Data Science

Relation Analysis

Relationship Analysis

Example: Wage Data

A large data regarding the wages for a group of employees from the eastern region of India is given.

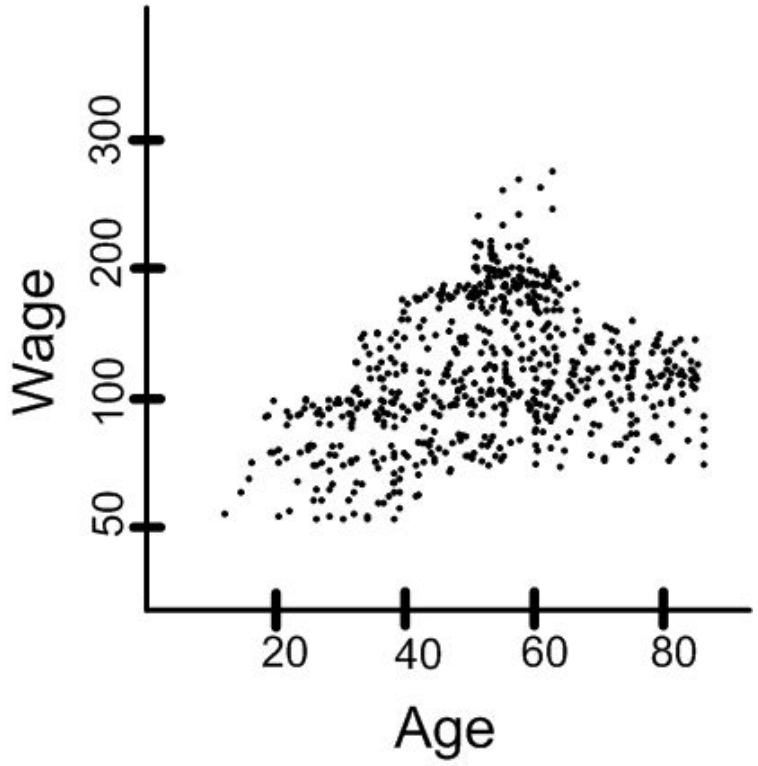
In particular, we wish to understand the following relationships:

- *Employee's age and wage:* How wages vary with ages?
- *Calendar year and wage:* How wages vary with time?
- *Employee's age and education:* Whether wages are anyway related with employees' education levels?

Relationship Analysis

Example: Wage Data

- Case I. Wage versus Age
  - From the data set, we have a graphical representations, which is as follows:



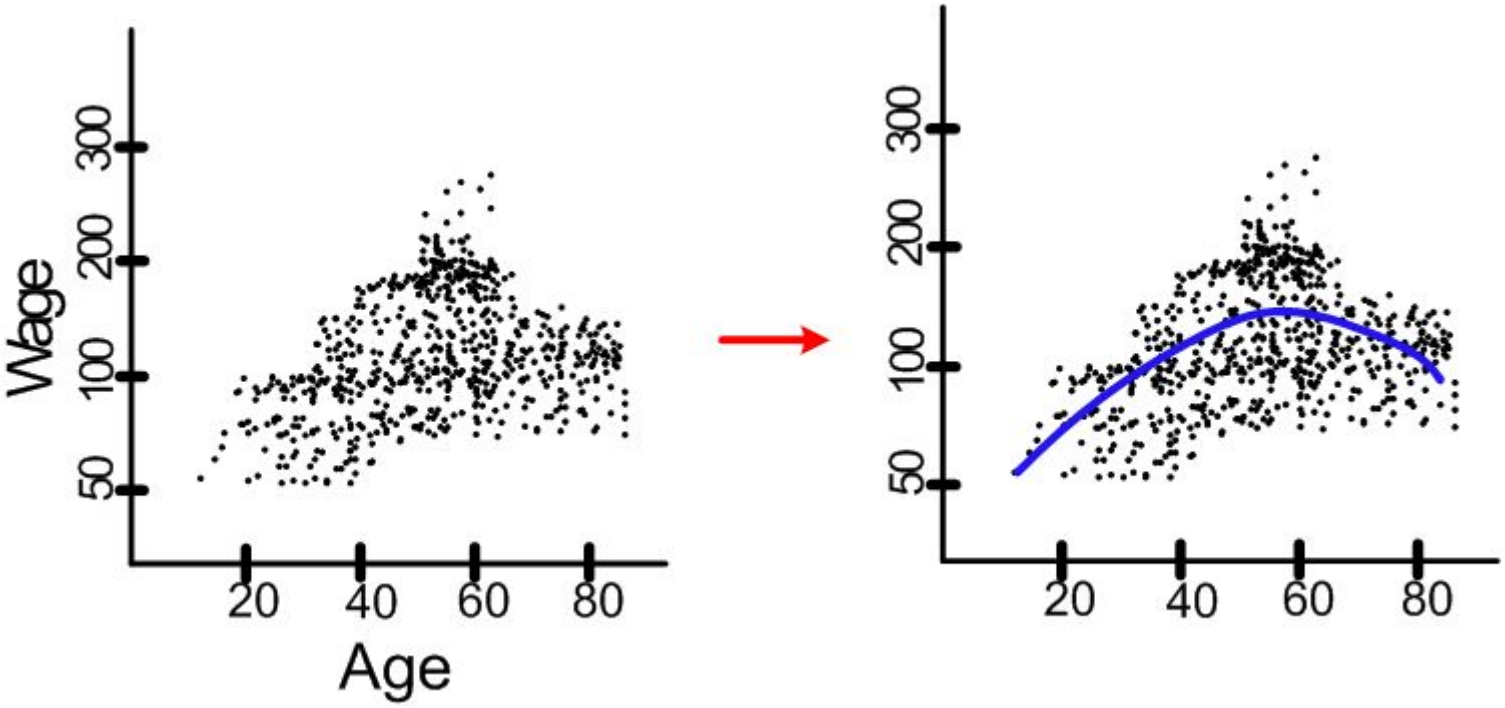
?

How wages vary with ages?

Relationship Analysis

Example: Wage Data

- *Employee's age and wage:* How wages vary with ages?

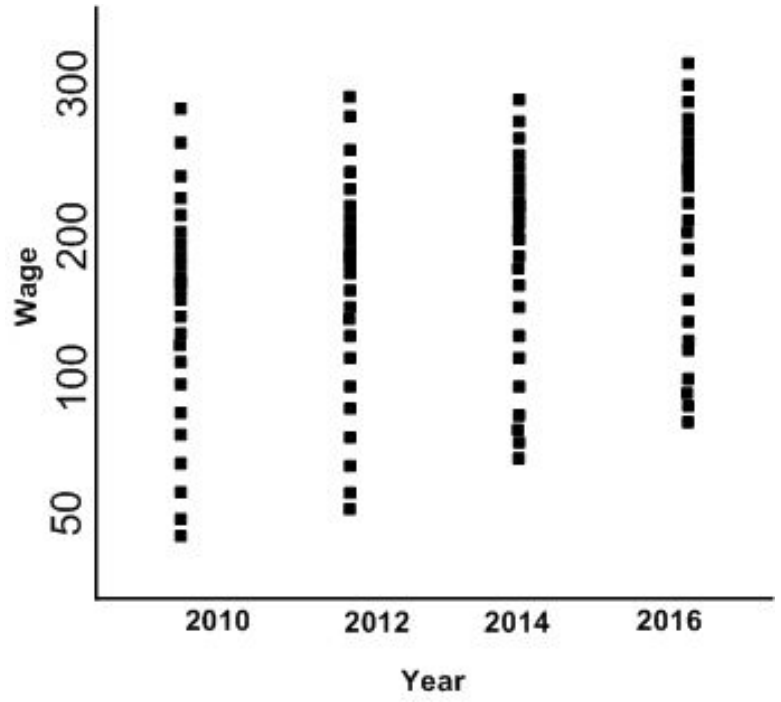


Interpretation: On the average, wage increases with age until about 60 years of age, at which point it begins to decline.

Relationship Analysis

Example: Wage Data

- Case II. Wage versus Year
  - From the data set, we have a graphical representations, which is as follows:



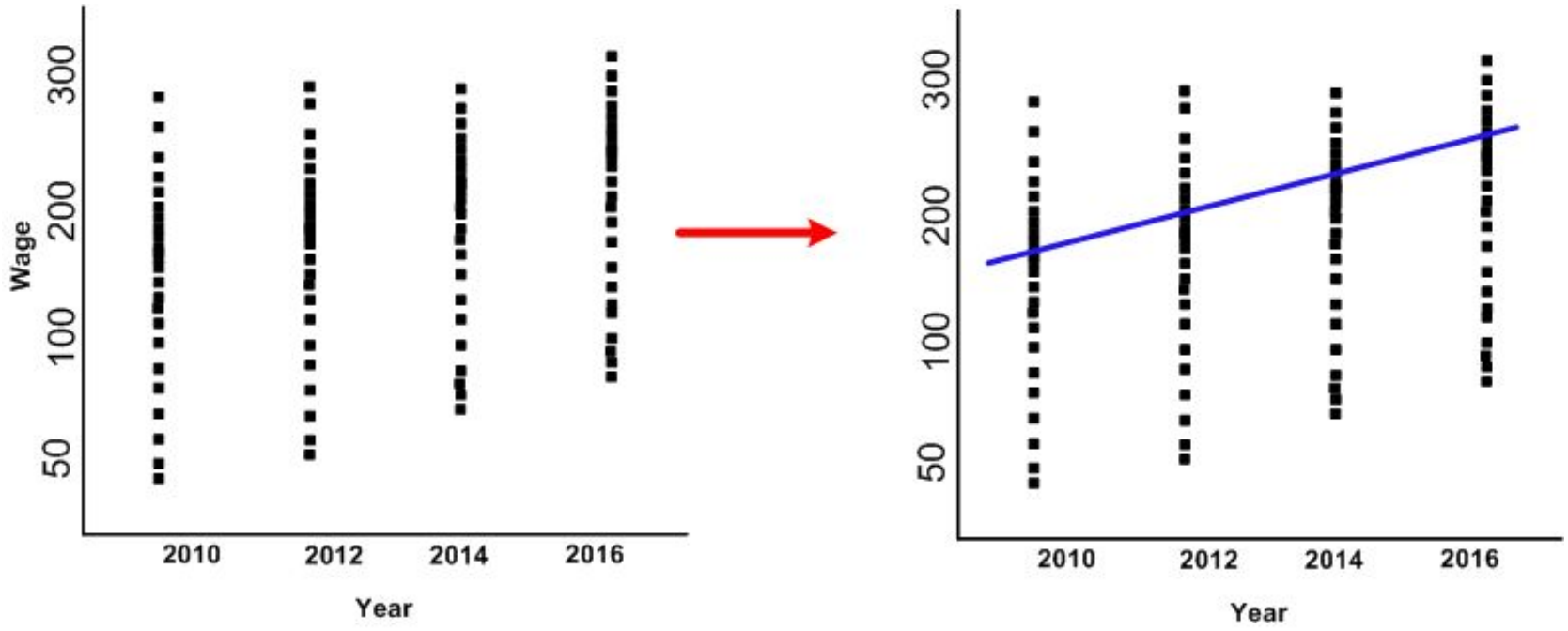
?

How wages vary with time?

Relationship Analysis

Example: Wage Data

- *Wage and calendar year:* How wages vary with years?



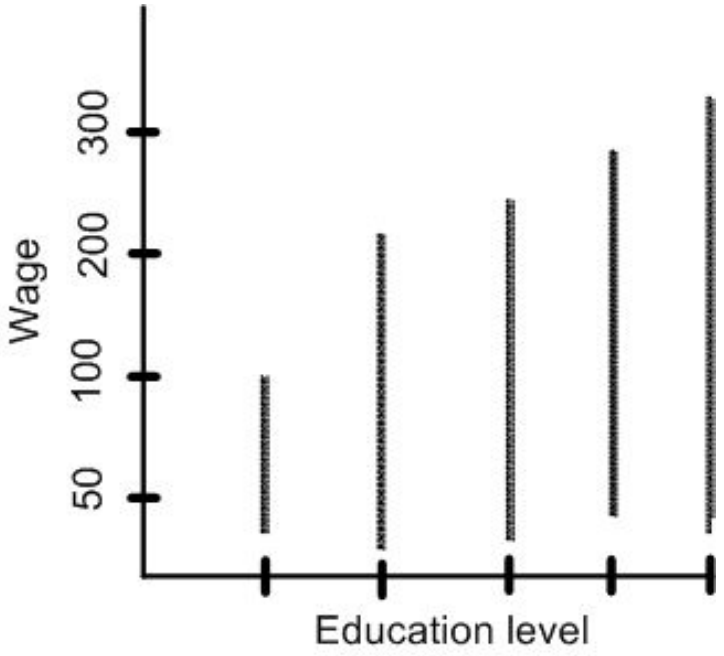
Interpretation: There is a slow but steady increase in the average wage between 2010 and 2016.

.

Relationship Analysis

Example: Wage Data

- Case III. Wage versus Education
  - From the data set, we have a graphical representations, which is as follows:



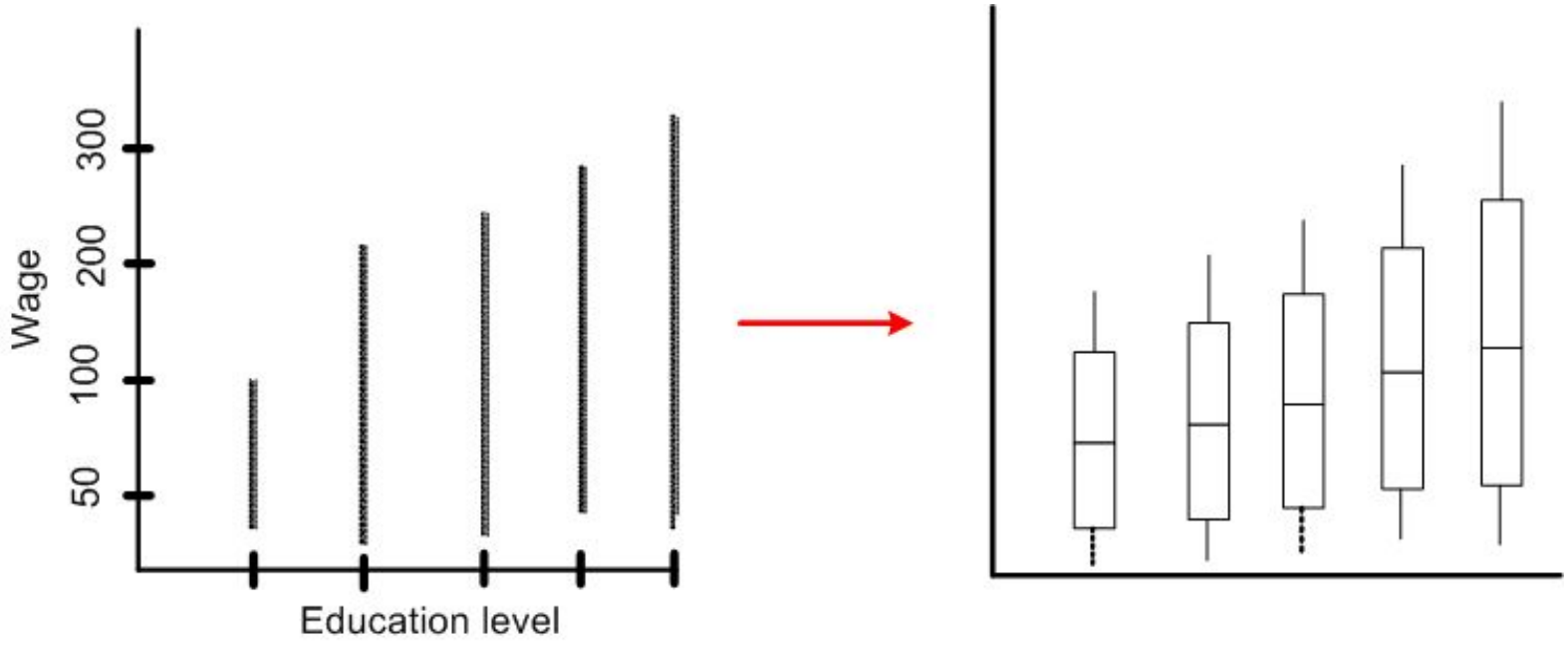
?

Whether wages are related with education?

