



Deep Dive into Linear Regression



Pre-requisites

Hope you have gone through the self-learning content for this session on the PRISM portal.



Recap



By the End of This session, You Will:

- Learn Linear regression Model Evaluation Metrics
- Understand Linear regression model performance metrics
- Explore Model diagnostic plots
- Perform a case study on Linear regression

What's In It For Me?

Gain a deep understanding of various evaluation metrics used to measure the performance of Linear Regression models.

Explore diagnostic plots that provide insights into your model's behavior, enabling you to identify strengths and areas for improvement.

Engage in a comprehensive case study involving Linear Regression, applying your newfound knowledge to solve practical problems.

Acquire hands-on experience through activities and projects that reinforce your learning and build a portfolio of practical skills.

Poll Time

Q. What can you expect to gain from doing this session on Linear Regression and Model Evaluation Metrics?

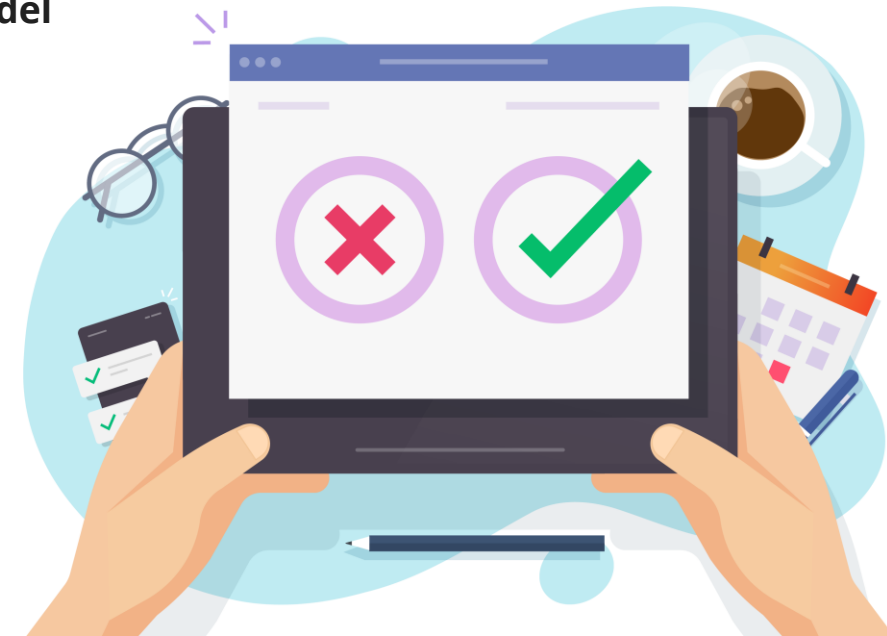
- a. Advanced topics in mathematics and theoretical concepts only
- b. Practical skills in building Linear Regression models, understanding model performance, and diagnostic techniques
- c. A certificate for participating in the course, but no real-world application
- d. Limited engagement with course instructors and no hands-on projects



Poll Time

Q. What can you expect to gain from doing this session on Linear Regression and Model Evaluation Metrics?

- a. Advanced topics in mathematics and theoretical concepts only
- b. Practical skills in building Linear Regression models, understanding model performance, and diagnostic techniques**
- c. A certificate for participating in the course, but no real-world application
- d. Limited engagement with course instructors and no hands-on projects





Linear Regression : Model Evaluation Metrics

Model evaluation Metrics

Model evaluation metrics refer to specific measurements and criteria used to assess the performance and quality of a linear regression model.

These metrics help you understand how well the model fits the data, how accurate its predictions are, and whether it meets your modeling goals.

The choice of evaluation metrics depends on the nature of the problem and the objectives of the analysis.

Here are some common model evaluation metrics used in linear regression: R^2 , Adjusted R^2 , and AIC.

R²

Definition

R-squared, also known as the coefficient of determination, is used to understand how well a linear regression model fits your data. R-squared measures how well your model's predictions match the actual values.

Usage

It is used to assess the proportion of the variance in dependent variable explained by independent variables.

Formula

Here's the simple intuitive formula for R-squared:

$$R^2 = 1 - \text{Unexplained Variability} / \text{Total Variability}$$

Significance

Here variability refers to the extent to which data points or values in a dataset differ from each other. Unexplained variability is like diversity that the model is not able to predict properly.

Range of R²

The range of R-squared is between 0 and 1. A higher R-squared indicates that a larger proportion of the variance in the dependent variable is explained by the independent variables.

