Supervised Learning: From Prediction to Interpretation

Supervised learning forms the cornerstone of predictive analytics, leveraging human-annotated examples to build robust models.

Core Concept: Learning the Mapping

The objective is to learn a function $y = f(X)$ that maps input features (**X**) to an output variable (**y**) based on labeled training data.

Regression: Continuous Outcomes

Predicting a continuous numerical value, such as house prices, temperature forecasts, or future stock values.

Classification: Categorical Labels

Predicting a discrete category or class, such as spam/not-spam, disease/no-disease, or digit recognition (0-9).

The Linear Family of Models

- **Linear Regression:** The fundamental baseline, modeling the relationship as a straight line or hyperplane.
- **Polynomial Regression:** Extends linearity to capture complex, non-linear feature interactions.
- **Regularized Regression (Ridge/Lasso):** Techniques used to prevent overfitting, improving model generalization.
- **Logistic Regression:** A powerful, interpretable linear model used exclusively for *classification* tasks.

Linear Regression: Modeling Continuous Relationships

Linear Regression is the simplest model, assuming a linear relationship exists between the input features and the target variable.

Objective Function

The primary goal is to find the optimal parameter vector θ (coefficients) that defines the line of best fit, minimising the distance between predicted and true values.

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

Learning the Parameters: Two Core Methods

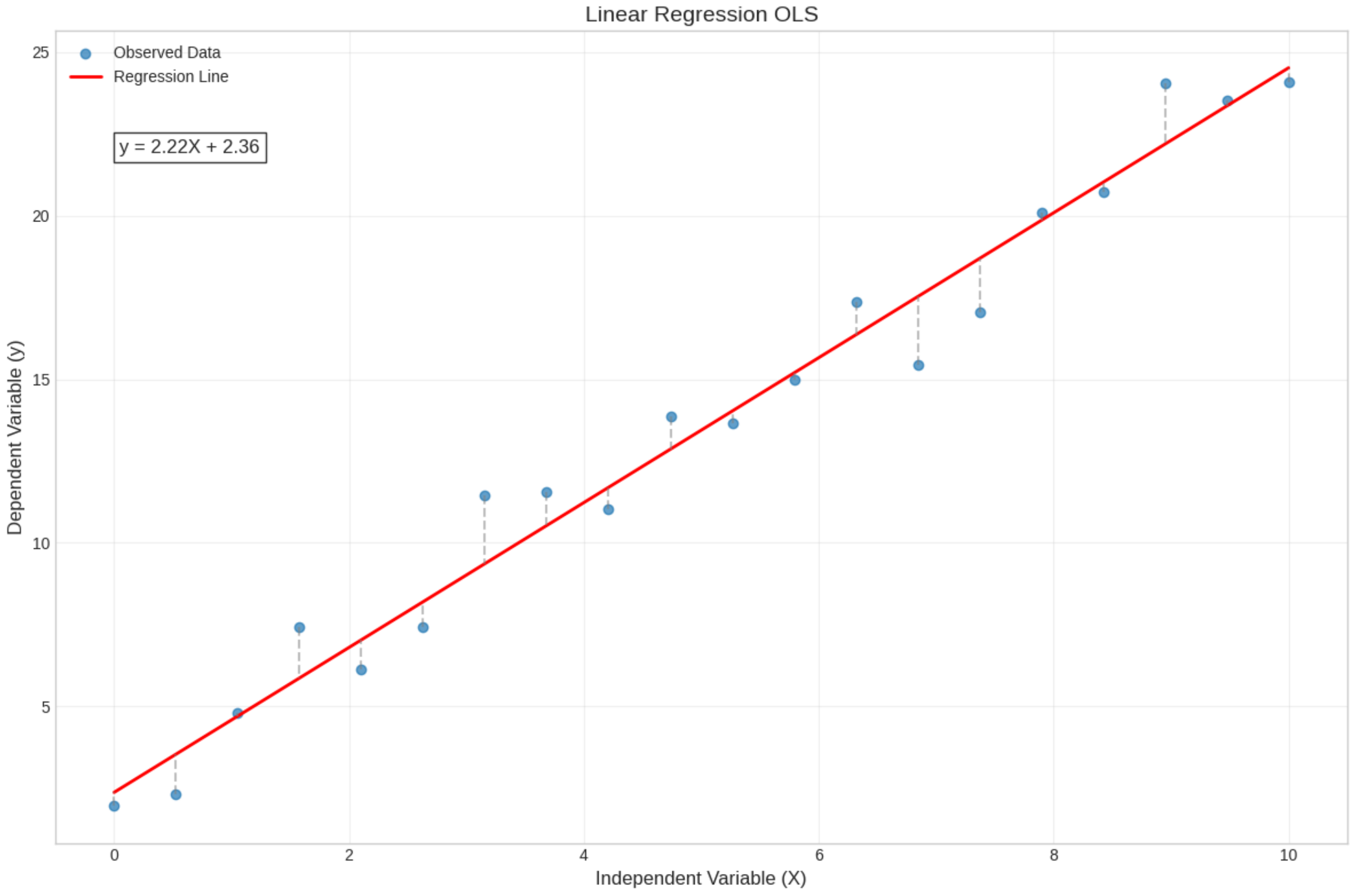
1. Ordinary Least Squares (OLS)

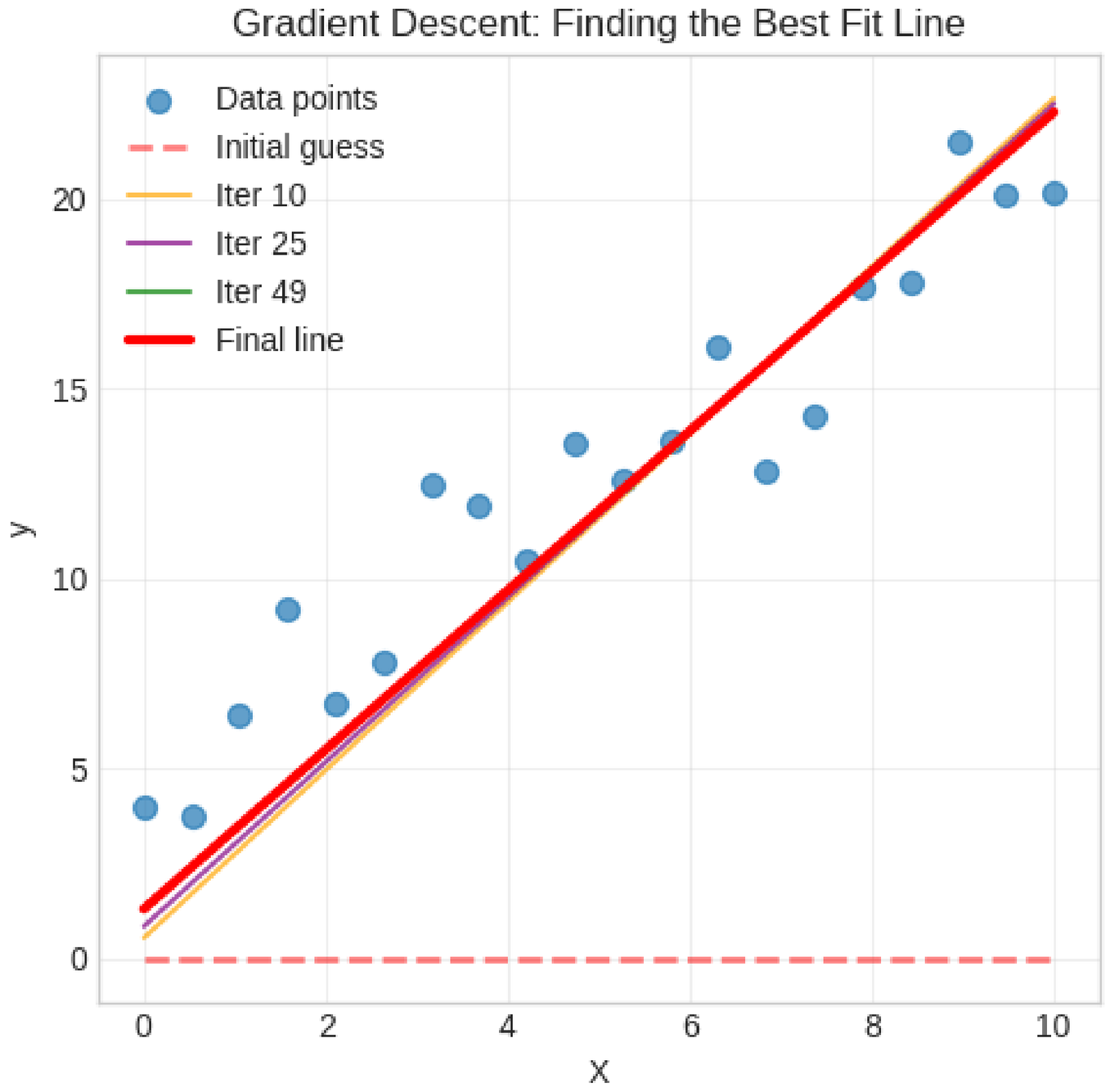
An analytical, closed-form solution derived by setting the gradient of the **Sum of Squared Errors (SSE)** loss function to zero. It's fast and direct for smaller datasets but computationally costly for very large feature matrices.

