Name: Jigar Siddhpura **SAPID:** 60004200155
DIV: C/C2 **Branch:** Computer Engineering

DMW EXP 4

.

AIM: Implementation of Linear Regression

1. Single Variate

2. Multi Variate

THEORY:

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering and the number of independent variables being used.
Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique fnds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression. In the fgure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

Hypothesis function for Linear Regression :

$$y = \theta_1 + \theta_2.x$$

While training the model we are given : x:input training data (univariate – one

input variable(parameter)) y:labels to data (supervised learning)

When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by fnding the best θ1 and θ2 values. θ1: intercept

θ2: coeffcient of x

Once we find the best θ1 and θ2 values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

COST FUNCTION(J):

By achieving the best-fIt regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum. So, it is very important to update the θ1 and θ2 values, to reach the best value that minimize the error between predicted y value (pred) and true y value (y).

$$minimize \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

$$J = \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

Cost function(J) of Linear Regression is the Root Mean Squared Error (RMSE) between predicted y value (pred) and true y value (y).

CODE:

```
from google.colab import drive
drive.mount('/content/gdrive')

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import mean_squared_error

df =
pd.read_csv('/content/gdrive/MyDrive/DMW/datasets/StudentsPerformance.csv')
df.head()

df['final_score'] = df.apply(lambda x: (x['math score'] + x['reading score']
+ x['writing score'])/3, axis=1)

df2 = pd.get_dummies(df, columns=['gender','lunch','parental level of
education','race/ethnicity','test preparation course'])
df2 = df2.drop(['math score','reading score','writing score'],axis=1)

# multi-variate
y = df2['final_score']
X = df2.drop(['final_score'],axis=1)
xtrain, xtest, ytrain, ytest =
train_test_split(X,y,test_size=0.25,random_state=10)
sns.boxplot(data=df2['final_score'],orient='h')
```

```
model = LinearRegression()
model.fit(xtrain,ytrain)
score = model.score(xtest,ytest)
print(score)
ypred = model.predict(xtest)

sns.scatterplot(data=df,x=df['reading score'],y=df['final_score'])

#regression line
m,b = np.polyfit(x=df['reading score'],y=df['final_score'],deg=1)
X = df['reading score']
plt.plot(X, m*X+b)

# univariate
X_uni = df['reading score']
y_uni = df['final_score']
x_uni_train, x_uni_test, y_uni_train, y_uni_test =
train_test_split(X_uni,y_uni,test_size=0.25,random_state=10)
x_uni_train = x_uni_train.values.reshape(-1,1)
x_uni_test = x_uni_test.values.reshape(-1,1)
uni_model = LinearRegression()
uni_model.fit(x_uni_train,y_uni_train)
uni_score = uni_model.score(x_uni_test,y_uni_test)
print(uni_score)
y_uni_pred = uni_model.predict(x_uni_test)
```

OUTPUT:

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	female	group B	bachelor's degree	standard	none	72	72	74
1	female	group C	some college	standard	completed	69	90	88
2	female	group B	master's degree	standard	none	90	95	93
3	male	group A	associate's degree	free/reduced	none	47	57	44
4	male	group C	some college	standard	none	76	78	75

head() of the database

	gender	race/ethnicity	parental level of education	lunch	test preparation course	final_score
0	female	group B	bachelor's degree	standard	none	72.666667
1	female	group C	some college	standard	completed	82.333333
2	female	group B	master's degree	standard	none	92.666667
3	male	group A	associate's degree	free/reduced	none	49.333333
4	male	group C	some college	standard	none	76.333333

df.head() after adding a final score column

	final_score	gender_female	gender_male	lunch_free/reduced	lunch_standard	parental level of education_associate's degree	parental level of education_bachelor's degree	parental level of education_high school	parental level of education_master's degree	parental level of education_some college	parental educatic high
0	72.666667	1	0	0	1	0	1	0	0	0	
1	82.333333	1	0	0	1	0	0	0	0	1	
2	92.666667	1	0	0	1	0	0	0	1	0	
3	49.333333	0	1	1	0	1	0	0	0	0	
4	76.333333	0	1	0	1	0	0	0	0	1	

df.head() after applying One-hot encoding to the dataset

Considering Multivariate Linear Regression

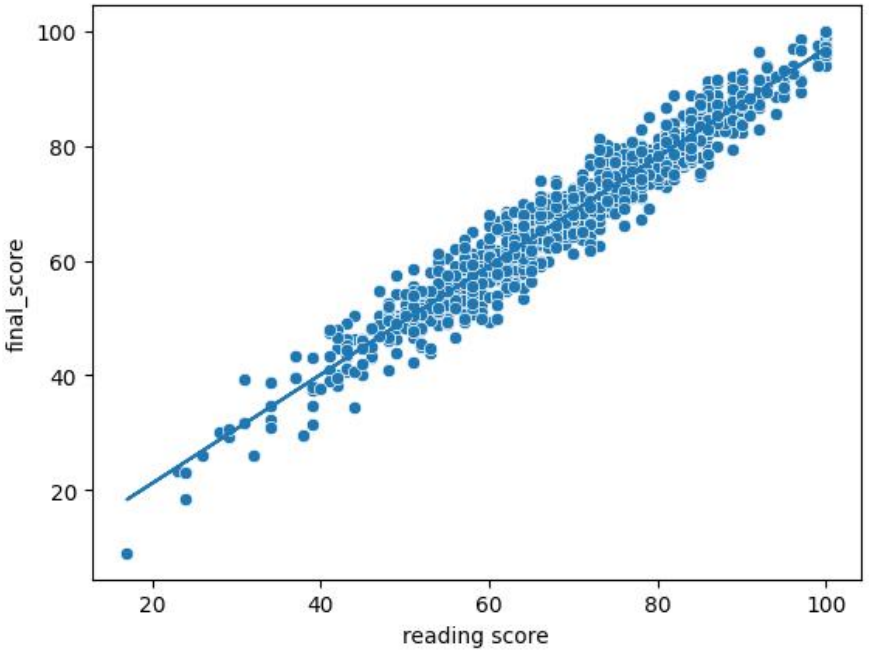
Prediction Score

```
score = model.score(xtest,ytest)
print(score)
ypred = model.predict(xtest)
```

```
0.19674412629893356
```

Now considering Univariate Linear Regression with Reading Score as the feature

Scatter plot



Prediction Score of Univariate LR

```
uni_score = uni_model.score(x_uni_test,y_uni_test)
print(uni_score)
y_uni_pred = uni_model.predict(x_uni_test)
```

```
0.944431815987387
```

CONCLUSION:

We have implemented Multivariate and Univariate Linear
Regression on a dataset and have observed the differences in their AccuracyScore and Mean Squared Errors. We observe 19.67% accuracy in the case of
Multivariate whereas in the case ofUnivariate, the accuracy score is
94.43%

94.21% and the Mean Squared Error is 109.48. Therefore we can conclude that using Multivariate Linear Regression is better than using Univariate but nevertheless the efciency of Univariate is still great.
