Multiple Linear Regression

Team Members:
Amey Mahendra Thakur,
Jithin Gijo Varghese,
Ritika Agarwal

Instructor: Dr. Yasser Alginahi

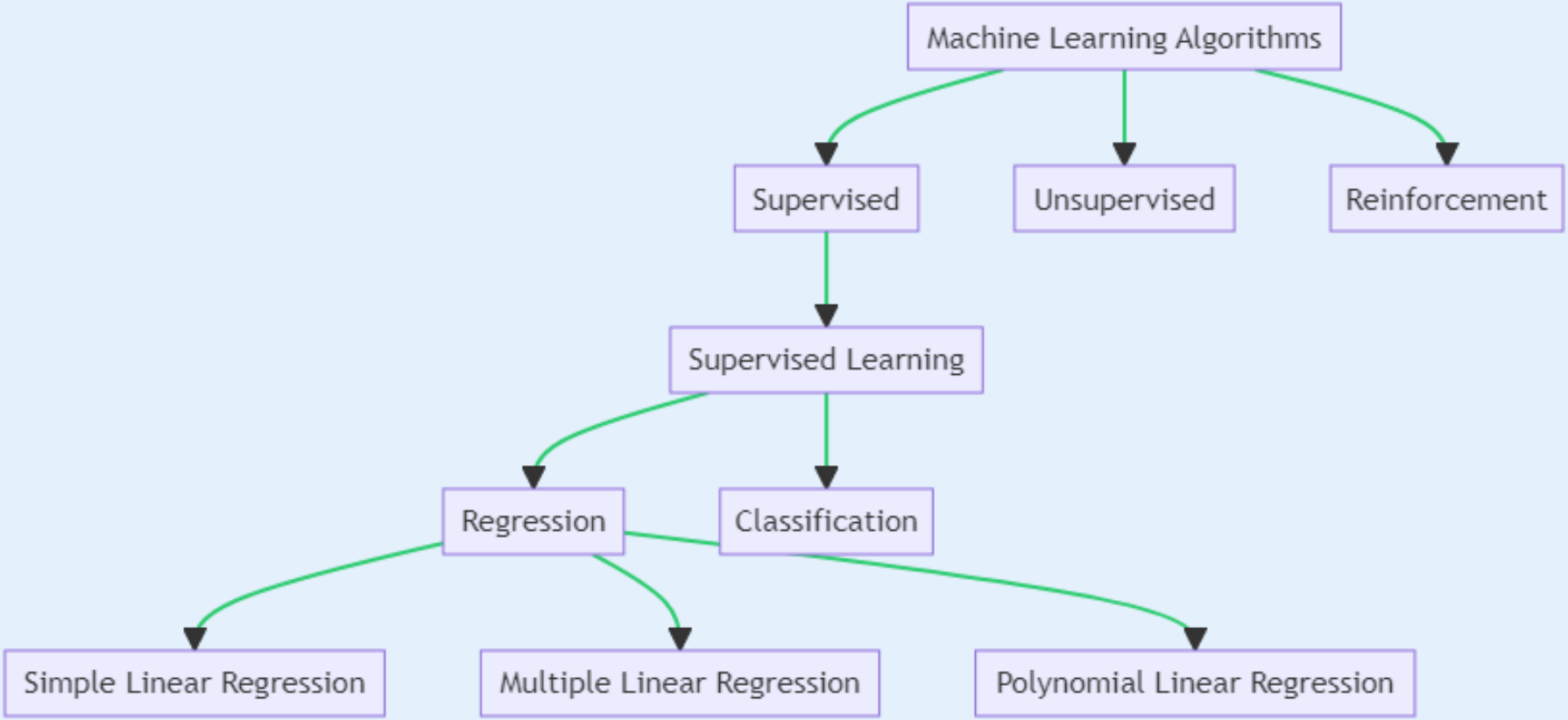
Date: 29th September 2023

University of Windsor

AGENDA

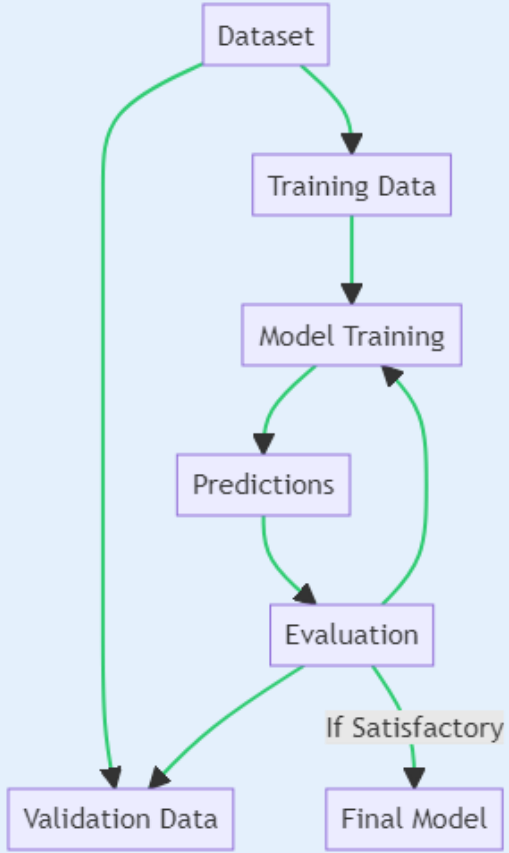
- Hierarchy of ML Algorithms
- Simple Linear Regression
- Linear vs. Multiple Regression
- Multiple Linear Regression
- Mathematical Concept
- Benefits and Limitations
- Real-life Applications
- Summary

Hierarchy of Machine Learning Algorithms



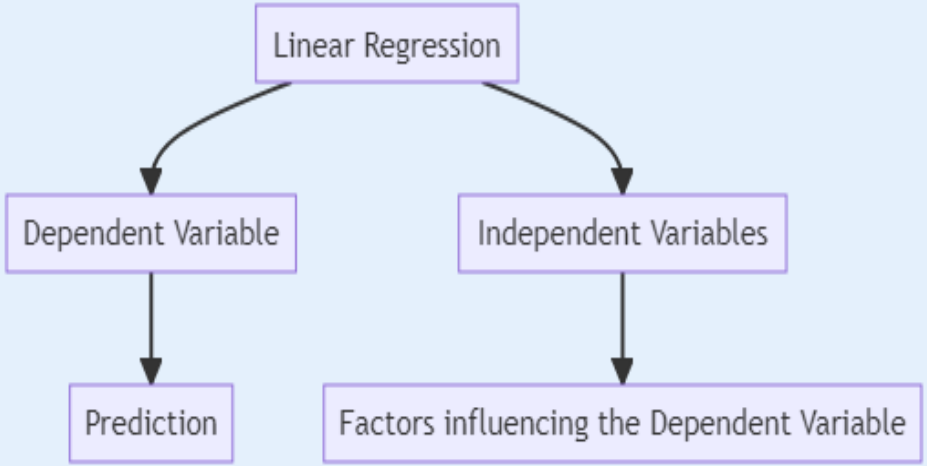
Supervised Learning Workflow

- Dataset is split into Training Data and Validation Data.
- Training Data is used for Model Training.
- The trained model makes Predictions.
- Predictions are Evaluated using the Validation Data.
- If the evaluation is satisfactory, we get the Final Model, else the model is retrained.



Simple Linear Regression

Linear regression is a statistical technique that finds the best-fitting straight line to show how a dependent variable (the one we want to predict, often denoted as "Y") is related to one independent variables (the factor or input we use to make predictions, often denoted as "X").

