CoGrammar

Welcome to this session

Skills Bootcamp:

Introduction to Machine Learning (Theory)

The session will start shortly...

Questions? Drop them in the chat.
We'll have dedicated moderators answering questions.

Skills Bootcamp Data Science Housekeeping

- The use of disrespectful language is prohibited in the questions, this is a supportive, learning environment for all - please engage accordingly. **(Fundamental British Values: Mutual Respect and Tolerance)**
- No question is daft or silly - **ask them!**
- There are **Q&A sessions** midway and at the end of the session, should you wish to ask any follow-up questions. We will be answering questions as the session progresses as well.
- If you have any questions outside of this lecture, or that are not answered during this lecture, please do submit these for upcoming Academic Sessions. You can submit these questions here: **Questions**

Skills Bootcamp Data Science Housekeeping

- For all **non-academic questions**, please submit a query: ***www.hyperiondev.com/support***
- **Report a safeguarding incident: *www.hyperiondev.com/safeguardreporting***
- We would love your feedback on lectures: ***Feedback on Lectures.***
- Find all the lecture **content** in your **Lecture Backpack** on GitHub.
- If you are hearing impaired, kindly use your computer's function through Google chrome to enable captions.

Safeguarding & Welfare

We are committed to all our students and staff feeling safe and happy; we want to make sure there is always someone you can turn to if you are worried about anything.

If you are feeling upset or unsafe, are worried about a friend, student or family member, or you feel like something isn't right, speak to our safeguarding team:

Ian Wyles
Designated Safeguarding Lead

Simone Botes

Nurhaan Snyman

Rafiq Manan

Ronald Munodawafa

Tevin Pitts

Scan to report a safeguarding concern

or email the Designated Safeguarding Lead:
Ian Wyles
safeguarding@hyperiondev.com

CoGrammar HyperionDev

Skills Bootcamp Progression Overview

✅ Criterion 1 - Initial Requirements

Specific achievements within the first two weeks of the program.

To meet this criterion, students need to, by no later than **01 December 2024 (C11)** or **22 December 2024 (C12):**

- **Guided Learning Hours** (GLH)**:** Attend a minimum of 7-8 GLH per week (lectures, workshops, or mentor calls) for a total minimum of **15 GLH**.
- **Task Completion:** Successfully complete the **first 4 of the assigned tasks**.

✅ Criterion 2 - Mid-Course Progress

Progress through the successful completion of tasks within the first half of the program.

To meet this criterion, students should, by no later than **12 January 2025 (C11)** or **02 February 2025 (C12):**

- **Guided Learning Hours** (GLH)**:** Complete at least **60 GLH**.
- **Task Completion :** Successfully complete the **first 13 of the assigned tasks**.

CoGrammar

Skills Bootcamp Progression Overview

✅ Criterion 3 – End-Course Progress

Showcasing students' progress nearing the completion of the course.

To meet this criterion, students should:

- **Guided Learning Hours** (GLH)**:** Complete the **total minimum required GLH,** by the **support end date.**
- **Task Completion : Complete all mandatory tasks**, including any necessary resubmissions, by the end of the bootcamp, **09 March 2025 (C11)** or **30 March 2025 (C12).**

✅ Criterion 4 - Employability

Demonstrating progress to find employment.

To meet this criterion, students should:

- **Record an Interview Invite:** Students are required to record proof of invitation to an interview by **30 March 2025 (C11)** or **04 May 2025 (C12)**.
  - **South Holland Students** are required to proof and interview by **17 March 2025.**
- **Record a Final Job Outcome :** Within 12 weeks post-graduation, students are required to record a job outcome.

CoGrammar

Learning Outcomes

- Discuss machine learning and its applications.
- Define simple linear regression and its purpose.
- Interpret the mathematical equation and assumptions of simple linear regression.
- Evaluate simple linear regression models.

Lecture Overview

- Discuss Machine Learning and its practical applications.
- Explore the types of Machine Learning: supervised, unsupervised, and reinforcement learning.
- Discuss the applications, assumptions, and the mathematics behind Simple Linear Regression.
- Introduce the concept of performance metrics.

Introduction to Machine Learning

CoGrammar

Real-World Application of Machine Learning

Imagine you're a data scientist for a music streaming service.

You have user data (listening history, genres, user demographics, etc.). The management wants to know:

- What type of music should we recommend to users?
- Which users are likely to subscribe to premium plans?
- How can we predict peak usage times to optimize server performance?

Using Machine Learning, you can build models to predict user behavior, identify patterns in preferences, and optimise services to improve user satisfaction and company revenue.

Introduction to Machine Learning

- Machine learning is a way of teaching computers to learn and improve from experience without being explicitly programmed.
- It allows computers to automatically learn and adapt based on data.

