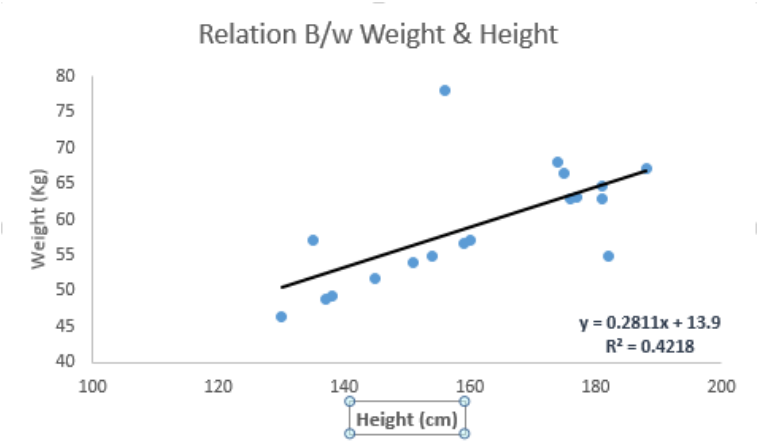
Module-3

1. Linear Regression

It is one of the most widely known modeling technique. Linear regression is usually among the first few topics which people pick while learning predictive modeling. In this technique, the dependent variable is continuous, independent variable(s) can be continuous or discrete, and nature of regression line is linear.

Linear Regression establishes a relationship between **dependent variable (Y)** and one or more **independent variables (X)** using a **best fit straight line** (also known as regression line).

It is represented by an equation **Y=a+b*X + e**, where a is intercept, b is slope of the line and e is error term. This equation can be used to predict the value of target variable based on given predictor variable(s).



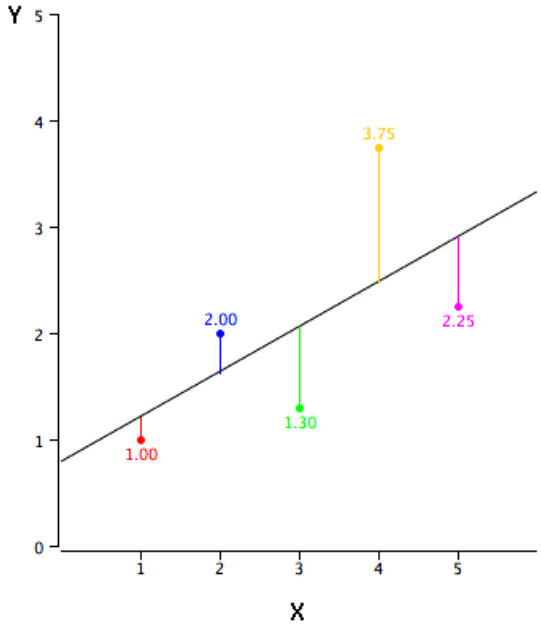
The difference between simple linear regression and multiple linear regression is that, multiple linear regression has (>1) independent variables, whereas simple linear regression has only 1 independent variable. Now, the question is "How do we obtain best fit line?".

How to obtain best fit line (Value of a and b)?

This task can be easily accomplished by Least Square Method. It is the most common method used for fitting a regression line. It calculates the best-fit line for the observed data by

minimizing the sum of the squares of the vertical deviations from each data point to the line. Because the deviations are first squared, when added, there is no cancelling out between positive and negative values.

$$\min_{w} ||Xw - y||_2^2$$



We can evaluate the model performance using the metric **R-square**. To know more details about these metrics, you can read: Model Performance metrics Part 1, Part 2 .

Important Points:

- There must be **linear relationship** between independent and dependent variables
- Multiple regression suffers from **multicollinearity, autocorrelation, heteroskedasticity**.
- Linear Regression is very sensitive to **Outliers**. It can terribly affect the regression line and eventually the forecasted values.
- Multicollinearity can increase the variance of the coefficient estimates and make the estimates very sensitive to minor changes in the model. The result is that the coefficient estimates are unstable

- In case of multiple independent variables, we can go with **forward selection**, **backward elimination** and **step wise approach** for selection of most significant independent variables.

Linear Regression Formula

Linear regression is the most basic and commonly used predictive analysis. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.

There are several linear regression analyses available to the researcher.

Simple linear regression

- One dependent variable (interval or ratio)
- One independent variable (interval or ratio or dichotomous)

Multiple linear regression

- One dependent variable (interval or ratio)
- Two or more independent variables (interval or ratio or dichotomous)

Logistic regression

- One dependent variable (binary)
- Two or more independent variable(s) (interval or ratio or dichotomous)

Ordinal regression

- One dependent variable (ordinal)
- One or more independent variable(s) (nominal or dichotomous)

Multinomial regression

- One dependent variable (nominal)
- One or more independent variable(s) (interval or ratio or dichotomous)

Discriminant analysis

- One dependent variable (nominal)
- One or more independent variable(s) (interval or ratio)

Formula for linear regression equation is given by:

$$y = a + bx$$

a and *b* are given by the following formulas:

$$a\,(intercept) = \frac{\sum y \sum x^2 - \sum x \sum xy}{(\sum x^2) - (\sum x)^2}$$

$$b\,(slope) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

Where,
x and *y* are two variables on the regression line.
b = Slope of the line.
a = *y*-intercept of the line.
x = Values of the first data set.
y = Values of the second data set.

