Medical Data Analysis with Python Week 14: Regression Models

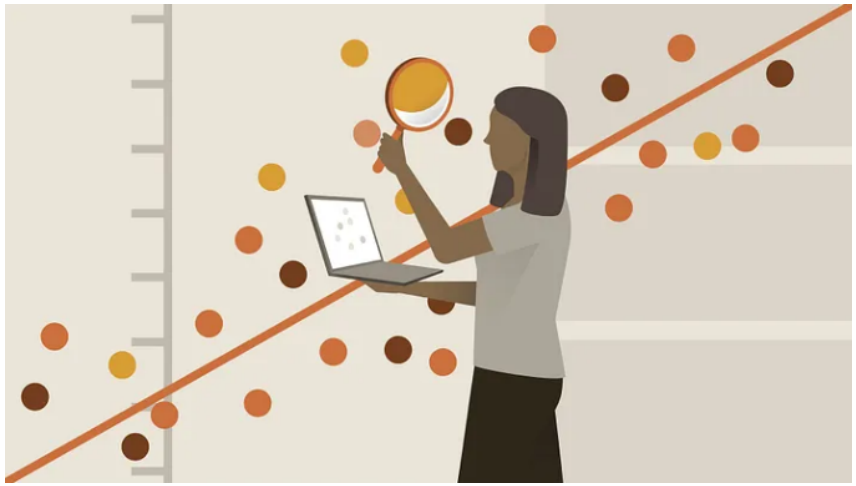
Assist. Prof. Huseyin TUNC

December 6, 2023

Content

What is Regression?

- Regression serves as a statistical method employed to establish and analyze connections between a dependent variable and one or multiple independent variables.
- Its primary objective lies in unraveling how variations in independent variables correlate with changes observed in the dependent variable.
- This method plays a fundamental role in predicting outcomes, modeling trends, and understanding the impact of diverse factors on the observed data patterns.



