[Regression]

Simple Linear Regression

Prof. Sandeep Kumar

Yardi School of Artificial Intelligence
Department of Electrical Engineering
Indian Institute of Technology Delhi
New Delhi, India

ksandeep@iitd.ac.in

Prof. Manabendra Saharia

Yardi School of Artificial Intelligence
Department of Civil Engineering
Indian Institute of Technology Delhi
New Delhi, India

msaharia@iitd.ac.in

https://aimlindustry.iitd.ac.in/

By, Yardi School of Artificial Intelligence (ScAI)
Indian Institute of Technology Delhi

भारतीय प्रौद्योगिकी संस्थान दिल्ली
Indian Institute of Technology Delhi

Outline of the talk

01 Introduction to Machine Learning Paradigms

- Supervised vs. Unsupervised Learning
- Regression vs Correlation
- Key Terminologies

02 Fundamentals of Regression Analysis

- Simple Linear Regression
- Coefficients, Residuals and Errors in Regression
- Predicting Sales from Advertisement Data

03 Multiple Linear Regression

- Expanding Linear Regression to Multiple Variables
- Variable Interaction and Synergy Effects

04 Advanced Techniques and Nonlinear Regression

- Polynomial Regression – Extending Linear Regression to handle Nonlinear Relationships
- Model Selection and Regularization

Learning Outcomes from this Lecture

→ **Simple Linear Regression Fundamentals:** Understand the formulation of simple linear regression, including its assumptions and mathematical underpinnings.

→ **Coefficient Estimation:** Learn how least squares method is used for estimating regression coefficients.

→ **Critical Evaluation of Relationships:** Evaluate the strength, direction, and nature (linear or nonlinear) of relationships between variables.

→ **Predictive Modeling:** Use regression models to make predictions and assess their accuracy.

Hands-On Demonstration

Python demo showcasing Simple Linear Regression and the Least Squares Method, featuring a) built-in functions, and b) detailed, step-by-step implementation. Using an advertising budget and sales dataset.

Outline of the talk

01 Learning Paradigms

- Supervised vs. Unsupervised Learning
- Regression vs Correlation
- Key Terminologies

02 Fundamentals of Regression Analysis

- Simple Linear Regression
- Coefficients, Residuals and Errors in Regression
- Predicting Sales from Advertisement Data

03 Multiple Linear Regression

- Expanding Linear Regression to Multiple Variables
- Variable Interaction and Synergy Effects

04 Advanced Techniques and Nonlinear Regression

- Polynomial Regression – Extending Linear Regression to handle Nonlinear Relationships
- Model Selection and Regularization

The Machine Learning Paradigm

Example: Segregation of fruits

- **Task:**
  - Suppose you have a basket filled with fresh fruits, including apples, bananas, cherries, and grapes. Your task is to arrange the fruits by type, gathering all the apples together, all the bananas together, all the cherries together, and all the grapes together.

Supervised vs Unsupervised Learning

Example: Segregation of fruits

- **Task:**
  - Suppose you have a basket filled with fresh fruits, including apples, bananas, cherries, and grapes. Your task is to arrange the fruits by type, gathering all the apples together, all the bananas together, all the cherries together, and all the grapes together.

Supervised vs Unsupervised Learning

Example: Segregation of fruits

- **Task:**
  - Suppose you have a basket filled with fresh fruits, including apples, bananas, cherries, and grapes. Your task is to arrange the fruits by type, gathering all the apples together, all the bananas together, all the cherries together, and all the grapes together.

- **Case 1:**
  - You already know: Shape (parametrize shape?), Color
  - Train data: Pre-classified data
  - Goal: Learn from the pre-classified data and predict on new unclassified fruits.
  - This type of learning is called as ***supervised learning.***

Supervised vs Unsupervised Learning

Example: Segregation of fruits

- **Task:**
  - Suppose you have a basket filled with fresh fruits, including apples, bananas, cherries, and grapes. Your task is to arrange the fruits by type, gathering all the apples together, all the bananas together, all the cherries together, and all the grapes together.

- **Case 2:**
  - In this case, *you know nothing about* the fruits, you are seeing them for the first time!
  - How will you arrange fruits of the same type together?
  - One approach is to consider various characteristics of a fruit and divide them based on that.
  - ... ...
  - Suppose you divide the fruits based on **color** first.
  - Now you take another physical characteristic, say, ***size***. The grouping will then be:
  - ... ...

Supervised vs Unsupervised Learning

Example: Segregation of fruits

- **Task:**
  - Suppose you have a basket filled with fresh fruits, including apples, bananas, cherries, and grapes. Your task is to arrange the fruits by type, gathering all the apples together, all the bananas together, all the cherries together, and all the grapes together.

