ELEC-8900-57 SPECIAL TOPICS: MACHINE LEARNING

IN-CLASS PRESENTATION ON MULTIPLE LINEAR REGRESSION ALGORITHM

Faculty of Engineering

Prepared by:

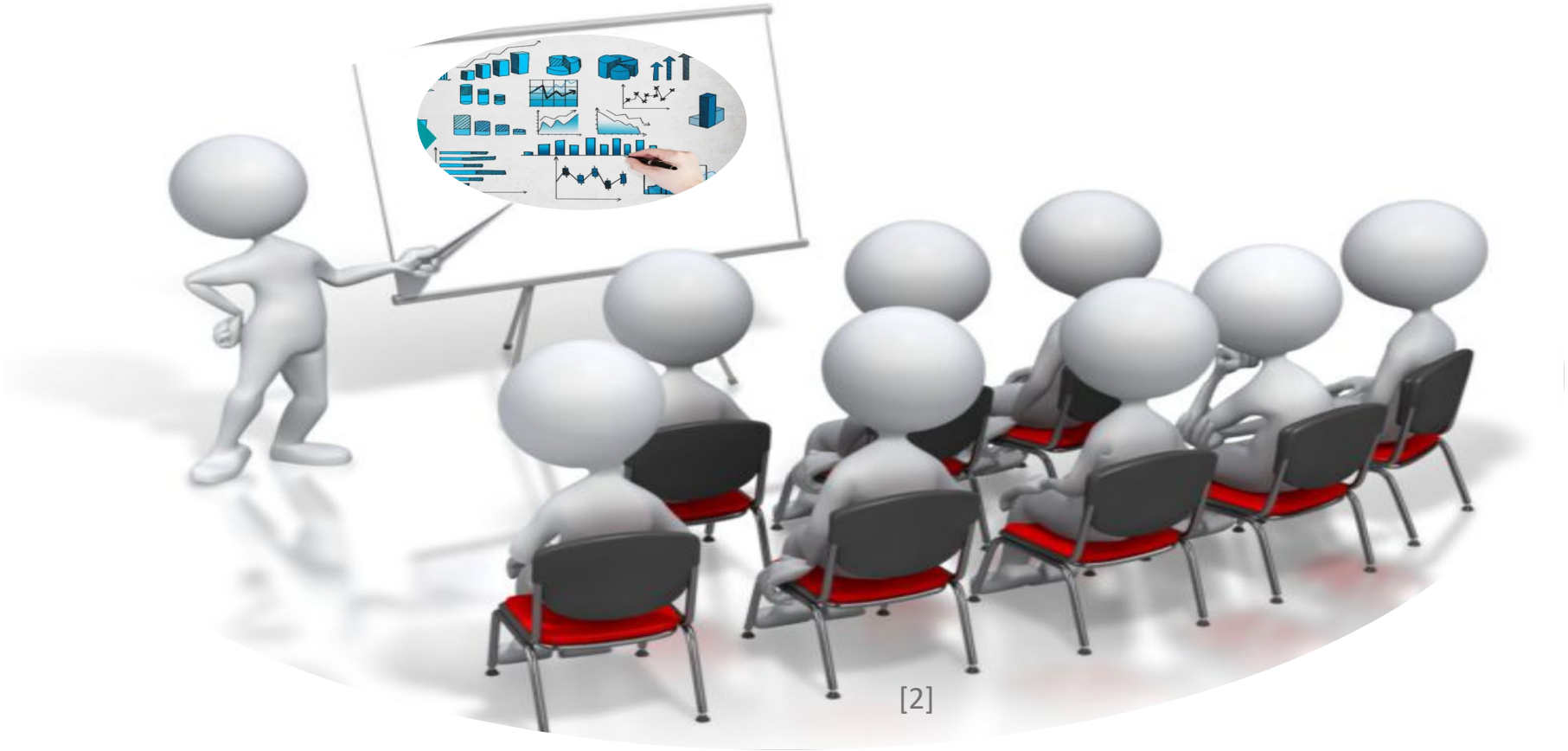
Altamash Yar Khan

Prince Dwumah-Boadi

Peter Imasuen

Submitted to: Dr. Yasser M. Alginahi

Date of Submission: October 6, 2023



[2]

ALTAMASH YAR KHAN

Linear Regression Concept in Machine Learning

- A form of supervised learning with both dependent (output) and independent (input) variables clearly defined

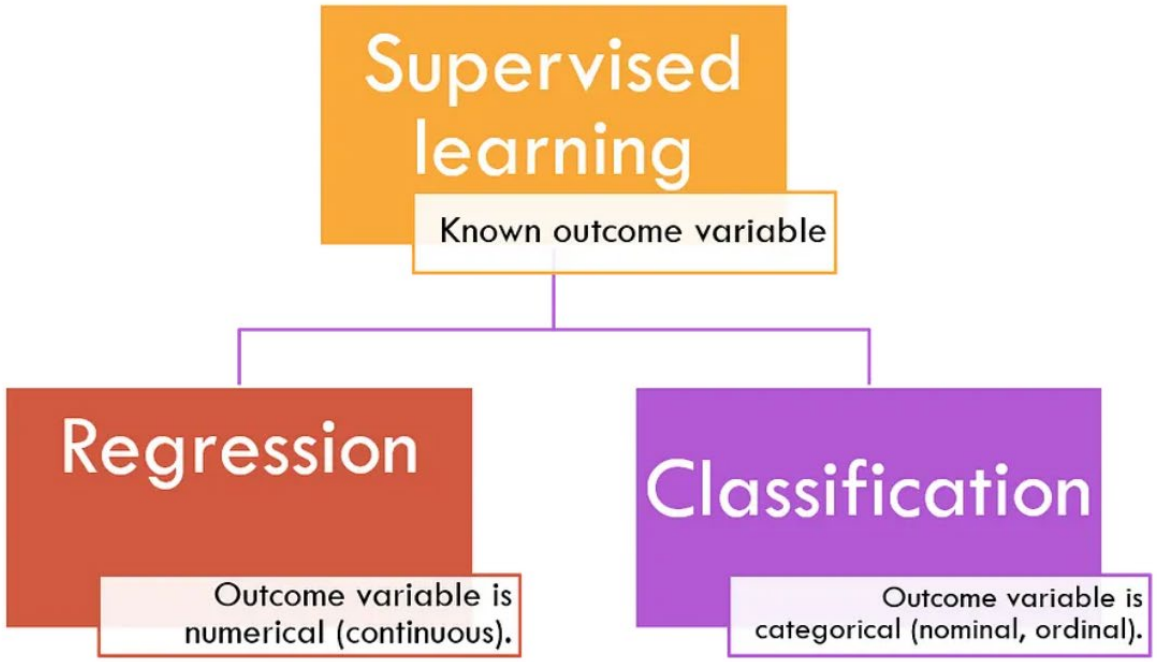


Figure 1. Types of supervised learning models [3]