Relationship Analysis

Example: Wage Data

- *Wage and education level:* Whether wages vary with employees' education levels?



Interpretation: On the average, wage increases with the level of education.

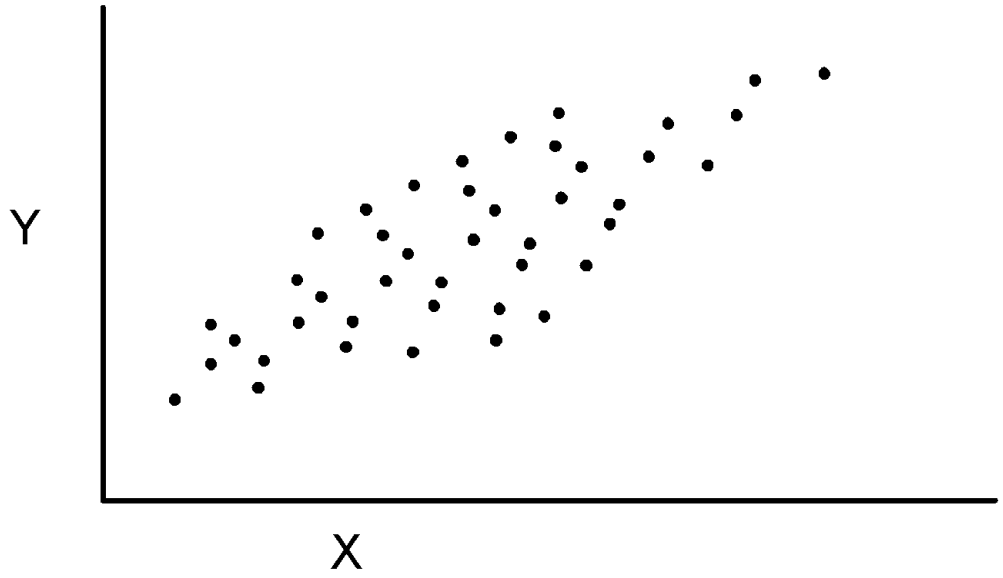
Relationship Analysis

Given an employee's wage can we predict his age?

Whether wage has any association with both year and education level?

etc….

An Open Challenge!

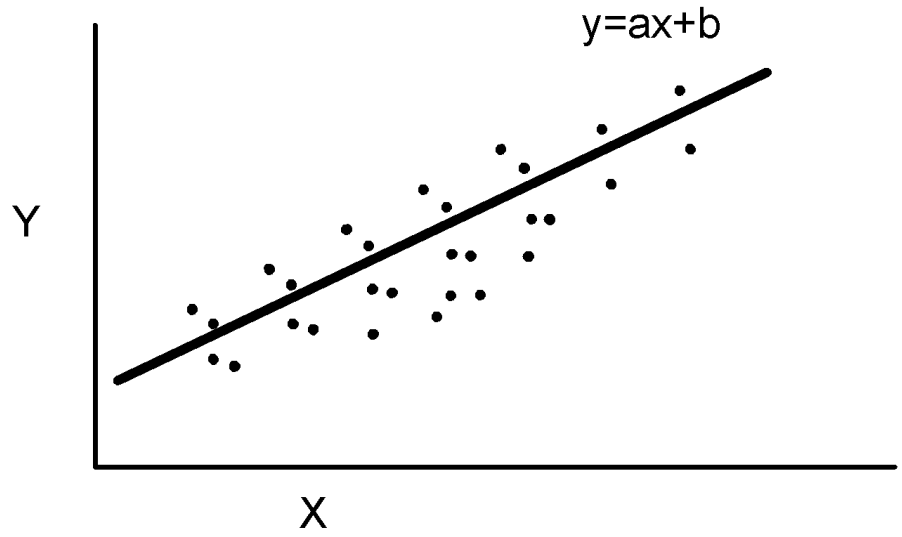


Suppose there are countably infinite points in the *XY plane*. We need a huge memory to store all such points.

Is there any way out to store this information with a least amount of memory?

Say, with two values only.

Yes...



Just decide the values of **a** and **b**
(as if storing one point's data only!)

Note: Here, tricks was to find a relationship among all the points.

Measures of Relationship

- *Univariate population:* The population consisting of only one variable.

Temperature	20	30	21	18	23	45	52

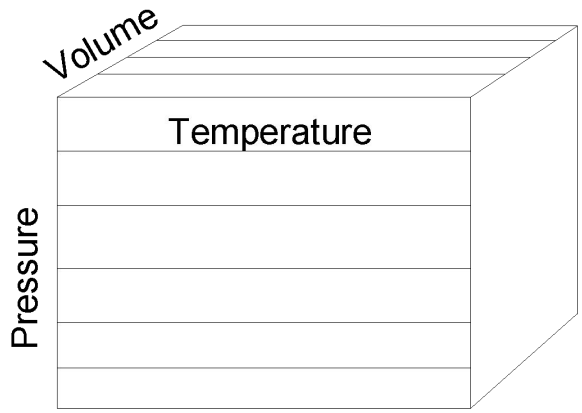
Here, statistical measures are suffice to find a relationship.

- *Bivariate population:* Here, the data happen to be on two variables.

Pressure	1	1.1	... ...	0.8
Temperature	35	41		29

Measures of Relationship

- *Multivariate population:* If the data happen to be one more than two variable.



? If we add another variable say viscosity in addition to Pressure, Volume or Temperature?

Measures of Relationship

In case of bivariate and multivariate populations, usually, we have to answer two types of questions:

Q1: Does there exist correlation (i.e., association) between two (or more) variables?

If yes, of what degree?

Q2: Is there any cause and effect relationship between the two variables (in case of bivariate population) or one variable in one side and two or more variables on the other side (in case of multivariate population)?

If yes, of what degree and in which direction?

To find solutions to the above questions, two approaches are known.

Correlation Analysis

Regression Analysis

Correlation Analysis

Correlation Analysis

In statistics, the word correlation is used to denote some form of association between two variables.

Example: **Weight** is correlated with **height**

Example:

A	a_1	a_2	a_3	a_4	a_5	a_6	... ...
B	b_1	b_2	b_3	b_4	b_5	b_6	... ...

The correlation may be positive, negative or zero.

Positive correlation: If the value of the attribute A increases with the increase in the value of the attribute B and vice-versa.

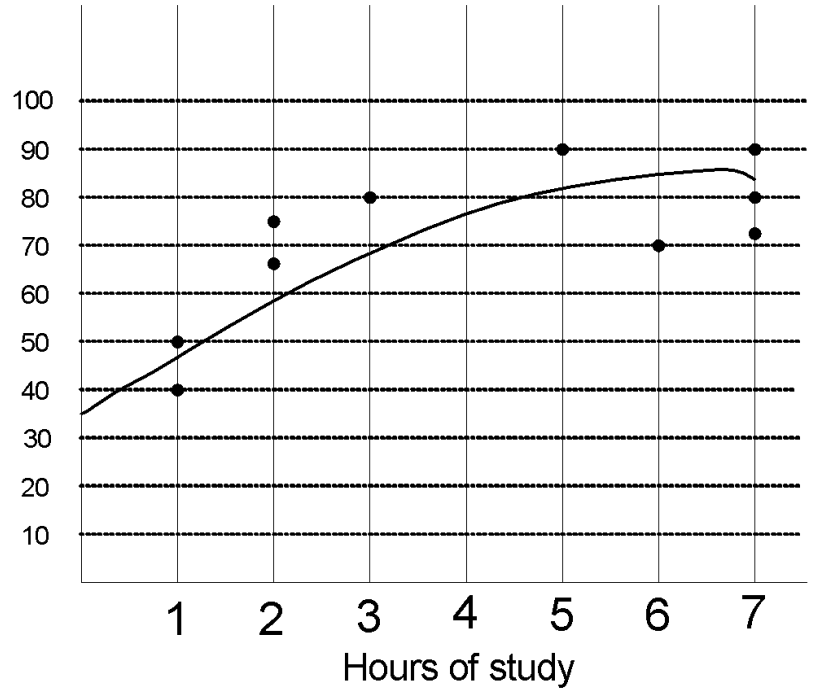
Negative correlation: If the value of the attribute A decreases with the increase in the value of the attribute B and vice-versa.

Zero correlation: When the values of attribute A varies at random with B and vice-versa.

Correlation Analysis

In order to measure the degree of correlation between two attributes.

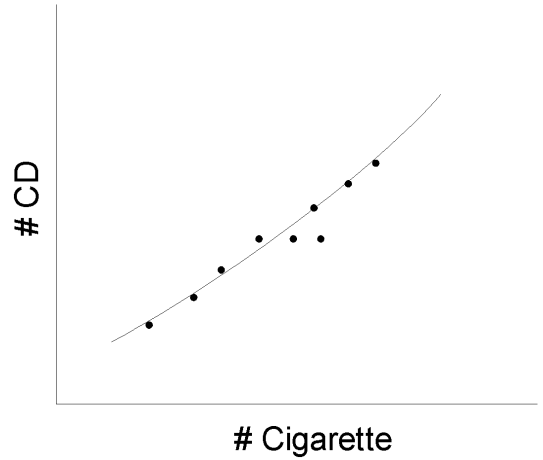
Hours Study	Exam Score
3	80
5	90
2	75
6	80
7	90
1	50
2	65
7	85
1	40
7	100



Correlation Analysis

Do you find any correlation between *X* and *Y* as shown in the table?.

No. of CD's sold in shop X	25	30	35	42	48	52	56
No. of cigarette sold in Y	5	7	9	10	11	11	12

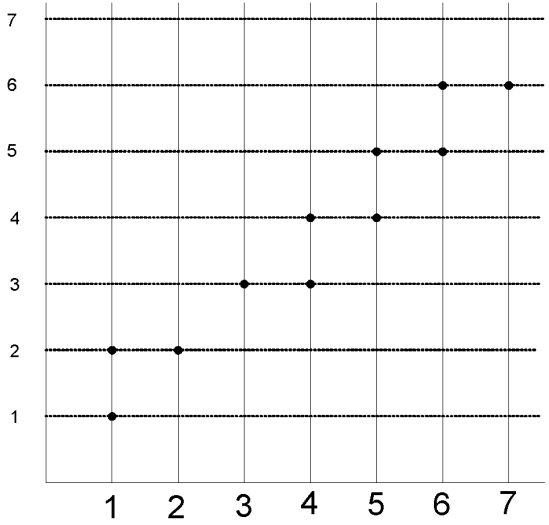


Note:

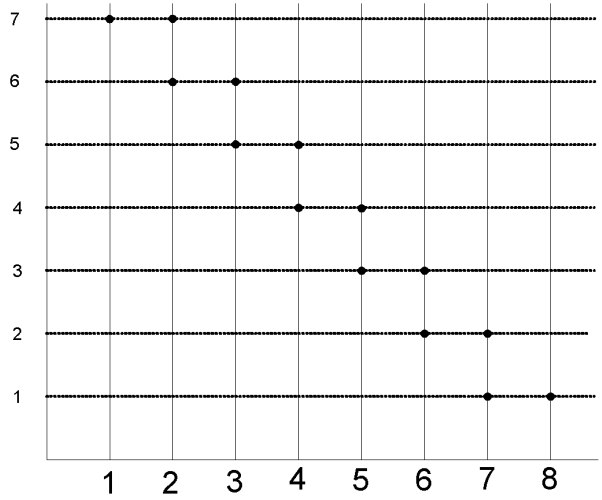
In data analytics, correlation analysis make sense only when relationship make sense. There should be a cause-effect relationship.

Correlation Analysis

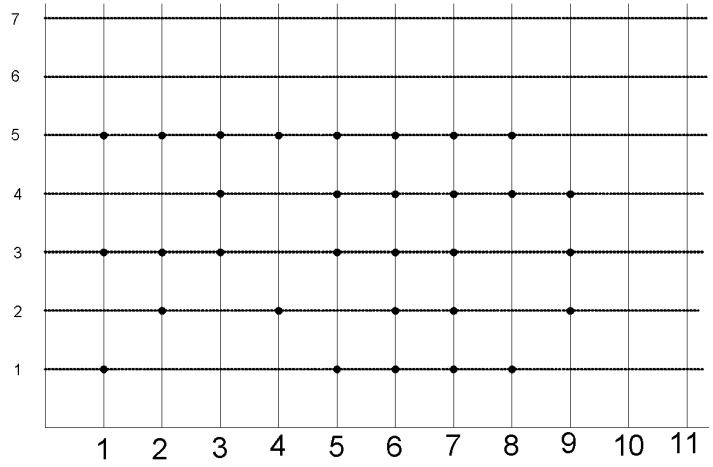
Positive correlation



Negative correlation



Zero correlation

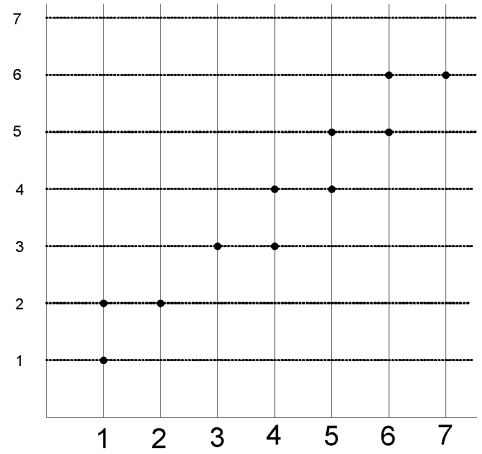


Correlation Coefficient

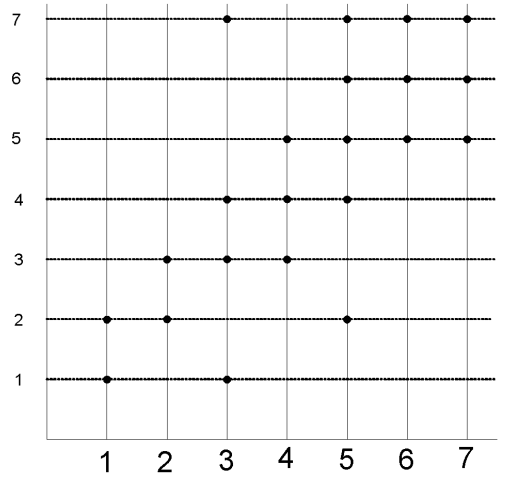
- **Correlation coefficient is used to measure the degree of association.**
- It is usually denoted by *r*.
- The value of *r* lies between +1 and -1.
- Positive values of *r* indicates positive correlation between two variables, whereas, negative values of *r* indicate negative correlation.
- $r = +1$ implies perfect positive correlation, and otherwise.
- The value of *r* nearer to +1 or -1 indicates high degree of correlation between the two variables.
- $r = 0$ implies, there is no correlation