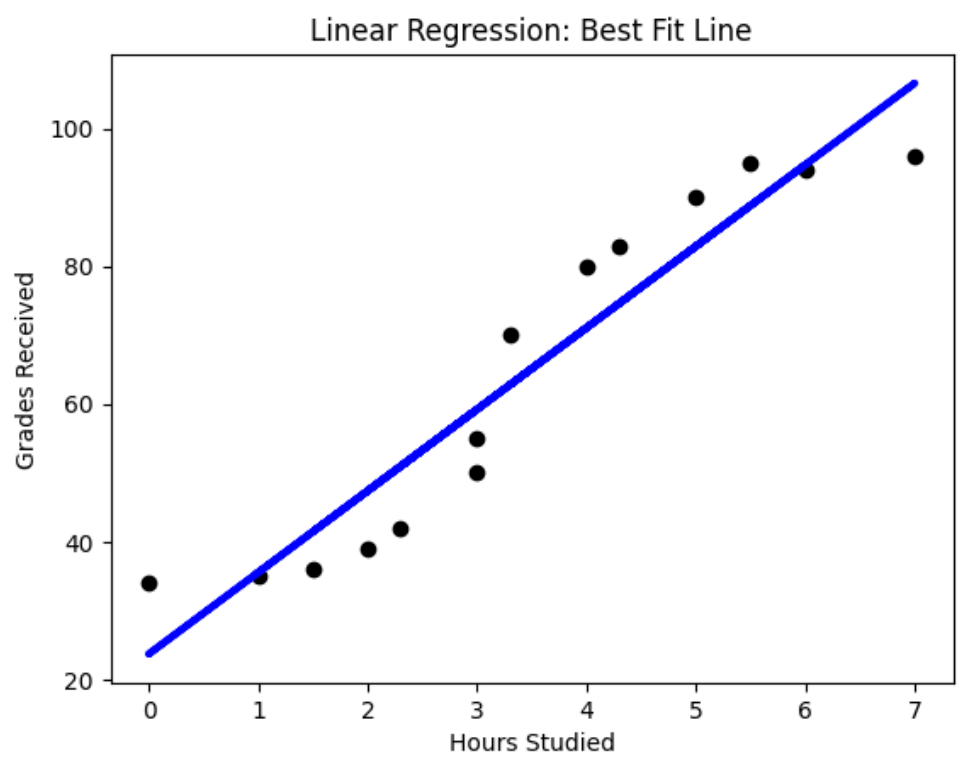


Relationship between Hours Studied and Grades Received

Hours Studied	Grades Received
1.0	35
3.0	55
2.3	42
6.0	94
1.5	36
7.0	96
5.0	90

Hours Studied	Grades Received
3.3	70
4.0	80
2.0	39
3.0	50
0,0	34
5.5	95
4.3	83

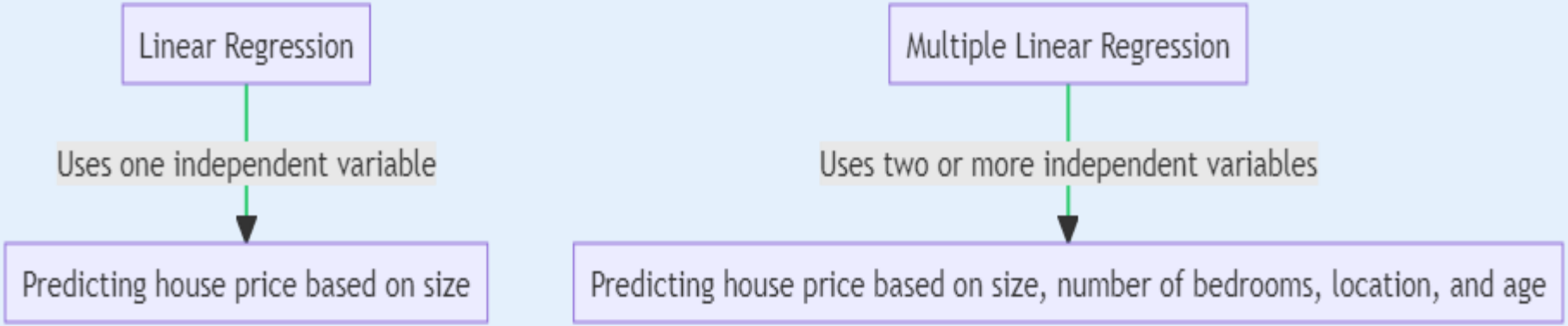
Correlation between Hours Studied and Grades Received



https://colab.research.google.com/drive/1m1JjcqGr_TyyqG2y29Ya9GY5epE8kuoU?usp=sharing