Types of Regression

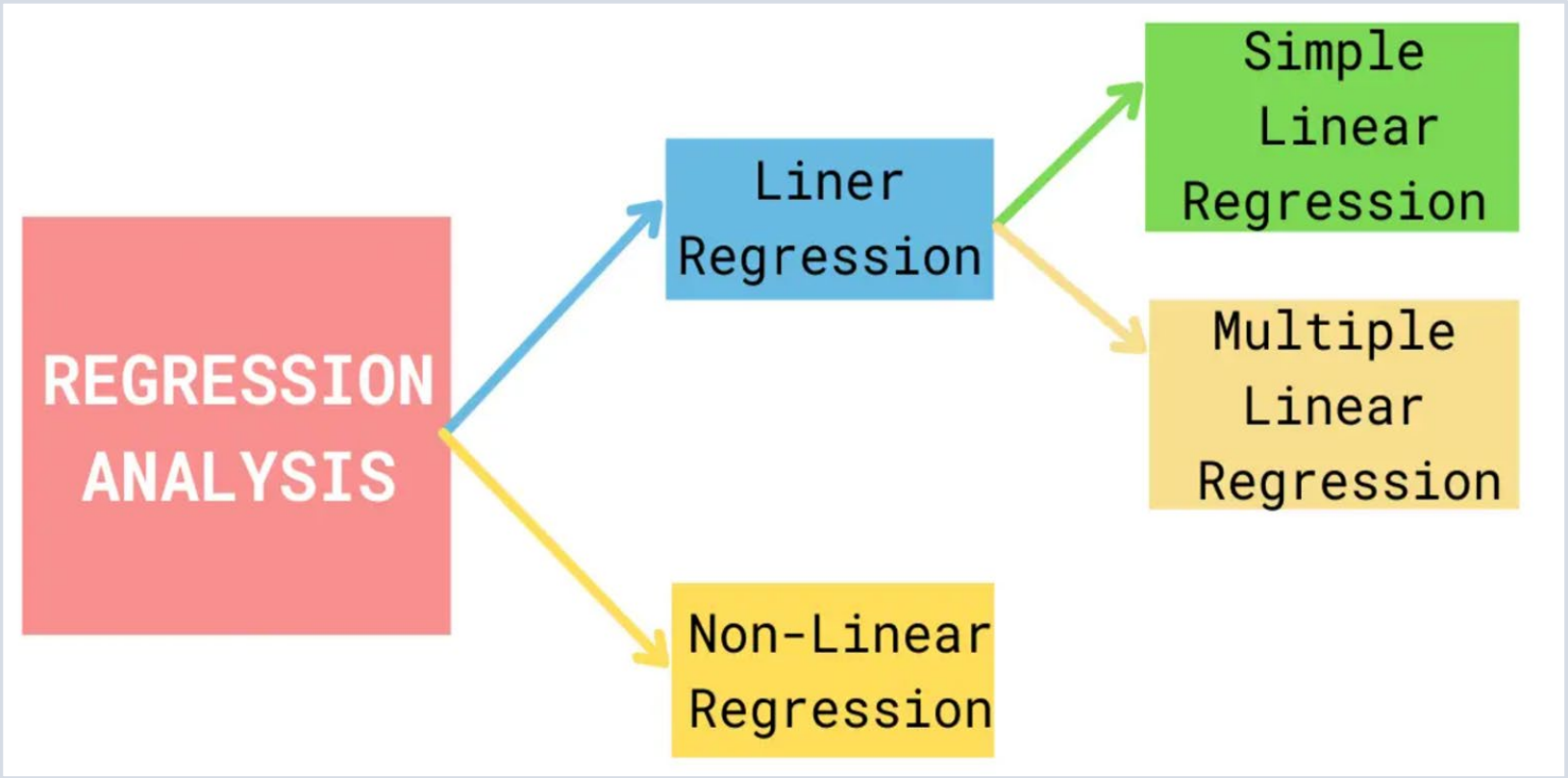


Figure 2. Types of regression [4]

Simple Linear Regression

- Has one dependent and one independent variable
- What can you tell about the model by observing the graph?
- In general, the closer the data points are to the regression line, the more accurate the final prediction.
- If there is a high degree of deviation between the data points and the regression line, the slope will provide less accurate predictions

- The model to be estimated from sample data is:

$$Y_i = b_0 + b_1 X_i + e_i$$

Residual (random error from the sample)

- The actual estimated from the sample

Estimated (or predicted) Y value for observation i

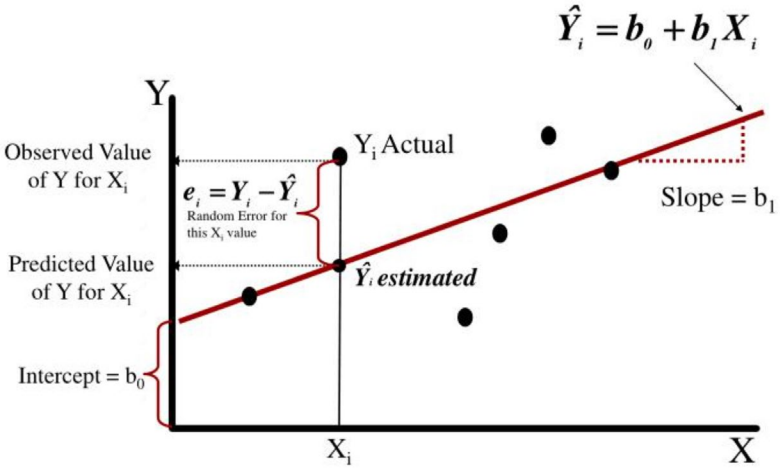
Estimate of the regression intercept

Estimate of the regression slope

Value of X for observation i

$$\hat{Y}_i = b_0 + b_1 X_i$$

– Where $e_i = Y_i - \hat{Y}_i$



Figures 3 & 4. Simple linear regression [5]

