

Supervised Learning: Regression

- **Regression** is a type of supervised learning method that uses an algorithm to understand the relationship between dependent and independent variables.
- Regression models are helpful for predicting numerical values based on different data points, such as sales revenue projections for a given business.
- Regression is a supervised machine learning technique which is used to predict continuous values.
- The three main metrics that are used for evaluating the trained regression model are variance, bias and error. If the variance is high, it leads to overfitting and when the bias is high, it leads to underfitting.
- Based on the number of input features and output labels, regression is classified as linear (one input and one output), multiple (many inputs and one output) and multivariate (many outputs).

EXAMPLE OF REGRESSION

- Real estate price prediction is a problem that can be solved by supervised learning or, more specifically, by regression.
- Let us delve a little deeper into the problem. New City is the primary hub of the commercial activities in the country.
- In the last couple of decades, with increasing globalization, commercial activities have intensified in New City.
- Together with that, a large number of people have come and settled in the city with a dream to achieve professional growth in their lives.
- As an obvious fall-out, a large number of housing projects have started in every nook and corner of the city.
- But the demand for apartments has still outgrown the supply.
- To get benefit from this boom in real estate business,
- Karen has started a digital market agency for buying and selling real estates (including apartments, independent houses, town houses, etc.).
- Initially, when the business was small, she used to interact with buyers and sellers personally and help them arrive at a price quote – either for selling a property (for a seller) or for buying a property (for a buyer).
- Her long experience in real estate business helped her develop an intuition on what the correct price quote of a property could be – given the value of certain standard parameters such as area (sq. m.) of the property, location, floor, number of years since purchase, amenities available, etc.
- However, with the huge surge in the business, she is facing a big challenge.
- She is not able to manage personal interactions as well as setting the correct price quote for the properties all alone.
- She hired an assistant for managing customer interactions.
- But the assistant, being new in the real estate business, is struggling with price quotations.
- How can Karen solve this problem?
- Fortunately, Karen has a friend, Frank, who is a data scientist with in-depth knowledge in machine learning models. Frank comes up with a solution to Karen's problem.

- He builds a model which can predict the correct value of a real estate if it has certain standard inputs such as **area (sq. m.) of the property, location, floor, number of years since purchase, amenities available, etc.**
- He used a regression model to solve Karen's real estate price prediction problem.
- Problems related to prediction of numerical value can be solved using the regression model.
- In the context of regression, dependent variable (Y) is the one whose value is to be predicted, e.g. the price quote of the real estate in the context of Karen's problem.
- This variable is presumed to be functionally related to one (say, X) or more independent variables called predictors.
- In the context of Karen's problem, Frank used area of the property, location, floor, etc. as predictors of the model that he built.
- The dependent variable depends on independent variable(s) or predictor(s).
- Regression is essentially finding a relationship (or) association between the dependent variable (Y) and the independent variable(s) (X), i.e. to find the function 'f' for the association $Y = f(X)$.

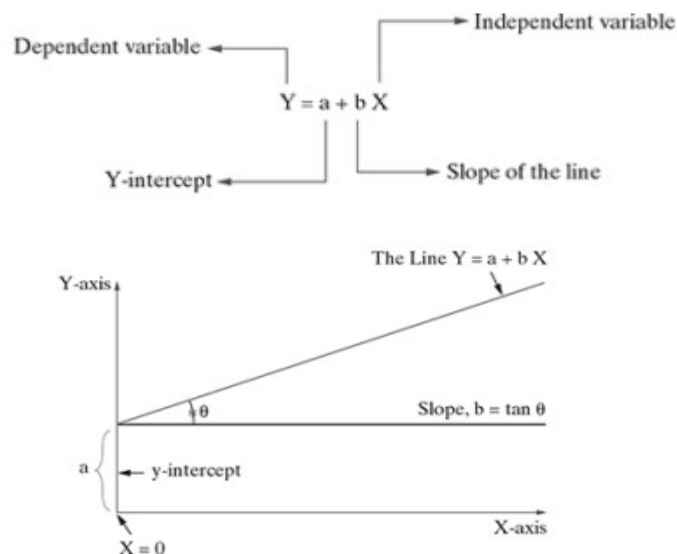
COMMON REGRESSION ALGORITHMS

The most common regression algorithms are

1. Simple linear regression
2. Multiple linear regression
3. Polynomial regression
4. Multivariate adaptive regression splines
5. Logistic regression
6. Maximum likelihood estimation (least squares)

Simple Linear Regression,

- Simple linear regression is the simplest regression model which involves only one predictor.
- This model assumes a linear relationship between the dependent variable and the predictor variable



- In the context of Karen's problem, if we take Price of a Property as the dependent variable and the Area of the Property (in sq. m.) as the predictor variable, we can build a model using simple linear regression.