Correlation Coefficient

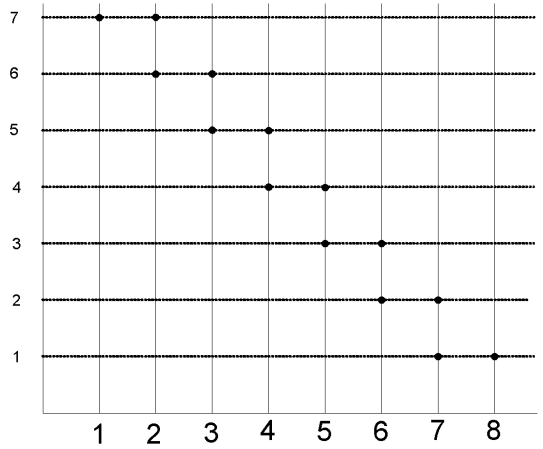
High Positive Correlation



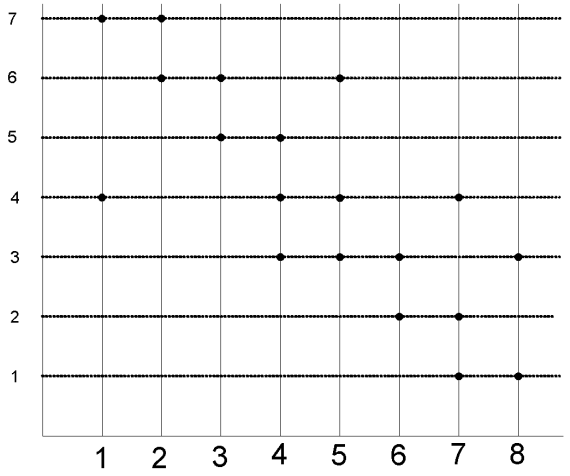
Low Positive Correlation



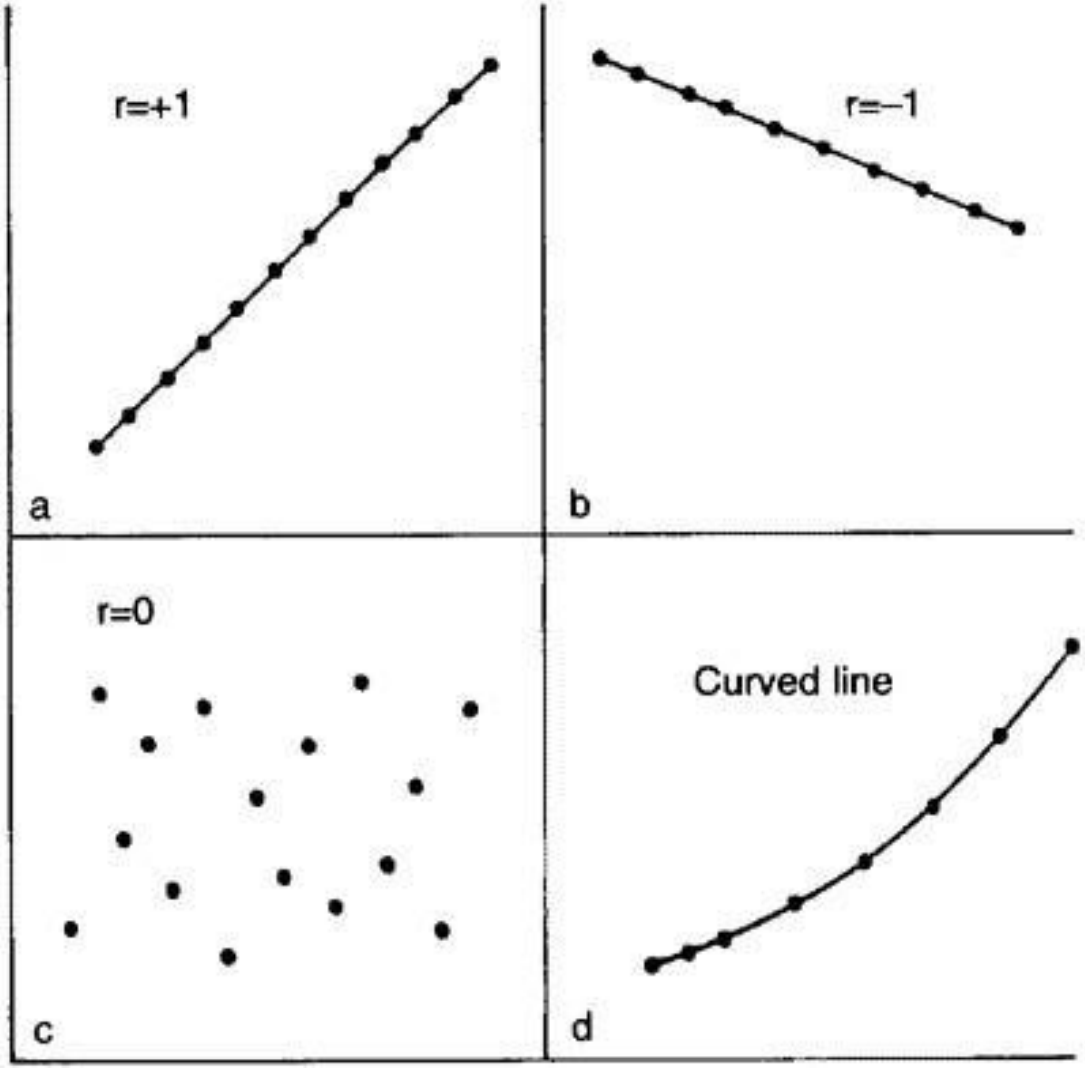
High Negative Correlation



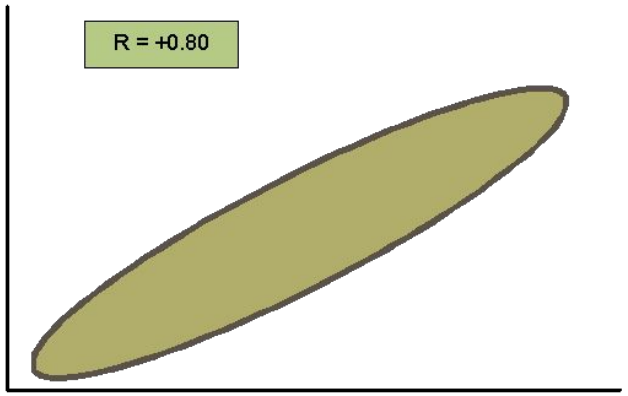
Low Negative Correlation

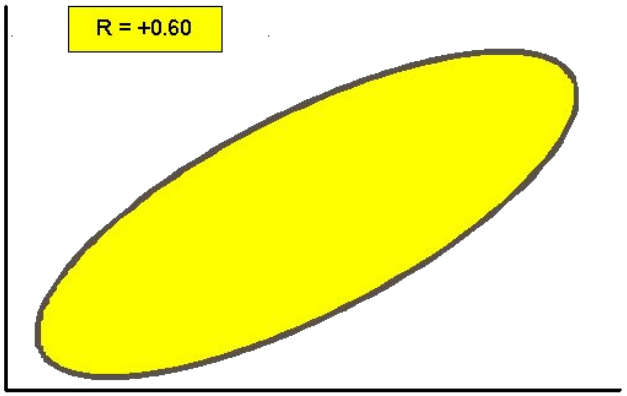


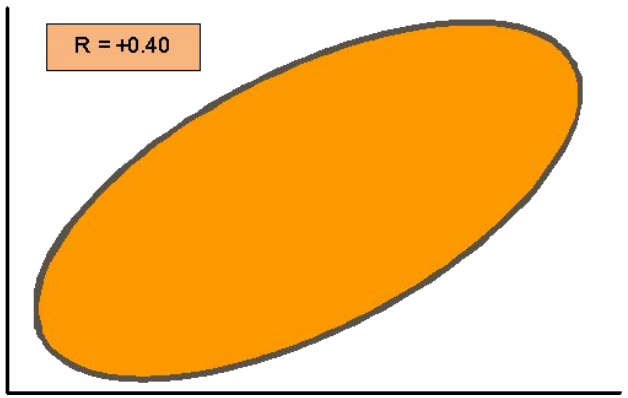
Correlation Coefficient

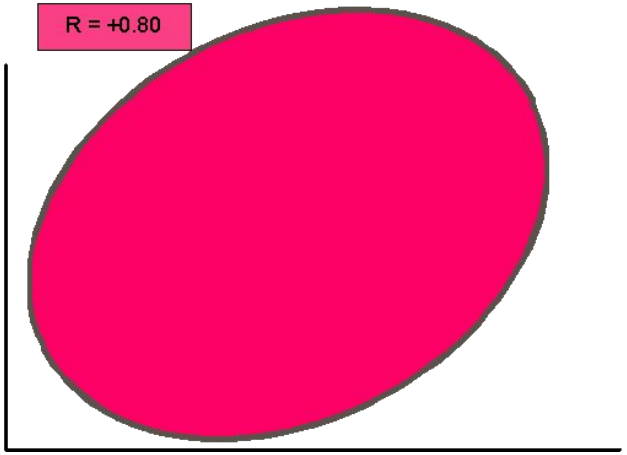


Correlation Coefficient









Measuring Correlation Coefficients

There are three methods known to measure the correlation coefficients

Karl Pearson's coefficient of correlation

This method is applicable to find correlation coefficient between two **numerical** attributes

Charles Spearman's coefficient of correlation

This method is applicable to find correlation coefficient between two **ordinal** attributes

Chi-square coefficient of correlation

This method is applicable to find correlation coefficient between two **categorical** attributes

Pearson's Correlation Coefficient

Karl Pearson's Correlation Coefficient

This is also called **Pearson's Product Moment Correlation**

Definition : **Karl Pearson's correlation coefficient**

Let us consider two attributes are X and Y.

The Karl Pearson's coefficient of correlation is denoted by r^* and is defined as

$$r^* = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1).\sigma_X.\sigma_Y}$$

where $X_i = \text{i} - \text{th value of } X - \text{variable}$

$\bar{X} = \text{mean of } X$

$Y_i = \text{i} - \text{th value of } Y - \text{variable}$

$\bar{Y} = \text{mean of } Y$

$n = \text{number of pairs of observation of } X \text{ and } Y$

$\sigma_X = \text{standard deviations of } X$

$\sigma_Y = \text{standard deviation of } Y$

Karl Pearson's coefficient of Correlation

Example : Correlation of Gestational Age and Birth Weight

A small study is conducted involving 17 infants to investigate the association between gestational age at birth, measured in weeks, and birth weight, measured in grams.

Infant ID #	Gestational Age (wks)	Birth Weight (gm)
1	34.7	1895
2	36.0	2030
3	29.3	1440
4	40.1	2835
5	35.7	3090
6	42.4	3827
7	40.3	3260
8	37.3	2690
9	40.9	3285
10	38.3	2920
11	38.5	3430
12	41.4	3657
13	39.7	3685
14	39.7	3345
15	41.1	3260
16	38.0	2680
17	38.7	2005

Karl Pearson's coefficient of Correlation

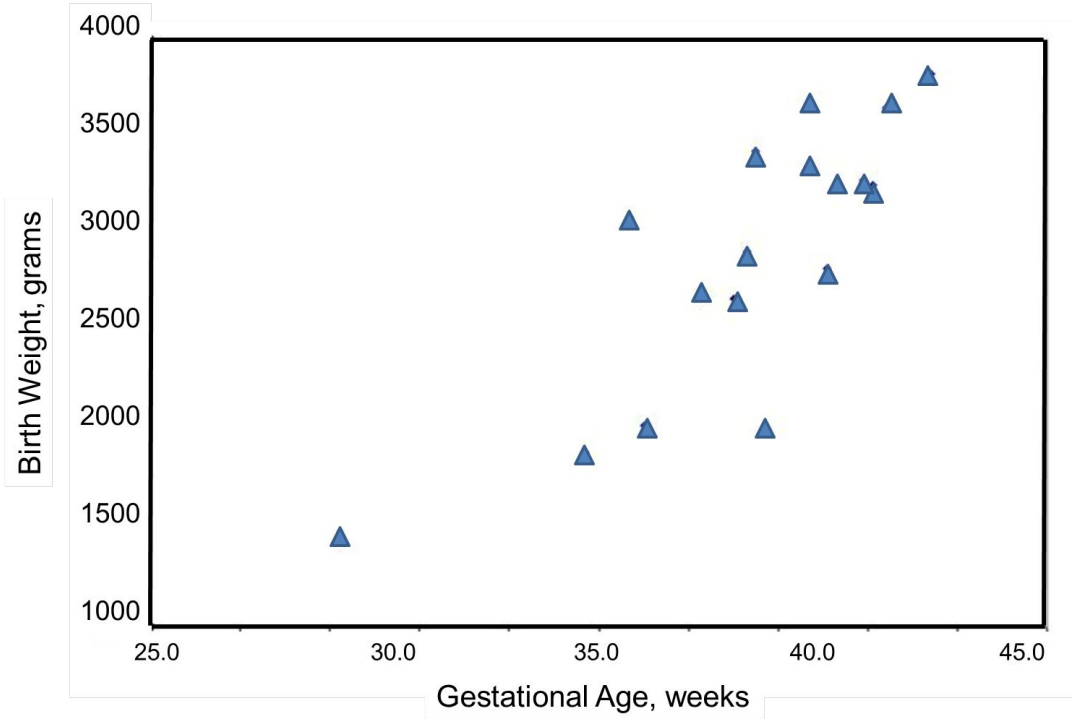
Example 7.1: Correlation of Gestational Age and Birth Weight

We wish to estimate the association between gestational age and infant birth weight.

In this example, birth weight is the dependent variable and gestational age is the independent variable. Thus *Y* = birth weight and *X* = gestational age.

The data are displayed in a scatter diagram in the figure below.

Example



Karl Pearson's coefficient of Correlation

Example 7.1: Correlation of Gestational Age and Birth Weight

For the given data, it can be shown the following

$$\overline{X} = \frac{\Sigma X}{n} = \frac{652.1}{17} = 38.4.$$

$$\overline{Y} = \frac{\Sigma Y}{n} = \frac{49{,}334}{17} = 2902.$$

$$s_x^2 = \frac{\Sigma(X - \overline{X})^2}{n-1} = \frac{159.45}{16} = 10.0.$$

$$s_y^2 = \frac{\Sigma(Y - \overline{Y})^2}{n-1} = \frac{7{,}767{,}660}{16} = 485{,}578.8.$$

$$r^* = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{s_x . s_y} = 0.82$$

Conclusion: The sample's correlation coefficient indicates a strong positive correlation between Gestational Age and Birth Weight.

Karl Pearson's coefficient of Correlation

Example 7.1: Correlation of Gestational Age and Birth Weight

- Significance Test
  - To test whether the association is merely apparent, and might have arisen by chance use the *t* test in the following calculation

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

  - Number of pair of observation is 17. Hence,

$$t = 0.82\sqrt{\frac{17-2}{1-0.82^2}} = 1.44$$

  - Consulting the t-test table, at degrees of freedom 15 and for $\alpha = 0.05$, we find that t = 1.753. Thus, the value of Pearson's correlation coefficient in this case may be regarded as highly significant.

Rank Correlation Coefficient

Charles Spearman's Correlation Coefficient

- This correlation measurement is also called **Rank correlation**.
- This technique is applicable to determine the degree of correlation between two variables in case of ordinal data.
- We can assign rank to the different values of a variable with ordinal data type.

Example:

Height: [VS S L T VT]
1 2 3 4 5

T – shirt: [XS S L XL XXL]
11 12 13 14 15

Rank assigned

Charles Spearman's Correlation Coefficient

Definition : **Charles Spearman's correlation coefficient**

The rank correlation can be defined as

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

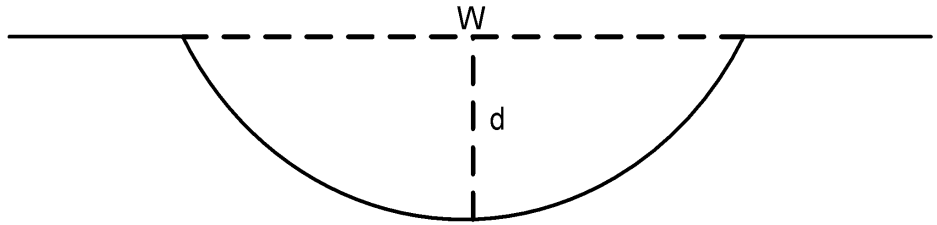
where d_i = Difference between ranks of i^{th} pair of the two variables

n = Number of pairs of observations

The Spearman's coefficient is often used as a statistical methods to aid either proving or disproving a hypothesis.

Charles Spearman's Coefficient of Correlation

Example : The hypothesis that the depth of a river **does not progressively increase** with the width of the river.



A sample of size 10 is collected to test the hypothesis, using Spearman's correlation coefficient.

Sample#	Width in m	Depth in m
1	0	0
2	50	10
3	150	28
4	200	42
5	250	59
6	300	51
7	350	73
8	400	85
9	450	104
10	500	96

Charles Spearman's Coefficient of Correlation

Step 1: Assign rank to each data. It is customary to assign rank 1 to the largest data, and 2 to next largest and so on.

Note: If there are two or more samples with the same value, the mean rank should be used.