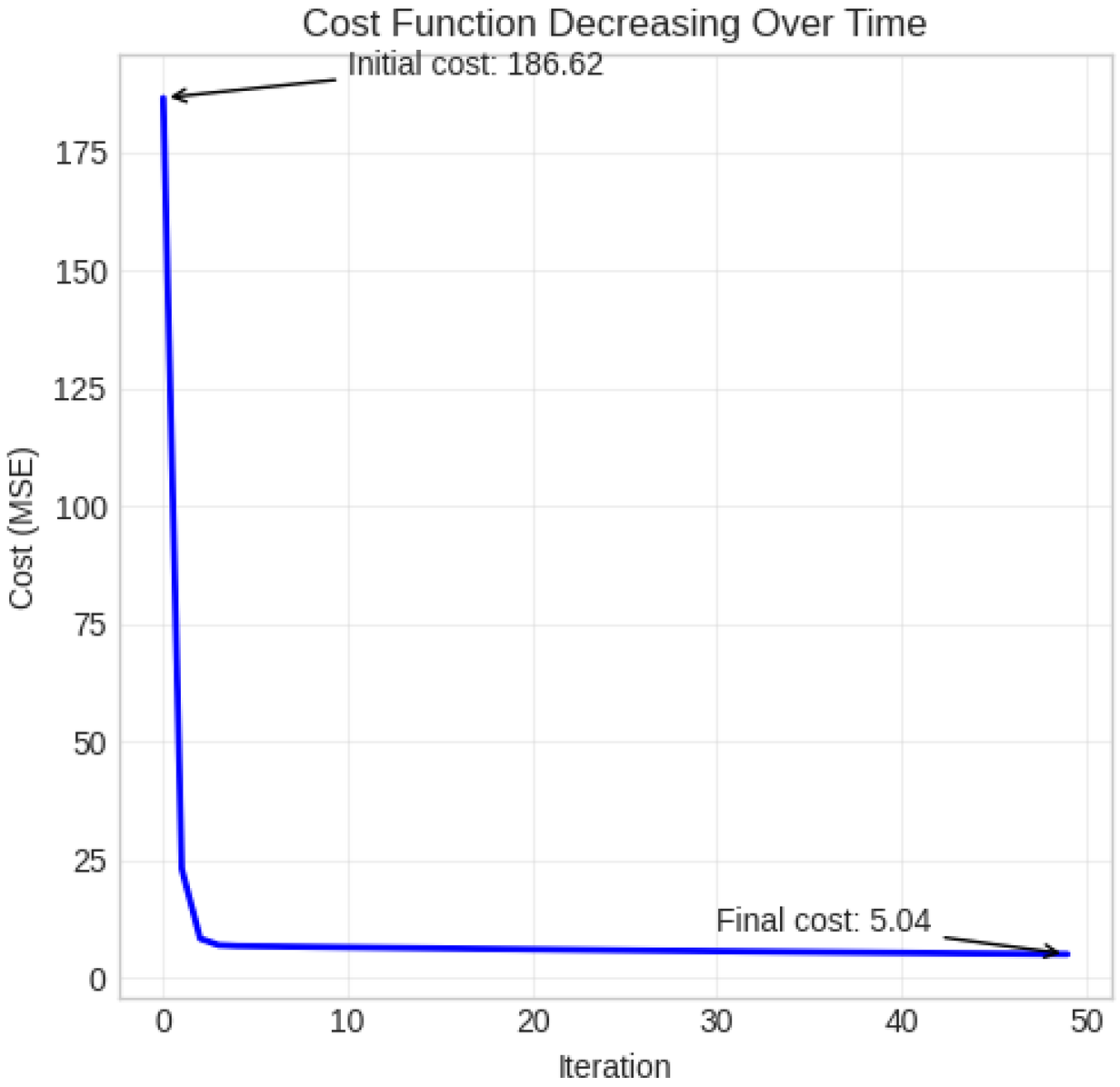
2. Gradient Descent (GD)

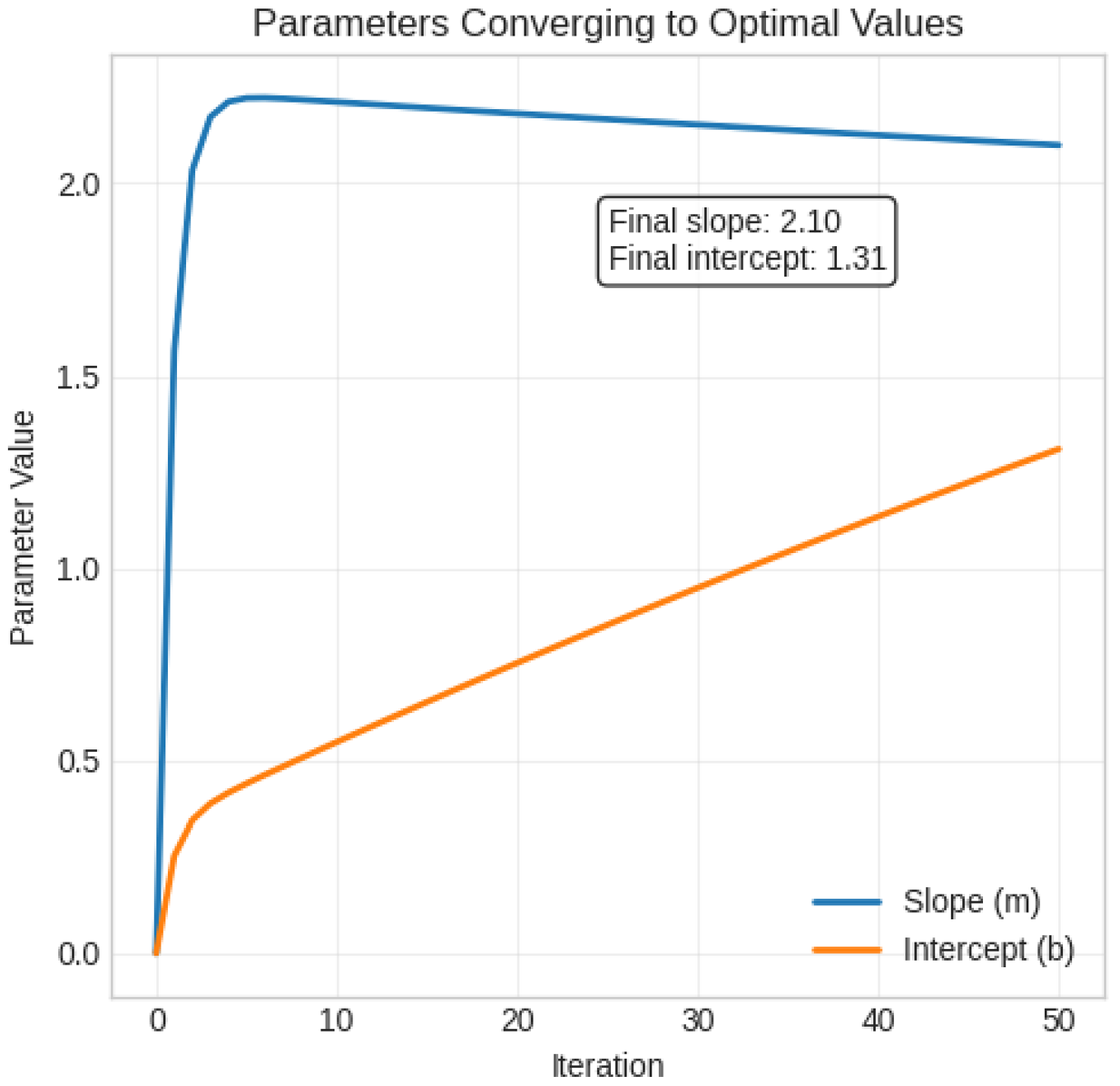
An iterative optimization algorithm that updates parameters by calculating the gradient of the loss function and moving in the direction of steepest descent. Essential for large-scale data and complex models where OLS is infeasible.

- The loss function for Linear Regression is convex, meaning Gradient Descent is guaranteed to find the global minimum.


Linear Regression OLS
Observed Data
Regression Line
y = 2.22X + 2.36
Dependent Variable (y)
Independent Variable (X)
25
20
15
10
5
0
2
4
6
8
10


Gradient Descent: Finding the Best Fit Line
Data points
Initial guess
Iter 10
Iter 25
Iter 49
Final line
y
X
0
5
10
15
20
0
2
4
6
8
10


Cost Function Decreasing Over Time
Initial cost: 186.62
Final cost: 5.04
Cost (MSE)
Iteration
0
25
50
75
100
125
150
175
0
10
20
30
40
50


Parameters Converging to Optimal Values
Final slope: 2.10
Final intercept: 1.31
Slope (m)
Intercept (b)
Parameter Value
Iteration
0.0
0.5
1.0
1.5
2.0
0
10
20
30
40
50

Polynomial Regression & The Power of Regularization

Polynomial Regression: Non-Linearity

This technique extends the model by including powers of features $e.g., x^2, x^3$). While it allows the model to capture complex, non-linear trends, it inherently increases model complexity.

The Danger: Using high-degree polynomials often leads to **overfitting,**

Regularization: Enforcing Simplicity

Regularization modifies the loss function by adding a penalty term based on the magnitude of the coefficients (θ).This forces the model to choose simpler, more generalizable solutions.

Comparing L1 and L2 Penalties

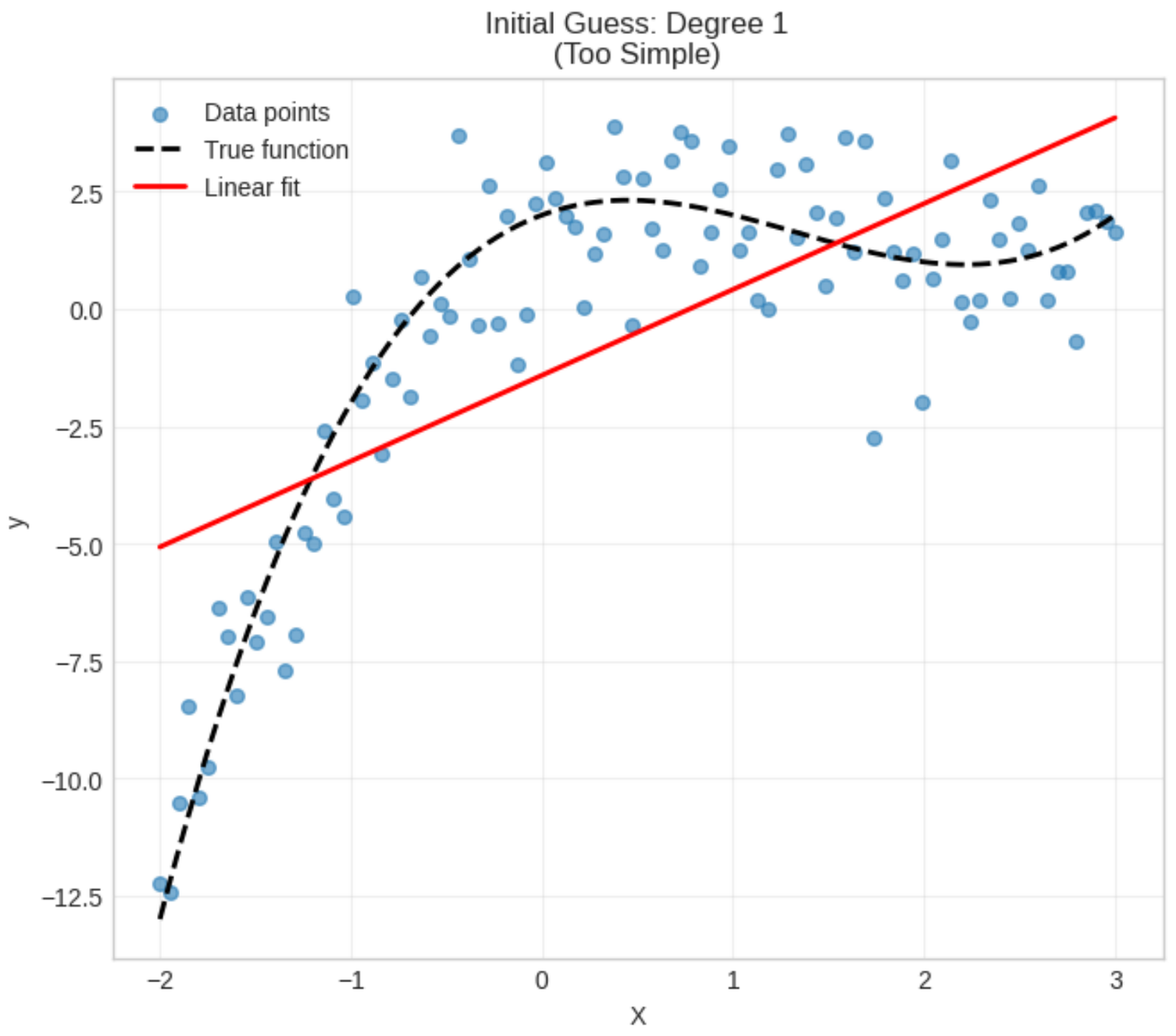
L2 Regularization (Ridge)