$$\text{Price}_{\text{property}} = f(\text{Area}_{\text{property}})$$

- Assuming a linear association, we can reformulate the model as

$$\text{Price}_{\text{property}} = a + b \cdot \text{Area}_{\text{property}}$$

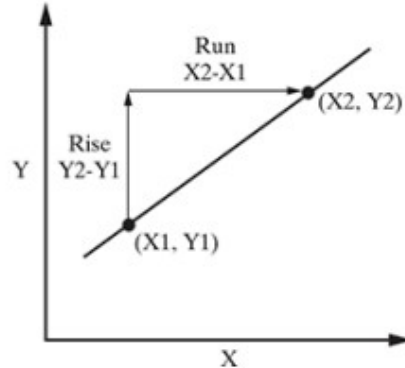
- where 'a' and 'b' are intercept and slope of the straight line, respectively.
- Straight lines can be defined in a slope– intercept form

$$Y = (a + bX),$$
- where a = intercept and b = slope of the straight line.
- The value of intercept indicates the value of Y when X = 0. It is known as 'the intercept or Y intercept' because it specifies where the straight line crosses the vertical or Y-axis

Slope of the simple linear regression model

- Slope of a straight line represents how much the line in a graph changes in the vertical direction (Y-axis) over a change in the horizontal direction (X-axis)

$$\text{Slope} = \text{Change in Y} / \text{Change in X}$$
- Rise is the change in Y-axis ($Y_2 - Y_1$) and Run is the change in X-axis ($X_2 - X_1$). So, slope is represented as



$$\text{Slope} = \frac{\text{Rise}}{\text{Run}} = \frac{Y_2 - Y_1}{X_2 - X_1}$$

- Let us find the slope of the graph where the lower point on the line is represented as (-3, -2) and the higher point on the line is represented as (2, 2).
- $(X_1, Y_1) = (-3, -2)$ and $(X_2, Y_2) = (2, 2)$

$$(X_1, Y_1) = (-3, -2) \text{ and } (X_2, Y_2) = (2, 2)$$

$$\text{Rise} = (Y_2 - Y_1) = (2 - (-2)) = 2 + 2 = 4$$

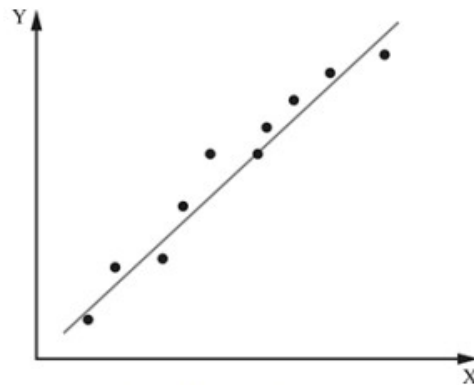
$$\text{Run} = (X_2 - X_1) = (2 - (-3)) = 2 + 3 = 5$$

$$\text{Slope} = \text{Rise}/\text{Run} = 4/5 = 0.8$$

There can be two types of slopes in a linear regression model: positive slope and negative slope. Different types of regression lines based on the type of slope include

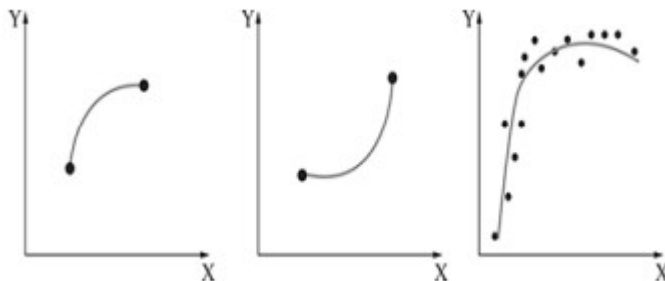
1. Linear positive slope
2. Curve linear positive slope
3. Linear negative slope
4. Curve linear negative slope

Linear positive slope A positive slope always moves upward on a graph from left to right

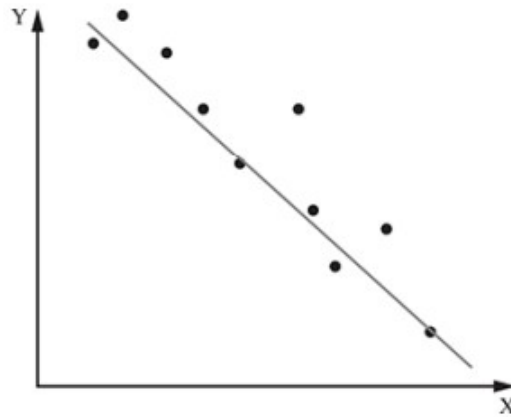


Curve linear positive slope Curve linear positive slope Curves in the following graphs slope upward from left to right.

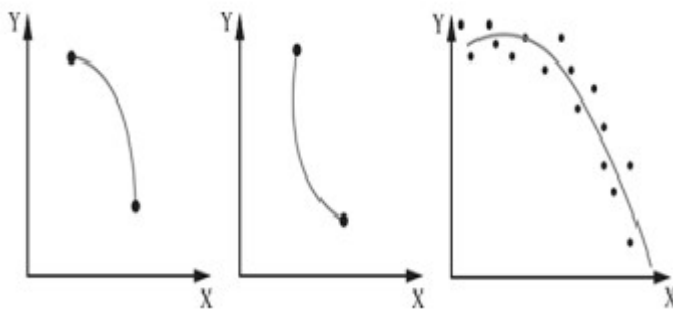
$$\text{Slope} = (Y - Y) / (X - X) = \text{Delta } (Y) / \text{Delta}(X)$$



- Slope for a variable (X) may vary between two graphs, but it will always be positive; hence, the above graphs are called as graphs with curve linear positive slope.
- **Linear negative slope** A negative slope always moves downward on a graph from left to right. As X value (on X-axis) increases, Y value decreases



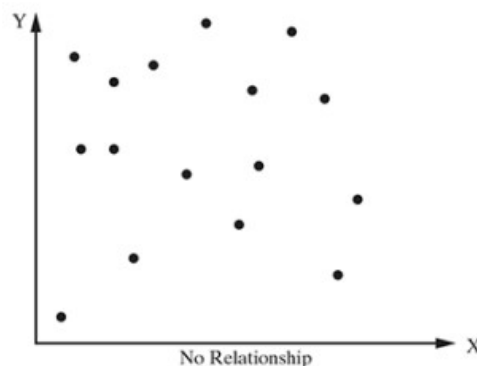
Curve linear negative slope Curves in the following graphs slope downward from left to right.