5

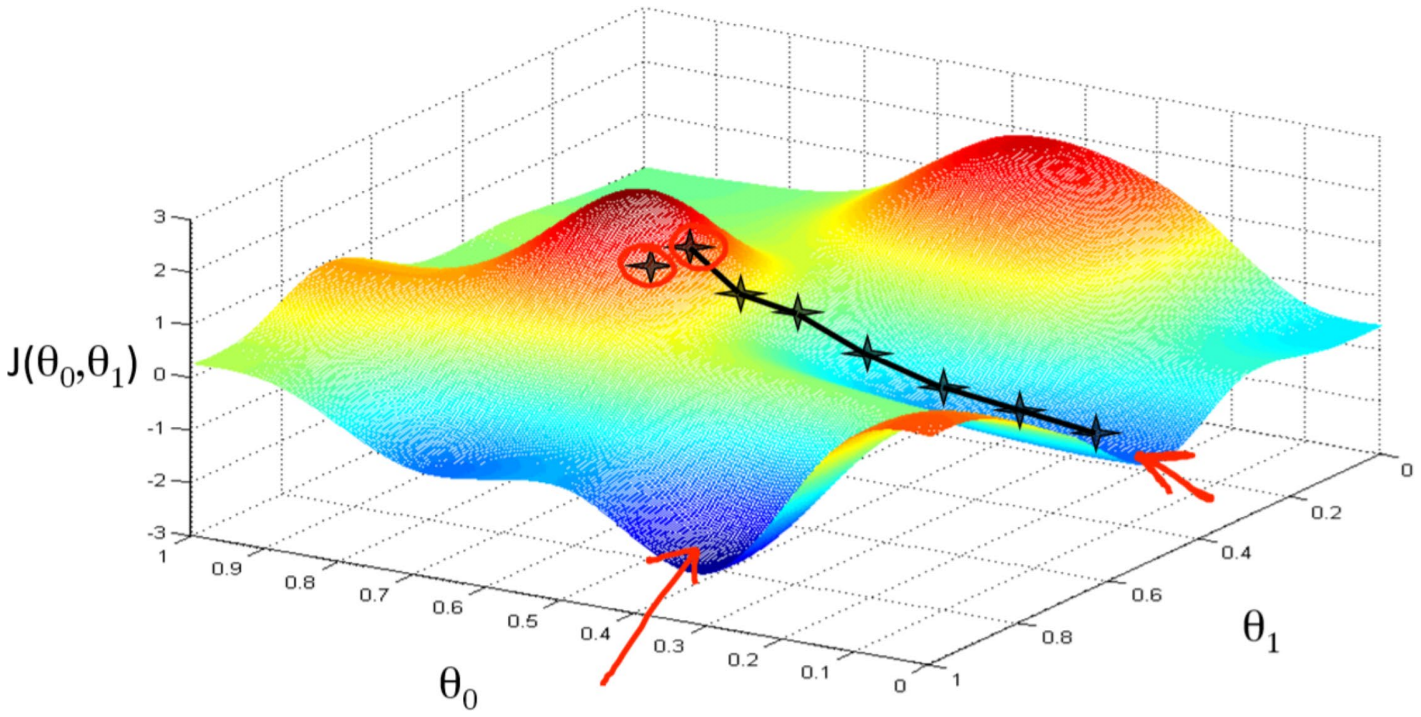
Why would one use a multiple linear regression over a simple linear regression?

- A dependent variable is rarely explained by only one variable in real life.
- **Example - Predicting House Prices involves many variables:**
  - Size: The size of the house in square feet.
  - Number of rooms: The number of bedrooms in the house.
  - Crime rate: The crime rate in the neighbourhood.
  - Proximity to Schools: The distance to the nearest schools.
  - Proximity to Transportation: The distance to public transportation.

Unlocking the Concept of Multiple Linear Regression Using the Housing Market Example: A Visual Journey [14]

7

Mathematical intuition of multiple linear regression

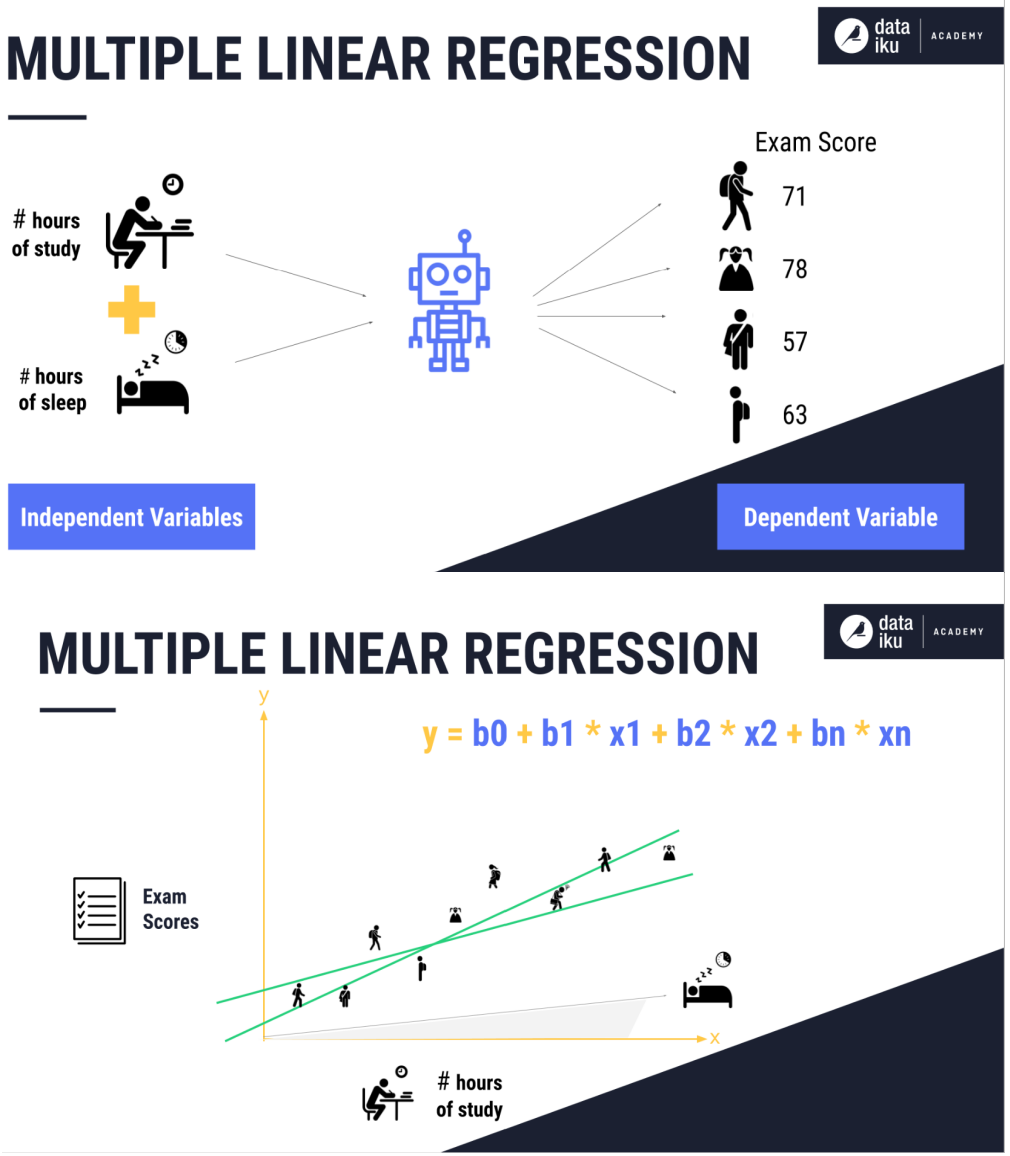


[7]