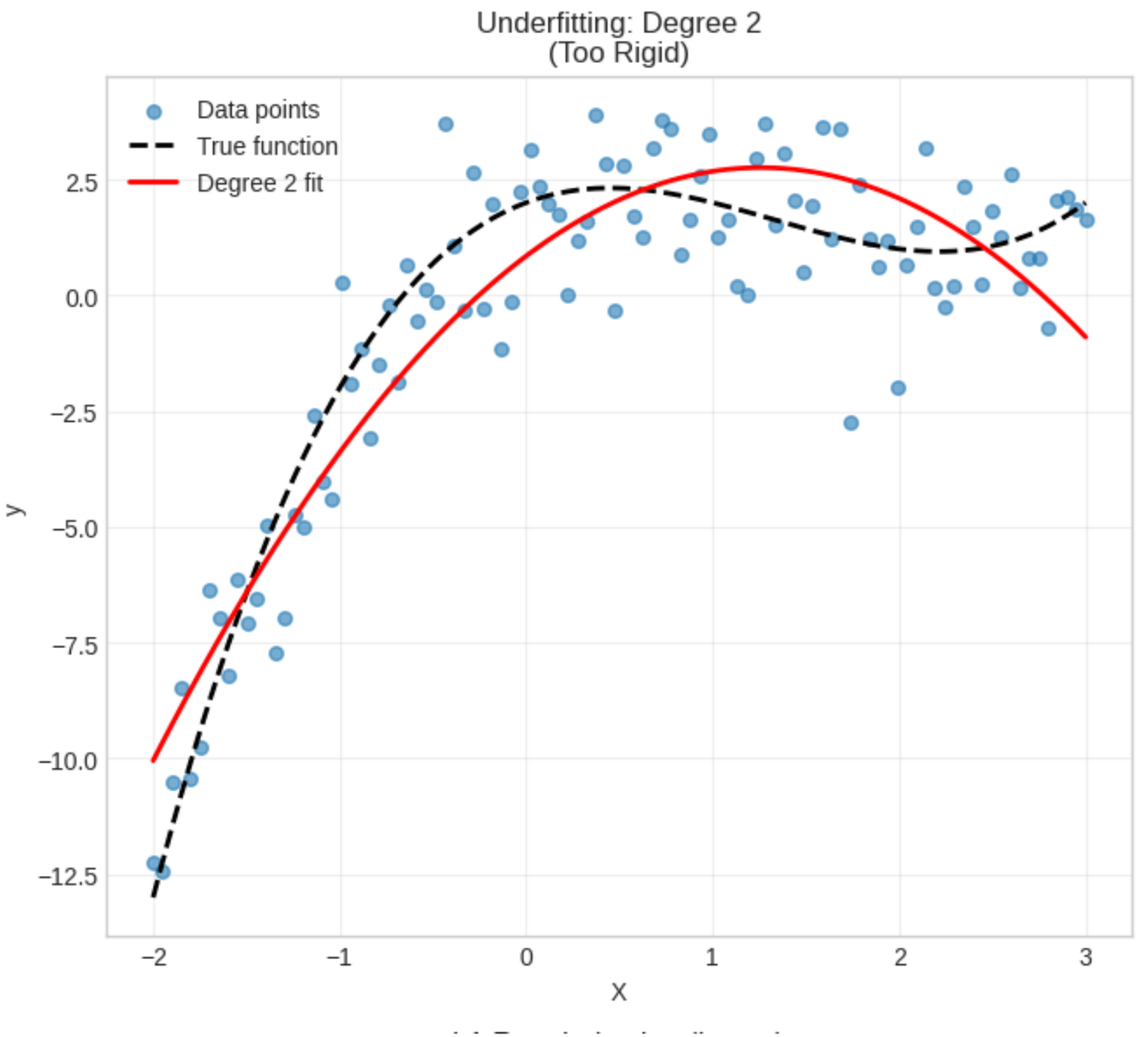
Adds a penalty proportional to the **squared magnitude** of the coefficients $\sum \theta_i^2$. This shrinks the coefficient values towards zero, making the model less sensitive to specific data points, but rarely sets them exactly to zero.

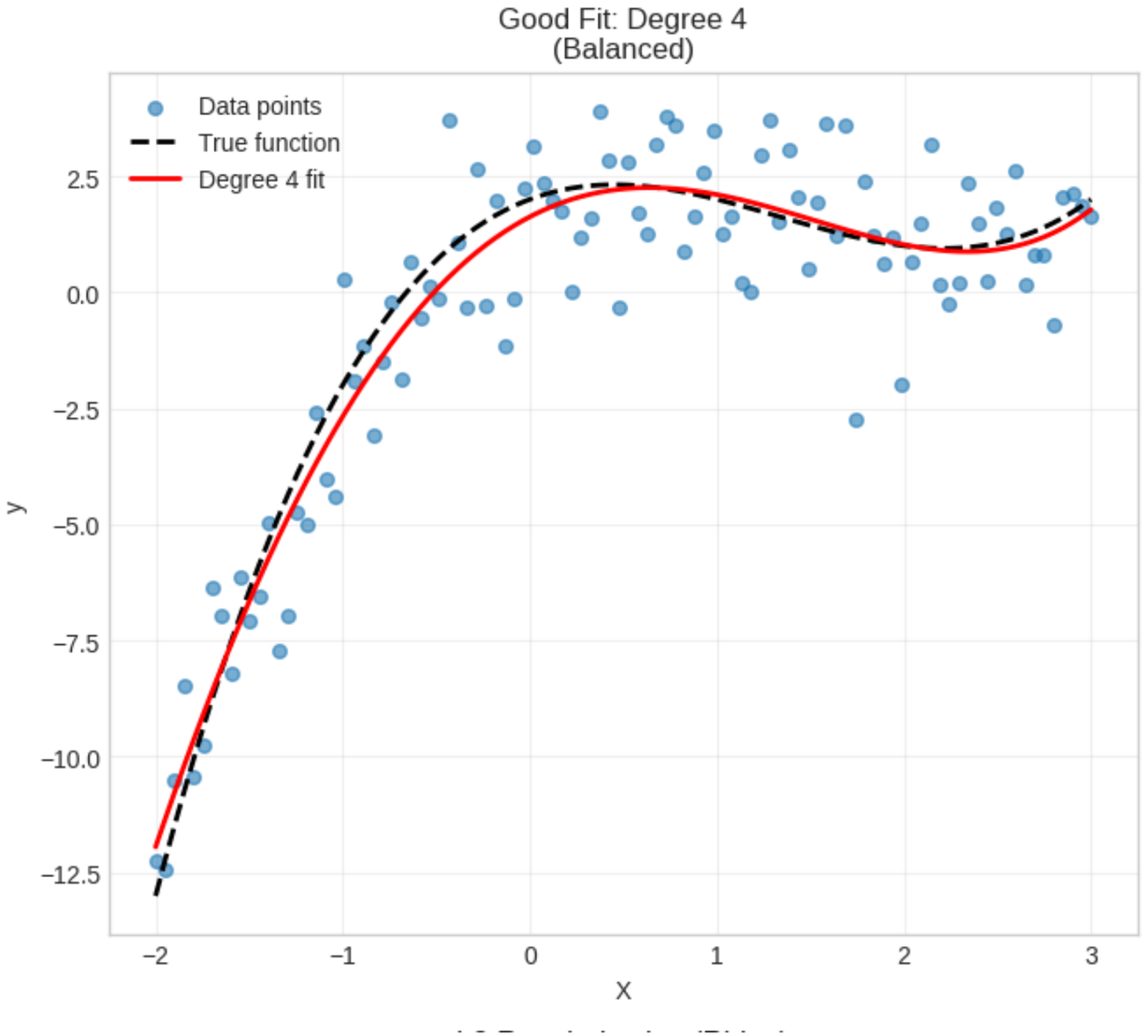
L1 Regularization (Lasso)

