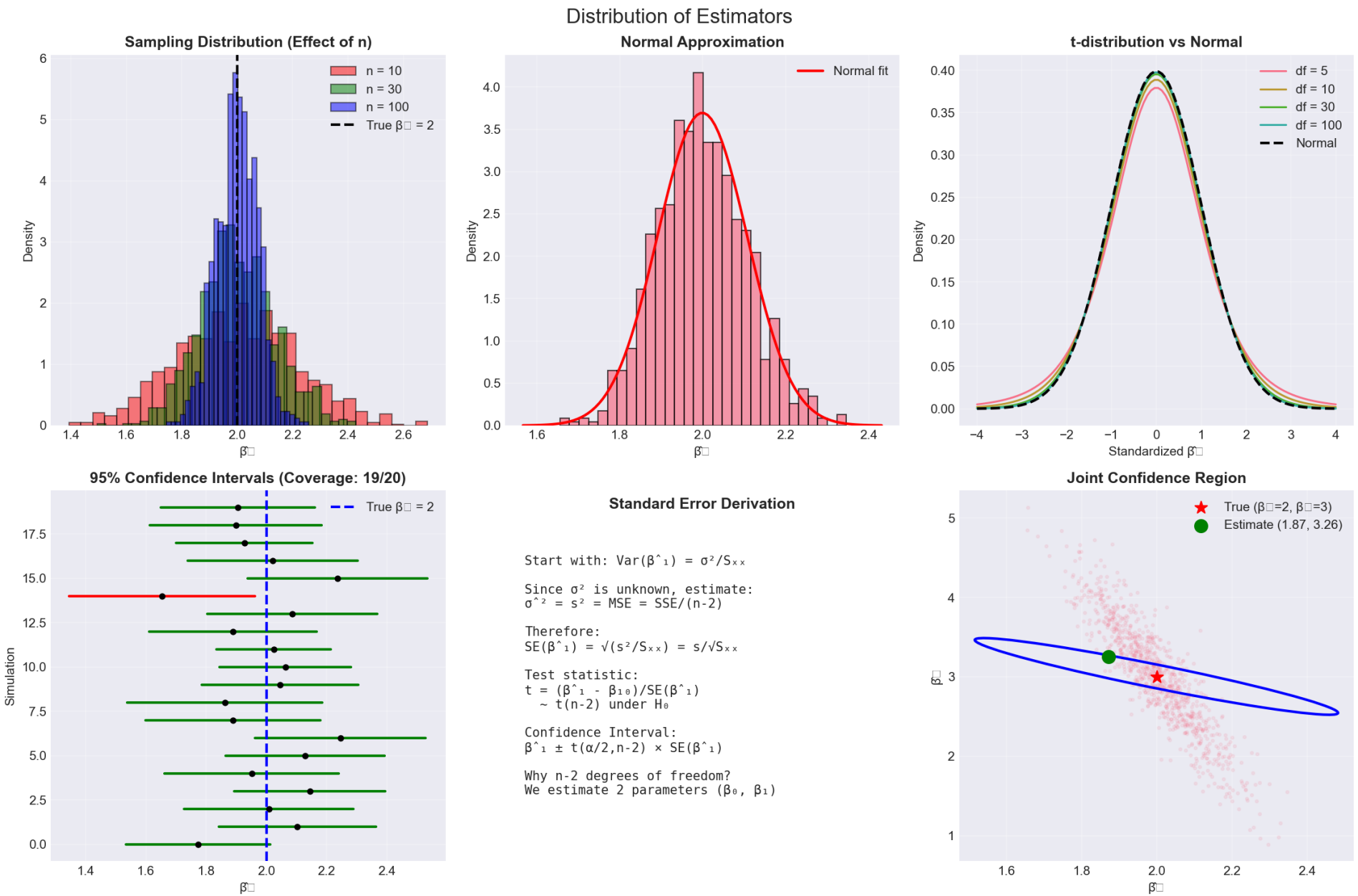
$\hat{\beta} \sim N(\beta, \sigma^2(X^TX)^{-1})$

(under normality)