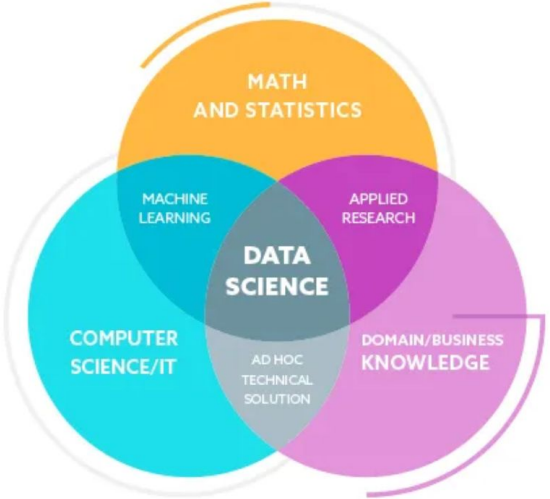


Image source:
https://ai.plainenglish.io/data-science-vs-artificial-intelligence-vs-machine-learning-vs-deep-learning-50d3718d51e5

Types of Machine Learning

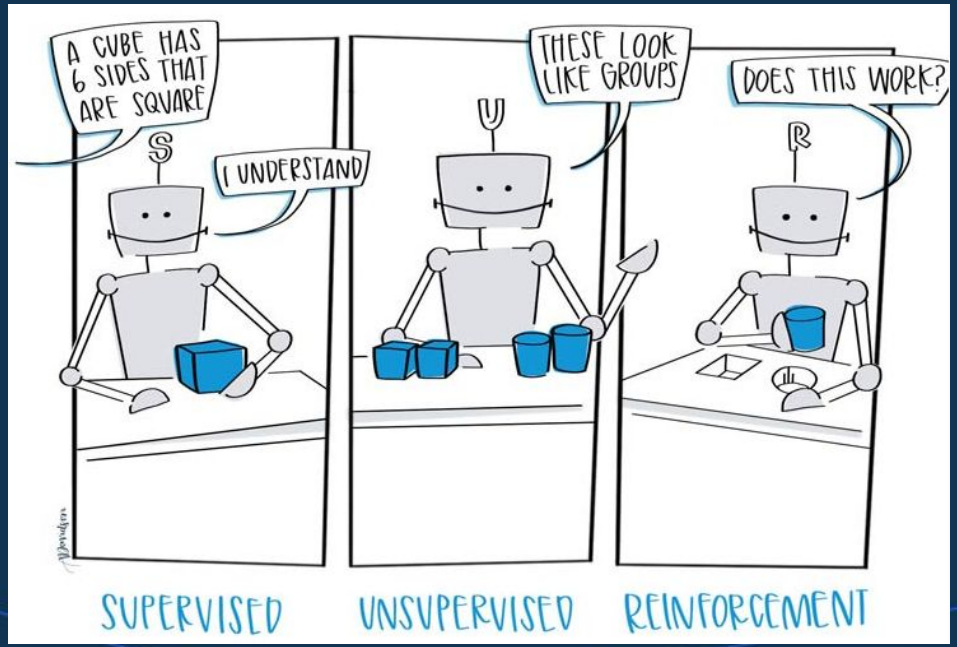
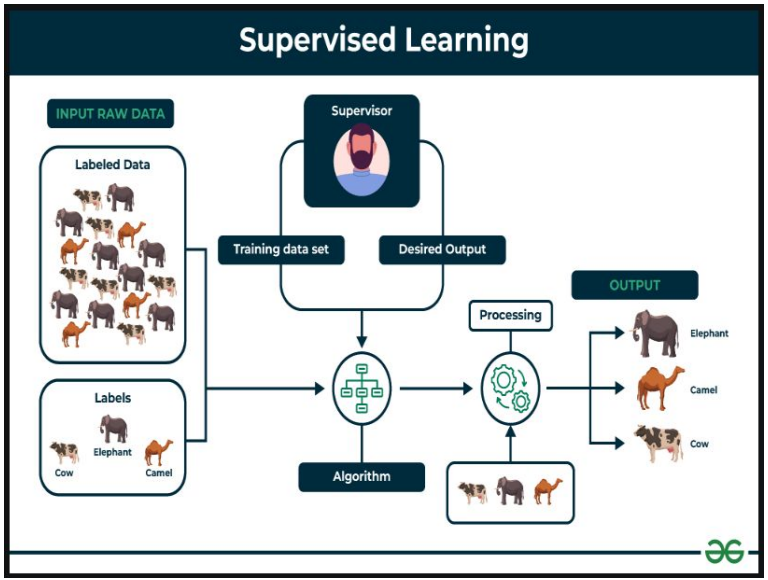