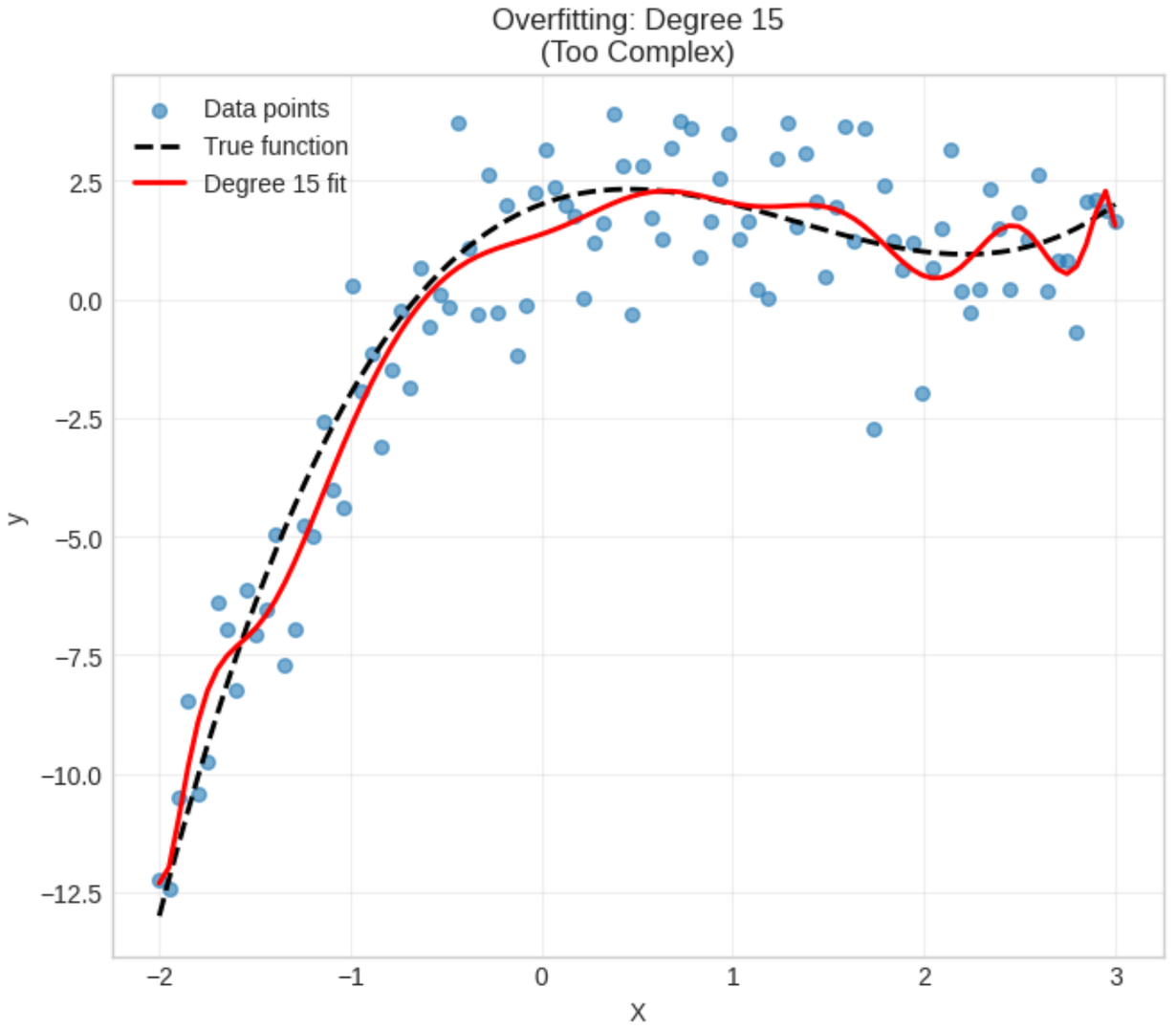
Adds a penalty proportional to the **absolute value** of the coefficients $(\sum|\theta\_i|)$. The geometry of the L1 penalty often forces some coefficients to be exactly zero, effectively performing **automatic feature selection**.

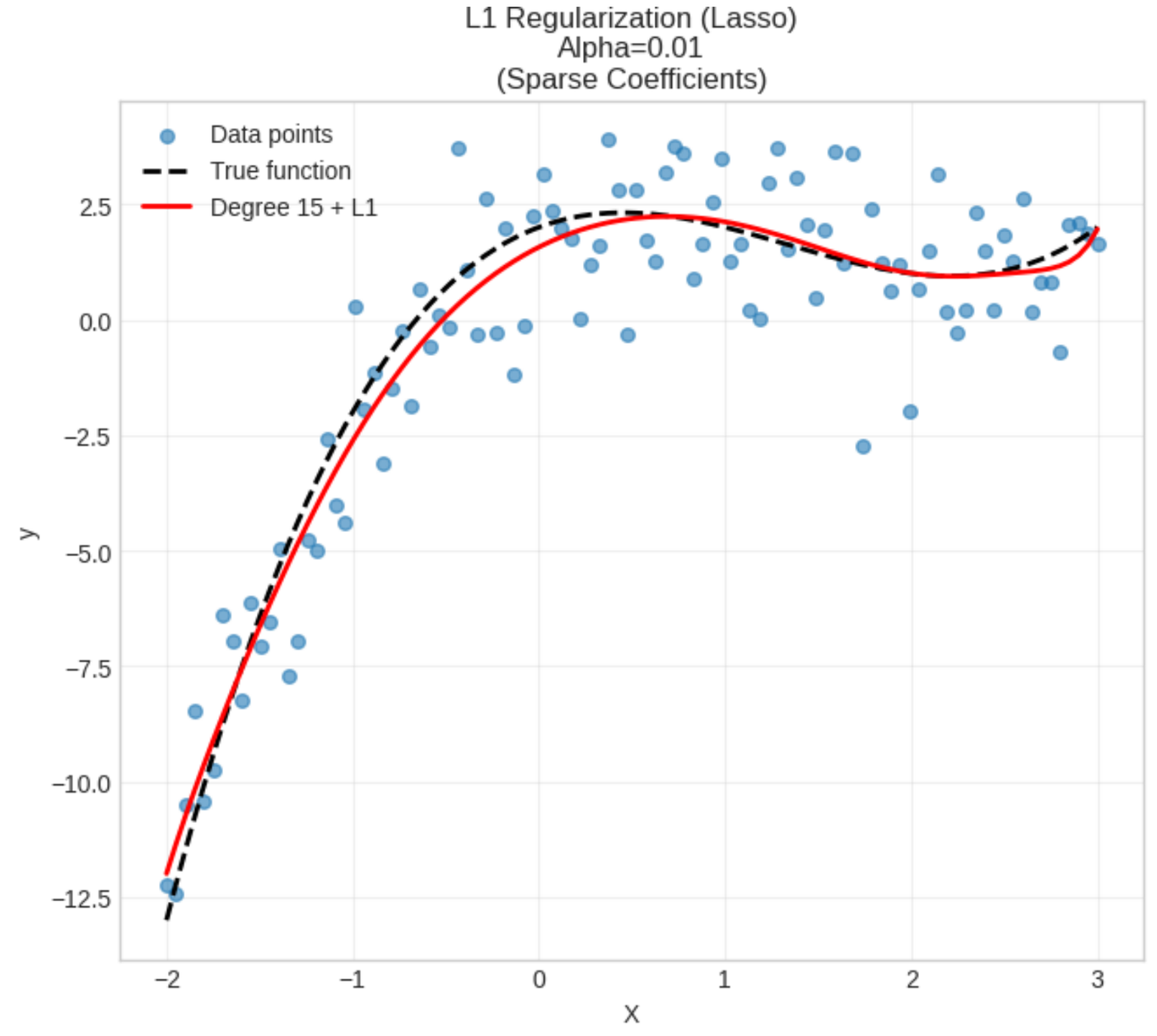
Hyperparameter Alpha λ: The parameter that controls the strength of the regularization penalty.
A higher alpha means a stronger penalty and a simpler model.

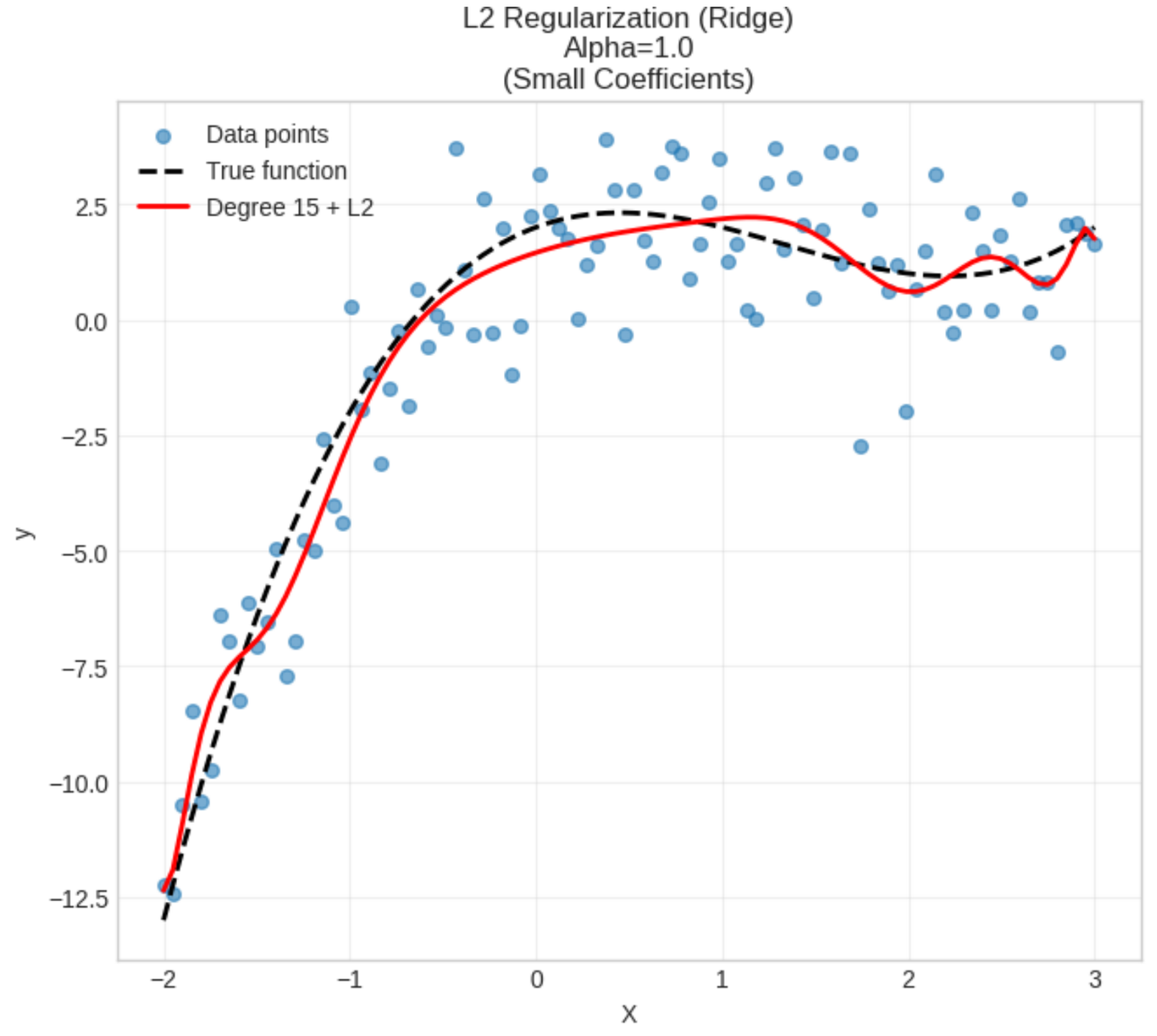

Initial Guess: Degree 1
(Too Simple)
Data points
True function
Linear fit
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−12.5
y
−2
−1
0
1
2
3
X

