Slide 3
Hello everyone!

- Today, we're talking about how computers learn from data, focusing on a method called Supervised Learning.
- There are three main ways computers learn: Supervised, Unsupervised, and Reinforcement Learning.
- We’re only looking at Supervised Learning today because we’re exploring something called Multiple Linear Regression, which falls under this category.

What is Supervised Learning?
- In Supervised Learning, we use data that already has answers to teach the computer.
- It’s like a teacher helping a student learn.
- The two main jobs of Supervised Learning are Regression and Classification.

Focusing on Regression:
- Regression helps us predict numbers, like scores or prices.
- There are three kinds of Regression: Simple Linear, Multiple Linear, and Polynomial.
- Today, we’re zooming in on Multiple Linear Regression.

Multiple Linear Regression:
- It’s like Simple Linear Regression but deals with more things at once.
- It helps us understand how different things relate to one another and make predictions.

Slide 4

We're going to talk about how we use data to teach a computer model to make predictions.

- We start with a whole bunch of data, but we don't use it all at once. We split it into two parts: Training Data and Validation Data.

Training Data:

- Think of Training Data like the lessons in school. It's used to teach the model.
- The model learns from this data and starts making predictions.

Validation Data:

- Once our model has learned and is making predictions, we need to check how well it's doing.
- This is where Validation Data comes in. It's like a test or exam for the model.
- We use Validation Data to evaluate the predictions the model makes.

Evaluation and Final Model:

- If the model does well on the test, great! We have our Final Model.
- But, if it doesn't do so well, it's back to the drawing board. We retrain the model using the Training Data.
- We keep doing this until we get a model that we're happy with.

So, it's a bit like learning in school – study, take a test, and if you don't do well, study some more until you get it right!

Slide 5
We're diving into a concept called Linear Regression. It's a way we can show the relationship between two things: one we want to predict (we call this "Y") and one we use to make that prediction (we call this "X").

What is Linear Regression?

- Imagine you have a bunch of points on a graph, and you want to draw a straight line that fits them best. That's what Linear Regression does!
- It finds the best-fitting straight line to show how our dependent variable "Y" (the one we want to predict) is related to our independent variable "X" (the one we use to make predictions).

Why do we use it?

- By drawing this line, we can make guesses about what "Y" could be, based on different values of "X".
- For example, if "Y" is the price of a house, and "X" is its size, we can use Linear Regression to predict the price of a house based on its size!

So, Linear Regression helps us see the connection between two things and make predictions using a straight line on a graph!

Slide 6
We have a table in front of us that shows a dataset. This dataset is telling us a story about the relationship between the number of hours students studied and the grades they received.

What Does the Data Show?

- When we look at the data points, we can see a hint that there might be a positive relationship. This means that as the number of study hours goes up, the grades might go up too!
- In simpler terms, it's like saying, the more you study, the better your grades could be.

Why is this Important?

- This dataset is like a starting point for us. It gives us an idea, a hint, that there might be a connection between study time and grades.

Next Steps:

- By doing linear regression, we can explore this relationship more. We can find out if studying more really leads to better grades, and if so, how much of a difference it makes.
- This helps us understand how study time can affect academic performance and can be really useful for students to plan their study time better!

Slide 7

- We have two main things we're looking at: the independent variable and the dependent variable.
- The independent variable is plotted on the the x-axis or horizontal axis, while the dependent variable is plotted on the ordinate (also called the y-axis or vertical axis).
- The dependent variable is the one whose value changes as a result of changes in the independent variable. In our example, study time was the independent variable, grade received was the dependent one.
- The grade depended on the amount of studying done, not vice versa.
- When we look at our graph, we can see a pattern: more hours studied generally seems to go with higher grades.
- This pattern is what we're interested in. It gives us a clue that there might be a relationship between study time and grades.

Slide 8

- Two ways we can make predictions using data: Linear Regression and Multiple Linear Regression. They sound similar, but they have a key difference!

