Salary Prediction: Understanding Linear Regression

This presentation explores how linear regression can be used to accurately estimate a person's salary, a valuable tool for hiring, budgeting, and ensuring fair compensation.

Building a Predictive Model

The Goal

To build a model that accurately estimates a person's salary.

Our Tool

Linear Regression, a powerful statistical method to find relationships between factors and salary.

Identifying Key Factors

Dependent Variable (Y)

The value we want to predict: **Salary**.

Independent Variable (X)

The primary factor for prediction: **Years of Experience**.

What is Linear Regression?

Linear Regression models the relationship between variables by fitting a straight line to observed data.

Understanding Influence

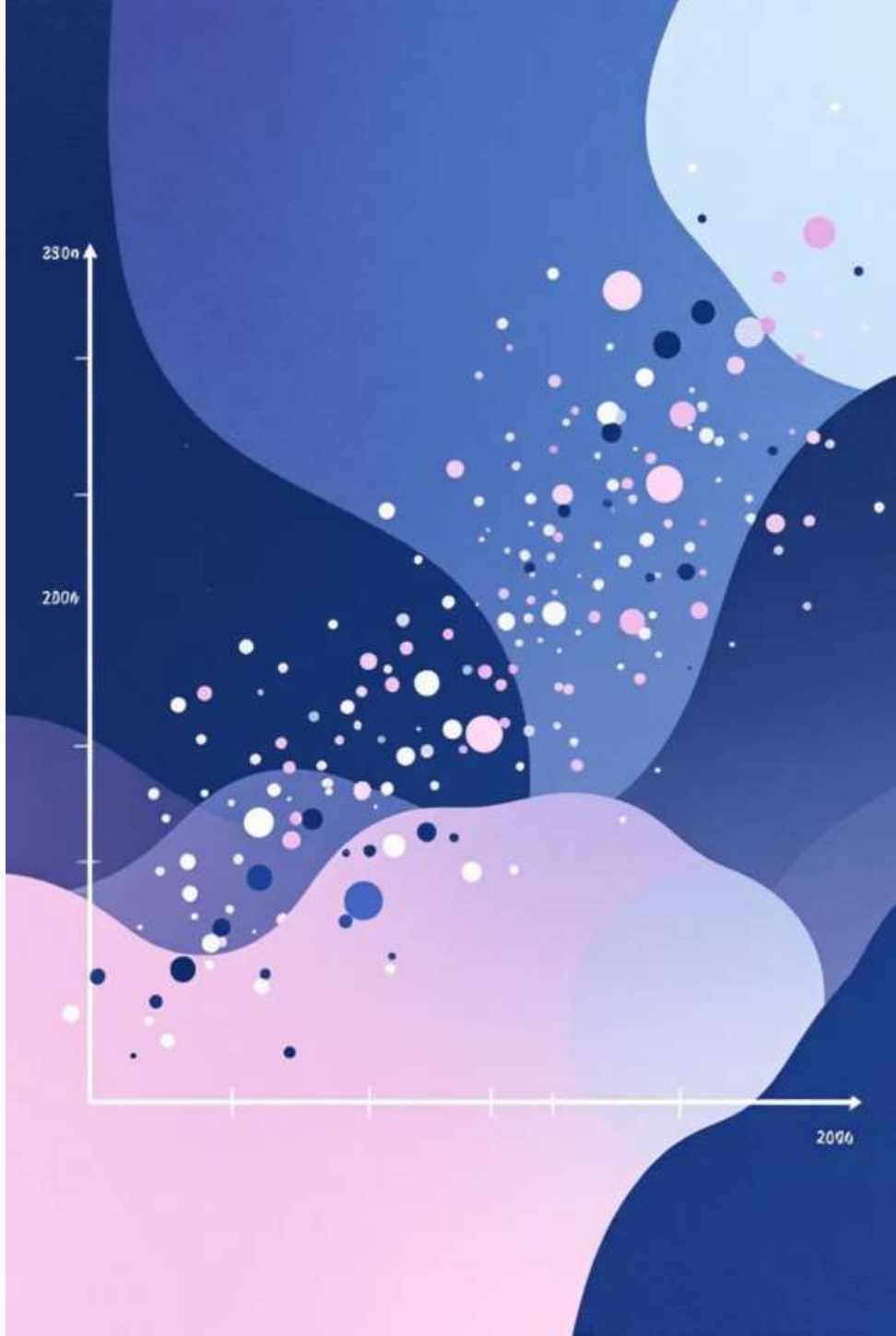
It shows how an independent variable (e.g., experience) influences a dependent variable (e.g., salary).

Best-Fit Line

The goal is to find the single "best-fit line" that most accurately describes the data.

Making Predictions

This line is then used to make predictions for new, unseen data points.



The Math Explained

$$Y = b_0 + b_1X + \varepsilon$$

1

Y: Predicted Value

The salary we want to find.

2

b_0: Intercept

Starting salary with zero experience.

3

b_1: Slope

Change in salary for each additional year of experience.

4

X: Independent Variable

Number of years of experience.

5

ε: Error

The part of the salary not explained by experience.

Finding the Best-Fit Line: OLS

The model finds the best line using **Ordinary Least Squares (OLS)**.

