CONTENTS

Assumptions of Linear Regression: Highlight the key assumptions of linear regression, such as linearity, independence, homoscedasticity, and normality of residuals.

Machine Learning Internship

Name : Mayank Sharma
Platform: Mentorness
Position : Machine Learning Intern
Duration : 1 Month
Starting Date : Monday, 08 April, 2024
Location : Remote

Introduction to Linear Regression

Linear regression is a fundamental statistical technique used to analyze and model the relationship between a dependent variable and one or more independent variables. This slide provides an overview of the key assumptions that underlie the application of linear regression.

by Mayank Sharma

Assumption 1: Linearity

Linear Relationship

The key assumption of linearity in linear regression is that the relationship between the independent and dependent variables is linear. This means the change in the dependent variable is proportional to the change in the independent variable.

Scatterplot Analysis

Analysts can visually inspect a scatterplot of the data to assess the linearity assumption. A clear linear pattern in the scatterplot indicates the linearity assumption is met.

Linearity Test

Statistical tests like the lack-of-fit test can also be used to formally evaluate the linearity assumption. These tests compare the linear model to more complex models to check if a linear relationship is appropriate.

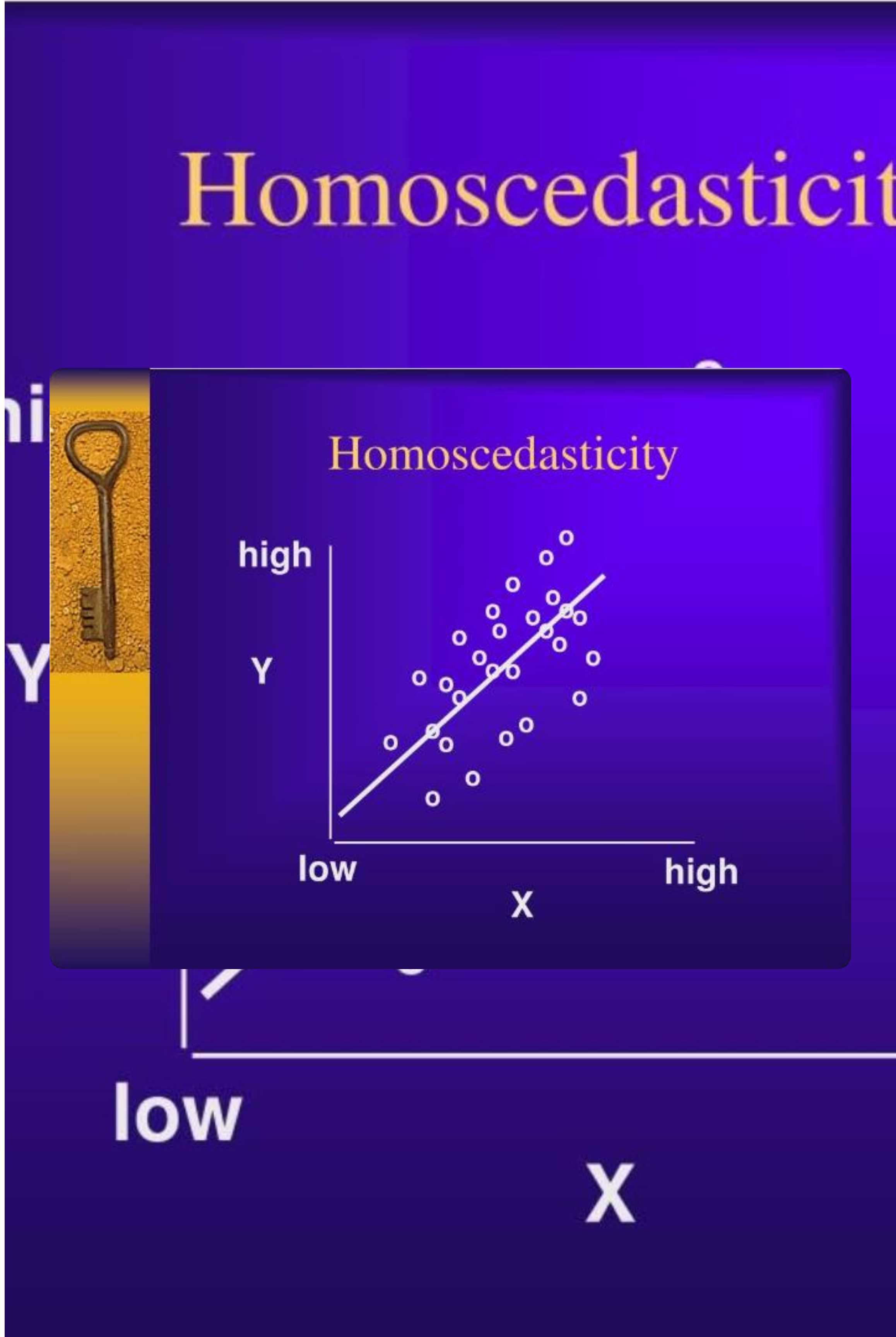
Assumption 2: Independence of Observations