Understanding R² calculation

- The mathematical formula for calculating R-squared is as follows:

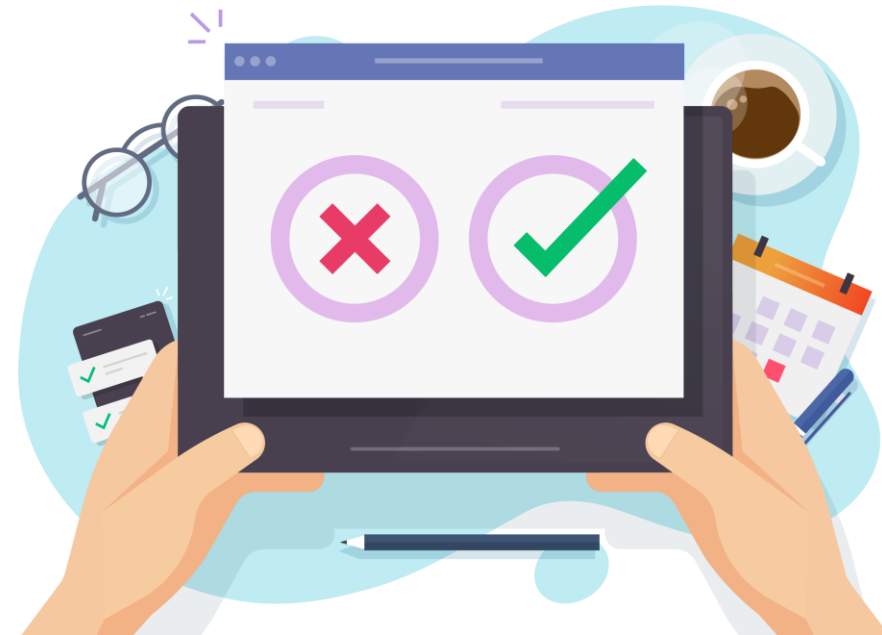
$$R^2 = 1 - SSR / SST$$

- *SSR* is the sum of squared residuals, which represents the sum of the squared differences between actual values and predicted values for each observation.
- *SST* is the total sum of squares, which represents the sum of squared differences between actual values and the mean of actual values.
- If the model's predictions are much better than the simple average, R-squared will be closer to 1.
- If the model doesn't improve upon the simple average much, R-squared will be closer to 0.

Poll Time

Q. What does R-squared (R^2) measure in the context of linear regression?

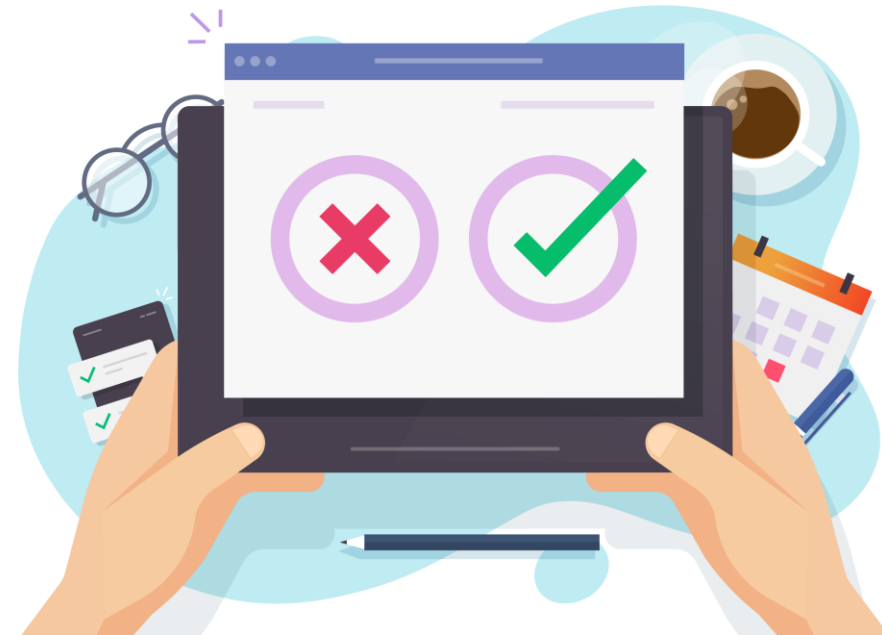
- a. The number of predictors used in the regression model
- b. The strength of the relationship between independent and dependent variables
- c. The probability that the regression model is accurate
- d. The mean value of the dependent variable



Poll Time

Q. What does R-squared (R^2) measure in the context of linear regression?

- a. The number of predictors used in the regression model
- b. The strength of the relationship between independent and dependent variables**
- c. The probability that the regression model is accurate
- d. The mean value of the dependent variable



Adjusted R²

- Adjusted R-squared is a modified version of the R-squared metric used in linear regression.
- Adjusted R-squared considers the number of independent variables in the model and adjusts the R-squared value accordingly.
- The formula for calculating Adjusted R-squared is:
Where:
 - R^2 is the regular R-squared value.
 - n is the number of observations in the dataset. p is number of independent variables in the model.
- The adjustment penalizes the R-squared value when more predictors are added, helping to account for the potential increase in R-squared due to randomness.
- Adjusted R^2 helps prevent overfitting. Overfitting occurs when a model fits the training data too closely, including noise and random fluctuations. A higher number of variables can lead to overfitting, resulting in a higher R-squared but poorer predictive performance on new data.

$$\text{Adjusted } R^2 = 1 - \frac{(1-R^2)(n-1)}{n-p-1}$$

AIC : Akaike's Information Criteria

- AIC is a statistical measure used for model selection and comparison in the context of linear regression and other statistical models.
- It helps balance model complexity and goodness of fit by penalizing models with a larger number of parameters.
- AIC is used to compare different models to determine which one provides the best trade-off between fitting the data well and avoiding overfitting.
- When comparing models, the model with the lowest AIC is often preferred because it provides a good balance between fit and complexity.
- In linear regression, AIC can help you choose the most appropriate subset of variables for your model. It guides you in selecting a model that captures important relationships while avoiding unnecessary complexity.