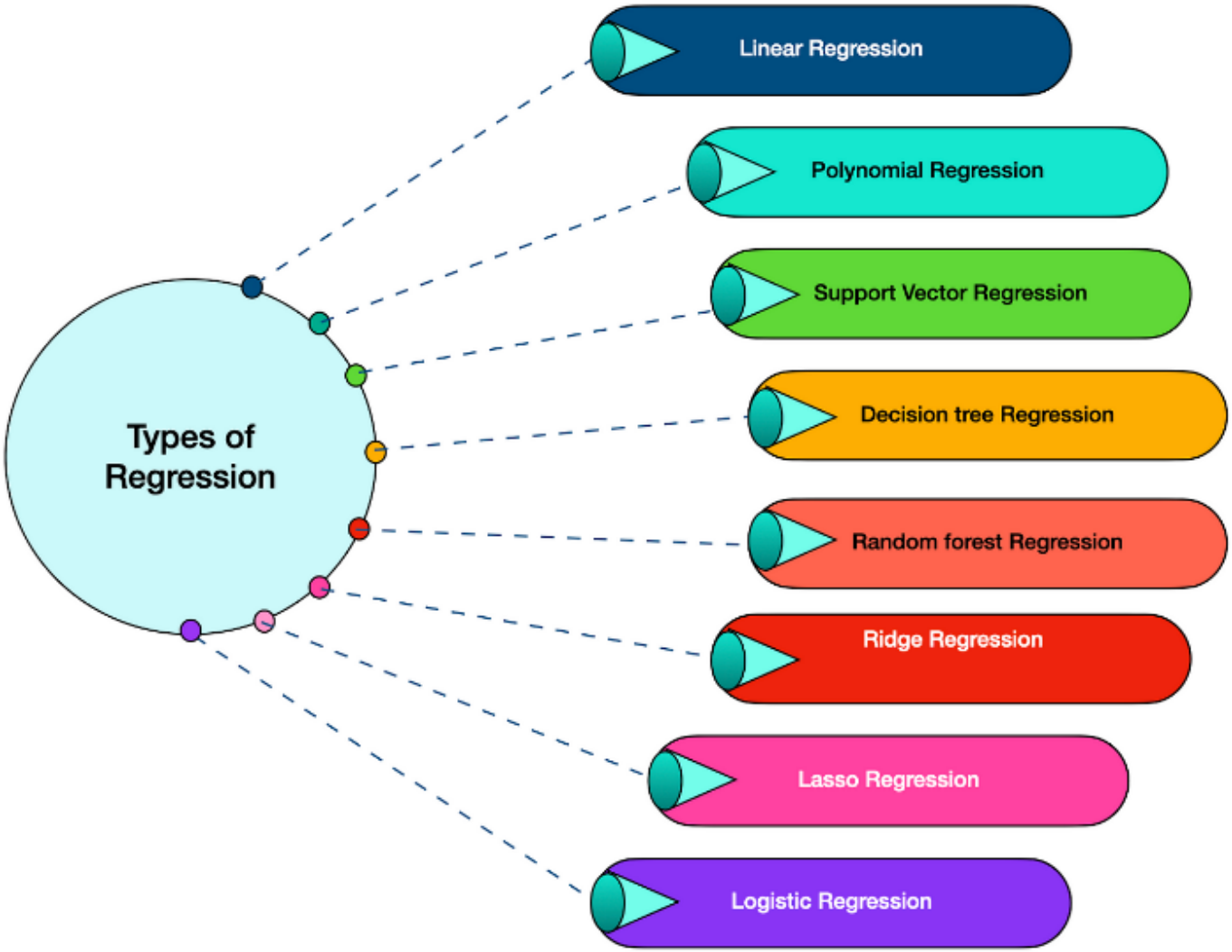
Significance in Health Data Analysis

- **Role of Regression in Health Data Analysis**
  - In the domain of health data analysis, regression assumes a critical role by unraveling the intricate relationships among diverse health-related factors and outcomes.
  - It serves as a powerful tool in predicting health-related outcomes, shedding light on complex disease patterns, and assessing the impact of interventions or risk factors.
  - By analyzing health data through regression models, researchers and healthcare professionals can gain insights into how various factors—ranging from lifestyle and environmental influences to genetic predispositions—affect health outcomes.

Applications of Regression in Healthcare

- **Predicting Disease Risk:**
  - Regression models are utilized to predict the risk of diseases based on various factors like age, genetics, lifestyle, and medical history.
- **Treatment Outcome Prediction:**
  - Regression analysis helps forecast the outcome of treatments or interventions for different diseases or medical conditions.
- **Healthcare Resource Planning:**
  - Regression techniques aid in planning healthcare resources by predicting patient admission rates, hospital bed occupancy, or demand for medical services.
- **Clinical Decision Support Systems:**
  - Regression models are integrated into clinical decision support systems to assist healthcare professionals in making informed decisions about patient care.
- **Drug Efficacy Assessment:**
  - Regression analysis is used to evaluate the effectiveness of drugs or treatments based on patient data and medical trials.

Types of Regression



Linear Regression

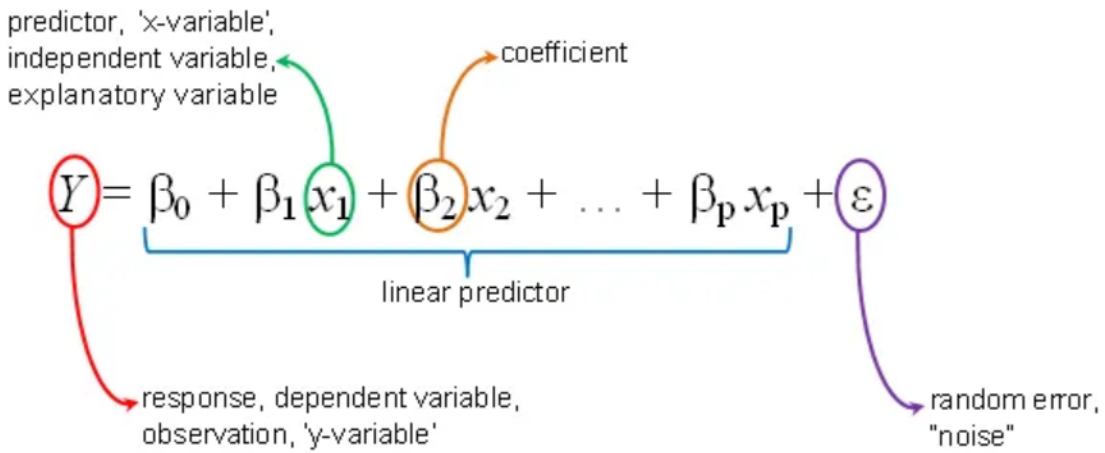
- **Linear Regression:** Linear regression is a statistical technique used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between these variables.
- **Formula:** The formula for a simple linear regression model is:

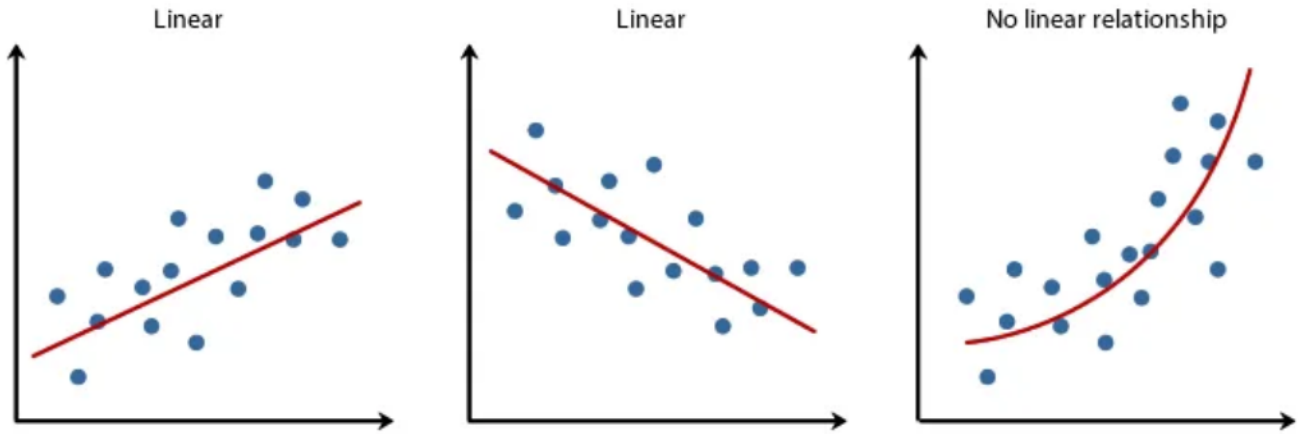
$$y = \beta_0 + \beta_1 x + \epsilon$$

where:

  - y is the dependent variable.
  - β_0 is the y-intercept (constant term).
  - β_1 is the slope coefficient for the independent variable x.
  - ϵ is the error term.

- **Application:** This method is widely used in various fields, including healthcare, for predicting outcomes, analyzing trends, and determining the strength of relationships between variables.


predictor, 'x-variable',
independent variable,
explanatory variable
coefficient
$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$
linear predictor
response, dependent variable,
observation, 'y-variable'
random error,
"noise"


