Correlation

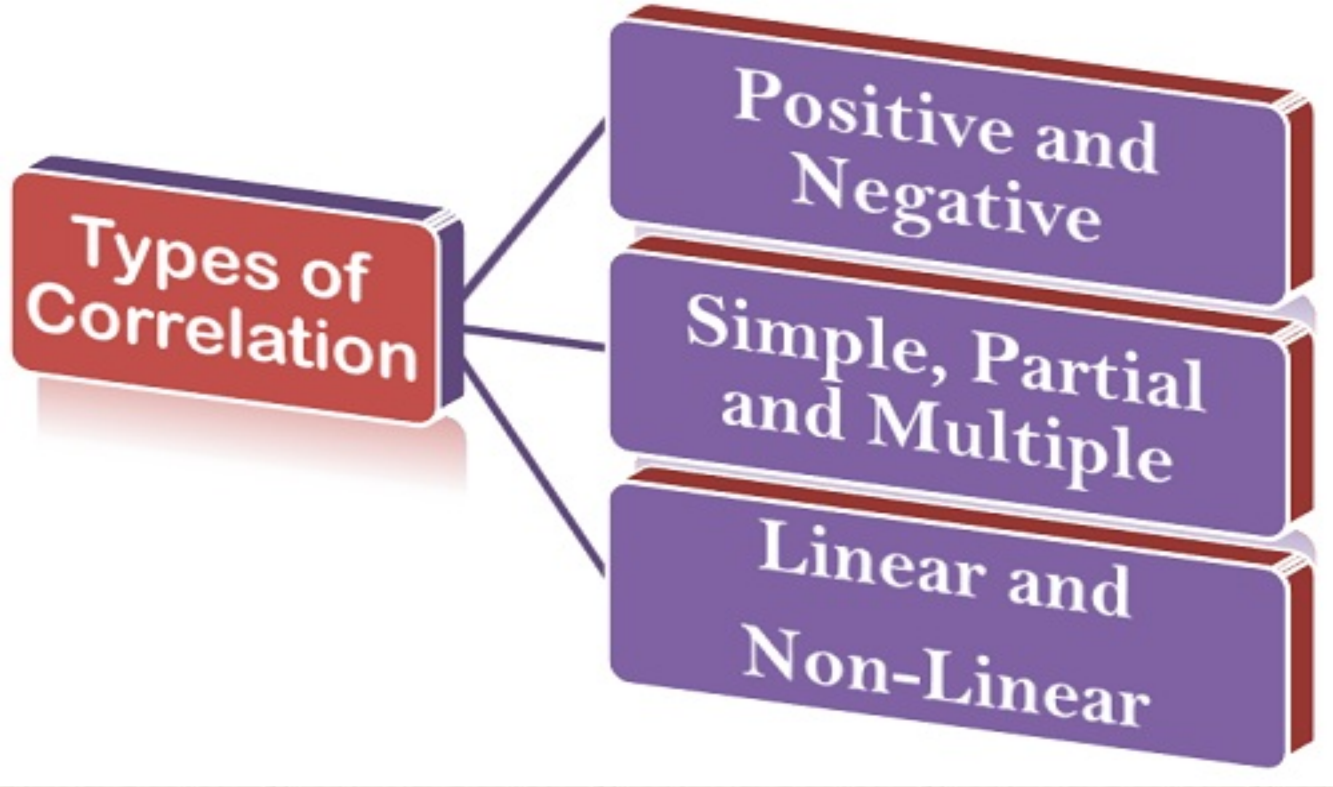
Learning Objectives

After completing this chapter, you should be able to

- Understand the different types of correlation
- Test hypotheses and construct confidence intervals on the Correlation coefficients.
- Learn about the different regression types in machine learning, including linear and logistic regression
- Understand how the method of least squares extends to fitting multiple regression models.
- Build regression models with polynomial terms.
- Understand how the gradient descent optimization works

Correlation

A **correlation** is a tool used to measure relationship between two or more variables. The data can be represented by the ordered pairs (x, y) where x is the **independent** (or **explanatory**) **variable**, and y is the **dependent** (or **response**) **variable**.



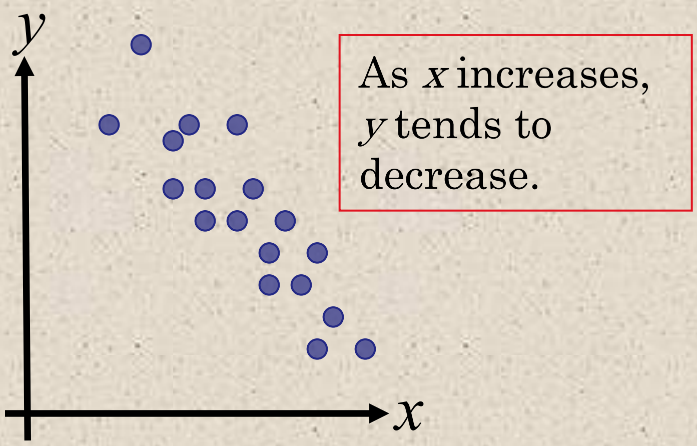
Types of Correlation

Positive and Negative

Simple, Partial and Multiple

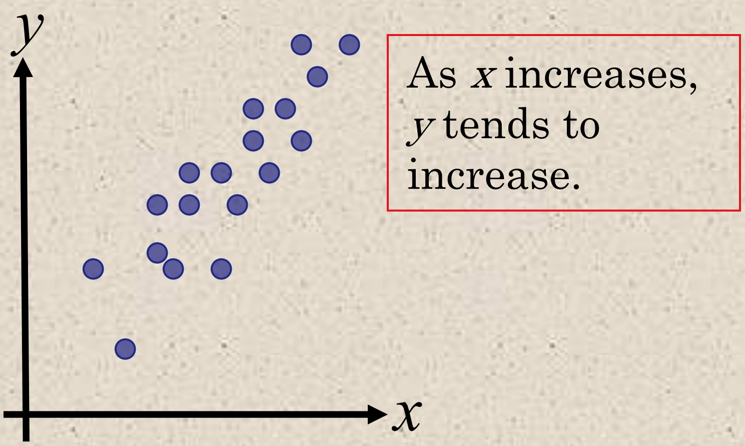
Linear and Non-Linear

Types of Correlation



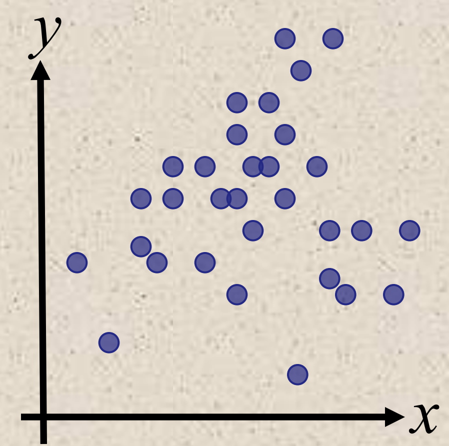
As *x* increases, *y* tends to decrease.

Negative Linear Correlation

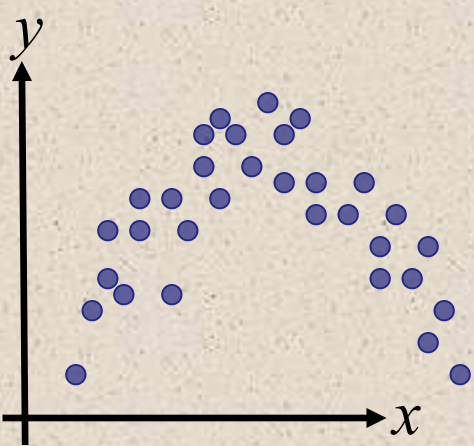


As *x* increases, *y* tends to increase.