Pop Quiz

Q. If a model has a higher Adjusted R-squared compared to another model, it means _____.

- a. The first model is better regardless of the number of predictors
- b. The first model is better at fitting the data, but might be overfitting
- c. The second model is better because it has more predictors
- d. The first model is better because it explains a larger proportion of the variance



Pop Quiz

Q. If a model has a higher Adjusted R-squared compared to another model, it means _____.

- a. The first model is better regardless of the number of predictors
- b. The first model is better at fitting the data, but might be overfitting
- c. The second model is better because it has more predictors**
- d. The first model is better because it explains a larger proportion of the variance







Linear Regression: Model performance Metrics

Model Performance Metrics

- A model performance metric in linear regression refers to a numerical measure or statistic that is used to evaluate and quantify how well a linear regression model is performing.
- These metrics help to understand :
 - How Accurate are the Predictions?
 - How Well Does the Model Fit the Data?
 - Is the Model Overfitting or Underfitting?
 - Which Variables are Important?
- Common model performance metrics in linear regression include:
 - SSE (Sum of Squared Errors)
 - MSE (Mean of Squared Errors)
 - RMSE (Root Mean Squared Error)
 - MAE (Mean Absolute Error)
 - MAPE (Mean Absolute Percentage Error)

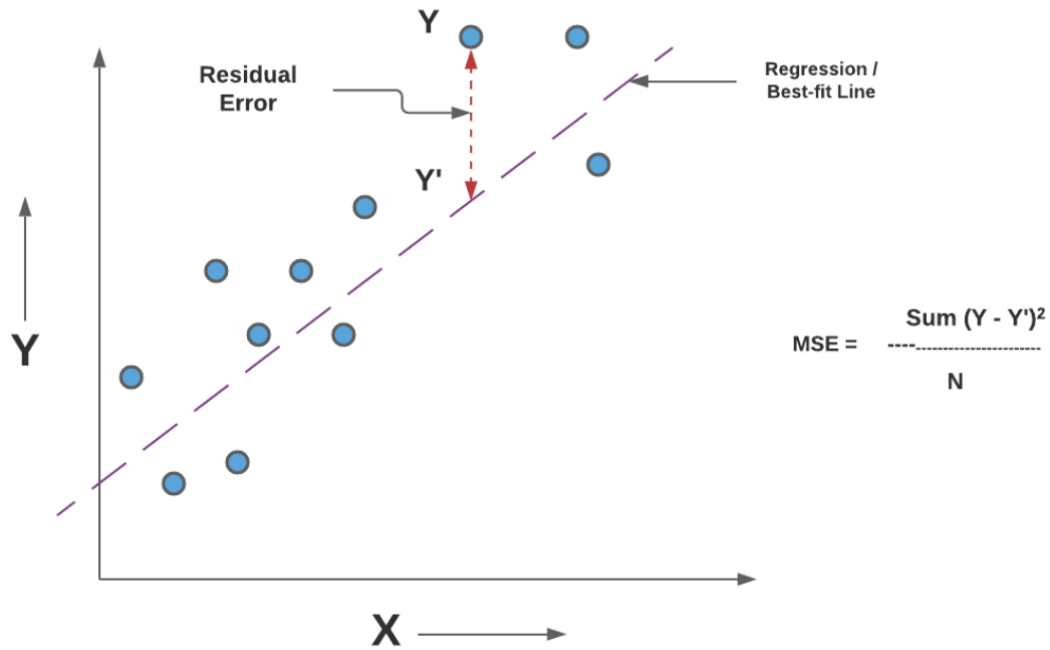
SSE

- SSE stands for "Sum of Squared Errors". It represents the sum of the squared differences between actual values and predicted values generated by a linear regression model.
- Mathematically, the formula for calculating SSE is as follows:

Where:

- n is the number of observations (data points).
 - y_i is the actual value of the dependent variable for the i th observation.
 - \hat{y}_i is the predicted value of the dependent variable for the i th observation.
- $$\text{SSE} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$
- SSE represents the sum of the squared vertical distances between each observed data point and its corresponding predicted value on the regression line.
 - SSE doesn't provide a direct measure of model's complexity. Models with more parameters can achieve lower SSE values, but they might be overfitting the noise in the data.

MSE



- MSE stands for "Mean of Squared Errors". It represents average squared difference between the predicted values and the actual values.
- Mathematically, the formula for calculating MSE is as follows:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Where:

- n is the number of observations (data points).
- y_i is the actual value of the dependent variable for the i th observation.
- \hat{y}_i is the predicted value of the dependent variable for the i th observation.