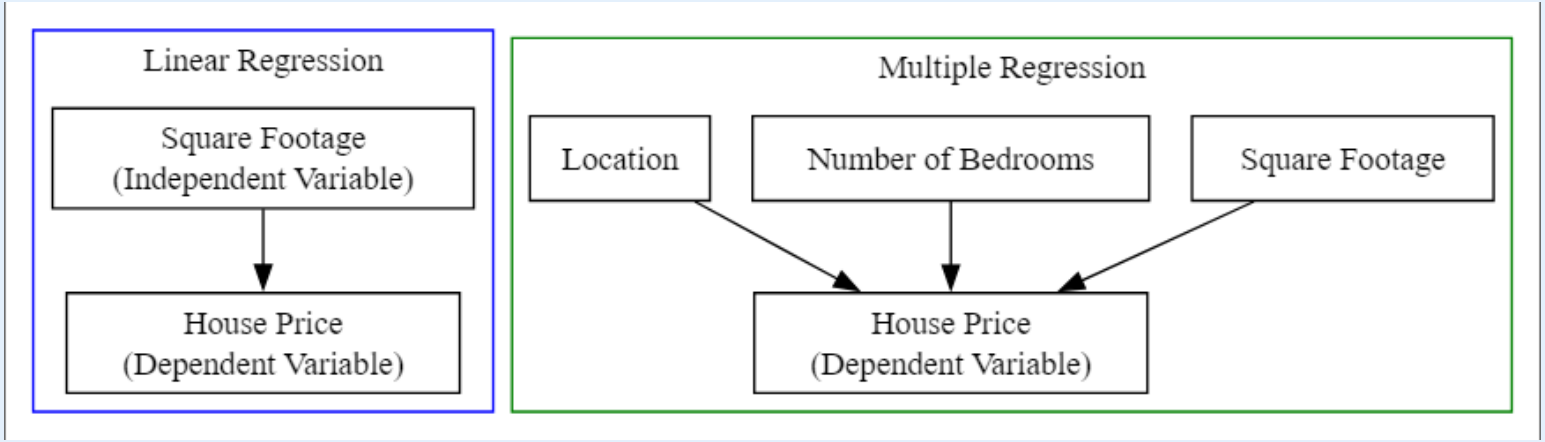
Linear vs. Multiple Linear Regression

Linear Regression and Multiple Linear Regression are both statistical techniques used to predict the value of a dependent variable based on the values of independent variable(s), but they differ in the number of independent variables used.



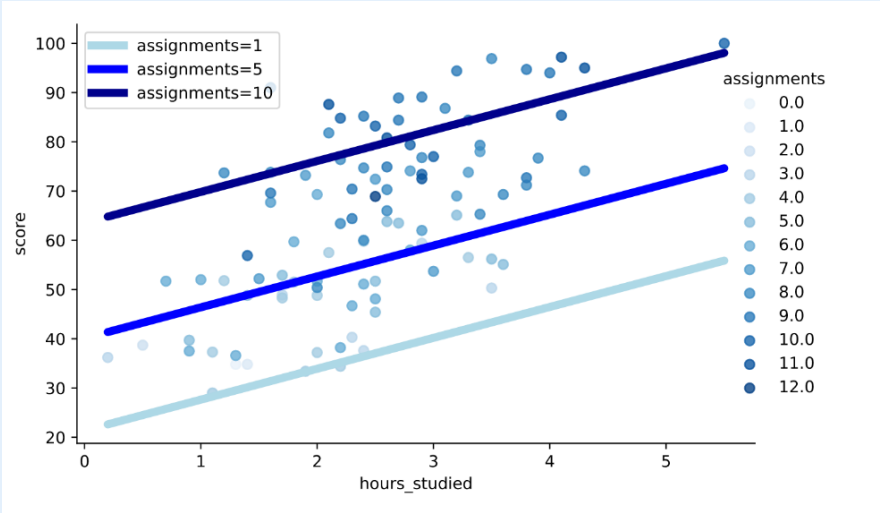
Linear vs. Multiple Linear Regression

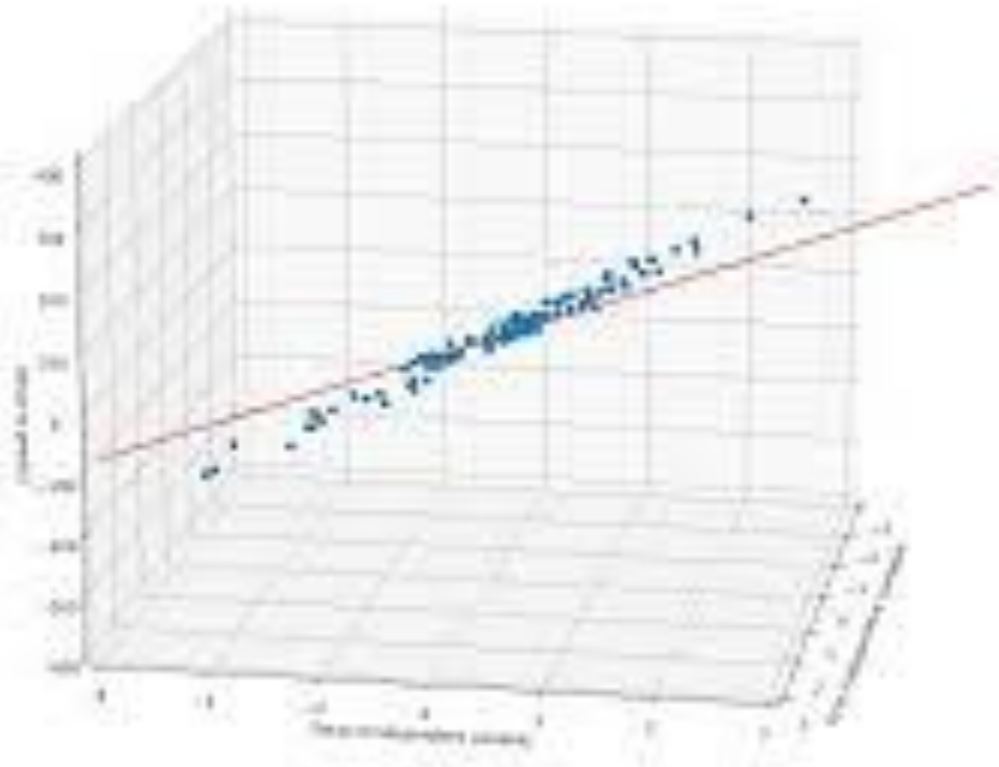
- In Linear Regression, the house price (dependent variable) is predicted based on one independent variable, which is the square footage of the house.
- In Multiple Regression, the house price (dependent variable) is predicted based on multiple independent variables, such as square footage, number of bedrooms, and location.



