Unit II Regression

Introduction- Regression, Need of Regression, Difference between Regression and Correlation

What is Regression?

Regression is a statistical method used to establish a relationship between a dependent variable (target variable) and one or more independent variables (predictor variables). It helps us understand how the independent variables influence the dependent variable.

Example: Predicting house prices based on factors like size, location, and number of bedrooms.

Need for Regression

Regression analysis is crucial in various fields due to its variables and how changes in independent variables impact the dependent variable.

- Control: Assess the impact of controlling or manipulating independent variables on the dependent variable.
- **Decision variables, they serve different purposes:

Feature	Regression	Correlation
Goal	Predict dependent variable	Measure strength of association
Output	Equation or model	Correlation coefficient (-1 to 1)
Causation	Can imply causation (with careful analysis)	No implication of causation

Export to Sheets

two variables, while regression builds a model to predict one variable based on another.

Would you like to delve deeper into different types of regression or discuss real-world applications?

Types of Regression: Univariate vs. Multivariate, Linear vs. Nonlinear, Simple Linear vs. Multiple Linear, Bias-Variance tradeoff, Overfitting and Underfitting

Types of Regression

Based on the number of variables

- **Univariate Regression: Involves predicting a single dependent variable based on one independent variable.**

- Multivariate Regression: Predicts a single dependent variable based on multiple independent variables.

Based on the relationship between variables

- Linear Regression: Assumes a linear relationship between the dependent and independent variables.
- Non-linear Regression: Handles relationships that are not linear, often using polynomial or other complex functions.

Combining the above

- Simple Linear Regression: Predicts a single dependent variable based on one independent variable using a linear model.
- Multiple Linear Regression: Predicts a single dependent variable based on multiple independent variables using a linear model.

Bias-Variance Trade-off

This concept is crucial in model building.

- Bias: The error introduced by approximating a real-world problem by a simpler model. High bias can lead to underfitting.
- Variance: The model's sensitivity to fluctuations in the training data. High variance can lead to overfitting.

The ideal model balances bias and variance.

Overfitting and Underfitting

- Overfitting: Occurs when a model is too complex and captures noise in the data instead of the underlying pattern. It performs well on training data but poorly on new data.
- Underfitting: Happens when a model is too simple and cannot capture the underlying pattern in the data. It performs poorly on both training and new data.

Regression Techniques - Polynomial Regression, Stepwise Regression, Decision Tree Regression, Random Forest Regression

Regression Techniques

Polynomial Regression

- Assumption: Non-linear relationship between dependent and independent variables.

- Approach: Fits a polynomial curve to the data.
- Advantages: Flexible to capture complex patterns.
- Disadvantages: Prone to overfitting, especially with high-degree polynomials.

Stepwise Regression

- Goal: Builds a model by sequentially adding or removing predictors.
- Methods: Forward selection, backward elimination, or both.
- Advantages: Helps in feature selection, reduces model complexity.
- Disadvantages: Can be unstable, might not find the optimal model.

Decision Tree Regression

- Approach: Creates a tree-like model where each internal node represents a feature, branches represent decision rules, and leaves represent predicted values.
- Advantages: Handles both numerical and categorical data, easy to interpret.
- Disadvantages: Prone to overfitting, sensitive to small changes in data.

Random Forest Regression

- Ensemble technique: Combines multiple decision trees.
- Advantages: Reduces overfitting, handles various data types, high accuracy.

. Disadvantages: Can be computationally expensive, less interpretable than a single decision tree

Support Vector Regression, Ridge Regression, Lasso Regression, ElasticNet Regression, Bayesian Linear Regression.

Advanced Regression Techniques

These techniques often address challenges like overfitting, multicollinearity, and model complexity.

Support Vector Regression (SVR)

- Based on: Support Vector Machines (SVM)
- Approach: Finds the optimal hyperplane to fit the data points within a specified epsilon-tube.
- Advantages: Effective in handling non-linear relationships through kernel trick, robust to outliers.
- Disadvantages: Can be computationally expensive for large datasets, tuning hyperparameters is crucial.

Ridge Regression

- **Goal: Addresses multicollinearity by adding a penalty term to the loss function.**
- **Approach: Shrinks coefficients towards zero but doesn't eliminate them completely.**
- **Advantages: Improves model stability, reduces overfitting.**
- **Disadvantages: Doesn't perform feature selection.**

Lasso Regression

- **Goal: Performs feature selection and regularization.**
- **Approach: Adds L1 penalty to the loss function, which can shrink coefficients to zero.**
- **Advantages: Creates sparse models (many coefficients become zero), interpretable.**
- **Disadvantages: Can be unstable, might not select the correct features in some cases.**

ElasticNet Regression

- **Combines: Ridge and Lasso regression.**
- **Approach: Uses a combination of L1 and L2 penalties.**
- **Advantages: Balances feature selection and stability, often performs better than Ridge or Lasso alone.**
- **Disadvantages: More complex than Ridge or Lasso.**

Bayesian Linear Regression

- **Approach: Treats model parameters as random variables with probability distributions.**
- **Advantages: Incorporates uncertainty in predictions, provides probabilistic interpretations.**
- **Disadvantages: Computationally intensive, requires careful modeling of prior distributions.**

Evaluation Metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean .Squared Error (RMSE),R-squared ,Adjusted R-squared.

Mean Squared Error (MSE)

- Definition: The average of the squared differences between predicted and actual values.
- Formula:

- MSE = Σ(y_pred- y_actual)^2 / n

where:

  - y_pred is the predicted value
  - y_actual is the actual value
  - n is the number of data points

- Characteristics:
-
  - Penalizes larger errors more severely due to squaring.
  - Sensitive to outliers.
  - Differentiable, making it suitable for optimization algorithms.

Mean Absolute Error (MAE)

- Definition: The average of the absolute differences between predicted and actual values.
- Formula:
- MAE = Σ|y_pred- y_actual| / n
- Characteristics:
  - Treats all errors equally.
  - Less sensitive to outliers compared to MSE.
  - Not differentiable at zero, which can pose challenges for some optimization algorithms.

Root Mean Squared Error (RMSE)

- Definition: The square root of the average of the squared differences between the predicted and actual values.
- Interpretation: Measures the average magnitude of the error in your predictions. A lower RMSE indicates better model performance.
- Formula:
- RMSE = sqrt(Σ(y_pred- y_actual)^2 / n)

where:

  - y_pred is the predicted value
  - y_actual is the actual value
  - n is the number of data points

R-squared

- Definition: Represents the proportion of variance in the dependent variable that is explained by the independent variables.
- Interpretation: Measures how well the regression model fits the observed data. A higher R-squared indicates a better fit.
- Formula:
- R^2 = 1- (SS_res / SS_tot)

  where:

  - SS_res is the sum of squared residuals
  - SS_tot is the total sum of squares

Adjusted R-squared

- Definition: Similar to R-squared but penalizes the addition of unnecessary predictors.
- Interpretation: Provides a more realistic assessment of the model's performance, especially when comparing models with different numbers of predictors.
- Formula:
- Adjusted R^2 = 1- [(1-R^2)(n-1)/(n-p-1)]

where:

- n is the number of data points
- p is the number of predictors