PETER IMASUEN

Multiple Linear Regression

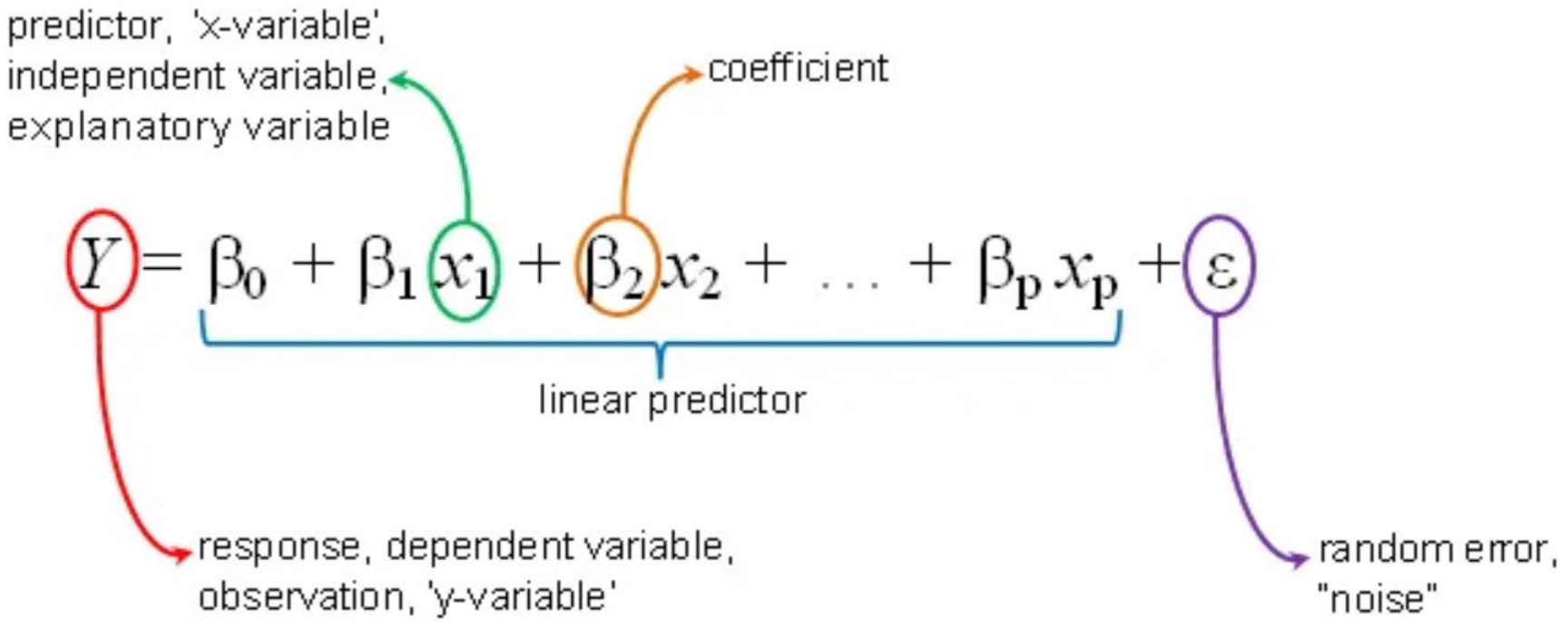
- Multiple regression is a statistical technique that aims to predict a variable of interest from several other variables.
- One dependent, Multiple Independent Variables



Figures 5 & 6. Multiple linear regression [6]

University of Windsor

General Equation for Multiple Linear Regression



Multiple linear regression formula with p independent variables [7]

Ordinary Least Squares

- Preferred technique for estimating the values of the coefficients for more than one input variable.
- Seeks to minimise the sum of the squared residuals.

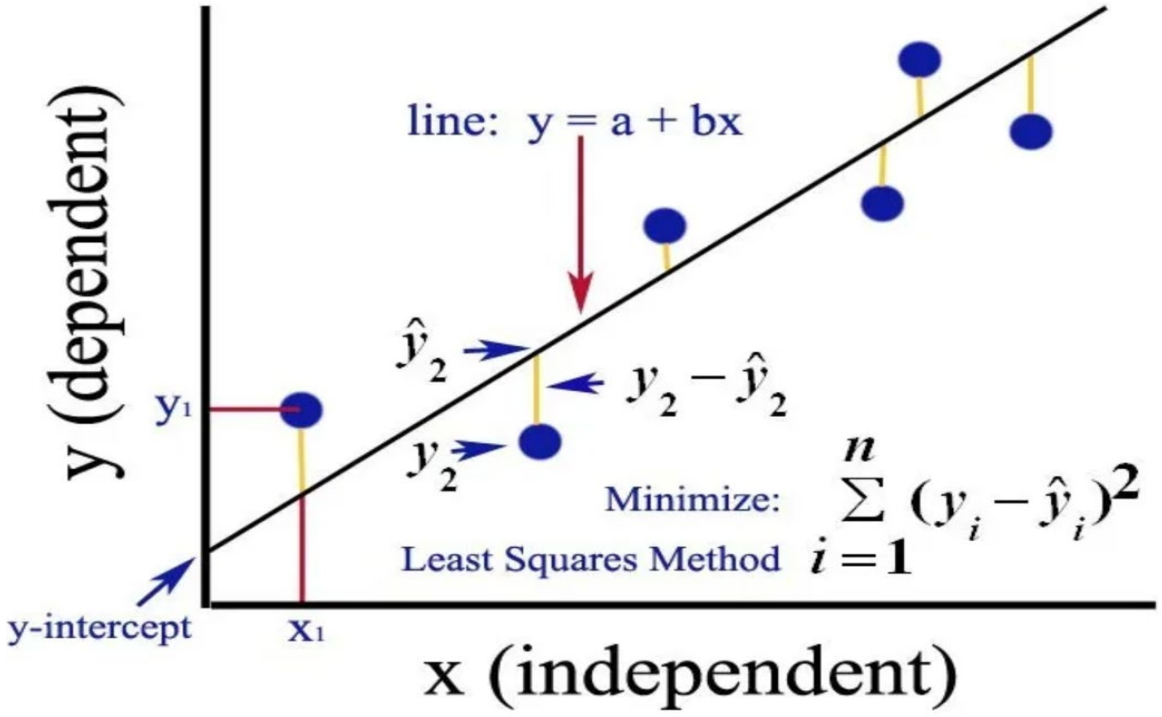
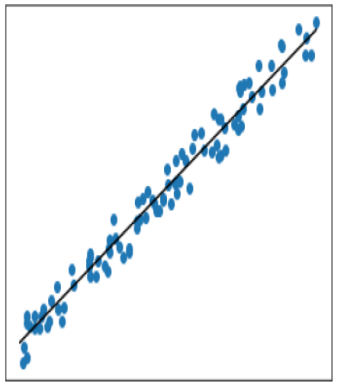