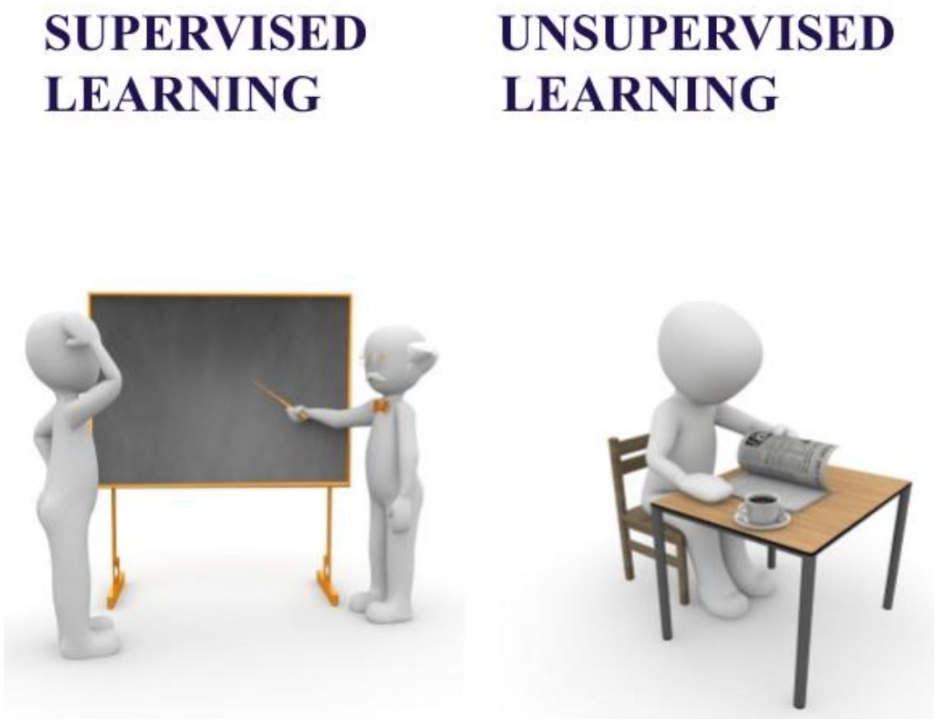
- **Case 2:**
  - ... ...

  - Red color and big size: Apple
  - Red color and small size: Cherry
  - Green color and big Size: Banana
  - Green color and small Size: Grapes
  - This type of learning is ***unsupervised learning***

Supervised vs Unsupervised Learning

Supervised vs Unsupervised Learning

- In supervised learning, the desired outputs are provided which are used to train the machine
- In unsupervised learning no desired outputs are provided, instead the data is analyzed and studied through clustering, mining associations, reduce dimensionality, etc. into different classes