- $\text{Slope} = (Y - Y) / (X - X) = \text{Delta}(Y) / \text{Delta}(X)$
- Slope for a variable (X) may vary between two graphs, but it will always be negative; hence, the above graphs are called as graphs with curve linear negative slope.

No relationship graph

Scatter graph shown in Figure indicates 'no relationship' curve as it is very difficult to conclude whether the relationship between X and Y is positive or negative.



Error in simple regression

- The regression equation model in machine learning uses the slope–intercept format in algorithms.
- X and Y values are provided to the machine, and it identifies the values of a (intercept) and b (slope) by relating the values of X and Y.
- However, identifying the exact match of values for a and b is not always possible.

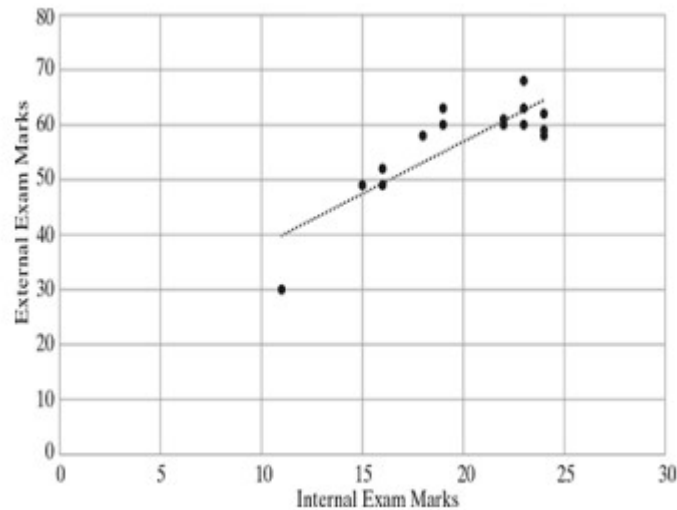
- There will be some error value (ϵ) associated with it.
- This error is called marginal or residual error.
 - $Y = (a + bX) + \epsilon$

Example of simple regression A college professor believes that if the grade for internal examination is high in a class, the grade for external examination will also be high.

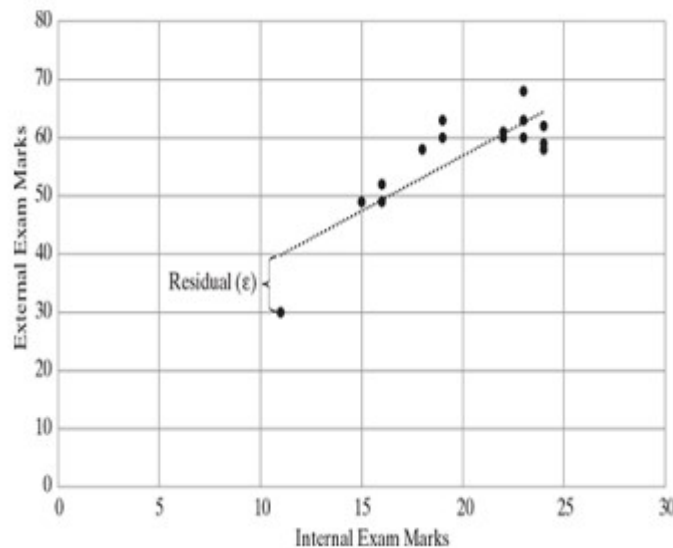
A random sample of 15 students in that class was selected, and the data is given below:

Internal Exam	15	23	18	23	24	22	22	19	19	16	24	11	24	16	23
External Exam	49	63	58	60	58	61	60	63	60	52	62	30	59	49	68

- A scatter plot was drawn to explore the relationship between the independent variable (internal marks) mapped to X-axis and dependent variable (external marks) mapped to Y-axis as depicted as



- As you can observe from the above graph, the line (i.e. the regression line) does not predict the data exactly). Instead, it just cuts through the data.
- Some predictions are lower than expected, while some others are higher than expected.
- Residual is the distance between the predicted point (on the regression line) and the actual point as depicted



- In simple linear regression, the line is drawn using the regression formula. $Y = (a + bX) + \epsilon$
- If the values of 'a' and 'b', are known then it is easy to predict the value of Y for any given X by using the above formula.
- A straight line is drawn as close as possible over the points on the scatter plot.
- **Ordinary Least Squares (OLS)** is the technique used to estimate a line that will minimize the error (ϵ), which is the difference between the predicted and the actual values of Y. This means summing the errors of each prediction or, more appropriately, the Sum of the Squares of the Errors

$$(\text{SSE}) \left(\text{i.e. } \sum_i \epsilon_i^2 \right).$$

- It is observed that the SSE is least when b takes the value

$$b = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2} = \frac{\text{cov}(X, Y)}{\text{Var}(X)}$$