Figure 7. Ordinary Least Squares [7]

Model Performance Evaluation Metrics

R-squared

- R^2: quantifies the variance in target values explained by the features
  - Values range from 0 to 1
- High R^2:



- Low R^2:

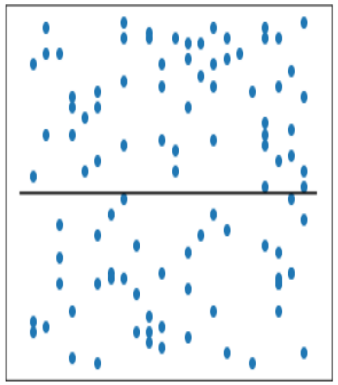


Figure 8. A graph to explain the concept of an R-squared [8]

Sum Squared Regression Error

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}}$$

Sum Squared Total Error

Sum Over All The Data Points — Square The Result

$$SS_{Regression} = \sum (y_i - y_{Regression})^2$$

Sum Squared Regression Error — Each Data Point — Regression Value

Sum Over All The Data Points — Square The Result

$$SS_{Total} = \sum (y_i - \bar{y})^2$$

Sum Squared Total Error — Each Data Point — Mean Value

An R-squared formula [9]

Adjusted R-Squared

- It only considers the features which are important for the model
- An improvement to R-squared which gives the illusion of a good model when the features increase

$$\text{Adjusted } R^2 = 1 - \frac{(n-1)}{[n-(k+1)]}(1-R^2)$$

Formula 9-6

where n = sample size

k = number of independent (x) variables

Adjusted R-squared formula [10]

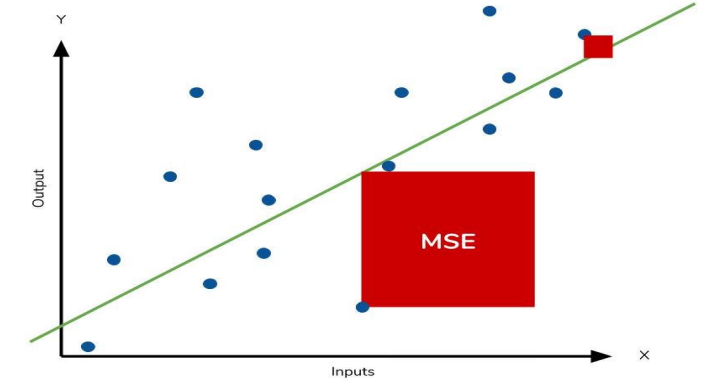
Mean squared error and root mean squared error

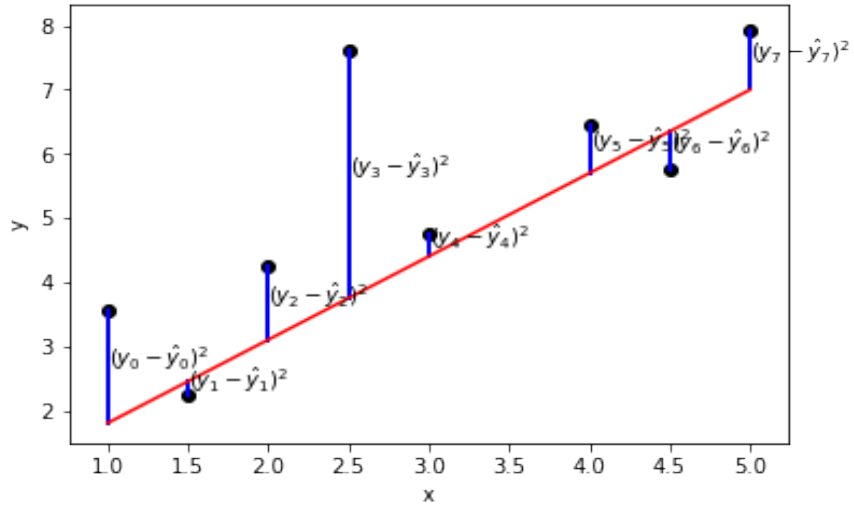
$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- MSE is measured in target units, squared

$$RMSE = \sqrt{MSE}$$

- Measure $RMSE$ in the same units at the target variable





Figures 9 & 10: MSE and RMSE [8]

14

Mean Absolute Error (MAE)

- This is used to measure the average absolute differences between predicted and actual values. It measures the prediction error.

$$MAE = \frac{1}{n}\sum\left|y - \hat{y}\right|$$

Divide by the total number of data points

Predicted output value

Actual output value

Sum of

The absolute value of the residual

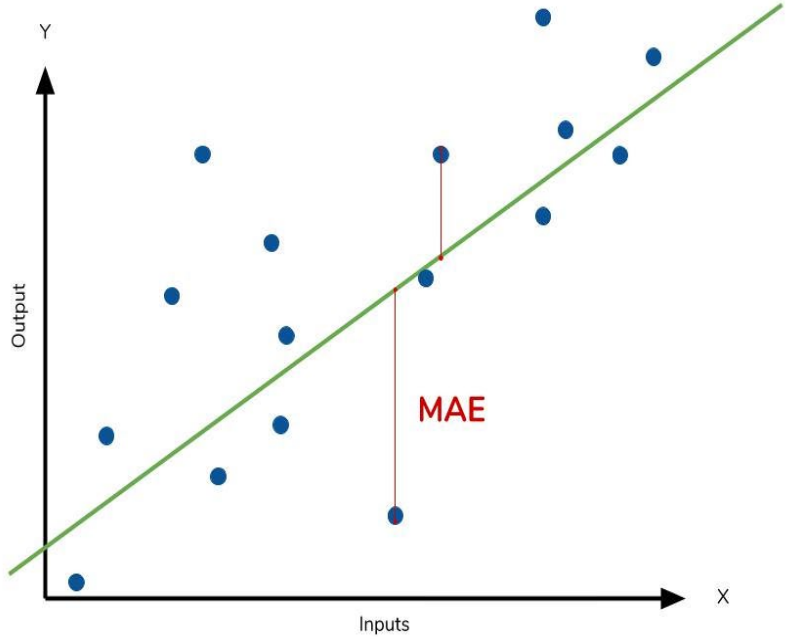


Figure 11. MAE [11]

15

OLS MODEL ASSUMPTIONS

- **Linearity** model is linear in parameters.
  F-test hypothesis testing or a scatter plot can be used to confirm this.