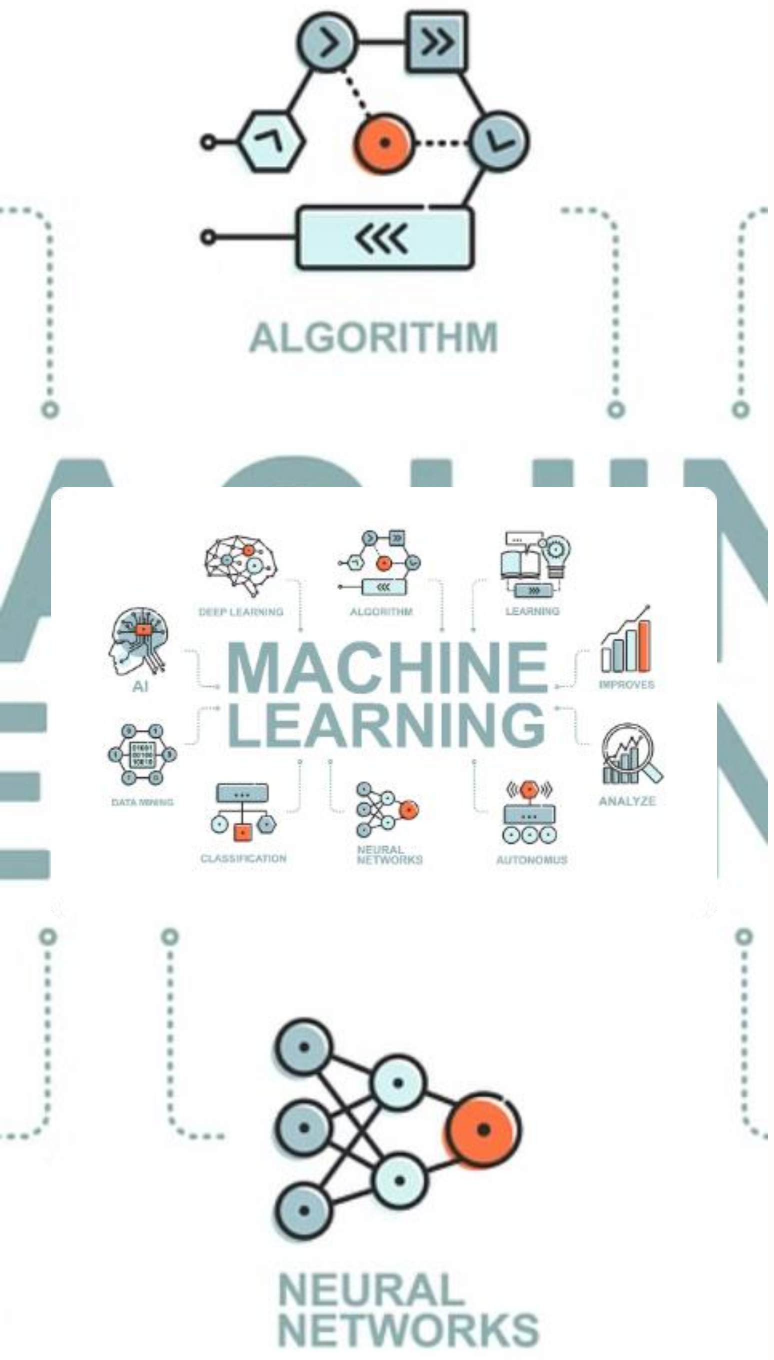
1. The **linear regression model** assumes that the observations in the data are **independent** of one another.
2. This means that the value of one observation does not depend on the value of any other observation in the dataset.
3. Violations of this **independence assumption** can lead to **biased** and **inefficient** regression estimates, compromising the validity of the analysis.