- The corresponding value of 'a' calculated using the above value of 'b' is

$$a = \bar{Y} - b\bar{X}$$

*Calculation summary*Sum of $X = 299$ Sum of $Y = 852$ Mean $X, M_X = 19.93$ Mean $Y, M_Y = 56.8$ Sum of squares (SS_X) = 226.9333

Sum of products (SP) = 429.8

Regression equation = $\hat{y} = bX + a$

$$b = \frac{SP}{SS_X} = \frac{429.8}{226.93} = 1.89395$$

$$a = M_Y - bM_X = 56.8 - (1.89 \times 19.93) = 19.0473$$

$$\hat{y} = 1.89395X + 19.0473$$

- Hence, for the above example, the estimated regression equation is constructed on the basis of the estimated values of a and b

$$\hat{y} = 1.89395X + 19.0473$$

So, in the context of the given problem, we can say

Marks in external exam = $19.04 + 1.89 \times (\text{Marks in internal exam})$

or,

$$M_{\text{Ext}} = 19.04 + 1.89 \times M_{\text{Int}}$$

		Step 2		Step 3	Step 5
X	Y	X- mean (X)	Y- Mean (Y)	$(X_i - \bar{X})(Y_i - \bar{Y})$	$(X_i - \bar{X})^2$
15	49	-4.09	-7.8	-31.954	16.7281
23	63	3.07	6.2	19.034	9.4249
18	58	-1.09	1.2	-1.306	1.1881
23	60	3.07	3.2	9.824	9.4249
24	58	4.07	1.2	4.884	16.5649
22	61	2.07	4.2	8.694	4.2849
22	60	2.07	3.2	6.624	4.2849
19	63	-0.09	6.2	-0.556	0.0081
19	60	-0.09	3.2	-0.296	0.0081
16	52	-3.09	-4.8	14.864	9.5481
24	62	4.07	5.2	21.164	16.5649
11	30	-8.09	-26.8	216.812	65.4481
24	59	4.07	2.2	8.954	16.5649
16	49	-3.09	-7.8	-24.154	9.5481
23	68	3.07	11.2	34.384	9.4249
19.9	56.8		$\Sigma(X_i - \bar{X})(Y_i - \bar{Y})$	429.8	226.9335

Step 7: Divide (step4 / step6)

$$b = 429.28 / 226.93 = 1.89$$

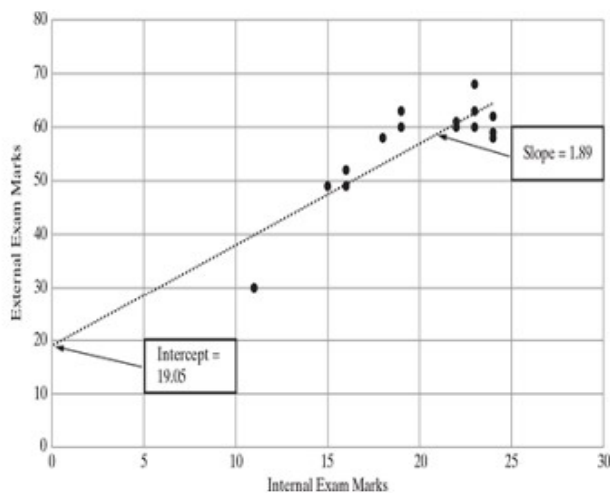
Step 8: Calculate a using the value of b

$$a = \bar{Y} - b\bar{X}$$

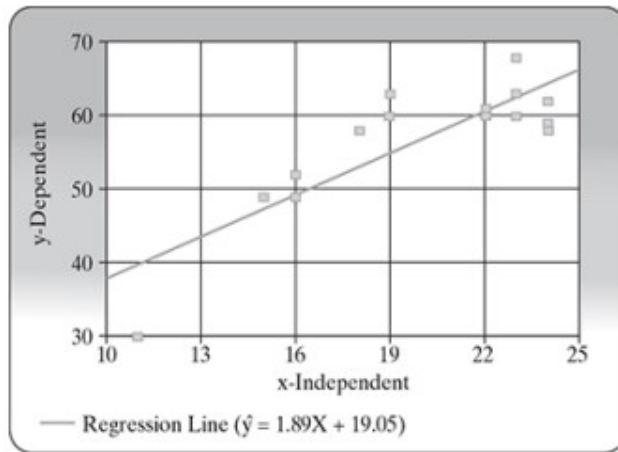
$$a = 56.8 - 1.89 \times 19.9$$

$$a = 19.05$$

The model built above can be represented graphically as an extended version a zoom-in version