Solved Examples

Question: Find linear regression equation for the following two sets of data:

x	2	4	6	8
y	3	7	5	10

Solution:

Construct the following table:

x	y	x^2	xy
2	3	4	6
4	7	16	28
6	5	36	30
8	10	64	80
$\sum x$ = 20	$\sum y$ = 25	$\sum x^2$ = 120	$\sum xy$ = 144

$$b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$b = \frac{4\times144 - 20\times25}{4\times120 - 400}$$

b = 0.95

$$a = \frac{\sum y \sum x^2 - \sum x \sum xy}{n(\sum x^2) - (\sum x)^2}$$

$$a = \frac{25\times120 - 20\times144}{4(120) - 400}$$

a = 1.5

Linear regression is given by:

y = a + bx

y = 1.5 + 0.95 x

Steps in building Regression Model:

1. Identify problem statement
2. Collect the data set
3. Prepare the data
4. Identify the relationship between variables(dependent and independent0
5. Split training and test data
6. Model building
7. Performance evaluation
8. Present the findings

Model diagnostics

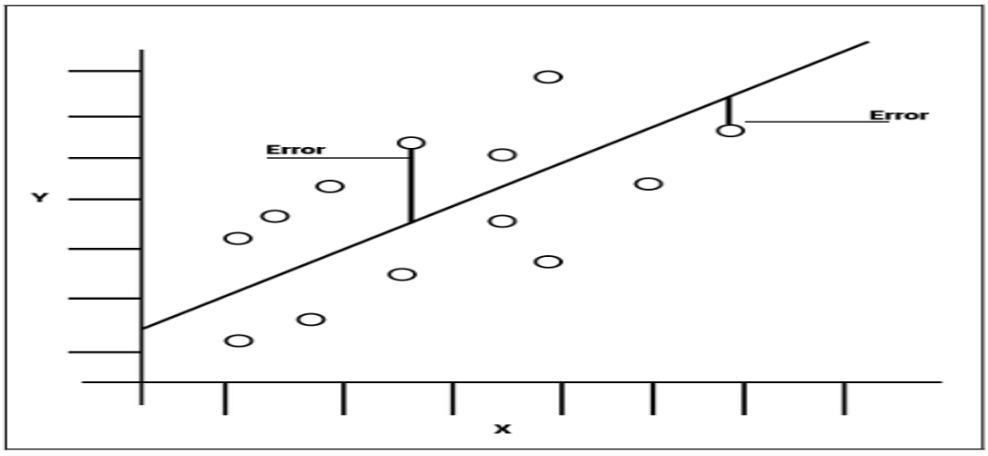
After building a machine learning model, the next step is to evaluate the model performance and understand how good our model is against a benchmark model. The evaluation metric to be used would depend upon the type of problem you are trying to solve —whether it is a supervised or unsupervised problem, and if it is a classification or a regression task.

Diagnostics are used to evaluate the model assumptions and figure out whether or not there are observations with a large, undue influence on the analysis. They can be used to optimize the model by making sure the model you use is actually appropriate for the data you are analyzing. There are many ways to assess the validity of a model using diagnostics. *Diagnostics* is an overarching name that covers the other topics under model assumptions. It may include exploring the model's basic statistical assumptions, examining the structure of a model by considering more, fewer, or different explanatory variables, or looking for data that is poorly represented by a model such as outliers or that have a large imbalanced effect on the regression model's prediction.

Some of the diagnostics metrics for evaluating the model

The Mean Squared Error measures how close a regression line is to a set of data points. It is a risk function corresponding to the expected value of the squared error loss.

Mean square error is calculated by taking the average, specifically the mean, of errors squared from data as it relates to a function.



A larger MSE indicates that the data points are dispersed widely around its central moment (mean), whereas a smaller MSE suggests the opposite. A smaller MSE is preferred because it indicates that your data points are dispersed closely around its central moment (mean). It reflects the centralized distribution of your data values, the fact that it is not skewed, and, most importantly, it has fewer errors (errors measured by the dispersion of the data points from its mean).

Lesser the MSE => Smaller is the error => Better the estimator.

The Mean Squared Error is calculated as:

$MSE = (1/n) * \Sigma(actual - forecast)2$

where:

- Σ – a symbol that means “sum”
- n – sample size
- actual – the actual data value
- forecast – the predicted data value

Calculate Mean Square Error Using Excel

Now, you will learn how you can calculate the MSE using Excel.

Suppose you have the sales data of a product of all the months.

Step 1: Enter the actual and forecasted data into two separate columns.

Month	Actual	Forecasted
January	67	70
February	50	44
March	36	38
April	74	44
May	84	64
June	84	80
July	64	54
August	34	44
September	23	43
October	72	90
November	62	56
December	42	38

Step 2: Calculate the squared error of each data

The squared error is calculated by (actual – forecast)2

Month	Actual	Forecasted	Squared Error
January	67	70	9
February	50	49	1
March	36	38	4
April	74	76	4
May	84	83	1
June	84	80	16
July	64	67	9
August	34	30	16
September	23	20	9
October	72	75	9
November	62	60	4
December	42	38	16

Step 3: Calculate the Mean Squared Error

Month	Actual	Forecasted	Squared Error	
January	67	70	9	
February	50	49	1	
March	36	38	4	
April	74	76	4	
May	84	83	1	
June	84	80	16	
July	64	67	9	
August	34	30	16	
September	23	20	9	
October	72	75	9	
November	62	60	4	
December	42	38	16	
			8.166666667	**MSE**

MSE = (1/12) * (98) = 8.166

The MSE for this model is 8.17.

Root Mean Squared Error (RMSE)and Mean Absolute Error (MAE) are metrics used to evaluate a Regression Model. These metrics tell us how accurate our predictions are and, what is the amount of deviation from the actual values.