Linear Regression:

- Linear Regression is like looking at the world with one eye open. We use only one independent variable to make predictions.
- It's all about understanding how one thing (like the size of a house) can help us predict something else (like the price of the house).
- So, if we were using Linear Regression, we'd only be looking at how the size of the house affects its price.

Multiple Linear Regression:

- Now, Multiple Linear Regression is like opening both eyes. We use two or more independent variables to make predictions.
- This means we're looking at how a bunch of different things (like size, number of bedrooms, location, and age) together affect the price of the house.
- It gives us a fuller picture and helps us make more accurate predictions because we're considering more factors.

Why Does This Matter?

- Knowing the difference between these two methods is important because it helps us decide how many factors we need to consider to make good predictions.
- If we want a simple view, we might use Linear Regression. But if we need to consider many things at once, Multiple Linear Regression is our friend!

Slide 9

We're breaking down the difference between Linear Regression and Multiple Regression in predicting values, using the example of predicting house prices.

Linear Regression:

- In Linear Regression, we're like detectives using one clue to solve a case. We use one independent variable to predict the value of the dependent variable.
- Imagine we want to guess the price of a house. In Linear Regression, we would only use one factor, like the square footage of the house, to make our guess.
- So, it's pretty straightforward – we look at the size of the house and use that to predict the price!

Multiple Regression:

- Now, Multiple Regression is like having multiple clues to solve a mystery. We use more than one independent variable to predict the value of the dependent variable.
- In the case of predicting house prices, we wouldn't just look at the size. We'd also consider the number of bedrooms, the location, and other factors to make a more informed guess.
- It's like putting together different pieces of a puzzle to see the whole picture and make a better prediction!

In Summary:

- Linear Regression: One clue (Square Footage) to predict the house price.
- Multiple Regression: Multiple clues (Square Footage, Number of Bedrooms, Location, etc.) to predict the house price.
- Understanding these differences helps us choose the right approach depending on how many factors we need to consider for making accurate predictions!

Slide 10

It visualizes a regression model that predicts a student's score based on two factors: hours studied and the number of assignments completed.

Understanding the Model:

- The model we are looking at is 'score = hours_studied + assignments'. This means we are trying to predict a student's score by looking at how many hours they've studied and how many assignments they've completed.
- In simpler terms, we're saying a student's score depends on both their study time and the work they've put into their assignments.

What's in the Image?

- In the image, we see three different regression lines. Each line represents a different value of assignments completed.
- This means we're looking at how the relationship between study hours and scores changes when students complete different numbers of assignments.
- For example, one line might show the scores of students who completed five assignments, another for those who completed ten, and so on.

Slide 12

Breaking Down the Equation:

- The equation for Multiple Linear Regression looks like this: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon$.
- Here, y is what we're trying to predict – it's called the dependent variable.
- The β values are the coefficients – they tell us how important each independent variable is in predicting y.
- The x values are our independent variables or features – these are the different factors we are using to make our prediction.
- And lastly, ϵ is the error term – it accounts for any difference between our prediction and the actual value of y.

Slide 13

What is R-squared?

- Imagine we're trying to guess the prices of houses using their size and the number of bedrooms.
- We gather data on lots of houses and use this information to build a regression model – think of it as a fancy equation – to make our guesses.
- R-squared is like a report card for our model. It tells us how well our model is doing at predicting house prices based on the factors we're considering.

How Does It Work?

- If R-squared is close to 1, like 0.9, that's like getting an A on the report card! It means our model is explaining 90% of the differences in house prices using the size and number of bedrooms. Our model is doing a great job!
- But, if R-squared is closer to 0, like 0.2, that's not so good. It's like getting a low grade because our model is only explaining 20% of the differences in house prices. We might need to study more or consider other factors!

Slide 14
The p-value! It's like a detective tool that helps us figure out which factors really matter when we're predicting something, like house prices.