Image source: https://ceralytics.com/3-types-of-machine-learning/

CoGrammar

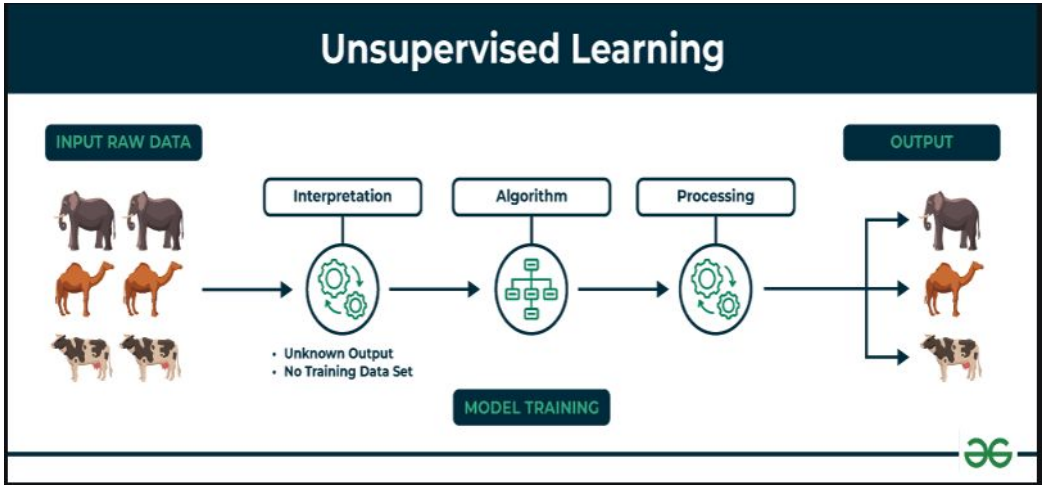
Types of Machine Learning

Supervised learning: The computer learns from labelled data, where both input and output data are provided.



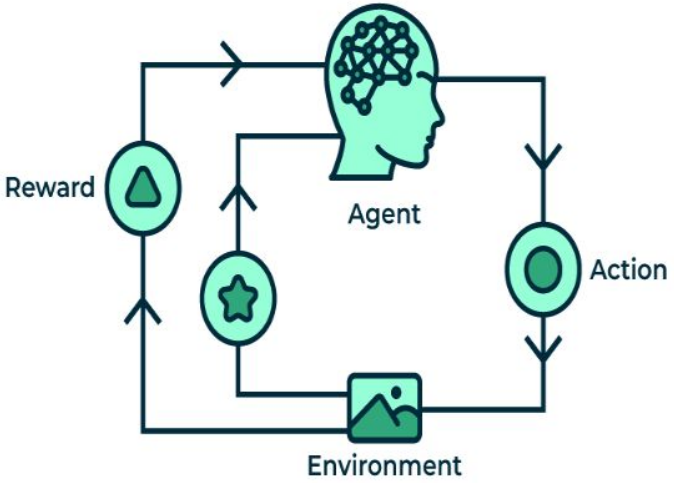
Types of Machine Learning

Unsupervised learning: The computer learns from unlabeled data, discovering hidden patterns or structures on its own.



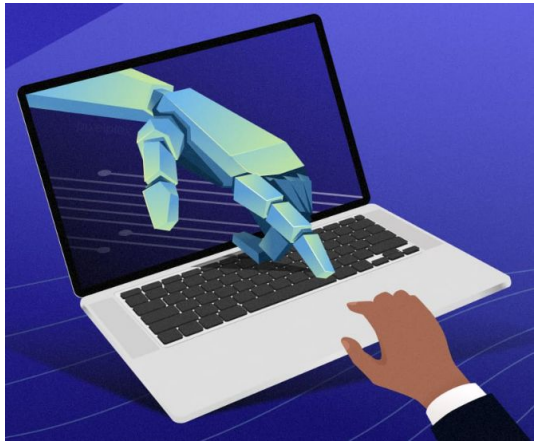
Types of Machine Learning

Reinforcement learning: The computer learns through interaction with an environment, receiving rewards or penalties for its actions.



More Applications of Machine Learning

- **Spam email filtering:** Identifying and separating spam emails from regular emails.
- **Image recognition:** Recognizing objects, faces, or scenes in images.
- **Recommender systems:** Suggesting products based on user preferences.



Supervised Learning

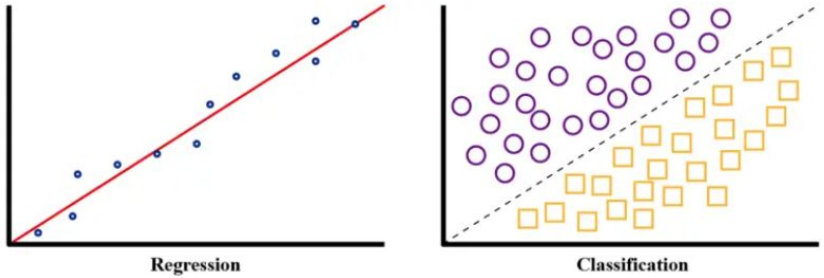
CoGrammar

Supervised Learning

- In **supervised learning**, the algorithm learns from **labelled data**, which consists of **input-output pairs**.
- The goal is to learn a function that maps input data to the correct output labels.