Extended version of the regression graph



Zoom-in regression line

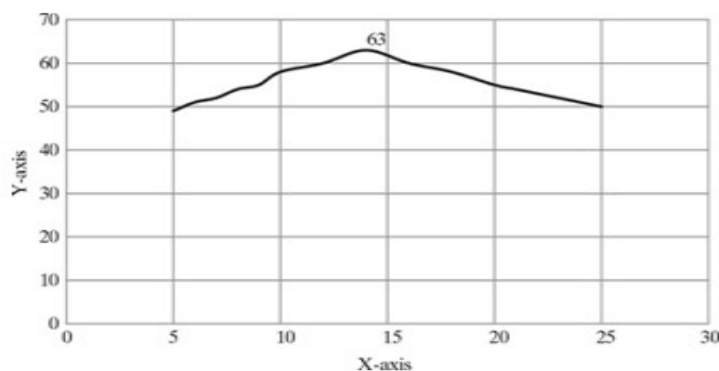
Slope measures the estimated change in the average value of Y as a result of a one-unit change in X. Here, slope = 1.89 tells us that the average value of the external examination marks increases by 1.89 for each additional 1 mark in the internal examination

OLS algorithm

- Step 1: Calculate the mean of X and Y
- Step 2: Calculate the errors of X and Y
- Step 3: Get the product
- Step 4: Get the summation of the products
- Step 5: Square the difference of X
- Step 6: Get the sum of the squared difference
- Step 7: Divide output of step 4 by output of step 6 to calculate 'b'
- Step 8: Calculate 'a' using the value of 'b'

Maximum and minimum point of curves

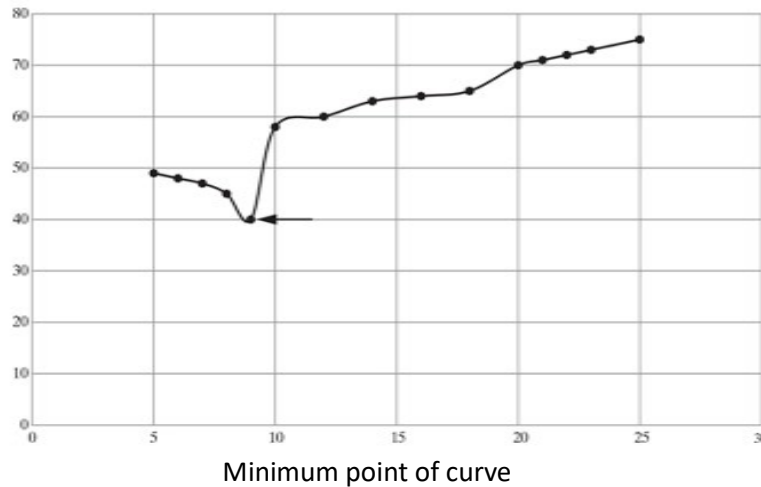
- Maximum and minimum points on a graph are found at points where the slope of the curve is zero.
- It becomes zero either from positive or negative value.
- The maximum point is the point on the curve of the graph with the highest y-coordinate and a slope of zero.
- The minimum point is the point on the curve of the graph with the lowest y-coordinate and a slope of zero.



Maximum point of curve

Point 63 is at the maximum point for this curve Point 63 is at the highest point on this curve.

It has a greater y-coordinate value than any other point on the curve and has a slope of zero.



Multiple Linear Regression

- In a multiple regression model, two or more independent variables, i.e. predictors are involved in the model.
- In a simple linear regression Price of a Property is considered as the dependent variable and the Area of the Property (in sq. m.) as the predictor variable.
- However, location, floor, number of years since purchase, amenities available, etc. are also important predictors which should not be ignored.
- Thus, if we consider Price of a Property (in \$) as the dependent variable and Area of the Property (in sq. m.), location, floor, number of years since purchase and amenities available as the independent variables, we can form a multiple regression equation as shown below:

$$\text{Price}_{\text{property}} = f(\text{Area}_{\text{property}}, \text{location}, \text{floor}, \text{Ageing}, \text{Amenities})$$

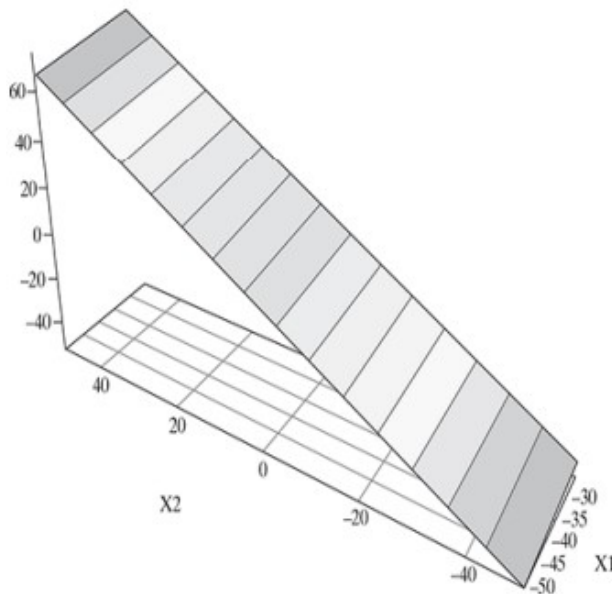
- The simple linear regression model and the multiple regression model assume that the dependent variable is continuous.
- The following expression describes the equation involving the relationship with two predictor variables, namely X_1 and X_2 .

$$\hat{Y} = a + b_1 X_1 + b_2 X_2$$

- The model describes a plane in the three-dimensional space of \hat{Y} , X_1 , and X_2 .
- Parameter 'a' is the intercept of this plane. Parameters ' b_1 ' and ' b_2 ' are referred to as partial regression coefficients.
- Parameter b_1 represents the change in the mean response corresponding to a unit change in X_1 when X_2 is held constant.
- Parameter b_2 represents the change in the mean response corresponding to a unit change in X_2 when X_1 is held constant.