Technically, RMSE is the **R**oot of the **M**ean of the **S**quare of **E**rrors and MAE is the **M**ean of **A**bsolute value of **E**rrors. Here, errors are the differences between the predicted values (values predicted by our regression model) and the actual values of a variable. They are calculated as follows :

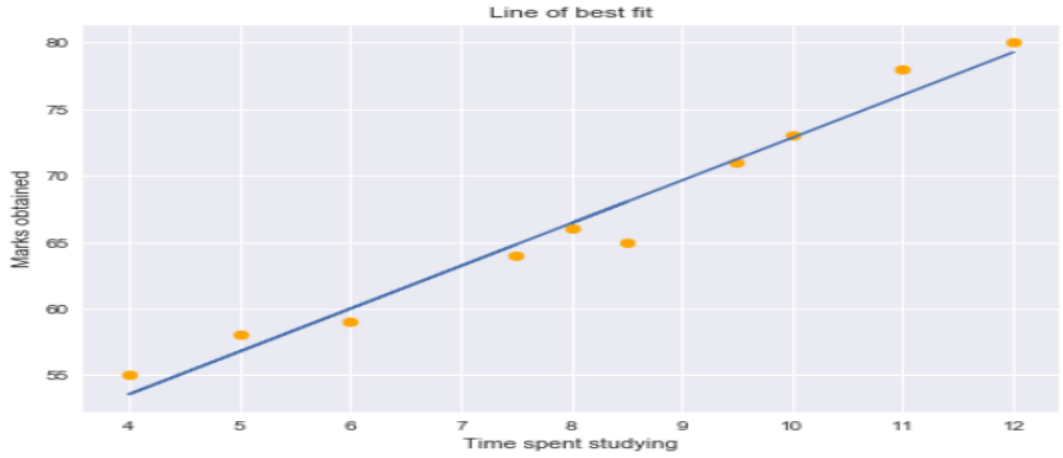
$$RMSE = \sqrt{\frac{\sum (y_i - y_p)^2}{n}}$$

$$MAE = \frac{|(y_i - y_p)|}{n}$$

y_i = actual value
y_p = predicted value
n = number of observations/rows

Residual Sum of Squares(R^2)

To understand the concepts clearly, we are going to take up a simple regression problem. Here, we are trying to predict the 'Marks Obtained' based on the amount of 'Time Spent Studying'. The **time** spent studying will be our **independent variable** and the **marks achieved** in the test is our **dependent** or **target variable**.



Residual plots tell us whether the regression model is the right fit for the data or not. It is actually an assumption of the regression model that there is no trend in residual plots. To study the assumptions of linear regression in detail, I suggest going through this great article!

Using the residual values, we can determine the sum of squares of the residuals also known as **Residual sum of squares** or RSS.

$$RSS = \sum (y_i - \hat{y}_i)^2$$

The lower the value of RSS, the better is the model predictions. Or we can say that – a regression line is a line of best fit if it minimizes the RSS value. But there is a flaw in this – RSS is a scale variant statistic. Since RSS is the sum of the squared difference between the actual and predicted value, the value depends on the

scale of the target variable.

Example:

Consider your target variable is the revenue generated by selling a product. The residuals would depend on the scale of this target. If the revenue scale was taken in "Hundreds of Rupees" (i.e. target would be 1, 2, 3, etc.) then we might get an RSS of about 0.54 (hypothetically speaking).

But if the revenue target variable was taken in "Rupees" (i.e. target would be 100, 200, 300, etc.), then we might get a larger RSS as 5400. Even though the data does not change, the value of RSS varies according to the scale of the target. This makes it difficult to judge what might be a good RSS value.

So, can we come up with a better statistic that is scale-invariant? This is where R-squared comes into the picture.

Understanding R-squared statistic

R-squared statistic or coefficient of determination is a scale invariant statistic that gives the proportion of variation in target variable explained by the linear regression model.

Overview

- Understand the concept of R-squared and Adjusted R-Squared
- Get to know the key differences between R-Squared and Adjusted R-squared

Introduction

When I started my journey in Data Science, the first algorithm that I explored was Linear Regression. After understanding the concepts of Linear Regression and how the algorithm works, I was really excited to use it and make predictions on a problem statement. I am sure most of you would have done the same. But once we have predicted the values, what is next?

Then comes the tricky part. Once we have built our model, the next step was to evaluate its performance. Needless to say, the task of model evaluation is a pivotal one and highlights the shortcomings of our model. Choosing the most appropriate Evaluation Metric is a crucial task. And, I came across two important metrics: R-squared and Adjusted R-squared apart from MAE/ MSE/ RMSE. What is the difference between these two? Which one should I use?

R-squared and Adjusted R-squared are two such evaluation metrics that might seem confusing to any data science aspirant initially. Since they both are extremely important to evaluate regression problems, we are going to understand and compare them in-depth. They both have their pros and cons which we will be discussing in detail in this article.