Assumption 3: Homoscedasticity

The assumption of homoscedasticity states that the variance of the residuals (errors) should be constant across all levels of the independent variable. This means the spread of the data points around the regression line should be approximately equal.

Violations of this assumption, known as heteroscedasticity, can lead to biased standard errors and impact the reliability of statistical inferences drawn from the model.





Assumption 4: Normality of Residuals

1 Defining Normality

The normality assumption states that the residuals, or the differences between the observed values and the predicted values, should be normally distributed.

2 Importance of Normality

The normality assumption is crucial because it ensures the validity of statistical inference, such as hypothesis testing and confidence interval estimation.

3 Checking for Normality

Analysts can use various techniques, such as visual inspection of histograms, Q-Q plots, or statistical tests like the Shapiro-Wilk test, to assess the normality of residuals.

Importance of Checking Assumptions

Ensures Validity

Checking the assumptions of linear regression is crucial to ensure the validity and reliability of the analysis. Violations of these assumptions can lead to biased and unreliable results.

Guides Interpretation

Understanding the assumptions helps interpret the regression results correctly and draw meaningful conclusions about the relationships between the variables.

Informs Diagnostics

Assumption checking provides insights into potential issues in the data or the model, guiding further diagnostics and remedial actions.

Supports Predictions

Adherence to assumptions enables more accurate predictions using the regression model, which is crucial for decision-making and forecasting.

Consequences of Violating Assumptions

1

Biased Parameter Estimates

If assumptions are violated, the regression coefficients will be biased, leading to incorrect interpretations and predictions.

2

Inefficient Standard Errors

Violation of assumptions can result in inefficient standard errors, making it difficult to draw valid statistical inferences.

3

Unreliable Hypothesis Testing

Violating assumptions can lead to invalid hypothesis testing, resulting in incorrect conclusions about the significance of the relationships.

4

Poor Predictive Accuracy

When assumptions are violated, the model's ability to accurately predict future outcomes is compromised, reducing its practical value.

Techniques for Checking Assumptions

Exploratory Data Analysis

Visualize the data, check for outliers, and assess the distribution of variables to identify potential violations of assumptions.

Statistical Tests

Conduct formal statistical tests, such as the Shapiro-Wilk test for normality or the Breusch-Pagan test for homoscedasticity, to quantify the degree of assumption violation.

Residual Plots

Inspect residual plots to visually assess the linearity, independence, and homoscedasticity of the model's errors.

Visual Inspection

Examine the data visually, such as scatter plots and histograms, to identify potential issues with the assumptions.

Remedial Measures for Assumption Violations

1

Transformation

Apply appropriate data transformations to address linearity or homoscedasticity issues.

2

Variable Selection

Identify and remove independent variables that violate assumptions.