- **Normality** error distribution is normal with mean 0 and constant variance.
  Histograms, Kernel Density Estimate (KDE) plots, or
  Quantile-Quantile (Q-Q) plots can be used to confirm this.

- **Homoscedasticity** The residuals have constant variance at every level of the predicted value on the x-axis. A residual plot can be used to confirm this.

OLS MODEL ASSUMPTIONS

- **No Autocorrelation**

  The residuals are independent. In particular, there's no correlation between consecutive residuals in time series. Durbin Watson test can be used in testing this.
  The value of the test lies between 0 to 4. If the value of the test is 2, then there is no autocorrelation.

- **No Multicollinearity**

  No correlation should be there between the independent variables. A correlation matrix or heatmap or VIF score can be used to check this. If the VIF score is greater than 5, then the variables are highly correlated.

APPLICATIONS OF MULTIPLE LINEAR REGRESSION

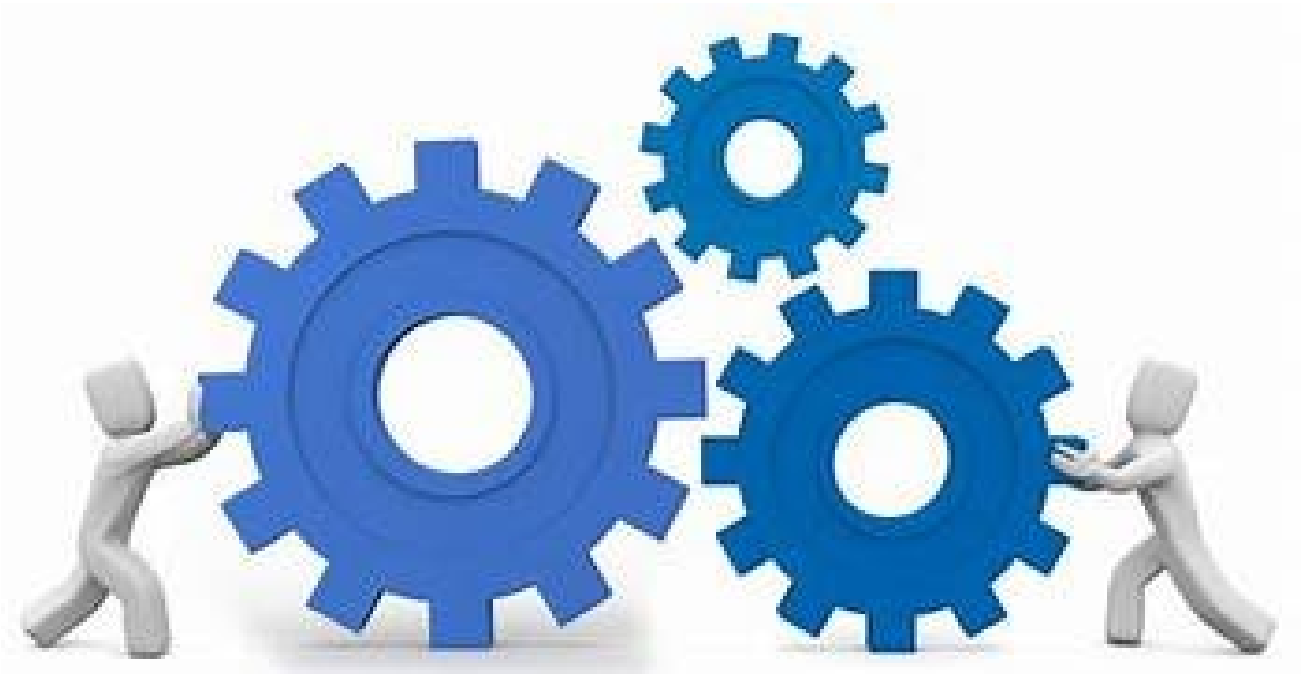
- **Various Fields Utilize Multiple Linear Regression**
  - Multiple Linear Regression finds applications in a wide range of fields due to its versatility and ability to model complex relationships. Some key application areas include:
- **Economics**
  - In economics, it's used to analyze factors affecting economic indicators like GDP, inflation, and employment rates.
  - Example: Studying the impact of interest rates, government spending, and consumer confidence on economic growth.
- **Finance**
  - In finance, it's employed to understand stock price movements, asset pricing, and risk assessment.
  - Example: Predicting stock prices based on various financial indicators like earnings, interest rates, and market volatility.
- **Social Sciences**
  - Social scientists use it to explore the determinants of social phenomena.
  - Example: Analyzing the factors influencing educational attainment, such as parental income, school quality, and family structure.

APPLICATIONS OF MULTIPLE LINEAR REGRESSION

- **Marketing**
  - In marketing, it aids in predicting sales, customer behavior, and market trends.
  - Example: Predicting product sales based on advertising expenditure, product features, and consumer demographics.
- **Environmental Sciences**
  - In environmental sciences, it's applied to understand the effects of multiple environmental factors.
  - Example: Modeling the impact of temperature, precipitation, and pollution levels on plant growth.
- **Medicine and Healthcare**
  - In healthcare, it helps predict patient outcomes, disease risk, and treatment effectiveness.
  - Example: Predicting patient readmission rates based on factors like age, comorbidities, and treatment protocols.
- **Engineering**
  - Engineers use it for various purposes, such as predicting equipment performance and optimizing processes.
  - Example: Predicting the wear and tear of machinery based on operating conditions and maintenance schedules.