Data	20	25	25	25	30
Assign rank	5	4	3	2	1
Final rank	5	3	3	3	1

Charles Spearman's Coefficient of Correlation

Step 2: The contingency table will look like

Sample	Width	Width r	Depth	Depth r	d	d^2
1	0	10	0	10	0	0
2	50	9	10	9	0	0
3	150	8	28	8	0	0
4	200	7	42	7	0	0
5	250	6	59	5	1	1
6	300	5	51	6	−1	1
7	350	4	73	4	0	0
8	400	3	85	3	0	0
9	450	2	104	1	1	1
10	500	1	96	2	−1	1

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6 \times 4}{10 \times 99}$$

$$r_s = 0.9757$$

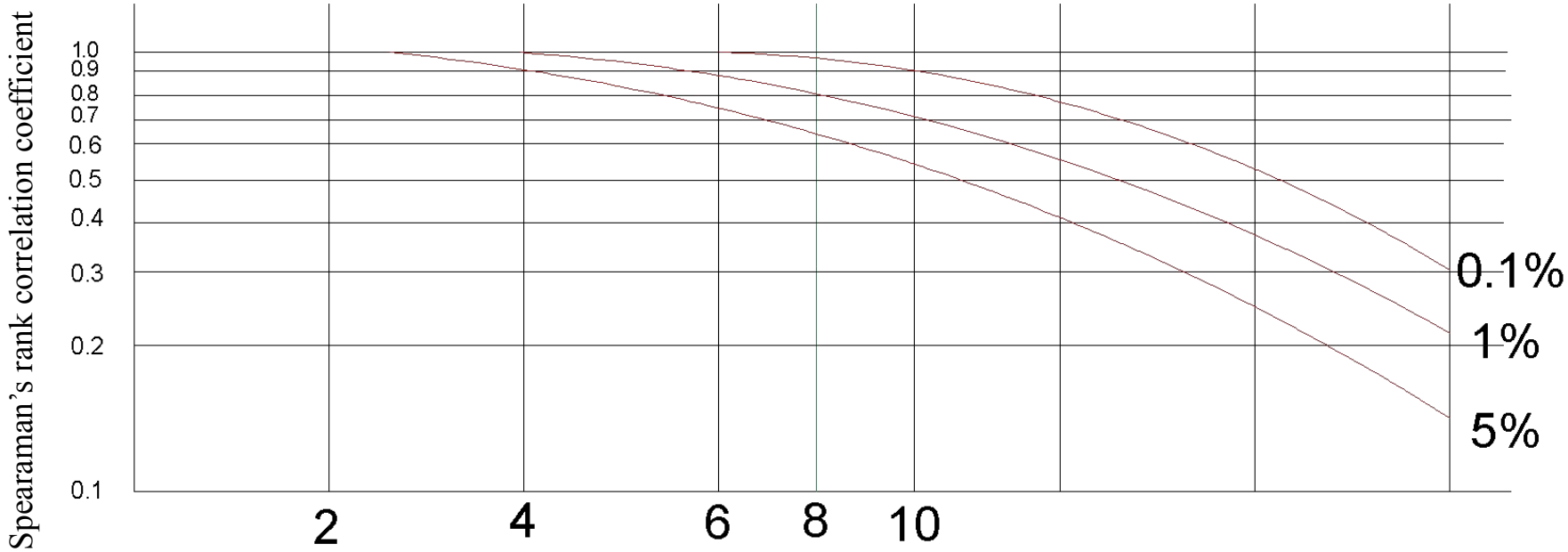
$$\sum d^2 = 4$$

Charles Spearman's Coefficient of Correlation

Step 3: To see, if this r_s value is significant, the Spearman's rank significance table (or graph) must be consulted.

Note: The degrees of freedom for the sample $= n - 2 = 8$

Assume, the significance level $= 0.1\%$



Charles Spearman's Coefficient of Correlation

Step 4: Final conclusion

From the graph, we see that $r_s = 0.9757$ lies above the line at 8 and 0.01% significance level. Hence, there is a greater than 99% chance that the relationship is significant (i.e., not random) and hence the hypothesis should be rejected.

Thus, we can reject the hypothesis and conclude that in this case, depth of a river **progressively increases** the further with the width of the river.

χ2-Correlation Analysis

Chi-Squared Test of Correlation

- This method is also alternatively termed as Pearson's χ^2–test or simply χ^2 -test
- This method is applicable to categorical (discrete) data only.

  - Suppose, two attributes A and B with categorical values

    $A = a_1, a_2, a_3, \ldots\ldots, a_m$ and

    $B = b_1, b_2, b_3, \ldots\ldots, b_n$

    having m and n distinct values.

A	a_1	a_2	a_3	a_1	a_5	a_1	… …
B	b_1	b_2	b_3	b_1	b_5	b_1	… …

Between whom we are to find the correlation relationship.

χ^2 –Test Methodology

Contingency Table

Given a data set, it is customary to draw a contingency table, whose structure is given below.

	b_1	b_2	-----	b_j	------	b_n	**Row Total**
a_1							
a_2							
⋮							
a_i							
⋮							
a_m							
Column Total							**Grand Total**

χ^2 –Test Methodology

Entry into Contingency Table: Observed Frequency

In contingency table, an entry O_{ij} denotes the event that attribute A takes on value a_i and attribute B takes on value b_j (i.e., $A = a_i$, $B = b_j$).

A	a_i	a_2	a_3	a_i	a_5	a_i	... ...
B	b_j	b_2	b_3	b_j	b_5	b_j	... ...

	b_1	b_2	-----	b_j	------	b_n	**Row Total**
a_1							
a_2							
⋮							
a_i				O_{ij}			
⋮							
a_m							
Column Total							**Grand Total**

χ^2 –Test Methodology

Entry into Contingency Table: Expected Frequency

In contingency table, an entry e_{ij} denotes the expected frequency, which can be calculated as

$$e_{ij} = \frac{Count(A = a_i) \times Count(B = b_j)}{Grand\ Total} = \frac{A_i \times B_j}{N}$$

	b_1	b_2	-----	b_j	------	b_n	Row Total
a_1							
a_2							
⋮							
a_i				e_{ij}			A_i
⋮							
a_m							
Column Total				B_j			N

χ^2 – Test

Definition : χ^2-Value

The χ^2 value (also known as the Pearson's χ^2 test) can be computes as

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

where o_{ij} is the **o**bserved frequency

e_{ij} is the **e**xpected frequency

χ^2 – Test

The cell that contribute the most to the χ^2 value are those whose actual count is very different from the expected.

The χ^2 statistics tests the hypothesis that *A* and *B* are independent. The test is based on a significance level, with $(n\text{-}1) \times (m\text{-}1)$ degrees of freedom., with a contingency table of size $n \times m$

If the hypothesis can be rejected, then we say that *A* and *B* are statistically related or associated.

χ^2 – Test

Example : Survey on Gender versus Hobby.

Suppose, a survey was conducted among a population of size 1500. In this survey, gender of each person and their hobby as either "book" or "computer" was noted. The survey result obtained in a table like the following.

GENDER	HOBBY
.....	.....
.....	.....
M	Book
F	Computer
.....	.....
.....	.....
.....	.....

We have to find if there is any association between Gender and Hobby of a people, that is, we are to test whether "gender" and "hobby" are correlated.

χ^2 – Test

Example : Survey on Gender versus Hobby.

From the survey table, the observed frequency are counted and entered into the contingency table, which is shown below.

		GENDER		
		Male	Female	Total
HOBBY	Book	250	200	450
	Computer	50	1000	1050
	Total	300	1200	**1500**

χ^2 – Test

Example : Survey on Gender versus Hobby.

From the survey table, the expected frequency are counted and entered into the contingency table, which is shown below.

		GENDER		
		Male	Female	Total
HOBBY	Book	90	360	450
	Computer	210	840	1050
	Total	300	1200	**1500**

χ^2 – Test

- Using equation for χ^2 computation, we get

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840}$$
$$= 507.93$$

- This value needs to be compared with the tabulated value of χ^2 (available in any standard book on statistics) with 1 degree of freedom (for a table of $m \times n$, the degrees of freedom is $(m-1)\times(n-1)$; here $m = 2$, $n = 2$).

- For 1 degree of freedom, the χ^2 value needed to reject the hypothesis at the 0.01 significance level is 10.828. Since our computed value is above this, we reject the hypothesis that "Gender" and "Hobby" are independent and hence, conclude that the two attributes are *strongly correlated* for the given group of people.

χ^2 – Test

Example 7.4: Hypothesis on "accident proneness" versus "driver's handedness".

- Consider the following contingency table on car accidents among left and right-handed drivers' of sample size 175.
- Hypothesis is that *"fatality of accidents is independent of driver's handedness"*

		HANDEDNESS		
FATALITY		Left-Handed	Right-Handed	Total
	Non-Fatal	8	141	149
	Fatal	3	23	26
	Total	11	164	175

- Find the correlation between Fatality and Handedness and test the significance of the correlation with significance level 0.1%.

Regression Analysis

Regression Analysis

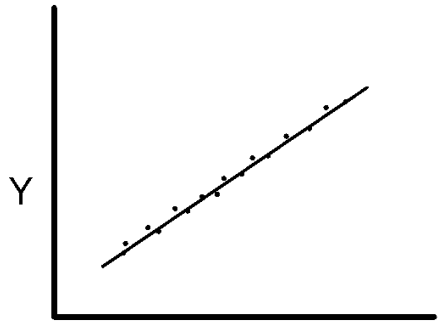
The regression analysis is a statistical method to deal with the formulation of mathematical model depicting relationship amongst variables, which can be used for the purpose of prediction of the values of dependent variable, given the values of independent variables.

Classification of Regression Analysis Models

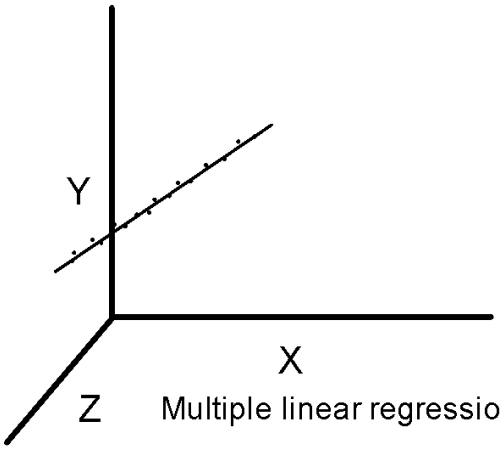
Linear regression models

1. Simple linear regression
2. Multiple linear regression

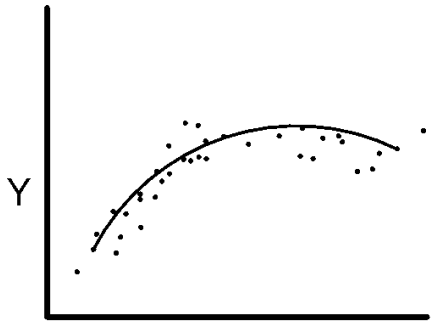
Non-linear regression models



Simple linear regression



Multiple linear regression

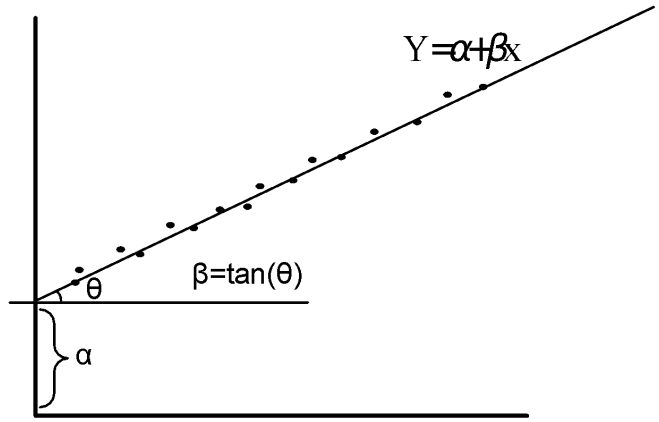


Non-linear regression

Simple Linear Regression Model

In simple linear regression, we have only two variables:

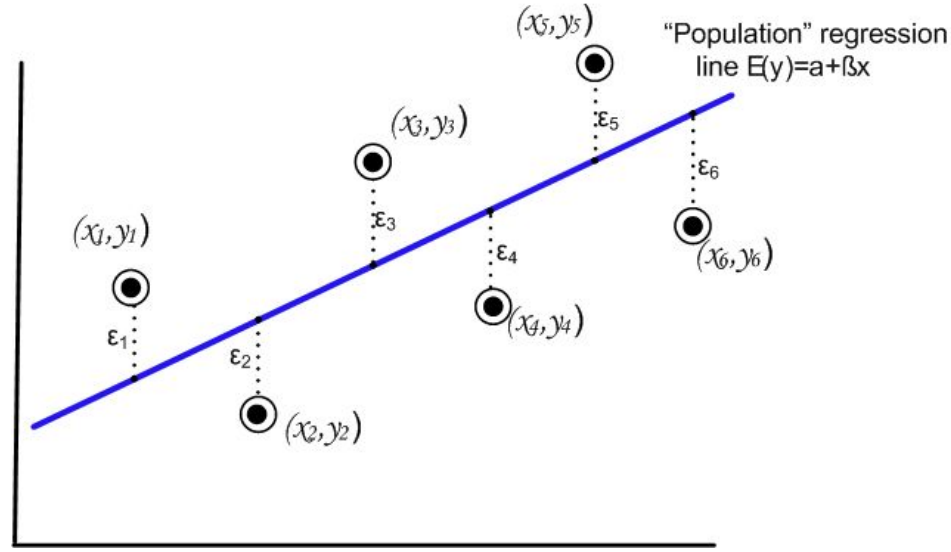
- Dependent variable (also called Response), usually denoted as Y.
- Independent variable (alternatively called Regressor), usually denoted as x.
- A reasonable form of a relationship between the Response Y and the Regressor x is the linear relationship, that is in the form $Y = \alpha + \beta x$



Note:

- There are infinite number of lines (and hence α_s *and* β_s)
- The concept of regression analysis deal with finding the best relationship between Y and x (and hence best fitted values of α *and* β) quantifying the strength of that relationship.

Regression Analysis



Given the set $[(x_i, y_i), i = 1,2, \ldots \ldots, n]$ of data involving n pairs of (x, y) values, our objective is to find "true" or population regression line such that $Y = \alpha + \beta x + \in$

Here, $\in$ is a random variable with $E(\in) = 0$ and $var(\in) = \sigma^2$. The quantity σ^2 is often called the **error variance**.

Note:

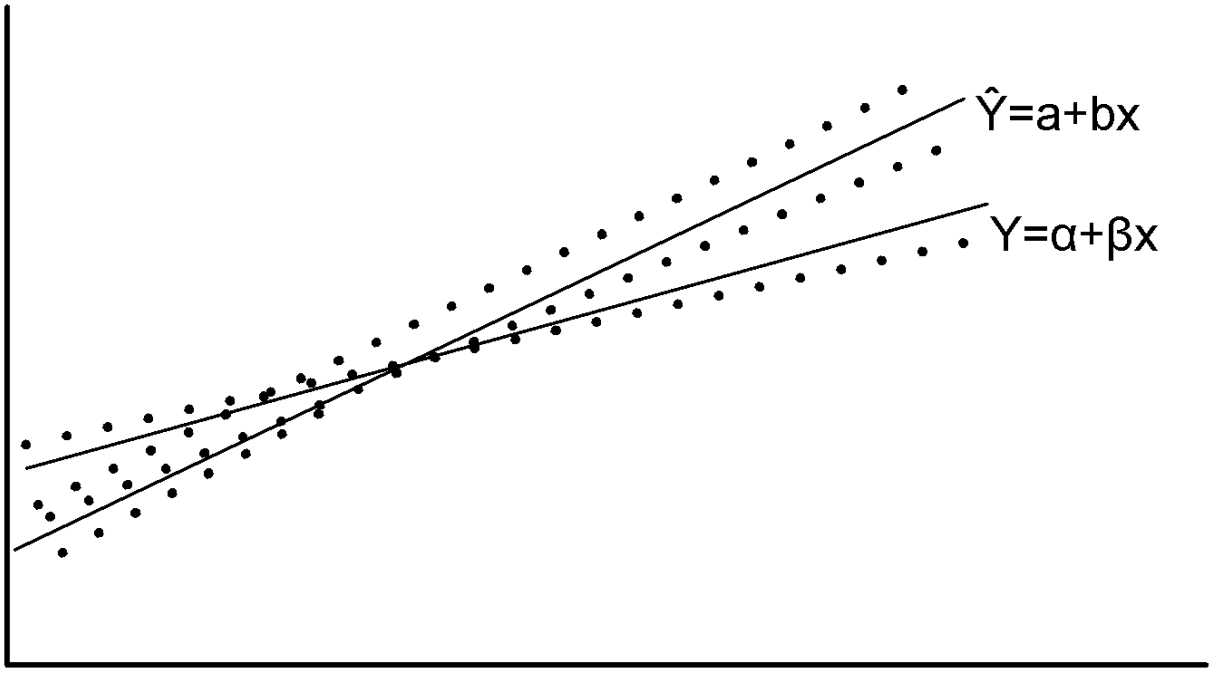
- $E(\in) = 0$ implies that at a specific x, the y values are distributed around the "true" regression line $Y = \alpha + \beta x$ (i.e., the positive and negative errors around the true line is reasonable).
- α and β are called **regression coefficients**.
- α and β values are to be estimated from the data.