Pop Quiz

Q. Which of the following is true about SSE?

- a. It measures the average absolute difference between observed values and predicted values
- b. It is always a positive value
- c. Lower SSE values indicate worse model fit
- d. It is the sum of the squared differences between predicted values and observed values



Pop Quiz

Q. Which of the following is true about SSE?

- a. It measures the average absolute difference between observed values and predicted values
- b. It is always a positive value
- c. Lower SSE values indicate worse model fit
- d. It is the sum of the squared differences between predicted values and observed values**



RMSE

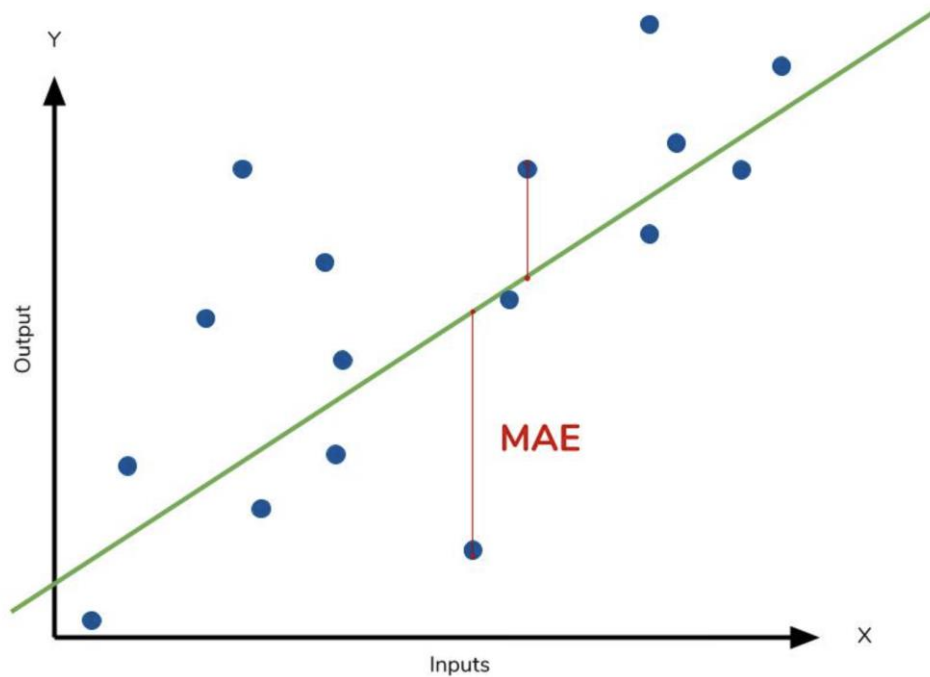
- RMSE stands for "Root Mean Squared Error," and it's a widely used metric in statistics and machine learning to assess the accuracy of predictions made by a model.
- RMSE is closely related to Mean Squared Error (MSE), but it provides a more interpretable measure by taking the square root of the MSE.
- Mathematically, the formula for calculating RMSE is as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

Where:

- n is the number of observations (data points).
 - y_i is the actual value of the dependent variable for the i th observation.
 - \hat{y}_i is the predicted value of the dependent variable for the i th observation.
- RMSE is in the same unit as the dependent variable, making it more interpretable and relatable to the original data.

MAE



- MAE stands for "Mean Absolute Error," and is calculated by taking average of absolute differences between predicted values and actual values.
- Mathematically, the formula for calculating RMSE is as follows:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

- Unlike squared differences in MSE and RMSE, MAE uses the absolute value of the differences, which means it treats positive and negative errors equally.

MAPE

- MAPE stands for "Mean Absolute Percentage Error," and is calculated by taking average of absolute percentage differences between predicted values and actual values, expressed as a percentage.
- Mathematically, the formula for calculating RMSE is as follows:

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100$$

Where:

- n is the number of observations (data points).
- y_i is the actual value of the dependent variable for the i th observation.
- \hat{y}_i is the predicted value of the dependent variable for the i th observation.
- It measures the average magnitude of errors as a percentage of the actual values.
- MAPE can be problematic when actual values are close to zero, as the percentage division may lead to undefined values or large percentages.