Positive Linear Correlation

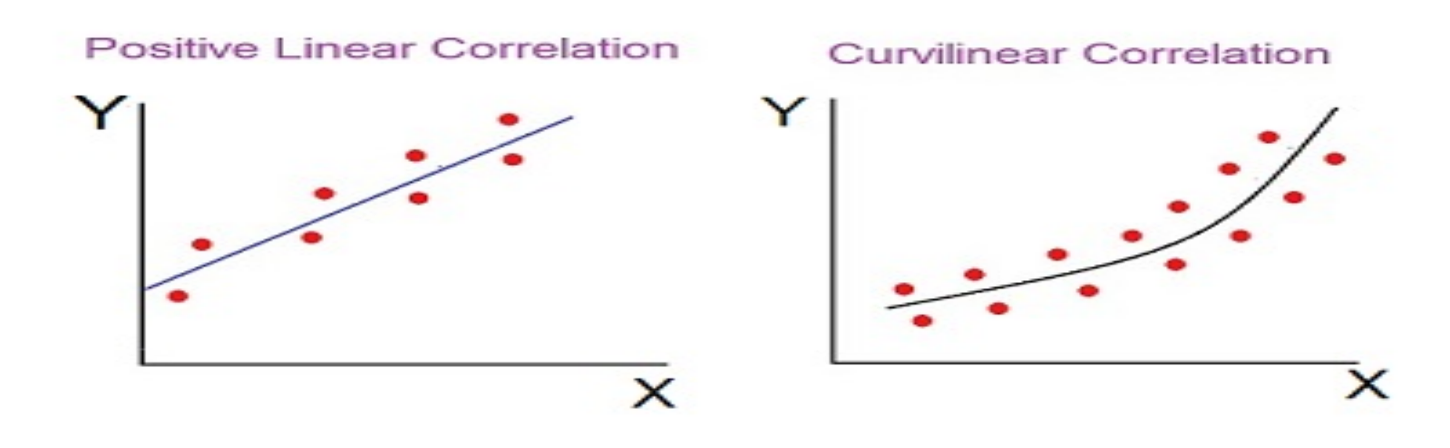


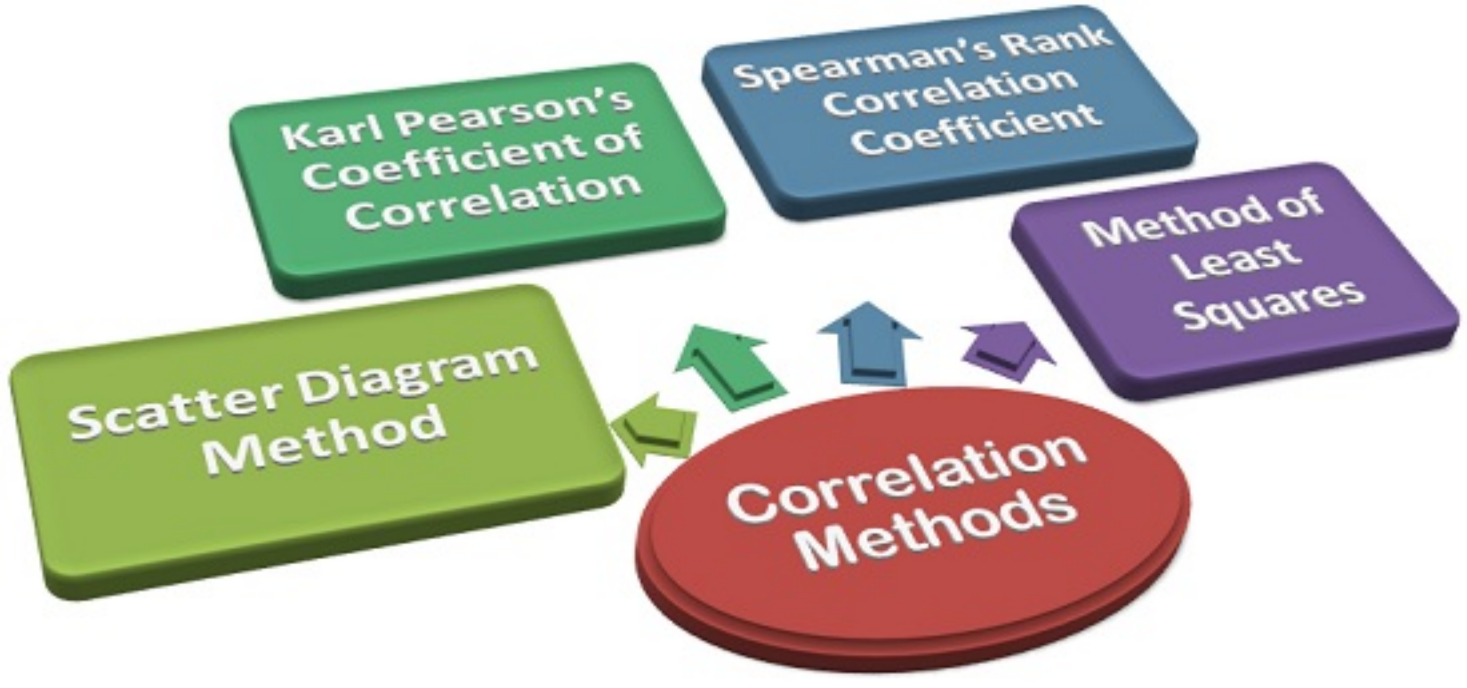
No Correlation



Nonlinear Correlation

Methods of Determining Correlation





Karl Pearson's Coefficient of Correlation

Spearman's Rank Correlation Coefficient

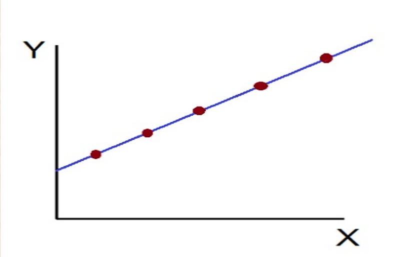
Method of Least Squares

Scatter Diagram Method

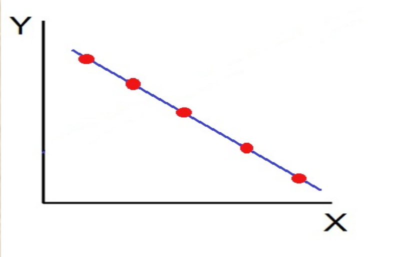
Correlation Methods

Scatter Diagram Method

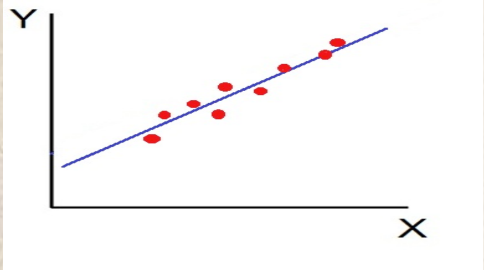
The **Scatter Diagram Method** is the simplest method to study the correlation between two variables wherein the values for each pair of a variable is plotted on a graph in the form of dots thereby obtaining as many points as the number of observations.



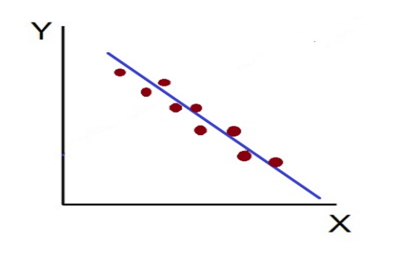
Perfect Positive Correlation(r-=+1)



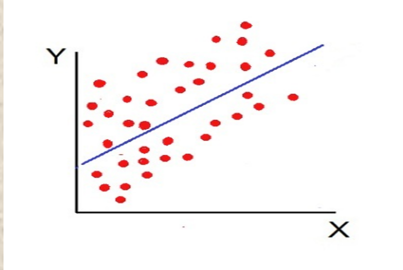
Perfect Negative Correlation(r-=-1)



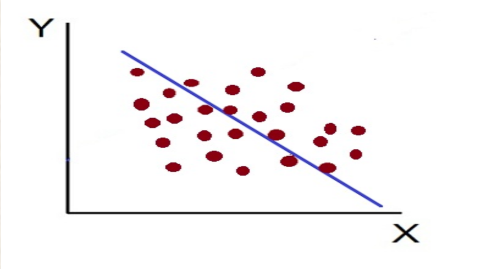
High Degree of +ve Correlation(r-=+ High)



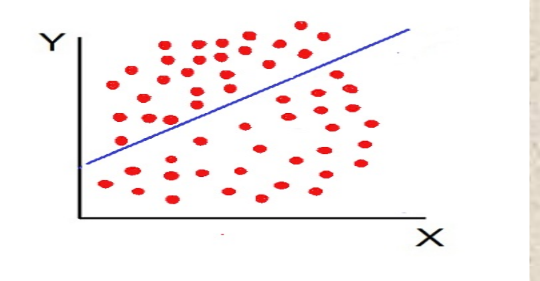
High Degree of -ve Correlation(r-=- High)



low Degree of +ve Correlation(r-=+ low)



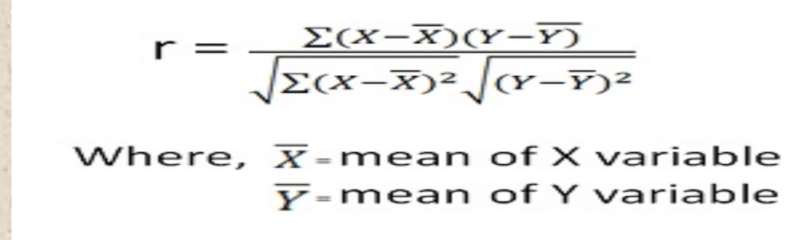
low Degree of -ve Correlation(r-=- low)



No Correlation (r=0)

Karl Pearson’s Coefficient of Correlation

The **Karl Pearson’s Coefficient of Correlation** is widely use mathematical method wherein the numerical expression is used to calculate the degree and direction of the relationship between linear related variables.



$$r = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{\sqrt{\Sigma(X-\overline{X})^2}\sqrt{(Y-\overline{Y})^2}}$$

Where, $\overline{X}$ = mean of X variable
$\overline{Y}$ = mean of Y variable

Properties of Coefficient of Correlation:

- The value of the coefficient of correlation (r) always lies between **±1**. Such as:
    - r=+1, perfect positive correlation
    - r=-1, perfect negative correlation
    - r=0, no correlation
- The coefficient of correlation is independent of the **origin and scale.** By origin, it means subtracting any non-zero constant from the given value of X and Y the vale of “r” remains unchanged. By scale it means, there is no effect on the value of “r” if the value of X and Y is divided or multiplied by any constant.