Consider the following example of a multiple linear regression model with two predictor variables, namely X_1 and X_2

$$\hat{Y} = 22 + 0.3X_1 + 1.2X_2$$



Multiple regression plane

Multiple regression for estimating equation when there are 'n' predictor variables is as follows:

$$\hat{Y} = a + b_1X_1 + b_2X_2 + b_3X_3 + \dots + b_nX_n$$

While finding the best fit line, we can fit either a polynomial or curvilinear regression. These are known as polynomial or curvilinear regression, respectively

Assumptions in Regression Analysis

1. The dependent variable (Y) can be calculated / predicated as a linear function of a specific set of independent variables (X's) plus an error term (ϵ).
2. The number of observations (n) is greater than the number of parameters (k) to be estimated, i.e. $n > k$.
3. Relationships determined by regression are only relationships of association based on the data set and not necessarily of cause and effect of the defined class.
4. Regression line can be valid only over a limited range of data. If the line is extended (outside the range of extrapolation), it may only lead to wrong predictions.
5. If the business conditions change and the business assumptions underlying the regression model are no longer valid, then the past data set will no longer be able to predict future trends.
6. Variance is the same for all values of X (homoskedasticity).
7. The error term (ϵ) is normally distributed. This also means that the mean of the error (ϵ) has an expected value of 0.

8. The values of the error (ϵ) are independent and are not related to any values of X. This means that there are no relationships between a particular X, Y that are related to another specific value of X, Y.

Given the above assumptions, the OLS estimator is the Best Linear Unbiased Estimator (BLUE), and this is called as **Gauss-Markov Theorem**

Main Problems in Regression Analysis

- In multiple regressions, there are two primary problems:
 1. Multicollinearity and
 2. Heteroskedasticity.

Multicollinearity

- Two variables are perfectly collinear if there is an exact linear relationship between them.
- Multicollinearity is the situation in which the degree of correlation is not only between the dependent variable and the independent variable, but there is also a strong correlation within (among) the independent variables themselves.
- A multiple regression equation can make good predictions when there is multicollinearity, but it is difficult for us to determine how the dependent variable will change if each independent variable is changed one at a time.
- When multicollinearity is present, it increases the standard errors of the coefficients.
- By overinflating the standard errors, multicollinearity tries to make some variables statistically insignificant when they actually should be significant (with lower standard errors).
- One way to gauge multicollinearity is to calculate the Variance Inflation Factor (VIF), which assesses how much the variance of an estimated regression coefficient increases if the predictors are correlated.
- If no factors are correlated, the VIFs will be equal to 1.
- The assumption of no perfect collinearity states that there is no exact linear relationship among the independent variables.
- This assumption implies two aspects of the data on the independent variables.
- First, none of the independent variables, other than the variable associated with the intercept term, can be a constant.
- Second, variation in the X's is necessary.
- In general, the more variation in the independent variables, the better will be the OLS estimates in terms of identifying the impacts of the different independent variables on the dependent variable

Heteroskedasticity

- Heteroskedasticity refers to the changing variance of the error term.
- If the variance of the error term is not constant across data sets, there will be erroneous predictions.
- In general, for a regression equation to make accurate predictions, the error term should be independent, identically (normally) distributed (iid).
- Mathematically, this assumption is written as

$$\begin{aligned} \text{var}(u_i|X) &= \sigma^2 \quad \text{and} \\ \text{cov}(u_i u_j | X) &= 0 \quad \text{for } i \neq j \end{aligned}$$

- where ‘var’ represents the variance, ‘cov’ represents the covariance, ‘u’ represents the error terms, and ‘X’ represents the independent variables.

- This assumption is more commonly written as

$$\begin{aligned}\text{var}(u_i) &= \sigma^2 \quad \text{and} \\ \text{cov}(u_i, u_j) &= 0 \quad \text{for } i \neq j.\end{aligned}$$

Improving Accuracy of the Linear Regression Model

- The concept of bias and variance is similar to accuracy and prediction.
- Accuracy refers to how close the estimation is near the actual value, whereas prediction refers to continuous estimation of the value.
 - High bias = low accuracy (not close to real value)
 - High variance = low prediction (values are scattered)
 - Low bias = high accuracy (close to real value)
 - Low variance = high prediction (values are close to each other)
- A regression model which is highly accurate and highly predictive; then, the overall error of model will be low, implying a low bias (high accuracy) and low variance (high prediction).
- This is highly preferable.
- Similarly, if the variance increases (low prediction), the spread of the data points increases, which results in less accurate prediction.
- As the bias increases (low accuracy), the error between our predicted value and the observed values increases.
- Therefore, balancing out bias and accuracy is essential in a regression model.
- In the linear regression model, it is assumed that the number of observations (n) is greater than the number of parameters (k) to be estimated, i.e. $n > k$, and in that case, the least squares estimates tend to have low variance and hence will perform well on test observations.
- However, if observations (n) is not much larger than parameters (k), then there can be high variability in the least squares fit, resulting in overfitting and leading to poor predictions. If $k > n$, then linear regression is not usable. This also indicates infinite variance, and so, the method cannot be used at all.

