Analysis of the suitability of the Linear Regression model for house price prediction

By
O.G.V.V Bandara

Department of Mathematics
University of Ruhuna
Matara.

2023

Acknowledgement

This project was done as a passion project to explore the linear regression models and it's different variants. Also to assess the suitability of the linear regression models for certain data sets and what type of data sets and data this model is applicable.

Abstract

The project explores the two linear regression models: "simple linear regression" and "multiple linear regression". The machine learning algorithm was developed in python. The Scikit-learn library and it's functionalities were used in this project.

Contents

Chapter 1

Introduction

1.1 Background of study

1.1.1 Simple Linear Regression

The simple linear regression model consists of 2 variables β_0, β_1

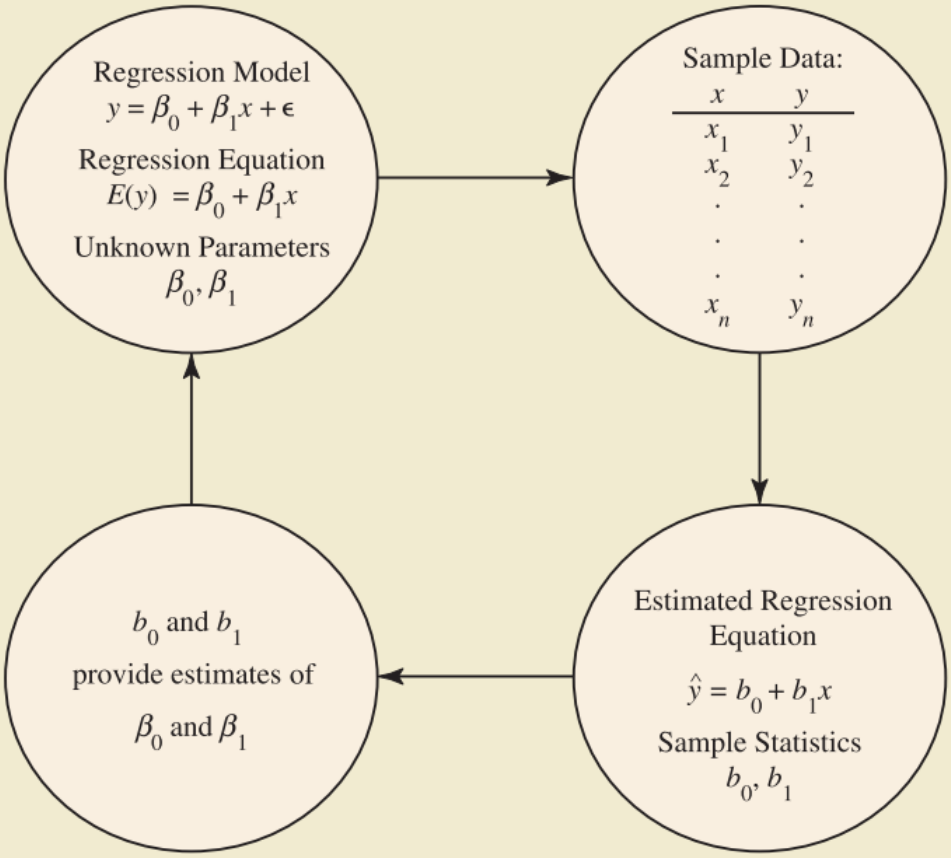
$$y = \beta_0 + \beta_1 * x + \epsilon \tag{1.1}$$

β_0: parameter of model
β_1: parameter of model
ϵ: random variable(error term accounts for the variability in y that cannot be accounted for by the linear relationship between x and y)

When the population statistics are unknown the parameters have to be estimated using sample data. The new equation with the estimated values is known as the "estimated regression equation". The estimated parameters using sample data are
$\hat{\beta}_0$ and $\hat{\beta}_1$ are replaced for β_0 and β_1

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 * x \tag{1.2}$$

The process of estimation of the parameters is done in the following process


Regression Model
$y = \beta_0 + \beta_1 x + \epsilon$
Regression Equation
$E(y) = \beta_0 + \beta_1 x$
Unknown Parameters
β_0, β_1
Sample Data:
x y
x_1 y_1
x_2 y_2
x_n y_n
Estimated Regression Equation
$\hat{y} = b_0 + b_1 x$
Sample Statistics
b_0, b_1
b_0 and b_1
provide estimates of
β_0 and β_1