OLS draws a line that **minimizes the total squared distance** between each data point and the line itself. These distances are called "residuals" or errors. OLS squares them to remove negative signs and finds the line with the smallest possible sum of these squared errors.

Expanding the Model

Simple Linear Regression

Uses only **one** independent variable (e.g., Experience → Salary).

Multiple Linear Regression

Uses **two or more** independent variables (e.g., Experience + Education → Salary). This is generally more accurate.

Is the Model Any Good?

R^2

R-squared

Percentage of salary change explained by model factors. Higher is better (e.g., 0.8 means 80% explained).

RMSE

Root Mean Square Error

Average error of prediction in original units (e.g., dollars). Lower is better (e.g., $5,000 off).

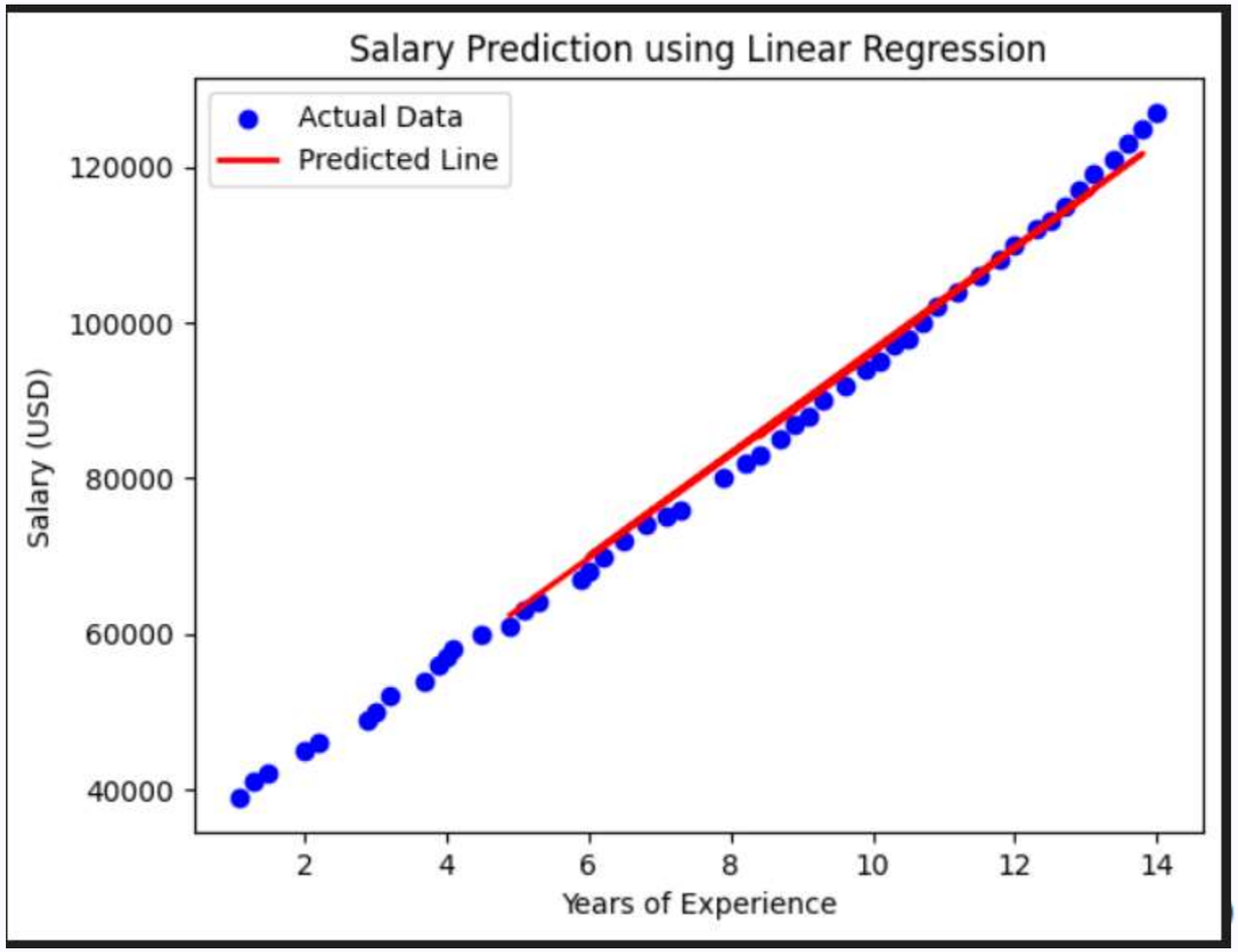
Important Assumptions & Pitfalls

Model Assumptions

- Linearity: Relationship between variables is linear.
- Independence: Errors are independent.
- Homoscedasticity: Errors are evenly spread.
- Normality: Errors are normally distributed.

Common Pitfall: Overfitting

When a model learns the noise in training data, failing to predict new data. A core data science concept.



Key Takeaways

Best-Fit Line

Linear Regression predicts salary from factors like experience.

Simple vs. Multiple

Use simple (one factor) or multiple (many factors) models for accuracy.

Evaluate

Use R-squared and RMSE to check accuracy and reliability.

Questions?

Thank you for your attention.