Pop Quiz

Q. RMSE is calculated as the square root of which of the following metrics?

- a. MSE
- b. MAE
- c. MAPE
- d. SSE



Pop Quiz

Q. RMSE is calculated as the square root of which of the following metrics?

- ☒ a. **MSE**
- b. MAE
- c. MAPE
- d. SSE

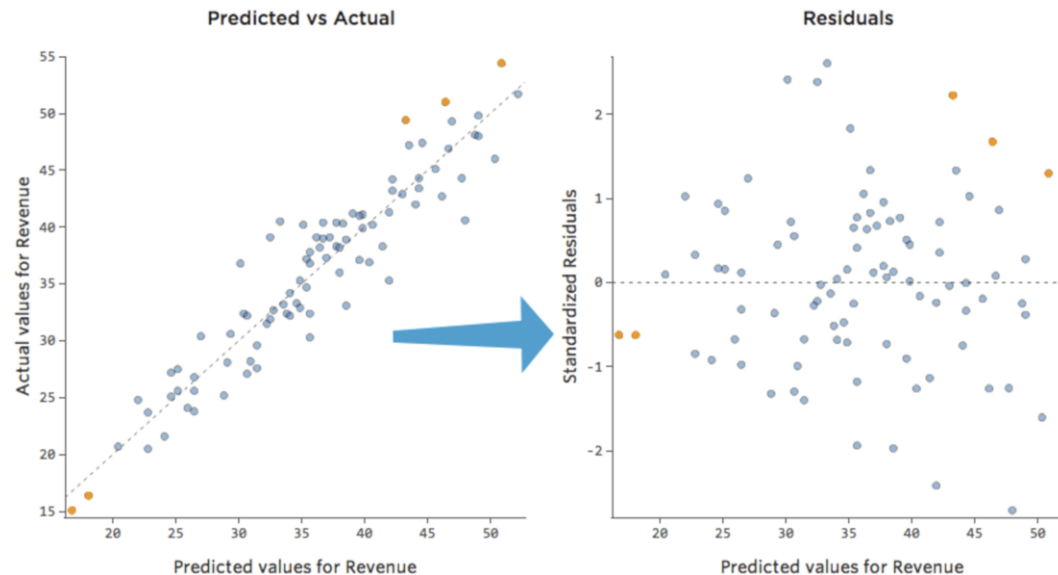




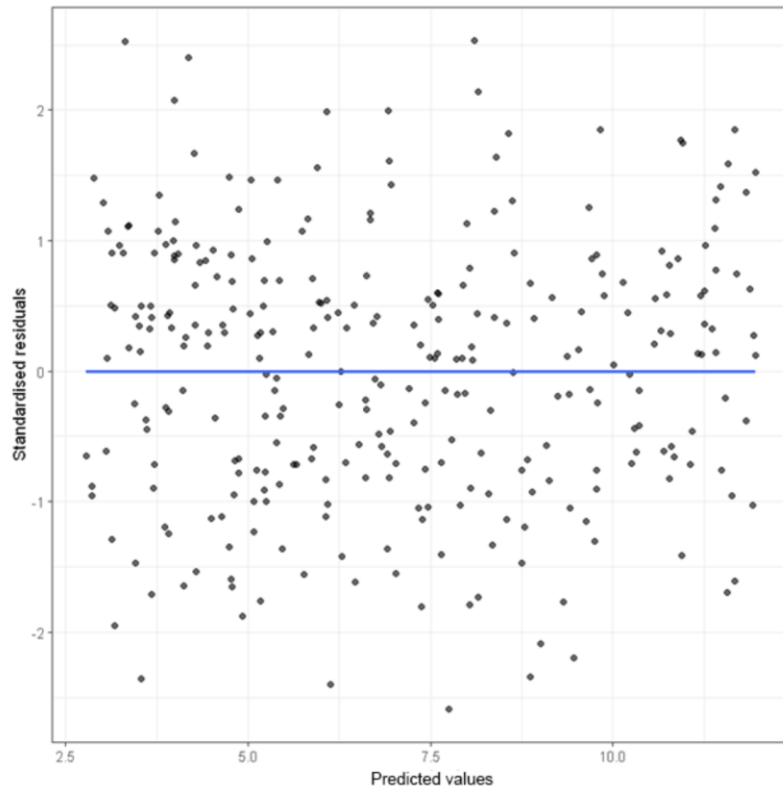
Linear Regression : Model Diagnostic Plots

Checking OLS assumptions using Residual plots

- Checking OLS (Ordinary Least Squares) assumptions using residual plots is a crucial step in validating the assumptions of a linear regression model.
- Residual plots help you identify potential issues or violations of the assumptions, allowing you to make informed decisions about the model's reliability and validity.



Residual plots checks



- If the points are randomly scattered around the horizontal line at 0, the linearity assumption is likely satisfied.
- If there is no discernible pattern or trend in the plot, then the residuals are not dependent on the order of observation meaning no autocorrelation.
- Look for a consistent spread of residuals along the horizontal line at 0. If the spread remains roughly constant, homoscedasticity is likely met.
- Create a histogram of the residuals and compare it to a normal distribution. If the histogram closely resembles a normal distribution, the assumption is satisfied.

Pop Quiz

Q. In a residual vs. fitted values plot, if the residuals are randomly scattered around the horizontal line at 0, it suggests:

- a. Heteroscedasticity
- b. Homoscedasticity
- c. Multicollinearity
- d. Normality of residuals



Pop Quiz

Q. In a residual vs. fitted values plot, if the residuals are randomly scattered around the horizontal line at 0, it suggests:

- a. Heteroscedasticity
- b. Homoscedasticity**
- c. Multicollinearity
- d. Normality of residuals





Activity 1

Pre-requisites:

Familiarity with calculations of linear regression evaluation metrics.

Scenario:

You are a data scientist at a real estate firm, and you are analyzing a dataset of houses to understand how well a linear regression model explains the variation in housing prices using the R-squared and adjusted R-squared metrics.

Data:

Housing Prices (in thousands of dollars): [250, 300, 180, 220, 350, 380, 400]

Predicted Prices (from a linear regression model): [260, 320, 200, 240, 340, 370, 410]

Number of Observations (n): 7

Expected Outcome:

Calculate and compare the R-squared and adjusted R-squared values for your linear regression model. These metrics will provide insights into how much of the variation in housing prices is explained by the model.

Steps:

- 1) Calculate R-squared value using actual and predicted house price values
- 2) Calculate adjusted R-squared value using actual and predicted house price values
- 3) Compare and interpret of R-squared and adjusted R-squared values to understand their implications in terms of model fit and explanatory power

Activity 2

Pre-requisites:

Familiarity with calculations of linear regression performance metrics.