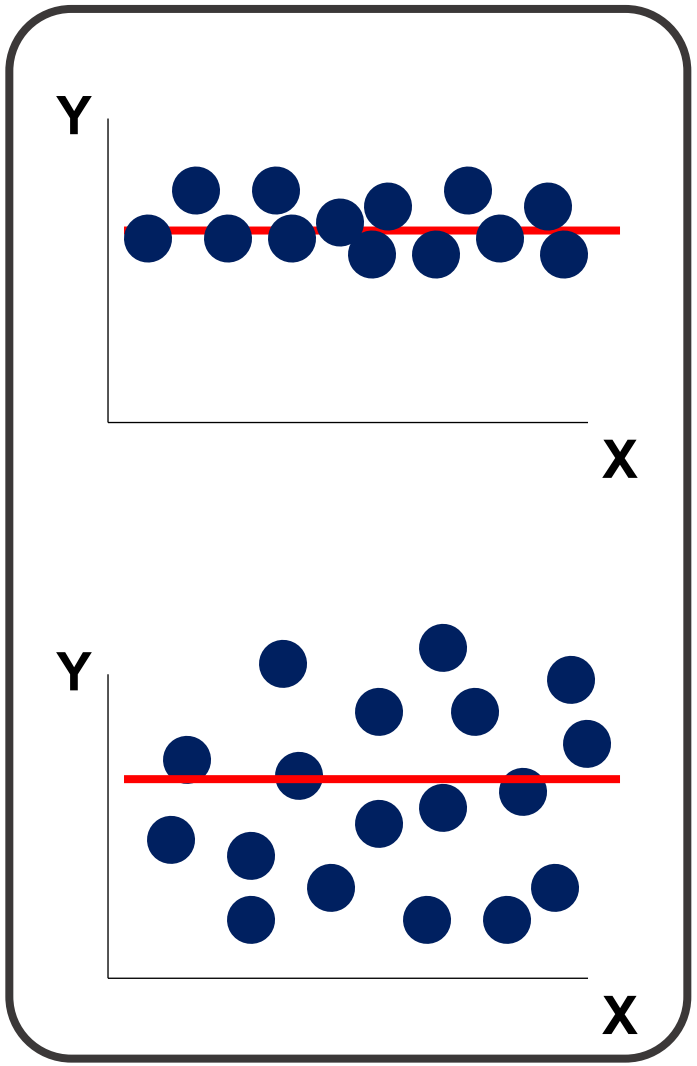


Source

Regression and Correlation

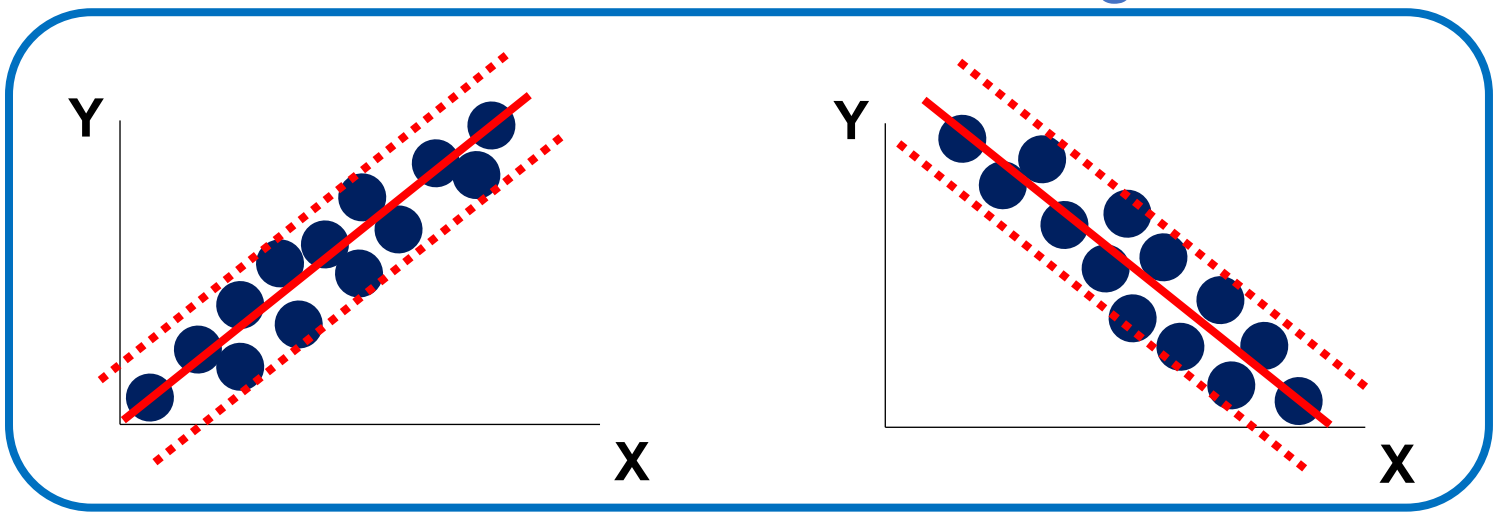
CORRELATION	REGRESSION
• Correlation simply describes the strength and direction of the relationship between two variables • There is no distinction between the input and output variables, and both are treated as equal in importance in describing the relationship between them	• Regression seeks to determine the direction of the relationship and predict the value of one variable based on the value of another • One variable is treated as the dependent variable (the outcome variable) and the other variable is treated as the independent variable (the predictor variable)