Types of Supervised Learning

- **Regression:** Predicting continuous numerical values, such as house prices or stock prices.
- **Classification:** Predicting discrete categories or classes, such as whether an email is spam or not.



Types of Supervised Learning Algorithms

- **Linear regression:** Fitting a straight line to data points to make predictions.
- **Logistic regression:** Predicting binary outcomes, such as yes/no or true/false.
- **Decision trees:** Making decisions based on a series of questions or conditions.
- **Support vector machines (SVM):** Finding the best boundary to separate different classes.
- **Neural networks:** Mimicking the structure and function of the human brain to learn complex patterns.

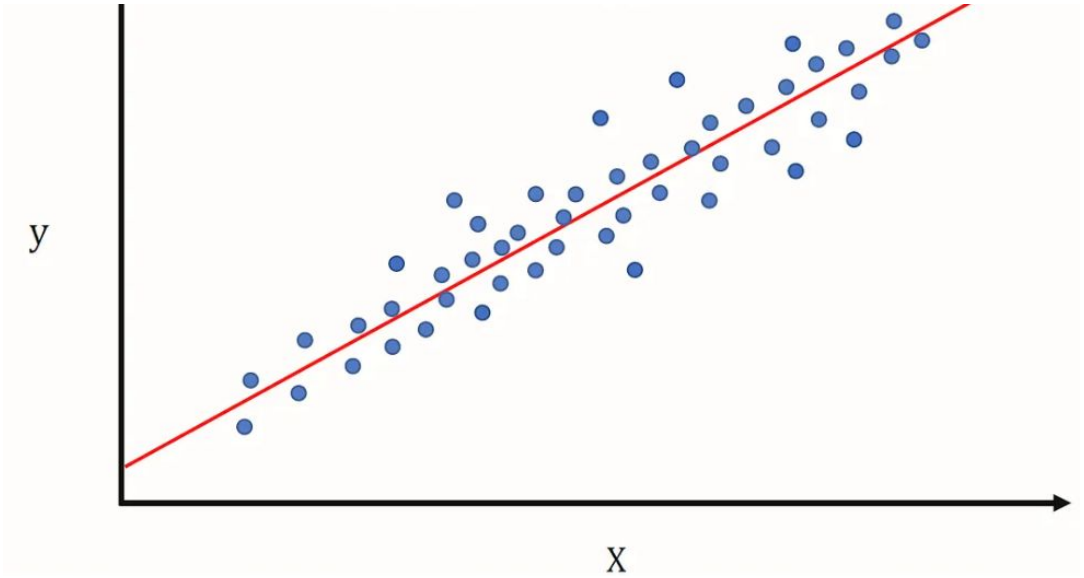
Supervised Learning: Simple Linear Regression

CoGrammar

Simple Linear Regression

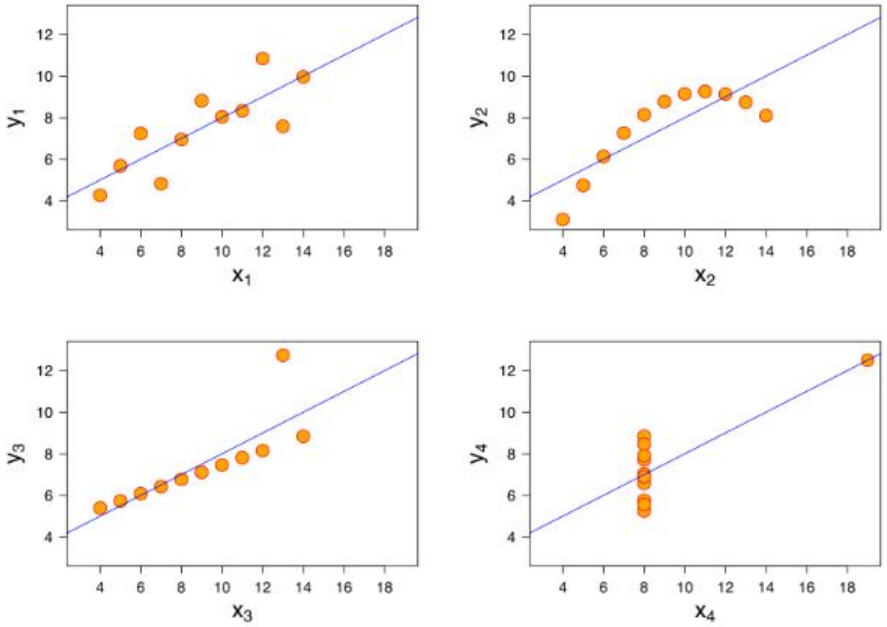
- **Simple linear regression** is a method to study the relationship between two variables: an **independent variable** (x) and a **dependent variable** (y).
- It helps us understand **how changes in the independent variable affect the dependent variable.**

Simple Linear Regression



Simple Linear Regression

- This image shows **Anscombe's Quartet**, illustrating the importance of **data visualisation** in regression analysis.
- While all four datasets have nearly identical statistical properties (variance, correlation, and regression line), their **visual patterns differ significantly**.
- **Always plot your data to ensure regression assumptions are valid.**