Least square fit method

The least squares method is a procedure for using sample data to find the estimated regression equation.

$$RSS = e_1^2 + e_2^2 + ... + e_n^2 \tag{1.3}$$

the values can be given as

$$e_1 = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 * x_1) \tag{1.4}$$

$$e_2 = (y_2 - \hat{\beta}_0 - \hat{\beta}_2 * x_2) \tag{1.5}$$

$$... \tag{1.6}$$

$$e_n = (y_n - \hat{\beta}_0 - \hat{\beta}_n * x_n) \tag{1.7}$$

The least square approach chooses the $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS.then

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x_i})(y_i - \bar{y_i})}{\sum_{i=1}^{n}(x_i - \bar{x_i})^2} \tag{1.8}$$

the new regression line

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 * \bar{x} \tag{1.9}$$

where

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \tag{1.10}$$

and

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1.11}$$

1.1.2 Multiple Linear Regression

The regression coefficient β_1 is interpreted as the expected change in Y associated with a 1-unit increase in x_1 while $x_2, ..., x_p$ are held fixed.

$$y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + ... + \beta_p * x_p + \epsilon \tag{1.12}$$

β_0: intercept
p:number of predictors
ϵ:error term

considering our model for house price prediction, there are multiple numerical variables. only numerical variables can be considered as independent variables and missing values of each variables have to replaced with the mean value or median.

$$price = \beta_0 + \beta_1 * crimeRate + ... + \beta_{16} * avgDist \tag{1.13}$$

Estimating regression coefficients

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y_i})^2 \tag{1.14}$$

1.2 Objectives

- The main objective of this experiment is to assess the suitability of The simple linear regression model and the multiple linear regression model for predicting the house price.
- Another objective is to identify the use of the scikit-learn library in developing machine learning algorithms for various models including the simple linear regression model and the multiple linear regression model.