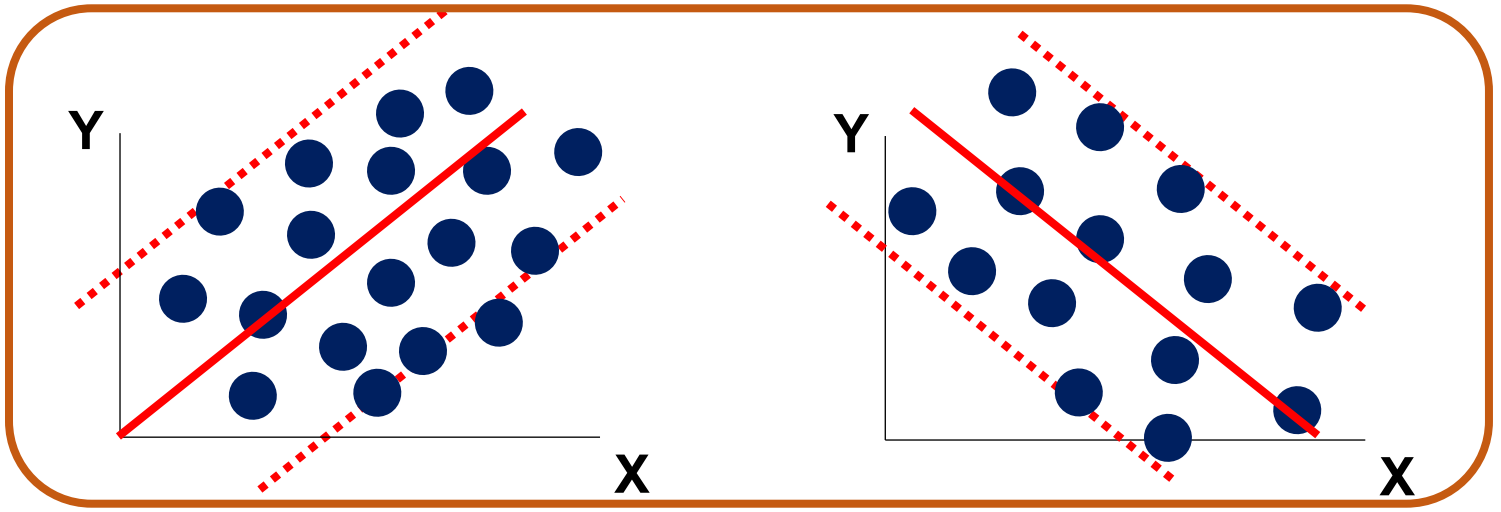
Relationship between Input and Output



No Relation

Strong Relation

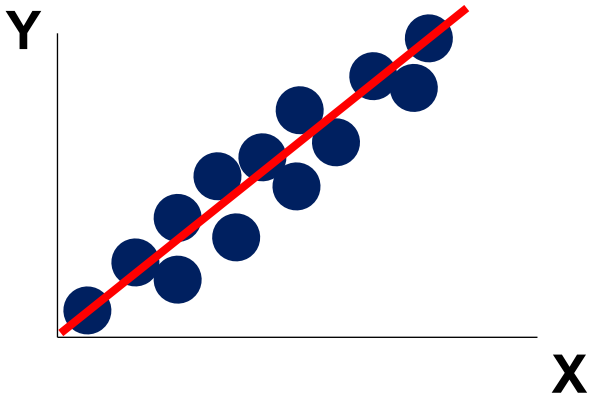


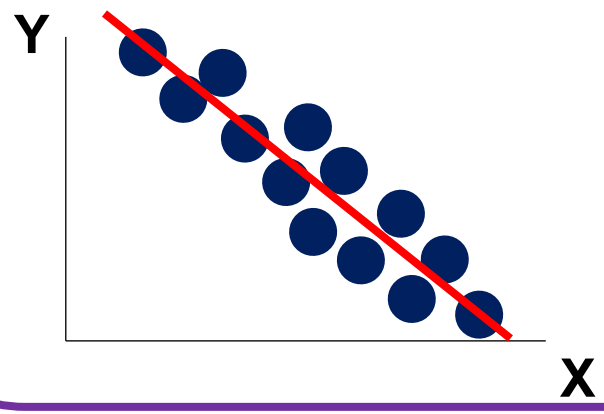


Weak Relation

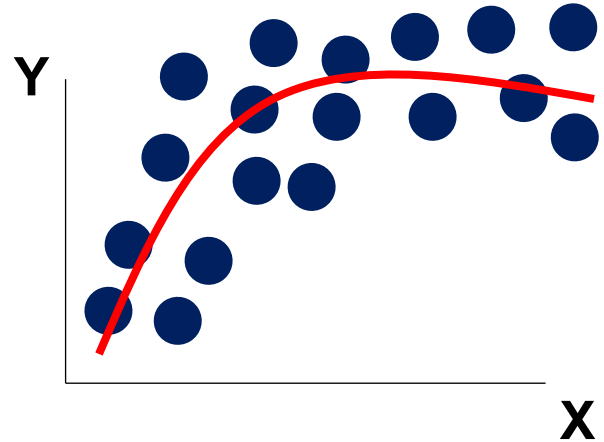
Relationship between Input and Output

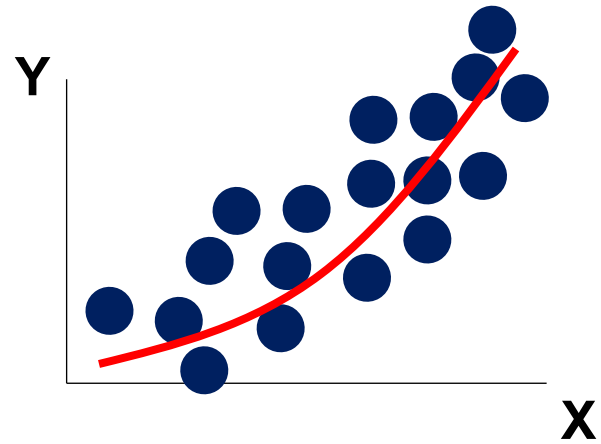
Linear Relation





Nonlinear Relation





Terminologies used in Regression

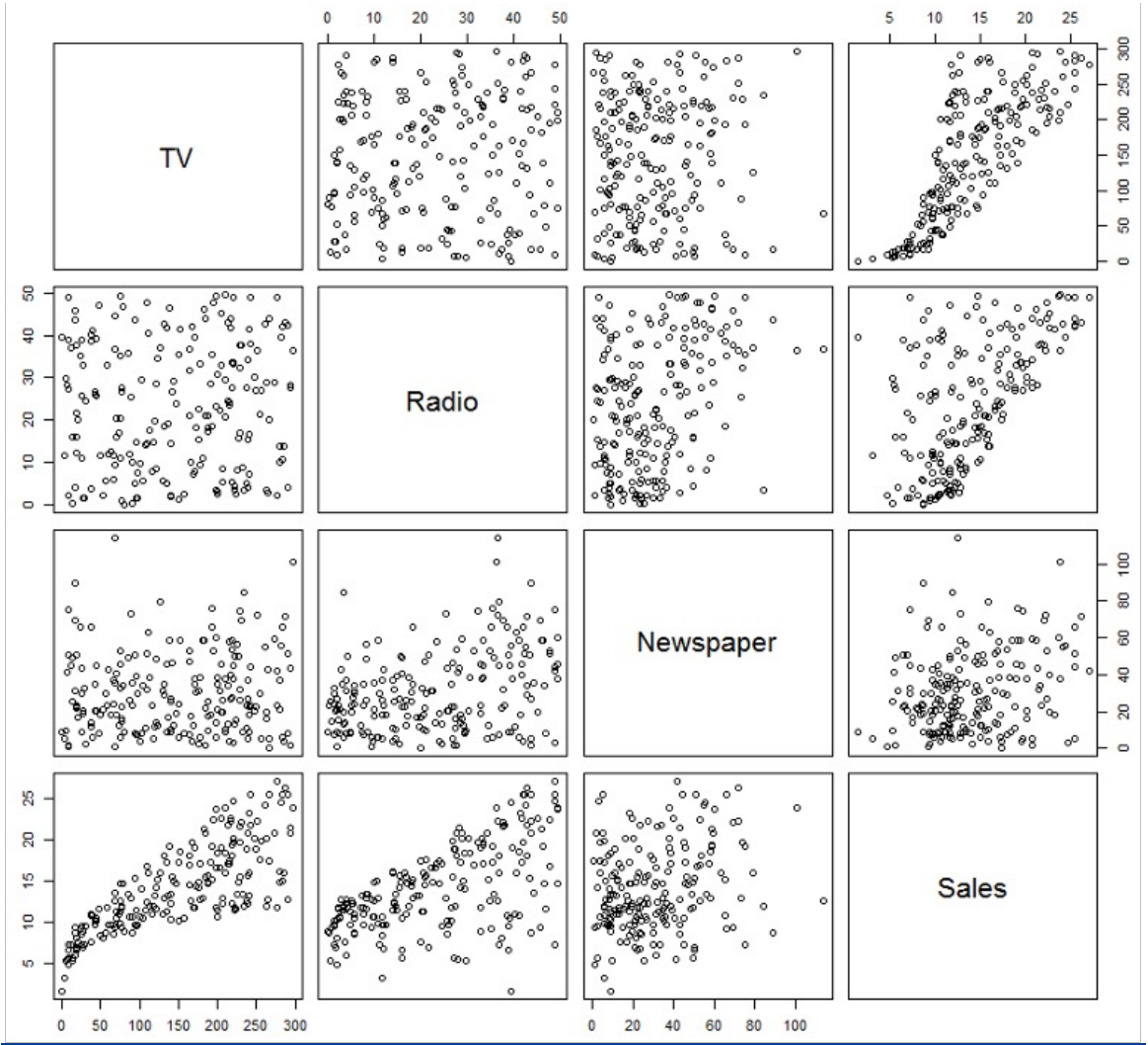
Terminologies

- **Dependent variable (Outcome/Target):**
  - a dependent variable is a variable that is being studied or measured and is expected to be affected by one or more independent variables. It is often denoted by "Y" and is the outcome or response variable in a study.