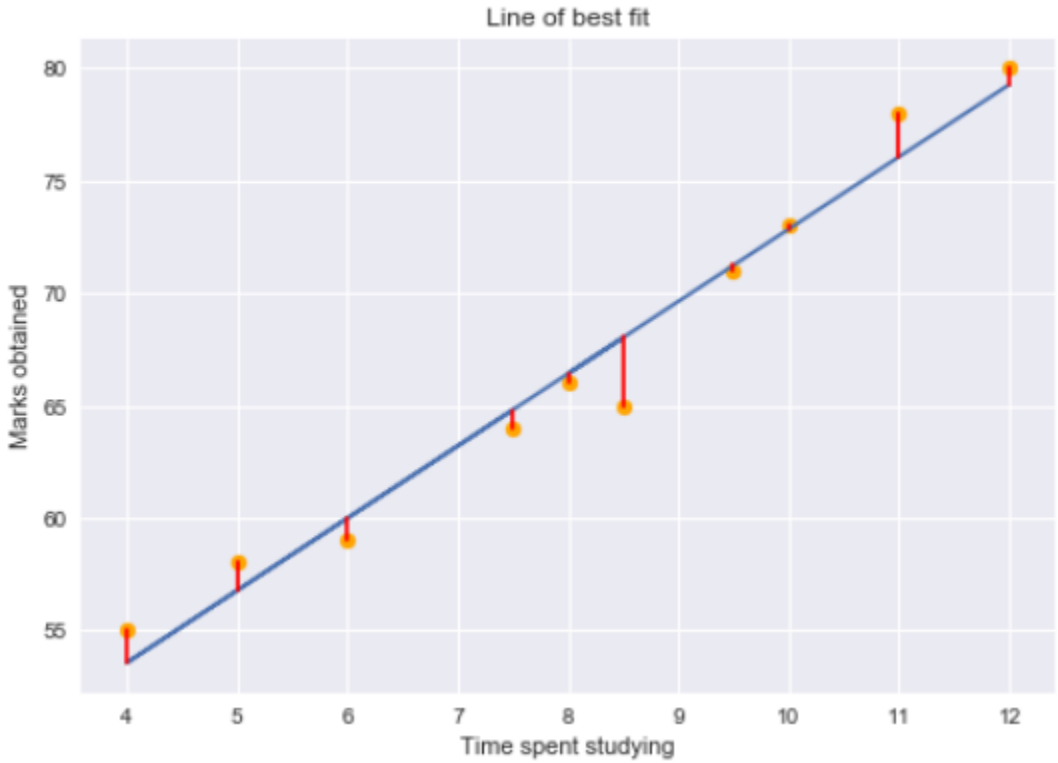
Note: To understand R-Squared and Adjusted R-Squared, you must have a good understanding of Linear Regression. Please refer to our free course –

Residual Sum of Squares

To understand the concepts clearly, we are going to take up a simple regression problem. Here, we are trying to predict the 'Marks Obtained' based on the amount of 'Time Spent Studying'. The **time** spent studying will be our **independent variable** and the **marks achieved** in the test is our **dependent** or **target variable**.

Residual for a point in the data is the difference between the actual value and the value predicted by our linear regression model.

$$Residual = actual - predicted = y\_\hat{y}$$



Residual plots tell us whether the regression model is the right fit for the data or not. It is actually an assumption of the regression model that there is no trend in residual plots. To study the assumptions of linear regression in detail, I suggest going through this great article!

Using the residual values, we can determine the sum of squares of the residuals also known as **Residual sum of squares** or RSS.

$$RSS = \sum (y_i - \hat{y}_i)^2$$

The lower the value of RSS, the better is the model predictions. Or we can say that – a regression line is a line of best fit if it minimizes the RSS value. But there is a flaw in this – RSS is a scale variant statistic. Since RSS is the sum of the squared difference between the actual and predicted value, the value depends on the scale of the target variable.

Example:

Consider your target variable is the revenue generated by selling a product. The residuals would depend on the scale of this target. If the revenue scale was taken in "Hundreds of Rupees" (i.e. target would be 1, 2, 3, etc.) then we might get an RSS of about 0.54 (hypothetically speaking).

But if the revenue target variable was taken in "Rupees" (i.e. target would be 100, 200, 300, etc.), then we might get a larger RSS as 5400. Even though the data does not change, the value of RSS varies according to the scale of the target. This makes it difficult to judge what might be a good RSS value.

So, can we come up with a better statistic that is scale-invariant? This is where R-squared comes into the picture.

Understanding R-squared statistic

R-squared statistic or coefficient of determination is a scale invariant statistic that gives the proportion of variation in target variable explained by the linear regression model.

This might seem a little complicated, so let me break this down here. In order to determine the proportion of target variation explained by the model, we need to first determine the following-

1. **Total Sum of Squares**

   Total variation in target variable is the sum of squares of the difference between the actual values and their mean.

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

1. TSS or Total sum of squares gives the total variation in Y. We can see that it is very similar to the variance of Y. While the variance is the average of the squared sums of difference between actual values and data points, TSS is the total of the squared sums.

   Now that we know the total variation in the target variable, how do we determine the proportion of this variation explained by our model? We go back to RSS.

2. **Residual Sum of Squares**

   As we discussed before, RSS gives us the total square of the distance of actual points from the regression line. But if we focus on a single residual, we can say that it is the distance that is not captured by the regression line. Therefore, RSS as a whole gives us the variation in the target variable that is **not explained** by our model.

1. **Calculate R-Squared**

Now, if TSS gives us the total variation in Y, and RSS gives us the variation in Y not explained by X, then **TSS-RSS gives us the variation in Y that is explained by our model**! We can simply divide this value by TSS to get the proportion of variation in Y that is explained by the model. And this our **R-squared statistic**!

R-squared = (TSS-RSS)/TSS
= Explained variation/ Total variation
= 1 – Unexplained variation/ Total variation

So R-squared gives the degree of variability in the target variable that is explained by the model or the independent variables. If this value is 0.7, then it means that the independent variables explain 70% of the variation in the target variable.

R-squared value always lies between 0 and 1. A higher R-squared value indicates a higher amount of variability being explained by our model and vice-versa.

If we had a really low RSS value, it would mean that the regression line was very close to the actual points. This means the independent variables explain the majority of variation in the target variable. In such a case, we would have a really high R-squared value.

$$\uparrow \text{R-squared} = 1 - \frac{\text{RSS}}{\text{TSS}} \downarrow$$

What is Adjusted R-squared?