Scenario:

As a data enthusiast, you have collected temperature predictions from two different weather apps for a specific day. Your goal is to determine which app provides more reliable temperature forecasts using linear regression evaluation metrics.

Data:

Actual temperature for the day: 72°F

App A's predicted temperatures: [70°F, 74°F, 73°F, 71°F]

App B's predicted temperatures: [68°F, 72°F, 75°F, 70°F]

Expected Outcome:

Calculate and compare MSE, RMSE, and MAE for both weather apps. These metrics will help you assess which app's predictions are closer to the actual temperatures.

Steps:

- 1) Calculate MSE value for both apps using actual and predicted temperature values
- 2) Calculate RMSE value for both apps using actual and predicted temperature values
- 3) Calculate MAE value for both apps using actual and predicted temperature values
- 4) Compare and interpret of MSE, RMSE and MAE values
- 5) Conclude to determine which weather app provides more accurate temperature predictions

Summary



R-squared (R^2) and adjusted R-squared are significant model performance metrics that help assess variance explanation of regression models.



Adjusted R-squared looks for the number of predictors and adjusts for the potential inclusion of irrelevant predictors, offering a more accurate representation of the model's explanatory power.



Model performance metrics like MAE, MSE, RMSE help in quantifying the accuracy of predictive models.



Model diagnostic plots help detect potential model assumption deviations.

Next Session:

Case Study on Linear Regression

THANK YOU

Please complete your assessments and review the self-learning content for this session on the **PRISM** portal.



knowledgehut
upGrad



Case Study on Linear Regression



Pre-requisites

Hope you have gone through the self-learning content for this session on the PRISM portal.



By the End of This Session, You Will:

- Learn the significance of linear regression in making predictions about real world problems
- Explore real world data using pre-processing steps before creating linear regression model
- Check if the assumptions of the linear regression are meeting or not
- Find best performing model by comparing model evaluation and performance metrics



Recap

Poll Time

Q. What is the significance of linear regression?

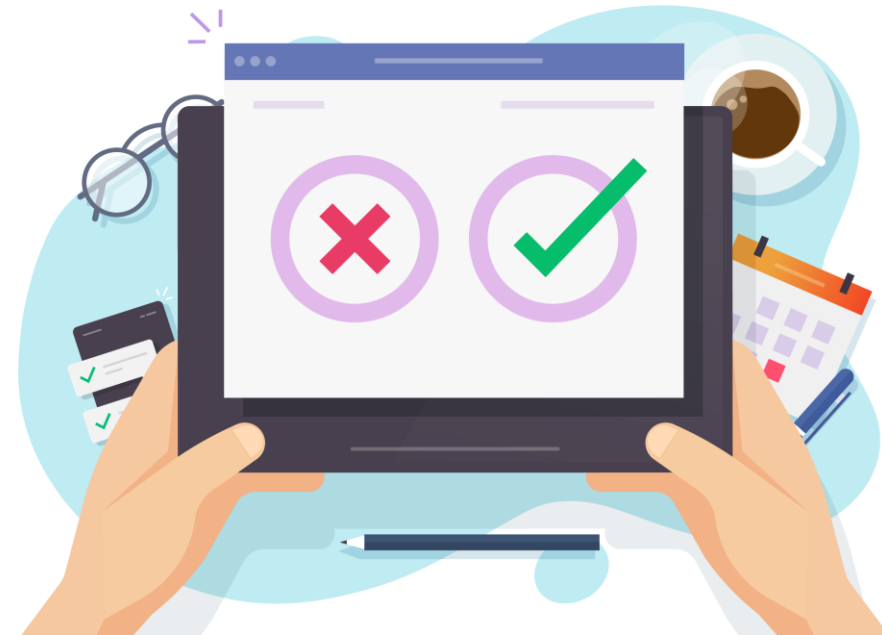
- a. Linear regression is used solely for visualizing data patterns
- b. Linear regression helps in predicting future events with absolute certainty
- c. Linear regression provides a statistical method to model relationships between variables
- d. Linear regression is used to categorize data into discrete groups



Poll Time

Q. What is the significance of linear regression?

- a. Linear regression is used solely for visualizing data patterns
- b. Linear regression helps in predicting future events with absolute certainty
- c. Linear regression provides a statistical method to model relationships between variables**
- d. Linear regression is used to categorize data into discrete groups