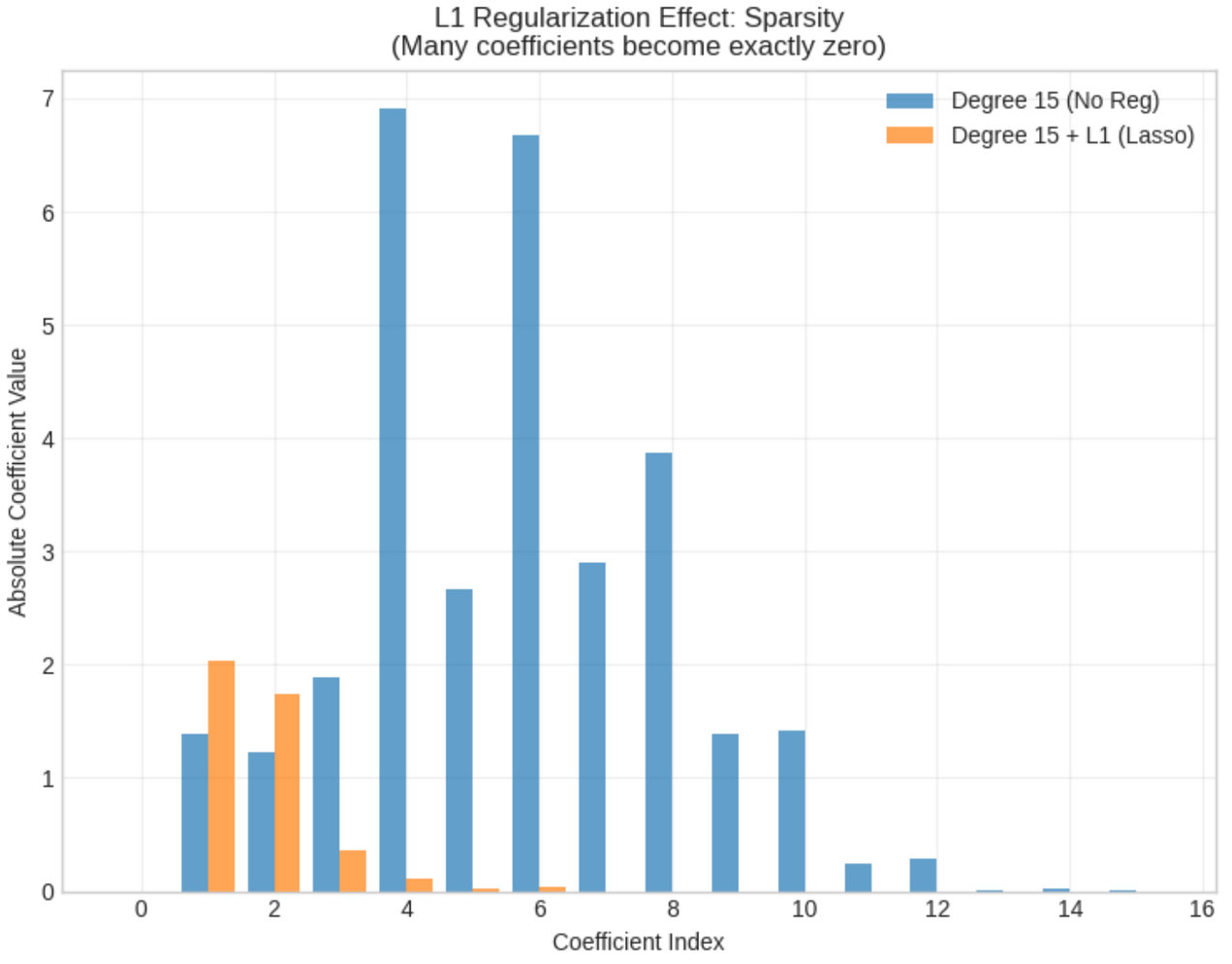

Underfitting: Degree 2
(Too Rigid)
Data points
True function
Degree 2 fit
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−12.5
y
−2
−1
0
1
2
3
X

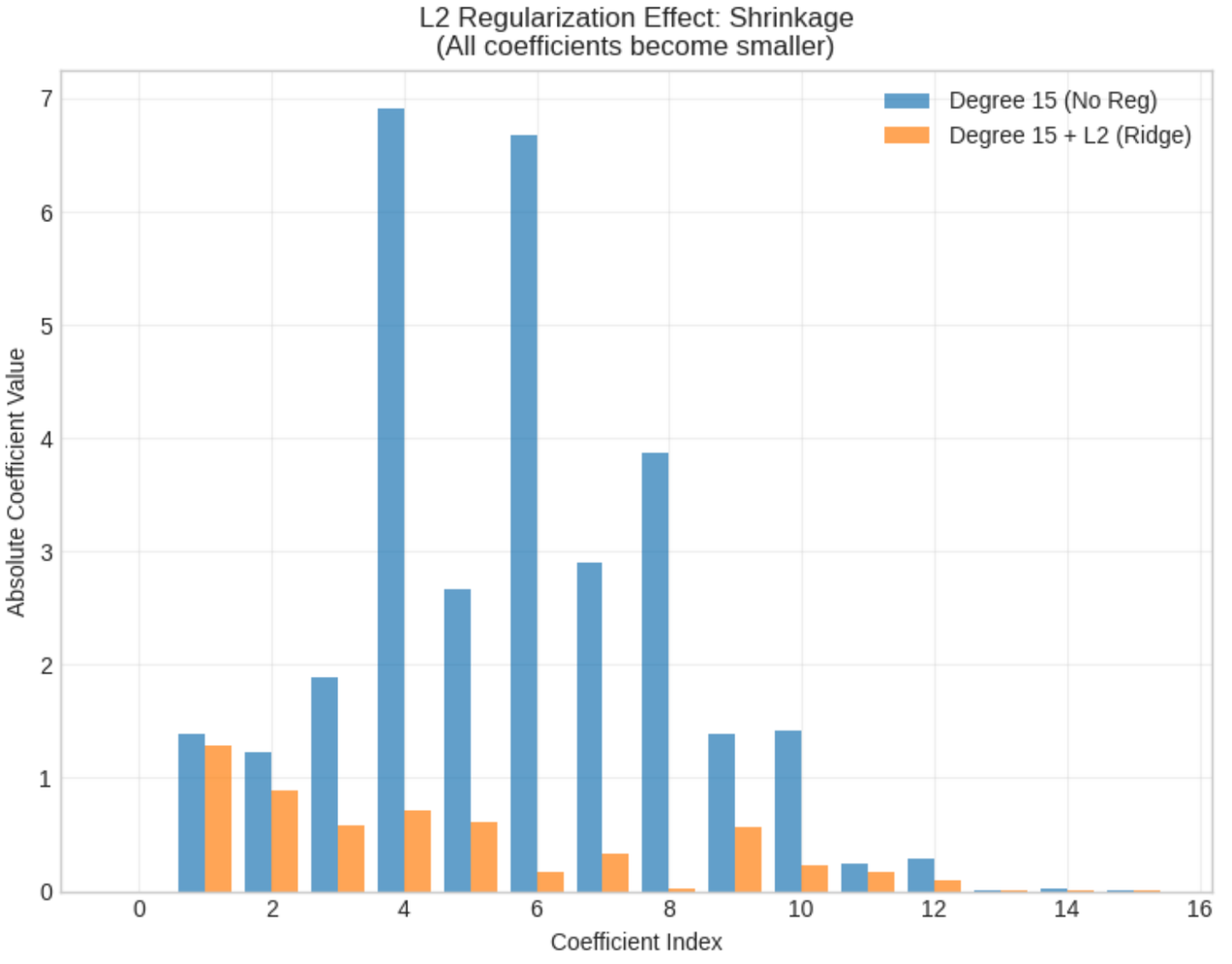

Good Fit: Degree 4
(Balanced)
Data points
True function
Degree 4 fit
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−12.5
y
−2
−1
0
1
2
3
X


Overfitting: Degree 15
(Too Complex)
Data points
True function
Degree 15 fit
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−12.5
y
−2
−1
0
1
2
3
X


L1 Regularization (Lasso)
Alpha=0.01
(Sparse Coefficients)
Data points
True function
Degree 15 + L1
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−12.5
y
−2
−1
0
1
2
3
X


L2 Regularization (Ridge)
Alpha=1.0
(Small Coefficients)
Data points
True function
Degree 15 + L2
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−12.5
y
−2
−1
0
1
2
3
X


L1 Regularization Effect: Sparsity
(Many coefficients become exactly zero)
Degree 15 (No Reg)
Degree 15 + L1 (Lasso)
Absolute Coefficient Value
Coefficient Index
0
1
2
3
4
5
6
7
0
2
4
6
8
10
12
14
16


L2 Regularization Effect: Shrinkage
(All coefficients become smaller)
Degree 15 (No Reg)
Degree 15 + L2 (Ridge)
Absolute Coefficient Value
Coefficient Index
0
1
2
3
4
5
6
7
0
2
4
6
8
10
12
14
16

Logistic Regression: A Linear Classifier

A crucial model for classification, Logistic Regression serves as a baseline for measuring performance in binary (two-class) problems.

Estimating Probability

Unlike Linear Regression, which predicts a value, Logistic Regression models the probability $P(y = 1|X)$ that an observation belongs to the positive class.