Adjusted R-squared is a statistical measure used to evaluate the goodness of fit of a regression model. It provides insights into how well the model explains the variability in the data.

Unlike the standard R-squared, which simply tells you the proportion of variance explained by the model, Adjusted R-squared takes into account the number of predictors (independent variables) in the model.

The advantage of Adjusted R-squared is that it penalizes the inclusion of unnecessary variables. This means that as you add more predictors to the model, the Adjusted R-squared value will only increase if the new variables significantly improve the model's performance.

Let's have a look at the formula for adjusted R-squared to better understand its working.

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

Here,

- **n** represents the number of data points in our dataset
    - **k** represents the number of independent variables, and
    - **R** represents the R-squared values determined by the model.

So, if R-squared does not increase significantly on the addition of a new independent variable, then the value of Adjusted R-squared will actually decrease.

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

Multiple Linear Regression by Hand (Step-by-Step)

Multiple linear regression is a method we can use to quantify the relationship between two or more predictor variables and a response variable.

Y=B0+B1X1+B2X2

X1,X2-Independent variables

Y-target variable

B11,B2-regression coefficients

This tutorial explains how to perform multiple linear regression by hand.

Example: Multiple Linear Regression by Hand

Suppose we have the following dataset with one response variable *y* and two predictor variables X_1 and X_2:

y	X_1	X_2
140	60	22
155	62	25
159	67	24
179	70	20
192	71	15
200	72	14
212	75	14
215	78	11

Use the following steps to fit a multiple linear regression model to this dataset.

Step 1: Calculate X_1^2, X_2^2, X_1y, X_2y and X_1X_2.

	y	X_1	X_2		X_1^2	X_2^2	X_1y	X_2y	X_1X_2
	140	60	22		3600	484	8400	3080	1320
	155	62	25		3844	625	9610	3875	1550
	159	67	24		4489	576	10653	3816	1608
	179	70	20		4900	400	12530	3580	1400
	192	71	15		5041	225	13632	2880	1065
	200	72	14		5184	196	14400	2800	1008
	212	75	14		5625	196	15900	2968	1050
	215	78	11		6084	121	16770	2365	858
Mean	181.5	69.375	18.125	**Sum**	38767	2823	101895	25364	9859
Sum	1452	555	145						

Step 2: Calculate Regression Sums.

Next, make the following regression sum calculations:

- $\Sigma x_1^2 = \Sigma X_1^2 - (\Sigma X_1)^2 / n = 38{,}767 - (555)^2 / 8 =$ **263.875**
- $\Sigma x_2^2 = \Sigma X_2^2 - (\Sigma X_2)^2 / n = 2{,}823 - (145)^2 / 8 =$ **194.875**
- $\Sigma x_1 y = \Sigma X_1 y - (\Sigma X_1 \Sigma y) / n = 101{,}895 - (555*1{,}452) / 8 =$ **1,162.5**
- $\Sigma x_2 y = \Sigma X_2 y - (\Sigma X_2 \Sigma y) / n = 25{,}364 - (145*1{,}452) / 8 =$ **-953.5**
- $\Sigma x_1 x_2 = \Sigma X_1 X_2 - (\Sigma X_1 \Sigma X_2) / n = 9{,}859 - (555*145) / 8 =$ **-200.375**

	y	X_1	X_2		X_1^2	X_2^2	X_1y	X_2y	X_1X_2
	140	60	22		3600	484	8400	3080	1320
	155	62	25		3844	625	9610	3875	1550
	159	67	24		4489	576	10653	3816	1608
	179	70	20		4900	400	12530	3580	1400
	192	71	15		5041	225	13632	2880	1065
	200	72	14		5184	196	14400	2800	1008
	212	75	14		5625	196	15900	2968	1050
	215	78	11		6084	121	16770	2365	858
Mean	181.5	69.375	18.125	**Sum**	38767	2823	101895	25364	9859
Sum	1452	555	145						
				Reg Sums	263.875	194.875	1162.5	-953.5	-200.375

Step 3: Calculate b_0, b_1, and b_2.

The formula to calculate b_1 is: $[(\Sigma x_2^2)(\Sigma x_1 y) - (\Sigma x_1 x_2)(\Sigma x_2 y)] / [(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2]$

Thus, $\mathbf{b_1} = [(194.875)(1162.5) - (-200.375)(-953.5)] / [(263.875)(194.875) - (-200.375)^2]$
$= \mathbf{3.148}$

The formula to calculate b_2 is: $[(\Sigma x_1^2)(\Sigma x_2 y) - (\Sigma x_1 x_2)(\Sigma x_1 y)] / [(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2]$

Thus, $\mathbf{b_2} = [(263.875)(-953.5) - (-200.375)(1152.5)] / [(263.875)(194.875) - (-200.375)^2] = \mathbf{-1.656}$

The formula to calculate b_0 is: $y - b_1X_1 - b_2X_2$

Thus, $\mathbf{b_0} = 181.5 - 3.148(69.375) - (-1.656)(18.125) = \mathbf{-6.867}$

Step 5: Place b_0, b_1, and b_2 in the estimated linear regression equation.

The estimated linear regression equation is: $\hat{y} = b_0 + b_1 * x_1 + b_2 * x_2$

In our example, it is $\mathbf{\hat{y} = -6.867 + 3.148x_1 - 1.656x_2}$

How to Interpret a Multiple Linear Regression Equation

Here is how to interpret this estimated linear regression equation: $\hat{y} = -6.867 + 3.148x_1 – 1.656x_2$

b_0 = -6.867. When both predictor variables are equal to zero, the mean value for y is -6.867.

b_1 = 3.148. A one unit increase in x_1 is associated with a 3.148 unit increase in y, on average, assuming x_2 is held constant.

b_2 = -1.656. A one unit increase in x_2 is associated with a 1.656 unit decrease in y, on average, assuming x_1 is held constant.