Accuracy of linear regression can be improved using the following three methods:

1. Shrinkage Approach
2. Subset Selection
3. Dimensionality (Variable) Reduction

Shrinkage (Regularization) approach

- By limiting (shrinking) the estimated coefficients, the variance can be reduced at the cost of a negligible increase in bias.
- This can in turn lead to substantial improvements in the accuracy of the model.
- Few variables used in the multiple regression model are in fact not associated with the overall response and are called as irrelevant variables; this may lead to unnecessary complexity in the regression model.

- This approach involves fitting a model involving all predictors. However, the estimated coefficients are shrunk towards zero relative to the least squares estimates.
- This shrinkage (also known as regularization) has the effect of reducing the overall variance. Some of the coefficients may also be estimated to be exactly zero, thereby indirectly performing variable selection.
- The two best-known techniques for shrinking the regression coefficients towards zero are
 1. ridge regression
 2. lasso (Least Absolute Shrinkage Selector Operator)
- Ridge regression performs L2 regularization, i.e. it adds penalty equivalent to square of the magnitude of coefficients
- Minimization objective of ridge = $LS\ Obj + \alpha \times (\text{sum of square of coefficients})$
- Lasso regression performs L1 regularization, i.e. it adds penalty equivalent to the absolute value of the magnitude of coefficients.
- Minimization objective of ridge = $LS\ Obj + \alpha \times (\text{absolute value of the magnitude of coefficients})$

Subset selection

- Identify a subset of the predictors that is assumed to be related to the response and then fit a model using OLS on the selected reduced subset of variables.
- There are two methods in which subset of the regression can be selected:
 1. Best subset selection (considers all the possible (2^k))
 2. Stepwise subset selection
 1. Forward stepwise selection (0 to k)
 2. Backward stepwise selection (k to 0)
- In best subset selection, we fit a separate least squares regression for each possible subset of the k predictors.
- For computational reasons, best subset selection cannot be applied with very large value of predictors (k).
- The best subset selection procedure considers all the possible (2^k) models containing subsets of the p predictors.
- The stepwise subset selection method can be applied to choose the best subset.
- There are two stepwise subset selection: 1. Forward stepwise selection (0 to k) 2. Backward stepwise selection (k to 0)
- Forward stepwise selection is a computationally efficient alternative to best subset selection. Forward stepwise considers a much smaller set of models, that too step by step, compared to best set selection.
- Forward stepwise selection begins with a model containing no predictors, and then, predictors are added one by one to the model, until all the k predictors are included in the model.
- In particular, at each step, the variable (X) that gives the highest additional improvement to the fit is added.
- Backward stepwise selection begins with the least squares model which contains all k predictors and then iteratively removes the least useful predictor one by one.

Dimensionality reduction (Variable reduction)

- In dimensionality reduction, predictors (X) are transformed, and the model is set up using the transformed variables after dimensionality reduction.
- The number of variables is reduced using the dimensionality reduction method.
- Principal component analysis is one of the most important dimensionality (variable) reduction techniques.

Polynomial Regression Model

- Polynomial regression model is the extension of the simple linear model by adding extra predictors obtained by raising (squaring) each of the original predictors to a power.
- For example, if there are three variables, X, X^2 , and X^3 are used as predictors.
- This approach provides a simple way to yield a non-linear fit to data.

$$f(x) = c_0 + c_1X + c_2X^2 + c_3X^3$$

Example: Let us use the below data set of (X, Y) for degree 3 polynomial.

Internal Exam (X)	15	23	18	23	24	22	22	19	19	16	24	11	24	16	23
External Exam (Y)	49	63	58	60	58	61	60	63	60	52	62	30	59	49	68

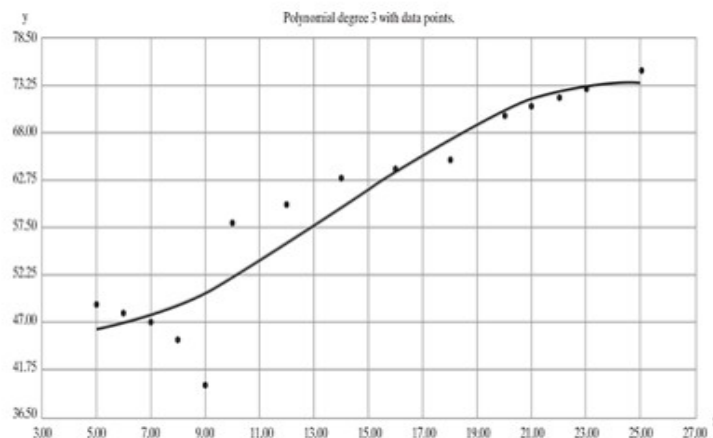
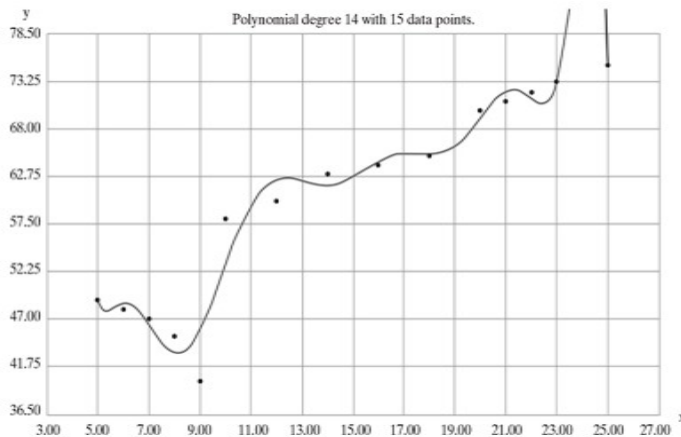


FIG. 8.16 Polynomial regression degree 3

- The regression line is slightly curved for polynomial degree 3 with the above 15 datapoints.