Spearman’s Rank Correlation Coefficient

The Spearman’s Rank Correlation Coefficient is the non-parametric statistical measure used to study the strength of association between the two ranked variables.

$$R = \frac{(1 - 6\sum D^2)}{N\,(N^2 - 1)} = \frac{(1 - 6\sum D^2\,)}{N^3\ - N}$$

The value of R lies between ±1 such as:

- R =+1, there is a complete agreement in the order of ranks and move in the same direction.
- R=-1, there is a complete agreement in the order of ranks, but are in opposite directions.
- R =0, there is no association in the ranks.

Equal Ranks or Tie in Ranks: In case the same ranks are assigned to two or more entities, then the ranks are assigned on an average basis. Such as if two individuals are ranked equal at third position, then the ranks shall be calculated as:

$$R = 1 - \frac{6\left\{\sum D^2 \ + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2 + \ldots\ldots\ldots\right\}}{N^3 - N}$$

Where m = number of items whose ranks are common.

Method of Least Squares

The **Method of Least Squares** is another mathematical method that tells the degree of correlation between the variables by using the square root of product of two regression coefficient that of x on y and y on x.

$$r = \sqrt{b_{xy} X b_{yx}}$$

Correlation Coefficient

Example:
The following data represents the number of hours 12 different students watched television during the weekend and the scores of each student who took a test the following Monday.

a.) Display the scatter plot.
b.) Calculate the correlation coefficient *r*.
c.) Spearman's Rank Correlation Coefficient r.
d.) Method of Least Squares

Hours, x	0	1	2	3	3	5	5	5	6	7	7	10
Test score, y	96	85	82	74	95	68	76	84	58	65	75	50

Testing a Population Correlation Coefficient

Once the sample correlation coefficient r has been calculated, we need to determine whether there is enough evidence to decide that the population correlation coefficient ρ is significant at a specified level of significance.

One way to determine this is to use Table.

If $|r|$ is greater than the critical value, there is enough evidence to decide that the correlation coefficient ρ is significant.

n	$\alpha = 0.05$	$\alpha = 0.01$
4	0.950	0.990
5	0.878	0.959
6	0.811	0.917
7	0.754	0.875

For a sample of size $n = 6$, ρ is significant at the 5% significance level, if $|r| > 0.811$.

Testing a Population Correlation Coefficient

Finding the Correlation Coefficient ρ

	In Words	*In Symbols*
1.	Determine the number of pairs of data in the sample.	Determine n.
2.	Specify the level of significance.	Identify α.
3.	Find the critical value.	Use Table 11 in Appendix B.
4.	Decide if the correlation is significant.	If $\lvert r \rvert >$ critical value, the correlation is significant. Otherwise, there is not enough evidence to support that the correlation is significant.
5.	Interpret the decision in the context of the original claim.	

Testing a Population Correlation Coefficient

Example:
The following data represents the number of hours 12 different students watched television during the weekend and the scores of each student who took a test the following Monday.

The correlation coefficient $r \approx -0.831$.

Hours, x	0	1	2	3	3	5	5	5	6	7	7	10
Test score, y	96	85	82	74	95	68	76	84	58	65	75	50

Is the correlation coefficient significant at $\alpha = 0.01$?

Continued.

Testing a Population Correlation Coefficient

Example continued:

$r \approx -0.831$

$n = 12$

$\alpha = 0.01$

Table

n	$\alpha = 0.05$	$\alpha = 0.01$
4	0.950	0.990
5	0.878	0.959
6	0.811	0.917
10	0.632	0.765
11	0.602	0.735
12	0.576	0.708
13	0.553	0.684

$|r| > 0.708$

Because, the population correlation is significant, there is enough evidence at the 1% level of significance to conclude that there is a significant linear correlation between the number of hours of television watched during the weekend and the scores of each student who took a test the following Monday.

Hypothesis Testing for ρ

A hypothesis test can also be used to determine whether the sample correlation coefficient r provides enough evidence to conclude that the population correlation coefficient ρ is significant at a specified level of significance.

A hypothesis test can be one tailed or two tailed.

$$\begin{cases} H_0\text{: } \rho \geq 0 & \text{(no significant negative correlation)} \\ H_a\text{: } \rho < 0 & \text{(significant negative correlation)} \end{cases}$$

Left-tailed test

$$\begin{cases} H_0\text{: } \rho \leq 0 & \text{(no significant positive correlation)} \\ H_a\text{: } \rho > 0 & \text{(significant positive correlation)} \end{cases}$$

Right-tailed test

$$\begin{cases} H_0\text{: } \rho = 0 & \text{(no significant correlation)} \\ H_a\text{: } \rho \neq 0 & \text{(significant correlation)} \end{cases}$$

Two-tailed test

Hypothesis Testing for ρ

The *t*-Test for the Correlation Coefficient

A ***t*-test** can be used to test whether the correlation between two variables is significant. The **test statistic** is r and the **standardized test statistic**

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

follows a t-distribution with $n - 2$ degrees of freedom.

In this session, only two-tailed hypothesis tests for ρ are considered.

Hypothesis Testing for ρ

Using the t-Test for the Correlation Coefficient ρ

In Words	*In Symbols*
1. State the null and alternative hypothesis.	State H_0 and H_a.
2. Specify the level of significance.	Identify α.
3. Identify the degrees of freedom.	d.f. = $n - 2$
4. Determine the critical value(s) and rejection region(s).	Use Table in Appendix .

Hypothesis Testing for ρ

Using the t-Test for the Correlation Coefficient ρ

In Words	*In Symbols*
5. Find the standardized test statistic.	$t = \dfrac{r}{\sqrt{\dfrac{1-r^2}{n-2}}}$
6. Make a decision to reject or fail to reject the null hypothesis.	If t is in the rejection region, reject H_0. Otherwise fail to reject H_0.
7. Interpret the decision in the context of the original claim.	

Hypothesis Testing for ρ

Example:
The following data represents the number of hours 12 different students watched television during the weekend and the scores of each student who took a test the following Monday.

The correlation coefficient $r \approx -0.831$.

Hours, x	0	1	2	3	3	5	5	5	6	7	7	10
Test score, y	96	85	82	74	95	68	76	84	58	65	75	50

Test the significance of this correlation coefficient significant at $\alpha = 0.01$?

Continued.

Hypothesis Testing for ρ

Example continued:

H_0: $\rho = 0$ (no correlation) H_a: $\rho \neq 0$ (significant correlation)

The level of significance is $\alpha = 0.01$.

Degrees of freedom are d.f. = 12 – 2 = 10.

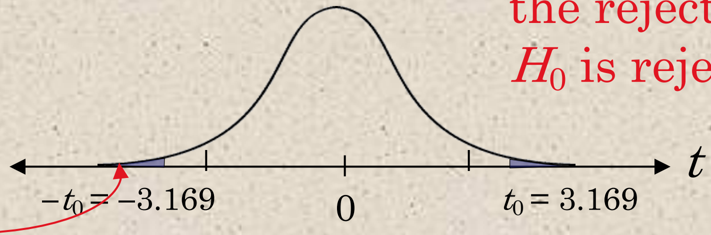
The critical values are $-t_0 = -3.169$ and $t_0 = 3.169$.

The standardized test statistic is

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{-0.831}{\sqrt{\frac{1-(-0.831)^2}{12-2}}}$$

$$\approx -4.72.$$

The test statistic falls in the rejection region, so H_0 is rejected.



At the 1% level of significance, there is enough evidence to conclude that there is a significant linear correlation between the number of hours of TV watched over the weekend and the test scores on Monday morning.