Colinearity:

What Is Multicollinearity?

Multicollinearity is a statistical phenomenon that occurs when two or more independent variables in a regression model are highly correlated with each other. In other words, multicollinearity indicates a strong linear relationship among the predictor variables. This can create challenges in the regression analysis because it becomes difficult to determine the individual effects of each independent variable on the dependent variable accurately.

This means that one independent variable can be predicted from another in a regression model. For example, sets like height and weight, household income and water consumption, mileage and the price of a car, study time and leisure time, etc.

Let me take a simple example from our everyday life to explain this. Colin loves watching television while munching on chips. The more television he watches, the more chips he eats, and the happier he gets!

- Multicollinearity is a statistical term that describes the correlation between multiple independent variables in a model.
- When two variables have a correlation coefficient of either +1.0 or -1.0, they are considered perfectly collinear.

- The presence of multicollinearity among independent variables can lead to less dependable statistical conclusions.

The Problem With Having Multicollinearity

Multicollinearity can be a problem in a regression model when using algorithms such as OLS (ordinary least squares) in statsmodels. This is because the estimated regression coefficients become unstable and difficult to interpret in the presence of multicollinearity. Statsmodels is a Python library that provides a range of tools for statistical analysis, including regression analysis.

When multicollinearity is present, the estimated regression coefficients may become large and unpredictable, leading to unreliable inferences about the effects of the predictor variables on the response variable. Therefore, it is important to check for multicollinearity and consider using other regression techniques that can handle this problem, such as ridge regression or principal component regression. For example, let's assume that in the following linear equation:

$Y = W0+W1*X1+W2*X2$

Coefficient W1 is the increase in Y for a unit increase in X1 while keeping X2 constant. But since X1 and X2 are highly correlated, changes in X1 would also cause changes in X2, and we would not be able to see their individual effect on Y.

The regression coefficient, also known as the beta coefficient, measures the strength and direction of the relationship between a predictor variable (X) and the response variable (Y). In the presence of multicollinearity, the regression coefficients become unstable and difficult to interpret because the variance of the coefficients becomes large. This results in wide confidence intervals and increased variability in the predicted values of Y for a given value of X. As a result, it becomes challenging to determine the individual contribution of each predictor variable to the response variable and make reliable inferences about their effects on Y.

> " This makes the effects of X1 on Y difficult to distinguish from the effects of X2 on Y. "

Multicollinearity may not affect the accuracy of the machine-learning model as much. But we might lose reliability in determining the effects of individual features in your model – and that can be a problem when it comes to interpretability.

What Causes Multicollinearity?

Multicollinearity could occur due to the following problems:

- Multicollinearity could exist because of the problems in the dataset at the time of creation. These problems could be because of poorly designed experiments, highly observational data, or the inability to manipulate the data.
  For example, determining the electricity consumption of a household from the household income and the number of electrical appliances. Here, we know that the number of electrical appliances in a household will increase with household income. However, this cannot be removed from the dataset.
- Multicollinearity could also occur when new variables are created which are dependent on other variables.
  For example, creating a variable for BMI from the height and weight variables would include redundant information in the model, and the new variable will be a highly correlated variable.
- Including identical variables in the dataset.
  For example, including variables for temperature in Fahrenheit and temperature in Celsius.
- Inaccurate use of dummy variables can also cause a multicollinearity problem. This is called the Dummy variable trap.
  For example, in a dataset containing the status of marriage variable with two unique values: ‘married’, and ’single’. Creating dummy variables for both of them would include redundant information. We can make do with only one variable containing 0/1 for ‘married’/’single’ status.
- Insufficient data, in some cases, can also cause multicollinearity problems.

Detecting Multicollinearity Using a Variance Inflation Factor (VIF)

Let’s try detecting multicollinearity in a dataset to give you a flavor of what can go wrong.

I have created a dataset determining the salary of a person in a company based on the following features:

- Gender (0 – female, 1- male)
- Age
- Years of service (Years spent working in the company)
- Education level (0 – no formal education, 1 – under-graduation, 2 – post-graduation)
    - VIF determines the strength of the correlation between the independent variables. It is predicted by taking a variable and regressing it against every other variable. “
    - or
    - VIF score of an independent variable represents how well the variable is explained by other independent variables.
    - R^2 value is determined to find out how well an independent variable is described by the other independent variables. A high value of R^2 means that the variable is highly correlated with the other variables. This is captured by the VIF, which is denoted below:
    - $$VIF = \frac{1}{1-R^2}$$
    - So, the closer the R^2 value to 1, the higher the value of VIF and the higher the multicollinearity with the particular independent variable.

Residual analysis:

Residual analysis is used to assess the appropriateness of a linear regression model by defining residuals and examining the residual plot graphs.