Linear
Linear
No linear relationship

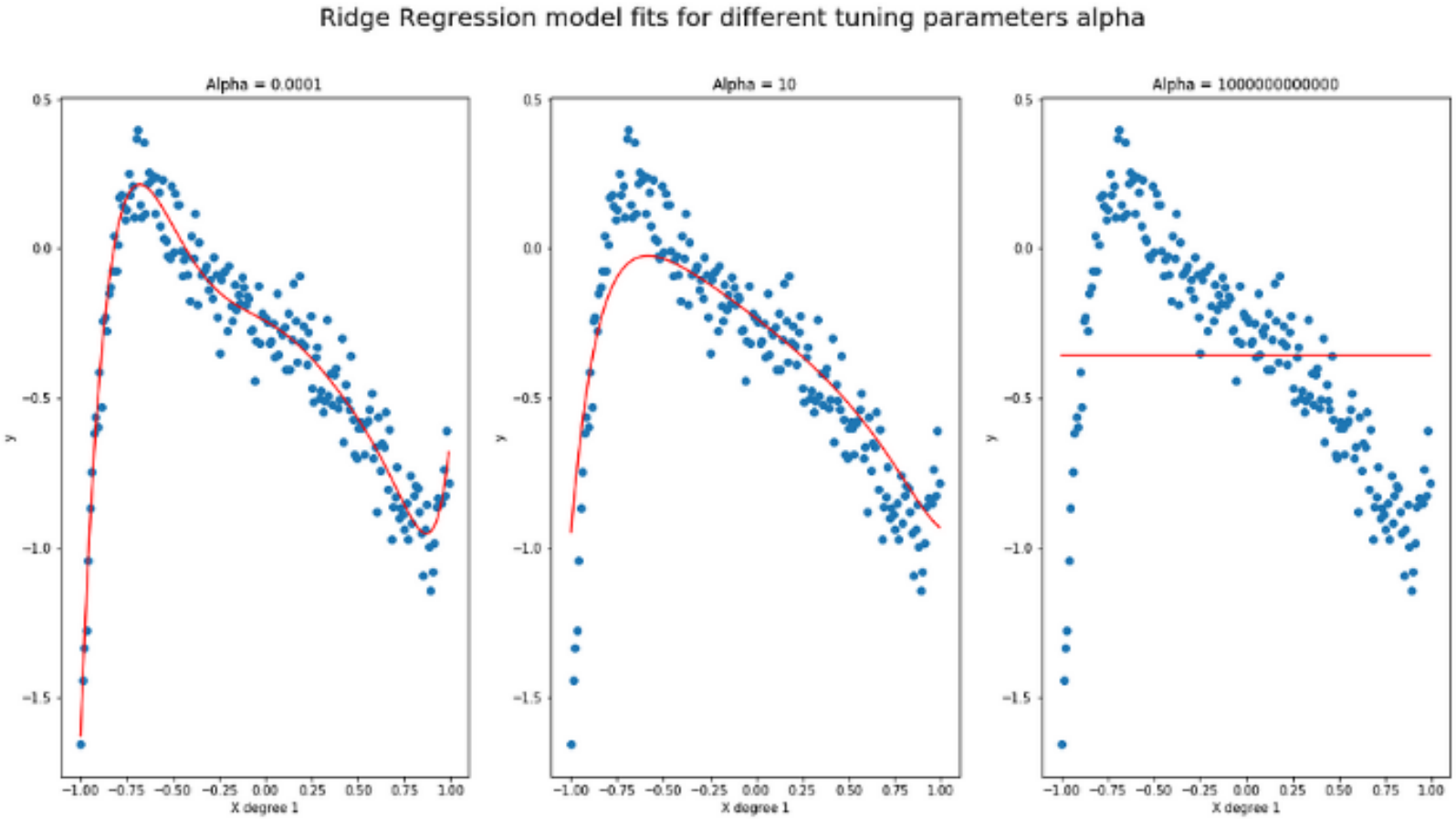
Ridge Regression

- **Ridge Regression:** Ridge regression is a regularization technique used to mitigate multicollinearity in multiple regression models.
- **Formula:** The formula for ridge regression is:

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

where $\hat{\beta}^{ridge}$ represents the ridge regression coefficients, λ is the regularization parameter, y_i are the observed values, x_{ij} are the predictor variables, and β_j are the coefficients.

Ridge Regression



Linear vs. Ridge Regression: Similarities and Differences

Similarities:

- Both models are linear, used for predicting outcomes.
- Utilize predictor variables to estimate the dependent variable.
- Aim to minimize the error between observed and predicted values.

Differences:

- **Regularization:** Ridge regression includes a regularization term (λ) to manage overfitting and multicollinearity.
- **Coefficient Estimation:** Linear regression uses least squares, while ridge regression adds a penalty to the coefficients.
- **Usage:** Ridge regression is better for multicollinearity or when controlling model complexity is needed.