CoGrammar

Source:
https://en.wikipedia.org/wiki/Linear_regression#/media/File:Anscombe's_quartet_3.svg

Purpose of Simple Linear Regression

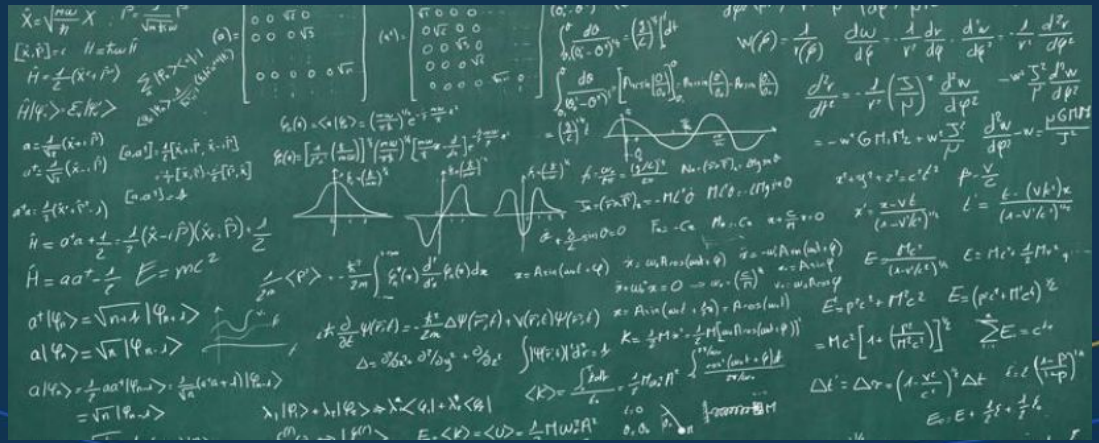
- To find the **best-fitting straight line** that describes the **relationship** between x and y.
- This line can be used to make **predictions** about the dependent variable based on new values of the independent variable.

Applications of Simple Linear Regression

- **Sales forecasting:** Predicting future sales based on historical data.
- **Price prediction:** Estimating the price of a product based on its features.
- **Trend analysis:** Identifying trends or patterns in data over time.

CoGrammar

Math Behind Simple Linear Regression



CoGrammar

Math Behind Simple Linear Regression

- Simple Linear Regression builds on the **equation of a straight line** ($y = mx + c$) by incorporating statistical components.
- Understanding this equation is essential to interpret the relationship between variables and make predictions.
- The equation is written as: $\mathbf{y = \beta_0 + \beta_1 x + \varepsilon}$
  - β_0 is the **intercept**, representing the predicted value of **y** when **x = 0**.
  - β_1 is the **slope**, indicating how much **y changes** for a **one-unit increase in x**.
  - ε is the **error term**, accounting for the variability in **y** that **cannot be explained by x**.