Linear Regression

Regression

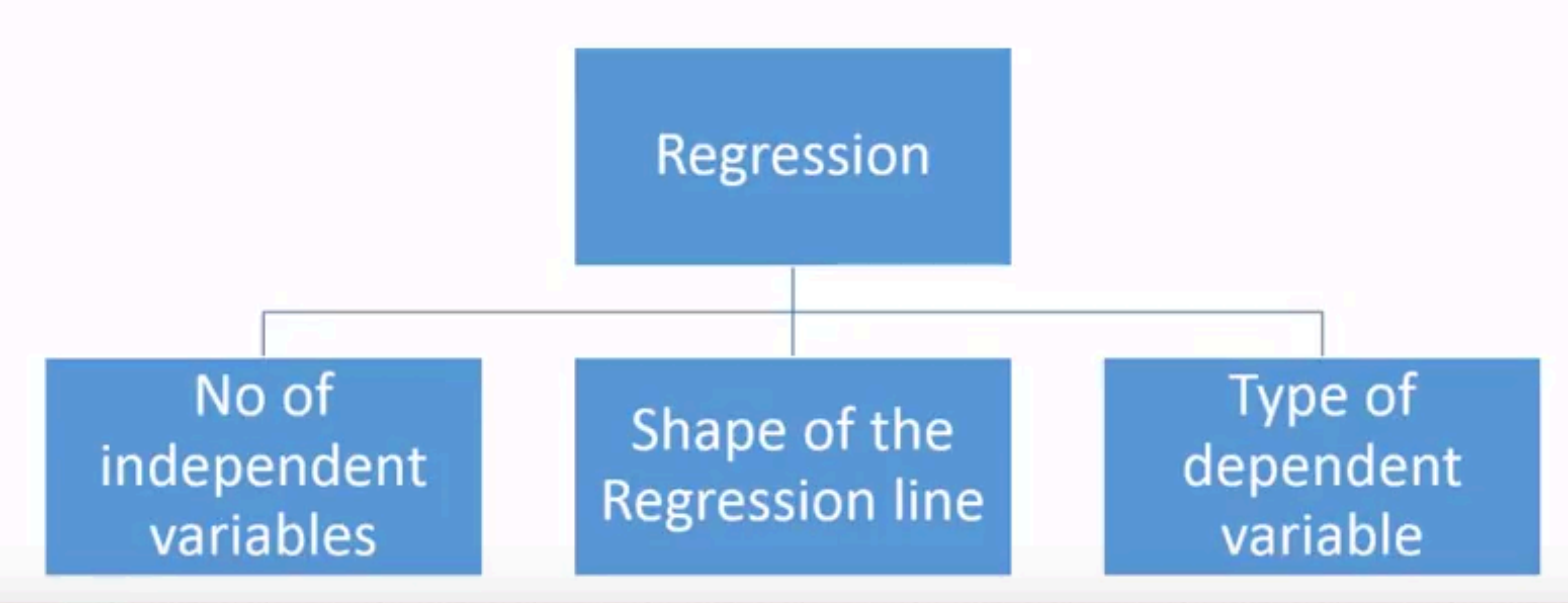
Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable (s) (predictor).

is a supervised learning technique

Advantage of Regression Analysis

- It is used to find the trends in data.
- It helps to predict real/continuous values.
- Finding the causal effect relationship
- Determine the most important factor, the least important factor, and how each factor is affecting the other factors.

Types of Regression



Regression

No of independent variables

Shape of the Regression line

Type of dependent variable

Terminologies Related to the Regression Analysis

Dependent Variable: The main factor in Regression analysis which we want to predict or understand is called the dependent variable. It is also called **target variable**.

Independent Variable: The factors which affect the dependent variables or which are used to predict the values of the dependent variables are called independent variable, also called as a **predictor**.

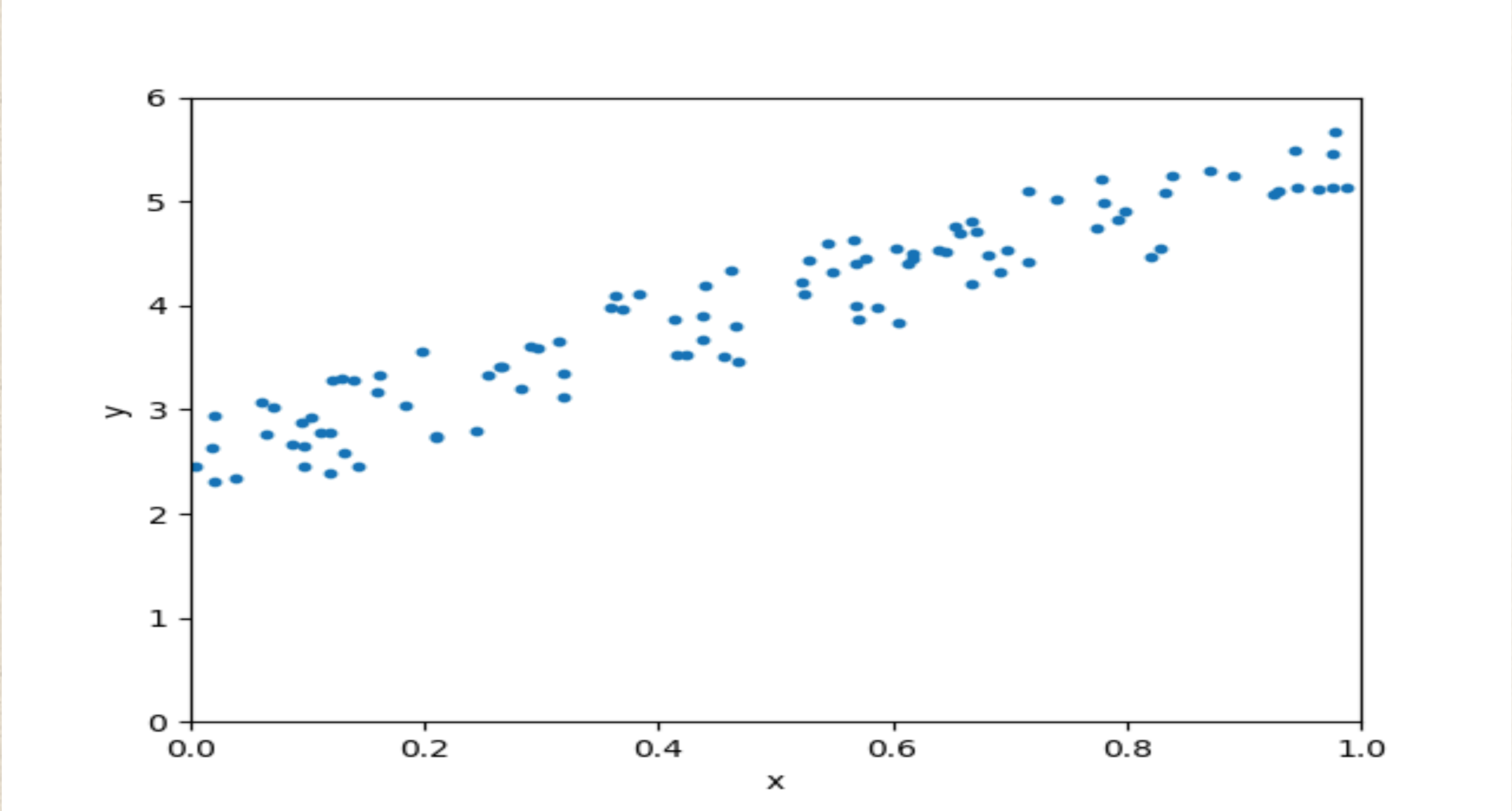
Outliers: Outlier is an observation which contains either very low value or very high value in comparison to other observed values. An outlier may hamper the result, so it should be avoided.

Multicollinearity: If the independent variables are highly correlated with each other than other variables, then such condition is called Multicollinearity. It should not be present in the dataset, because it creates problem while ranking the most affecting variable.

Underfitting and Overfitting: If our algorithm works well with the training dataset but not well with test dataset, then such problem is called **Overfitting**. And if our algorithm does not perform well even with training dataset, then such problem is called **underfitting**.

Regression Line

Draw a best fit line which shows the exact relationship



Regression Line

A **regression line**, also called a **line of best fit**, is the line for which the sum of the squares of the residuals is a minimum.

Hypothesis of Linear Regression

$$Y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_n x_n$$

Where :

- Y is the predicted value
- θ_0 is the bias term
- $\theta_1,\ldots,\theta_n$ are the model parameters
- x_1, $x_2,\ldots,x_n$ are the feature values.