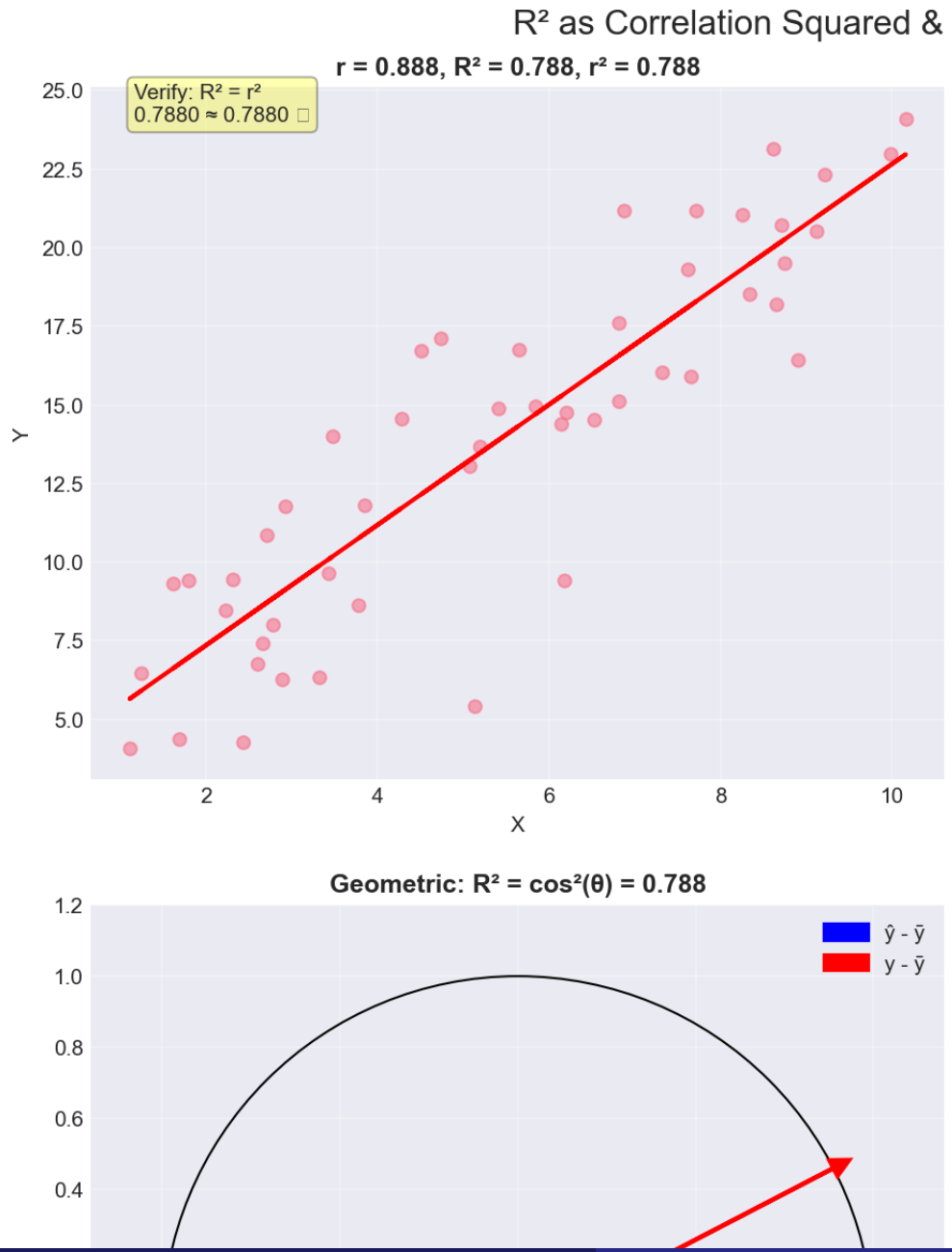
□ **Sufficient:**

Uses all information in data

Distribution of Estimators



R^2 as Correlation Squared



Proof: $R^2 = r^2$ (Simple Regression)

Correlation coefficient:
$r = \Sigma(x_i - \bar{x})(y_i - \bar{y}) / \sqrt{[\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2]}$
$= S_{x□} / (S_x \times S_□)$

Regression slope:
$\hat{\beta}_1 = \Sigma(x_i - \bar{x})(y_i - \bar{y}) / \Sigma(x_i - \bar{x})^2$
$= S_{x□} / S_x^2$

Therefore: $\hat{\beta}_1 = r \times (S_□ / S_x)$

$R^2 = SSR/SST = \hat{\beta}_1^2 \times S_x^2 / S_□^2$
$= [r \times (S_□ / S_x)]^2 \times (S_x^2 / S_□^2)$
$= r^2$

QED: In simple regression, $R^2 = r^2$

Adjusted R^2 Derivation

Problem with R^2:
Always increases with more predictors
(even if they're irrelevant)