True versus Fitted Regression Line

- The task in regression analysis is to estimate the regression coefficients α and β.
- Suppose, we denote the estimates a for α and b for β. Then the fitted regression line is

$$\hat{Y} = a + bx$$

where $\hat{Y}$ is the predicted or fitted value.

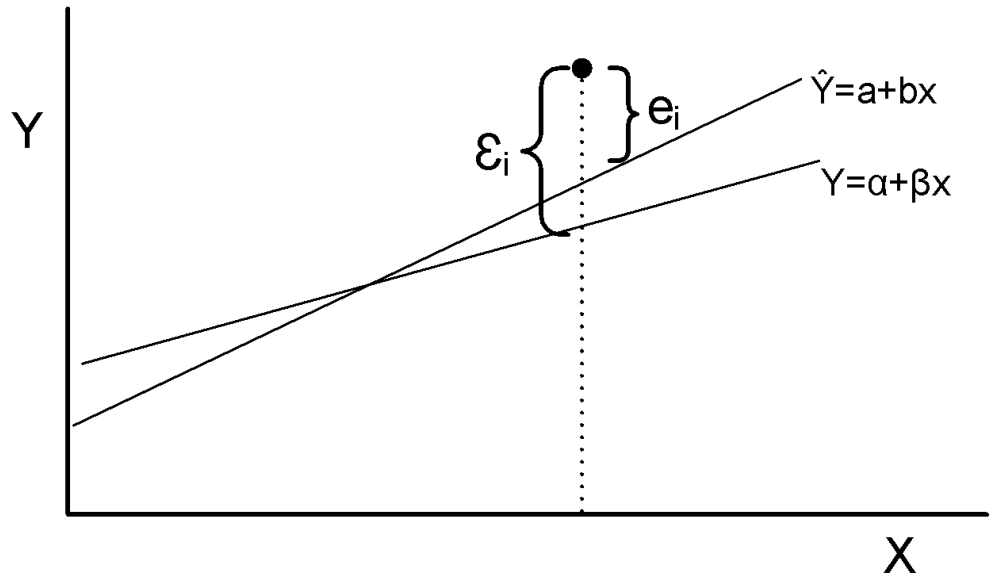


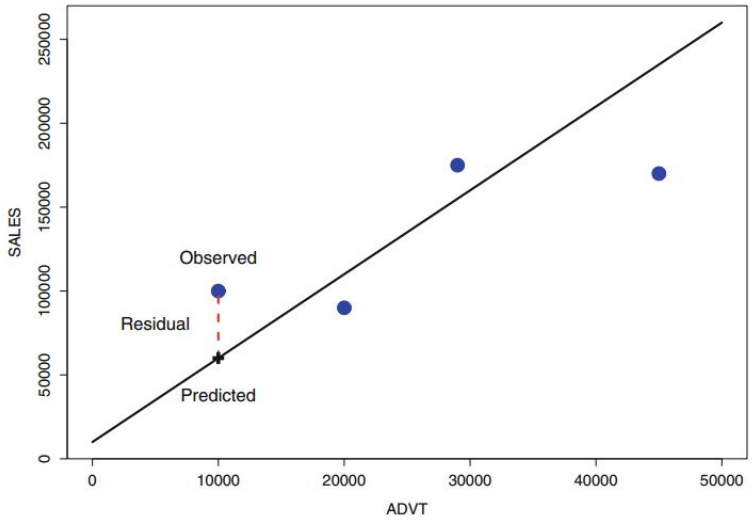
Least Square Method to estimate α and β

This method uses the concept of residual. A residual is essentially an error in the fit of the model $\hat{Y} = a + bx$. Thus, i^{th} residual is

$$e_i = Y_i - \hat{Y}_i, i = 1,2,3, \ldots \ldots, n$$

Residual = observed – predicted





Least Square method

- The residual sum of squares is often called **the sum of squares of the errors** about the fitted line and is denoted as SSE

$$\text{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

- We are to minimize the value of SSE and hence to determine the parameters of a and b.
- Differentiating SSE with respect to a and b, we have

$$\frac{\partial(SSE)}{\partial a} = -2 \sum_{i=1}^{n} (y_i - a - bx_i)$$

$$\frac{\partial(SSE)}{\partial b} = -2 \sum_{i=1}^{n} (y_i - a - bx_i).x_i$$

For minimum value of SSE, $\frac{\partial(SSE)}{\partial a} = 0$

$$\frac{\partial(SSE)}{\partial b} = 0$$

Least Square method to estimate α and β

Thus we set

$$na + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

These two equations can be solved to determine the values of a and b, and it can be calculated that

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$a = \bar{y} - b\bar{x}$$

Example

	Student	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
	1	95	85	17	8	289	64	136
	2	85	95	7	18	49	324	126
	3	80	70	2	-7	4	49	-14
	4	70	65	-8	-12	64	144	96
	5	60	70	-18	-7	324	49	126
Sum		390	385			730	630	470
Mean		78	77					

The regression equation is a linear equation of the form: $\hat{y} = b_0 + b_1 x$. To conduct a regression analysis, we need to solve for b_0 and b_1. Computations are shown below.

$b_1 = \Sigma [(x_i - \bar{x})(y_i - \bar{y})] / \Sigma [(x_i - \bar{x})^2]$ $b_1 = 470/730 = 0.644$	$b_0 = \bar{y} - b_1 * \bar{x}$ $b_0 = 77 - (0.644)(78) = 26.768$

Therefore, the regression equation is: $\hat{y} = 26.768 + 0.644x$.

Watch Out!

$$\text{Sales} = \$10{,}000 + 5 \times \text{Advertising}$$

So, if advertising = 1000, sales = 15000

How about…

If sales = 20000, advertising =?

Watch Out!

$$\text{Sales} = \$10{,}000 + 5 \times \text{Advertising}$$

So, if advertising = 1000, sales = 15000

How about…

If sales = 20000, advertising =?

Mistake!

Our model is $\hat{y} = b_0 + b_1 x$ and we want y from x !!

Slope from y to x is *not* reciprocal of slope from x to y

Aim of Regression Analysis

Find the coefficients and interpret the results

You don't need to be superman to find the coefficients

- It is all about one mouse click
- One line of code in R

What is more important is to be able to interpret the results

Typical Output

```
Call:
lm(formula = SALES ~ ADVT)

Residuals:
     Min       1Q   Median       3Q      Max
-31.0945  -9.9708   0.4255   9.6146  21.7419

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   51.849     27.990   1.852   0.0768 .
ADVT           7.527      2.741   2.746   0.0115 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.51 on 23 degrees of freedom
Multiple R-squared: 0.2469,     Adjusted R-squared: 0.2142
F-statistic:  7.54 on 1 and 23 DF,  p-value: 0.01151
```

Part I

Part II

Regression Coefficients

Always be skeptical and ask yourself if the answer is reasonable

Multiple R Squared

Often referred to as R Squared

Measures the overall quality of a regression model

How well the idealized model tracks the actual data

We got R^2 of 0.2469

What does it mean?

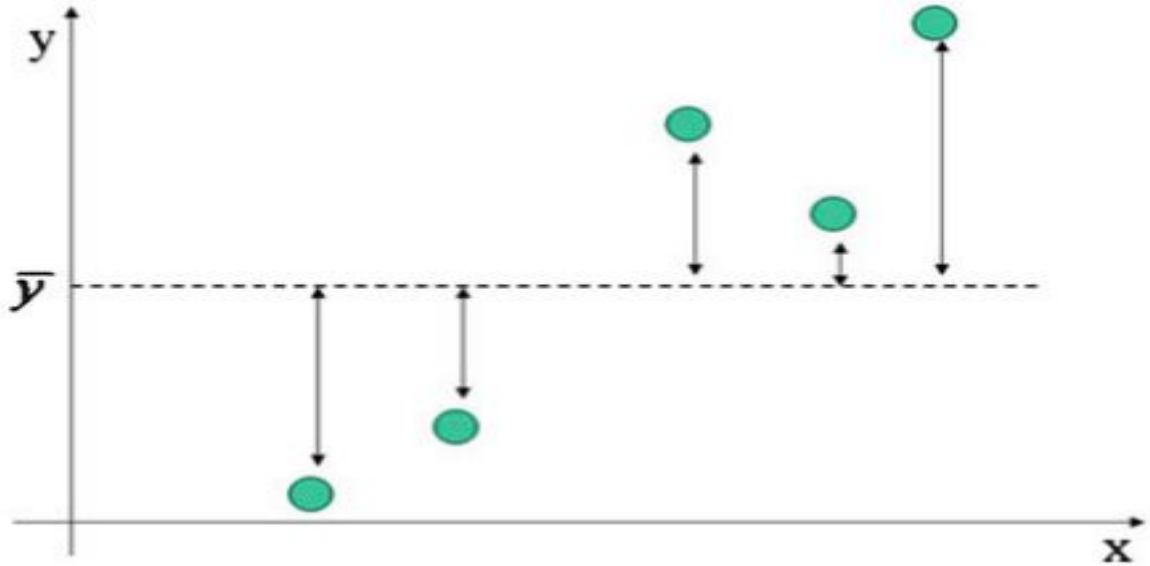
Standard Error of Regression

The problem though is that the standard error is in units of the dependent variable, and on its own is difficult to interpret as being big or small.

The fact that it is expressed in the squares of the units makes it a bit more difficult to comprehend

What do we compare the standard error to in order to determine how good our regression is? How big is big?

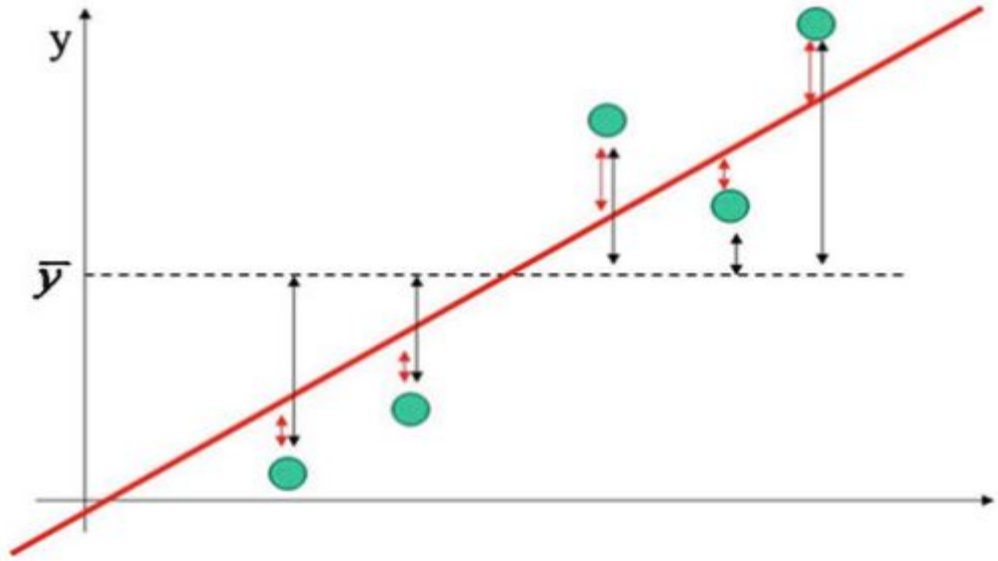
“Average Model”



How good or bad is it?
Variability: Total Sum of Squares (SST)

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Our Regression Model



$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Error sum of squares = SSE (Sum of squares due to error)
How much the model has helped reduce the uncertainty?
SSR = SST – SSE (Sum of Squares due to Regression)
This quantifies the uncertainty that regression model was able to model away

R^2 : Measure of Quality of Fit

- A quantity R^2, is called **coefficient of determination** is used to measure the proportion of variability of the fitted model.
- We have $SSE = \sum_{i=1}^{n}(y_i - \hat{y})^2$
- It signifies the **variability due to error**.
- Now, let us define the total corrected sum of squares, defined as

$$\text{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

SST measures the uncertainty in predicting Y when X is not considered

- SST represents the variation in the response values. The R^2 is

$$R^2 = 1 - \frac{SSE}{SST} \qquad R\text{-}squared = \frac{SSR}{SST}$$

Note:

- If fit is perfect, all residuals are zero and thus $R^2 = 1.0$ (very good fit)
- If SSE is only slightly smaller than SST, then $R^2 \approx 0$ (very poor fit)

Interpretation of R^2

R^2 of 0.2469 means our regression model explains 24.69% of total variability in sales

- Higher the better
- Is 0.2469 good or bad?
  - Depends upon the context
  - Scientific experiments 0.8 - 0.9
  - Observational studies and surveys 0.3-0.5 or even lower is ok

There is no single benchmark that applies equally to all situations

It is the first part of a regression that many people look at because, along with the scatterplot, it tells whether the regression model is even worth thinking about

Extreme Cases!

One major company developed a method to differentiate between proteins

To do so, they had to distinguish between regressions with of 99.99% and 99.98%.

For this application, 99.98% was not high enough!

The president of a financial services company
reports that although his regressions give below 2%, they are highly successful because those used by his competition are even lower

If…

… you get an R^2 of 0

It means that none of the variance in the data is in the model; all of it is still in the residuals

Useless model

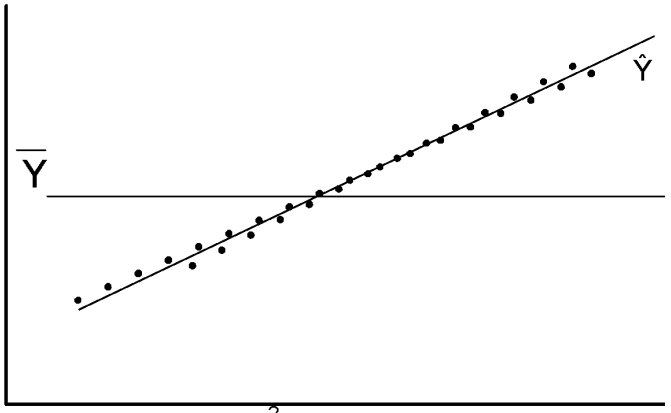
If…

…you get R^2 of 1,

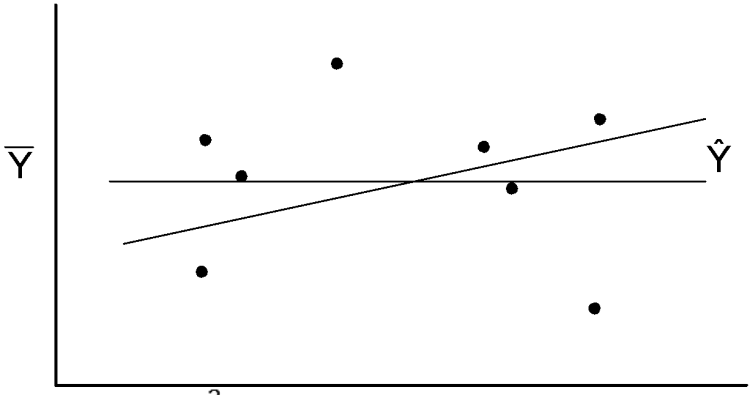
You may have discovered a new law of Physics

But it's much more likely that you accidentally regressed two variables that measure the same thing ☺

R^2 : Measure of Quality of Fit



$R^2 \approx 1.0$ (Very good fit)



$R^2 \approx 0$ (Very poor fit)

Comparing R^2

First model explained 24.69% of total variability in sales data

Second model explains 45.2%

Certainly second model fits the data better

Caution!

- R^2 will never decrease, even if you keep adding non sense variables to the model
- We could have added the weight of each region's sales manager to the model and *R*-squared would not have decreased!

Thus, we should not overly rely on *R*-squared alone when comparing models

Adjusted R^2

Adjusted R-squared penalizes the model for the inclusion of nonsense variables, and hence we can use it to compare models with different variables

The ***adjusted R^2*** takes into account the number of independent variables in the model, and the sample size, and provides a more accurate assessment of the reliability of the model

It is better to use Adjusted R-squared when there are multiple variables in the regression model. This would allow us to compare models with differing numbers of independent variables.