Case Study on Linear Regression

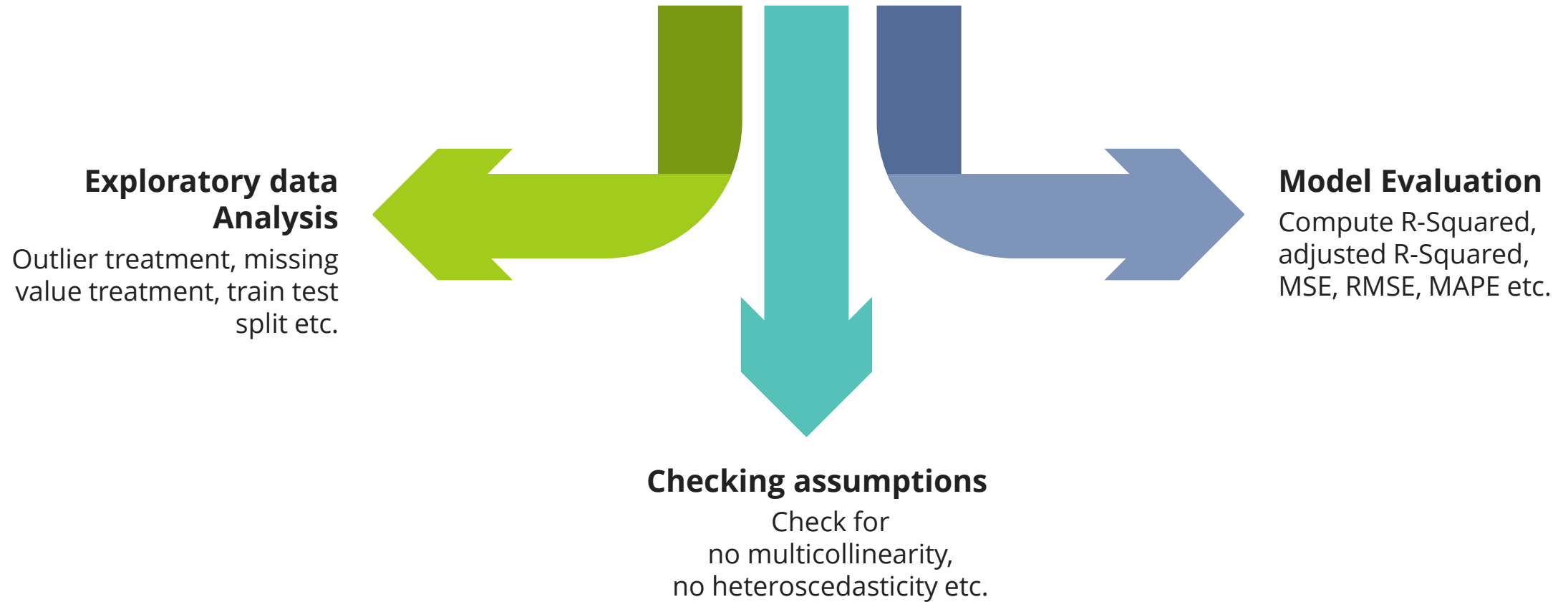


Case Study – Problem Statement

Problem Statement

- One of the Indian companies sells the old/refurbished cars to the customers.
- It is looking to predict the car price based on the various features of the car.
- To achieve this goal, they require an extensive analysis of their historical data having specifications of various cars and their selling price.
- The dataset comprises several columns, including the car's company, name of the car model, kms driven, selling price, etc., and information on 1700+ cars.
- Based on the features of the car, this company would like to predict the car price and quote to the customers who are looking to sell their car based on their car's features.
- In this business problem, they need to your help to apply the linear regression technique to create a model that can predict car price and help the company take care of future customers.

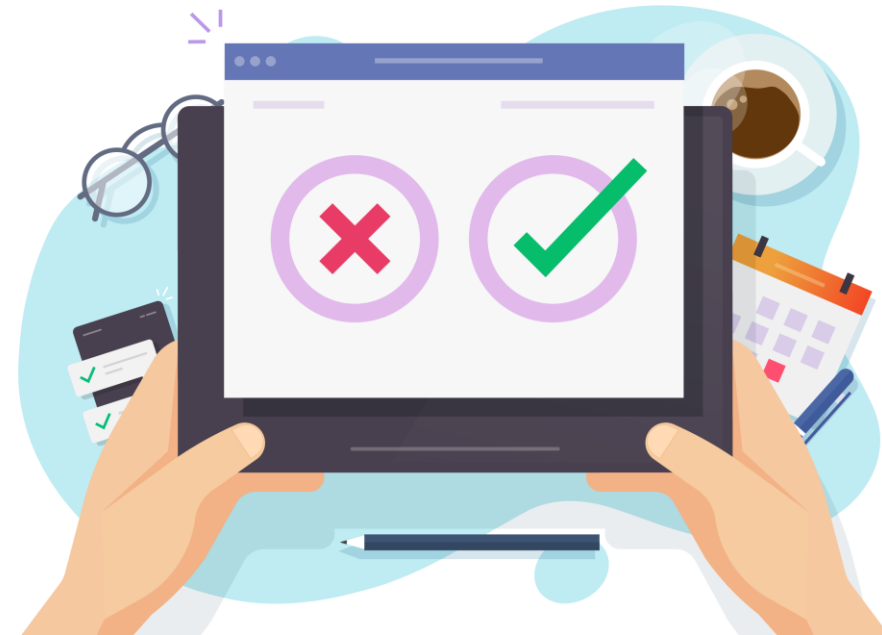
Areas to Focus



Poll Time

Q. Which of the following metrics gives the average percentage difference between predicted and actual values in a linear regression model?

- a. R-squared (R^2)
- b. Mean Absolute Percentage Error (MAPE)
- c. Root Mean Squared Error (RMSE)
- d. Adjusted R-squared



Poll Time

Q. Which of the following metrics gives the average percentage difference between predicted and actual values in a linear regression model?