Multiple Linear Regression

Multiple linear regression is a statistical technique that examines the relationship between a dependent variable (the outcome we want to predict or explain) and two or more independent variables (factors that may influence the dependent variable), modeling how they collectively influence the dependent variable through a linear equation.





Multiple Linear Regression

In multiple linear regression, we model the relationship between multiple independent variables (features) and a dependent variable.

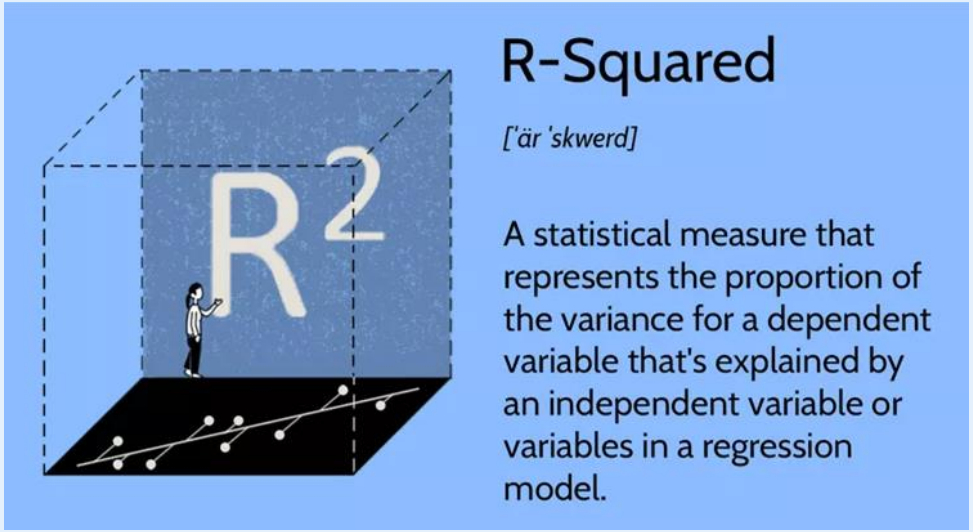
The equation for multiple linear regression is:

$$y=\beta0+\beta1x1+\beta2x2+\ldots+\beta nxn+\epsilon$$

- y is the dependent variable (what we are trying to predict).
- $\beta0$ is the y-intercept.
- $\beta1,\beta2,\ldots,\beta n$ are the coefficients of the independent variables.
- $x1,x2,\ldots,xn$ are the independent variables (features).
- ϵ is the error term.

Understanding R-squared (R^2)

- R-squared (R^2) measures how well a regression model fits data.
- It ranges from 0 to 1, with 1 meaning a perfect fit.
- Higher R^2 values indicate better model fit.
- R^2 doesn't explain variable significance.