Adding a New Factor:

- Imagine we're still trying to predict house prices, but this time, we add a third factor – the distance to the nearest school.
- Now, we have three factors: size, number of bedrooms, and distance to the school. We're curious to see which of these really affect house prices.

Understanding p-values:

- Each of these factors gets its own p-value, which is like a clue telling us how important that factor is.
- If the p-value is very low (below 0.05), it's like finding strong evidence! It means that factor, like size or bedrooms, probably has a big impact on house prices.
- But, if the p-value is high (above 0.05), it's like the evidence isn't strong enough. For example, the distance to the school might not be that important in predicting house prices.

Slide 15

R-squared and p-values Together:

- R-squared gives us a big picture view – it tells us how well our whole model, with all the factors, is doing at explaining house prices.
- The p-values are like zooming in – they tell us which specific factors in our model are really making a difference and which might not be that important.

Slide 16

Let's consider a simple example where we are trying to predict the price of a house (dependent variable) based on two independent variables: the size of the house (in square feet) and the number of bedrooms. In this case, the multiple linear regression equation would be:

Imagine you're trying to guess how much a house costs. You look at two things: how big the house is (its size in square feet) and how many bedrooms it has.

In this guessing game:

- Bo is like your starting point of guessing. It's the price you'd guess if the house had no size and no bedrooms (which sounds silly, but it helps with the math!).
- B1 tells you how much more money you'd guess for each extra square foot of size.
- B2 tells you how much more money you'd guess for each extra bedroom.
- Size (sq ft) and Number of Bedrooms are the details you know about the house.
- € (Epsilon) is like the difference between your guess and the actual price. Nobody's perfect, so it accounts for any mistakes in the guess!
- Price (House Price) is what you're trying to guess – the actual cost of the house.
- So, when you put it all together in the equation: Price=Bo+B1(Size)+B2(Bedrooms)+€

It's like saying:

- Start with your base guess (Bo),
- Add a bit for each square foot of the house (B1 times Size),
- Add a bit for each bedroom (B2 times Bedrooms),
- And then, there might be a little bit of difference (€) because guessing isn't perfect!
- And that's how you'd make your best guess at the house price using the size and the number of bedrooms!

Slide 17

1. Linear Relationship:

    First off, we assume there's a linear relationship between the dependent (what we're trying to predict) and independent variables (what we're using to make predictions). This means as one goes up or down, the other one does so in a straight-line manner.

2. No Multicollinearity:

    Secondly, we assume that the independent variables are not highly correlated with each other. This means that each variable brings something unique to the table and isn't just repeating what another variable is saying.

3. Constant Variance of Residuals:

    Third, we assume that the variance of the residuals (the differences between predicted and actual values) is constant. This means that our model's accuracy doesn't change drastically for different levels of the dependent variable.

4. Independence of Data Points:

    Fourth, each data point should be independent; they shouldn't depend on each other. It's like saying each piece of information we have is its own separate clue and isn't influenced by the other clues.

5. Normal Distribution of Variables:

    Lastly, we assume that all variables should be normally distributed. This means that the values of our variables tend to cluster around the average in a certain way, forming a bell-shaped curve.

Why are These Assumptions Important?

- These assumptions are crucial because if they hold true, we can trust our model more. They help ensure that our predictions are reliable and that we're not just seeing patterns that aren't really there.

So, by checking these assumptions, we're making sure we're playing the prediction game by the rules and that our results can be trusted!

Slide 18

1. Predictive Power:

    One of the standout benefits is its predictive power. It allows us to make educated guesses about the dependent variable based on several independent variables. This ability is a treasure in fields like finance, economics, and social sciences, helping professionals make informed decisions.

2. Quantifying Relationships:

    Multiple Linear Regression shines in quantifying the relationships between variables. It helps us understand which factors significantly impact the outcome and in what direction, allowing us to pinpoint the key drivers behind the dependent variable.

