Analytics Vidhya

#AV10DAYSOFMLCODE

DAY 1: LINEAR REGRESSION

Analytics Vidhya

Linear Regression

- It is used to estimate real values (cost of houses, number of calls, total sales etc.) based on continuous variable(s).
- Here, we establish relationship between independent and dependent variables by fitting a best line. This best fit line is known as regression line and represented by a linear equation

$$Y = a * X + b.$$

- The best way to understand linear regression is to relive this experience of childhood.

- Let us say, you ask a child in fifth grade to arrange people in his class by increasing order of weight, without asking them their weights!
- *What do you think the child will do?*
- He / she would likely look (visually analyze) at the height and build of people and arrange them using a combination of these visible parameters. This is linear regression in real life!
- The child has actually figured out that height and build would be correlated to the weight by a relationship, which looks like the equation in the next slide.

Analytics Vidhya

$$Y = a * X + b$$

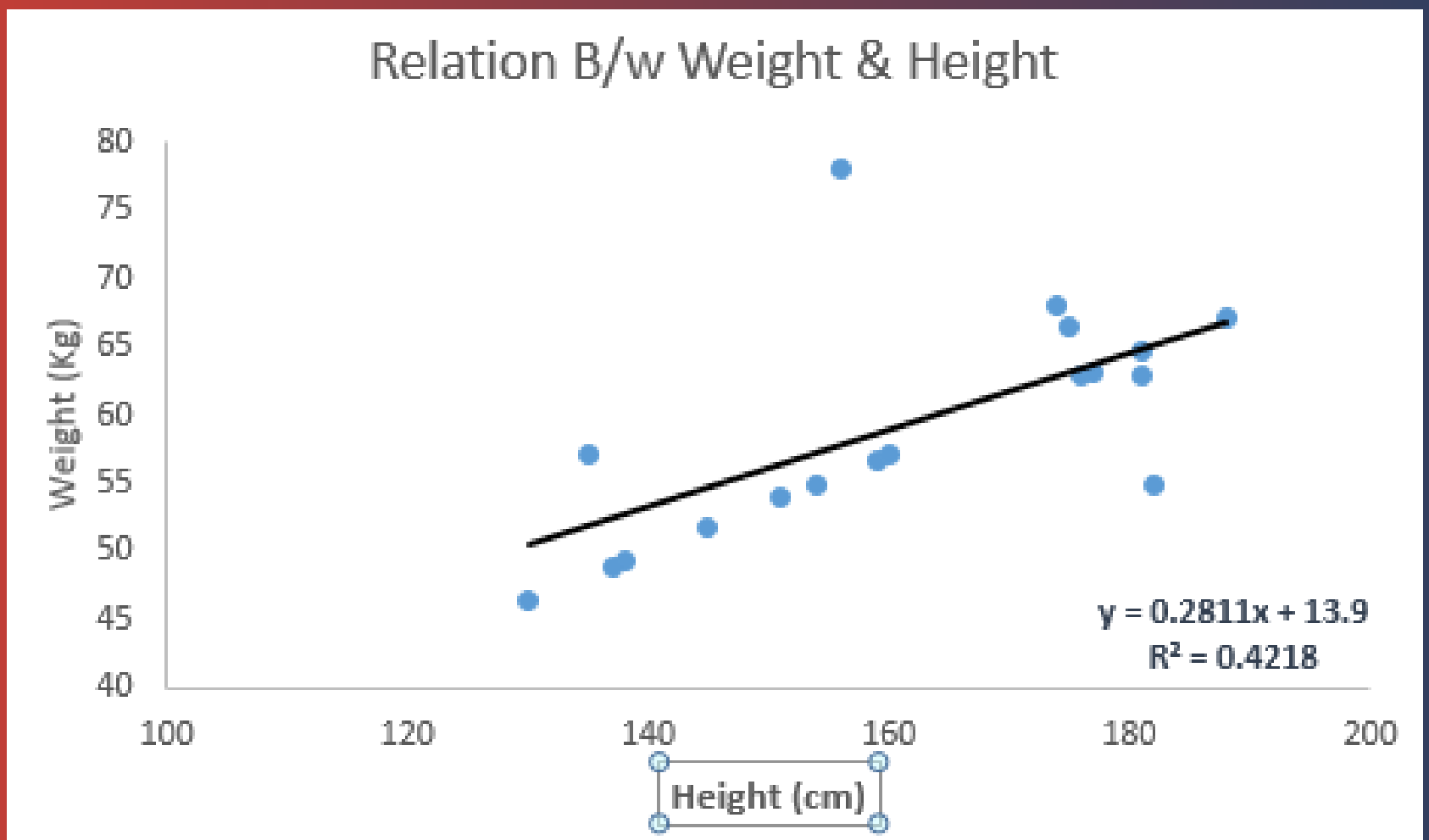
In this equation:

- Y – Dependent Variable
- a – Slope
- X – Independent variable
- b – Intercept

These coefficients a and b are derived based on minimizing the sum of squared difference of distance between data points and regression line.

Analytics Vidhya

- Look at the below example. Here we have identified the best fit line having linear equation y=0.2811x+13.9. Now using this equation, we can find the weight, knowing the height of a person.



Analytics Vidhya

Linear Regression

Simple Linear Regression

Simple Linear Regression is characterized by one independent variable.

Multiple Linear Regression

Multiple Linear Regression(as the name suggests) is characterized by multiple (more than 1) independent variables.

While finding the best fit line, you can fit a polynomial or curvilinear regression. And these are known as polynomial or curvilinear regression.

Analytics Vidhya

We're assuming that you've split the data into training and testing phase

```
# importing required libraries
from sklearn.linear_model import LinearRegression
import pandas as pd

# read the train and test dataset
train_data = pd.read_csv('train.csv')
test_data = pd.read_csv('test.csv')

# seperate the independent and target variable on training data
train_x = train_data.drop(columns=['Item_Outlet_Sales'],axis=1)
train_y = train_data['Item_Outlet_Sales']

# separate the independent and target variable on training data
test_x = test_data.drop(columns=['Item_Outlet_Sales'],axis=1)
test_y = test_data['Item_Outlet_Sales']
```

The dataset has been taken from https://datahack.analyticsvidhya.com

Analytics Vidhya

```
#Create the object of the Linear Regression model
model = LinearRegression()

# fit the model with the training data
model.fit(train_x,train_y)

# coefficients of the trained model
print('\nCoefficient of model :', model.coef_)

# intercept of the model
print('\nIntercept of model',model.intercept_)

# predict the target on the test dataset
predict_train = model.predict(train_x)
print('\nItem_Outlet_Sales on training data',predict_train)

# predict the target on the testing dataset
predict_test = model.predict(test_x)
print('\nItem_Outlet_Sales on test data',predict_test)
```

Analytics Vidhya

```
#Load Train and Test datasets
#Identify feature and response variable(s) and values must be numeric and numpy arrays
x_train <- input_variables_values_training_datasets

y_train <- target_variables_values_training_datasets

x_test <- input_variables_values_test_datasets

x <- cbind(x_train,y_train)

# Train the model using the training sets and check score
linear <- lm(y_train ~ ., data = x)

summary(linear)

#Predict Output
predicted= predict(linear,x_test)
```