Linear Regression: Theory and Case Study

August 30, 2024

Introduction to Linear Regression

- Linear regression is a statistical method for modeling the relationship between a dependent variable and one or more independent variables.
- It is widely used in predictive modeling, data analysis, and machine learning.
- Common applications include predicting housing prices, stock market trends, and sales forecasting.

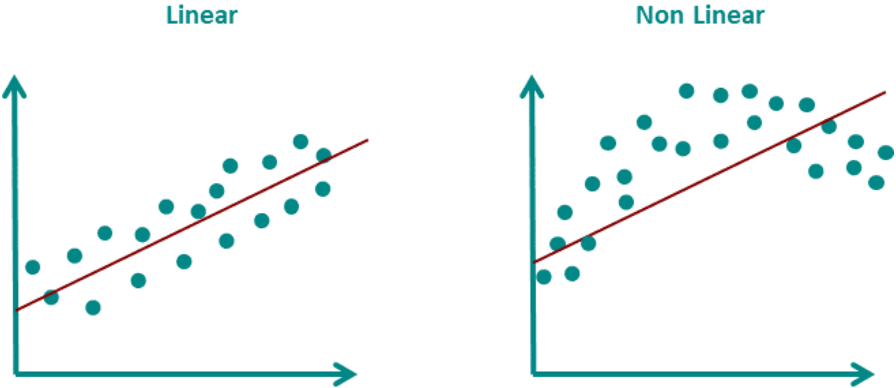
Linear Regression Model

Model Equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon \tag{1}$$

- y: Dependent variable (response)
- x_i: Independent variables (predictors)
- β_0: Intercept term
- β_i: Coefficients (weights)
- ϵ: Error term (residuals)

Linear Regression: Illustration



Assumptions of Linear Regression

- Linearity: The relationship between predictors and the response is linear.
- Independence: The residuals are independent.
- Homoscedasticity: The residuals have constant variance.
- Normality: The residuals are normally distributed.

Ordinary Least Squares (OLS)

Objective: Minimize the sum of squared residuals (errors).

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 \tag{2}$$

where:

- $\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_n x_{in}$
- $\hat{\beta}$: Estimated coefficients

Case Study: Predicting Housing Prices

Dataset: Boston Housing Dataset

- Features: Number of rooms, crime rate, property tax rate, etc.
- Target: Median value of owner-occupied homes ($1000s)

Objective: Predict the median housing price based on various features.

Model Training and Evaluation

Training the Model:

- Split data into training and test sets
- Fit the linear regression model using training data

Evaluation Metrics:

- Mean Squared Error (MSE): Measures the average squared difference between actual and predicted values.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3}$$

- R-squared (R^2): Proportion of variance in the dependent variable that is predictable from the independent variables.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4}$$

Residual Analysis

Residuals: The differences between observed and predicted values.

- Residual plot should show no patterns for a good fit.
- Check for homoscedasticity and normality of residuals.

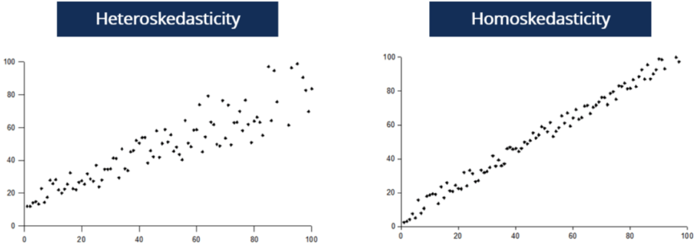


Figure: Residual plot for linear regression model

Conclusion

- Linear regression is a powerful tool for predicting continuous outcomes.
- Assumptions must be checked to ensure model validity.
- Case study demonstrated the practical application of linear regression to predict housing prices.
- Future work can include regularization techniques like Ridge and Lasso regression.