Residual

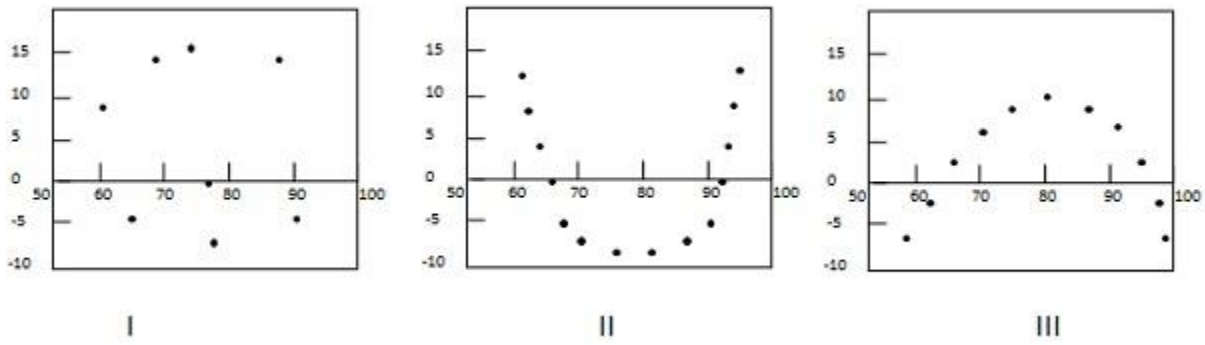
Residual(e$_i$) refers to the difference between observed value(y) vs predicted value (y^). Every data point has one residual.
residual=observedValue−predictedValuee=y−y^

Residual Plot

A residual plot is a graph in which residuals are on tthe vertical axis and the independent variable is on the horizontal axis. If the dots are randomly dispersed around the horizontal axis then a linear regression model is appropriate for the data; otherwise, choose a non-linear model.

Types of Residual Plot

Following example shows few patterns in residual plots.



In first case, dots are randomly dispersed. So linear regression model is preferred. In Second and third case, dots are non-randomly dispersed and suggests that a non-linear regression method is preferred.

Example

Problem Statement:

Check where a linear regression model is appropriate for the following data.

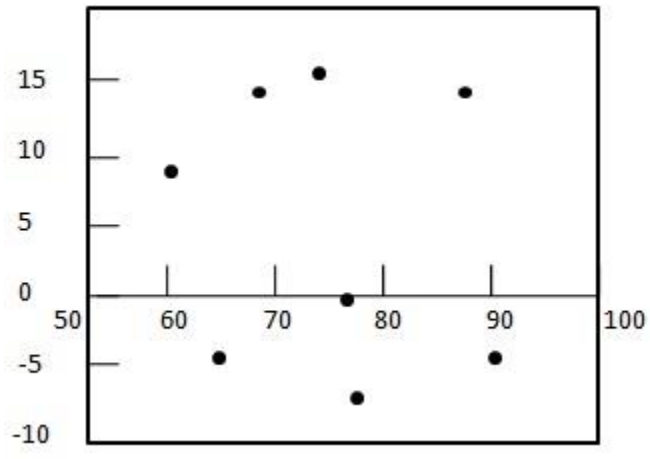
x****	60	70	80	85	95
y**** **(Actual Value)**	70	65	70	95	85
y^****^ **(Predicted Value)**	65.411	71.849	78.288	81.507	87.945

Solution:

Step 1: Compute residuals for each data point.

x̄	60	70	80	85	95
ȳ **(Actual Value)**	70	65	70	95	85
ŷ **(Predicted Value)**	65.411	71.849	78.288	81.507	87.945
e **(Residual)**	4.589	-6.849	-8.288	13.493	-2.945

Step 2: - Draw the residual plot graph.



Step 3: - Check the randomness of the residuals.

Here residual plot exibits a random pattern - First residual is positive, following two are negative, the fourth one is positive, and the last residual is negative. As pattern is quite random which indicates that a linear regression model is appropriate for the above data.

Detecting influencers

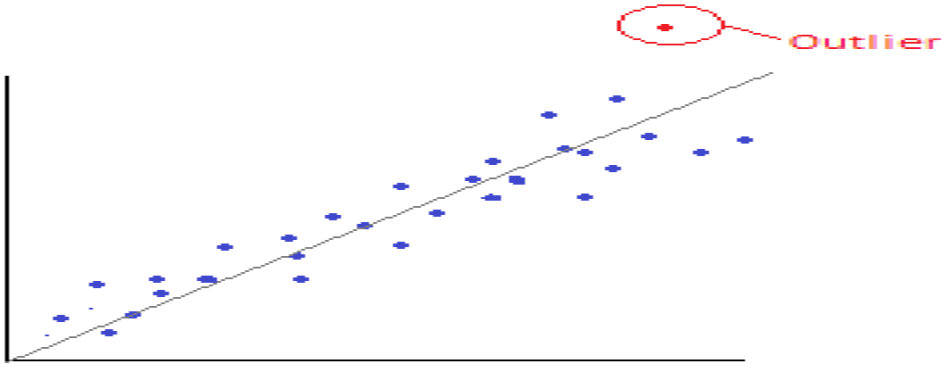
It is easy to mistake these points with “outliers”, however, they have different definitions. Not all outliers are considered influential points. In fact, in some cases, the presence of outliers, although unusual, may not change the regression line.

For example, if you have the data points: (1,1), (5,5), (6,6), (3,3), and (500,500), you can consider the last point as an outlier, but the regression life remains unchanged.

Instead, we should breakdown these extreme values into *extreme y-values (high residuals/outliers)* and *extreme x-values (high leverage*). In some cases, the observation may have both high residuals and high leverage.

OUTLIERS

An outlier is an observation with extreme y-values. Because the extreme values occur in the dependent or target variable, these observations have high residuals.



LEVERAGE

Leverage is a measure of how far the value of a predictor variable (e.g. independent or usually the x-variable) from the mean of that variable.

We now look at how to detect potential outliers that have an undue influence on the multiple regression model. Keep in mind that since we are dealing with a multi-dimensional model, there may be data points that look perfectly fine in any single dimension but are multivariate outliers. E.g. for the general population, there is nothing unusual about a 6-foot man or a 125-pound man, but a 6-foot man that weighs 125 pounds is unusual.