Multiple Linear Regression

- When more than one variable are independent variable, then the regression can be estimated as a multiple regression model
- When this model is linear in coefficients, it is called multiple linear regression model
- If k-independent variables $x_1, x_2, x_3 \ldots\ldots\ldots, x_k$ are associated, the multiple linear regression model is given by

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_k x_k + \in$$

- And the estimated response is obtained as

$$\hat{y}_i = b_0 + b_1 x_1 + b_2 x_2 + b_k x_k$$

Multiple Linear Regression

Estimating the coefficients

Let the data points given to us is

$$(x_{1i},\ x_{2i},\ x_{3i}, \dots\dots\dots\dots\dots\dots, x_{ki}, y_i) \quad i = 1,2, \dots\dots\dots, n, \qquad n > k$$

where y_i is the observed response to the values $x_{1i},\ x_{2i},\ x_{3i}, \dots\dots\dots\dots\dots\dots, x_{ki}$ of k independent variables $x_1,\ x_2,\ x_3, \dots\dots\dots\dots\dots\dots, x_k$.

Thus,

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_k x_{ki} + \in_i$$

and

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + b_k x_{ki} + e_i$$

where $\in_i$ and e_i are the random error and residual error, respectively associated with true response y_i and fitted response $\hat{y}_i$.

Using the concept of **Least Square Method** to estimate $b_0, b_1, b_2, \dots, b_k$, we minimize the expression

$$\text{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Multiple Linear Regression

- Differentiating SSE in turn with respect to $b_0, b_1, b_2, \ldots, b_k$ and equating to zero, we generate the set of $(k+1)$ normal estimation equations for multiple linear regression.

$$\mathrm{n b_0} + \mathrm{b_1} \sum_{i=1}^{n} x_{1i} + \mathrm{b_2} \sum_{i=1}^{n} x_{2i} + \ldots + \mathrm{b_k} \sum_{i=1}^{n} x_{\mathrm{k}i} = \sum_{i=1}^{n} y_i$$

$$\mathrm{b_0} \sum_{i=1}^{n} x_{1i} + \mathrm{b_1} \sum_{i=1}^{n} x_{1\mathrm{i}}^2 + \mathrm{b_2} \sum_{i=1}^{n} x_{1i}.x_{2i} + \ldots + \mathrm{b_k} \sum_{i=1}^{n} x_{1i}.x_{\mathrm{k}i} = \sum_{i=1}^{n} x_i.y_i$$

$$\ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots$$

$$\ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots$$

$$\mathrm{b_0} \sum_{i=1}^{n} x_{\mathrm{k}i} + \mathrm{b_1} \sum_{i=1}^{n} x_{\mathrm{k}i}.x_{1i} + \mathrm{b_2} \sum_{i=1}^{n} x_{\mathrm{k}i}.x_{2i} + \ldots + \mathrm{b_k} \sum_{i=1}^{n} \mathrm{x}_{ki}^2 = \sum_{i=1}^{n} x_i.y_i$$

- The system of linear equations can be solved for $b_0, b_1, \ldots, b_k$ by any appropriate method for solving system of linear equations.
- Hence, the multiple linear regression model can be built.

Non Linear Regression Model

- When the regression equation is in terms of *r*-degree, *r*>1, then it is called nonlinear regression model. When more than one independent variables are there, then it is called Multiple Non linear Regression model. Also, alternatively termed as polynomial regression model. In general, it takes the form

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_r x^r + \epsilon$$

- The estimated response is obtained as

$$\hat{y} = b_0 + b_1 x + b_2 x^2 + \ldots + b_r x^r$$

Solving for Polynomial Regression Model

Given that (x_i, y_i); $i = 1,2,\ldots,n$ are n pairs of observations. Each observations would satisfy the equations:

$$y_i = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_r x^r + \epsilon_i$$

and

$$\hat{y}_i = b_0 + b_1 x + b_2 x^2 + \ldots + b_r x^r + e_i$$

where, r is the degree of polynomial

ϵ_i = is the i^{th} random error

e_i = is the i^{th} residual error

Note: The number of observations, n, must be at least as large as $r+1$, the number of parameters to be estimated.

The polynomial model can be transformed into a general linear regression model setting $x_1 = x, x_2 = x^2, \ldots, x_n = x^r$. Thus, the equation assumes the form:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_r x^r + \epsilon_i$$

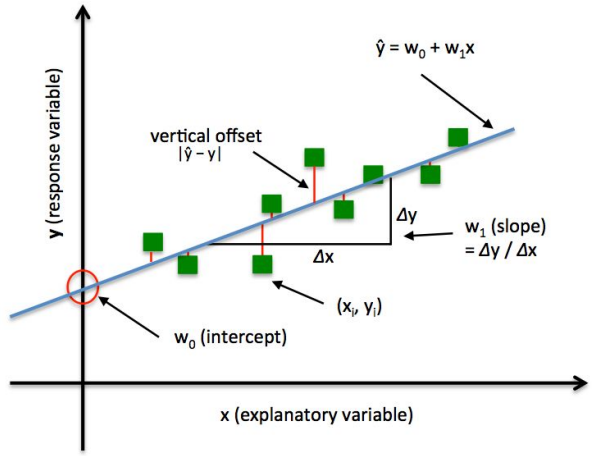
$$\hat{y}_i = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_r x_r + e_i$$

This model then can be solved using the procedure followed for multiple linear regression model.

LRM

Perform regression using one of the following approaches:

- Simple linear regression (SLR) ordinary least squares (OLS) non-iterative method
- Using an optimization algorithm



- Simple linear regression (SLR) ordinary least squares (OLS) non-iterative method

$$observed\ data \rightarrow \quad y = b_0 + b_1 x + \varepsilon$$

$$predicted\ data \rightarrow \quad y' = b_0 + b_1 x$$

$$error \rightarrow \quad \varepsilon = y - y'$$

$$minimize \sum (y - \bar{y})^2$$

LRM

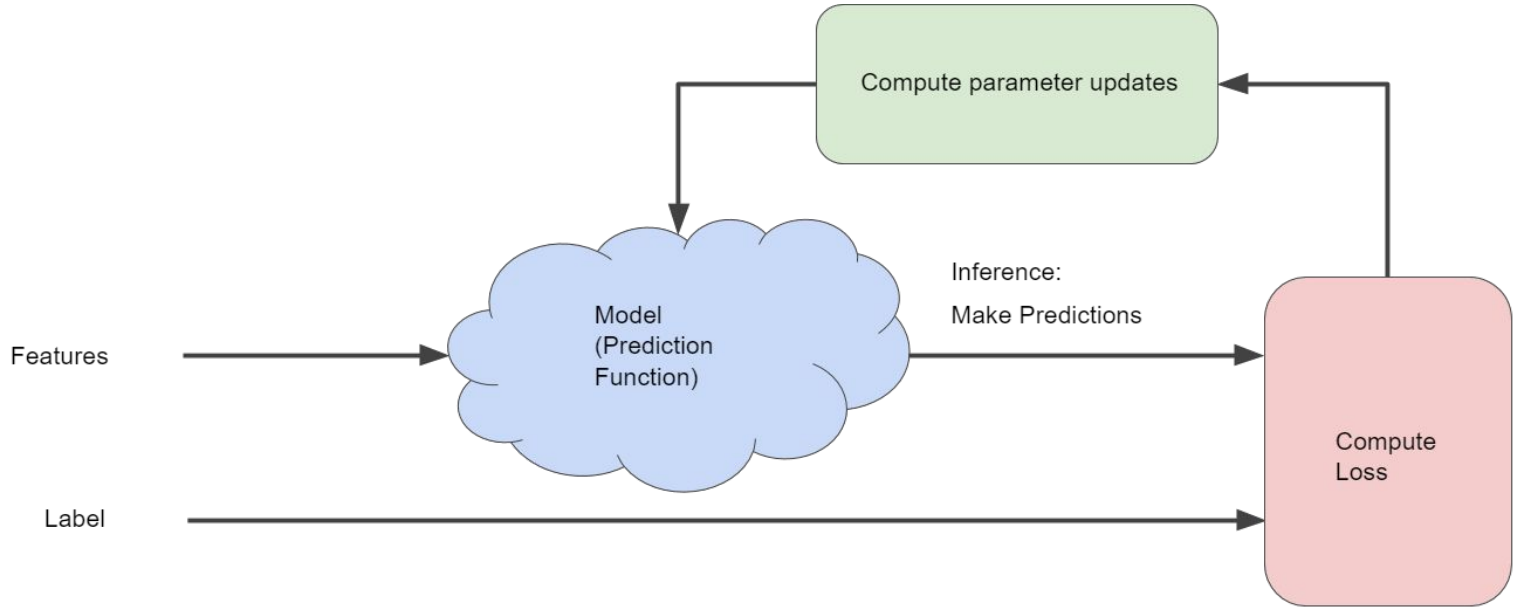
$$y = w_0 x_0 + w_1 x_1 + \ldots + w_m x_m = \sum_{j=0}^{m} \mathbf{w}^\top \mathbf{x}$$

Using an optimization algorithm

where *y* is the response variable, ***x*** is an *m*-dimensional sample vector, and ***w*** is the weight vector (vector of coefficients). Note that w_0 represents the y-axis intercept (bias) of the model and therefore $x_0=1$

Gradient Descent (GD)

Using the Gradient Descent (GD) optimization algorithm, the weights are updated incrementally after each epoch (= pass over the training dataset).



LRM : Gradient Descent (GD)

Using an optimization algorithm

The cost function $J(\cdot)$, the sum of squared errors (SSE), can be written as:

$$J(\mathbf{w}) = \frac{1}{2}\sum_i (\text{target}^{(i)} - \text{output}^{(i)})^2$$

The magnitude and direction of the weight update is computed by taking a step in the opposite direction of the cost gradient

$$\Delta w_j = -\eta \frac{\partial J}{\partial w_j},$$

where η is the learning rate

LRM : Gradient Descent (GD)

The weights are then updated after each epoch via the following update rule:

$$\mathbf{w} := \mathbf{w} + \Delta\mathbf{w},$$

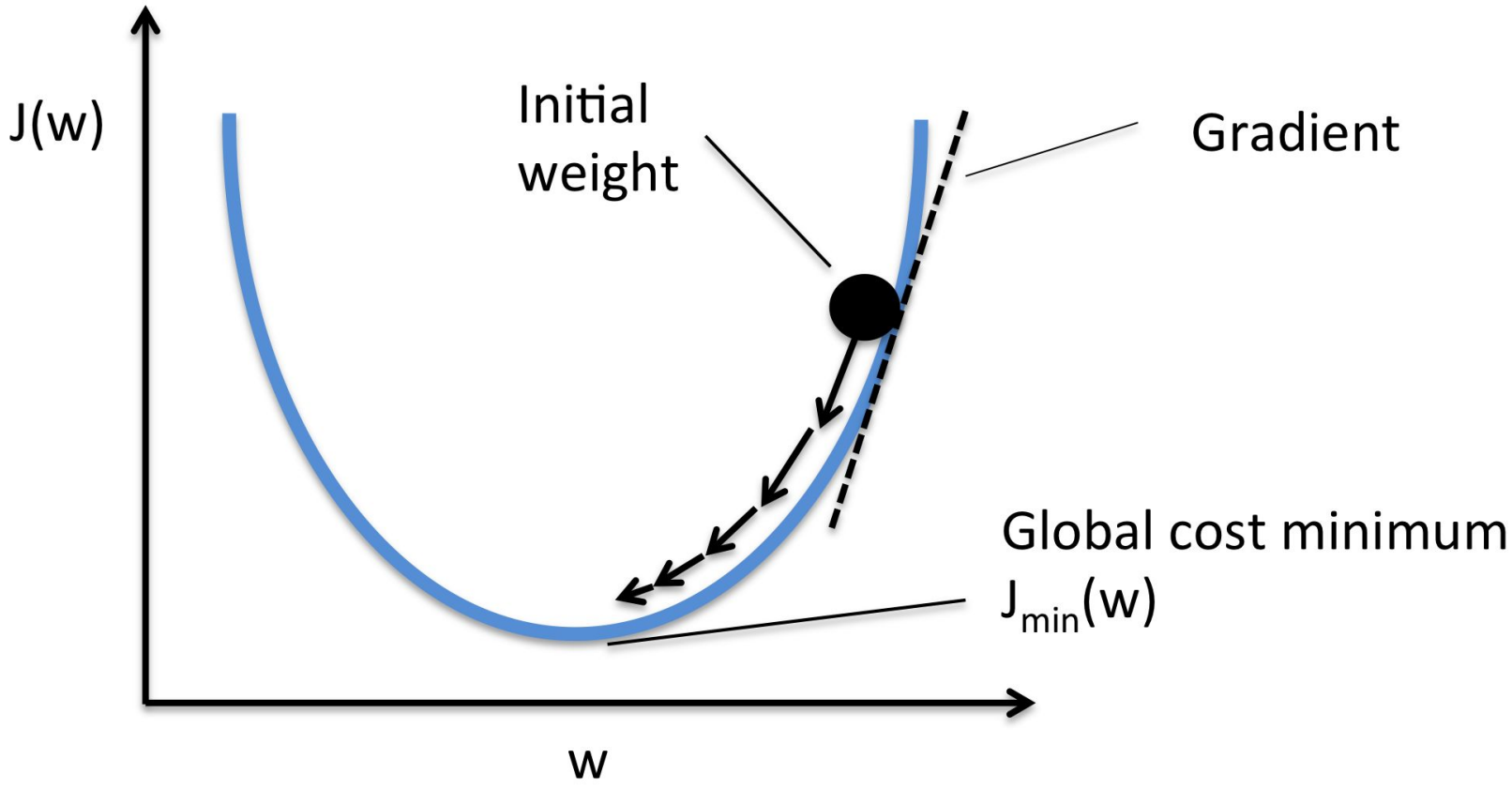
where **Δw** is a vector that contains the weight updates of each weight coefficient *w*, which are computed as follows:

$$\Delta w_j = -\eta \frac{\partial J}{\partial w_j}$$

$$= -\eta \sum_i (\text{target}^{(i)} - \text{output}^{(i)})(-x_j^{(i)})$$

$$= \eta \sum_i (\text{target}^{(i)} - \text{output}^{(i)}) x_j^{(i)}.$$

LRM : Gradient Descent (GD)



cost function with only a single weight coefficient

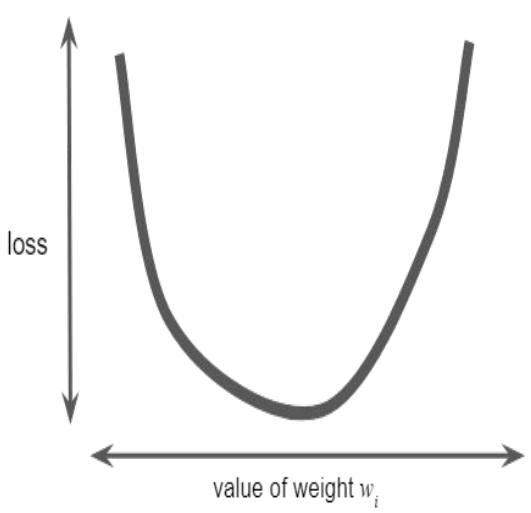
LRM：Gradient Descent (GD)

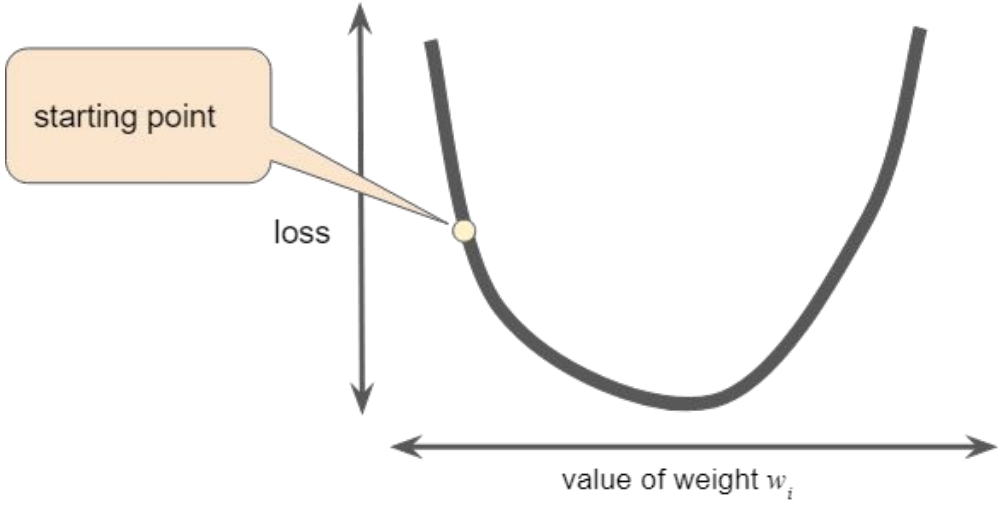
Regression problems yield convex loss vs. weight plots y'=w0+w1x

Convex problems have only one minimum; that is, only one place where the slope is exactly 0. That minimum is where the loss function converges.