3. Control for Confounding Factors:

    This technique is adept at controlling for confounding factors. By including relevant variables in our model, we can account for other influences that might affect the outcome, ensuring our results are robust and reliable.

4. Model Interpretability:

    The clarity of Multiple Linear Regression is another perk. The coefficients in the equation tell us about each variable's contribution, making it easier to interpret the findings and understand the dynamics at play.

5. Assumption Testing:

    Lastly, it offers a suite of diagnostic tools, such as residual analysis and multicollinearity detection, allowing us to assess the model's quality and ensure the assumptions hold, which is crucial for the reliability of our findings.

Slide 19

1. Linearity Assumption:

   The model assumes a linear relationship between the independent and dependent variables. If this isn't true, our model might not be accurate, and the results could be misleading.

2. Multicollinearity:

   When the independent variables are highly correlated, it's like they're speaking over each other, making it hard to hear each one's individual contribution. This can lead to unstable estimates and make it challenging to interpret the results.

3. Overfitting:

   Including too many variables can make the model too tailored to the training data, like wearing a suit that's too tight. It might not perform well with new, unseen data, limiting its generalizability.

4. Assumption Violations:

   The model relies on several assumptions, like normality of residuals and constant variance. If these are not met, it's like building a house on shaky foundations – the results might not be reliable.

5. Limited Handling of Categorical Variables:

   The model is best with continuous variables. Handling categorical ones requires extra steps, which can complicate the model and add to the challenge.

6. Data Requirements:

   Lastly, the model needs a good amount of data to be reliable. With too few data points, the results might be as shaky as a boat in a storm.

Slide 20

1. Stock Price Prediction:
    - In the financial world, predicting stock prices is like trying to catch a moving train. Multiple Regression comes in handy here!
    - Analysts use it to predict a stock's future price by considering various independent variables like company earnings, interest rates, and market volatility.
    - By analyzing how these factors influence stock prices, investors can make more informed decisions about buying or selling stocks, aiming to maximize returns and minimize risks.

2. Economic Forecasting:
    - Shifting gears to economics, Multiple Regression is a key player in forecasting economic indicators.
    - Economists use this technique to predict variables like GDP growth, inflation rates, and unemployment rates. They consider a myriad of factors such as government spending, consumer spending, and trade balances.
    - By understanding the relationships between these variables, policymakers and economists can anticipate economic trends, formulate policies, and make recommendations to steer the economy in the right direction.

Slide 21

1. Medical Research:
    - In the realm of medical research, Multiple Regression is a valuable ally. Researchers use it to investigate how various factors such as age, genetics, and lifestyle contribute to health outcomes or the risk of diseases.
    - By analyzing the relationships between these factors, scientists can identify risk factors, develop preventive strategies, and contribute to the advancement of personalized medicine, ultimately aiming to improve health and well-being.

2. Hospital Readmission Prediction:
    - Shifting to hospital settings, Multiple Regression plays a crucial role in predicting patient readmissions. It helps healthcare professionals estimate the likelihood of a patient returning to the hospital within a specific time frame.
    - By considering factors like medical history, comorbidities, and hospital procedures, healthcare providers can identify high-risk patients, optimize care plans, and implement interventions to reduce readmissions, enhancing patient outcomes and healthcare efficiency.

Slide 22

1. Sales Forecasting:
    - In the dynamic world of sales, predicting future trends is crucial. Multiple Regression steps in as a forecasting wizard, helping companies anticipate future sales.
    - By considering various variables like advertising spending, pricing, and seasonality, businesses can gauge market trends, optimize their strategies, and ensure they are well-positioned to meet consumer demands and maximize profits.

2. Market Research:
    - Shifting to market research, Multiple Regression is like a magnifying glass, helping companies analyze the impact of different marketing strategies.
    - By examining how various elements like advertising channels and product features influence consumer purchasing behavior, businesses can fine-tune their marketing mix, identify what resonates with consumers, and craft strategies that enhance brand appeal and drive sales.