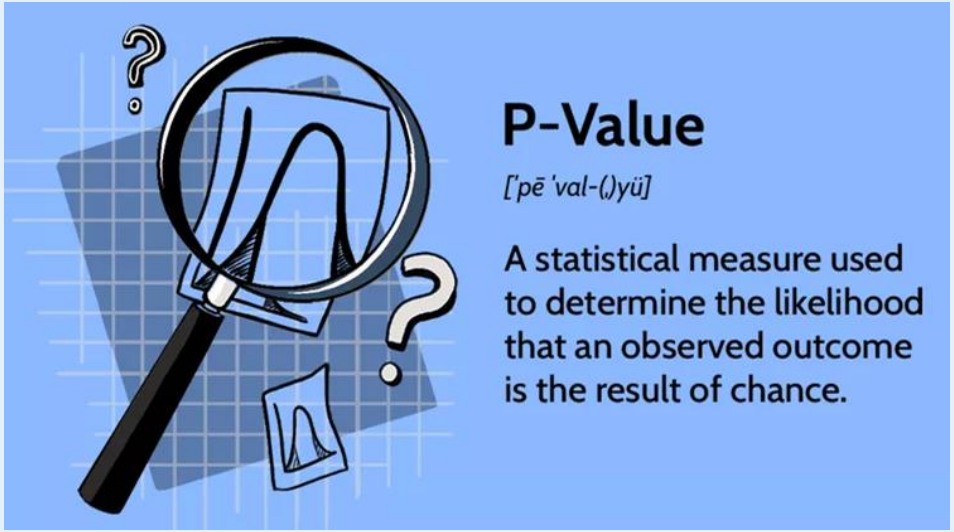


Understanding R-squared (R^2)

- Formula:
- R2 = Total Variance/Explained Variance
- =1− Total Variance/Residual Variance

Understanding p-values

- P-values test variable significance in a regression model.
- Low p-values (<0.05) mean significance.
- High p-values suggest insignificance.
- Use p-values to decide variable inclusion.

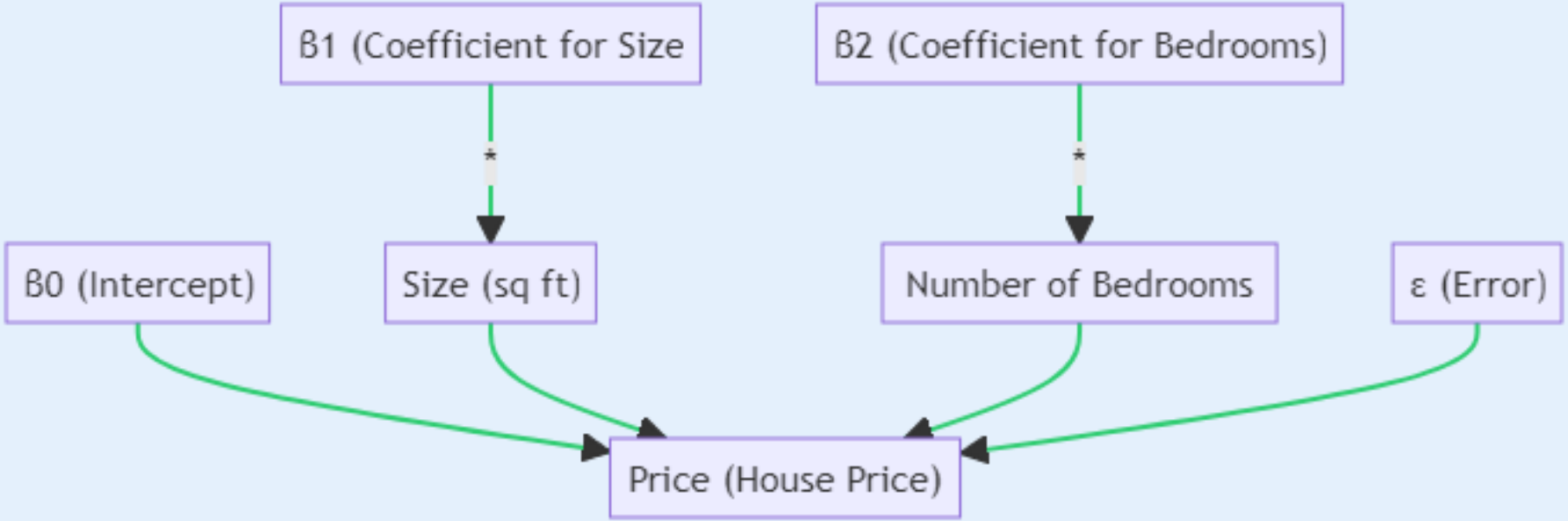


Purposes of R-squared and p-values

- **R-squared** measures overall model goodness of fit.
  - Evaluates how well the model explains dependent variable variation.
  - Higher R^2 indicates better overall fit (0 to 1).

- **p-values** assess individual variable significance.
  - Determines if each variable contributes significantly to the model.
  - Low p-values (<0.05) imply significance.

House Price Prediction with Multiple Regression



$$Price=\beta0+\beta1(Size)+\beta2(Bedrooms)+\epsilon$$

Assumptions of Multiple Linear Regression

- A linear relationship between the dependent and independent variables.
- The independent variables are not highly correlated with each other.
- The variance of the residuals is constant (i.e., differences between predicted and actual values).
- Each data point shouldn't depend on others; they should be separate.
- All variables should be normally distributed.