Chapter 2

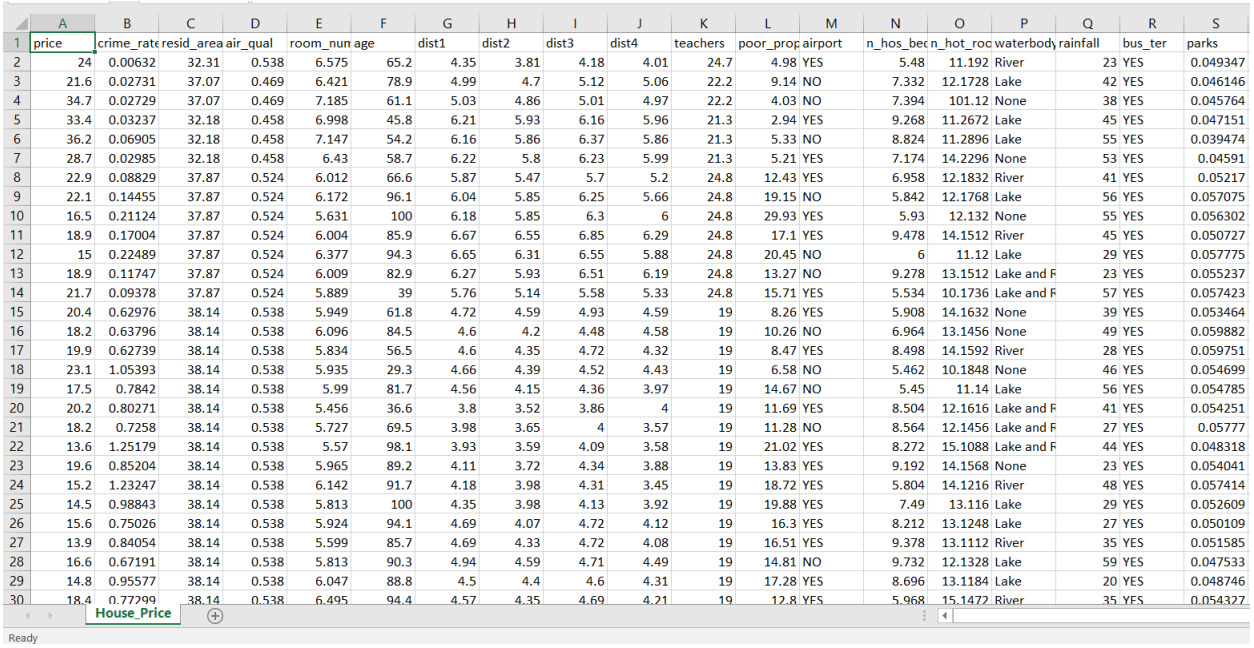
Problem Statement

When selecting a suitable residence the buyer has to consider many factors. considering each factor the buyer has to assess whether the house is suitable for purchase or worth the value. Through the use of machine learning the objective is to ease this process for the user by providing various statistics and predictions such that the buyer can assess his/her decisions accurately. The machine learning models can consider a large data set and compare the data in a larger scope when the buyer can only do so in a small spectrum.

Chapter 3

Methodology

The methods used here are the "Simple Linear Regression" models and the "Multiple Linear Regression" models. Initially a data set was obtained. The data set was exported as a .csv format.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	price	crime_rate	resid_area	air_qual	room_num	age	dist1	dist2	dist3	dist4	teachers	poor_prop	airport	n_hos_bed	n_hot_roo	waterbody	rainfall	bus_ter	parks
2	24	0.00632	32.31	0.538	6.575	65.2	4.35	3.81	4.18	4.01	24.7	4.98	YES	5.48	11.192	River	23	YES	0.049347
3	21.6	0.02731	37.07	0.469	6.421	78.9	4.99	4.7	5.12	5.06	22.2	9.14	NO	7.332	12.1728	Lake	42	YES	0.046146
4	34.7	0.02729	37.07	0.469	7.185	61.1	5.03	4.86	5.01	4.97	22.2	4.03	NO	7.394	101.12	None	38	YES	0.045764
5	33.4	0.03237	32.18	0.458	6.998	45.8	6.21	5.93	6.16	5.96	21.3	2.94	YES	9.268	11.2672	Lake	45	YES	0.047151
6	36.2	0.06905	32.18	0.458	7.147	54.2	6.16	5.86	6.37	5.86	21.3	5.33	NO	8.824	11.2896	Lake	55	YES	0.039474
7	28.7	0.02985	32.18	0.458	6.43	58.7	6.22	5.8	6.23	5.99	21.3	5.21	YES	7.174	14.2296	None	53	YES	0.04591
8	22.9	0.08829	37.87	0.524	6.012	66.6	5.87	5.47	5.7	5.2	24.8	12.43	YES	6.958	12.1832	River	41	YES	0.05217
9	22.1	0.14455	37.87	0.524	6.172	96.1	6.04	5.85	6.25	5.66	24.8	19.15	NO	5.842	12.1768	Lake	56	YES	0.057075
10	16.5	0.21124	37.87	0.524	5.631	100	6.18	5.85	6.3	6	24.8	29.93	YES	5.93	12.132	None	55	YES	0.056302
11	18.9	0.17004	37.87	0.524	6.004	85.9	6.67	6.55	6.85	6.29	24.8	17.1	YES	9.478	14.1512	River	45	YES	0.050727
12	15	0.22489	37.87	0.524	6.377	94.3	6.65	6.31	6.55	5.88	24.8	20.45	NO	6	11.12	Lake	29	YES	0.057775
13	18.9	0.11747	37.87	0.524	6.009	82.9	6.27	5.93	6.51	6.19	24.8	13.27	NO	9.278	13.1512	Lake and R	23	YES	0.055237
14	21.7	0.09378	37.87	0.524	5.889	39	5.76	5.14	5.58	5.33	24.8	15.71	YES	5.534	10.1736	Lake and R	57	YES	0.057423
15	20.4	0.62976	38.14	0.538	5.949	61.8	4.72	4.59	4.93	4.59	19	8.26	YES	5.908	14.1632	None	39	YES	0.053464
16	18.2	0.63796	38.14	0.538	6.096	84.5	4.6	4.2	4.48	4.58	19	10.26	NO	6.964	13.1456	None	49	YES	0.059882
17	19.9	0.62739	38.14	0.538	5.834	56.5	4.6	4.35	4.72	4.32	19	8.47	YES	8.498	14.1592	River	28	YES	0.059751
18	23.1	1.05393	38.14	0.538	5.935	29.3	4.66	4.39	4.52	4.43	19	6.58	NO	5.462	10.1848	None	46	YES	0.054699
19	17.5	0.7842	38.14	0.538	5.99	81.7	4.56	4.15	4.36	3.97	19	14.67	NO	5.45	11.14	Lake	56	YES	0.054785
20	20.2	0.80271	38.14	0.538	5.456	36.6	3.8	3.52	3.86	4	19	11.69	YES	8.504	12.1616	Lake and R	41	YES	0.054251
21	18.2	0.7258	38.14	0.538	5.727	69.5	3.98	3.65	4	3.57	19	11.28	NO	8.564	12.1456	Lake and R	27	YES	0.05777
22	13.6	1.25179	38.14	0.538	5.57	98.1	3.93	3.59	4.09	3.58	19	21.02	YES	8.272	15.1088	Lake and R	44	YES	0.048318
23	19.6	0.85204	38.14	0.538	5.965	89.2	4.11	3.72	4.34	3.88	19	13.83	YES	9.192	14.1568	None	23	YES	0.054041
24	15.2	1.23247	38.14	0.538	6.142	91.7	4.18	3.98	4.31	3.45	19	18.72	YES	5.804	14.1216	River	48	YES	0.057414
25	14.5	0.98843	38.14	0.538	5.813	100	4.35	3.98	4.13	3.92	19	19.88	YES	7.49	13.116	Lake	29	YES	0.052609
26	15.6	0.75026	38.14	0.538	5.924	94.1	4.69	4.07	4.72	4.12	19	16.3	YES	8.212	13.1248	Lake	27	YES	0.050109
27	13.9	0.84054	38.14	0.538	5.599	85.7	4.69	4.33	4.72	4.08	19	16.51	YES	9.378	13.1112	River	35	YES	0.051585
28	16.6	0.67191	38.14	0.538	5.813	90.3	4.94	4.59	4.71	4.49	19	14.81	NO	9.732	12.1328	Lake	59	YES	0.047533
29	14.8	0.95577	38.14	0.538	6.047	88.8	4.5	4.4	4.6	4.31	19	17.28	YES	8.696	13.1184	Lake	20	YES	0.048746
30	18.4	0.77299	38.14	0.538	6.495	94.4	4.57	4.35	4.69	4.21	19	12.8	YES	5.968	15.1472	River	35	YES	0.054327