The above hypothesis can also be represented by

$$Y = \theta^T x$$

where

θ is the model's parameter vector including the bias term θ_0

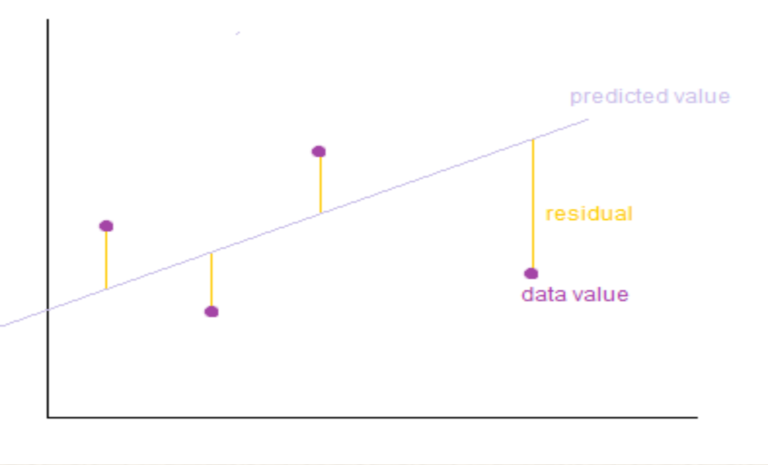
x is the feature vector with $x_0 = 1$

Training a Linear Regression Model

Calculating/Finding the parameters, so that the model **best fits the data.**

Determining the best fit line :

- The line for which the the error between the predicted values and the observed values is minimum is called the best fit line or the regression line.
- These errors are also called as ***residuals***.



Training a Linear Regression Model

Cost Function

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h(x^i) - y^i)^2$$

Hypothesis Function h(x)

$$h(x) = \theta_0 + \theta_1 x_1 + ... + \theta_n x_n$$

m is the total number of training examples in the data-set.

Gradient Descent

- is a generic optimization algorithm used in many machine learning algorithms.
- It iteratively tweaks the parameters of the model in order to minimize the cost function.

Gradient Descent

- Random initialization of model parameters
- measure how the cost function changes with change in it's parameters. Therefore compute the partial derivatives of the cost function w.r.t to the parameters θ_0, θ_1, … , θ_n

$$\frac{\partial J(\theta)}{\partial \theta_0} = \frac{1}{m}\sum_{i=1}^{m}(h(x^i) - y^i)$$

$$\frac{\partial J(\theta)}{\partial \theta_1} = \frac{1}{m}\sum_{i=1}^{m}(h(x^i) - y^i)x_1^i$$

similarly, the partial derivative of the cost function w.r.t to any parameter can be denoted by

$$\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{m}\sum_{i=1}^{m}(h(x^i) - y^i)x_j^i$$

Gradient Descent

The partial derivatives of all parameters at once can be computed using:

$$\begin{bmatrix} \frac{\partial J(\theta)}{\partial \theta_0} \\ \frac{\partial J(\theta)}{\partial \theta_1} \\ \vdots \\ \frac{\partial J(\theta)}{\partial \theta_n} \end{bmatrix} = \frac{1}{m} x^T (h(x) - y)$$

- After computing the derivative update the parameters as given below

$$\theta_0 = \theta_0 - \frac{\alpha}{m} \sum_{i=1}^{m} (h(x^i) - y^i)$$

$$\theta_1 = \theta_1 - \frac{\alpha}{m} \sum_{i=1}^{m} (h(x^i) - y^i) x_1^i$$

where α is the learning parameter.

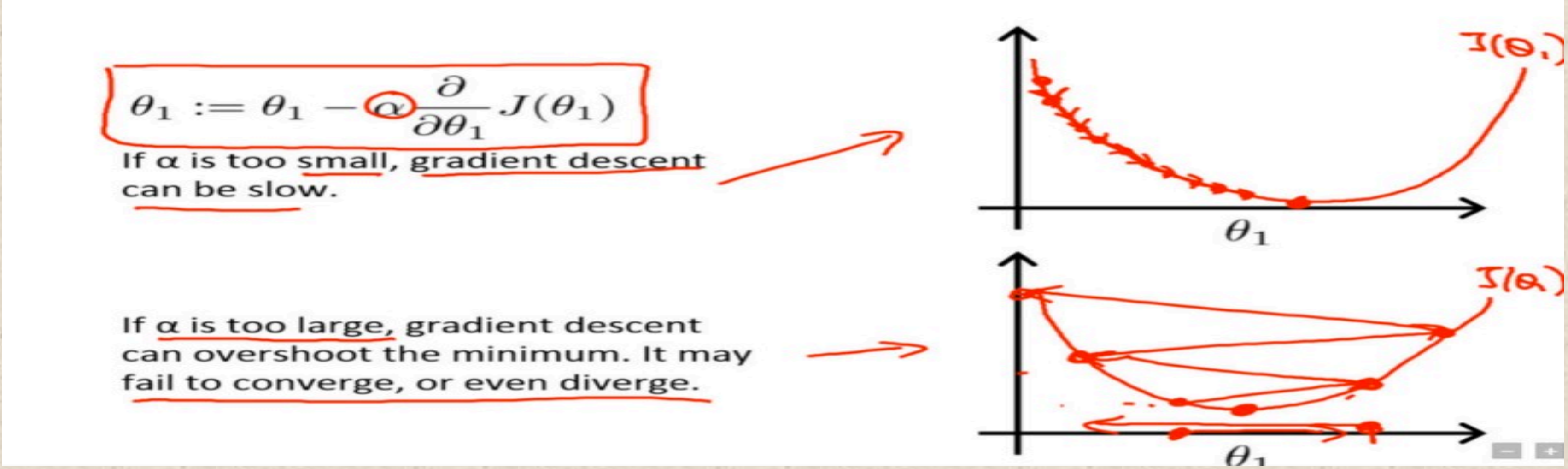
Gradient Descent

All the parameters can be updated at once using

$$\begin{bmatrix}\theta_0\\ \theta_1\\ \vdots\\ \theta_n\end{bmatrix} = \begin{bmatrix}\theta_0\\ \theta_1\\ \vdots\\ \theta_n\end{bmatrix} - \alpha \begin{bmatrix}\frac{\partial J(\theta)}{\partial \theta_0}\\ \frac{\partial J(\theta)}{\partial \theta_1}\\ \vdots\\ \frac{\partial J(\theta)}{\partial \theta_n}\end{bmatrix}$$

- Repeat the steps 2,3 until the cost function converges to the minimum value.

If the value of α is too small, the cost function takes larger time to converge. If α is too large, gradient descent may overshoot the minimum and may finally fail to converge.



$$\boxed{\theta_1 := \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_1)}$$

If α is too small, gradient descent can be slow.

If α is too large, gradient descent can overshoot the minimum. It may fail to converge, or even diverge.

Evaluating the performance of the model

RMSE is the square root of the average of the sum of the squares of residuals.

RMSE is defined by

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(h(x^i) - y^i)^2}$$

R^2 score or the **coefficient of determination** explains how much the total variance of the dependent variable can be reduced by using the least square regression.

R^2 is determined by

$$R^2 = 1 - \frac{SS_r}{SS_t}$$

SS_t is the total sum of errors if we take the mean of the observed values as the predicted value.

$$SS_t = \sum_{i=1}^{m}(y^i - \bar{y})^2$$

SS_r is the sum of the square of residuals

$$SS_r = \sum_{i=1}^{m}(h(x^i) - y^i)^2$$

Regression Line

Example 1:
The following data represents the number of hours 12 different students watched television during the weekend and the scores of each student who took a test the following Monday.

a.) Display the scatter plot.

b.)Apply the gradient decent optimization and Find the equation of the regression line after 5 iteration.

c.)Determine RMSE

d.) Determine the coefficient of determination (R^2)

e.) Use the equation to find the expected test score for a student who watches 9 hours of TV.