Multiple Linear Regression

[12]



IMPLEMENTATION FLOW CHART

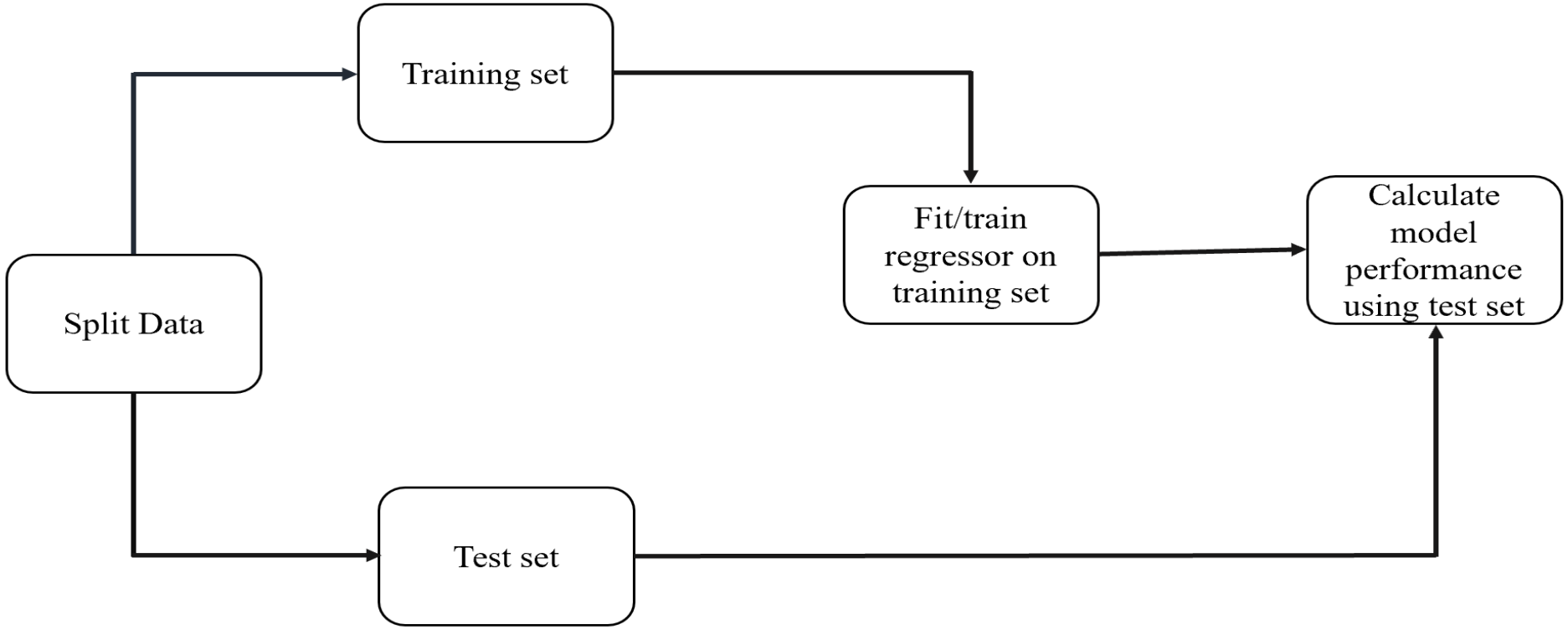


Figure 12. Implementation work flow chart [8]

21

SAMPLE DATASET

- The dataset illustrates how sales are influenced by the type of advertisement and the corresponding advertising expenses [13].

In [3]:
```
## Call the dataframe and do basic checks
data.head()
```

Out[3]:

	TV	Radio	Sales
0	230.1	37.8	22.1
1	44.5	39.3	10.4
2	17.2	45.9	9.3
3	151.5	41.3	18.5
4	180.8	10.8	12.9

In [4]:
```
#checking the statistics of the numerical variables
data.describe()
```

Out[4]:

	TV	Radio	Sales
count	20.000000	20.000000	20.00000
mean	119.310000	27.740000	13.49500
std	84.178425	15.442573	5.43289
min	8.600000	2.100000	4.80000
25%	54.250000	17.400000	9.60000
50%	108.850000	32.850000	12.15000
75%	196.500000	39.375000	17.67500
max	281.400000	48.900000	24.40000

- The dataset comprises 20 rows and includes 2 features (TV and Radio) along with 1 target variable (Sales).

In [31]:
```
#checking the shape of the dataset
data.shape
```

Out[31]: (20, 3)

In [5]:
```
#checking basic info about the data
data.info()

```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20 entries, 0 to 19
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   TV      20 non-null     float64
 1   Radio   20 non-null     float64
 2   Sales   20 non-null     float64
dtypes: float64(3)
memory usage: 608.0 bytes
```

PREDICTING SALES

Loading Data into Pandas

```
#importing library
#loading the data
import pandas as pd
data=pd.read_csv('media_marketing.csv')
print(data.head())
```

```
      TV  Radio  Sales
0  230.1   37.8   22.1
1   44.5   39.3   10.4
2   17.2   45.9    9.3
3  151.5   41.3   18.5
4  180.8   10.8   12.9
```

Splitting Data into X and y

```
# X variable contains all features
X=data[['TV','Radio']]

#y variable is the target variable
y=data.Sales
```

Splitting Data into X and y

```
# X variable contains all features
X=data[['TV','Radio']]

#y variable is the target variable
y=data.Sales
```

Splitting Data into Train and Test Set

```
# Traning and testing data creation
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.3,random_state=42)

#Checking the shapes of train and test data
print(X_train.shape,X_test.shape,y_train.shape,y_test.shape)
```

```
(14, 2) (6, 2) (14,) (6,)
```

Model Creation

```
#import the necessary library package
from sklearn.linear_model import LinearRegression

# object creation
Lin_Reg_all=LinearRegression()

#training of linear regression
Lin_Reg_all.fit(X_train,y_train)

#predicting y values
y_predict=Lin_Reg_all.predict(X_test)
```

THE REGRESSION MODEL

The Model Coefficients and Intercept

```
print('The coefficients of TV and Radio are: ',Lin_Reg_all.coef_)
print('The intercept of the regression model is: ',Lin_Reg_all.intercept_)
print('The error term of the model is: ',sum(residual))
```