The first stage in gradient descent is to pick a starting value (a starting point) for w1.
simply set w1 to 0 or pick a random value
calculate the gradient of the loss curve at the starting point
gradient of the loss is equal to the derivative (slope) of the curve

$$\text{slope} = \frac{\text{change in } y}{\text{change in } x}$$



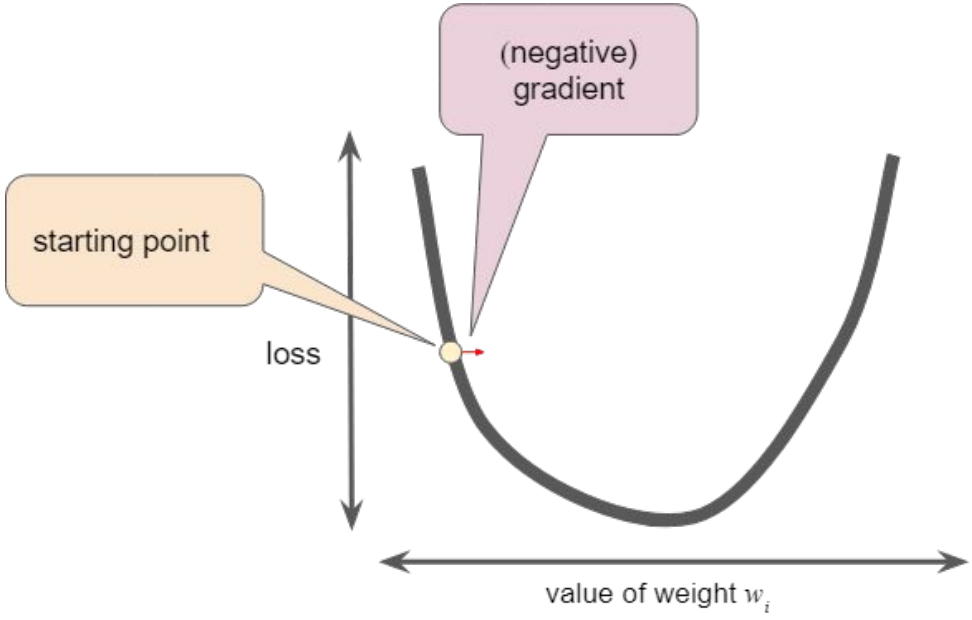


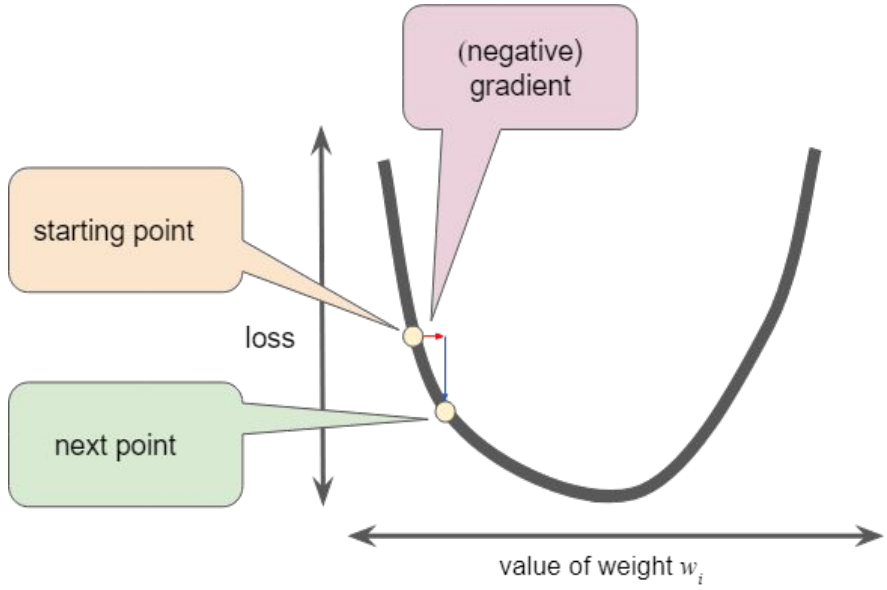
ref:https://developers.google.com/machine-learning/crash-course/reducing-loss/gradient-descent

LRM : Gradient Descent (GD)

Regression problems yield convex loss vs. weight plots

The gradient descent algorithm takes a step in the direction of the negative gradient in order to reduce loss as quickly as possible.





To determine the next point along the loss function curve, the gradient descent algorithm adds some fraction of the gradient's magnitude to the starting point

The gradient descent then repeats this process, edging ever closer to the minimum.

LRM : Gradient Descent (GD)

y'=w0+w1x

When performing gradient descent, generalize the process to tune all the model parameters simultaneously.

calculate the gradients with respect to both w1 and w0

Next, modify the values of w1 and w0 based on their respective gradients.

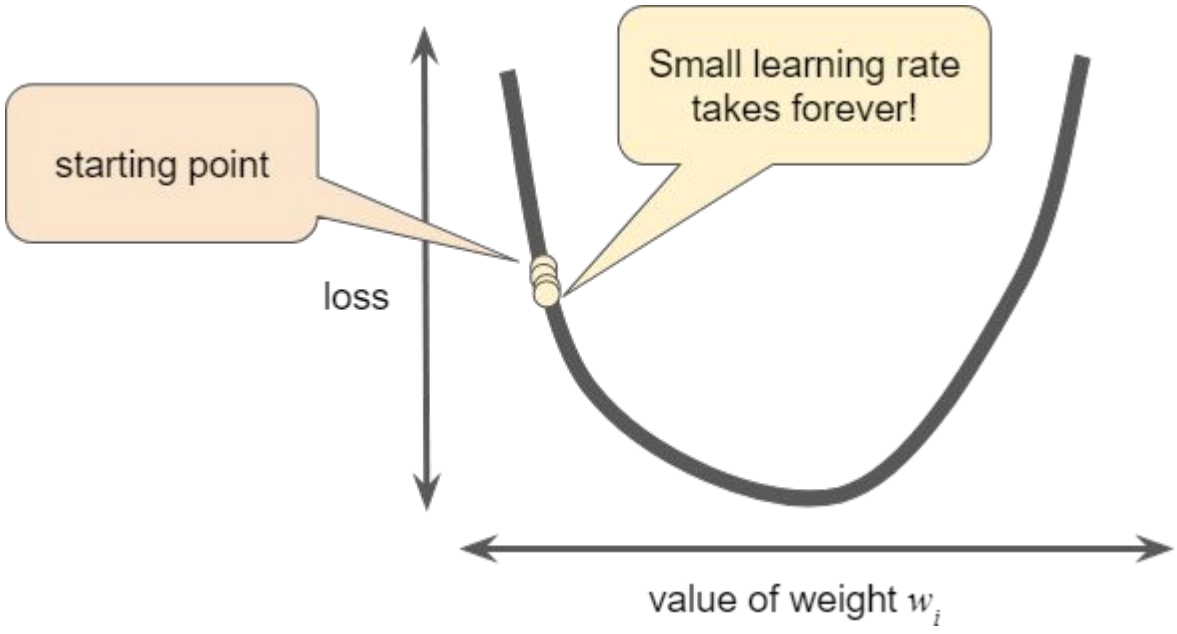
Then repeat these steps until we reach minimum loss

LRM : Gradient Descent (GD)

Reducing Loss: Learning Rate

the gradient vector has both a direction and a magnitude. Gradient descent algorithms multiply the gradient by a scalar known as the **learning rate** (also sometimes called **step size**) to determine the next point

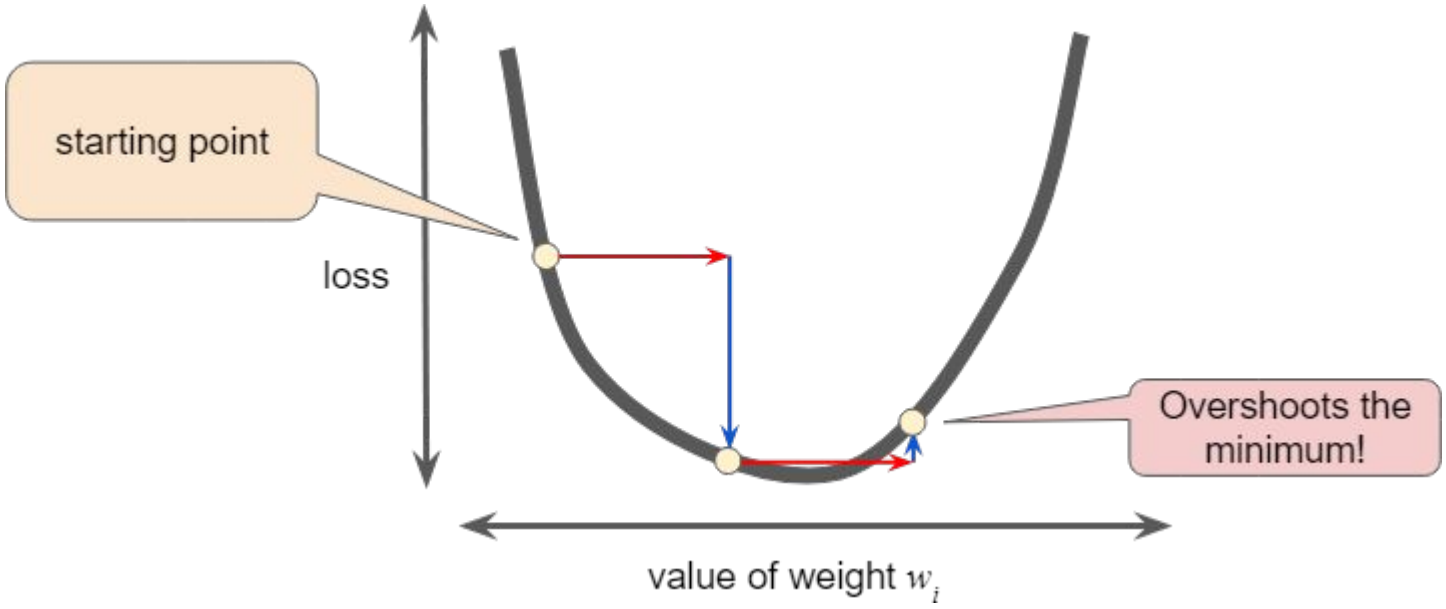
For example, if the gradient magnitude is 2.5 and the learning rate is 0.01, then the gradient descent algorithm will pick the next point 0.025 away from the previous point.



a learning rate that is too small, learning will take too long:

LRM : Gradient Descent (GD)

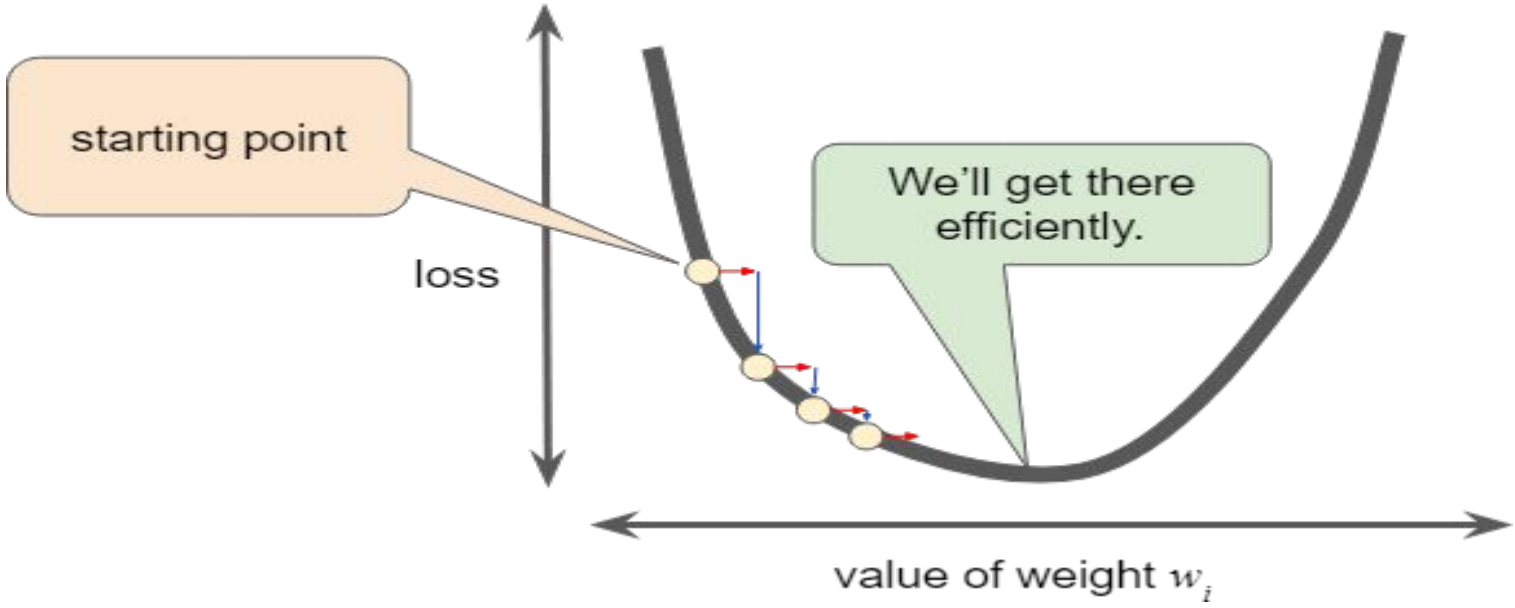
Reducing Loss: Learning Rate



a learning rate that is too large, the next point will perpetually bounce haphazardly across the bottom of the well

LRM : Gradient Descent (GD)

Reducing Loss: Learning Rate



Learning rate is just right.

DEMO

LRM : Stochastic Gradient Descent (SGD)

In gradient descent, a **batch** is the total number of examples you use to calculate the gradient in a single iteration

a batch can be enormous. A very large batch may cause even a single iteration to take a very long time to compute.

Stochastic gradient descent (**SGD**) --it uses only a single example (a batch size of 1) per iteration. The term "stochastic" indicates that the one example comprising each batch is chosen at random.

SGD; sometimes also referred to as *iterative* or *on-line* GD , we **don't** accumulate the weight updates

- for one or more epochs:
  - for each weight j
    - $w_j := w + \Delta w_j$, where: $\Delta w_j = \eta \sum_i (\text{target}^{(i)} - \text{output}^{(i)}) x_j^{(i)}$

← GD

SGD →

- for one or more epochs, or until approx. cost minimum is reached:
  - for training sample i:
    - for each weight j
      - $w_j := w + \Delta w_j$, where: $\Delta w_j = \eta(\text{target}^{(i)} - \text{output}^{(i)}) x_j^{(i)}$

LRM : Stochastic Gradient Descent (SGD)

There are several different flavors of SGD. Let's take a look at the three most common variants:

A)

- randomly shuffle samples in the training set
  - for one or more epochs, or until approx. cost minimum is reached
    - for training sample *i*
      - compute gradients and perform weight updates

B)

- for one or more epochs, or until approx. cost minimum is reached
  - randomly shuffle samples in the training set
    - for training sample *i*
      - compute gradients and perform weight updates

C)

- for iterations *t*, or until approx. cost minimum is reached:
  - draw random sample from the training set
    - compute gradients and perform weight updates

In scenario A, shuffle the training set only one time in the beginning; whereas in scenario B, shuffle the training set after each epoch to prevent repeating update cycles.

In both scenario A and scenario B, each training sample is only used once per epoch to update the model weights

In scenario C, draw the training samples randomly with replacement from the training set. If the number of iterations *t* is equal to the number of training samples, learn the model based on a *bootstrap sample* of the training set.