- **Independent variable (Predictor/Explanatory):**
  - an independent variable is a variable that is being manipulated or controlled by the researcher in a study. It is often denoted by "X" and is the variable that is hypothesized to cause changes in the dependent variable.

- **Regression analysis is used to:**
  - Predict the value of a dependent variable based on the value of at least one independent variable
  - Explain the impact of changes in an independent variable on the dependent variable

Advertisement Sales Data



Scatter plot for the advertisement data

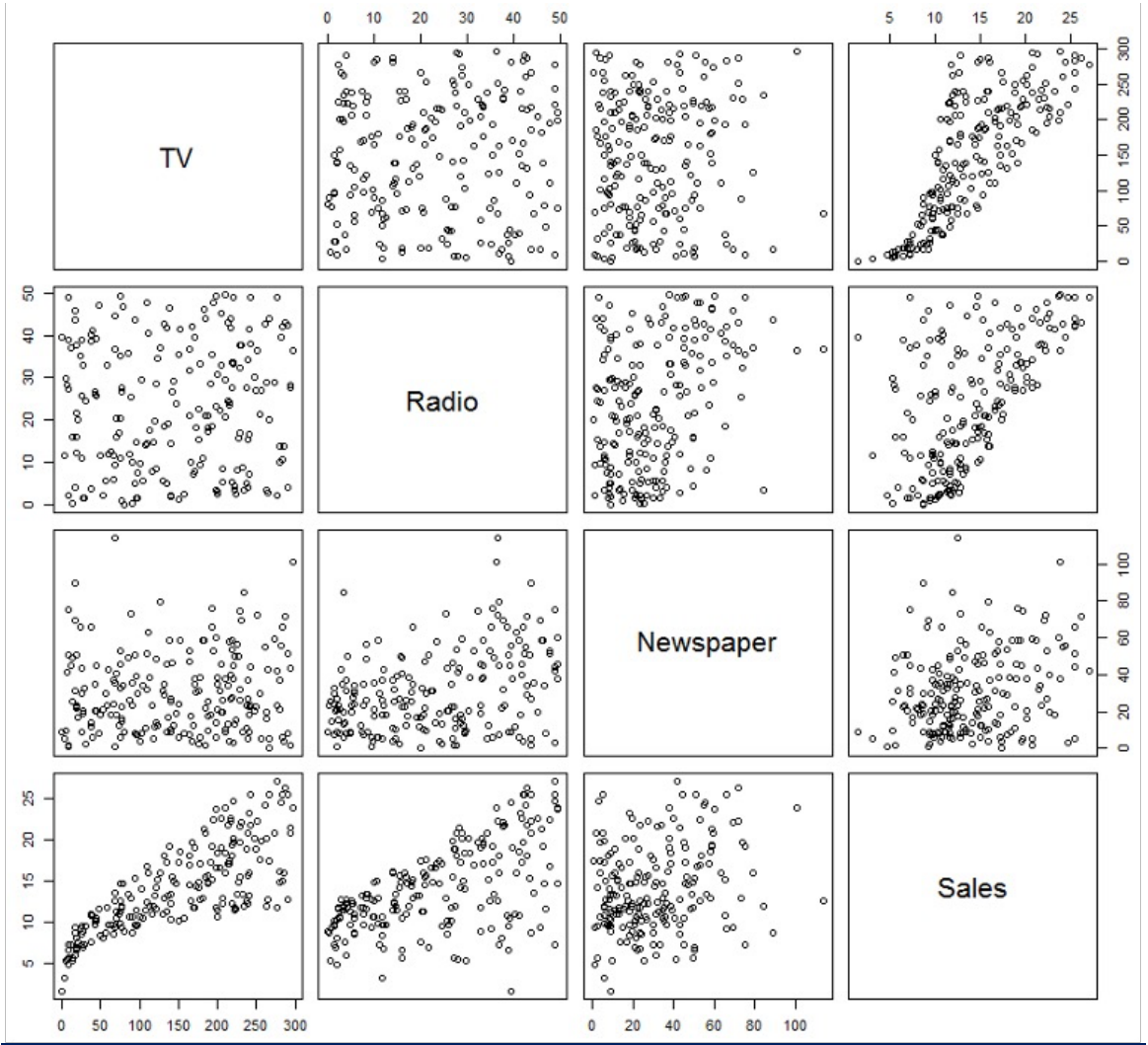
Source: Introduction to Statistical Learning