Hours, x	0	1	2	3	3	5	5	5	6	7	7	10
Test score, y	96	85	82	74	95	68	76	84	58	65	75	50

Regression Line

Example 2:

Consider the table below. It shows three performance measures for five students.

a.) Display the scatter plot.

b.)Apply the gradient decent optimization and Find the equation of the regression line after 5 iteration.

c.)Determine RMSE

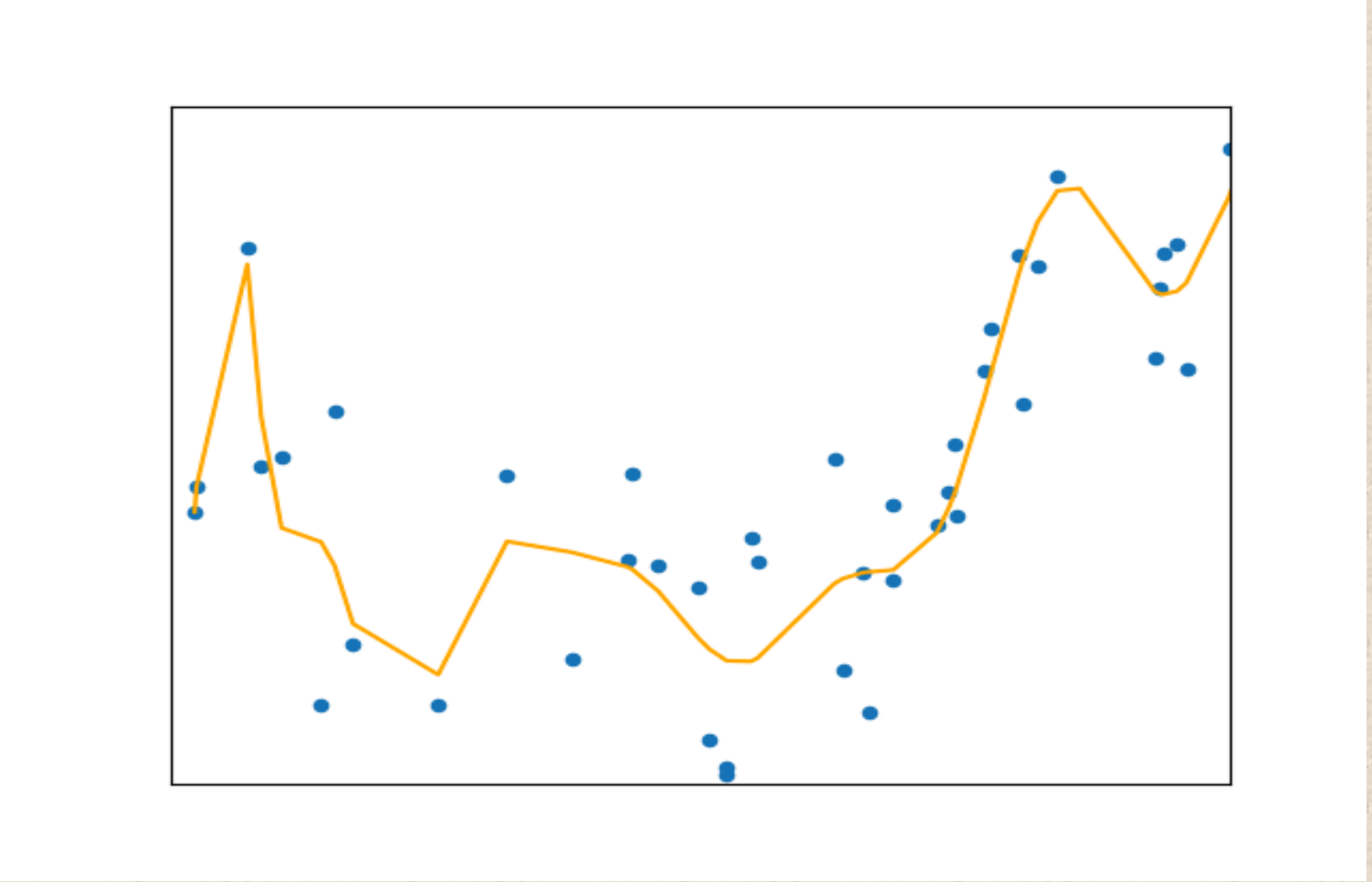
d.) Determine the coefficient of determination (R^2)

e.) Predict the test score when Study hours = 5, and IQ = 50.

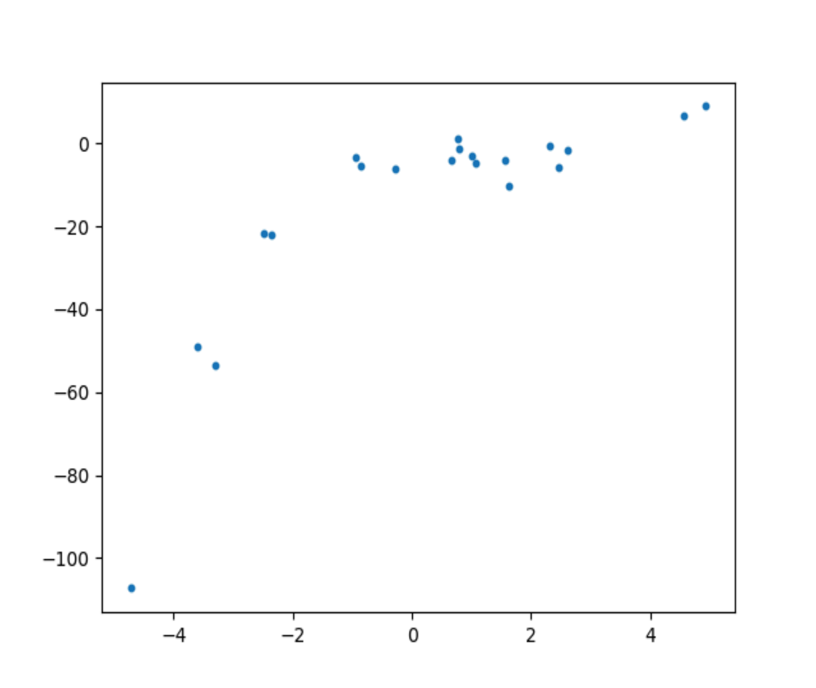
Study hours	40	30	20	0	10
IQ	110	120	100	90	80
Test score, *y*	100	90	80	70	60

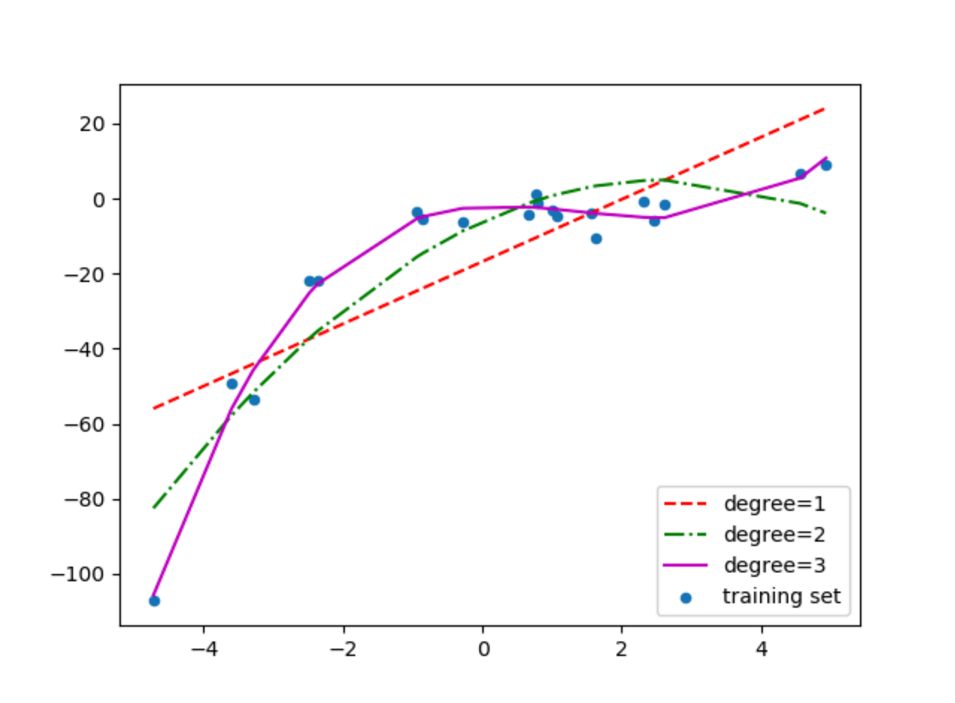
Polynomial Regression

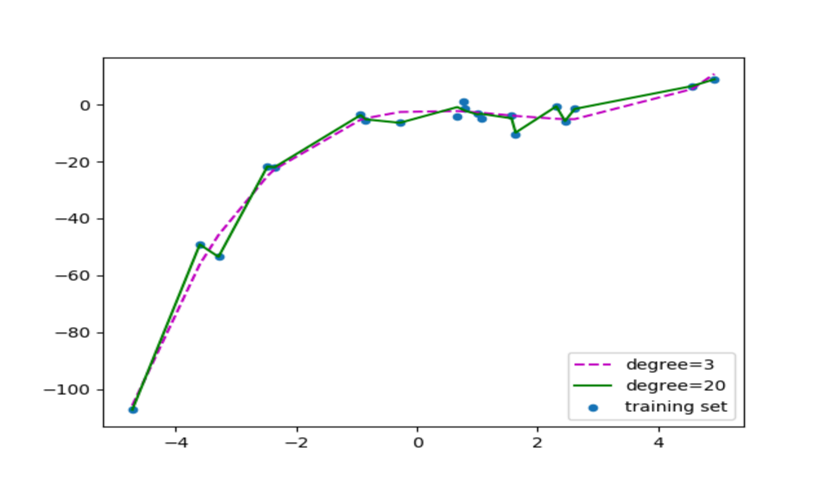
Why Polynomial Regression



Why Polynomial Regression







Why Polynomial Regression

Given $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ best fit $y = a_0 + a_1 x + \ldots + a_m x^m$ $(m \leq n-2)$ to a given data set.