Benefits of Multiple Linear Regression

- **Predictive Power**
  - Predict dependent variable based on multiple independent variables.
- **Quantifying Relationships**
  - Measures strength and direction of relationships.
- **Control for Confounding Factors**
  - Accounts for other influencing variables.
- **Model Interpretability**
  - Coefficients offer insights into variable impact.
- **Assumption Testing**
  - Provides diagnostic tools for model quality.

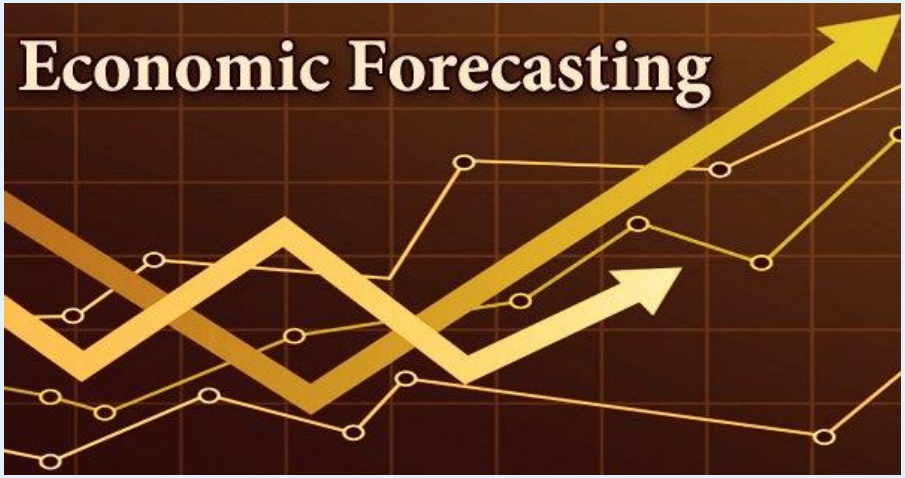
Limitations of Multiple Linear Regression

- **Linearity Assumption**
  - Assumes linear relationships.
- **Multicollinearity**
  - High correlations among variables can lead to instability.
- **Overfitting**
  - Too many variables can cause poor generalization.
- **Assumption Violations**
  - Relies on normality, homoscedasticity, and independence.
- **Handling Categorical Variables**
  - Requires encoding, adding complexity.
- **Data Requirements**
  - Needs a large dataset for reliable results.

Real-Life Applications of Multiple Regression

- Economics and Finance:

Stock Price Prediction



Economic Forecasting

Real-Life Applications of Multiple Regression

- Healthcare:

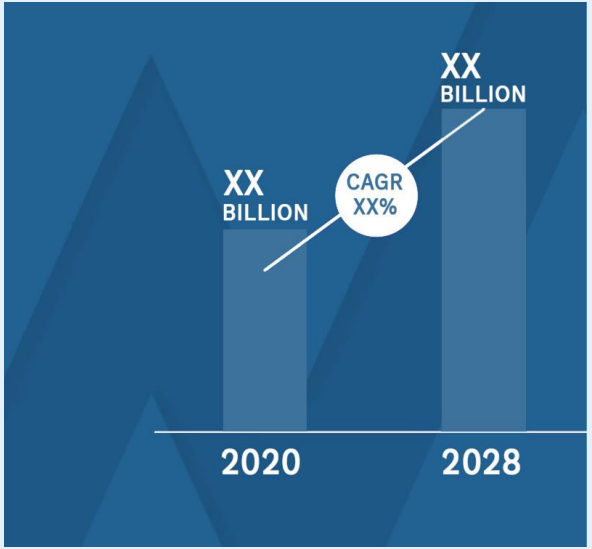
Medical Research

Hospital Readmission Prediction

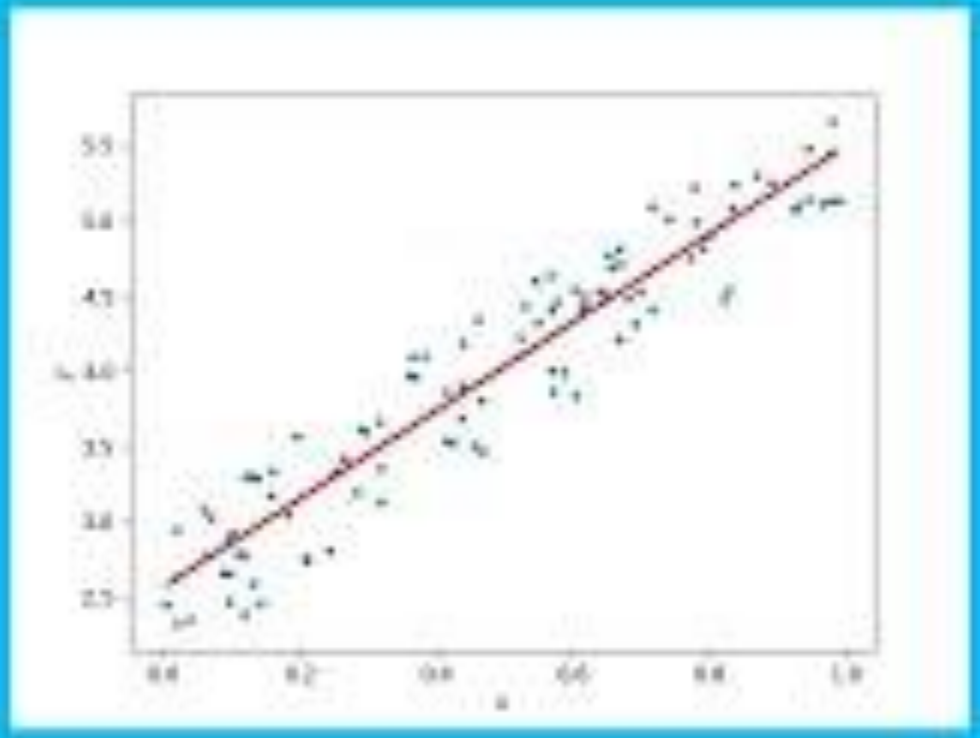
Real-Life Applications of Multiple Regression

- Marketing:

Sales Forecasting



Market Research



Multiple Linear Regression in Python sklearn

Summary

- Hierarchy: Explored ML algorithm hierarchy.
- Simple Linear Regression: Basics of modeling relationships.
- Linear vs. Multiple Regression: Single vs. multiple predictor variables.
- Multiple Linear Regression: Complex modeling with multiple predictors.
- Mathematical Concept: Key equations behind regression.
- Benefits and Limitations: Pros and cons of multiple regression.
- Real-life Applications: Practical use cases across domains.

References

[1] M. Batta, "Machine learning algorithms - a review," https://www.researchgate.net/profile/Batta-Mahesh/publication/344717762_Machine_Learning_Algorithms_-A_Review/links/5f8b2365299bf1b53e2d243a/Machine-Learning-Algorithms-A-Review.pdf. [Accessed: Sep. 25, 2023]. DOI: 10.21275/ART20203995.

[2] S. I. Bangdiwala, "Regression: Simple linear," *International Journal of Injury Control and Safety Promotion*, vol. 25, no. 1, pp. 113–115, 2018, doi: 10.1080/17457300.2018.1426702.

[3] M. Tranmer, J. Murphy, M. Elliot, and M. Pampaka, "Multiple Linear Regression," 2nd ed., 2020. https://hummedia.manchester.ac.uk/institutes/cmist/archive-publications/working-papers/2020/multiple-linear-regression.pdf. [Accessed: Sep. 25, 2023].

[4] R. Goldstein, "Regression methods in biostatistics: Linear, logistic, survival and repeated measures models," *Technometrics*, vol. 48, no. 1, pp. 149–150, 2006, doi: 10.1198/tech.2006.s357.

[5] M. N. Williams, C. A. G. Grajales, and D. Kurkiewicz, "Assumptions of Multiple Regression: Correcting Two Misconceptions," *Practical Assessment, Research, and Evaluation*, vol. 18, Nov. 2019. doi: https://doi.org/10.7275/55hn-wk47. [Accessed: Sep. 25, 2023].

References

[6] A. E. Maxwell, "Limitations on the use of the multiple linear regression model," *British Journal of Mathematical and Statistical Psychology*, vol. 28, no. 1, pp. 51–62, 1975, doi: 10.1111/j.2044-8317.1975.tb00547.x. [Accessed: Sep. 25, 2023].

[7] J. Fernando, "R-squared: Definition, calculation formula, uses, and limitations," Investopedia, https://www.investopedia.com/terms/r/r-squared.asp. [Accessed: Sep. 25, 2023].

[8] B. Beers, "P-value: What it is, how to calculate it, and why it matters," Investopedia, https://www.investopedia.com/terms/p/p-value.asp. [Accessed: Sep. 25, 2023].

[9] "Multiple Linear Regression in Python - Sklearn," 2022, https://youtu.be/wH_ezgftiy0 [Accessed: Sep. 25, 2023].

[10] M. Ralston, “Multiple regression,” SAGE Publications Inc, https://us.sagepub.com/en-us/nam/multiple-regression/book262446 (accessed Sep. 25, 2023).

Thank You