House_Price

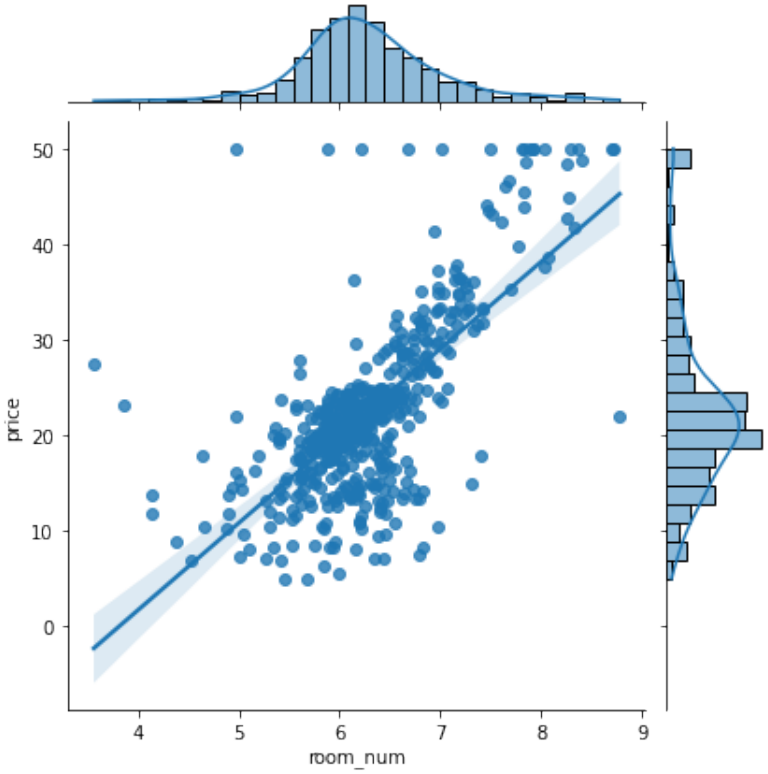
Ready

the data set is very large and a small portion is only displayed above.