The Sigmoid Function

The linear combination of features $(\theta^T X)$is passed through the logistic (sigmoid) function, which squashes the output into the range [0, 1].

Decision Boundary

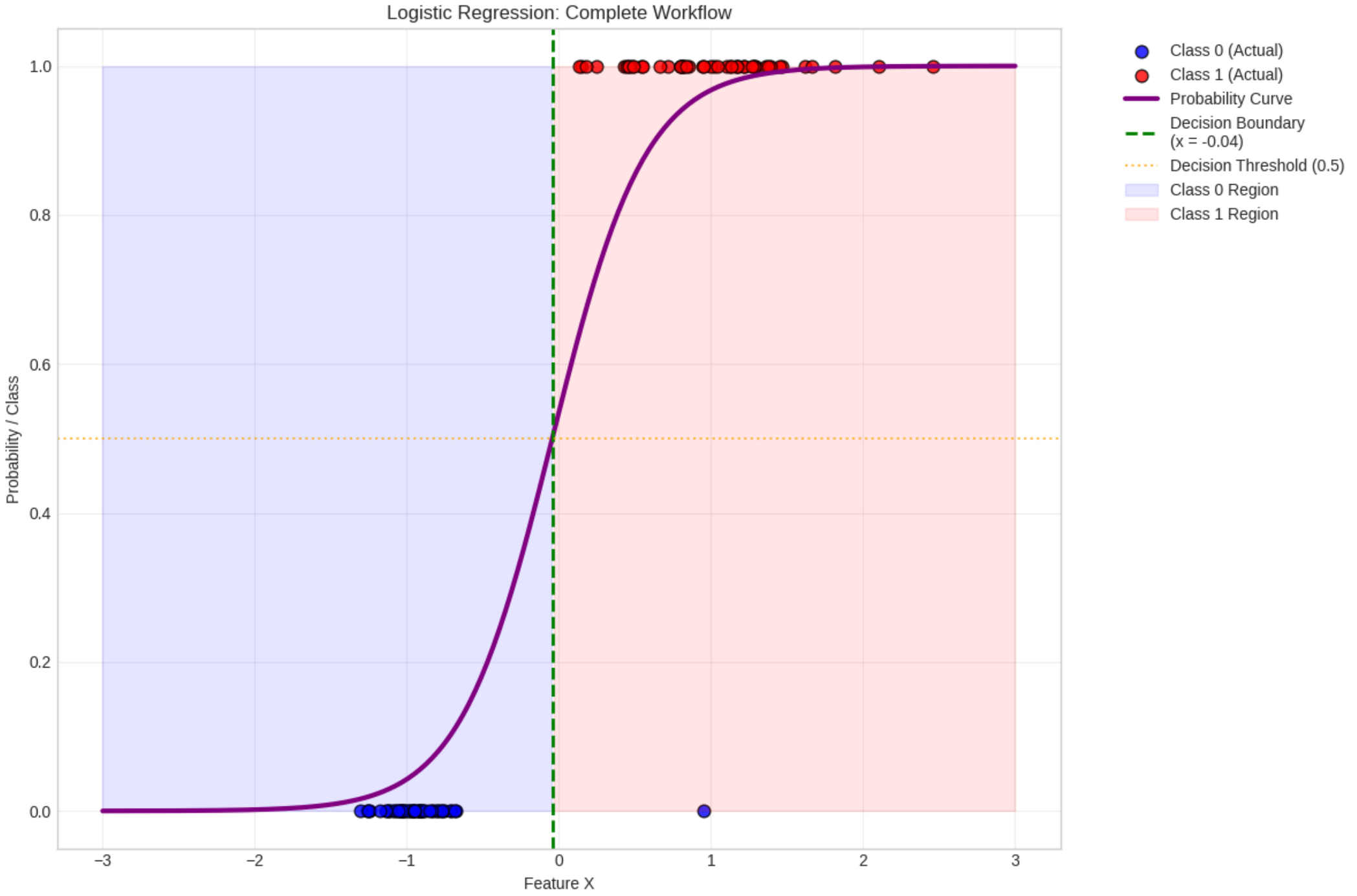
A prediction is made by applying a threshold (e.g., 0.5). If the calculated probability is ≥ 0.5, the model predicts class 1; otherwise, class 0. The boundary itself remains linear.

The Logistic Function

$$P(y=1|X) = \frac{1}{1+e^{-(\theta^T X)}}$$

The model parameters (θ) are typically learned using **Maximum Likelihood Estimation (MLE)**, which seeks to maximise the likelihood of observing the training data given the model parameters.

> The greatest strength of Logistic Regression lies in its interpretability: the coefficients θ_i can be converted into odds ratios to understand the marginal impact of each feature on the likelihood of the positive outcome.


Logistic Regression: Complete Workflow
Probability / Class
Feature X
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3
Class 0 (Actual)
Class 1 (Actual)
Probability Curve
Decision Boundary
(x = -0.04)
Decision Threshold (0.5)
Class 0 Region
Class 1 Region

Evaluating Model Performance: Core Regression Metrics

Objective evaluation metrics are essential for comparing models, tuning hyperparameters, and understanding prediction quality. We focus on three fundamental metrics for continuous prediction tasks.

1

Mean Absolute Error (MAE)

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- Measures the average magnitude of error, regardless of direction.
- Units are the same as the target variable (highly interpretable).
- Less sensitive to outliers compared to MSE.

2

Mean Squared Error (MSE)

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

- Measures the average squared difference between true and predicted values.
- Strongly penalises large errors (outliers have a significant impact).
- Units are squared, making it less intuitive for business interpretation.

3

Root Mean Squared Error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- The square root of MSE, returning the metric to the original units.
- Maintains the strong outlier penalty from MSE while retaining unit interpretability.
- The most commonly reported metric in many regression contexts.

R-Squared and Adjusted R-Squared: Explaining Variance

Beyond simple error magnitude, R-squared metrics quantify how well the model accounts for the variability in the observed data.

R-Squared (R^2)

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

The **Coefficient of Determination** measures the proportion of the variance in the dependent variable that is predictable from the independent variables.

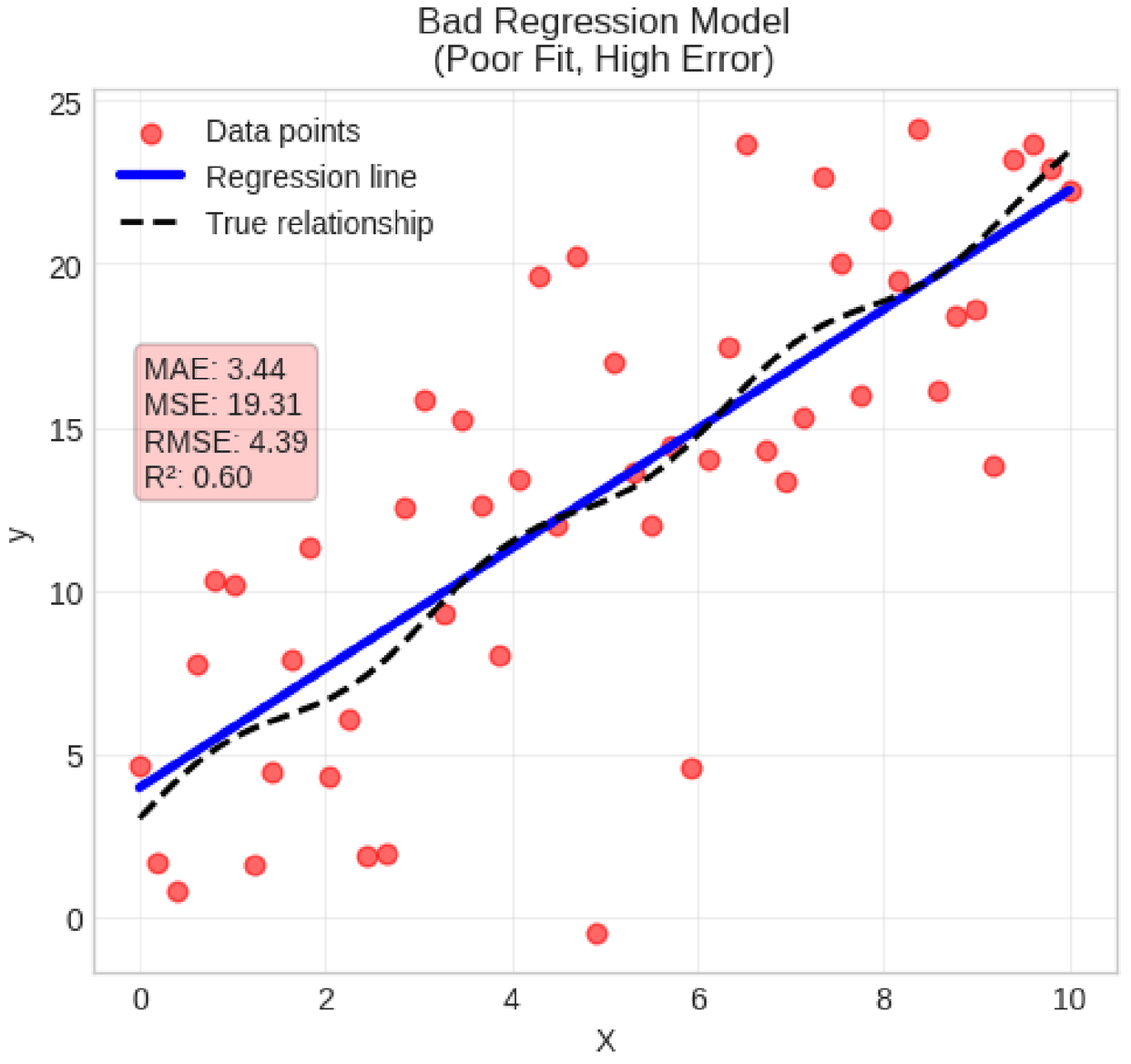
- **$R^2 = 1$:** Perfect model fit; all variance explained.
- **$R^2 = 0$:** Model performs no better than simply predicting the mean $(\bar{y})$.
- **$R^2 < 0$:** Model is performing worse than the simple mean baseline.