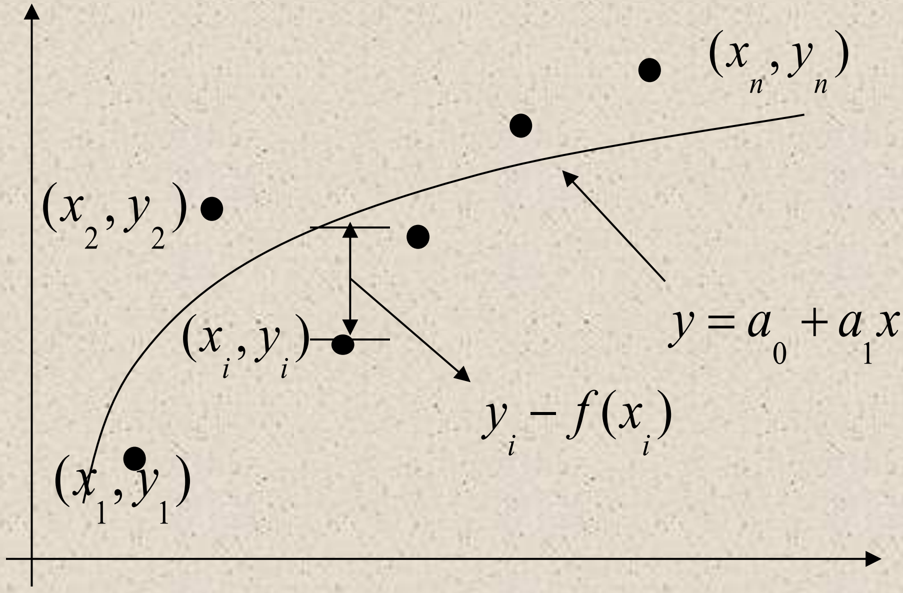


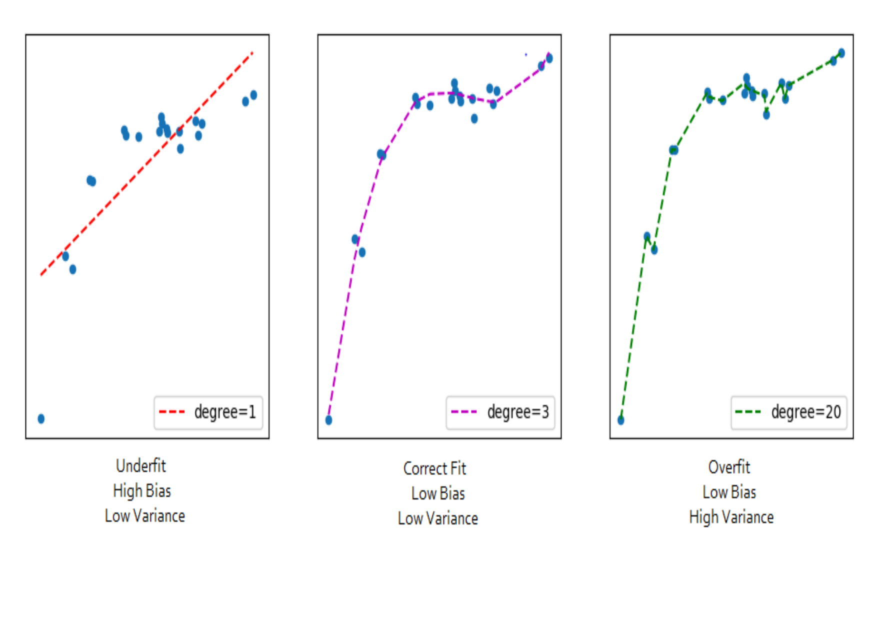
Figure. Polynomial model for nonlinear regression of y vs. x data

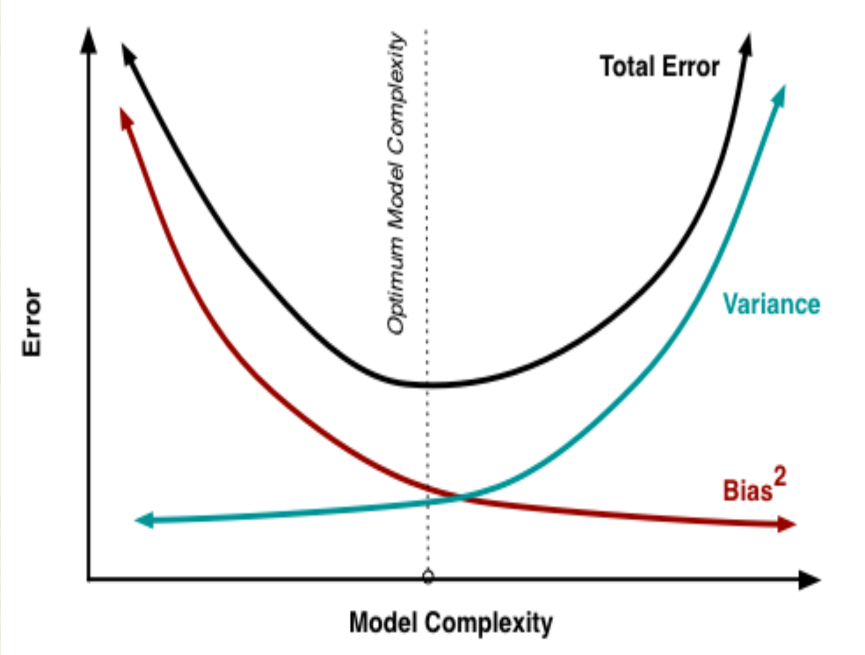
Training a Polynomial Regression Model cont.

It is the same with the previous approach.

How do we choose an optimal model?

The Bias vs Variance trade-off





Example 2-Polynomial Model

Regress the thermal expansion coefficient vs. temperature data to a second order polynomial.

Table. Data points for temperature vs α

Temperature, T (ºF)	Coefficient of thermal expansion, α (in/in/ºF)
80	6.47×10^{-6}
40	6.24×10^{-6}
−40	5.72×10^{-6}
−120	5.09×10^{-6}
−200	4.30×10^{-6}
−280	3.33×10^{-6}
−340	2.45×10^{-6}

a.) Display the scatter plot.

b.)Apply the gradient decent optimization and Find the equation of the regression line after 5 iteration.

c.)Determine RMSE

d.) Determine the coefficient of determination (R^2)

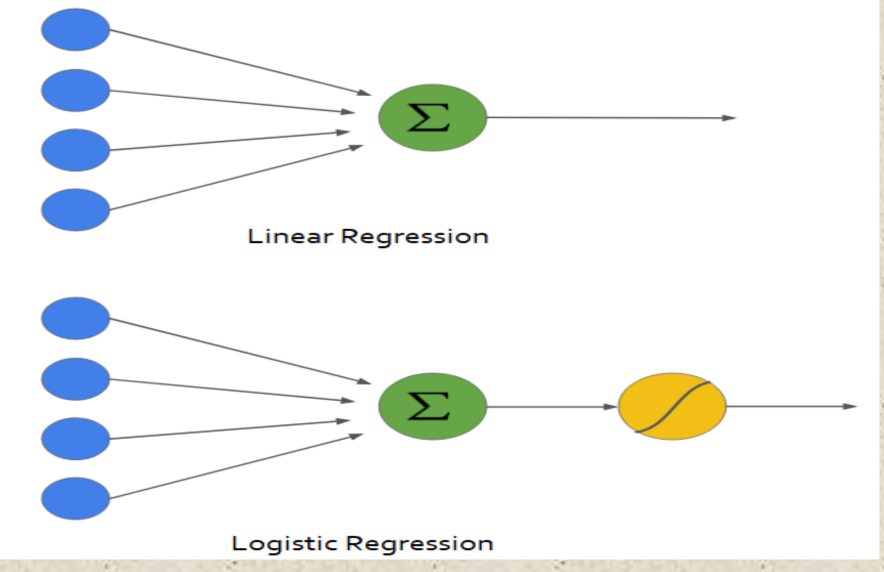
e.) Use the equation to find the expected test score for a student who watches 9 hours of TV.