- a. R-squared (R^2)
- b. Mean Absolute Percentage Error (MAPE)**
- c. Root Mean Squared Error (RMSE)
- d. Adjusted R-squared







Hands-on: Case Study Questions

Poll Time


Q. Which metric considers the relative size of prediction errors to the actual values in a linear regression model?

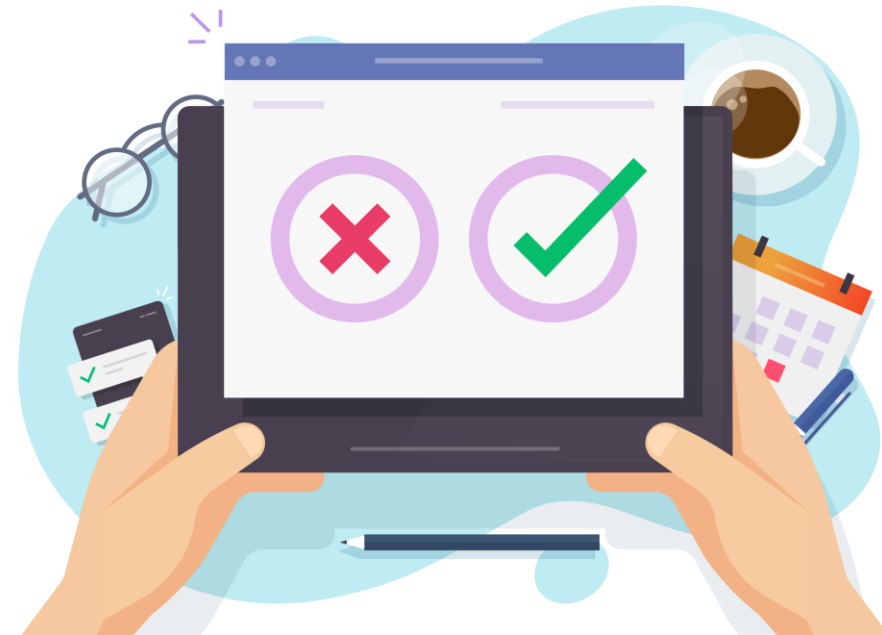
- a. R-squared (R^2)
- b. Mean Absolute Error (MAE)
- c. Root Mean Squared Error (RMSE)
- d. Mean Absolute Percentage Error (MAPE)



Poll Time

Q. Which metric considers the relative size of prediction errors to the actual values in a linear regression model?

- a. R-squared (R^2)
- b. Mean Absolute Error (MAE)
- c. Root Mean Squared Error (RMSE)
-  d. **Mean Absolute Percentage Error (MAPE)**





Activity 1

Pre-requisites:

Familiarity with Python pandas library and linear regression concepts.

Scenario:

You are a data enthusiast, and you have a very small dataset. You want to load it in Python using the pandas library to build a linear regression model. You will calculate the R-squared and Adjusted R-squared values to evaluate the model's goodness of fit.

Data:

```
import pandas as pd
df = pd.DataFrame({'Study Hours': [2, 3, 4, 5, 6], 'Exam Score': [65, 78, 82, 90, 95]})
```

Expected Outcome:

Built a linear regression model using the dataset. Calculated R-squared and Adjusted R-squared values and gain insights into the model's fit.

Steps:

- 1) Load and explore the dataset.
- 2) Split the dataset into the feature (study hours) and target (exam score) variables.
- 3) Build a linear regression model using `LinearRegression` class from scikit-learn.
- 4) Calculate R-squared and adjusted R-squared values.
- 5) Compare and interpret of R-squared and adjusted R-squared values to understand their implications in terms of model fit.

Activity 2

Pre-requisites:

Familiarity with Python pandas library and linear regression evaluation metrics.

Scenario:

Imagine you are a data analyst working for a marketing agency. Your client, a retail company, is seeking insights into the relationship between advertising expenses and sales revenue. They believe that understanding this correlation will help them optimize their advertising budget allocation to maximize revenue.

Data:

```
import pandas as pd
df = pd.DataFrame({'Advertising Expense (in thousands)': [2, 3, 4, 5, 6], 'Sales Revenue (in thousands)': [50, 75, 100, 120, 150]})
```

Expected Outcome:

Built a linear regression model using the dataset. By quantifying the model's accuracy using various evaluation metrics like RMSE, MSE etc., you will enable your client to make informed decisions about their marketing strategies.

Steps:

- 1) Load and explore the dataset
- 2) Split the dataset into the feature (advertising expense) and target (sales revenue) variables.
- 3) Build a linear regression model using the `LinearRegression` class from scikit-learn.
- 4) Calculate different evaluation metrics like SSE, MSE, RMSE, MAE, MAPE in python.
- 5) Interpret these metrics to quantify the accuracy and reliability of the model's predictions.
- 6) Discuss the implications of the evaluation metrics on the company's decision-making process.

Summary



During pre-processing, missing value treatment helps ensure data completeness and outlier treatment helps to prevent skewing of the model.



Checking assumptions of linear regression like linearity, homoscedasticity, and normality of residuals helps to ensure the validity of model results.



Linear regression helps in identifying key variables impacting the variable we are trying to predict.



R-Squared and Adjusted R-Squared values help assess the model's explained variance.

Session Feedback



Next Session:
Regularization

THANK YOU!

Please complete your assessments and review the self-learning content for this session on the **PRISM** portal.