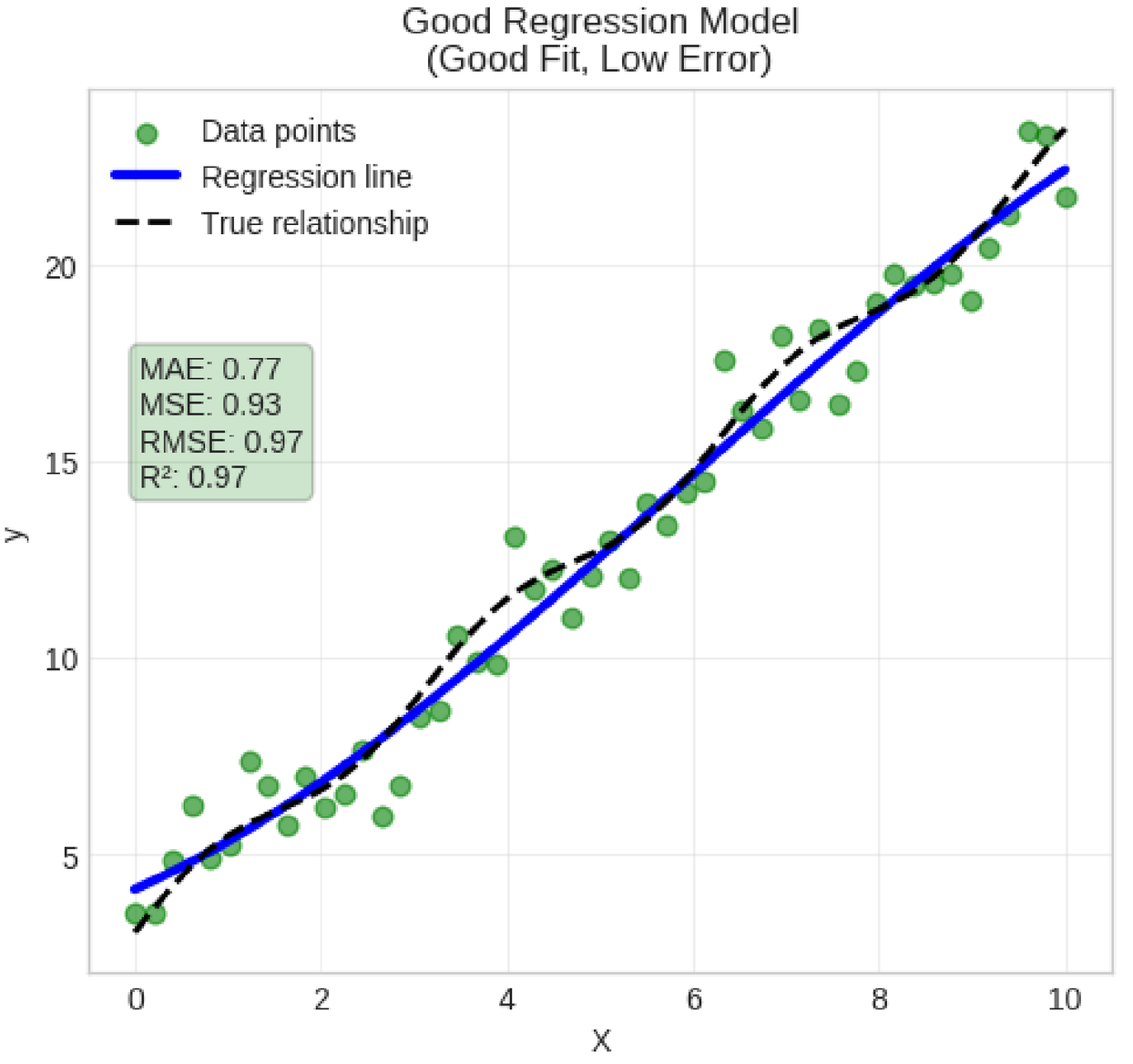
Adjusted R-Squared (R^2_{adj})

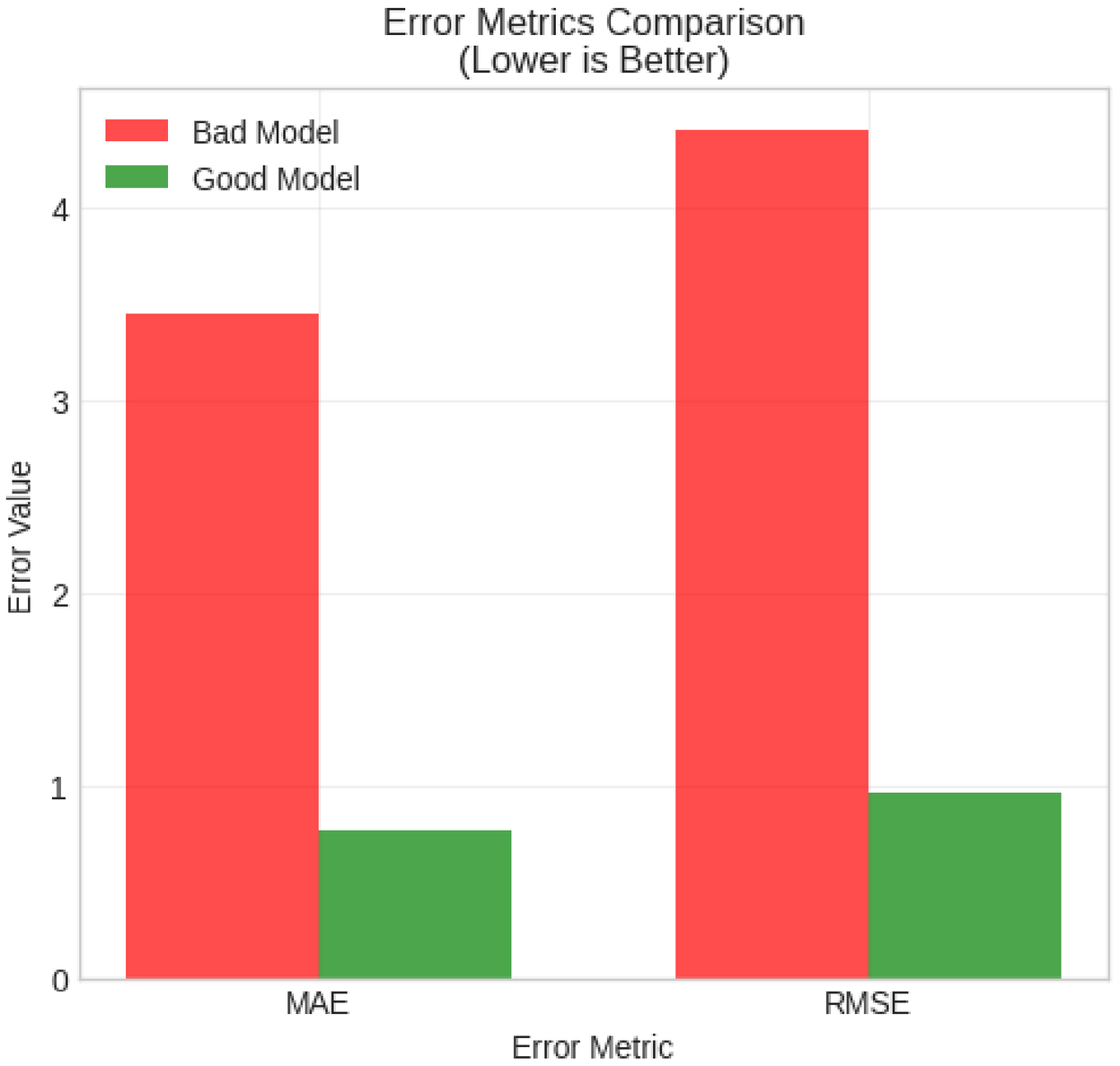
$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

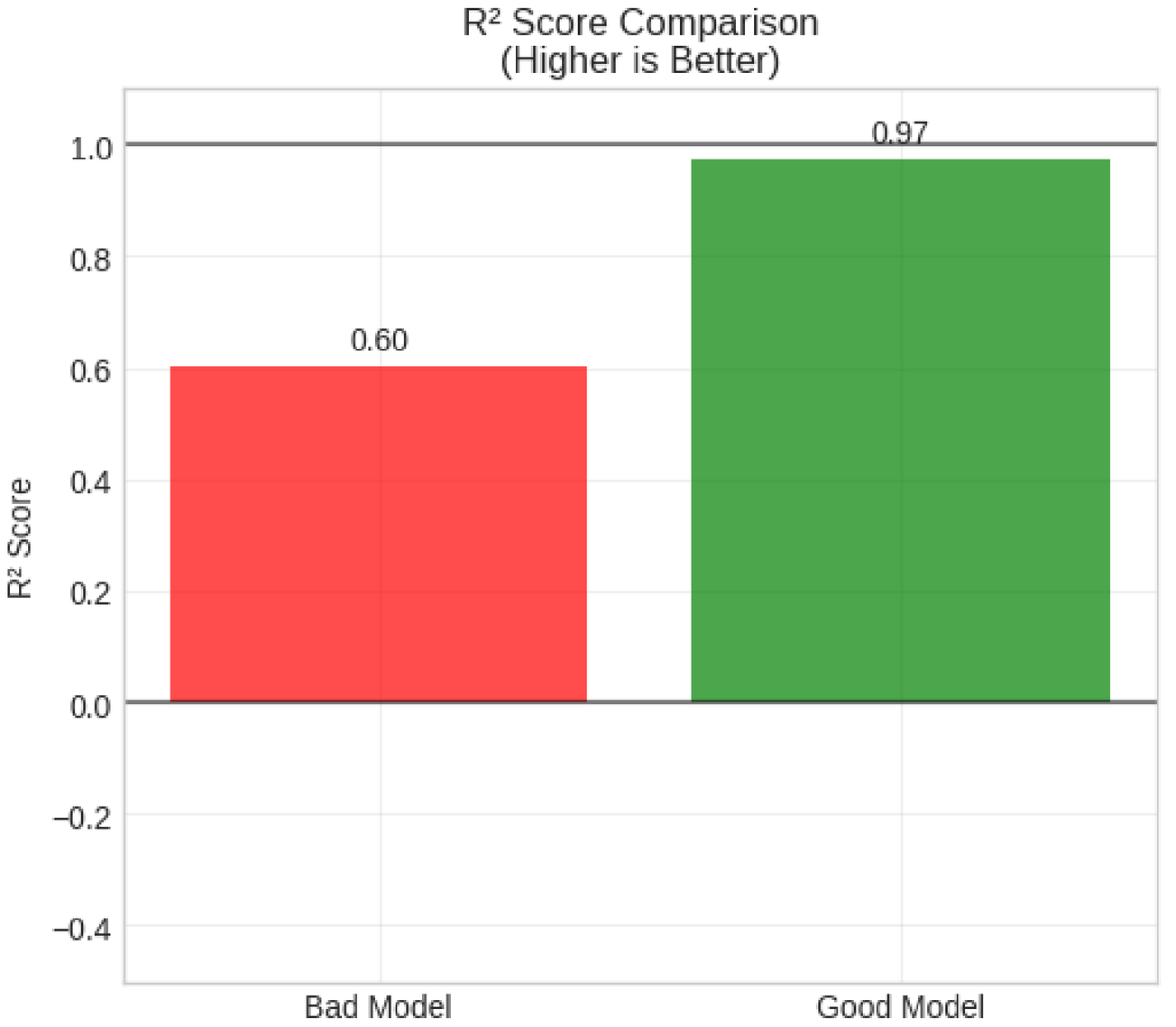
Standard R^2 always increases or stays the same when new features (p) are added, even if they are irrelevant. The Adjusted R^2 corrects for this by penalising the inclusion of extraneous variables.