LRM : Mini-Batch Gradient Descent (MB-GD)

Mini-Batch Gradient Descent (MB-GD) a compromise between batch GD and SGD.

In MB-GD, update the model based on smaller groups of training samples; instead of computing the gradient from 1 sample (SGD) or all *n* training samples (GD)

compute the gradient from $1 < k < n$ training samples (typically between 10 and 1,000 examples, chosen at random; a common mini-batch size is $k=50$).

MB-GD converges in fewer iterations than GD because we update the weights more frequently

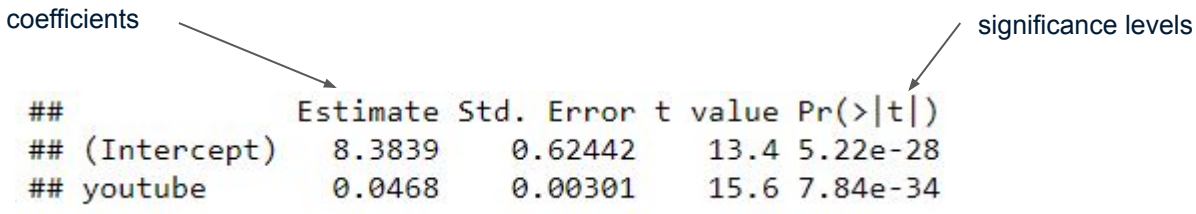
MB-GD utilize vectorized operation, which typically results in a computational performance gain over SGD.

Simple Linear Regression

build a simple linear model to predict sales units based on the advertising budget spent on youtube

```
sales = w0 + w1*youtube
```

```
model <- lm(sales ~ youtube, data = train.data)
summary(model)$coef
```




```
sales = 8.38 + 0.046*youtube
```

for each new youtube advertising budget, predict the number of sale units.

For example:

- For a youtube advertising budget equal zero, we can expect a sale of 8.38 units.
- For a youtube advertising budget equal 1000, we can expect a sale of 8.38 + 0.046*1000 = 55 units.

Simple Linear Regression

The level of statistical significance is often expressed as a *p*-value between 0 and 1.

The smaller the p-value, the stronger the evidence that you should reject the null hypothesis.

Null hypothesis: the predictor is not meaningful model

- A *p*-value less than 0.05 (typically ≤ 0.05) is statistically significant. It indicates strong evidence against the null hypothesis, as there is less than a 5% probability the null is correct.. Therefore, reject the null hypothesis, and accept the alternative hypothesis.

  However, this does not mean that there is a 95% probability that the hypothesis is true

- A *p*-value higher than 0.05 (> 0.05) is not statistically significant and indicates strong evidence for the null hypothesis. This means retain the null hypothesis and reject the alternative hypothesis

Multiple Linear Regression

Build a model to predict sales based on the budget invested in three advertising medias: youtube, facebook and newspaper.

sales = w0 + w1*youtube + w2*facebook + w3*newspaper

```
model <- lm(sales ~ youtube + facebook + newspaper,  data = train.data)
summary(model)$coef
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.39188    0.44062   7.698 1.41e-12
## youtube      0.04557    0.00159  28.630 2.03e-64
## facebook     0.18694    0.00989  18.905 2.07e-42
## newspaper    0.00179    0.00677   0.264 7.92e-01
```

- Estimate: the intercept and the coefficient estimates associated to each predictor variable
- Std.Error: the standard error of the coefficient estimates. This represents the accuracy of the coefficients. The larger the standard error, the less confident we are about the estimate.
- t value: the t-statistic, which is the coefficient estimate (column 2) divided by the standard error of the estimate (column 3)
- Pr(>|t|): The p-value corresponding to the t-statistic. The smaller the p-value, the more significant the estimate is.

Multiple Linear Regression

Before using a model for predictions,
assess the statistical significance of the model

```
summary(model)
```

```
##
## Call:
## lm(formula = sales ~ ., data = train.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.412  -1.110   0.348   1.422   3.499
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.39188    0.44062    7.70  1.4e-12 ***
## youtube      0.04557    0.00159   28.63  < 2e-16 ***
## facebook     0.18694    0.00989   18.90  < 2e-16 ***
## newspaper    0.00179    0.00677    0.26     0.79
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.12 on 158 degrees of freedom
## Multiple R-squared:  0.89,   Adjusted R-squared:  0.888
## F-statistic:  427 on 3 and 158 DF,  p-value: <2e-16
```

The p-value for newspaper is 0.79. In other words, there's 79% chance that this predictor is not meaningful for the regression.

The summary outputs shows 6 components, including:

Call. Shows the function call used to compute the regression model.

Residuals. Provide a quick view of the distribution of the residuals, which by definition have a mean zero. Therefore, the median should not be far from zero, and the minimum and maximum should be roughly equal in absolute value.

Coefficients. Shows the regression coefficients and their statistical significance. Predictor variables, that are significantly associated to the outcome variable, are marked by stars.

Residual standard error (RSE), **R-squared** (R2) and the **F-statistic** are metrics that are used to check how well the model fits to data.

Multiple Linear Regression

The first step in interpreting the multiple regression analysis is to examine the F-statistic and the associated p-value, at the bottom of model summary.

```
## F-statistic:  427 on 3 and 158 DF,  p-value: <2e-16
```

p-value of the F-statistic is < 2.2e-16, which is highly significant. This means that, at least, one of the predictor variables is significantly related to the outcome variable

The model with zero predictor variables is also called **"Intercept Only Model"**.
F – Test for overall significance compares a intercept only regression model with the current model

The Hypothesis for F-Test for significance can be constructed as –

H0 : The fit of intercept only model and the current model is same. i.e. Additional variables do not provide value taken together

Ha : The fit of intercept only model is significantly less compared to our current model. i.e. Additional variables do make the model significantly better.

Multiple Linear Regression

```
## F-statistic:  427 on 3 and 158 DF,  p-value: <2e-16
```

$$F = \frac{explained\ variation/(k-1)}{unexplained\ variation/(n-k)}$$

$$F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}$$

n = 162 (Total number of observations)

k = 4 (no of variables + 1 for intercept)

So degrees of freedom that we get are

DF Numerator = (k-1) = 3 – Matches with our DF

DF Denominator = (n-k) = (162 – 4) = 158– Matches with our DF

P Value of F Statistic 427 for DF 3 and 158 is extremely small, i.e smaller that 0.001 so we can reject H0 and say that overall addition of variables is significantly improving the model.

Which in a way implies that by adding those extra variables we were able to improve the fit of our model significantly.

Multiple Linear Regression

How is F Statistic different from R Squared ?

R squared provides a measure of strength of relationship between predictors and response variable and it does not comment on whether the relationship is statistically significant.

F Statistic gives us a power to judge whether that relationship is statistically significant in other words it comments on whether or R^2 is significant or not.

What should we do with F statistic in Regression model ?

- If F-statistic is significant that gives us extra confidence on the R^2 value that we have got .
- In case we get insignificant F-Statistic or if p values for F are greater that level of significance (say 0.05 or 0.01) then personally we would stay away from that model since we will not be able to confidently comment on the R^2 values

A large F-statistic will corresponds to a statistically significant p-value ($p < 0.05$)

Multiple Linear Regression

Coefficients significance

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.39188    0.44062   7.698 1.41e-12
## youtube      0.04557    0.00159  28.630 2.03e-64
## facebook     0.18694    0.00989  18.905 2.07e-42
## newspaper    0.00179    0.00677   0.264 7.92e-01
```

To see which predictor variables are significant, examine the coefficients table, which shows the estimate of regression coefficients and the associated t-statistic p-values.

For a given the predictor, the t-statistic evaluates whether or not there is significant association between the predictor and the outcome variable, that is whether the coefficient of the predictor is significantly different from zero.

It can be seen that, changing in youtube and facebook advertising budget are significantly associated to changes in sales while changes in newspaper budget is not significantly associated with sales.

For a given predictor variable, the coefficient can be interpreted as the average effect on y of a one unit increase in predictor, holding all other predictors fixed.

For example, for a fixed amount of youtube and newspaper advertising budget, spending an additional 1 000 dollars on facebook advertising leads to an increase in sales by approximately 0.1869*1000 = 187 sale units, on average.

newspaper is not significant in the multiple regression model

This means that, for a fixed amount of youtube and newspaper advertising budget, changes in the newspaper advertising budget will not significantly affect sales units.

As the newspaper variable is not significant, it is possible to remove it from the model:

Multiple Linear Regression

Coefficients significance

```
model <- lm(sales ~ youtube + facebook, data = train.data)
summary(model)
```

```
##
## Call:
## lm(formula = sales ~ youtube + facebook, data = train.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.481  -1.104   0.349   1.423   3.486
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.43446    0.40877     8.4  2.3e-14 ***
## youtube      0.04558    0.00159    28.7  < 2e-16 ***
## facebook     0.18788    0.00920    20.4  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.11 on 159 degrees of freedom
## Multiple R-squared:  0.89,   Adjusted R-squared:  0.889
## F-statistic:  644 on 2 and 159 DF,  p-value: <2e-16
```

✔ Finally, our model equation can be written as follow: `sales = 3.43+ 0.045youtube + 0.187facebook`.

Multiple Linear Regression

Model accuracy (*goodness-of-fit)*

Once you identified that, at least, one predictor variable is significantly associated to the outcome, continue the diagnostic by checking how well the model fits the data.

The overall quality of the linear regression fit can be assessed using the following three quantities:

1. Residual Standard Error (RSE),
2. R-squared (R2) and adjusted R2,
3. F-statistic

```
## Residual standard error: 2.11 on 159 degrees of freedom
## Multiple R-squared:  0.89,   Adjusted R-squared:  0.889
## F-statistic:  644 on 2 and 159 DF,  p-value: <2e-16
```

1. Residual standard error (RSE).

The RSE corresponding to the prediction error, represents roughly the average difference between the observed outcome values and the predicted values by the model. The lower the RSE the best the model fits to our data.

Dividing the RSE by the average value of the outcome variable will give you the prediction error rate, which should be as small as possible.

In our example, using only youtube and facebook predictor variables, the RSE = 2.11, meaning that the observed sales values deviate from the predicted values by approximately 2.11 units in average.

This corresponds to an error rate of 2.11/mean(train.data$sales) = 2.11/16.77 = 13%, which is low.

Multiple Linear Regression

```
## Residual standard error: 2.11 on 159 degrees of freedom
## Multiple R-squared:  0.89,   Adjusted R-squared:  0.889
## F-statistic:  644 on 2 and 159 DF,  p-value: <2e-16
```

2. R-squared and Adjusted R-squared:

The R-squared ranges from 0 to 1 and represents the proportion of variation in the outcome variable that can be explained by the model predictor variables.

a regression equation explain 89 % variation of observed values around mean

For a simple linear regression, R-squared is the square of the Pearson correlation coefficient between the outcome and the predictor variables.

In multiple linear regression, the R-squared represents the correlation coefficient between the observed outcome values and the predicted values.

The higher the better the model??

However, a problem with the R-squared, is that, it will always increase when more variables are added to the model, even if those variables are only weakly associated with the outcome .

A solution is to adjust the R-squared by taking into account the number of predictor variables.

Multiple Linear Regression

```
## Residual standard error: 2.11 on 159 degrees of freedom
## Multiple R-squared:  0.89,   Adjusted R-squared:  0.889
## F-statistic:  644 on 2 and 159 DF,  p-value: <2e-16
```

2. R-squared and Adjusted R-squared:

The adjustment in the “Adjusted R Squared” value in the summary output is a correction for the number of variables included in the predictive model.

- An adjusted R-squared that is close to 1 indicates that a large proportion of the variability in the outcome has been explained by the regression model.
- A number near 0 indicates that the regression model did not explain much of the variability in the outcome.

In our example, the adjusted R2 is 0.88, which is good.

$$Adjusted\ R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

N = Sample Size

p = number of predictors

Multiple Linear Regression

Is Low R^2 always bad ?
NO.
Desirable range of R^2 is highly domain dependent.
Model to predict Human behavior is seldom very precise and hence lower R^2 is expected.
Where as for models in medicine and pharma R^2 values above 90% are very common.

Is High R^2 always good ?
NO.
If you have High R^2 but you have some inherent Residual pattern or if residuals are not normally distributed then the model is not considered good enough.

Multiple Linear Regression

3. F-Statistic:

F-statistic gives the overall significance of the model. It assess whether at least one predictor variable has a non-zero coefficient.

In a simple linear regression, this test is not really interesting since it just duplicates the information given by the t-test, available in the coefficient table.

The F-statistic becomes more important with **multiple predictors** as in multiple linear regression.

A large F-statistic will corresponds to a statistically significant p-value ($p < 0.05$).

In our example, the F-statistic equal 644 producing a p-value of 1.46e-42, which is highly significant.

```
## Residual standard error: 2.11 on 159 degrees of freedom
## Multiple R-squared:  0.89,   Adjusted R-squared:  0.889
## F-statistic:  644 on 2 and 159 DF,  p-value: <2e-16
```

Multiple Linear Regression

Making predictions

Make predictions using the test data in order to evaluate the performance of our regression model.

The procedure is as follow:

1. Predict the sales values based on new advertising budgets in the test data
2. Assess the model performance by computing:

   The prediction error RMSE (Root Mean Squared Error): representing the average difference between the observed known outcome values in the test data and the predicted outcome values by the model. The lower the RMSE, the better the model.

   The R-squared:, representing the correlation between the observed outcome values and the predicted outcome values. The higher the R-squared, the better the model.

Multiple Linear Regression

```
# Make predictions
predictions <- model %>% predict(test.data)
# Model performance
# (a) Compute the prediction error, RMSE
RMSE(predictions, test.data$sales)
```

```
## [1] 1.58
```

```
# (b) Compute R-square
R2(predictions, test.data$sales)
```

```
## [1] 0.938
```

From the output above, the R2 is 0.93, meaning that the observed and the predicted outcome values are highly correlated, which is very good.

The prediction error RMSE is 1.58, representing an error rate of 1.58/mean(test.data$sales) = 1.58/17 = 9.2%, which is good.

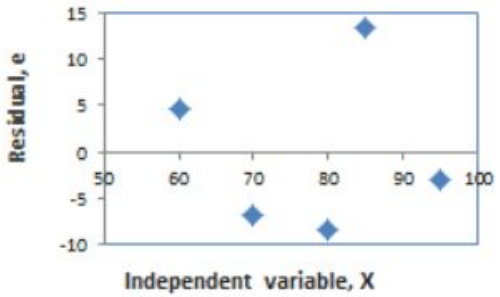
Residual Plots

A **residual plot** is a graph that shows the residuals on the vertical axis and the independent variable on the horizontal axis. If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a nonlinear model is more appropriate.

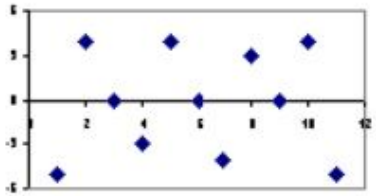
The table below shows inputs and outputs from a simple linear regression analysis.

x	y	$\hat{y}$	e
60	70	65.411	4.589
70	65	71.849	-6.849
80	70	78.288	-8.288
85	95	81.507	13.493
95	85	87.945	-2.945

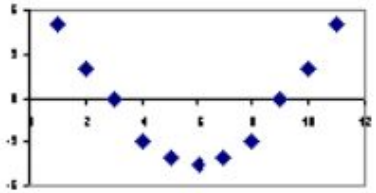
And the chart below displays the residual (e) and independent variable (X) as a residual plot.



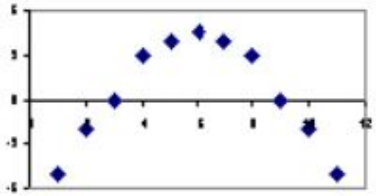
Below, the residual plots show three typical patterns. The first plot shows a random pattern, indicating a good fit for a linear model.



Random pattern



Non-random: U-shaped



Non-random: Inverted U

The other plot patterns are non-random (U-shaped and inverted U), suggesting a better fit for a nonlinear model.

Reference

- The detail material related to this lecture can be found in

  The Elements of Statistical Learning, Data Mining, Inference, and Prediction (2nd Edn.), Trevor Hastie, Robert Tibshirani, Jerome Friedman, Springer, 2014.