Math Behind Simple Linear Regression

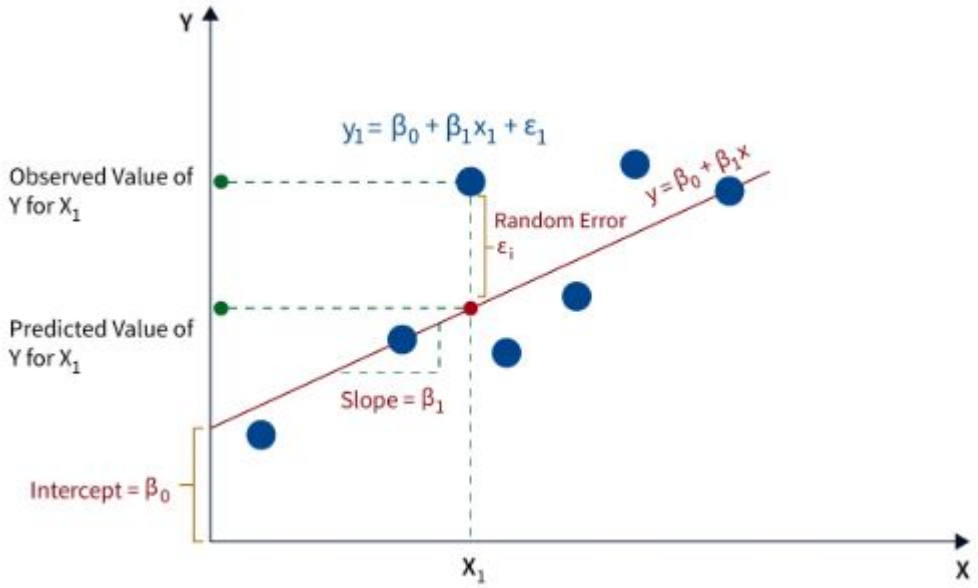


Image source: https://www.scaler.com/topics/regression-in-data-mining/

Let's take a break

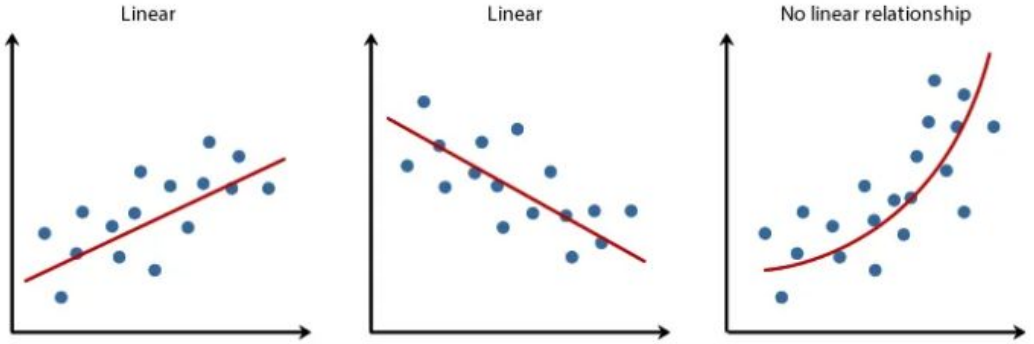
CoGrammar

Assumptions and Limitations of Simple Linear Regression

CoGrammar

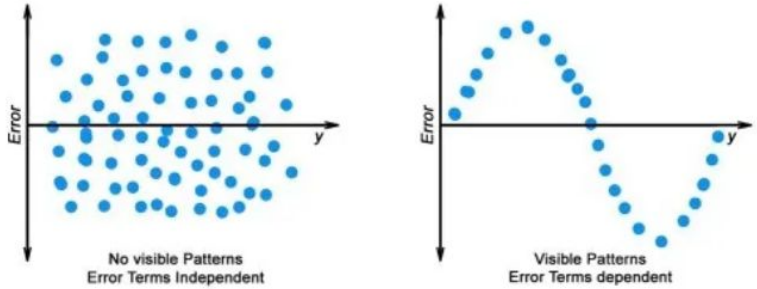
Assumptions & Limitations of Simple Linear Regression

- **Linearity:** The relationship between x and y should be linear.



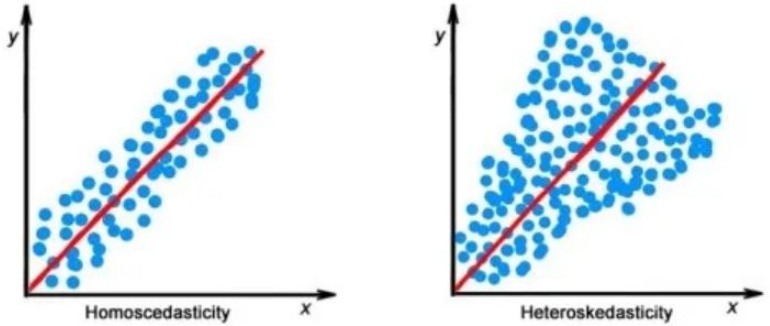
Assumptions & Limitations of Simple Linear Regression

- **Independence:** The observations should be independent of each other.



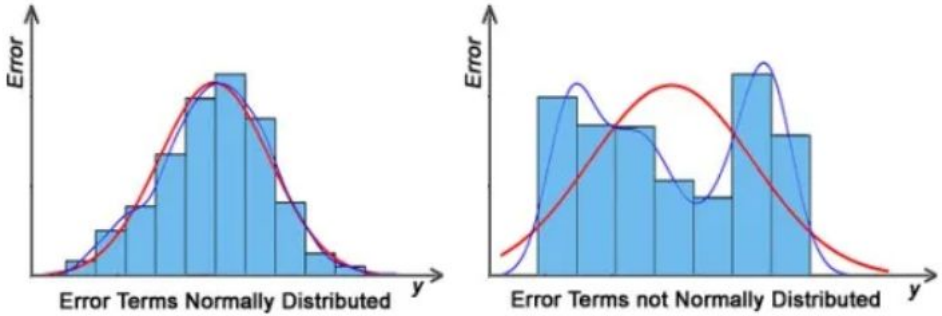
Assumptions & Limitations of Simple Linear Regression

- **Homoscedasticity:** The variability of y should be constant across all values of x.



Assumptions & Limitations of Simple Linear Regression

- **Homoscedasticity:** The variability of y should be constant across all values of x.