The sales advertisement dataset consists of a set of variables that describe the advertising budgets and sales figures for a company. The data contains 200 observations of 4 variables:

1. TV advertising budgets
2. Radio advertising budgets
3. Newspaper advertising budgets, and
4. sales figures.

The dataset is used here to demonstrate the effectiveness of linear regression models in predicting sales figures based on advertising budgets.

Advertisement Sales Data



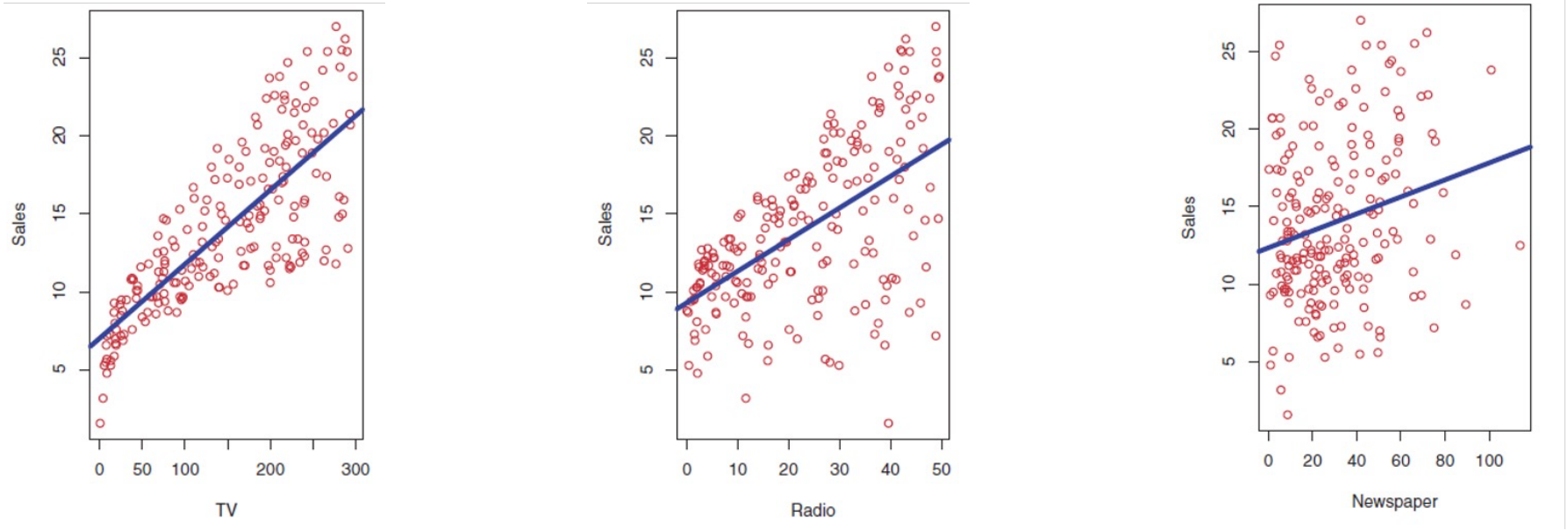
Scatter plot for the advertisement data

Source: Introduction to Statistical Learning

- **Objective:** To build a Linear Regression Model to predict Sales from TV, Radio, Newspaper
- **Dependent variable:** Sales
- **Independent variables:** TV, Radio, Newspaper

	TV	radio	newspaper	sales
TV	1.00000000	0.05480866	0.05664787	0.7822244
radio	0.05480866	1.00000000	0.35410375	0.5762226
newspaper	0.05664787	0.35410375	1.00000000	0.2282990
sales	0.78222442	0.57622257	0.22829903	1.0000000

As a Data Scientist, you have been asked to suggest an advertising/marketing plan that increases sales. How do you recommend?



Linear fits on the ***Advertising data set***; The plot displays sales, in thousands of units, as a function of *TV, radio, and newspaper budgets*, in thousands of dollars, for 200 different markets. In each plot we show the simple least squares fit of sales to that variable, as described. In other words, each blue line represents a simple model that can be used to predict sales using TV, radio, and newspaper, respectively

What kind of questions can you answer?