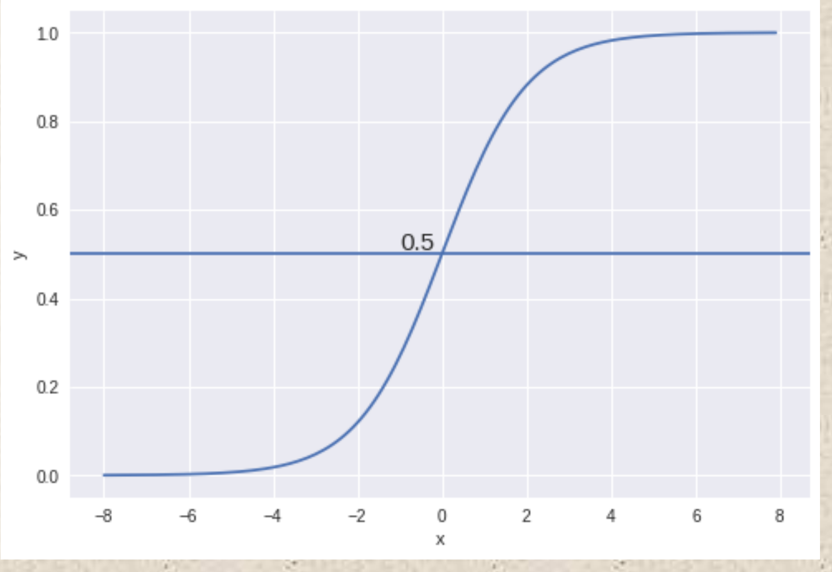
Figure. Data points for thermal expansion coefficient vs temperature.

Logistic Regression

- It is used for predicting the categorical dependent variable using a given set of independent variables.
- **is used for solving the classification problems**.
- Unlike Linear Regression, the dependent variable can take a limited number of values only i.e, the dependent variable is **categorical**.

Logistic Regression





Assumptions for Logistic Regression

- The dependent variable must be categorical
- The independent variables(features) must be independent (to avoid multicollinearity).

Logistic Regression

Hypothesis and Cost Function

A Linear Regression model can be represented by the equation.

When the the sigmoid function is applied to the output of the linear regression $h(x) = \theta^T x$

where the sigmoid function is represented by,

$$h(x) = \sigma(\theta^T x)$$

The hypothesis for logistic regression then becomes,

$$h(x) = \frac{1}{1 + e^{-\theta^T x}}$$

$$h(x) = \begin{cases} > 0.5, & \text{if } \theta^T x > 0 \\ < 0.5, & \text{if } \theta^T x < 0 \end{cases}$$

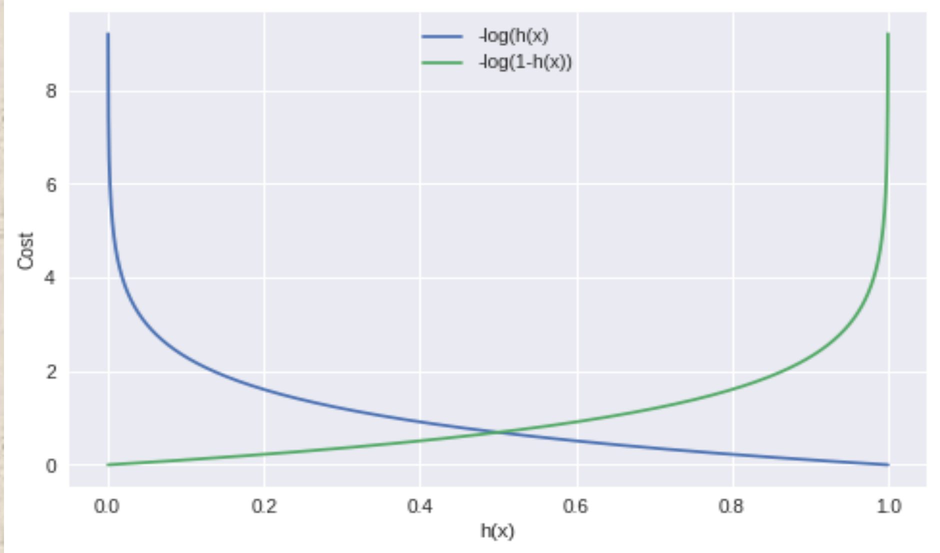
Logistic Regression

Cost Function

The cost function for a single training example can be given by:

$$cost = \begin{cases} -log(h(x), & \text{if } y = 1 \\ -log(1 - h(x)), & \text{if } y = 0 \end{cases}$$

Cost function intuition



Logistic Regression

Cost Function

We can combine both of the equations using:

$$cost(h(x), y) = -ylog(h(x)) - (1-y)log(1-h(x))$$

The cost for all the training examples denoted by *J(θ)* can be computed by taking the average over the cost of all the training samples

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{i}log(h(x^{i})) + (1-y^{i})log(1-h(x^{i}))]$$

where ***m*** is the number of training samples.

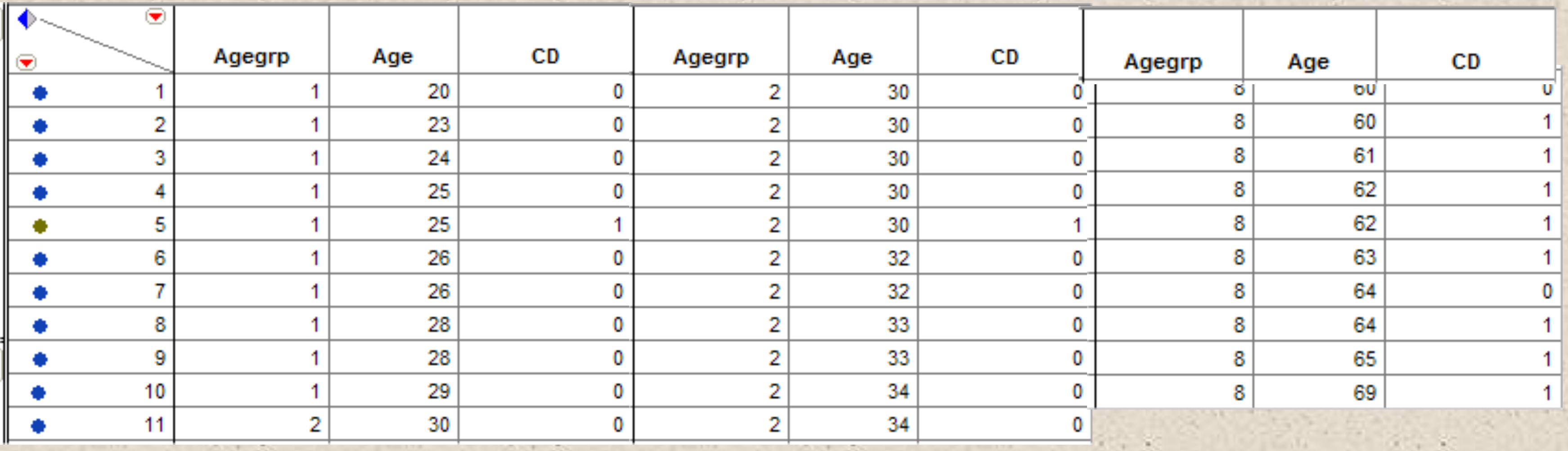
We will use gradient descent to minimize the cost function. The gradient w.r.t any parameter can be given by

$$\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{m}\sum_{i=1}^{m}(h(x^{i}) - y^{i})\ x_j^{i}$$

The equation is similar to what we achieved in Linear Regression, only h(x) is different in both the cases

Logistic Regression Example

Example: Coronary Heart Disease (CD) and Age In this study sampled individuals were examined for signs of CD (present = 1 / absent = 0) and the potential relationship between this outcome and their age (yrs.) was considered.



	Agegrp	Age	CD	Agegrp	Age	CD	Agegrp	Age	CD
1	1	20	0	2	30	0	8	60	0
2	1	23	0	2	30	0	8	60	1
3	1	24	0	2	30	0	8	61	1
4	1	25	0	2	30	0	8	62	1
5	1	25	1	2	30	1	8	62	1
6	1	26	0	2	32	0	8	63	1
7	1	26	0	2	32	0	8	64	0
8	1	28	0	2	33	0	8	64	1
9	1	28	0	2	33	0	8	65	1
10	1	29	0	2	34	0	8	69	1
11	2	30	0	2	34	0			

This is a portion of the raw data for the 100 subjects who participated in the study.

a.) Display the scatter plot.

b.)Apply the gradient decent optimization and Find the equation of the logistic regression line after 2 iteration.

c.) Use the equation to find the CD test for a subject whose age 78.