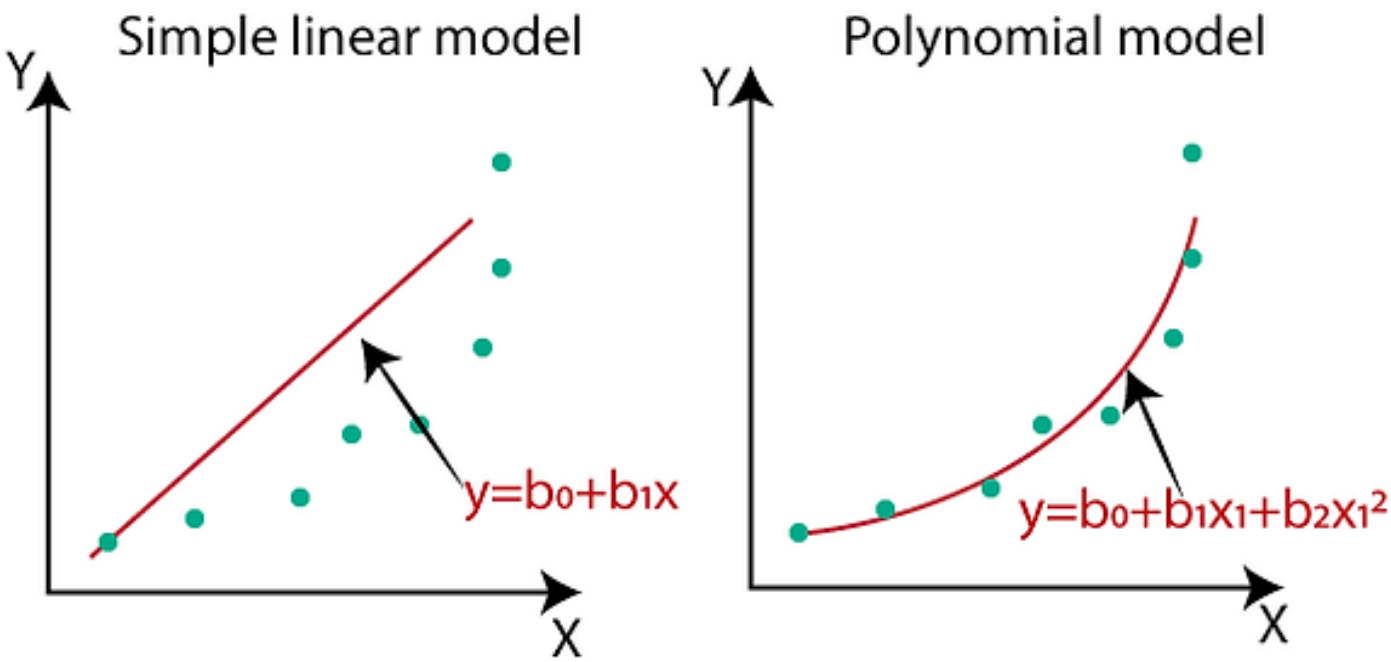
Polynomial Regression

- **Polynomial Regression:** Polynomial regression is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modeled as an n-th degree polynomial.
- **Formula:** The formula for polynomial regression with an n-th degree polynomial is:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_n x^n + \epsilon$$

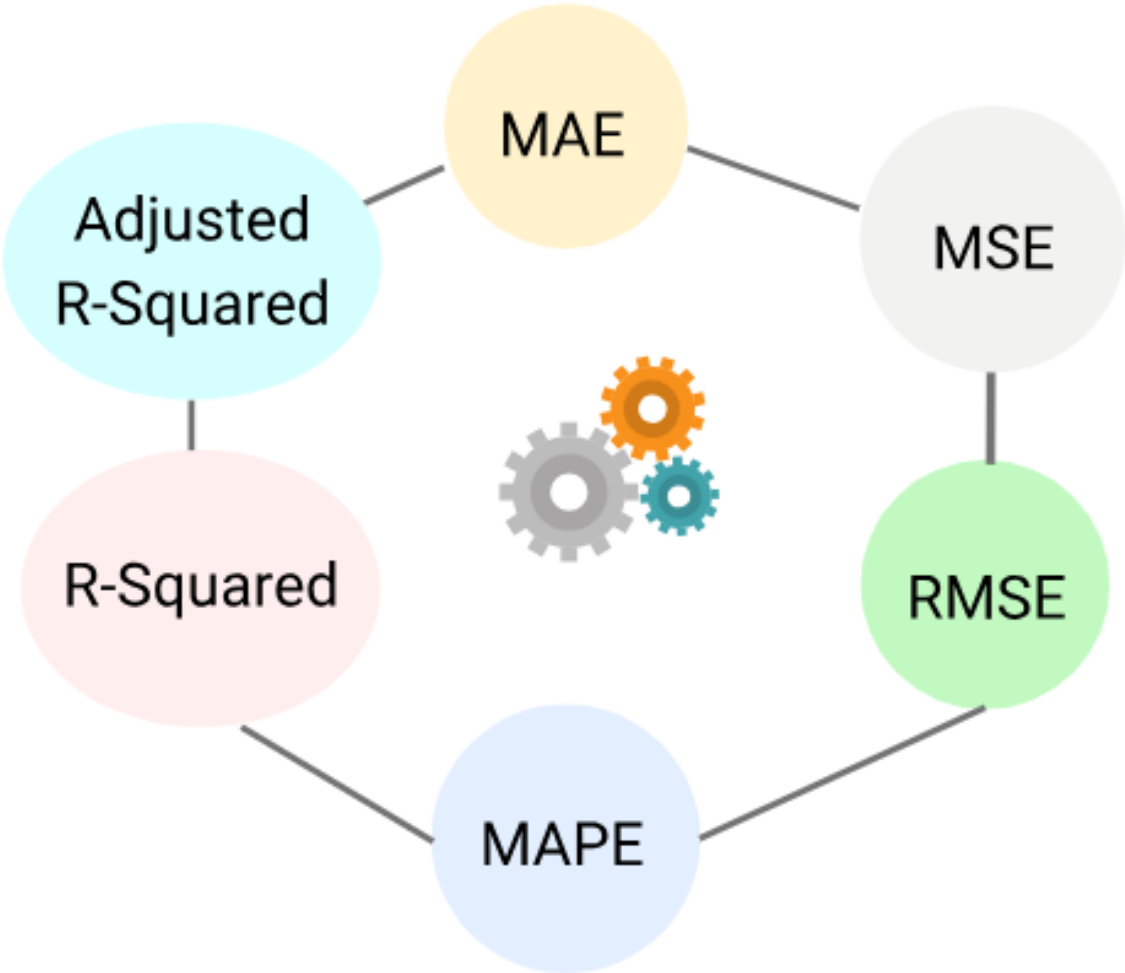
where y is the dependent variable, x is the independent variable, $\beta_0, \beta_1, \ldots, \beta_n$ are the coefficients, ϵ represents the error term, and n is the degree of the polynomial.