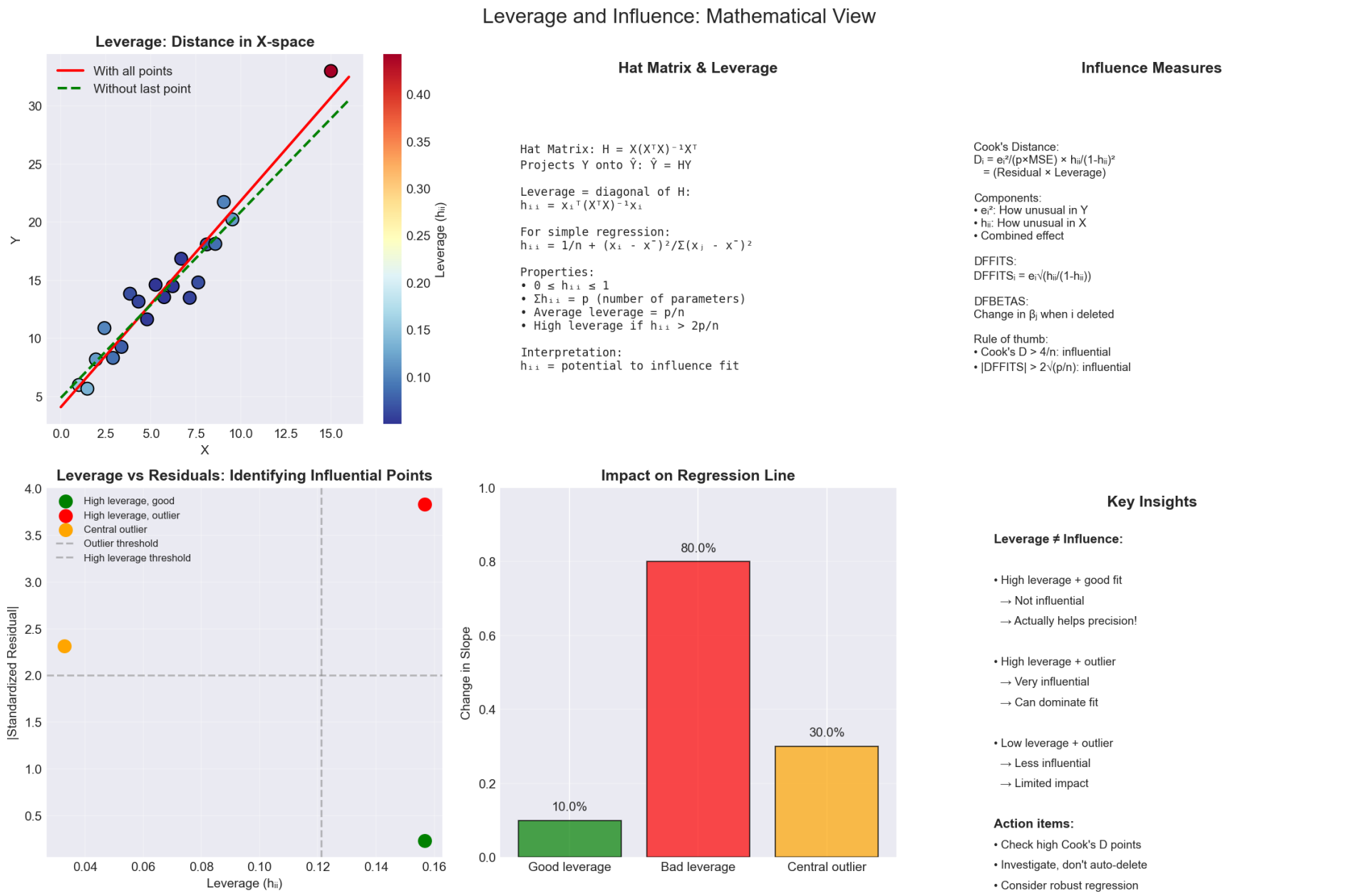
Solution: Adjusted R^2
Penalizes for number of predictors