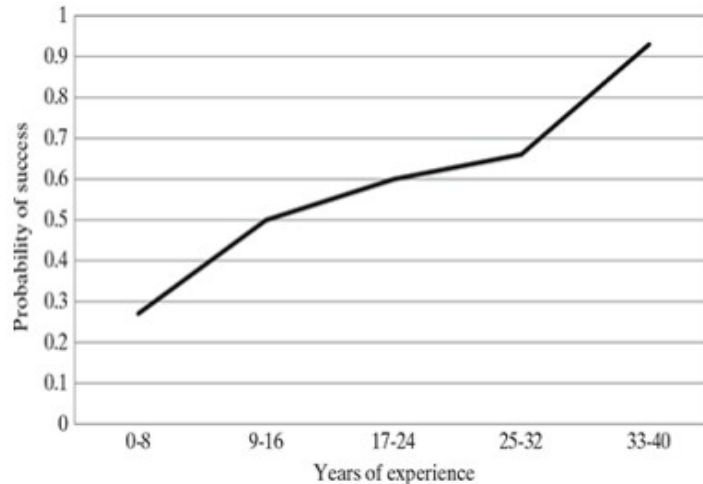
- The regression line will curve further if we increase the polynomial degree
- At the extreme value, the regression line will be overfitting into all the original values of X

Logistic Regression

- Logistic regression is both classification and regression technique depending on the scenario used.
- Logistic regression (logit regression) is a type of regression analysis used for predicting the outcome of a categorical dependent variable similar to OLS regression.
- In logistic regression, dependent variable (Y) is binary (0,1) and independent variables (X) are continuous in nature.
- The probabilities describing the possible outcomes (probability that $Y = 1$) of a single trial are modelled as a logistic function of the predictor variables.
- In the logistic regression model, there is no R to gauge the fit of the overall model; however, a chi-square test is used to gauge how well the logistic regression model fits the data.
- The goal of logistic regression is to predict the likelihood that Y is equal to 1 (probability that $Y = 1$ rather than 0) given certain values of X.
- That is, if X and Y have a strong positive linear relationship, the probability that a person will have a score of $Y = 1$ will increase as values of X increase.
- So, probabilities are predicted rather than the scores of the dependent variable.
- The logistic function, always takes values between zero and one.
- The logistic formulae are stated in terms of the probability that $Y = 1$, which is referred to as P. The probability that Y is 0 is $1 - P$.

$$\ln\left(\frac{P}{1-P}\right) = a + bX$$

$$\ln(p/1 - p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \varepsilon$$



Assumptions in logistic regression

The following assumptions must hold when building a logistic regression model:

- There exists a linear relationship between logit function and independent variables
- The dependent variable Y must be categorical (1/0) and take binary value, e.g. if pass then Y = 1; else Y = 0
- The data meets the 'iid' criterion, i.e. the error terms, ε , are independent from one another and identically distributed
- The error term follows a binomial distribution [n, p] n = # of records in the data p = probability of success (pass, responder)

Maximum Likelihood Estimation

- The coefficients in a logistic regression are estimated using a process called Maximum likelihood Estimation (MLE).
- Likelihood function means suppose a fair coin outcome flips equally heads and tails of the same number of times.
- If the is tossed coin 10 times, it is expected to get five times Head and five times Tail.
- Let us consider the probability of getting only Head as an outcome; it is $5/10 = 0.5$
- Whenever this number (P) is greater than 0.5, it is said to be in favour of Head.
- Whenever P is lesser than 0.5, it is said to be against the outcome of getting Head.
- Let us represent 'n' flips of coin as $X_1, X_2, X_3, \dots, X_n$.
- Now X_i can take the value of 1 or 0.
- $X_i = 1$ if Head is the outcome
- $X_i = 0$ if Tail is the outcome
- When Bernoulli distribution used to represents each flip of the coin

$$f(x_i|\theta) = \theta^{x_i}(1 - \theta)^{1-x_i}$$

- Each observation X_i is independent and also identically distributed (iid), and the joint distribution simplifies to a product of distributions.

$$f(x_1, \dots, x_n | \theta) = \prod_{i=1}^n f(x_i | \theta) = \theta^{x_1} (1 - \theta)^{1-x_1} \dots \theta^{x_n} (1 - \theta)^{1-x_n} = \theta^{\#H} (1 - \theta)^{n-\#H},$$

- where #H is the number of flips that resulted in the expected outcome (heads in this case).
- The likelihood equation is

$$L(\theta | x) = \prod_{i=1}^n f(x_i | \theta)$$

MLE is about predicting the value for the parameters that maximizes the likelihood function.

$$\log L(\theta | x) = \sum_{i=1}^n \log f(x_i | \theta)$$