The data was analyzed using the dataframe analysis tools of the "pandas" libraries in python.

Simple linear regression model

Since simple linear Regression only uses two variables, the dependent and the independent, the independent variable was considered as $room - num$ and the dependent variable was *price*.The relationship between these two variables can be visualized through a jointplot.



A code was developed in python using the "scikit learn" library in order to obtain the machine learning algorithm for linear regression and obtain predictions.

Multiple linear regression model

The same data set "house price" was used for analysis. The number of dependent variables remain the same but the number of independent variables increased in multiple linear regression. Instead of a 2 dimensional relationship between independent and dependent variables, here we can observe a higher dimensional relationship which is either a plane or 3D object or of higher dimension etc.

3.0.1 error calculation

Mean squared error

In statistics, the mean squared error or mean squared deviation of an estimator measures the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value.

$$MSE = \frac{1}{n} * \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3.1}$$

MSE=mean squared error
n= number of data points
y_i=observed values
$\hat{y}_i$=predicted values

R squared error

the error is a measure of the accuracy of the model fitting. This is a statistical measure of how well the model fits the data. This is importent when the statistical model is used for prediction.

$$R^2 = 1 - \frac{sumsquaredregression(SSR)}{Totalsumofsquares(SST} \tag{3.2}$$

$$= 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \tag{3.3}$$

Chapter 4

Conclusion

4.0.1 Simple linear regression

Mean squared error

```
#mean squared error
from sklearn.metrics import mean_squared_error
mse=mean_squared_error(y_test,y_pred)
print("mean square errror is:",mse)
```

the mean squared error is: 46.630761887451136 the low mean squared error implies that the model is accurate.

Root mean squared error

```
#root mean squared error
import numpy as np
rmse=np.sqrt(mse)
print("root mean squared error is:",rmse)
```

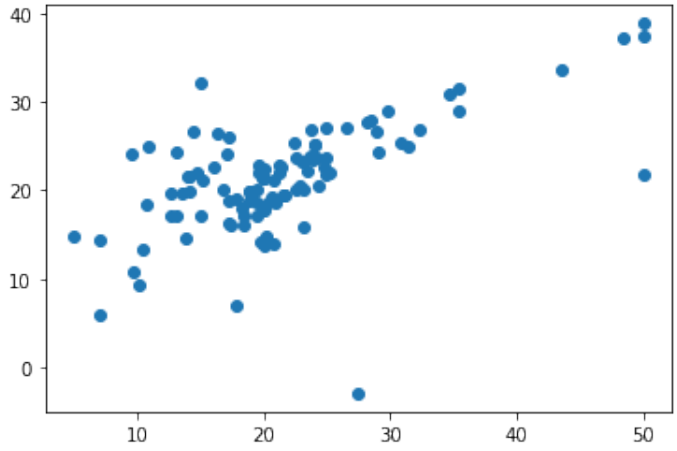
the root mean squared error is: 6.828672044215562

R squared error

```
#r squared error
from sklearn.metrics import r2_score
r2=r2_score(y_test,y_pred)
print("r squared error is:",r2)
```

the R squared error is: 0.3676692756441877 the R squared error is closer to zero implies that the model requires furthur adjustment or replacement of fitting.