```
The coefficients of TV and Radio are:  [0.04748281 0.1615382 ]
The intercept of the regression model is:  3.3550354306286057
The error term of the model is:  -0.12557388552418125
```

The multiple linear regression equation

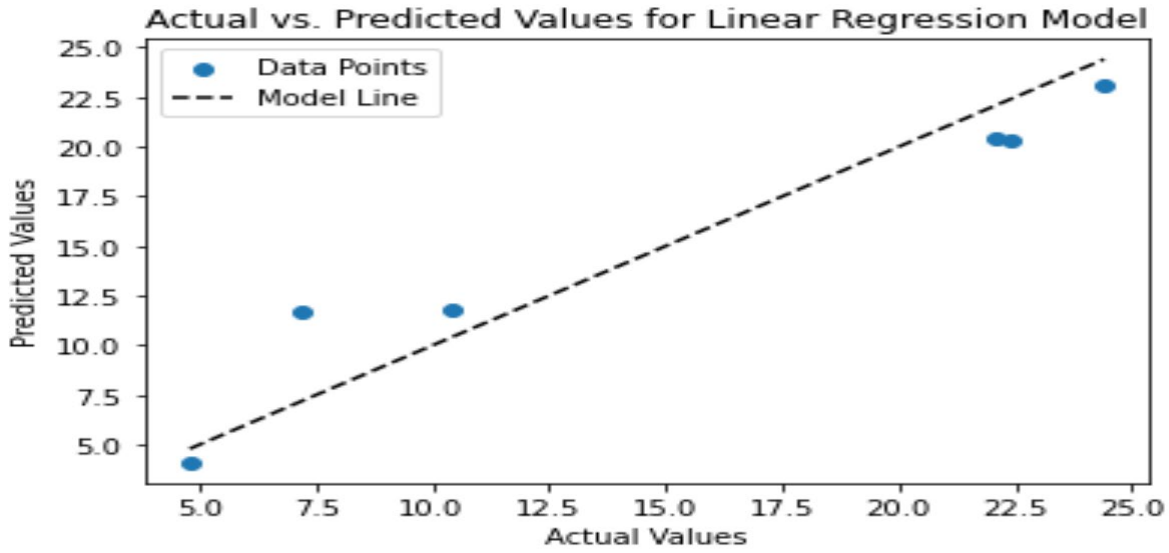
Sales = 0.04748281*TV + 0.1615382*Radio + 3.3550354306286057 – 0.12557388552418125

Plotting of Residuals

```
# Calculate residuals
residual  = y_test - y_predict

#importing relevant library
import matplotlib.pyplot as plt
%matplotlib inline

# Plot the actual vs. predicted values
plt.scatter(y_test, y_predict, label="Data Points")
plt.plot([min(y_test), max(y_test)], [min(y_test), max(y_test)], '--k', label="Model Line")
plt.xlabel("Actual Values")
plt.ylabel("Predicted Values")
plt.title("Actual vs. Predicted Values for Linear Regression Model")
plt.legend()
plt.show()
```



24

Model Performance Evaluation

```python
#importing relevant library
from sklearn.metrics import r2_score,mean_squared_error,mean_absolute_error

#Checking the R2 value of the model
r2score=r2_score(y_test,y_predict)
print('The R-squared value of the model is: ',r2score*100)

## calculation of adjusted r2 score
adjusted_r2 = 1-(1-r2score)*(6-2)/(6-2-1)
print('The adjusted R-squared value of the model is: ',adjusted_r2*100)

#Calculating the mean squared error
mse = mean_squared_error(y_test,y_predict,squared = True)
print('The mean squared error of the model is ',mse)

#calculating the root mean squared error
rmse = mean_squared_error(y_test,y_predict,squared = False)
print('The mean squared error of the model is ',rmse)

#Calculating the mean absolute error
mae = mean_absolute_error(y_test,y_predict)
print('The mean absolute error of the model is ',mae)
```

```
The R-squared value of the model is:  91.75064812125913
The adjusted R-squared value of the model is:  89.00086416167883
The mean squared error of the model is  5.21480487527971
The mean squared error of the model is  2.283594726583443
The mean absolute error of the model is  1.940346129459319 6
```


25

APPENDIX

Implementation Files

Jupyter Notebook file

C:\Users\user\
oads\Semester_3\I

Dataset

Microsoft Excel
ma Separated Valu

26

REFERENCES

[1] B. Mutea, “Linear regression in Python with Scikit-learn (With examples, code, and notebook),” *Mlnuggets*, Sep. 08, 2022. https://www.machinelearningnuggets.com/python-linear-regression/ (accessed Oct. 05, 2023).

[2] “Stick figure Presenter meeting,” *PresenterMedia*. https://www.presentermedia.com/powerpoint-clipart/stick-figure-presenter-meeting-pid-3268 (accessed Oct. 05, 2023).

[3] S. Srivastava, “Supervised learning — the What, When, Why, Good and Bad (Part 1),” *Medium*, Mar. 30, 2022. Accessed: Oct. 05, 2023. [Online]. Available: https://towardsdatascience.com/supervised-learning-the-what-when-why-good-and-bad-part-1-f90e6fe2a606

[4] E. Informer, “Regression Analysis (Types, Uses, and Tips),” *ERP Information*, Apr. 25, 2023. https://www.erp-information.com/regression-analysis.html (accessed Oct. 05, 2023).

[5] Thor, “PPT - Introduction to Regression Analysis, Chapter 13, PowerPoint Presentation - ID:3195508,” *SlideServe*, Aug. 14, 2014. https://www.slideserve.com/thor/introduction-to-regression-analysis-chapter-13 (accessed Oct. 05, 2023).

[6] “Concept | Regression algorithms — Dataiku Knowledge Base.” https://knowledge.dataiku.com/latest/ml-analytics/ml-concepts/concept-regression.html (accessed Oct. 06, 2023).

[7] A. Chakure, “Linear Regression and its Mathematical implementation,” *Medium*, Dec. 10, 2021. Accessed: Oct. 06, 2023. [Online]. Available: https://medium.datadriveninvestor.com/linear-regression-and-its-mathematical-implementation-29d520a75ede

REFERENCES

[8] "Regression - Introduction to regression," *DataCamp*. Accessed: Oct. 02, 2023. [Online]. Available: https://campus.datacamp.com/courses/supervised-learning-with-scikit-learn/regression-6320c92e-31c3-48fb-9382-6a9169125722?ex=1

[9] "What does negative R-squared mean?" https://stats.stackexchange.com/questions/183265/what-does-negative-r-squared-mean (accessed Oct. 06, 2023).

[10] W. Paul, *Created by Erin Hodgess, Houston, Texas*. 2017. [Online]. Available: https://slideplayer.com/slide/8485605/

[11] "A guide on regression error metrics (MSE, RMSE, MAE, MAPE, sMAPE, MPE) with Python code," *Amir Masoud Sefidian - Sefidian Academy*, Apr. 21, 2023. https://sefidian.com/2022/08/18/a-guide-on-regression-error-metrics-with-python-code/ (accessed Sep. 25, 2023).

[12] Grewalr, "Reddit - Dive into anything." https://www.reddit.com/user/GREWALR1/comments/idolao/multiple_linear_regression_model_python/ (accessed Oct. 06, 2023).

[13] "Advertising dataset," *Kaggle*, May 04, 2023. https://www.kaggle.com/datasets/tawfikelmetwally/advertising-dataset (accessed Oct. 06, 2023).

[14] codebasics, "Machine Learning Tutorial Python - 3: Linear Regression Multiple variables," *YouTube*. Jul. 04, 2018. Accessed: Sep. 23, 2023. [Online]. Available: https://www.youtube.com/watch?v=J_LnPL3Qg70