3

Model Adjustment

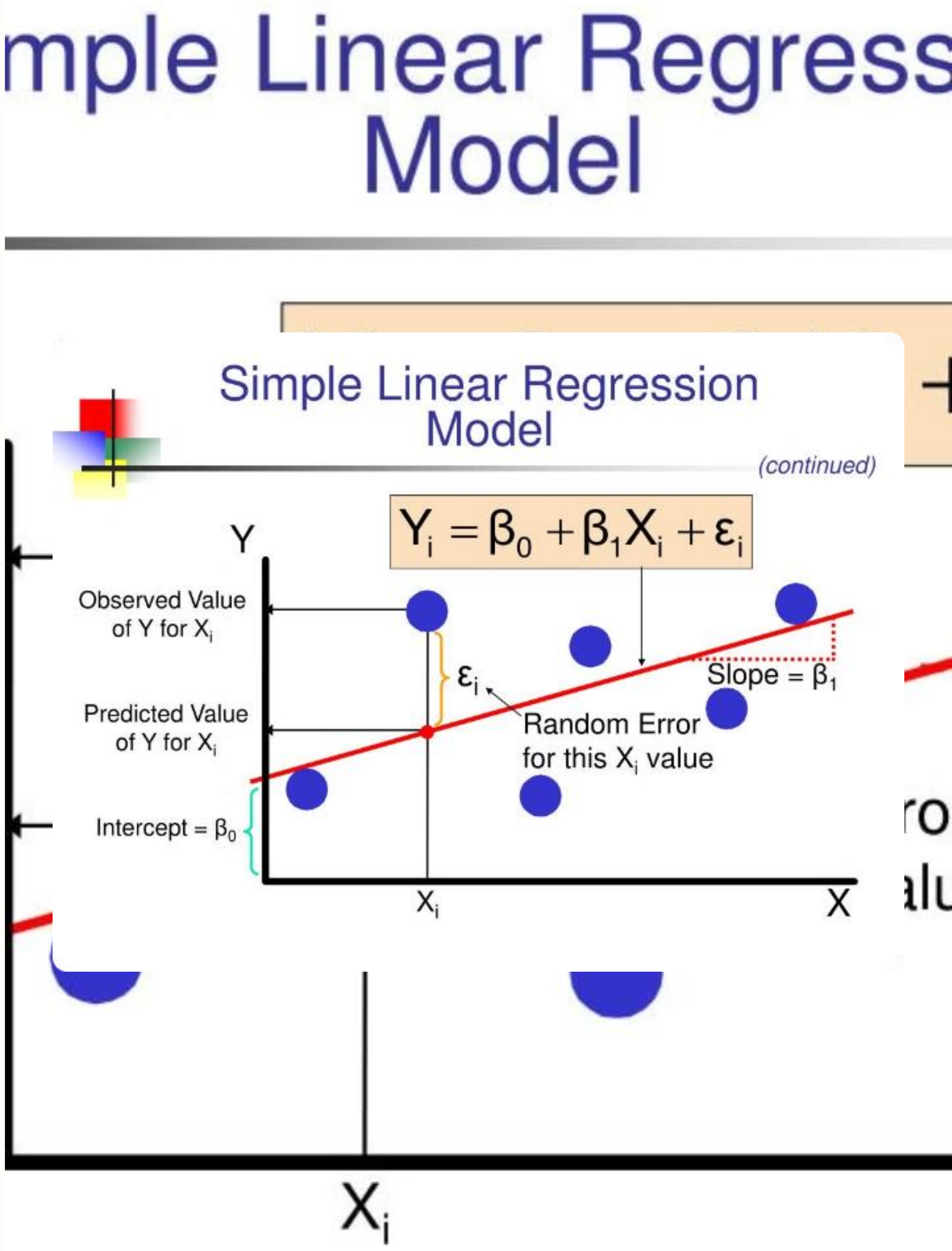
Modify the regression model to accommodate assumption violations, e.g. using robust or non-parametric methods.

When the assumptions of linear regression are violated, there are several remedial measures that can be taken to address the issues. These include transforming the data, carefully selecting variables, and adjusting the regression model itself to better fit the underlying data structure.

Conclusion and Key Takeaways

In conclusion, the key assumptions of linear regression are critical for ensuring the validity and reliability of the model. Adherence to these assumptions, such as linearity, independence, homoscedasticity, and normality of residuals, is essential for making accurate predictions and drawing meaningful insights from the data.

By thoroughly checking and addressing any violations of these assumptions, researchers and data analysts can have confidence in the robustness of their linear regression analyses. Ignoring these assumptions can lead to biased and unreliable results, underscoring the importance of diligent diagnostic testing and appropriate remedial measures.



mple Linear Regress
Model
Simple Linear Regression Model
(continued)
$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
Y
Observed Value of Y for X_i
ε_i
Slope = β_1
Predicted Value of Y for X_i
Random Error for this X_i value
Intercept = β_0
X_i
X
ro
alu
X_i

Thank You