Learning Objective

In this module, you will learn about linear regression and how to perform regression given a data.

Learning Objective:

1. Perform Regression Test

Regression Test

Regression analysis is a set of statistical processes for estimating the relationships between a dependent variable and independent variables.

`Dependent variable:`(DV) is what is impacted by the independent variable. To clarify, it reflects the change made by independent variable.

`Independent variable`: (IV) is what we expect will influence of variable. In other words, this represent the variable that changes on its own.In other words, it is independent of other variables in the study.

`Regression:` a statistical technique for finding the best-fitting straight line to describe the variable relationships for a set of data.

`Regression line:` the resulting best-fitting straight line.

`Regression equation:` the resulting best-fitting linear equation.

Regression

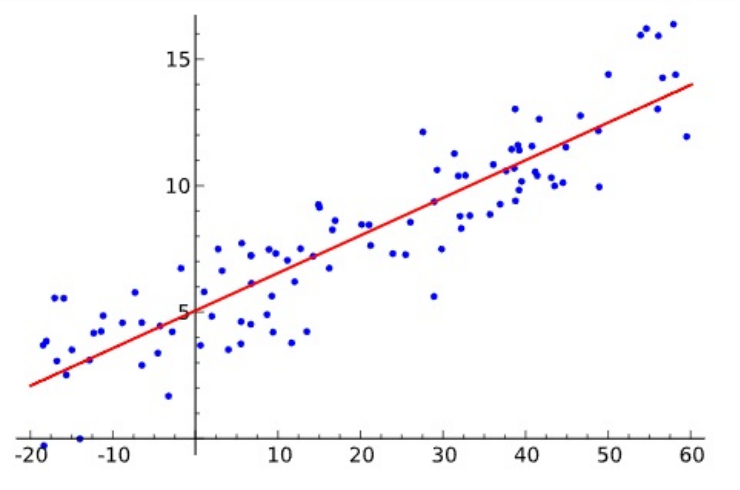
Linear Regression

Simple and Multiple Linear Regression

If you determine that your data sets are associated or correlated with each other, you can perform linear regression on them. Doing so enables you to **predict** additional data based on that regression model. There is **single linear** when you have one **independent** variable (x) and one **dependent** variable (y). And there is **multiple linear** regression when you have **multiple** independent variables (x's) but only one dependent variable (y). You can think of a *dependent* variable as an event that is **influenced** by the *independent* variable. For example, when comparing how much sunlight a plant receives versus its growth in length, it is probably better to say that the plant's growth is more likely dependent on the amount of sunlight than the reverse. Thus, the plant's length is the dependent variable and the amount of sunlight is the independent variable.

> When thinking of Linear regression:
> * The independent variable is denoted as x.
> * The dependent variable, is denoted as y.

Linear Regression finds out a linear relationship between x (input) and y(output). If we put all of our x values, the linear regression line will give us the line of best fit.



regre

Lets talk about the graph above. So the blue dots are our actual data points and the red light is the line of best fit for said graph. A linear regression model assumes that the relationship is linear. So if we are given an y we can use our linear regression model to predict the y. The difference is between our actual and the value on the line at that point is `error term`.

We will be using sklearn. To help us with the linear regression model in python.

For example, if you have the following data:

```
# create our data
month = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]
spend =[1000, 4000, 5000, 4500, 3000, 4000, 9000, 11000, 15000,
        12000, 7000, 3000]
sales =[9914, 40487, 54324, 50044, 34719, 42551, 94871, 118914,
        158484, 131348, 78504, 36284]


# the lists, with columns specified
df = pd.DataFrame(list(zip(month, spend,sales)),
               columns =['month', 'spend','sales'])
```

We are going to perform a simple linear regression where we have spend and sales as our IV and DV. Lets label them.

```
# Defining our independent and dependent variables
y=np.array(df['spend'])
x=np.array(df['sales']).reshape(-1, 1)
```

Now that we have our x and our y. We are going to create our model and then find the coefficient of determination. The coefficient of determination or `R-squared` examines how the difference of the two variables compare to each other. `R-square` is a value between `0.0` and `1.0`. Where `1.0` represent a perfect fit. The closer to 0.0 the value is the less reliable we can consider our model to be.

```


# create our regression model
model = LinearRegression().fit(x,y)
# get our r_squared
r_sq = model.score(x, y)
print('coefficient of determination:', r_sq)
```

Now that we know our coefficient of determination, is close to 1.0. We can make predictions,that will be somewhat close to the actual value. Lets test it out

Feel free to replace the 39000 with other number to check how close or accurate it is.

```
# Here can use the .predict and give it an x value to
print(model.predict(np.array([39000]).reshape(-1, 1)))
```

Logistic Regression

Logistic regression is performed when you have **categorical** data in the mix. For example, if you want to determine if sex is dependent on the height of person, you should use logistic regression.