Question 1: Is there any evidence of an association?

Q: Is there a relationship between advertising budget (TV, Radio, or Newspaper) and sales?

⇒ Our first goal should be to analyze the data provided top find an evidence of an association between expenditure and sales.

⇒ If the evidence for association is strong, spend more money on advertising

⇒ If the evident for association is weak, spend less or no money on advertising!

What kind of questions can you answer?

Question 2: How strong is the relationship?

Q: How strong is the relationship between advertising budget (TV, Radio, or Newspaper) and sales?

⇒ Assuming there is a relationship between advertising and sales, we have to determine the **strength** of this relationship.

⇒ In other words, given a certain advertising budget, can we predict sales with high levels of accuracy? That would be a strong relationship.

⇒ Or is the prediction of sales based on advertising only slightly better than a random guess. This would be a weak relationship.

What kind of questions can you answer?

Question 3: Which component is contributing to the output?

Q: Which advertising media (TV, Radio, or Newspaper) is contributing to sales?

$\Rightarrow$ Do all three media (TV, Radio, Newspaper) contribute to sales or just one or two of them.

$\Rightarrow$ To answer this, we must be able to separate out the individual effects of each medium when we have spent money on all three media.

What kind of questions can you answer?

Question 4: How accurate is the estimation of individual effect?

Q: How accurately can we estimate the impact of each medium (TV, Radio, Newspaper) on sales?

⇒ For every dollar spent on advertising in a particular medium, how much will sales increase?

⇒ How accurately can we predict the amount of increase of sales?

What kind of questions can you answer?

Question 5: How accurately can we predict the future?

Q: How accurately can we predict future sales given an advertising budget?

$\Rightarrow$ For any given level of advertising budget (TV, Radio, Newspaper), can we predict the sales?
$\Rightarrow$ What is the accuracy of this prediction?

What kind of questions can you answer?

Question 6: Is the relationship linear?

Q: Is there a linear relationship?

$\Rightarrow$ If there is approximately a straight-line relationship between advertising expenditure in the various media and sales, then linear regression can be an appropriate tool.

$\Rightarrow$ If not, then it may still be possible to transform the predictor or the response so that linear regression can be performed.

What kind of questions can you answer?

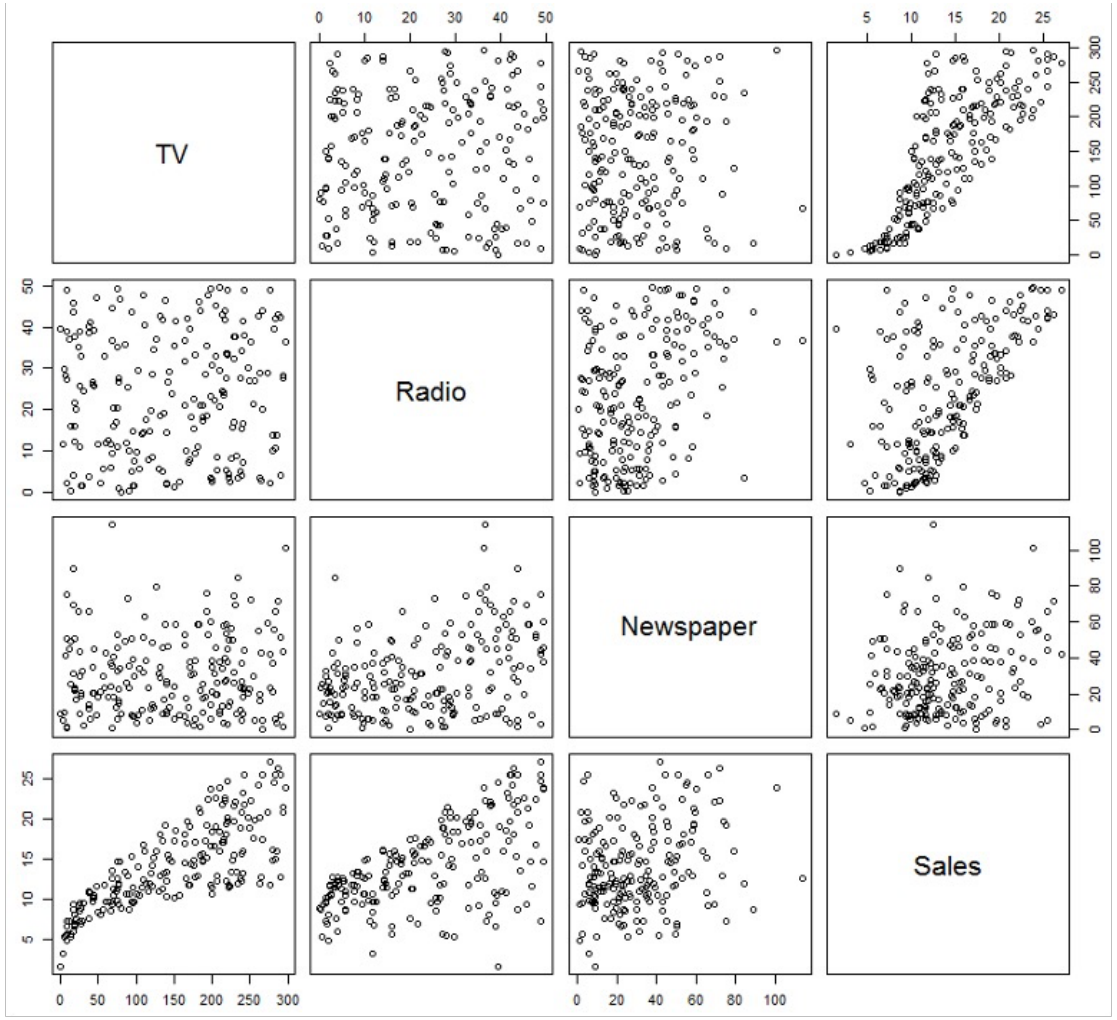
Question 7: Is there synergy among the components?