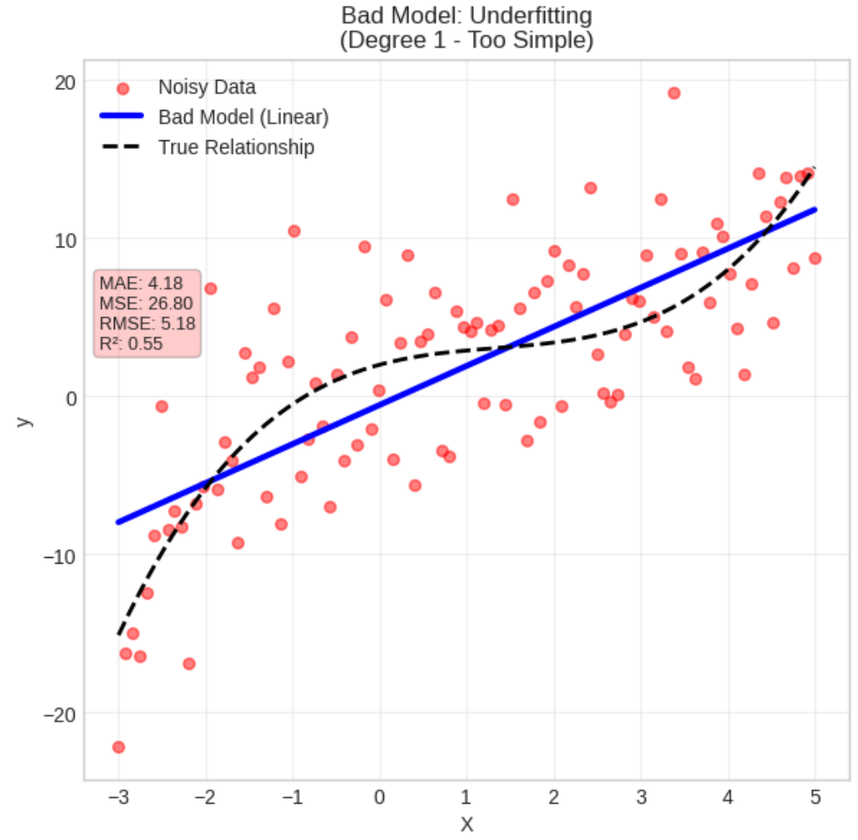
> The Adjusted R^2 is superior for comparing models that use a different number of predictor variables, as it reflects the model's actual generalisation capability.

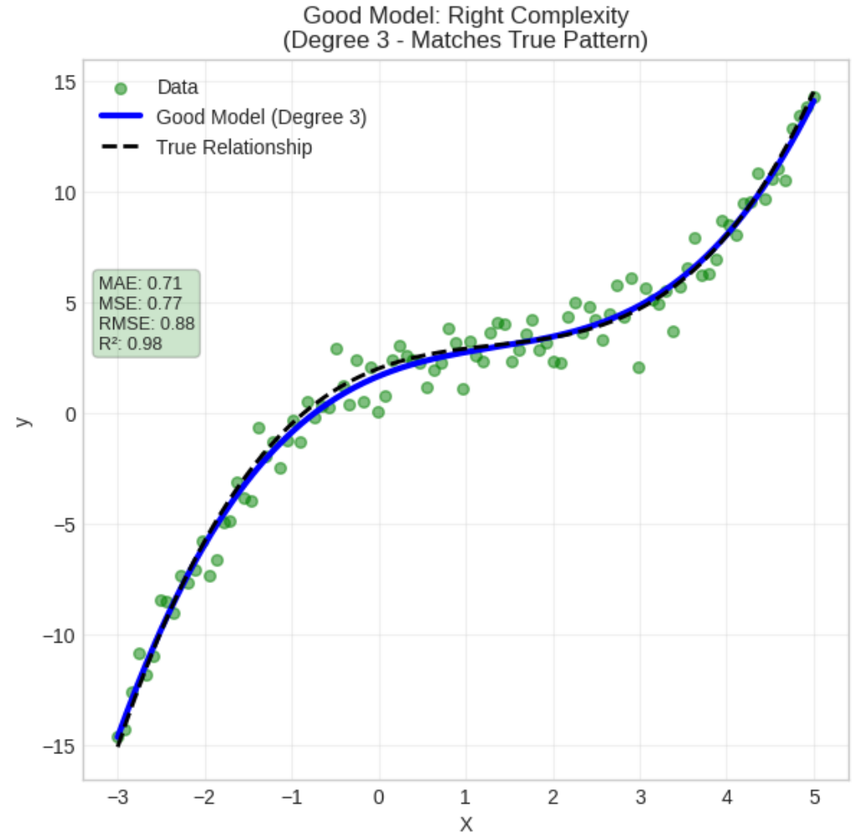

Bad Regression Model
(Poor Fit, High Error)
Data points
Regression line
True relationship
MAE: 3.44
MSE: 19.31
RMSE: 4.39
R²: 0.60
X
y

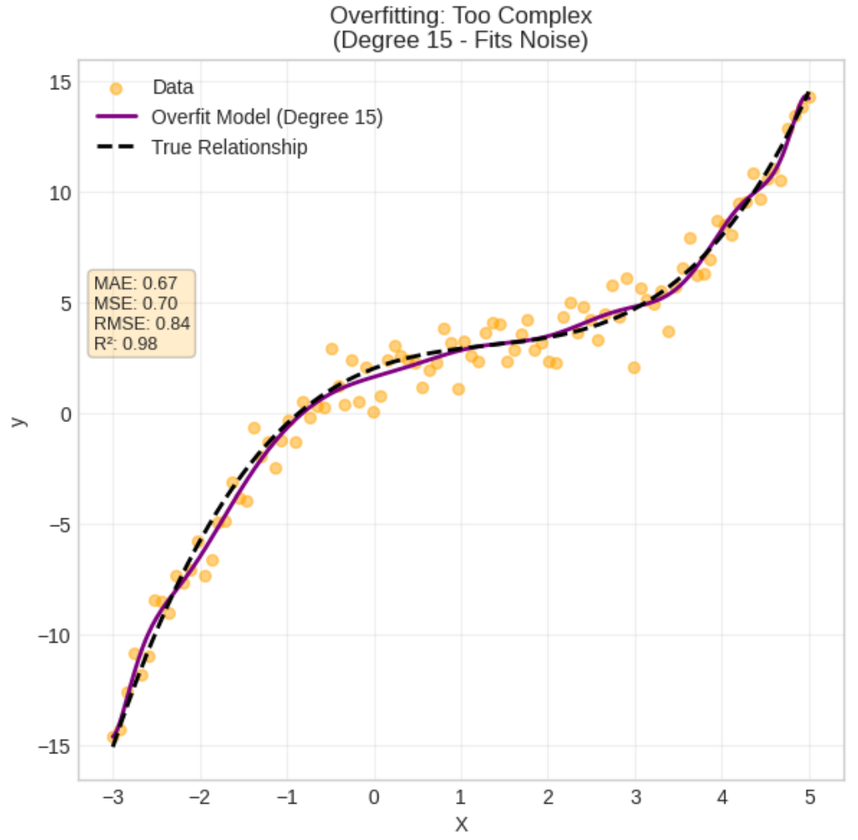

Good Regression Model
(Good Fit, Low Error)
Data points
Regression line
True relationship
MAE: 0.77
MSE: 0.93
RMSE: 0.97
R²: 0.97
X
y


Error Metrics Comparison
(Lower is Better)
Bad Model
Good Model
Error Value
0
1
2
3
4
MAE
RMSE
Error Metric


R² Score Comparison
(Higher is Better)
R² Score
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
0.60
0.97
Bad Model
Good Model


Bad Model: Underfitting
(Degree 1 - Too Simple)
Noisy Data
Bad Model (Linear)
True Relationship
MAE: 4.18
MSE: 26.80
RMSE: 5.18
R²: 0.55
X
y


Good Model: Right Complexity
(Degree 3 - Matches True Pattern)
Data
Good Model (Degree 3)
True Relationship
MAE: 0.71
MSE: 0.77
RMSE: 0.88
R²: 0.98
X
y


Overfitting: Too Complex
(Degree 15 - Fits Noise)
Data
Overfit Model (Degree 15)
True Relationship
MAE: 0.67
MSE: 0.70
RMSE: 0.84
R²: 0.98
X
y

Summary: A Comparative View of Regression Metrics

Selecting the appropriate metric depends on the problem context, particularly whether large errors should be heavily penalised.

Metric	Penalises Large Errors	Same Unit as Target	Sensitive to Outliers	Typical Range
MAE	No	Yes	Moderate	≥ 0
MSE	Yes	No (Squared)	High	≥ 0
$RMSE$	Yes	Yes	High	≥ 0
R^2	—	—	Moderate	$(-\infty, 1]$
$Adjusted\ R^2$	—	—	Moderate	$(-\infty, 1]$

Final Takeaway

For highly interpretable results, MAE or RMSE are preferred. For model comparison or focusing on predictive power (variance explained), use Adjusted R^2. Always report multiple metrics for a complete picture.