$R^2_{adj} = 1 - (1 - R^2) \times (n-1)/(n-p-1)$

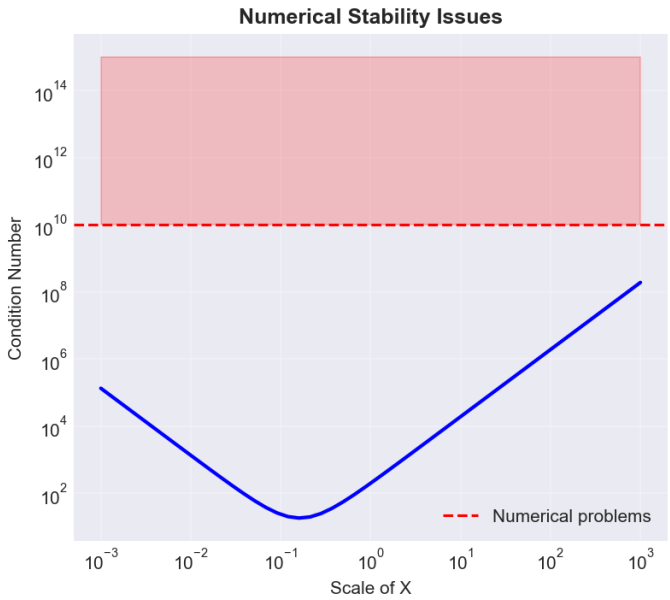
where p = number of predictors

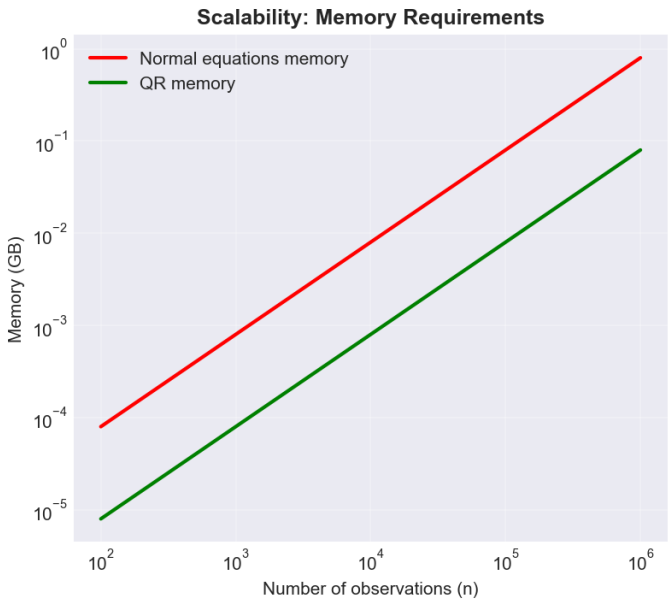
Alternative form:

Leverage and Influence - Mathematical View



Computational Considerations





Computational Considerations

QR Decomposition Alternative

Normal Equations:
$\hat{\beta} = (X^T X)^{-1} X^T Y$
- Forms $X^T X$ (squares condition number)
- Can be numerically unstable

QR Decomposition:
$X = QR$ where $Q^T Q = I$
$\hat{\beta} = R^{-1} Q^T Y$
- No $X^T X$ formation
- Better numerical stability
- Preferred in practice

Computational complexity:
- Normal: $O(np^2 + p^3)$
- QR: $O(np^2)$
- QR better for $n >> p$

Benefits of Standardization

Before Standardization:

X□: mean=4896, std=904
X□: mean=0.50, std=0.009

Condition number: 2.89e+09

After Standardization:

X□: mean=0, std=1
X□: mean=0, std=1

Condition number: 1.32

Large-Scale Solutions

For Big Data ($n >> 10^6$):

Stochastic Gradient Descent:
- Process mini-batches
- O(1) memory
- Convergence guaranteed

Coordinate Descent:
- Update one parameter at a time
- Good for sparse data

Out-of-core algorithms:
- Data doesn't fit in memory
- Stream processing
- Distributed computing

Modern approaches:
- Spark MLlib
- TensorFlow/PyTorch
- Cloud-based solutions

Practical Recommendations

□ **Always standardize predictors**
(unless interpretability critical)

□ **Use QR for p < 100**
Use iterative for p > 1000

□ **Check condition number**
> 10^{10} → numerical issues

□ **For big data:**
- Sample first for exploration
- Use SGD for final model
- Consider distributed computing

□ **Validate numerically:**
- Compare methods
- Check residuals sum to ~0
- Verify X^TResiduals ≈ 0

Key Takeaways from Mathematical Foundations

Building Blocks:

- Distance $\rightarrow$ SSE $\rightarrow$ Optimization
- Calculus $\rightarrow$ Normal Equations
- Linear Algebra $\rightarrow$ Matrix Form
- All lead to same solution!

Why These Methods?

- Mathematical elegance
- Computational efficiency
- Statistical optimality
- Practical interpretability

Deep Insights:

- SSE minimization = MLE under normality
- Geometry: Projection onto column space
- R^2 has multiple interpretations
- Leverage $\neq$ Influence

Practical Implications:

- Always check numerical stability
- Standardize when appropriate
- Understand your influential points
- Choose right algorithm for data size

Remember: The beauty of regression lies in how simple geometry, calculus, and linear algebra converge to solve practical problems!

Key insight from Computational Considerations