Leverage

Definition 1: The following parameters are indicators that a sample point ($x_{i1}, \ldots, x_{ik}, y_i$) is an outlier:

Distance – the residual

$e_i = y_i - \hat{y}_i$

is the measure of the distance of the *i*th sample point from the regression line. Points with large residuals are potential outliers.

Cook's distance and DFFITS

Definition 2: If we remove a point from the sample, then the equation for the regression line changes. Points that have the most **influence** produce the largest change in the equation of the regression line. A measure of this influence is called **Cook's distance**. For the *i*th point in the sample, Cook's distance is defined as

$$D_i = \frac{\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_{j(i)})^2}{(k+1)MS_E}$$

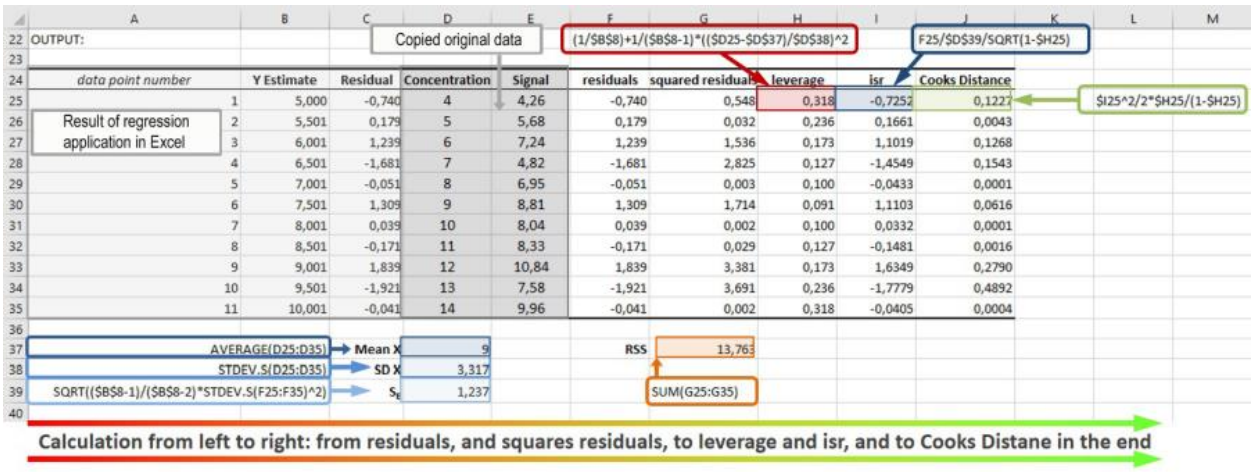
where $\hat{y}_{j(i)}$ is the prediction of y_j by the revised regression model when the point ($x, \ldots, x_{ik}, y_i$) is removed from the sample.

Another measure of influence is **DFFITS**, which is defined by the formula

$$\text{DFFITS}_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\sqrt{MS_{E(i)}h_i}}$$

Whereas Cook's distance is a measure of the change in the mean vector when the *i*th point is removed, DFFITS is a measure of the change in the *i*th mean when the *i*th point is removed.

These metrics calculations can be done in Excel also



data point number	Y Estimate	Residual	Concentration	Signal	residuals	squared residuals	leverage	isr	Cooks Distance
1	5,000	-0,740	4	4,26	-0,740	0,548	0,318	-0,7252	0,1227
2	5,501	0,179	5	5,68	0,179	0,032	0,236	0,1661	0,0043
3	6,001	1,239	6	7,24	1,239	1,536	0,173	1,1019	0,1268
4	6,501	-1,681	7	4,82	-1,681	2,825	0,127	-1,4549	0,1543
5	7,001	-0,051	8	6,95	-0,051	0,003	0,100	-0,0433	0,0001
6	7,501	1,309	9	8,81	1,309	1,714	0,091	1,1103	0,0616
7	8,001	0,039	10	8,04	0,039	0,002	0,100	0,0332	0,0001
8	8,501	-0,171	11	8,33	-0,171	0,029	0,127	-0,1481	0,0016
9	9,001	1,839	12	10,84	1,839	3,381	0,173	1,6349	0,2790
10	9,501	-1,921	13	7,58	-1,921	3,691	0,236	-1,7779	0,4892
11	10,001	-0,041	14	9,96	-0,041	0,002	0,318	-0,0405	0,0004

Question Bank:

1.Explain Linear Regression in detail.

2.Explain the steps in Building a Regression Model.

3.Explain Model Diagonstics.

4.Explain Multiple Linear Regression.

5.Explain the process of Developing Multiple Linear regression

6.Explain Residual analysis.

7.Explain Multi collinearity.

8.Explain Detecting Influencers in Linear Regression.

9. What Causes Multicollinearity.

10.What is adjusted R squared?

11. Examine whether the Linear Regression is appropriate for the following dataset.

X={10,20,30,40,50}

Y={12,13,14,15,16}

Ypred={13.33, 15.55 , 16.66 , 17.88 ,19.99}

12. Illustrate R^2 Score, MSE, MAE for the following dataset.

X={1,2,3,4}

Y={10,20,30,40}

$Y_{pred=}$ {11.7, 22.3 ,28.9 ,35.6}

13. Explain the mathematical implementation behind the Linear Regression Analysis with the following dataset.{x=(1,2,3),y=(3,4,2)}

14. Classify the various steps in Building a Regression Model with a suitable python code.

15. Discuss Regression Analysis with a suitable example.