Polynomial Regression



Regression Metrics

- Regression metrics are criteria used to evaluate the performance of regression models. These metrics measure the differences between the model's predicted values and the actual values or how well the model fits the data.
- Mean Squared Error (MSE) measures how far the model's predictions are from the actual values.
- R-Squared (R^2) indicates how much variance in the dependent variable is explained by the independent variables.
- Additionally, metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are commonly used in regression analysis. MAE measures the average absolute differences between predicted and actual values, while RMSE quantifies the average magnitude of the error by taking the square root of the average squared differences between predicted and actual values.


MAE
Adjusted
R-Squared
MSE
R-Squared
RMSE
MAPE

Mean Squared Error (MSE)

- **MSE in Regression:**
  - Mean Squared Error (MSE) quantifies the average squared difference between the actual and predicted values in a regression problem.
  - It is calculated as:

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

where n represents the number of observations, Y_i is the actual value, and $\hat{Y}_i$ is the predicted value.

R-Squared (R^2)

- **R-Squared in Regression:**
  - R-Squared measures the proportion of the variance in the dependent variable that is predictable from the independent variables.
  - The formula for R-Squared is:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

where Y_i represents the actual value, $\hat{Y}_i$ is the predicted value, and $\bar{Y}$ is the mean of the observed data.

- R2 SCORE

BAD MODEL

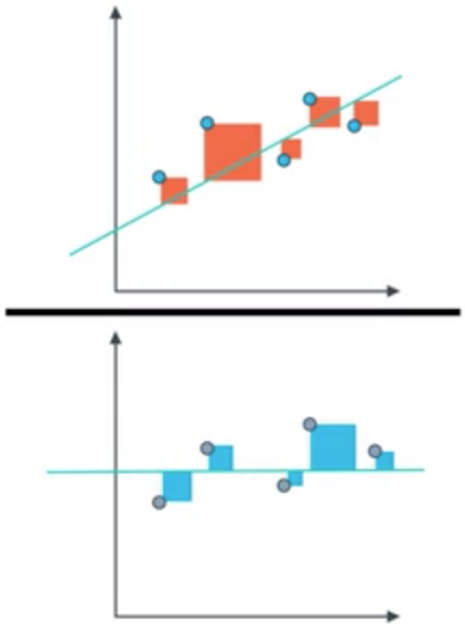
The errors should be similar.
R2 score should be close to 0.

GOOD MODEL

The mean squared error for the linear regression model should be a lot smaller than the mean squared error for the simple model.

R2 score should be close to 1.

$$R2 = 1 - \frac{\text{}}{\text{}}$$



Conclusion

- Regression analysis is a vital statistical tool in understanding relationships between variables, making predictions, and assessing the impact of factors on observed patterns.
- In the realm of healthcare, regression techniques play a crucial role in predicting disease risks, evaluating treatment outcomes, and planning healthcare resources efficiently.
- Different types of regression, such as ridge regression and polynomial regression, offer diverse approaches to modeling complex relationships in data.
- Evaluation metrics like Mean Squared Error (MSE) and R-Squared (R^2) provide quantitative measures to assess the performance of regression models.