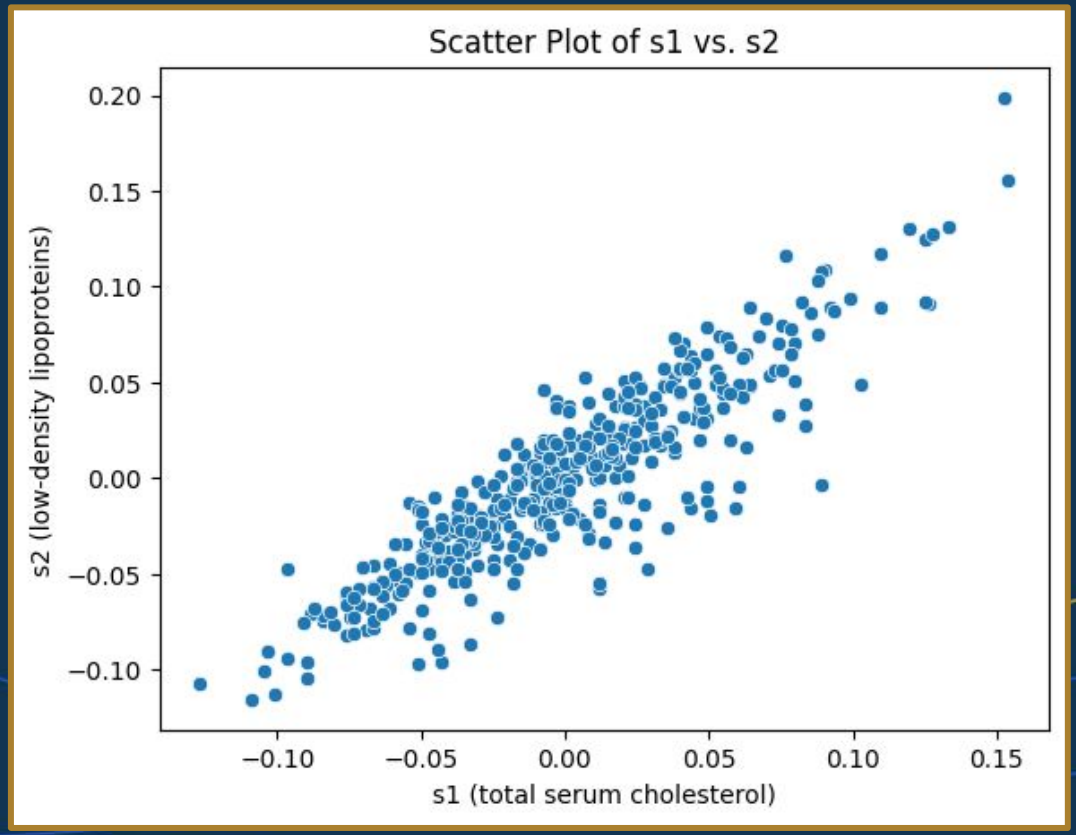
Implementing Simple Linear Regression

CoGrammar

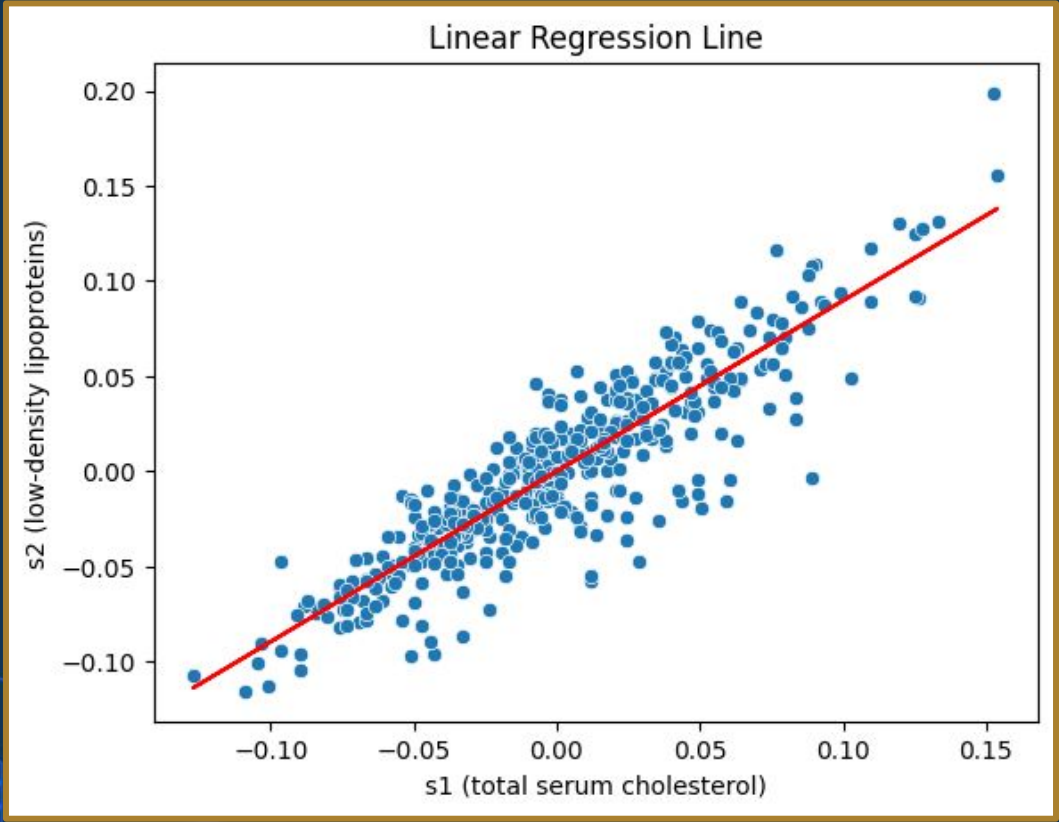
Simple Linear Regression

- **Scikit-learn** is a popular Python library for machine learning.
- It provides simple and efficient tools for data analysis and modelling.
- In the next slides, we will look at the results of a simple linear regression model that was fitted on the built-in **diabetes dataset** from scikit-learn.
- The dataset contains information about **various medical predictors** and a **quantitative measure of disease progression**.

Regression Code Results



Regression Code Results



Regression Code Results

```
Evaluation Metrics:
Mean Squared Error (MSE): 0.00044342882373426217
R-squared (R2) Score: 0.8040044599094562
```

Interpretation of Results

- **Scatter plot:**
  - The scatter plot visualises the relationship between s1 (total serum cholesterol) and s2 (low-density lipoproteins).
  - It helps **assess the linearity and spread of the data points**.
- **Linear regression line:**
  - The red line represents the **best-fit line** obtained from the linear regression model.
  - It shows the **predicted relationship** between s1 and s2 based on the trained model.

Evaluation Metrics

- **Mean Squared Error (MSE):**
  - MSE measures the average squared difference between the predicted and actual values.
  - A lower MSE indicates better model performance.
- **R-squared (R^2) score:**
  - R^2 represents the proportion of variance in the target variable that can be explained by the model.
  - An R^2 value closer to 1 indicates a better fit of the model to the data.

Evaluation Metrics

- **Accuracy** is another commonly used metric for evaluating the performance of a machine learning model, particularly in **classification problems**.

  Accuracy = (Number of correct predictions) / (Total number of predictions) * 100%

- While accuracy is more suitable for classification tasks, metrics like Mean Squared Error (MSE) and R-squared (R^2) are used for **regression problems**.

Parameter Tuning

- **Parameter tuning** is the process of finding the optimal values for a model's hyperparameters to improve its performance.
- **Hyperparameters are settings** that are not learned from the data but are set before training the model.
  - Examples of hyperparameters in linear regression include the **learning rate**, **regularisation strength**, and **the number of iterations**.

Polls Assessment

CoGrammar

Which of the following best describes the purpose of simple linear regression?

1. To classify data points into distinct categories
2. To estimate the relationship between two continuous variables
3. To cluster data points based on their similarities
4. To predict the probability of an event occurring

Which of the following best describes the purpose of simple linear regression?