Q: Is there synergy among the components?

$\Rightarrow$ Maybe spending $50,000 on television advertising and $50,000 on radio advertising results in more sales than allocating $100,000 to television or radio individually. In marketing, this is known as synergy effect, while in Statistics/ML, we call it an **interaction effect.**

Linear regression can be used to answer all these questions

Correlation among variables



Scatter plot for the advertisement data

Source: Introduction to Statistical Learning (book)

Correlation tells us about strength (scatter) and direction of the linear relationship between two quantitative variables.

In addition, we would like to have a ***numerical description*** of how both variables vary together. For instance, is one variable increasing faster than the other one? And we would like to make predictions based on that numerical description.

	TV	radio	newspaper	sales
TV	1.00000000	0.05480866	0.05664787	0.7822244
radio	0.05480866	1.00000000	0.35410375	0.5762226
newspaper	0.05664787	0.35410375	1.00000000	0.2282990
sales	0.78222442	0.57622257	0.22829903	1.0000000

Outline of the talk

01 Introduction to Machine Learning Paradigms

- Supervised vs. Unsupervised Learning
- Regression vs Correlation
- Key Terminologies

02 **Fundamentals of Regression Analysis**

- **Simple Linear Regression**
- **Coefficients, Residuals and Errors in Regression**
- **Predicting Sales from Advertisement Data**

03 Multiple Linear Regression

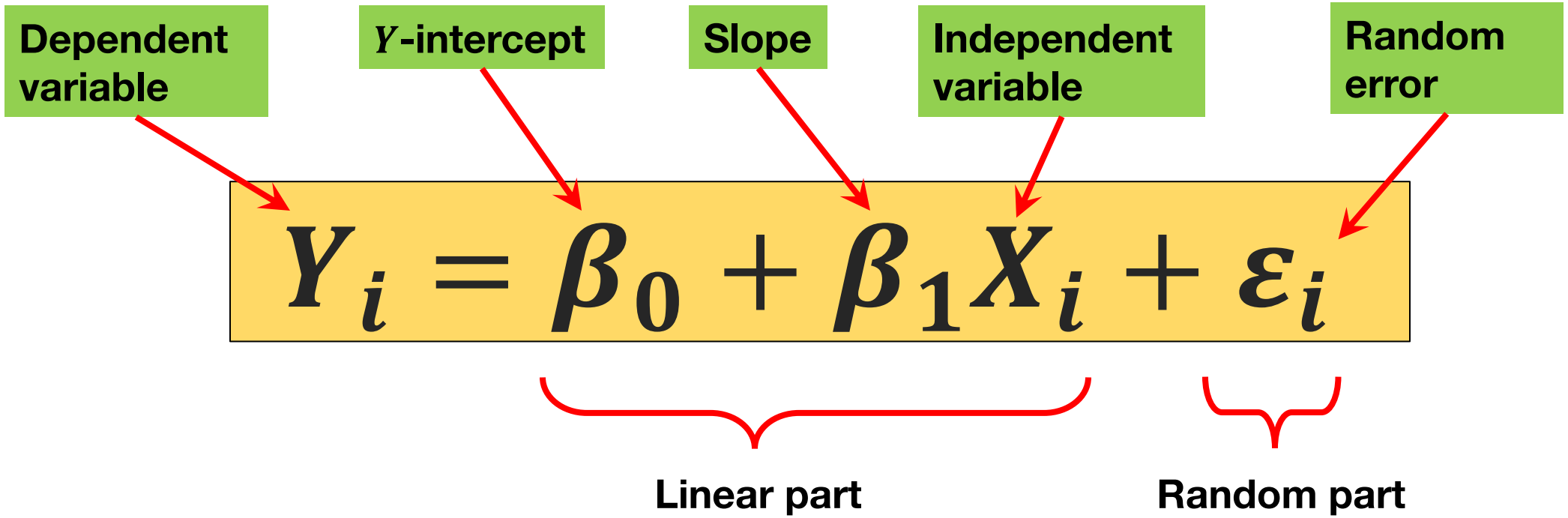
- Expanding Linear Regression to Multiple Variables
- Variable Interaction and Synergy Effects

04 Advanced Techniques and Nonlinear Regression

- Polynomial Regression – Extending Linear Regression to handle Nonlinear Relationships
- Model Selection and Regularization

Estimating f, where f represents the systematic information X provides about Y

Formulation



Simple Linear Regression

Formulation

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$