Visualizing the relationship between predicted dependent variable values and the test set dependent variable values



4.0.2 Multiple linear regression

Mean squared error

```
from sklearn.metrics import mean_squared_error
mse = mean_squared_error(y_test, y_pred)
print(mse)
```

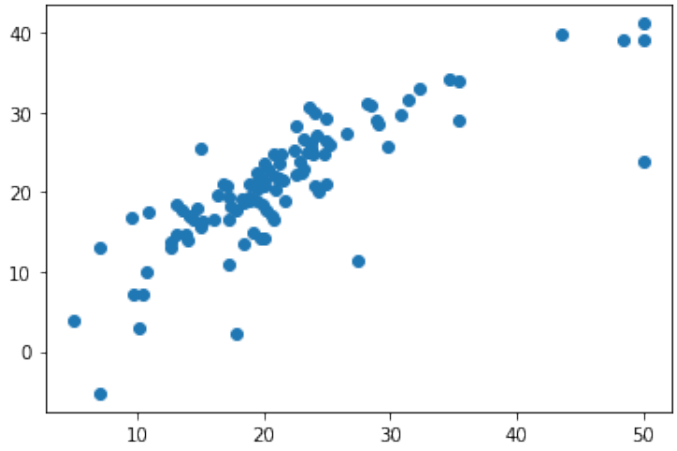
the mean squared error was obtained as:26.05088254457464
considering the values of price in the initial data, the prediction seam to vary only by a small value.

R squared error

```
from sklearn.metrics import r2_score
r2 = r2_score(y_test, y_pred)
print(r2)
```

The value of the R squared error was obtained to be:0.6467402040464607
The higher R squared value shows the model performance is good

Visualizing the relationship between predicted values of dependent variable and test values of dependent variable



Chapter 5

Appendix

5.0.1 Simple Linear Regression

```
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
import pandas as pd

#importing the csv file as a dataframe using pandas
df=pd.read_csv("House_Price.csv")
print(df)

#assigning the independent variable
X=df[["room_num"]]#x variable should be a two dimensional array

#assigning the dependent variable
y=df["price"]

#splitting the dataset into test set and train set
X_train,X_test,y_train,y_test=train_test_split(X,y, test_size=0.2, random_state=42)

linoB=LinearRegression() #the object for the linear regression function

#fitting the model
linoB.fit(X_train,y_train)

#see untercepts and coefficients
print(linoB.intercept_,linoB.coef_)#underscores mean that they are attributes of
linear regression model

#predicting the values of the test set
y_pred = linoB.predict(X_test)

#the predicted values
print(y_pred)
```

5.0.2 Multiple linear regression

```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import StandardScaler

df=pd.read_csv("House_Price.csv")
print(df)

#isnull() and any() functions from Pandas
will determine if there are missing or invalid values
print(df.isnull().any())

#prints the list of columns with missing values
print(df.dropna())

#replacing missing values with mean or median
df.fillna(df.mean(), inplace=True) #This will replace all NaN values in the DataFrame

#independent variables
X_multi=df.drop(["price","airport","waterbody","bus_ter"],axis=1)
#drop command will drop the price column and
retreive all other
#columns
#axis=1 specifies the columns

print(X_multi)

#creating dependent variable
y_multi=df["price"]

print(y_multi)

#splitting the data into test set and train set
X_train, X_test, y_train, y_test = train_test_split(X_multi,
y_multi, test_size=0.2, random_state=42)

#preprocessing the data by standardisation of test set and train set
scaler = StandardScaler() #assigning the StandardScaler function to scaler object
X_train = scaler.fit_transform(X_train) #train set transformation
X_test = scaler.transform(X_test) #test set transformation

#training the train data using linear regression
model = LinearRegression() #assigning LinearRegression function to object model
model.fit(X_train, y_train) #fitting the training set
```

```
#performing predictions on the test set
y_pred = model.predict(X_test)
print(y_pred)
```

Bibliography

[1] Politeknik NSC surabaya: Accounting (no date) Politeknik NSC Surabaya — Accounting. Available at: https://nscpolteksby.ac.id/ebook/book/accounting/ (Accessed: March 6, 2023).

[2] 13 multiple linear(regression(- university of Colorado Boulder (no date). Available at: https://www.colorado.edu/amath/sites/default/files/attached-files/lesson12$_m ultregression.pdf(Accessed : March6, 2023)$.

[3] Mean squared error (2022) Wikipedia. Wikimedia Foundation. Available at: https://en.wikipedia.org/wiki/Mean$_s quared_e rror(Accessed : March7, 2023)$.

[4] (no date) Numeracy, Maths and statistics - academic skills kit. Available at: https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/coefficient-of-determination-r-squared.html (Accessed: March 7, 2023).