1. To classify data points into distinct categories
2. **To estimate the relationship between two continuous variables**
3. To cluster data points based on their similarities
4. To predict the probability of an event occurring

What does the coefficient of determination (R-squared) represent in a linear regression model?

1. The average squared difference between predicted and actual values
2. The proportion of variance in the target variable explained by the model
3. The slope of the regression line
4. The intercept of the regression line

What does the coefficient of determination (R-squared) represent in a linear regression model?

1. The average squared difference between predicted and actual values
2. **The proportion of variance in the target variable explained by the model**
3. The slope of the regression line
4. The intercept of the regression line

Which Python library is commonly used for implementing machine learning algorithms, including linear regression?

1. NumPy
2. Pandas
3. Matplotlib
4. Scikit-learn

Which Python library is commonly used for implementing machine learning algorithms, including linear regression?

1. NumPy
2. Pandas
3. Matplotlib
4. **Scikit-learn**

What is the primary difference between supervised and unsupervised learning?

1. Supervised learning uses labelled data, while unsupervised learning uses unlabelled data
2. Supervised learning is used for classification, while unsupervised learning is used for regression
3. Supervised learning requires more computational resources than unsupervised learning
4. Supervised learning is more accurate than unsupervised learning

What is the primary difference between supervised and unsupervised learning?

1. **Supervised learning uses labelled data, while unsupervised learning uses unlabelled data**
2. Supervised learning is used for classification, while unsupervised learning is used for regression
3. Supervised learning requires more computational resources than unsupervised learning
4. Supervised learning is more accurate than unsupervised learning

Which of the following is an example of a regression problem?

1. Predicting the sentiment of a movie review (positive or negative)
2. Identifying the species of a flower based on its features
3. Estimating the price of a house based on its size and location
4. Grouping customers into different segments based on their purchasing behaviour

Which of the following is an example of a regression problem?

1. Predicting the sentiment of a movie review (positive or negative)
2. Identifying the species of a flower based on its features
3. **Estimating the price of a house based on its size and location**
4. Grouping customers into different segments based on their purchasing behaviour

Questions and Answers

CoGrammar

Thank you for attending

CoGrammar

SKILLS FOR LIFE
SKILLS BOOTCAMPS

Department for Education